Enhancing Emotional Awareness in Individuals with Dysarthria

Filippo Lampa
filippo23@ru.is

Elena Ovsiannikova
elenao23@ru.is

Selma Rimmelzwaan
selma23@ru.is

Abstract

Dysarthria, a speech sound disorder resulting from neurological injury of the motor–speech system's motor component, poses significant challenges to affected individuals, impacting their academic, social, and vocational lives due to reduced speech intelligibility. Computerized speech offers promising avenues for improvement, particularly in areas such as therapy, feedback, and inclusive voice assistants. This project aims to construct an end-to-end emotion recognition model specifically tailored for dysarthric speech, classifying emotional connotations to enhance emotional awareness. The research faces challenges in limited voice data and the broad spectrum of Dysarthria severity. Despite these constraints, the Torgo and UASpeech datasets show promising results in accurate speech recognition for dysarthric speech when merged and used to train Seq-2-Seq models such as Whisper. The paper addresses key research questions, including the suitability of ASR models for dysarthric speech recognition and strategies for data augmentation to improve system accuracy.

Index Terms: Dysarthria, Speech Disorder, Dysarthric Speech Recognition, Automatic Speech Recognition, Seq-2-Seq, Sentiment Analysis, Assistive Technology

1 Introduction

Dysarthria is a motor speech disorder resulting from brain damage or weakened muscles that are important for intelligible speech. It is one of the more common speech disorders in adults since it can result from a variety of different medical problems (Enderby, 2013). Dysarthria may arise from various neurological conditions, each impacting speech production in distinct ways. Common causes include stroke, traumatic brain injury, cerebral palsy, Parkinson's disease, multiple sclerosis, and amyotrophic lateral sclerosis (ALS). These conditions can lead to varying degrees of motor impairment, affecting the muscles used in speech production. Consequently, dysarthria presents in multiple forms, ranging from slurred and slow speech to rapid and mumbling articulation, depending on the underlying neurological cause. The resulting articulation issues are purely due to reduced motor control, meaning that dysarthria does not affect language formation or comprehension in other ways. The consequence is that dysarthric speakers generally know what and how to communicate, but are often still not understood, making this speech disorder especially frustrating for patients. The reduced intelligibility that often comes with dysarthria affects patients in their academic, social, and vocational lives (Doyle et al., 1997). Computerized speech could greatly improve the lives of patients suffering from dysarthria in many ways. Successful advances have been made in computerized therapy (Palmer et al., 2007), effective feedback (Bakker et al., 2019), inclusive voice assistants (Nicolao) (Jefferson, 2019), and many more areas. Specifically, computerized speech could help dysarthric speakers properly convey emotions and control the emotional connotation of speech. Due to differences in speed, pitch, volume, and rhythm, it can be challenging for dysarthric speakers to communicate their emotions properly, which greatly affects self-expression and confidence. Becoming more aware of the emotional connotation that a phrase has provides the opportunity for a patient to practice and improve on it, resulting in more satisfying interactions. Our aim for this project is to construct an end-to-end emotion recognition model that recognizes dysarthric speech and can classify it according to its emotional connotation. The goal of this research is to enhance the emotional awareness of dysarthric speakers and give them a tool to potentially improve their emotional intelligibility. To achieve this and create a practical tool, the project can be separated into two parts: recognizing dysarthric speech and performing sentiment analysis on said speech. In Section 2, we will outline the state of the art in the field of dysarthric automatic speech recognition, reporting the outcomes of our literature review. Section 3 contains the details about the work performed to realise the final system, from the choice of the dataset to train the speech recognition model to the phase of sentiment analysis. In Section 4 the outcomes will be reported, looking back at the design and implementation choices made and discussing both their impact on the final results and future experiments and research that could improve them. Finally, Section 5 provides a brief conclusion.

2 Related Works

Our study reckons on automatic speech recognition to transcribe dysarthric speech. (Qian and Xiao, 2023) performed a comprehensive survey, analysing 63 papers in the field of ASR dysarthric speech. They highlighted how, among the different approaches for ASR such as acoustic models and language-lexical models, sequence-to-sequence models can improve the accuracy of speech recognition for dysarthric speech, although they need larger amounts of data to be effectively trained. Various approaches have been tried to improve the accuracy of these Seq-2-Seq models. (Takashima et al., 2019) introduced an innovative end-to-end ASR framework that effectively recognizes dysarthric speech, even with a small dataset, by combining an acoustic and a language model. The former is shared among dysarthric speakers, while the latter is specific to each language, regardless of dysarthria. (Lin et al., 2020a) proposed a parameter restructuring method for the acoustic model, optimizing the use of limited data for dysarthric speech ASR training. In addition, (Lin et al., 2020b) employed knowledge distillation to significantly reduce phone error rates in dysarthric speech recognition compared to baseline methods. (Soleymanpour et al., 2021) took a distinct approach by implementing data augmentation techniques based on sub-word models, resulting in notable reductions in character and word error rates. Their method involved "prosodic transformation" and "time-feature masking." (Almadhor et al., 2023) developed a unique spatiotemporal dysarthric speech recognition system using spatial CNN and attention transformers. Their approach achieved high word recognition accuracy but faced challenges related to overfitting, impacting generalization. These research efforts collectively aim to improve the accessibility and accuracy of ASR for individuals with dysarthria. In our study on enhancing dysarthric speech recognition, we also had to focus on expanding limited datasets since there is a significant gap in available data with sufficient vocabulary and a variety of dysarthria. This issue is central to improving the accuracy and versatility of speech recognition models. We've researched different methods of data augmentation and found two key studies that align with our goals. (Geng et al., 2022) looked at several data augmentation techniques. They examined vocal tract length perturbation (VTLP), tempo perturbation, and speed perturbation, with a focus on the latter, to improve the performance of speech recognition systems. They used both normal and disordered speech in their experiments and incorporated learning hidden unit contributions (LHUC) for speaker adaptive training. Their approach, especially the application of speed perturbation, yielded a 2.92% absolute reduction in word error rate (WER) for a system trained on the UASpeech corpus. This suggests that simple manipulations of speech speed can have a significant impact on the system's effectiveness in recognizing disordered speech. Their findings showed that changing the speed of speech in training data can make recognition systems more accurate. This approach is straightforward and gives good results. (Harvill et al., 2021) used voice conversion to increase the number of words that ASR systems can recognize. They pointed out the limited vocabulary in existing ASR systems due to the scarcity of dysarthric speech data. By converting voices, they were able to train ASR systems to recognize words that weren't in the original training data, expanding the system's vocabulary. This indicates that through voice conversion, it is possible to create realistic examples of words not previously recorded by speakers with dysarthria, potentially improving recognition performance in practical settings. Both studies underscore the potential of data augmentation to address the challenge of limited dysarthric speech data. By synthesizing new words and altering the speed of existing recordings, these research efforts contribute significantly to the development of more inclusive and effective ASR models for individuals with dysarthria. The focus of this work, rather than trying to improve the already widely explored accuracy of ASR for dysarthric speech, is on delivering a complete end-to-end system able to give real-time feedback to the speaker and increase its awareness about how its speech could be perceived by others.

3 Methods

3.1 Overview

Being the core objective of the project the development of a comprehensive system to accurately identify emotions in dysarthric speech, our methodology involved a sequence-to-sequence ASR model used to transcribe the speech in input, and a rule-based model for sentiment analysis.

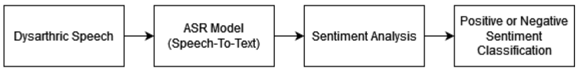

Figure 1: Overview of the system

In particular, given the content of the datasets and the large amount of data needed to obtain satisfying results in speech recognition as mentioned in Section 2, the process involved, before training, a pre-processing phase, whose steps were: data cleaning, transformation, augmentation.

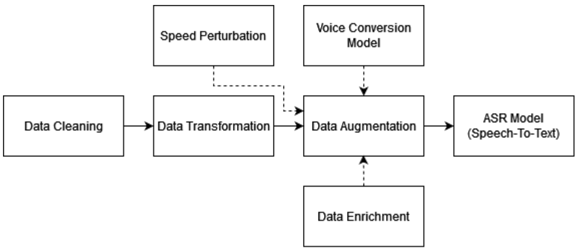

Figure 2: Overview of the ASR model's training

In the next sections, we will outline the datasets, model selection, pre-processing steps, and data augmentation strategies employed to build the model. All the code can be found on GitHub [1].

3.2 Datasets

For this research, we made use of two datasets: the Torgo database (Rudzicz et al., 2012) and UASpeech database (Kim et al., 2008). The Torgo dataset consists of speech by male and female patients with cerebral palsy or amyotrophic lateral sclerosis. The patients have different levels of dysarthria. The prompts used in this dataset consist of non-words, short words, restricted sentences (speakers are asked to read a specific, phoneme-rich sentence), and unrestricted sentences (speakers are asked to talk spontaneously about a series of pictures). The whole dataset contains approximately 2900 audio files, which is 18GB of dysarthric audio and transcriptions, equaling a total of 21 hours of speech. The UASpeech dataset, to which the University of Illinois gave us access, consists of audio by 15 speakers with Cerebral Palsy and 13 healthy speakers of the same age, comprising 102.7 hours of speech. The prompts used include digits, letters, computer commands, common words, and uncommon words. We used the speech of 4 different speakers with dysarthria and 9 healthy speakers. The intelligibility of the patients with dysarthria ranges from "Mid" to "High". Both male and female speakers are included in our experiments. This all summed up to a total of 69615 utterances of which, after pre-processing, we used 69254 of those utterances.

3.3 Pre-processing

3.3.1 Data Cleaning and Transformation

First, the data have been cleaned by removing erroneous or unnecessary entries, i.e. entries containing the value "xxx", meaning the audio did not have an associated transcription or rows where data were mistakenly placed in the wrong columns. Additionally, we identified certain recordings that were merely transcriptions of repetitive sounds, noted as instructions like "[repeat a sound three times]". These recordings, while potentially offering insights into speech patterns, were ultimately deemed unsuitable for our study's purposes due to their repetitive nature. Consequently, these repetitive sound recordings were excluded from our dataset. This decision was made to ensure that our analysis focused solely on meaningful and informative data. We then restructured the dataset to only contain the features we're interested in: gender, whether the speaker has dysarthria or not, the audio, and the prompt in text format. We then split it into 80% for training, 10% for testing, and 10% for validation, aiming to save as much data as possible for the training. The recordings in the UASpeech dataset contain some stationary noise, so we removed this using Noisereduce [2]. Additionally, we removed the silent audio present at the beginning and end of each recording. Finally, during training, the data collator class takes care of preparing mini-batches, applying padding to both audios and labels to make

[1] https://github.com/filippo-lampa/DysarthriaEmoEnhancer

[2] https://github.com/timsainb/noisereduce

them all the same length, and tokenizing the latter. We mask the padding in the labels with -100 to correct for loss calculation and beginning-of-sequence tokens are removed.

3.3.2 Data augmentation

Although the Torgo dataset is one of the biggest ones currently available on dysarthric speech, it is on its own not enough to deliver consistent results (Soleymanpour et al., 2023). Data augmentation is necessary to be able to acquire enough training data to train the speech recognition model. We attempted two methods of data augmentation. Initially, we tried to synthesize more Dysarthric speech by using voice conversion as outlined by (Harvill et al., 2021). After performing the preliminary operations such as applying noisereduce, trimming silence, extracting mel log spectrograms and aligning utterances using Dynamic Time Warp, we created and trained our voice conversion model consisting of 6 layers of multihead attention using TransformerEncoderLayer from PyTorch. However, after training the model for 150.000 iterations, the results obtained were of no use, presumably due to errors in the encoding and preparation phase of the audio data, where every mistake in setting the parameters has the potential to spoil the final result. In literature, successful data augmentation on dysarthric speech has also been achieved in past research by using speed perturbation (Geng et al., 2022). We sped up the audio files in the Torgo dataset by 0.9, 0.95, 1.05, 1.1. Speed perturbation turned out to not be enough to substantially improve our speech recognition model. Being the results from both data augmentation attempts not beneficial enough, to acquire enough training data, we decided to perform data enhancement by merging the UASpeech dataset (Kim et al., 2008), a dataset containing dysarthric spoken digits, letters, common and uncommon words, with the Torgo dataset mainly containing words and sentences, to be able to feed the model with a larger number and variety of data and improve generalization.

3.4 Model

Different pre-trained models have been tested before deciding which one to fine-tune with our data. AzureAI models allowed to get satisfying results by just uploading the data and letting the platform manage the rest. This black-box approach did not give us any control over the hyperparameters, leading us to discard this option, also considered the improvement over the other tested models being very small. A key factor in our decision-making process was the need for a more transparent and customizable approach, which led us to explore other options. Hugging Face, on the other side, allows customization and training of models at a lower level, and offers both CTC and Seq-2-Seq pre-trained models. Given the decent results achieved by the model after only a few training iterations on the Torgo dataset, and the reasonable amount of data available after the data enhancement phase, the choice fell on the Whisper-small model (Radford et al., 2023) from Hugging Face. The training was performed on a cluster using two Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz CPUs, setting the following hyperparameters: A more detailed list of hyperparameters

Optimizer	Adam
Train batch size	8
Gradient accumulation steps	2
Learning rate	1e-5
Max steps	500
FP16	True

Table 1: Hyperparameters for Whisper models

can be found on Hugging Face, at https://huggingface.co/FilippoLampa/dysarthria-emo-enhancer.

3.5 Sentiment analysis

To classify the recognized speech as positive or negative we used the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment lexicon (Hutto and Gilbert, 2014). VADER is a model that uses a dictionary of words or lexical features that are mapped to a sentiment score ranging from -4 to 4. To decide the sentiment score of a sentence, the sentiment score of each lexical feature in the given sentence is summed up. The sum is normalized to an overall sentiment score between -1 and 1, which we threshold to zero in order to state whether the word or sentence could be interpreted as positive or negative. In addition to the lexical features, VADER takes five heuristics into account when calculating the final sentiment score: the punctuation, capitalization, degree modifiers, the effect of the word “but”, and negation. The VADER lexicon is specifically curated for sentiment analysis on “microblog-like contexts” such as textual social media content. We decided to use the

VADER lexicon for our tool since it's computationally efficient and has proven to have a comparable performance to other high-end sentiment analysis models. It had for example an F1-score of 0.96 when tested on Tweets, making it a good choice in a scenario where dysarthric people should be able to use our system to get a feedback about ordinary sentences pronounced in informal contexts on a daily basis. In addition, the effectiveness of VADER in microblog-like contexts, characterized by brief and sometimes fragmented expressions, mirrors the communication style that might be used by individuals with dysarthria, who may prefer shorter, simpler sentences. This similarity suggests that VADER is well-equipped to handle the nuances and limitations of dysarthric speech, making it an ideal choice for our system, designed to provide immediate and accurate sentiment feedback to dysarthric individuals in their daily interactions.

4 Results and Discussion

4.1 Evaluation

To evaluate our ASR models, we focused on two main metrics: accuracy, which measures the model's ability to transcribe words correctly, and the Word Error Rate (WER), which quantifies the transcription errors. Our evaluation compared three versions of the Whisper model: the original base model, the version fine-tuned with the Torgo dataset, and the version enhanced with the combined Torgo and UASpeech datasets. Our evaluation involved the comparison of the three distinct iterations of the Whisper model, each designed to serve a unique purpose. The initial variant represented the Whisper base model, providing us with a foundational benchmark for assessing performance. The second model was the finely-tuned version, trained specifically on the Torgo dataset. The third one incorporated data from the merged dataset. As for the final model, also the one trained on the Torgo dataset can be found on Hugging Face [3] or retrieved directly using Hugging Face's Pipelines. The consistent application of hyperparameters across these models ensured that our comparison was fair and that any differences in performance were due to the models' training and not to variations in setup. The values of the hyperparameters are the ones previously listed in Table 1. To test the performances after the training, a custom test script has been written, which iteratively feeds the models with more than 5500 unseen words and sentences and evaluates their outputs. The generated transcriptions are compared with the expected ones. In this comparison, differences in capitalization or punctuation and the presence of leading/trailing spaces are not counted as errors, since they do not have an impact on the quality of the sentiment analysis. Following, the testing results for the three models are reported. The results shown in Table 2, not only underline the effectiveness of fine-tuning a generic pre-trained ASR model on dysarthric speech recognition, but also highlight the impact that data enhancement had on achieving lower WER and higher accuracy. Although the final model is able to better generalize and understand dysarthric speech, from both the rather low accuracy and a subjective evaluation performed by using the final sentiment analysis tool, it emerges that it is still not perfect and sometimes misinterpret words. It must be noted that the training process has been cut by the cluster after hitting the maximum time. As we can see in the following table, the training of the final model lasted roughly 29 hours, achieving a WER of 34.52. However, at the time of the interruption, training, validation and word error rate were still decreasing, suggesting that there was some margin of improvement before starting to overfit or showing a degradation in performances. Giving the model more time to train and using a GPU could, in addition to allow the loss lowering even more, give the chance to increase the number of iterations and make the learning rate smaller, which will result in a slower training and

	Accuracy	WER-rate
Whisper-base	0.23	383.91%
Whisper-Torgo	0.30	80.63%
Whisper-merged	0.68	32.28%

Table 2: Accuracy and WER-rate for the base Whisper model, the Whisper model fine-tuned on the Torgo dataset, and the Whisper model fine-tuned on both the Torgo and UASpeech dataset

WER Ortho	36.0478
WER	34.5269
Training runtime	106811.1634

Table 3: WER rates and training runtime for Whisper model fine-tuned on merged dataset

[3] https://huggingface.co/FilippoLampa/whisper-small-dv

a model better able to capture fine-grained features in the acoustic signals, reducing the number of misinterpreted words.

4.2 Ethical Considerations

In our project, we used datasets that are made available to the public for research purposes and are therefore expected to be used properly. This means making sure our research matches the purpose for which these datasets were shared. Our responsibility is to ensure that the data we use for training our model includes a good mix of men and women, covers different age groups, and represents various levels of dysarthria. This is important to make sure our model works well for everyone, no matter their gender, age, or how severe their dysarthria is. When we think about using the emotion recognition system in real life, we have to be considerate: the main goal is to help people with dysarthria communicate better, without taking away their independence or privacy. The system should be accessible to everyone, even those with other disabilities. Lastly, the ethical implications of artificial intelligence in healthcare and assistive technologies must be considered. This involves being open about how the system functions and having plans to address any mistakes or unexpected outcomes. By focusing on these ethical considerations, our goal is to ensure the system is both effective in its operation and considerate and beneficial for individuals with dysarthria.

5 Conclusions

In this project, we aimed to develop an end-to-end emotion recognition model specifically tailored for dysarthric speech. Dysarthria is a motor speech disorder that significantly impacts the lives of affected individuals, affecting their ability to communicate and express emotions. Our approach combined the Whisper ASR model with data from the Torgo and UASpeech datasets to improve the accuracy of speech recognition for dysarthric speech. Our results demonstrated that fine-tuning the Whisper model on dysarthric speech data, especially when supplemented with additional data, led to significant improvements in accuracy and reduced Word Error Rate (WER). While the final model showed promise in recognizing dysarthric speech, there is still room for improvement. Further refinement and training could lead to even better results.

6 Future Works

While we reached the goal of delivering an end-to-end system, we are keenly aware that there is significant room for improvement and further refinement in the domain of dysarthric speech recognition and emotion analysis. Since, as we saw, the amount and quality of data has a major impact on the final result, the data augmentation phase is worth being performed again with greater care and caution, comparing even more methodologies which can be found in literature. Larger and more diverse datasets may help the model generalize better to various levels of dysarthria severity and speech contexts. Also trying different splits of the dataset could have a major impact on the final accuracy of the model. Spending some time fine-tuning hyperparameters could also improve the results, including learning rate and training duration, as mentioned in Section 4, which could lead to better model convergence and reduced word error rates. Integration of Sentiment Analysis: Integrating the emotion recognition model with a sentiment analysis tool using the VADER lexicon provided a valuable addition to our system. Future work could involve enhancing and refining the sentiment analysis component, possibly with deep learning models trained on dysarthric speech. To make this system more practical and beneficial to dysarthric individuals, it would be important to gather user feedback from the already existing user interface. The system could exploit that feedback to adjust its suggestions through a reward mechanism, to provide a better help to users in improving their speech clarity and emotional expressiveness. Ethical Considerations: As with any assistive technology, ethical considerations are paramount. Ensuring that the system respects user privacy, provides appropriate assistance, and is accessible to a wide range of users is essential. In summary, this project represents a small step toward improving the lives of individuals with dysarthria. The development of an emotion recognition model tailored for dysarthric speech has the potential to enhance communication, emotional expression, and overall quality of life for those affected by this speech disorder. Continued research and development in this area can lead to further advancements and benefits for the dysarthric community.

References

Ahmad Almadhor, Rizwana Irfan, Jiechao Gao, Nasir Saleem, Hafiz Tayyab Rauf, and Seifedine Kadry. 2023. E2e-dasr: End-to-end deep learning-based dysarthric automatic speech recognition. *Expert Systems with Applications*, 222:119797.

Marjoke Bakker, Lilian Beijer, and Toni Rietveld. 2019. Considerations on effective feedback in computerized speech training for dysarthric speakers. *Telemedicine and e-Health*, 25(5):351–358.

Philip C Doyle, Herbert A Leeper, Ava-Lee Kotler, Nancy Thomas-Stonell, Charlene O'Neill, Marie-Claire Dylke, and Katherine Rolls. 1997. Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of rehabilitation research and development*, 34:309–316.

Pam Enderby. 2013. Disorders of communication: dysarthria. *Handbook of clinical neurology*, 110:273–281.

Mengzhe Geng, Xurong Xie, Shansong Liu, Jianwei Yu, Shoukang Hu, Xunying Liu, and Helen Meng. 2022. Investigation of data augmentation techniques for disordered speech recognition. *arXiv preprint arXiv:2201.05562*.

John Harvill, Dias Issa, Mark Hasegawa-Johnson, and Changdong Yoo. 2021. 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)*, pages 6428–6432. IEEE.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Madeline Jefferson. 2019. Usability of automatic speech recognition systems for individuals with speech disorders: Past, present, future, and a proposed model.

Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame. 2008. Dysarthric speech database for universal access research. In *Ninth Annual Conference of the International Speech Communication Association*.

Yuqin Lin, Longbiao Wang, Jianwu Dang, Sheng Li, and Chenchen Ding. 2020a. End-to-end articulatory modeling for dysarthric articulatory attribute detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7349–7353. IEEE.

Yuqin Lin, Longbiao Wang, Sheng Li, Jianwu Dang, and Chenchen Ding. 2020b. Staged knowledge distillation for end-to-end dysarthric speech recognition and speech attribute transcription. In *INTERSPEECH*, pages 4791–4795.

Rebecca Palmer, Pam Enderby, and Mark Hawley. 2007. Addressing the needs of speakers with longstanding dysarthria: computerized and traditional therapy compared. *International journal of language & communication disorders*, 42(S1):61–79.

Zhaopeng Qian and Kejing Xiao. 2023. A survey of automatic speech recognition for dysarthric speech. *Electronics*, 12(20).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. 2012. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46:523–541.

Mohammad Soleymanpour, Michael T Johnson, and Jeffrey Berry. 2021. Dysarthric speech augmentation using prosodic transformation and masking for subword end-to-end asr. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 42–46. IEEE.

Mohammad Soleymanpour, Michael T Johnson, Rahim Soleymanpour, and Jeffrey Berry. 2023. Accurate synthesis of dysarthric speech for asr data augmentation. *arXiv preprint arXiv:2308.08438*.

Yuki Takashima, Tetsuya Takiguchi, and Yasuo Ariki. 2019. End-to-end dysarthric speech recognition using multiple databases. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6395–6399. IEEE.