An Objective Evaluation Framework for Pathological Speech Synthesis

Bence Mark Halpern[1,2,3], *Julian Fritsch*[4,5], *Enno Hermann*[4,5], *Rob van Son*[1,3] *Odette Scharenborg*[2], *Mathew Magimai.-Doss*[4]

[1]University of Amsterdam, Amsterdam, The Netherlands
[2]Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
[3]Netherlands Cancer Institute, Amsterdam, The Netherlands
[4]Idiap Research Institute, Martigny, Switzerland
[5]École polytechnique fédérale de Lausanne (EPFL), Switzerland
Email: {b.halpern,r.v.son}@nki.nl,
{julian.fritsch, enno.hermann, mathew}@idiap.ch, o.e.scharenborg@tudelft.nl

Abstract

The development of pathological speech systems is currently hindered by the lack of a standardised objective evaluation framework. In this work, (1) we utilise existing detection and analysis techniques to propose a general framework for the consistent evaluation of synthetic pathological speech. This framework evaluates the voice quality and the intelligibility aspects of speech and is shown to be complementary using our experiments. (2) Using our proposed evaluation framework, we develop and test a dysarthric voice conversion system (VC) using CycleGAN-VC and a PSOLA-based speech rate modification technique. We show that the developed system is able to synthesise dysarthric speech with different levels of speech intelligibility.

1 Introduction

In recent years, there has been a growing interest in pathological speech processing [1]. It is a highly challenging area, as our understanding of speech is largely limited to "typical" and unimpaired speech. The advances made in detection and analysis of pathological speech are continuously improving our understanding of it. These efforts could be further accelerated with pathological speech synthesis - as speech processing has advanced through the synergy between "analysis" and "synthesis". For instance, if one considers the message component in the speech signal, automatic speech recognition (ASR) can be regarded as analysis and text-to-speech synthesis (TTS) can be regarded as synthesis. Such understandings have benefited both ASR and TTS and have driven innovation, e.g., [2–5].

Another motivation for the development of pathological speech synthesis is that it could assist in informed decision making for the medical conditions at the root of the pathological speech. For instance, oral cancer surgery results in changes to a speaker's voice. Availability of a synthesis model that can generate how the voice could sound after surgery could help the patients and clinicians to make informed decisions about the surgery and alleviate stress of the patients [6].

Pathological speech synthesis is a non-trivial task due to two main challenges. Firstly, state-of-the-art methods for TTS/voice conversion (VC) often rely upon the availability of a large amount of speech data from many different speakers and linguistic resources. In that sense, pathological TTS/VC is inherently under-resourced, as the collection of pathological speech is time-consuming and is governed by medical research ethical aspects.

Secondly, a crucial aspects in the development of TTS/VC systems is the evaluation of the naturalness of the synthesised speech. This is typically done by asking a group of naive human listeners to rate the synthesised speech on its naturalness [7]. The most common example of such an evaluation by naive listeners is the mean opinion score (MOS). MOS is a good evaluation measure for TTS of typical speech where the only deterioration in the speech is due to the decrease in naturalness. However, it is not appropriate for pathological speech, for several reasons. Firstly, a pathological VC system that receives better MOS scores than the reference pathological speech is likely not able to capture the characteristics of the pathological speech. Secondly, on the reverse, a VC system mimicking the pathology, albeit in an exaggerated manner, would likely produce a MOS score that is a lot lower than that of the reference. Thirdly, it is unknown whether non-expert listeners have sufficient expertise to evaluate pathological speech. Fourthly, involving speech-language pathologists for evaluation at every stage of development would make the evaluation process not only costly but it is also not certain that they can distinguish the unnaturalness of the synthesised speech from the characteristics of the pathology.

To alleviate this problem, the present paper

(a) proposes a new framework, based on the synthesis-analysis approach for ASR, for the consistent evaluation of limited-resource, synthesised pathological speech based on a combination of pathological speech detection and analysis methods.

(b) and, to that end, investigates a CycleGAN-based approach to convert healthy control speech to pathological speech of different levels of intelligibility.

Specifically, in this paper, we focus on the synthesis of dysarthric speech and its evaluation through pathological speech detection (SD) and automatic intelligibility assessment. We demonstrate the viability of the proposed framework through an investigation on the UASpeech corpus.

The paper is organised as follows: the UASpeech database used for training and evaluation is described in Section 2. The employed CycleGAN-VC method is described in Section 3. We present the proposed objective evaluation framework in Section 4 and the evaluation results in Section 5. Speech samples and a demo of the system is available online.[1]

2 Database

For all experiments, we use the UASpeech database [8], which contains 14 dysarthric speakers, 10 male and 4 female, and 455 unique isolated word recordings for each speaker. The speakers have received an intelligibility rating in the database (high, mid, low, very low) based on their average scores in a word transcription task rated by five non-expert listeners (subjective intelligibility score). In our experiments, low and very low intelligibility speakers are grouped together to increase statistical power. In addition, the database contains 9 male and 4 female unimpaired, control speakers. All recordings were done with multiple microphones and repetitions, however, only microphone 5 and the first repetition of each word was used in our experiments. We refer to the healthy speakers' utterances from the database as ground truth (GT) healthy, and the dysarthric utterances as ground truth dysarthric from now on.
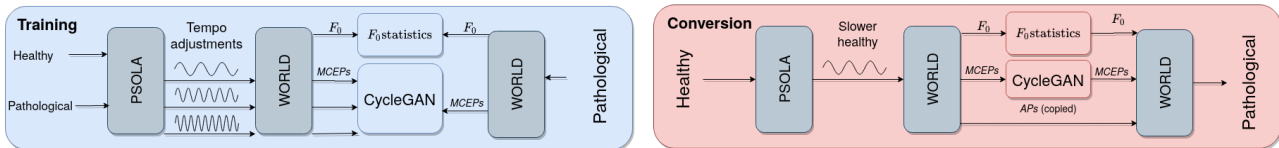
---

[1]https://karkirowle.github.io/publication/itg-2021-dysarthric-vc

Figure 1: Training (blue) and conversion (red) setup for dysarthric voice conversion.

3 Proposed conversion system

We train a voice conversion model for each dysarthric speaker in the database, using CM08 as a male source speaker and CF04 as a female source speaker. This means that – contrary to traditional VC – we are not only converting speaker traits, we are also aiming to change the intelligibility of the speech. The overall experimental setup is illustrated in Figure 1. In the following sections, we will explain the techniques used in detail.

3.1 Speech rate modification and analysis

VC systems (including CycleGAN-VC) do not change the speech rate, but for pathological speech applications, it is essential that VC systems convert this aspect of the speech: e.g., in the case of dysarthric speech, speech rate has been shown to be a reliable predictor of speech severity [9]. Therefore, prior to conversion, the tempo of the source speech is adjusted to the dysarthric speaker's tempo using the pitch synchronous overlap-add algorithm (PSOLA) [10] from Praat [11]. We've run preliminary tempo modification experiments with WSOLA (`sox` implementation) and phase vocoding (librosa [12] implementation) techniques too, but we found that PSOLA results in the most natural sounding speech. We would like to emphasise that, even though the target speaker's tempo is not generally known beforehand, for all practical purposes, the expected speech rate can be estimated if the severity is known. Such estimation has been already done in various works, such as [9, 13, 14].

Furthermore, a tempo-based data augmentation scheme is used to increase the dataset size in the CycleGAN-VC training, which means that the same speech signal is fed into the neural networks three times with three different speech rates, modified using the pitch synchronous overlap-add algorithm from Praat. The first version is the original unadjusted version of the speech signal, spoken at the rate of the control speaker. The second version is adjusted to the speech rate of the dysarthric speaker. The last version is adjusted to a rate exactly halfway in between the original unadjusted and the dysarthric adjusted. This tempo-based data augmentation scheme increases the size of the training set three-fold. We refer to the dataset obtained after the augmentation as the augmented database, and the speech adjusted to dysarthric speech rate as adjusted speech. Subsequently, the augmented database is preprocessed using the WORLD vocoder [15] to obtain the Mel-generalised cepstrum (MCEP), the pitch (F_0) and the aperiodicities (AP).

3.2 CycleGAN-VC training and conversion

Previous studies have found that CycleGAN-VC-based systems can improve the speech intelligibility of a dysarthric signal [16, 17]. For this reason, we are interested if a CycleGAN-VC based system can also model the deterioration of speech intelligibility too. An additional benefit of using the CycleGAN-VC is that the system is known to work with small amount (approximately 10 mins) of data, which can be non-parallel. This makes CycleGAN-VC lucrative for pathological speech applications which are inherently low-resourced.

A CycleGAN-VC system is trained with the augmented database [18][2] with the same hyperparameters and values (as

---

[2]The implementation used was `https://github.com/leimao/Voice_Converter_CycleGAN`

most hyperparameters are directly adapted from the original CycleGAN paper) for 1000 epochs [19]. During conversion of the healthy speech to dysarthric speech, only the healthy tempo-adjusted speech is used. The MCEPs are then converted using the trained model. The APs are copied. During training, the mean and standard deviation of F_0 values are calculated for both speakers (source and target). The source F_0 is mean-std normalised to the source speaker's F_0 distribution and mean-std unnormalised to the target speaker's F_0 distribution.

4 Objective evaluation framework

Pathological speech deviates from healthy speech along several different dimensions. Therefore, the proposed system has four parts, each of which evaluates the pathological speech signal generated by the VC from a different angle, together yielding a consistent evaluation measure: a pathological speech detector (Section 4.1), a voice quality measure (Section 4.2), an intelligibility assessment tool (Section 4.3) and an ASR system (Section 4.4).

Before the evaluation experiments, voice activity detection (VAD) is performed using Praat to avoid analysis of non-speech cues [11]. This VAD is manually checked for quality. Ideally, VAD would be done before training, however, our experiments found that in this case the CycleGAN-VC does not converge well due to the short duration of the utterances.

4.1 Voice quality measure: LTAS-LASSO-SD

Long time average spectrum (LTAS) has been successfully used as a voice quality measure [20] and predictor for various speech pathologies [21, 22]. Thus, an LTAS based pathological speech detector (SD) will be trained on detecting dysarthric speech using the voice quality cues provided by the LTAS-feature, assessed for generalisation using the GT samples, and then used to evaluate our converted (VC) samples.

The leave-one-out validation scheme used for the LTAS-LASSO-SD is illustrated in Figure 2. The main factors considered during the design of this validation scheme were that (1) the speaker who is evaluated should not be included in the training data (2) each speaker should be mapped with a single control speaker so that everyone is only left out during its own evaluation round and that the number of control and pathological speakers remain changed. Two male speakers had no control pairs, so we excluded one with very high intelligbility (M10) and one with very low (M01). The weakness of this experiment design is that male speakers are overrepresented in the dataset. However, the alternative would be one single speaker per gender per severity, which we found an inferior design.

First, an LTAS is extracted based on a 512-point FFT with Hann window and 128 sample frame shift with librosa [12] for the control, VC and original dysarthric speech.

The chosen detector model was a LASSO due to its automatic feature selection property. The sparsity penalty was $\alpha = 10^{-5}$. The model was trained for a maximum of 1000 iterations. The VC samples are then evaluated on these detector models in terms of accuracy. Furthermore, we investigate to what extent voice quality is influenced by the subjective intelligibility of utterances (taken from the UASpeech database) by performing a Pearson's correlation (r) between the subjective intelligibility scores and mean detection scores of the VC utterances.
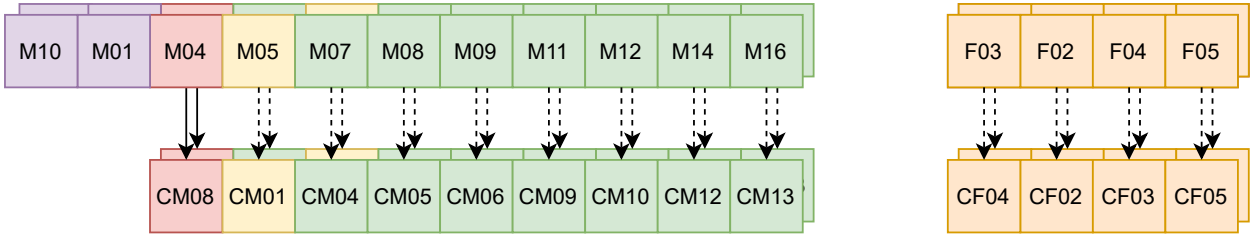
Figure 2: Leave-one-out validation scheme for LTAS-LASSO-SD. Purple colour shows the excluded speakers. The conversion pair (red) is excluded from the evaluation, and in each fold a different speaker pair (yellow) is left out. The green colour shows the speakers that we are training on. This is repeated for the two genders separately.

4.2 Intelligibility decrease measure: LTAS-SKL measure

The idea of the LTAS-SKL measure is the following: each word of a reference speaker is compared to the same word of all other speakers (control and pathological; VC and GT). The comparison is done on the distribution of LTAS energy bins using the symmetric Kullback-Leibler (SKL) divergence. The speakers are then grouped based on intelligibility and the results visualised using a box plot. This way we ensure that the difference in the utterances can only be due to the individual differences in speakers (spread) and due to the differences in intelligibility of the healthy ground-truth speech and the generated pathological speech (median). Here, a normalised LTAS with a 1024-point FFT with Hann window and 256 sample frame shift is extracted. A t-test is used to check the significance of the difference in intelligibility.

4.3 PPG-DTW-based utterance verification

The phonetic posteriorgram (PPG)-dynamic time warping (DTW) method [23] evaluates pathological utterances by matching them to a healthy reference utterance of the same word and aligning their PPG sequences with DTW. To verify an utterance, the DTW matching score is converted into a probability P_c, by plugging it into a logistic function, c denoting the control class; an utterance is hence verified correctly if $P_c \geq 0.5$. A speaker's intelligibility is then estimated in terms of the percentage correctly verified utterances, just as in human listening tests.

For our experiments, we used an SKL local cost function in the DTW implementation and posterior sequences of 45 context-independent phones (PPG). Spearman's ρ and Pearson's correlation r analyses between the intelligibility estimates of the dysarthric speech GT and the VC speech, show that the method performs at $\rho = .965$, $r = .854$, which is comparable to other state-of-the-art methods such as [24]. The evaluation is repeated on the WORLD vocoder resynthesised GT dysarthric signals to see if vocoding affects the intelligibility estimates.

4.4 ASR based intelligibility evaluation

As an additional objective intelligibility measure, we use the word error rate (WER) of an ASR system trained only on unimpaired control speakers. We examine whether the VC speech gives comparable results to the GT dysarthric data to show that the VC system learns to model the lower intelligibility of dysarthric speech.

We trained acoustic models with the Kaldi ASR toolkit [25] on block 2 and 3 of the UASpeech corpus, which contain distinct sets of words from block 1 that was used for VC. We trained the acoustic models on the data from all microphones and all control speakers except the 2 source speakers for VC (CM08, CF04), based on an existing Kaldi recipe [26] and decode with a unigram grammar that includes all words in the UASpeech corpus. To verify that the results are not specific to a single type of model, we compare both subspace Gaussian mixture models (SGMMs) [27] and neural networks trained with the sequence-discriminative lattice-free maximum mutual information (LF-MMI) loss [28], the state-of-the-art methods in the GMM and hybrid hidden Markov model/deep neural network ASR paradigms, respectively.

5 Results and discussion

5.1 Voice quality measure: LTAS-LASSO-SD

The LTAS-LASSO-SD detectors perform well on the held-out GT speakers, with the best held-out speaker having an accuracy of $93.15\% \pm 9.75$ (M14), the worst held-out speaker performing with a $90.84\% \pm 10.1$ (M05). This means that the proposed evaluation framework is able to detect the voice quality of dysarthric speech. On the VC samples the best held-out speaker accuracy is $100\% \pm 0$ (multiple speakers, i.e., M16, F02, F03, F04, F05) and the worst accuracy is $92.07\% \pm 6.05$ (M11). This means that the proposed VC system successfully mimics the voice quality aspects of dysarthric speech.

There is a linear relationship between the male speakers' detection scores and the subjective intelligibility scores from UASpeech (male; $r = -0.85$, p-value $= 0.003$), which shows that the intelligibility aspects of the synthesised dysarthric speech are closely related to the voice quality aspects in case of the male speakers. This relationship does not exist in the case of females (female; $r = -0.02$, p-value $= 0.98$). This means, while the voice quality aspects can be heavily influenced by the intelligibility of the speakers in the case of male speakers, the proposed voice quality measure should be used in complement to the other, intelligibility-based measures.

5.2 Intelligibility decrease measure: LTAS-SKL

In Figure 3, it can be seen that both for the GT (gray) and the VC (red) speakers the median of the LTAS increases (SKL-LTAS dB) as the intelligibility decreases. This increase in the median is significant, except for mid-low VC male and mid-low GT female. This means that VC performs well overall, however it is not that well able to generate dysarthric speech that differentiates between mid and low intelligibility.

5.3 PPG-DTW-based utterance verification

Table 5.3 shows that the correlations between the subjective and objective intelligibility scores are high, and are significant for all male speakers, meaning that the proposed VC model successfully captures characteristics of dysarthric speech that influence intelligibility for male speakers. The results also show high correlation for female speakers, though not significant, which might be due to the lower number of female speakers. Additionally, it can be seen that overall there is a correlation gap between the WORLD all (copy synthesis) and VC all (conversion). This indicates that conversion performance is not limited by the vocoder, but rather by the used model.
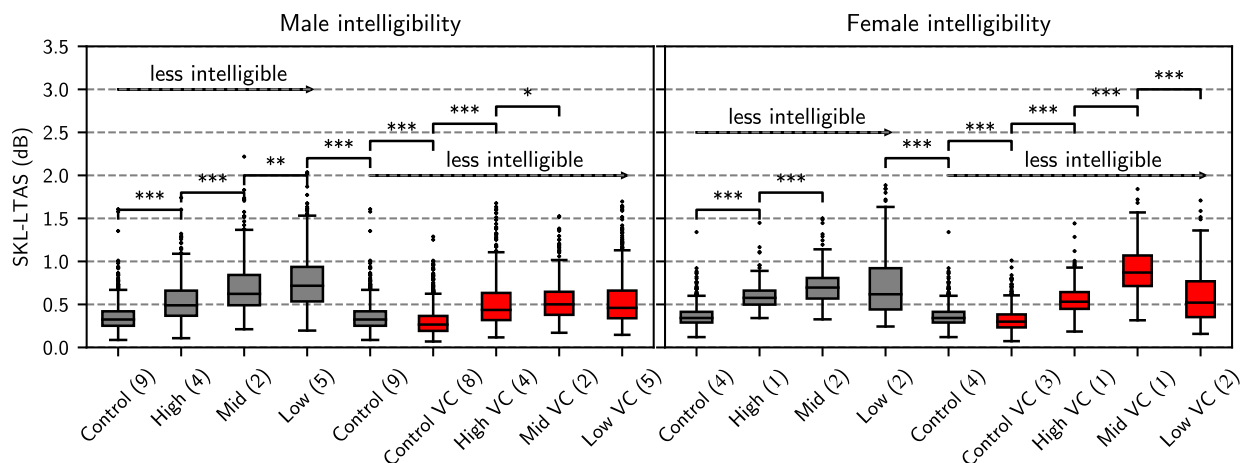
Figure 3: Result of the LTAS-SKL analysis. A grey interquartile range (IQR) indicates original samples, a red IQR indicates VC samples. The center line indicates the median. (***) $p < .001$ (**) $p < .01$ (*) $p < .05$, otherwise not significant. Number of speakers are written in parentheses. The ground truth control speakers of are plotted twice on each panel for easier comparison.

Data	ρ	p-value	r	p-value
WORLD all	.954	***	.836	***
VC all	.448	*	.481	*
WORLD male	.940	***	.807	**
VC male	.735	**	.594	*
WORLD female	1.0	***	.957	*
VC female	.774	0.229	.720	0.279

Table 1: Pearson's and Spearman's correlation between subjective intelligibility score (from UASpeech) and predicted intelligibility by PPG-DTW. (*) $p < .1$ (**) $p < .05$.
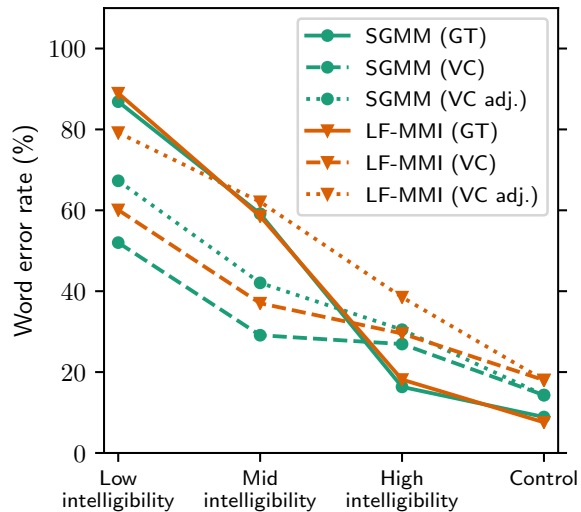

Figure 4: WER results grouped by intelligibility of the target speakers, comparing GT and VC speech using original and adjusted tempo (VC adj).

5.4 ASR-based intelligibility evaluation

Figure 4 shows the WER obtained with the two ASR systems for the low to highly intelligible dysarthric and the control speakers. As expected, the GT speech data (solid lines) exhibits a steep drop in error rates as intelligibility increases. The VC speech mirrors this decrease. When the speech rate is adjusted (dotted lines), the WERs resemble the ground truth more closely. This is consistent with previous work that has shown slower speaking rates of dysarthric speech to harm ASR performance [29]. However, the same pattern of fewer errors when intelligibility is higher is still apparent without tempo adjustment (dashed lines). This shows that the VC system learns to manipulate the speech in such a way that makes it harder for an ASR system trained only on control speech to recognise it correctly. Both when synthesising healty and dysarthric speech, we expect the VC speech to perform slightly worse than the GT because of the added noise introduced by the VC, which does not match the acoustic conditions of the ASR training data. We observe the same broad pattern for SGMM (green) and LF-MMI (orange) models, indicating that our analysis generalises across different types of acoustic models.

6 Conclusion

In this paper, we proposed an objective evaluation framework for pathological speech synthesis. The proposed framework (1) captures the voice quality aspects of pathological speech, and consistently shows the decrease in intelligibility of synthesised dysarthric speech compared to healthy speech using the (2) LTAS-SKL (3) PPG-DTW (4) and an ASR-based methods. Simultaneously, we have developed and evaluated a proof-of-concept CycleGAN-VC system which (1) mimics the voice quality of dysarthric speech, (2,3,4) can exhibit different levels of speech intelligibility (4) and shows that this decrease in intelligibility is independent of the decrease in speech rate.

7 Acknowledgements

References

[1] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone, and S. Narayanan, "Pathological speech processing: State-of-the-art, current challenges, and future directions," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6470–6474, IEEE, 2016.

[2] K. Tokuda, T. Masuko, J. Hiroi, T. Kobayashi, and T. Kitamura, "A very low bit rate speech coder using hmm-based speech recognition/synthesis techniques," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[3] K. Lee and R. V. Cox, "A very low bit rate speech coder based on a recognition/synthesis paradigm," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 482–491, 2001.

[4] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between hmm-based asr and tts," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1046–1058, 2010.

[5] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 301–308, IEEE, 2017.

[6] J. B. Epstein, S. Emerton, D. A. Kolbinson, N. D. Le, N. Phillips, P. Stevenson-Moore, and D. Osoba, "Quality of life and oral function following radiotherapy for head and neck cancer.," *Head Neck*, 1999.

[7] Z. Wu, Z. Xie, and S. King, "The blizzard challenge 2019," in *Proc. Blizzard Challenge workshop*, vol. 2019, 2019.

[8] 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 *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[9] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition.," in *Interspeech*, pp. 471–475, 2018.

[10] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[11] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2011.

[12] B. McFee, V. Lostanlen, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, J. Mason, D. Ellis, E. Battenberg, S. Seyfarth, R. Yamamoto, K. Choi, V. et. Reevich Morozov, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, and T. Kim, "librosa/librosa: 0.8.0," July 2020.

[13] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5836–5840, IEEE, 2019.

[14] C. Bhat, B. Vachhani, and S. Kopparapu, "Improving recognition of dysarthric speech using severity based tempo adaptation," in *Speech and Computer* (A. Ronzhin, R. Potapova, and G. Németh, eds.), (Cham), pp. 370–377, Springer International Publishing, 2016.

[15] 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.

[16] S. H. Yang and M. Chung, "Improving dysarthric speech intelligibility using cycle-consistent adversarial training," *arXiv preprint arXiv:2001.04260*, 2020.

[17] 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)*, pp. 1–5, IEEE, 2020.

[18] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2100–2104, 2018.

[19] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[20] S. Master, N. d. Biase, V. Pedrosa, and B. M. Chiari, "The long-term average spectrum in research and in the clinical practice of speech therapists," *Pró-Fono Revista de Atualização Científica*, vol. 18, no. 1, pp. 111–120, 2006.

[21] B. M. Halpern, R. van Son, M. van den Brekel, and O. Scharenborg, "Detecting and Analysing Spontaneous Oral Cancer Speech in the Wild," in *Proc. Interspeech 2020*, pp. 4826–4830, 2020.

[22] L. K. Smith and A. M. Goberman, "Long-time average spectrum in individuals with parkinson disease," *NeuroRehabilitation*, vol. 35, no. 1, pp. 77–88, 2014.

[23] J. Fritsch, G. Quer, and M. Magimai.-Doss, "Probabilistic symbol sequence matching and its application to pathological speech intelligibility assessment," Idiap-RR Idiap-RR-01-2021, Idiap, 1 2021.

[24] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6405–6409, IEEE, 2019.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," technical Report, 2011.

[26] Anonymous, "Handling Acoustic- and Pronunciation-Level Mismatch in Dysarthric Speech Recognition Systems through Model Combination," *In submission*, 2020.

[27] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model - A structured model for speech recognition," *Computer Speech & Language*, vol. 25, pp. 404–439, 2010.

[28] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Proc. Interspeech*, pp. 2751–2755, 2016.

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