Interspeech 2022
18-22 September 2022, Incheon, Korea

The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition

Luke Prananta[1], *Bence Mark Halpern*[1,2,3], *Siyuan Feng*[1], *Odette Scharenborg*[1]

[1]Multimedia Computing Group, Delft University of Technology, the Netherlands
[2]Netherlands Cancer Institute, Amsterdam, the Netherlands
[3]ACLC, University of Amsterdam, Amsterdam, the Netherlands

o.e.scharenborg@tudelft.nl

Abstract

In this paper, we investigate several existing and a new state-of-the-art generative adversarial network-based (GAN) voice conversion method for enhancing dysarthric speech for improved dysarthric speech recognition. We compare key components of existing methods as part of a rigorous ablation study to find the most effective solution to improve dysarthric speech recognition. We find that straightforward signal processing methods such as stationary noise removal and vocoder-based time stretching lead to dysarthric speech recognition results comparable to those obtained when using state-of-the-art GAN-based voice conversion methods as measured using a phoneme recognition task. Additionally, our proposed solution of a combination of MaskCycleGAN-VC and time stretching is able to improve the phoneme recognition results for certain dysarthric speakers compared to our time stretched baseline.

Index Terms: voice conversion, dysarthric speech recognition, time stretching, generative adversarial networks

1. Introduction

Dysarthria is an encapsulating term for various motor speech disorders in which the muscles that produce speech are weakened or damaged. The reduced motor capabilities of certain speech muscles result in speech that is less intelligible. Dysarthria can greatly reduce a person's quality of life and independence. Operating home appliances through voice could greatly improve these people's lives; however, dysarthric speech recognition performance is not good enough for practical applications [1].

Dysarthric speech recognition is usually tackled from one of two perspectives, namely, data augmentation and dysarthric speech enhancement. The aim of data augmentation is to improve the recognition performance of dysarthric speech by training automatic speech recognition (ASR) models with synthetic dysarthric data. Previous approaches have tried to generate dysarthric data using different neural models including Transformers [2] and generative adversarial networks (GANs) [3].

For dysarthric speech enhancement, voice conversion (VC) has been the dominant approach, although there are previous studies using time stretching, and formant synthesis [4]. VC's goal is to convert a source speaker's speech (here: dysarthric speech) to a target speaker's speech (here: healthy speech), while simultaneously retaining the linguistic content of the utterance. A VC task is either categorised as parallel or non-parallel. For parallel VC, identical utterances are available from the source and the target speakers during training. When parallel utterances are not available, the VC task is called non-parallel. Previous studies of dysarthric VC (DVC) have largely consisted of partial least squares regression- (PLS) [5], Gaussian mixture model (GMM) [4], or deep neural network-based (DNN) [6, 7] parallel methods. There are also some methods that incorporate non-parallel VC methods as part of a parallel VC system [8, 9].

Despite parallel models being able to synthesise highly natural speech using low amounts of data, non-parallel frameworks have higher practical use because it is substantially less difficult to collect non-parallel than parallel data. It is therefore not surprising that non-parallel CycleGAN-based approaches recently attracted some attention for DVC [10]. However, it is as yet unclear what variant of CycleGAN-VC is the most ideal for DVC as [11] found the DiscoGAN and CycleGAN-VC architectures comparable on objective metrics, while Mask-CycleGAN-VC [12] seems to be better than CycleGAN-VC. Our study not only serves to fill this gap in the understanding of CycleGAN-based VC but importantly, also investigates the efficiency of GAN-based VC methods compared to time stretching methods. Previous research has shown that the performance of plain time stretching is comparable to GMM-based parallel voice conversion on the measure of phoneme accuracy [4]; however, it is unclear how it compares to state-of-the-art GAN-based methods.

In this paper, we investigate the efficiency of GAN-based methods for dysarthric-to-normal speech conversion with the aim of enhancing dysarthric speech for improved dysarthric speech recognition, and compare their performance to that when using time stretching. Our research questions are as follows: (**RQ1**:) What aspects of CycleGAN-based non-parallel techniques are essential for enhancing dysarthric speech as indicated by improved dysarthric speech recognition performance (measured using the phoneme error rate; PER)? We investigate state-of-the-art solutions for DVC using CycleGAN-based models in the form of an ablation study. (**RQ2**:) How does the performance of state-of-the-art GAN-based methods for dysarthric-to-normal speech conversion compare to the performance when using time stretching?

2. Methodology

2.1. Dataset and voice conversion setup

We used the UASpeech dataset [17], which contains parallel word recordings of 15 dysarthric speakers and 13 normal control speakers. Each speaker produced 455 utterances. The subjective speech intelligibility of each speaker was judged by 5 non-expert American English native speakers.

For our experiments, we followed the same data and speaker split as used in [10], which is the following: During training, the dysarthric speech from four male speakers (M05, M08, M09 and M10) and four female speakers (F02, F03, F04 and F05) is used as source speech, while the healthy speech from four male healthy control speakers (CM05, CM08, CM09, CM10) and four

10.21437/Interspeech.2022-190

Table 1: *Overview of the key differences between the main GAN-based models tested in this paper, and the newly implemented intermediate models for the ablation study. 2-STEP stands for two-step adversarial loss, DTW stands for dynamic time warping, FIF DA stands for fill in the frame data augmentation.*

	Loss	Vocoder	2-STEP	DTW	FIF DA
CycleGAN-VC [13]	L_1	WORLD [14]	✗	✗	✗
DiscoGAN [10]	L_2	AHOCODER [15]	✗	✓	✗
CycleGAN-VC + DTW	L_1	WORLD [14]	✗	✓	✗
CycleGAN-VC + 2-STEP	L_1	WORLD	✓	✗	✗
CycleGAN-VC + DTW + 2-STEP	L_1	WORLD	✓	✓	✗
MaskCycleGAN-VC [12]	L_1	MelGAN [16]	✓	✗	✓

female healthy control speakers (CF02, CF03, CF04 CF05) is used as target speech. A leave-one-out cross-validation scheme is used for training and evaluating the models. Each model is trained with 1365 utterances from three different speakers, and evaluated with the 455 utterances from the remaining speaker. To be comparable with Purohit's results, we have used a parallel setup.

The speech data was denoised using a Python package called noisereduce [18], which performs stationary noise removal based on the first 0.5 s of each utterances. Please note that this denoising step is by default applied on the UASpeech dataset since December 2020. [10] was written before December 2020 and does not mention the use of a denoising preprocessing step, we therefore believe that the difference in applying the denoising step is the most important difference between our work and [10].

Preliminary experiments showed that the quality of converted utterances are sensitive to the amount of silence in the audio. Therefore, we also performed an energy-based silence trimming (30 dB) using librosa [19]. Moreover, we removed clicks from the audio using a simple heuristic (removing leading and trailing 0.2 s of the signal).

2.2. Experimental design

2.2.1. Ablation study design

In order to answer what aspects of the CycleGAN-based techniques are essential for enhancing dysarthric speech (**RQ1**), we compared the three CycleGAN-based models mentioned in the Introduction: **DiscoGAN** (Section 2.3.2), **CycleGAN-VC** (Section 2.3.1), and **MaskCycleGAN-VC** (Section 2.3.3). All models were trained using the same UASpeech training data. ASR performance was measured in terms of the PER (Section 2.5).

It is important to recognise that the three GAN models are conceptually very similar, and that they consist of highly similar building blocks; however, important differences also exist between these models. Table 1 gives an overview of the key differences between the GAN-based models tested in this work. These differences focus primarily on the use of a two-step adversarial loss (2-STEP), the use of dynamic time warping (DTW), and whether frame data augmentation (FIF DA) is used (explained in Section 2.3.3). In order to pinpoint the importance of these aspects of the GAN models on dysarthric speech recognition performance, we carried out an ablation study for which we created intermediate CycleGAN-based models where we added the 2-STEP loss and/or DTW to the CycleGAN-VC models:

CycleGAN-VC + DTW CycleGAN-VC with dynamic time warping (DTW is explained in Section 2.4)

CycleGAN-VC + 2-STEP CycleGAN-VC with two-step adversarial loss (Two-step adversarial loss is explained in Section 2.3.3)

CycleGAN-VC + DTW + 2-STEP CycleGAN-VC with parallel data, DTW and two-step adversarial loss

Finally, in order to investigate the role of time stretching (**RQ2**), we provided all the models (CycleGAN-VC, CycleGAN-VC + DTW, CycleGAN-VC + 2-STEP, CycleGAN-VC + 2-STEP + DTW, MaskCycleGAN-VC) with time stretched speech input (see Section 2.4), without retraining the VC models. We will denote these models with the shorthand **+ TS**, and we will refer to these as *time stretched models*.

2.3. GAN architectures

2.3.1. CycleGAN-VC

CycleGAN-based VC aims to convert acoustic features from domain $\mathbf{x} \in X$ to domain $\mathbf{y} \in Y$ using a neural network F as a forward-generator ($X \to Y$), and G as the backward-generator ($Y \to X$), and another set of neural networks D_X and D_Y as the discriminators. The generators and the discriminators are optimised with regards to three loss functions: an adversarial, a cycle-consistent, and an identity loss function.

Adversarial loss function: The aim of the adversarial loss function is to incentivise G to fool the discriminator D, while D is optimised to learn the difference between the distribution of samples generated by G (the "fakes", denoted by 0) and the distribution of real samples (the "reals", denoted by 1),

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D, X, Y) = \mathbb{E}_{\boldsymbol{y}\sim p_{\text{data}}(\boldsymbol{y})}[\log(0 - D(\boldsymbol{y}))] \\ + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log(1 - D(G(\boldsymbol{x})))].\end{aligned}$$

Cycle-consistent loss function: The intuition of this loss 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 minimise 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 $G(F(\boldsymbol{y}))$

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

Identity loss function: While the cycle-consistency loss constrains the structure of the mapping, on its own it does not suffice for preserving linguistic information [13]. Therefore an identity-mapping loss is used to improve preservation of linguistic information. This loss was recommended for the original CycleGAN as well, to preserve colour composition between the input and output images.

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

The complete CycleGAN loss is then defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{CycleGAN}}(G, F, D_X, D_Y, X, Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\ + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}(G, F).\end{aligned}$$

Our implementation of the CycleGAN-VC is built using PyTorch[1]. The CycleGAN-VC is used as the baseline model.

We set up our CycleGAN-VC model with a fixed set of hyper-parameters. The model configuration followed the configuration proposed in [13]. The model used the Mel-generalised cepstrum (MCEP) features provided by the the WORLD vocoder [14]. The pitch (F_0) features were log-speaker normalised during the conversion and the aperiodicities (AP) are simply copied.

[1] `https://github.com/karkirowle/cyclegan_pytorch`

2.3.2. DiscoGAN

DiscoGAN differs from CycleGAN-VC in the (1) application of DTW to temporally align the source acoustic features (see Section 2.4), (2) use of the AHOCODER instead of the WORLD vocoder, (3) a modification of the cycle-consistency loss function, called the mean squared error (MSE) cycle-consistency loss.

MSE cycle-consistency loss function: The MSE cycle-consistency loss function mimics the cycle-consistent loss function, except that it uses the L_1 norm instead of the L_2 norm:

$$\mathcal{L}_{\text{cycle}}(G, F) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\|F(G(\boldsymbol{x})) - \boldsymbol{x}\|_2] + \mathbb{E}_{\boldsymbol{y} \sim p_{\text{data}}(\boldsymbol{y})}[\|G(F(\boldsymbol{y})) - \boldsymbol{y}\|_2].$$

2.3.3. MaskCycleGAN-VC

MaskCycleGAN-VC differs from CycleGAN-VC by using (1) a different (MelGAN) vocoder, and therefore also a different feature representation; (2) a two step-step adversarial loss function; and (3) a data augmentation strategy called fill-in-the-frame data augmentation (FAF DA).

Two-step adversarial loss: This loss was introduced to address the over-smoothing caused by the cycle-consistency loss. Additional discriminators D'_X and D'_Y are used for a second adversarial loss for bidirectionally converted features. The loss is defined as follows:

$$\mathcal{L}_{\text{GAN}_2}(G, F, D', X) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log(0 - D'(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log(1 - D'(F(G(\boldsymbol{x}))))].$$

Fill-in-the-frame data augmentation: This data augmentation technique randomly sets a temporal region in the source mel-spectrogram to zero with a binary mask. This random masking conditions the MaskCycleGAN-VC on a secondary task, namely, filling in the missing frames, alongside the original conversion task. We used the implementation and same parameters provided here[2]for the MaskCycleGAN-VC [12] implementation.

2.4. Dynamic time warping and time stretching

DTW implements a time alignment of the MCEP features of the source and the target utterances such that optimal warping paths are founds between these representations. We used the most common L_2 based DTW implementation from librosa [19].

For time stretching, we used the phase vocoder-based method from librosa [19]. Phase vocoding resamples the magnitude short-time Fourier transform of the speech signal by linear interpolation, while simultaneously adjusting for the change in the phase. This results in a magnitude spectrogram with a smaller or larger number of analysis frames, corresponding to a contracted or a stretched speech signal, respectively. In our work, we always adjusted the dysarthric speech to the duration of the target (healthy) speaker's speech.

2.5. Evaluation: Phoneme recognition

The different models in the ablation study were evaluated on a phoneme recognition task. We used a pre-trained, HMM-based Kaldi ASR model with the same specifications as the one used in [10] for phoneme recognition.

The ASR was trained with the TIMIT dataset, which is an English read speech corpus specifically designed for acoustic-phonetic studies [20]. The UASpeech database does not come with phonemic transcriptions. We created these reference phoneme transcription using g2p-en[3], a tool for grapheme-to-phoneme conversion, in order to compute the PER.

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

3. Results and discussion

Table 2 gives an overview of the performance of all models. The PER results are shown for individual speakers separately and averaged over all speakers and blocked by model type (see table caption for an explanation). Note that empty cells refer to results that were not given or specified by [10].

3.1. RQ1: GAN modifications

To investigate what aspects of GAN models are important for enhancing dysarthric speech for improved dysarthric speech recognition, we first compare the results of the CycleGAN-VC baseline to the results of the newly created, intermediate CycleGAN-based models in the DN ∅ TS block of Table 2.

Comparing the row CycleGAN-VC with the other rows in the second block (DN ∅ TS) shows us that the addition of DTW (CycleGAN-VC+DTW) results in 0.5% absolute improvement in the case of the male speakers and 3.0% absolute improvement in the case of female speakers. The better performance of the CycleGAN-VC + DTW indicates that the DTW seems to slightly improve the temporal aspects of the dysarthric speech.

Both intermediate models including the 2-STEP (CycleGAN-VC + 2-STEP + (DTW)) have a performance that is on average worse compared to the CycleGAN-VC model. We hypothesise that the over-smoothing issue that the 2-STEP is supposed to alleviate might be essential for the naturalness of the synthesised speech, but not for enhancing the speech signal for improved dysarthric speech recognition.

The MaskCycleGAN-VC variant, which uses 2-STEP and FIF DA, is the best performing GAN-variant, with a 4.8% absolute improvement in the case of the male speakers, and 10.8% absolute improvement in the case of the female speakers. The better performance of the MaskCycleGAN-VC is likely due to 1) the use of an improved MelGAN vocoder instead of the WORLD vocoder, which also means that higher dimensional mel-spectrogram features are used instead of MCEPs. 2) the fill-in-the-frame data augmentation technique is used, increasing the amount of available data for training. We suspect that these two aspects are jointly responsible for the MaskCycleGAN-VC outperforming the other models but this question is left for future research.

To summarise, we observe that the different GAN architectures lead to relatively similar performances of enhanced dysarthric speech recognition, meaning that the different aspects do not have a major effect on the intelligibility of the converted speech as measured by the PER. The slightly better results for MaskCycleGAN-VC and CycleGAN-VC + DTW suggest that the most important aspects for the success of GAN models are the choice of vocoder, the FAF DA, and the application of DTW.

3.2. RQ2: Effectiveness of time stretching

To examine the effectiveness of time stretching, we first compared the recognition performance of the different models using time stretched speech (DN TS; third block of Table 2) to that of the different models without time stretched speech (DN ∅ TS; second block of Table 2), followed by a within-block comparison

[3] `https://github.com/Kyubyong/g2p`

Table 2: *Overview of the ASR performance in PER for all models for each speaker separately, listed per set of experiments. 'DN' denotes denoised dysarthric speech data is used as input. GT: models with only noise enhancement and no VC that serve as ground truth; DN ∅ TS: models trained on data that is enhanced using VC, no time stretching; DN TS: models trained on data that is enhanced using VC and that use time stretched dysarthric speech as input; P [10]: results taken from [10]. Percentages in parentheses indicate the subjective speech intelligibility taken from the UASpeech database.* ***Bold*** *highlights column-wise the best result for each set of models.*

	Model	M05 (58%)	M08 (93%)	M09 (86%)	M10 (93%)	Average (82.5%)	F02 (29%)	F03 (6%)	F04 (62%)	F05 (95%)	Average (48%)
GT	Control (Healthy)	47.9%	41.3%	51.9%	50.9%	48.0%	51.98%	57.2%	71.5%	46.7%	56.8%
	Dysarthric	96.1%	60.2%	66.1%	64.6%	71.8%	112.2%	89.3%	78.0%	84.7%	91.2%
DN ∅ TS	CycleGAN-VC	110.4%	**69.8%**	72.7%	80.5%	83.3%	131.1%	103.6%	89.3%	100.0%	106.1%
	CycleGAN-VC + DTW	108.7%	74.0%	**72.7%**	76.0%	82.8%	111.0%	105.6%	84.1%	111.2%	103.1%
	CycleGAN-VC + 2-STEP	110.7%	73.1%	74.1%	77.8%	84.0%	136.2%	103.8%	86.8%	100.0%	106.9%
	CycleGAN-VC + 2-STEP + DTW	114.0%	74.0%	77.9%	78.7%	86.2%	132.2%	107.3%	96.0%	103.4%	109.8%
	MaskCycleGAN-VC	**105.7%**	71.4%	74.1%	**62.9%**	**78.5%**	**119.1%**	**97.1%**	**75.5%**	**88.8%**	**95.3%**
DN TS	Dysarthric + TS	73.8%	**64.8%**	**66.4%**	**60.7%**	**66.4%**	80.4%	**76.9%**	72.6%	**63.7%**	73.4%
	CycleGAN-VC + TS	76.8%	67.8%	72.3%	72.9%	72.4%	80.2%	83.3%	79.3%	74.0%	79.2%
	CycleGAN-VC + DTW + TS	75.3%	71.6%	71.3%	70.9%	72.3%	81.2%	87.6%	77.7%	86.6%	83.3%
	CycleGAN-VC + 2-STEP + TS	78.7%	79.3%	77.8%	76.4%	78.1%	80.8%	83.3%	79.2%	75.3%	79.7%
	CycleGAN-VC + 2-STEP + DTW + TS	76.7%	75.6%	78.0%	76.4%	76.7%	81.8%	86.6%	84.5%	77.6%	82.6%
	MaskCycleGAN-VC + TS	**65.1%**	65.2%	71.7%	65.2%	66.8%	**72.4%**	81.0%	**70.6%**	69.0%	**73.2%**
P [10]	Control (Healthy)	-	-	-	-	64.7%	-	-	-	-	65.4%
	Dysarthric	-	-	-	-	77.9%	-	-	-	-	87.1%
	DNN	-	-	-	-	82.9%	-	-	-	-	75.7%
	DiscoGAN	-	-	-	-	**73.3%**	-	-	-	-	**71.1%**

of the models using time stretched speech to each other and to Dysarthric + TS speech.

We observe that all of the time stretched GAN-based models substantially outperform their non-time stretched counterparts, ranging from 5.9% (for CycleGAN-VC + 2-STEP) to above 9.0% (all other models). Moreover, the comparison of the time stretched models shows that the order of performance for the models is similar to the order of the performance of the models that do not use time stretched speech: The MaskCycleGAN-VC still outperforms the CycleGAN-VC baseline, however the improvements introduced by the DTW seem to be either marginal (in the case of males) or specific to certain speakers (in the case of females). We hypothesise (since both DTW and the time stretching aim to improve the temporal aspects of the speech) that after time stretching the DTW becomes unnecessary, which could explain the smaller improvement in recognition performance for DTW for time stretched dysarthric speech.

Surprisingly, we observe that simple time stretching of the dysarthric speech on average outperforms the best performing GAN-model (MaskCycleGAN-VC + TS), with the exception of speakers M05, F02, and F05, which are mid to high severity cases. This suggests that time stretching speech seems to be a better solution for improving dysarthric speech recognition than using purely GAN-based methods for the low and very high severity cases. Nevertheless, GAN-based methods can complement time stretching to improve the spectral structure of the dysarthric speech (**RQ2**) for the mid severity cases.

3.3. Comparison to state-of-the-art and DiscoGAN

The results of our models cannot be directly compared to those reported in [10] and their state-of-the-art DiscoGAN model because they did not apply denoising (see Section 2.1). However, we can compare the ground truth results: The comparison of our ground truth (GT) healthy speech and dysarthric speech (block 1) and Purohit's results (P) (block 4) in Table 2 shows that the denoising of the dysarthric speech overall led to a lower, thus better, average PER for healthy speech and for dysarthric speech from male speakers. We should also point out that our dysarthric male speech result is already better than Purohit's proposed DiscoGAN model.

The CycleGAN-VC - while being conceptually similar to the DiscoGAN - performs worse than DiscoGAN (see block 2). The observed differences are most likely due to the differences in the CycleGAN-VC and DiscoGAN setup (see also Table 1): (1) the AHOCODER [21] vocoder is used in [10], while we use the WORLD vocoder, (2) CycleGAN uses a single cycle-consistency loss while DiscoGAN uses individual reconstruction losses. However, these losses are conceptually similar, therefore we expect (2) to have only a minor influence on the results compared to the vocoder.

4. Conclusions

We investigated several existing and a new state-of-the-art generative adversarial network-based (GAN) voice conversion methods for enhancing dysarthric speech for improved dysarthric speech recognition. Our main finding is that time stretching dysarthric speech to the target (healthy) speaker's speaking rate improves dysarthric speech recognition performance to a level that is comparable, and even outperforms that of, existing state-of-the-art generative adversarial networks. The application of MaskCycleGAN-based voice conversion on time stretched speech yields results that are slightly better than pure time stretching, but only for mid to high severity speakers with dysarthria. The current performance of dysarthric speech enhancement is unfortunately still not good enough for practical use. The enhancement results proved to be highly dependent on the temporal structure of the speech, as demonstrated by the improved performance of the models using time stretching (+ TS) and dynamic time warping (+ DTW). Therefore, future work should focus on sequence-to-sequence based architecture, which have already shown to excel at improving the temporal structure of dysarthric speech for parallel voice conversion [22].

5. Acknowledgements

This project has received funding from the EU's H2020 research and innovation programme under MMSC grant agreement No 766287. The Department of Head and Neck Oncology and surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research.

6. References

[1] 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," in *Proc. Interspeech 2019*, 2019, pp. 3875–3879. [Online]. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2019/morovelazquez19_interspeech.pdf

[2] J. Harvill, D. Issa, M. Hasegawa-Johnson, and C. Yoo, "Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6428–6432.

[3] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6009–6013.

[4] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[5] R. Aihara, T. Takiguchi, and Y. Ariki, "Phoneme-Discriminative Features for Dysarthric Speech Conversion," in *Proc. Interspeech 2017*, 2017, pp. 3374–3378. [Online]. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2017/aihara17_interspeech.pdf

[6] C.-Y. Chen, W.-Z. Zheng, S.-S. Wang, Y. Tsao, P.-C. Li, and Y.-H. Lai, "Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-Based Voice Conversion System," in *Proc. Interspeech 2020*, 2020, pp. 4686–4690. [Online]. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2020/chen20s_interspeech.pdf

[7] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation," in *Proc. Interspeech 2019*, 2019, pp. 4115–4119. [Online]. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2019/biadsy19_interspeech.pdf

[8] W.-C. Huang, K. Kobayashi, Y.-H. Peng, C.-F. Liu, Y. Tsao, H.-M. Wang, and T. Toda, "A Preliminary Study of a Two-Stage Paradigm for Preserving Speaker Identity in Dysarthric Voice Conversion," in *Proc. Interspeech*, 2021, pp. 1329–1333. [Online]. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2021/huang21d_interspeech.pdf

[9] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, "High-intelligibility Speech Synthesis for Dysarthric Speakers with LPCnet-based TTS And CycleVAE-based VC," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7058–7062.

[10] 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," in *2020 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 7 2020, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9179511/

[11] 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.

[12] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5919–5923.

[13] T. Kaneko and H. Kameoka, "Cyclegan-VC: Non-parallel Voice Conversion Using Cycle-consistent Adversarial Networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[14] 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. 99, no. 7, pp. 1877–1884, 2016.

[15] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2013.

[16] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Mel-GAN: Generative Adversarial Networks for Conditional Waveform Synthesis," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. Interspeech 2008*, 2008, pp. 1741–1744. [Online]. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2008/kim08c_interspeech.pdf

[18] T. Sainburg, "timsainb/noisereduce: v1.0," Jun. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139

[19] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[21] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved HNM-based vocoder for statistical synthesizers," in *Proc. Interspeech 2011*, 2011, pp. 1809–1812. [Online]. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2011/erro11_interspeech.pdf

[22] W. Huang, B. M. Halpern, L. P. Violeta, O. Scharenborg, and T. Toda, "Towards identity preserving normal to dysarthric voice conversion," *CoRR*, vol. abs/2110.08213, 2021. [Online]. Available: https://arxiv.org/abs/2110.08213