Interspeech 2022
18-22 September 2022, Incheon, Korea

Improved ASR Performance for Dysarthric Speech Using Two-stage Data Augmentation

Chitralekha Bhat[1,2], *Ashish Panda*[1], *Helmer Strik*[2,3,4]

[1]TCS Research and Innovation, India
[2]Centre for Language and Speech Technology (CLST), Radboud University Nijmegen, The Netherlands
[3]Centre for Language Studies (CLS), Radboud University Nijmegen, The Netherlands
[4]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, The Netherlands

bhat.chitralekha@tcs.com,ashish.panda@tcs.com,helmer.strik@ru.nl

Abstract

Machine learning (ML) and Deep Neural Networks (DNN) have greatly aided the problem of Automatic Speech Recognition (ASR). However, accurate ASR for dysarthric speech remains a serious challenge. Dearth of usable data remains a problem in applying ML and DNN techniques for dysarthric speech recognition. In the current research, we address this challenge using a novel two-stage data augmentation scheme, a combination of static and dynamic data augmentation techniques that are designed by leveraging an understanding of the characteristics of dysarthric speech. Deep Autoencoder (DAE)-based healthy speech modification and various perturbations comprise static augmentations, whereas SpecAugment techniques modified to specifically augment dysarthric speech comprise the dynamic data augmentation. The objective of this work is to improve the ASR performance for dysarthric speech using the two-stage data augmentation scheme. An end-to-end ASR using a Transformer acoustic model is used to evaluate the data augmentation scheme on speech from the UA dysarthric speech corpus. We achieve an absolute improvement of **16%** in word error rate (WER) over a baseline with no augmentation, with a final WER of **20.6%**.

Index Terms: Dysarthric speech, ASR, Static and dynamic data augmentation, SpecAugment

1. Introduction

Dysarthria is a motor-speech disorder caused by either developmental or acquired health conditions such as cerebral palsy, Parkisons disease or head trauma. Performance of ASR systems and personal assistants has made great strides owing to the recent Machine learning (ML) and Deep Neural Networks (DNN) techniques, albeit this is not the case with the atypical dysarthric speech due to the inter-speaker and intra-speaker inconsistencies in the acoustic space as well as the sparseness of data. In order to capitalize on the current research on ML techniques for ASRs, such as the End-to-End (E2E) ASR systems, suitable and abundant data to build these systems is imperative. However, collection of dysarthric data is tedious, especially for speakers with severe dysarthria, on account of speaker muscle weakness and fatigue. Data augmentation policies designed specifically for dysarthric speech can act as a key factor in improving dyarthric ASR with limited intrusion on the speakers for data collection.

Data augmentation is a common approach employed to increase the amount of training data, especially for DNN training in order to avoid overfitting while generating robust DNN models. Several types of data augmentation techniques have been employed to vastly increase the quantity of matching training data [1, 2, 3], especially in atypical scenarios such as reverberant speech [4], child speech [5] and dysarthric speech [6, 7, 8]. Simple audio-level augmentations such as speed, pitch, tempo and volume perturbations applied directly on raw speech to increase the training data multi-fold [1, 5] have proven to be extremely effective. SpecAugment [9] and data augmentation techniques inspired by SpecAugment such as frame level SpecAugment [10] and usage of semantic masks [11] have been successful in improving ASR performance through dynamic data augmentation.

Research on data augmentation in the context of dysarthric speech is sparse since this involves a clear understanding of dysarthric speech patterns. Time and tempo-stretching of healthy speech-based data augmenation for improving speech recognition has been investigated in [8]. Two separate augmentation policies involving speed, tempo and vocal tract length perturbation (VTLP) applied on healthy and dysarthric speech showed significant improvement in the ASR performance [6]. A Deep convolution based generative adversarial network (DCGAN) was used for tempo and speed perturbations in addition to learning hidden unit contributions (LHUC) based speaker adaptation in [12]. In order to address ASR performance on severe dysarthric speech, speaker-dependent acoustic models based on phoneme-level speech tempo ratio between typical and speaker-specific dysarthric speech have been created to augment existing dysarthric speech [13]. A transformation of healthy speech to dysarthric speech using voice conversion-based techniques involving speaking rate modification, pitch modification and spectral feature transformation using adversarial training, have been employed to simulate training data using healthy speech [14]. Visual data augmentation techniques are applied on speech features that are extracted visually [15].

The objective of this work is to achieve improved ASR performance in terms of word error rate (WER) by using a two-stage data augmentation process that involves a static and dynamic augmentation of dysarthric speech data. We refer to the process of data augmentation prior to DNN training as static and data augmentation as a part of the DNN training as dynamic. Since the dynamic augmentation is done in real-time, the DNN system benefits from viewing diverse data at every epoch, which in turn strengthens the training process. Our main contribution is the use of DAE and SpecAugment [9] in novel

10.21437/Interspeech.2022-10335

ways to achieve dysarthric speech data augmentation. We introduce two new concepts in this paper, which improve the peformance of ASR for dysarthric speech. First, we train the DAE with dysarthric speech and use the bottleneck layer to generate dysarthric speech using healthy speech. Since this a reversal of how the DAE traditionally works, we refer to it as reverse DAE (R-DAE). Second, we leverage our knowledge of dysarthic speech to design a dynamic data augmentation method akin to SpecAugment, which we call Dysarthric SpecAugment (DSA) scheme. In DSA, we dynamically introduce breathiness, stutter and hypernasality to healthy speech, thereby increasing the diversity of the training data. We analyze the performance of an End-to-End Transformer-based ASR in terms of WER for scenarios with no dysarthric data being available for training as well as when some dysarthric speech data is available. We use the ESPnet toolkit [16] for all our ASR experiments. ASR performance is evaluated on UA dysarthric speech corpus.

We describe static and dynamic data augmentation techniques in section 2.1. We discuss the experimental set-up in section 3 and present the contributions of various combinations of data augmentation techniques on the WER in section 4. Finally we conclude with our observations and recommendations in section 5.

2. Two-stage Data Augmentation

Two-stage data augmentation involves data augmentation done in two steps. First, static data augmentation (SDA) techniques were applied to both healthy and dysarthric speech as described in section 2.1. The SDA data was subsequently subjected to dynamic augmentation. Dysarthric SpecAugment (DSA) was applied on healthy speech, SpecAugment was applied on dysarthric and R-DAE generated speech. Thus, robust ASR models for dysarthic speech recognition were trained using (1) SDA and DSA augmented healthy speech, (2) Speed and SpecAugment dysarthric speech and (3) Speed and SpecAugment R-DAE speech.

2.1. Static Data Augmentation

The data augmentation techniques applied prior to DNN training augmented the dysarthric speech data in a static manner. We have used two different types of static augmentation as described in the rest of this section.

2.1.1. Speed, tempo and volume perturbation (SDA)

Speed perturbation is a recommended method for data augmentation since it has been known to improve speech recognition as well as has a low implementation cost [1]. However, it is to be noted that speed perturbation affects both pitch and tempo of the original speech since it involves resampling of the original speech signal. In the current work, we have applied speed modifications to both healthy speech as well as dysarthric speech to provide three different versions, resulting in data augmentation by a factor of three.

Relationship between the tempo of healthy speech and dysarthric speech has been investigated and leveraged to improve the dysarthric speech recognition [17, 18, 6]. Typically dysarthric speech is slow and slurred, indicating a slower tempo as compared to healthy speech. We have applied three different tempo modifications on healthy speech to match the severity levels in the dysarthric speech corpus. Tempo perturbation does not alter the pitch of the speech being modified.

Training dysarthic speakers to increase loudness of speech in order to improve intelligibility, is a known therapy technique which results in a higher articulatory-acoustic working space as well as improved acoustic contrast for dysarthric speakers[19]. In order to match the characteristics of dysarthric speech, we have applied loudness modifications to healthy speech by reducing the loudness. Two different loudness factors have been used to generate two distinct versions of the healthy speech.

2.1.2. Reverse Deep Autoencoder (R-DAE)

Deep Denoising Autoencoders are traditionally used for improvement of speech recognition of noisy speech, by enhancing the noisy speech to match clean speech [20]. DAE-based dysarthric speech enhancement has been used as a tool for dysarthric speech enhancement, thereby improving the ASR performance for dysarthric speech[21, 22]. The purpose of the current research is to generate data to augment dysarthric speech, which is the reverse of enhancement. Hence we call this static data augmentation as reverse-DAE (R-DAE). We have used Time Delay Neural Network (TDNN)-DAE in an unconventional way, wherein dysarthric speech was used to train the bottleneck layer of the TDNN-DAE. Healthy speech was then used to generate speech akin to dysarthric speech. However, this technique is dependent on the availability of dyasrthric speech data. The configuration of the R-DAE is as described in [22].

2.2. Dynamic Data Augmentation

2.2.1. SpecAugment

SpecAugment is a data augmentation technique that is directly applied on the spectral speech features used for DNN training. The augmentation policy is designed to build a robust ASR by allowing for prediction of changes to data in the time direction, partial loss of information in the frequency direction, as well as due to loss of small segments of speech [9]. Towards this end, masks are constructed to dynamically mask or modify the information in the time and frequency directions. The width and location of the masks are determined randomly, ensuring that the DNN is exposed to a different version of the input speech at every epoch of the training process.

2.2.2. Dysarthric SpecAugment (DSA)

We have leveraged our understanding of dysarthric speech to design three DSA policies specific to dysarthric speech. These masks have been applied only on healthy speech in a manner similar to the SpecAugment process described above.

A. **Stutter mask**
   Stuttering is a speech disorder that manifests as either the arrested articulation of a syllable or clonic repetition of the same syllable [23]. Dysarthria is a motor-speech disorder and the resulting speech has some of the characteristics of a stutter.

   A mask was constructed along the time direction, wherein random and small segments of speech were repeated to emulate stuttering. Stutter mask was applied so that t consecutive time steps [$t0$, $t0 + t$) were repeated, where the mask width t was randomly chosen from a uniform distribution such that $t \sim U(0,T)$ where T is the stutter mask parameter and $t0$ is chosen from [0, $\tau - t$), where τ is the length of the utterance.

B. **Hypernasal mask**
   Hypernasality is a consequence of velopharyngeal dysfunction (VPD) or velopharyngeal incompetence (VPI)

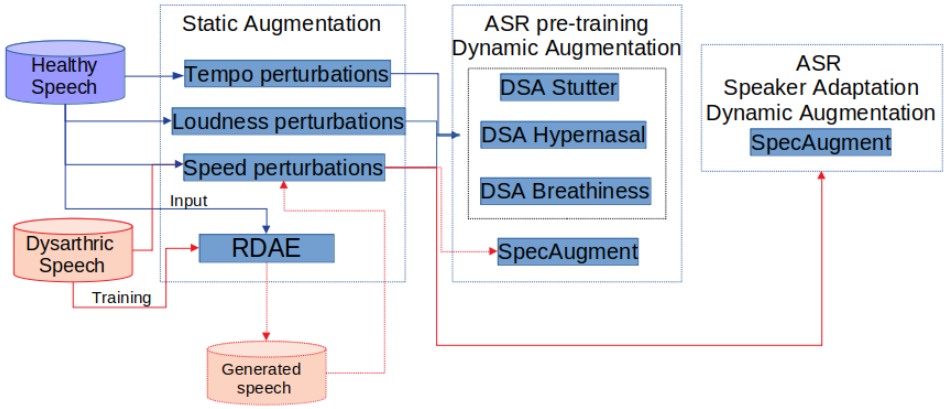

Figure 1: *Two-stage data augmentation for dysarthric speech recognition; continuous blue line indicates path for healthy speech, continuous red line for dysarthric speech and dashed red line for R-DAE generated data*

which manifests as excessive nasal resonance in speech. It is the outcome of improper closure of the soft palate, that regulates the airflow between the oral and nasal cavities. Hypernasality is a common occurrence in motor-speech disorders such as dysarthria [24]. Hypernasal speech exhibits significantly higher energy level at frequency bands centred at 630, 800 and 1000 Hz, and significantly lower amplitude for the band centred at 2500 Hz as compared to healthy speech [25, 26].

In order to simulate hypernasality in healthy speech, the SpecAugment was modified along the frequency direction. Channels of the Mel filter-bank corresponding to the frequency regions from 600 to 1600 Hz were identified as the first region and the second region corresponded to the frequencies around 2500 Hz. The energy level in f consecutive mel frequency channels around the first region was increased by three times and the amplitude in the second region was reduced by half. f consecutive Mel frequency channels for modification $[f0, f0 + f)$ were randomly chosen from a uniform distribution such that $f \sim U(0,F)$ where F is the hypernasal mask parameter and $f0$ is chosen from $[\nu1, \nu2 - f)$, where $\nu1$ and $\nu2$ correspond to the first and last Mel channels of the corresponding region.

C. **Breathiness (Noise) mask**

Dysarthria is often associated with disturbances of respiration, laryngeal function, airflow direction, and articulation resulting in breathy speech quality and reduced intelligibility. Breathiness is typically caused by glottal air leakage and acoustic measures related to breathiness are often used to distinguish between different physiological phonation conditions for pathological speech [27].

A scaled white noise mask was applied to healthy speech in order to replicate the presence of breathiness along both the time and frequency directions. The initial point and the width of the mask was chosen randomly as described in the stutter and hypernasal sections.

3. Experimental setup

3.1. Data

Data from Universal Access (UA) speech corpus [28] was used for training the End-to-End(E2E) Transformer-based ASR as well the R-DAE. UA dysarthric speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. Data was collected in three separate sessions for each speaker and categorized into three blocks B1, B2 and B3. The speech material contains 155 words that are common to all three blocks and 100 words that are distinct for each block. Blocks B1, B2 and B3 from healthy speakers and blocks B1 and B3 from dysarthric speakers were treated as training set and block B2 from dyasrthric speakers was treated as test set. We have not included uncommon words in the test set-up. The corpus also includes speech intelligibility rating for each dysarthric speaker, as assessed by five naive listeners.

ESPnet toolkit [16] was used as the E2E-Transformer based system to evaluate the ASR performance for data augmentation of dysarthric speech, along with a word-based language model. The training was conducted for **20 epochs**. *SoX* was used for speed, tempo and loudness perturbations. The options and factors are as mentioned below:

- *speed* option for speed perturbation with the factors 0.9, 1.0 and 1.1
- *tempo* option for tempo perturbation with the factors 0.7, 0.5 and 0.4 based on the factors mentioned in [17]
- *vol* option for loudness perturbation with the factors 0.7, 0.5

SpecAugment is the baseline SpecAugment method as discussed in [9]. SpecAugment was applied on both healthy as well as dysarthric speech. Three different dysarthric SpecAugment (DSA) techniques were devised as discussed in section 2.2. The details of the configurations for the DSA-masks are provided in table 1. Stutter, Hypersnasal and Breathiness masks were designed to augment healthy speech to match dysarthric speech. Hence they were applied only on healthy speech as a part of the ASR training. This model was then used as pre-trained ASR model for adaptation using dysarthric speech, to arrive at a final word error rate (WER).

The data augmenation sequence and manner of application can be visualized as shown in Figure 1. In order to demonstrate the benefits of R-DAE, SDA and DSA to data augmenation of dysarthic speech, we have examined the performance of E2E ASR using each of the techniques separately based on the set-ups discussed below:

Table 1: *Masks used in Dynamic Data augmenation*

System	Mask type
SpecAugment	time warp,time mask,freq mask
DSA-Stutter	time warp,noise mask, stutter mask
DSA-Hypernasal	time warp, time mask, hypernasal mask
DSA-Breathiness	time warp, frequency mask, noise mask
DSA-All	time warp, time mask, noise mask, stutter mask,hypernasal mask

- Healthy speech and R-DAE generated speech for scenarios of no augmenation, with SDA and with SpecAugment.
- DSA with pre-training on augmented healthy speech followed by speaker adaptation using dysarthic speech.
- Effect of different combinations of augmenation procedures along with speaker adaptation using dysarthic speech.

4. Results and Discussion

The E2E-Transformer based ASR was trained on a combination of augmented healthy speech and dysarthric speech and evaluated on dysarthric speech. We present the baseline WER wherein no augmentation techniques were applied. We then proceed to present the ASR performance for static augmentation and dynamic augmentations separately. Finally the ASR is evaluated for the two-stage augmentation.

Table 2: *A comparison of WER for Healthy speech and R-DAE generated speech*

System	Healthy Speech	R-DAE speech
No Augmentation	67.1	55.7
SDA-speed	62.1	51.5
SpecAugment	65.5	52.2

The objective of using R-DAE is to generate speech data with characteristics closer to dysarthric data as compared to healthy speech. In order to get an understanding of the R-DAE augmentation, we compare the performance of the ASR for both healthy speech and R-DAE generated speech for scenarios with no augmentation, SDA-speed and SpecAugment in the table 2. We observe that R-DAE speech is better matched to the dysarthric test data as compared to healthy speech since the WER is lower for all three scenarios.

Table 3: *WER for Dynamic data augmentation*

System	Speaker Ind.	Speaker adapted	Intelligibility			
			Very Low	Low	Mid	High
DSA-Stutter	64.6	32.9	75.8	35.5	29.9	14.0
DSA-Hypernasal	64.2	30.8	75.8	29.1	27.4	13.1
DSA-Breathiness	63.1	30.4	75.2	28.0	27.0	13.0
DSA-All	62.1	**29.0**	73.6	25.1	25.1	12.7

As mentioned in section 3, dynamic augmentations (DSA) specific to dysarthric speech have been applied only on healthy speech data, in order to achieve maximum matching between training and test data.The E2E-Transformer models are trained using DSA-applied healthy speech followed by speaker adaptation using dysarthric speech data. Tables 3 and 4 demonstrate that applying DSA improves the ASR performance across all the DSA techniques. We achieve the lowest WER when all the DSAs are applied together in the final DSA-All system. An absolute improvement of 5% is achieved for healthy data and 7.6% post speaker adaptation. We also examine the improvement at dysarthria intelligibility levels, that has been provided by the UA dysarthic speech corpus. While both SDA and DSA improve the WER across intelligibility levels, we note that SDA gives higher improvement for dysarthic speech with very low intelligibility. It can be observed from Tables 3 and 4 that each of the augmenation techniques applied, contribute to the matching of augmented healthy speech to dysarthric speech.

The overall system comprises both static and dynamic augmentations in cascade as shown in the figure 1. For the final system, a combination of (1) healthy speech SDA followed by DSA-All, (2) R-DAE speech with SDA and SpecAugment and (3) adapted using dysarthric speech SDA follwed by SpecAugment has been used to improve the overall WER of the E2E-Transformer ASR. Please note that all the experimental results reported in Tables 2, 3 and 4 correspond to a training set-up of 20 epochs. While, data augmenation plays a key role in the ASR outcome, significant gains were achieved using speaker adaptation. We achieve an absolute improvement of **16%** in word error rate (WER) over a baseline with no augmentation, with a final WER of **20.6%**.

Table 4: *Overall WER*

System	Overall WER	Intelligibility			
		Very low	Low	Mid	High
No Augmentation	36.6	78.9	40.8	36.1	16.1
SDA	29.0	67.9	29.8	21.8	9.1
DSA-All	29.0	73.6	25.1	25.1	12.7
SDA + DSA-all	26.3	63.1	25.7	22.0	7.1
SDA+DSA-All+R-DAE	**20.6**	61.5	18.9	14.6	6.4

5. Conclusion and Future work

This work proposes a novel two-stage data augmentation scheme to improve ASR performance in terms of WER. We discuss a novel two-stage *data augmentation* scheme to improve ASR performance by introducing two concepts (1) R-DAE and (2) Dysarthric SpecAugment to augment dysarthic speech data. The objective is to leverage our understanding of the acoustic characteristics of dysarthic speech and incorporate this information in the design of the R-DAE, SDA and DSA augmenation policies. Each of the augmentation processes contribute to the improvement of WER of the E2E ASR, with an absolute improvement of 16% in the WER.

SpecAugment and its variants [10, 11] can be explored further to simulate dysarthric speech. Factors and parameters used for both SDA and DSA can be tuned further to produce speech utterances pertaining to specific severity levels. It may be worth exploring the concepts of *Sequence-to-sequence learning (Seq2Seq)* and *Generative adversarial networks (GAN)* for data augmenation, and thereby a robust ASR for dysarthic speech.

6. References

[1] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[2] T.-S. Nguyen, S. Stueker, J. Niehues, and A. Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7689–7693.

[3] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[4] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[5] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition - the "Ethiopian" system for the SLT 2021 children speech recognition challenge," *CoRR*, vol. abs/2011.04547, 2020. [Online]. Available: https://arxiv.org/abs/2011.04547

[6] M. Geng, X. Xie, S. Liu, J. Yu, S. Hu, X. Liu, and H. Meng, "Investigation of data augmentation techniques for disordered speech recognition," *arXiv preprint arXiv:2201.05562*, 2022.

[7] 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.

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

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.

[10] X. Li, Y. Zhang, X. Zhuang, and D. Liu, "Frame-level specaugment for deep convolutional neural networks in hybrid ASR systems," in *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*. IEEE, 2021, pp. 209–214. [Online]. Available: https://doi.org/10.1109/SLT48900.2021.9383626

[11] C. Wang, Y. Wu, Y. Du, J. Li, S. Liu, L. Lu, S. Ren, G. Ye, S. Zhao, and M. Zhou, "Semantic Mask for Transformer Based End-to-End Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 971–975.

[12] Z. Jin, M. Geng, X. Xie, J. Yu, S. Liu, X. Liu, and H. Meng, "Adversarial data augmentation for disordered speech recognition," *arXiv preprint arXiv:2108.00899*, 2021.

[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)*, 2019, pp. 5836–5840.

[14] T. A. M. Celin, T. Nagarajan, and P. Vijayalakshmi, "Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 346–354, 2020.

[15] S. R. Shahamiri, "Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 852–861, 2021.

[16] 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 *Proc. Interspeech 2018*, 2018, pp. 2207–2211.

[17] C. Bhat, B. Vachhani, and S. Kopparapu, "Improving recognition of dysarthric speech using severity based tempo adaptation," in *International Conference on Speech and Computer*. Springer, 2016, pp. 370–377.

[18] 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)*. IEEE, 2019, pp. 5836–5840.

[19] K. Tjaden and G. E. Wilding, "Rate and loudness manipulations in dysarthria," 2004.

[20] P. G. Shivakumar and P. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in *In Proc. INTERSPEECH*, 2016, pp. 3743–3747.

[21] B. Vachhani, C. Bhat, B. Das, and S. K. Kopparapu, "Deep autoencoder based speech features for improved dysarthric speech recognition." in *Interspeech*, 2017, pp. 1854–1858.

[22] C. Bhat, B. Das, B. Vachhani, and S. K. Kopparapu, "Dysarthric speech recognition using time-delay neural network based denoising autoencoder." in *INTERSPEECH*, 2018, pp. 451–455.

[23] L. Rampello, L. Rampello, F. Patti, and M. Zappia, "When the word doesn't come out: A synthetic overview of dysarthria," *Journal of the Neurological Sciences*, vol. 369, pp. 354–360, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022510X16305391

[24] M. Saxon, A. Tripathi, Y. Jiao, J. M. Liss, and V. Berisha, "Robust estimation of hypernasality in dysarthria with acoustic model likelihood features," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 2511–2522, 2020.

[25] A. S. Lee, V. Ciocca, and T. L. Whitehill, "Acoustic correlates of hypernasality," *Clinical Linguistics & Phonetics*, vol. 17, no. 4-5, pp. 259–264, 2003, pMID: 12945600. [Online]. Available: https://doi.org/10.1080/0269920031000080091

[26] Y. Kozaki-Yamaguchi, N. Suzuki, Y. Fujita, H. Yoshimasu, M. Akagi, and T. Amagasa, "Perception of hypernasality and its physical correlates," *Oral Science International*, vol. 2, no. 1, pp. 21–35, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1348864305800047

[27] M. Frohlich, D. Michaelis, and H. Werner Strube, "Acoustic "breathiness measures" in the description of pathologic voices," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 2, 1998, pp. 937–940 vol.2.

[28] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1741–1744, Dec. 2008, iNTERSPEECH 2008 - 9th Annual Conference of the International Speech Communication Association ; Conference date: 22-09-2008 Through 26-09-2008.