

Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System

Seyed Reza Shahamiri^{ID}

Abstract—Dysarthria is a disorder that affects an individual's speech intelligibility due to the paralysis of muscles and organs involved in the articulation process. As the condition is often associated with physically debilitating disabilities, not only do such individuals face communication problems, but also interactions with digital devices can become a burden. For these individuals, automatic speech recognition (ASR) technologies can make a significant difference in their lives as computing and portable digital devices can become an interaction medium, enabling them to communicate with others and computers. However, ASR technologies have performed poorly in recognizing dysarthric speech, especially for severe dysarthria, due to multiple challenges facing dysarthric ASR systems. We identified these challenges are due to the alternation and inaccuracy of dysarthric phonemes, the scarcity of dysarthric speech data, and the phoneme labeling imprecision. This paper reports on our second dysarthric-specific ASR system, called Speech Vision (SV) that tackles these challenges by adopting a novel approach towards dysarthric ASR in which speech features are extracted visually, then SV learns to see the shape of the words pronounced by dysarthric individuals. This visual acoustic modeling feature of SV eliminates phoneme-related challenges. To address the data scarcity problem, SV adopts visual data augmentation techniques, generates synthetic dysarthric acoustic visuals, and leverages transfer learning. Benchmarking with other state-of-the-art dysarthric ASR considered in this study, SV outperformed them by improving recognition accuracies for 67% of UA-Speech speakers, where the biggest improvements were achieved for severe dysarthria.

Index Terms—Dysarthria, dysarthric speech recognition, deep learning, synthetic speech.

I. INTRODUCTION

DYSARTHRIA is a neurological motor speech disorder characterized by an individual's loss of control of their motor subsystems [1]. Symptoms of dysarthria can vary significantly depending on the underlying cause and severity that may result in speech produced be moderately slurred or severely unintelligible as the disease progresses [2].

Manuscript received January 19, 2021; revised April 15, 2021; accepted April 27, 2021. Date of publication April 30, 2021; date of current version May 7, 2021.

The author is with the Department of Electrical, Computer, and Software Engineering, Faculty of Engineering, The University of Auckland, Auckland 1142, New Zealand (e-mail: admin@rezanet.com).

Digital Object Identifier 10.1109/TNSRE.2021.3076778

Since dysarthric speech can be significantly unintelligible [3], a typical audience may face difficulties communicating with dysarthric speakers unless s/he has prior experience with such individuals. This substantially affects the abilities of dysarthric speakers in communicating with normal speakers. Furthermore, dysarthria can often accompany neurological conditions; thus, many people with dysarthria are also physically debilitated, which means interfacing with digital devices and computers via mouse, keyboard, and touchscreen may be challenging or impossible. For such individuals, automatic speech recognition (ASR) technologies can be a desirable alternative to enable them to interface with digital devices or become a communication intermediary [4]; ASR technologies can significantly improve the quality of life of dysarthric individuals via their applications in Augmentative/Alternative Communication (AAC) tools.

Dysarthria may significantly affect how phonemes are pronounced by dysarthric individuals, especially in severe dysarthria, making dysarthric speech quite different from normal speech. The phonemes pronounced by the affected individuals can be highly imprecise, with pitch pauses in vocalic segments and consonants' production inaccuracies. These alternations may also mask the discriminative acoustic attributes that ASR systems rely upon to recognize phonemes. Furthermore, because the effects of the disability vary from one subject to another, speech variations among dysarthric speakers are significantly more than normal speech. These differences make acoustic modeling components in standard ASR systems ineffective in mapping dysarthric speech signals to phonemes correctly; they need to deal with challenges caused by unusual and imprecise phonation, tempo and speed inconsistencies, sonorants random shifting of formant frequencies, etc. Thus, normal speech recognition systems have shown poor performance in recognizing dysarthric speech [5]. A review conducted in [6] suggests that while state-of-the-art normal ASR systems perform well on mild dysarthria, the performance degrades significantly as the condition worsens. The study shows how various characteristics of dysarthria at different severity levels affect the performance of standard ASR systems and reported the Word Error Rate (WER) measure was very high for the most severe cases.

Hence, tackling the dysarthric ASR problem requires speech recognition systems specially designed to recognize dysarthric speech. Studying the literature shows multiple

attempts to design dysarthric-specific ASR systems; however, these systems still often perform poorly, especially for severe dysarthria, since they mostly adopt standard acoustic modeling approaches used in normal ASR systems that rely on identifying phoneme sequences. Additionally, the increased variability of speech associated with dysarthria means building a dysarthric ASR requires significantly more dysarthric speech samples than building a normal ASR. The increased data is necessary for the ASR systems' acoustic models to comprehend dysarthric speech better and deal with the variability of dysarthric speech and phoneme alternations and masking issues. Nonetheless, this is not the case since the research community has a very limited amount of dysarthric speech data available to create accurate ASR systems because the subjects are not capable of producing enough data as their speech production muscles are also weakened, resulting in physical fatigue and frustration when they try to talk for longer periods.

In addition, since the subjects are typically not capable of producing accurate sounds, labeling the phonemes for dysarthric speech can be challenging and inaccurate. Although modern end-to-end deep learning ASR models do not require phoneme-level labeling given there are sufficient data available for them to learn, dysarthric ASR systems do not benefit from this advantage due to the data scarcity problem. Consequently, any attempt to design dysarthric ASR should consider solutions to overcome 1) the alternation and imprecision of phonemes in dysarthric speech, 2) the scarcity of dysarthric speech data, and 3) the dysarthric speech phoneme labeling inaccuracy. This paper reports on our progress to develop our second dysarthric-specific ASR system, called Speech Vision (SV), which continues from our first attempt published in 2014 [7]. This new system benefits from deep learning advancements, data augmentation, and synthetic speech generation ever since we developed our first system and addresses the challenges mentioned above.

SV is a whole-word, isolated speech dysarthric ASR. Instead of using a standard acoustic modeling component to recognize the sequence of phonemes, SV adopts a novel approach towards speech recognition in which speech features are presented visually. SV then learns to see the shape of the words pronounced by dysarthric individuals using our deep 2-dimensional Spatial Convolutional Neural Network (S-CNN) and recognizes individual words. It adopts visual data augmentation techniques to augment the available dysarthric speech data during training to help with tackling the data scarcity problem. We also applied transfer learning to maximize the effects of learning from healthy speech-visuals and transfer that knowledge to dysarthric speech without relying on sound-specific information, which further improved SV's odds with the limited dysarthric data available.. Furthermore, we experimented with speech generation systems to reconfigure a text-to-normal-speech system to produce dysarthric synthetic speech and used these extra speech samples during SV's training. Finally, because SV does not operate by recognizing phoneme features, it is not affected by the difficulties associated with the dysarthric speech phoneme labeling. SV has

been verified in detail and compared to other related dysarthric ASR systems in the literature.

II. RELATED WORK

This section surveys the state-of-the-art, dysarthric-specific, English ASR systems published in the literature. To avoid redundancy, we excluded reviewing those works we already reviewed in [7], [8].

Our first attempt to design a dysarthric ASR started with studying how to best present dysarthric speech features and finding the most effective MFCC setup [8]. The findings of that study were then used to design our first system that benefited from an active learning theory called multi-views multi-learners (MVML), in which several learners were used to distribute the complexity of pattern recognition problems [7]. We developed a dysarthric multi-networks speech recognizer (DM-NSR) based on a realization of MVML using an array of artificial neural networks (ANNs) capable of improving the tolerance of dysarthric speech. Trained on a vocabulary of 25 common words from UA-Speech dataset [9], the DM-NSR delivered considerable efficacy improvements over the baseline system across all severity levels of dysarthria. Nevertheless, the data scarcity problem resulted in a sharp accuracy reduction when we attempted to increase the vocabulary size since a speech recognition task's difficulty increases linearly as the vocabulary size gets larger [10], [11]. The limitations of DM-NSR led to the development of Speech Vision.

Between the development of DM-NSR and Speech Vision, there have been few other attempts to design dysarthric-specific ASR. The first attempt is [12], in which a whole-word speaker adaptive dysarthric ASR was designed and evaluated on UA-Speech speakers with a vocabulary size of 155 words. Based on whether an ASR system is open-set or closed-set speaker, ASR tasks are categorized into three categories. The first category is speaker-dependent (SD), in which the ASR is trained to recognize a specific speaker's speech. The second category is speaker-independent (SI), where the ASR recognizes speech uttered by any speaker, even if speakers' data were not given to the ASR during training. With speaker-adaptive (SA) ASR, a speaker-independent ASR is typically customized to learn the acoustic features of a specific speaker. With dysarthric ASR, SD and SA approaches are more favorable due to the increased speech variability between different dysarthric individuals and the existing data scarcity problem significantly impacting the performance of SI dysarthric ASR. Among SD and SA dysarthric ASR, SA seems to be more on-demand lately as the SI system can be initially trained using normal speech and then adopt to individual speakers with dysarthria [5].

The study conducted in [12] investigated the best baseline SI system to design an adaptive dysarthric ASR. It then proposed a hybrid adaptation approach based on maximum a posterior (MAP) combined with maximum likelihood linear regression (MLLR) to adjust the Hidden Markov Model (HMM) in the baseline ASR. Speech data were presented as 12-dimensional MFCCs in 25ms frames with a sliding window

of 10ms, and 15 UA-Speech speakers were considered in which block B1 and B3 utterances were used for training and B2 for testing. Despite achieving a high absolute average accuracy across all dysarthria severity levels for the SA models, the study was inconclusive due to the difficulties associated with selecting the baseline ASR. In particular, the authors concluded that different baseline systems should be considered for each dysarthria severity level by individually studying dysarthric acoustic characteristics to achieve the highest accuracies; hence, no best speaker-adaptation strategy was recommended. Furthermore, MAP and maximum likelihood-based ASR systems are effective when enough training data are available, which is not the case for dysarthric ASR [13].

Before deep learning (DL) based ASR approaches achieved near human-level performance to recognize normal speech, HMMs were amongst the most popular generative algorithms in ASR tasks. As such, multiple studies tried to apply them in developing dysarthric ASR, similar to [12]. Nevertheless, because the speech produced by dysarthric individuals is typically partial and incomplete, and due to the scarcity of dysarthric data, conventional HMM-based ASR approaches tend to perform poorly when given dysarthric speech. Hence, some studies were more interested in hybrid HMM-based ASR or customized HMMs. An example is a small vocabulary dysarthric ASR based on Generative Model-Driven Feature Learning in which conventional HMMs were trained by extra features produced by log-likelihood and transition probability Support Vector Machines (SVMs) in addition to the 39-dimensional MFCCs [13]. The system was trained on 15 dysarthric speakers' data provided by UA-Speech, but the vocabulary size was only 29 words. The study concluded that conventional HMMs did not perform well, but when log-likelihood SVM features were considered, an overall Word Recognition Accuracy (WRA) of 87.91% was achieved.

To improve acoustic features in dysarthric ASR, Vachhani *et al.* [14] trained autoencoders on normal speech to enhance dysarthric speech features. The authors also applied a severity-based tempo adaptation to modify the speech data. Once the features were augmented, HMM-based ASR systems were trained on UA-Speech speakers, but the authors did not mention the vocabulary size. In another attempt, the authors experimented with manually augmenting normal speech samples to capture dysarthric speech characteristics [15]. Particularly, they manipulated the duration of normal speech samples using both speed and tempo-based augmentation [16], creating 3,458 augmented speech samples, then trained an ASR with a vocabulary of 19 words from UA-Speech on both original and the augmented dysarthric speech. Nonetheless, an issue with both of these studies is in choosing the evaluation metric used to measure their ASR performance. Both studies evaluated their ASR systems with UA-Speech data, but the dataset only provides isolated word samples rather than continuous speech samples, yet the performance of their systems was evaluated by measuring Word Error Rate (WER). WER is measured for word sequence tasks based on the number of word substitutions, deletion, insertion, etc., in the prompts recognized by the ASR, where an isolated speech ASR does not produce a word sequence by default. However, the authors

did not explain how they trained and tested a continuous speech ASR using isolated speech data and measured WER instead of word recognition related metrics.

Another popular dysarthric speech corpus dataset is TORG [17] that contains continuous dysarthric speech samples. España-Bonet and Fonollosa [18] experimented with a hybrid DL and HMM-based continuous speech ASR trained on TORG data in which different DL algorithms were used to estimate the HMM state likelihoods. Speech features were presented as the fusion of 13-dimensional MFCCs with 25ms frames sliding each 15ms, Linear Discriminative Analysis transformation that extracted 40-dimensional frame sequences, and Maximum Likelihood Linear transformation to calculate the vector correlations. The authors built multiple ASR systems by selecting various DL algorithms such as the standard fully connected dense neural networks, CNNs, and Long Short-Term Memory neural nets. The best mean WER was achieved with the standard neural net, which was expected since the dysarthric data provided by TORG solely is not enough to train deep and complex DL models.

A similar study is [19], in which ways to improve hybrid HMM-based dysarthric ASR on TORG were studied. The authors experimented with finding and fine-tuning the required parameters of Gaussian Mixture Model (GMM) acoustic models and then applied their best performing configuration to training DL-HMM based ASR solutions where the neural net hyperparameters were provided by their previous study [20]. The neural net was a fully connected, dense model. The authors reported a 17.62% relative reduction of WER comparing to [18] across all speakers. Unlike previous studies, the authors reported that combining normal speech samples with dysarthric speech reduced the recognition accuracy for server and moderate dysarthria. Additionally, this study highlights the complexity of hybrid approaches where different algorithms are used in different stages of ASR, comparing to end-to-end DL models, as each of these algorithms needs to be trained and fine-tuned individually while end-to-end models are more straightforward to train; this highlight is also consistent with the current trend of using end-to-end models for normal ASR systems.

In another study, Takashima *et al.* [21] proposed another hybrid DL-HMM-based isolated-speech ASR where the acoustic model was a Convolutional Restricted Boltzmann Machine (CRBM) pre-trained on normal speech. The CRBM in this study mapped segment MFCC mel-maps extracted by a CNN to 54 phonemes, but the authors experimented with only one Japanese speaker diagnosed with athetoid cerebral palsy, and the severity of the impairment was not provided in the paper.

An AAC system composed of a continuous-speech, SD, HMM-based Dysarthric ASR, an error correction module, and speech synthesis system was proposed by Celin *et al.* [22]. In this study speech is initially recognized using the continuous ASR then corrected using a weighted finite state transducer (WFST) before being uttered by the synthesis system. The ASR was trained on 10 Nemours dysarthric speakers [23] and a Tamil corpus. The authors reported that the inclusion of the error correction component of the

AAC system proved to be highly effective in improving the overall performance.

The latest attempt at the time of this writing is[24], where a comparison between MFCCs, mel-frequency spectral coefficients, and perceptual linear prediction features extraction approaches was made to develop a dysarthric phoneme recognition system. Then, another comparison was made between CNN and Long-Short-Term Memory neural architectures and benchmarked with the conventional GMM-HMM-based approaches. This study presented a dysarthric phoneme recognizer trained and evaluated using 11 Nemours dysarthric speakers, and the authors reported the CNN trained with perceptual linear prediction features to recognize 44 phonemes achieved the highest accuracy of 82% for mild dysarthria. Although the results presented in this study is comprehensive, this study and other similar ones that rely on phoneme recognition or mapping to recognize dysarthric ASR do not address the dysarthric ASR issues mentioned before as they still need to rely on ambiguous and inconsistent dysarthric sounds and require phoneme labeling until the data scarcity problem is addressed. On the other hand, our proposed solution is not affected by these limitations, as explained in the next section. It is pertinent to note that[24] was excluded from our comparative study since a compete ASR was not proposed, and the phoneme accuracy measured was not comparable to WER or WRA – two criteria usually used to evaluate ASR efficacy.

III. SPEECH VISION (SV)

SV is a whole-word and isolated speech, deep learning ASR system specifically designed to recognize dysarthric speech that considers all three challenges highlighted in section 1. SV is explained in detail here.

A. Dysarthric Speech as Voicegrams

As explained before, the unreliability of phonemes uttered by dysarthric individuals is one of the reasons making conventional acoustic models representing the correlation between a speech signal and phoneme sequences inaccurate when given dysarthric speech. To overcome this challenge, we investigated other ways to present speech without relying on phoneme information solely. In this investigation, we were more interested in visual representations of speech so that visual-data augmentation approaches could also be leveraged to deal with the dysarthric acoustic data scarcity problem. To achieve this, we studied whether there are correlations between the shapes of dysarthric speech samples pronouncing the same word; such correlations could be learned via shape detectors such as CNNs.

During this investigation we realized that some correlations exist in voicegrams extracted from different dysarthric utterances of the same word. A voicegram is the visualization of spectrum frequencies and their dynamics over time, presented as a heat map. Fig. 1 depicts examples of waveforms and their voicegrams for two dysarthric speech samples, one pronouncing the word “Command” and the other “Word” by two different dysarthric speakers from UA-Speech. The colors seen on

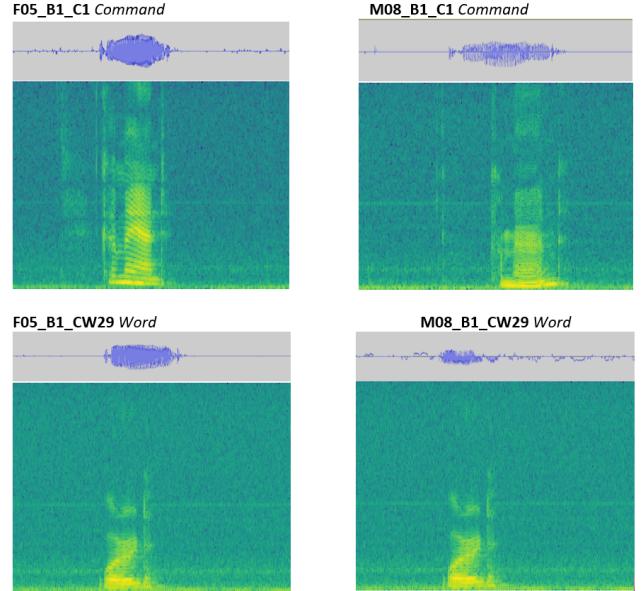


Fig. 1. Waveform and voicegram comparison.

the voicegrams in Fig. 1 indicate the intensity of the frequency at a specific time, with blue being low and yellow being high. As can be seen, the shapes created by the heat maps are similar for each word, while the same cannot be seen between the waveforms. These similarities are interesting: we expect to see similarities between phoneme-level voicegram shapes in normal speech. In contrast, since dysarthric phonemes tend to be inaccurate, their phoneme-level voicegrams have different shapes, which should have resulted in different word-level heat map shapes in return. However, this is not the case, as can be seen in Fig. 1.

Additionally, the presence of intense noise in utterance M08_B1_CW29 is obvious in the waveform of the signal but not in the voicegram that may facilitate the denoising tasks sometimes required to pre-process speech before fed to ASR. It is pertinent to note that the background noise (that exists in UA-Speech samples in varying degrees) is the reason for the greenish background instead of deep blue in Fig. 1 voicegrams.

Thus, unlike other prominent speech recognition systems in which recognizing phonemes is the core of their acoustic modeling, SV operates by extracting the speech features visually and learning to see the shape of the words pronounced by dysarthric speakers. To put it differently, instead of mapping speech signals to a sequence of phonemes and then use a language model and decoder to find the likelihood of the phoneme sequences presenting words, SV extracts word-level voicegrams for the given speech signals and visualizes them as RGB images highlighting the words’ shapes. Then, taking advantage of this visual representation, SV perceives dysarthric speech recognition as a word-shape-detector and uses our deep 2-dimensional Spatial Convolutional Neural Network (S-CNN) to learn and recognize the shapes of the words pronounced by dysarthric speakers.

B. Deep 2-Dimensional Spatial Convolutional Neural Network

Fig. 2 depicts the structure of the S-CNN. The input voicegram RGB images are resampled to 150×150 pixels before being fed to the network. The network is composed of eight convolutional layers in four sets, which each set is followed by a max pooling layer with pool size 2×2 . The convolutional layers started with applying 32 filters that were gradually increased to 256 filters in the last set.

Because the ratio of available data to the number of distinct classes (i.e., the vocabulary of 155 words) was quite small, SV faced a significant overfitting problem during our initial experiments. To resolve overfitting, Spatial 2D Dropout regularization [25] with a relatively large drop rate was applied in each set of convolutional layers to minimize the network's memorability and emphasize on generalizability instead. The main difference between standard dropout and spatial dropout is the latter drops the entire 2D features maps extracted from the voicegrams instead of individual pixels selected randomly. However, the last max pooling layer is followed by a standard dropout layer. The drop rate for all layers was set to 50%.

The activation function for all convolution layers was ReLU, but the output layer was softmax to promote one-of-many classification. The loss function was Categorical Cross Entropy, and the optimizer was ADADELTA [25]. The training data were given to the network in batches of 256 samples. The optimizer needed to adjust 1,877,403 trainable parameters. The configuration and hyperparameters were selected via grid search in which multiple setups were trialed, and then the best performing one was selected. The network was designed in TensorFlow and Keras.

C. Voicegram Data Augmentation

Addressing the scarcity of dysarthric data problem, we utilized visual-data augmentation [26] to create more voicegrams based on the existing speech data available during training. In particular, using this technique, we were able to artificially increase the number of training samples by creating modified versions of the voicegrams. The new images were created by shifting the width of voicegrams, sheering, and zooming through them.

D. Synthetic Dysarthric Speech Data Generation

In addition to the data augmentation explained above, SV also adopts normal-speech generation techniques to reconfigure a text-to-speech system trained to produce synthetic dysarthric speech samples. We experimented with Tacotron 2 [27], Ryuichi Yamamoto's implementation of Deep Voice 3 [28], and DC-TTS [29]. The first two systems did not produce naturally sounded dysarthric speech as they either needed more dysarthric samples to learn acoustic characteristics of the speakers or required significantly longer training time. Hence, DC-TTS was selected for this task; the system was initially trained on control speech and then fine-tuned to produce dysarthric speech leveraging transfer learning and neuron freezing. Several configurations of DC-TTS were considered and used to produce synthetic dysarthric speech, then a Mean

Opinion Score (MOS) analysis was conducted to rate the naturalness of the generated dysarthric speech and how similar the generated speech was to the original dysarthric speaker. The configuration with the highest average MOS was then chosen to produce synthetic voicegrams, and in conjunction with the original dysarthric speech augmented voicegrams, used to train SV.

E. Transfer Learning

Transfer learning is a machine learning technique where learning in a new task is accelerated by transferring knowledge from a related task. In deep learning, transfer learning happens when the knowledge learned by a model trained on a different dataset, usually with sufficient or more training data, is re-trained for another dataset; however, some trainable hyperparameters and neurons are usually frozen to preserve the knowledge obtained from the original dataset [30]. Speech Vision utilizes this technique to maximize the benefits of the control data (i.e., speech data collected from normal speakers uttering the same vocabulary) and further address the scarcity of dysarthric data. During this procedure, SV learns the basic acoustic-visual features of words in the vocabulary and then refines its knowledge when presented with dysarthric voicegrams. This was achieved by initially training the S-CNN shown in **Fig. 2** using the speech data collected from 12 normal speakers provided by the dataset (explained in the next section), while speech samples collected from one speaker were employed to validate the control-model. During this initial phase, no impaired speech sample was given to the model.

While using control data to pre-train ASR is a common practice in dysarthric ASR, SV freezes some neurons to ensure the overall word-shape-knowledge acquired by the S-CNN is not forgotten during the dysarthric refinements. Once the control-model was trained, we proceeded with applying transfer learning and re-training the model with the impaired speech samples during which the synaptic weights and biases of the top three sets of convolutional neurons were frozen so that only the hyperparameters of the last set were adjusted to learn dysarthric speech patterns. Freezing layers informs the network's optimization algorithm not to alter or update the hyperparameters of the frozen neurons and forces the S-CNN to move towards minimizing the loss function by only adjusting the hyperparameters of the unfrozen neurons. In addition, the drop rate was increased to 70% to further protect the network from overfitting due to the limited availability of dysarthric voicegrams – this increase of the drop rate only reflected on the unfrozen layers since the rest of the layers were non-trainable. **Fig. 3** illustrates which layers were frozen during Speech Vision's transfer learning step. The blue layers were frozen, i.e., non-trainable, while the green layers were trainable.

IV. EXPERIMENTS

A. Materials and Participants

Produced by the University of Illinois, UA-Speech [9] contains speech samples collected from 19 dysarthric subjects

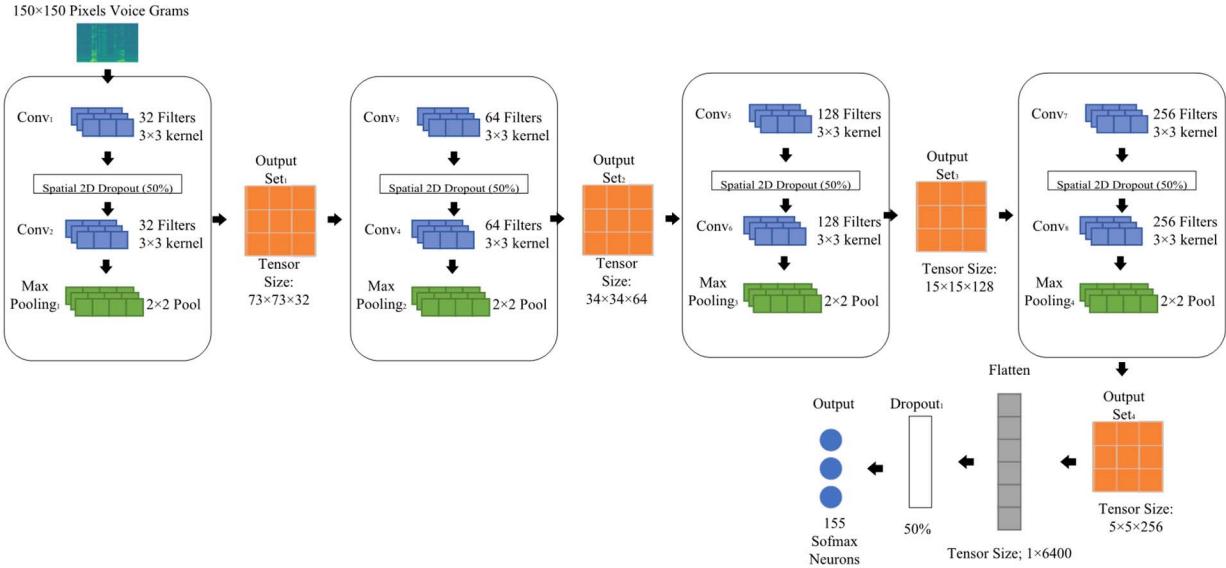


Fig. 2. The S-CNN architecture.

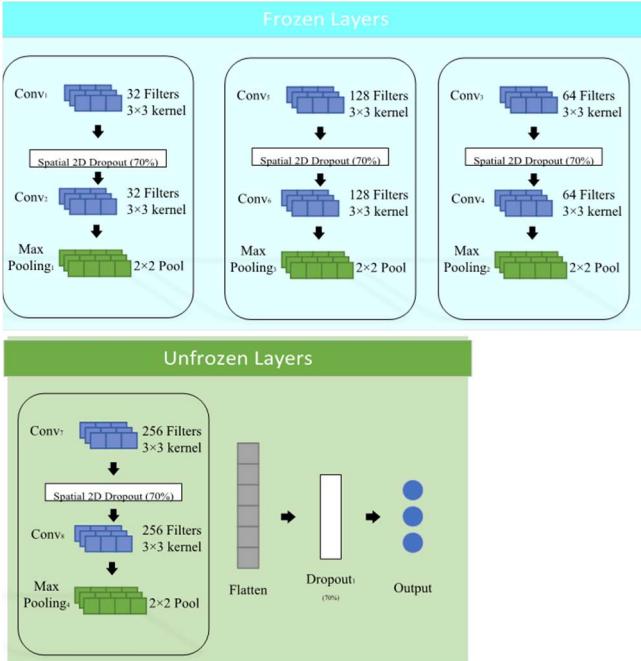


Fig. 3. Speech vision transfer learning.

with different speech intelligibility levels from as low as 2% to 95%. Speech intelligibility is the proportion of one's speech that can be comprehended by a normal listener, which is commonly used to present the severity of dysarthria in affected individuals [31]. Each subject was asked to utter a collection of commonly used words such as yes, no, up, down, etc., and each word was recorded individually. Furthermore, the dataset provides utterances of the same words obtained from 12 normal speakers as the control data. We utilized the utterances of 15 UA-Speech dysarthric speakers shown in Table I, and the vocabulary size was set to 155 words composed of the ten digits, 19 computer commands, 26 radio alphabets, and one hundred common words. However, the dataset did not supply

TABLE I
UA-SPEECH DYSARTHRIC PARTICIPANTS

Number	Participants	Gender	Age	Speech Intelligibility (%)	Intelligibility Level
1	M04	Male	>18	2	Very Low
2	F03	Female	51	6	
3	M12	Male	19	7	
4	M01	Male	>18	17	
5	M07	Male	58	28	Low
6	F02	Female	30	29	
7	M16	Male	40	43	
8	M05	Male	21	58	
9	M11	Male	48	62	Mild
10	F04	Female	18	62	
11	M09	Male	18	86	
12	M14	Male	44	90	
13	M10	Male	21	93	High
14	M08	Male	28	95	
15	F05	Female	22	95	

speech samples from the other four dysarthric speakers, so we could not consider them in this study.

The speech samples were recorded using a multichannel microphone array setup, and the speakers repeated each word three times; hence, three blocks of recordings per speaker are provided by UA-Speech. Each block contains seven wave files of the vocabulary words corresponding to a different recording channel. We applied blocks B1 and B1 audio samples (2179 utterances) to train each speaker model and then block B3 samples (1085 utterances) to test them.

B. Control-Model Training

Speech Vision was initially trained as a speaker-independent, whole-word ASR trained on the control data provided by eleven normal speakers from UA-Speech. The verification of this model was done by all utterances collected from UA-Speech control subject CM06, a male normal speaker. The best performing model delivered the training loss of 0.27 with 92% accuracy, and SI validation loss of 0.55 with

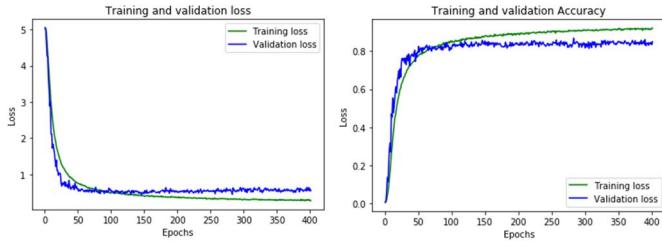


Fig. 4. Control-model training performance.

TABLE II
PROMPTS USED DURING THE MOS ANALYSIS

Number	Prompt	Set
1	"Well, he is nearly 93 years old."	A
2	"You wished to know all about my grandfather."	A
3	"When he speaks, his voice is just a bit cracked."	A
4	"Forward"	B
5	"Glitter"	B
6	"Oscar"	B
7	"These days a chicken leg is a rare dish."	C
8	"Rice is often served in round bowls"	C
9	"The juice of lemons makes fine punch."	C

85% accuracy given that CM06 speech was unforeseen for the model. This model was saved as the control-model and used to train the dysarthric SD models via the transfer learning procedure explained before. Fig. 4 portrays the training performance of the control model.

C. Generating Synthetic Dysarthric Speech and MOS Analysis

In order to identify the best configuration of DC-TTS transfer learning to generate synthetic dysarthric speech, we experimented with freezing different neural components of DC-TTS. DC-TTS is composed of multiple modules with various layers, such as convolution, highway, deconvolution, and CharEmbed layers. Each module's layers can be individually frozen, so their weights remained unchanged during the transfer learning procedure with the dysarthric speech samples.

After the initial screening, five different configurations were shortlisted, and for each configuration, three sets of phrases were generated, as shown in Table II. Set A contained three sentences that the model saw during training, whereas sets B and C had three prompts that were unforeseen for the model during training. Set B's prompts were all isolated words to evaluate the model's single word generation performance. The references to these prompts were supplied by TORG dataset [17] but omitted during training. The prompts from set C were selected from Harvard Sentences dataset [33] to gauge the model's performance across various sounds, but there was no corresponding dysarthric ground truth sample for these prompts as they were external to TORG.

Ten participants were invited to rate the generated dysarthric speech from each model configuration based on the prompts in Table II. As these participants had no dysarthric speech experience, a few samples from the original training data were played to help them better understand dysarthria. The participants were then asked to judge each generated synthetic dysarthric sample on two criteria: naturalness and similarity. Naturalness was defined for them as to how 'human' the

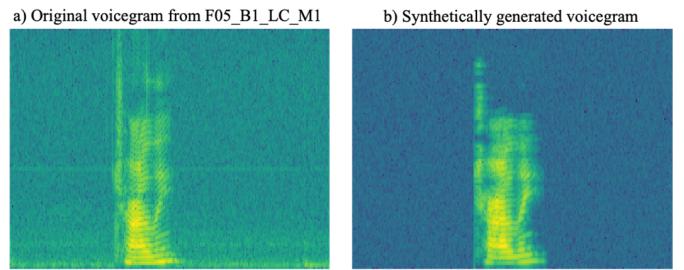


Fig. 5. Synthetic voicegram versus the original.

samples sounded, where a robotic sounded sample would score lower for naturalness. For 'similarity', the criterion was how the generated samples were similar to the original reference speaker (for sets A and B prompts). The generated sample should capture the typical speech characteristics of the speaker (i.e., pronunciation) as well as the dysarthric characteristics (for example, slurring, breathiness, and nasality). The ground truth samples for sets A and B were played alongside the generated samples to help with the comparison.

The best MOS scores were achieved when the last eight layers of each DC-TTS module were re-trained, but the rest remained frozen. Thus, this model was selected to generate the synthetic dysarthric speech samples from UA-Speech to train Speech Vision. Particularly, we re-trained DC-TTS for each dysarthric speaker in Table I separately, then used the trained model to produce synthetic speech for the entire 155 words in the vocabulary, adding one extra sample per word-speaker. Fig. 5 compares a sample of the generated voicegram with the original one for UA-Speech speaker F05 pronouncing "Charlie".

D. Speech Vision Dysarthric Training

Production of the dysarthric speaker adaptive models was done in two steps in which 30 SA versions of SV were produced. First, 15 SA versions of SV for each dysarthric speaker mentioned in Table I were generated, during which only the original data supplied by UA-Speech were used to train the models. Next, the generated synthetic speech data was also included in the training data in addition to the original blocks B1 and B2 dysarthric utterances, and then another 15 SA models were re-trained from the control-model and re-evaluated. This procedure was performed to measure the impact of the synthetic data on SV's performance. In both steps, the transfer learning procedure mentioned in section III.E was considered in which the control-model neurons were frozen, then the training of the model proceeded with the dysarthric speech data.

V. RESULTS AND DISCUSSION

A. Speech Vision Dysarthric SA ASR Results

Table III provides the testing results obtained from SV trained with each dysarthric speaker's data, including and excluding the synthetic data. As can be seen, the inclusion of the synthetically generated dysarthric speech improved SV's accuracy for 14 out of 15 dysarthric speakers comparing to those experiments in which only original utterances were used

TABLE III
DYSARTHRIC SPEECH VISION TESTING RESULTS

Intelligibility Level	Speaker	Word Recognition Accuracy (%)	
		Speech Vision (no synthetic data)	Speech Vision (with synthetic data)
Very Low	M04	10.94	17.03
	F03	32.47	36.5
	M12	40.65	46.78
	M01	28.39	34.32
<i>Very Low Intelligibility Average WRA (%)</i>		28.11	33.66
Low	M07	69.46	74.47
	F02	67.02	72.7
	M16	55.91	61.14
<i>Low Intelligibility Average WRA (%)</i>		64.13	69.44
Mild	M05	64.95	70.04
	M11	49.68	55.21
	F04	53.98	54.75
<i>Mild Intelligibility Average WRA (%)</i>		56.49	60
High	M09	85.16	87.31
	M14	86.45	88.15
	M10	89.68	90.21
	M08	86.67	88.71
	F05	94.41	93.3
<i>High Intelligibility Average WRA (%)</i>		88.47	89.54
Absolute Average WRA (%)		61.11	64.71

during training. The biggest improvement was for speaker M12 where an improvement of WRA 6.13% was achieved while the minimum improvement was 0.53% for speaker M10. WRA was defined as the ratio of correctly recognized words for each speaker model to the vocabulary size. The average improvements for each intelligibility levels were 5.54%, 5.31%, 3.8%, and 1.1% for very low, low, mild, and high intelligibility, respectively, which demonstrates better efficacy for dysarthric speakers with speech intelligibility of less than 60% with an average improvement of 5.4%. Specifically, severe dysarthria's improved performance is consistent with the similar experiments reported in the literature [15], [32]. Overall, using the synthetic dysarthric voicegrams delivered an absolute average WRA improvement of 3.60%.

Nonetheless, using the synthetic data did not significantly improve SV's performance for mild dysarthric severity with high speech intelligibility. In this category, the best WRA improvement was 2.15% for speaker M09, with an average improvement of 1.1%. The speaker who did not show any improvement was F05, classified as having mild dysarthria.

After listening to her speech samples provided by the dataset, we noticed that this speaker was highly intelligible and almost indistinguishable from normal speech. For this speaker, SV already achieved a 94.41% accuracy without the extra synthetic data; hence, we believe there was little to be gained by adding the additional generated samples. It is possible that the adverse effects of having synthetic utterances may have outweighed the benefits of adding the extra samples to the training set.

B. Performance Comparative Study

To select the baseline systems and benchmark SV's performance, the following conditions were set: 1) being whole-word and isolated speech dysarthric ASR, 2) the same dataset and speakers were considered, 3) a vocabulary size of at least 100 words was adopted, and 4) WRA was measured for each speaker. Thus, the dysarthric ASR systems in [12] (aka *Baseline #1*) and [33] (aka *Baseline #2*) were selected that can be directly compared with Speech Vision, although the vocabulary size in [33] was larger than SV and Baseline #1. Other dysarthric ASR systems did not satisfy these conditions; hence a direct comparison was either impossible or not informative due to the differences in the ASR tasks, datasets used, the performance criterion, or the vocabulary size. Fig. 6 plots the WRAs of these benchmark systems against SV results. It is noteworthy to mention that results from both speaker-dependent and speaker-adaptive dysarthric ASR experiments were provided in these baseline systems and included in our comparative study as well. However, since SV was initially trained on normal speech in an SI manner, we consider SV to be a speaker-adaptive ASR.

SV delivers the highest WRAs for 67% of the speakers compared to both SD and SA baseline systems. Between SD baseline 1 and 2, baseline 1 delivered better results for 13 of the speakers, but SV performed significantly better than SD baseline 1 with delivering higher WRAs for 73% of the speakers, improving baseline 1 SD accuracies by up to 260%. Nevertheless, baseline 1 speaker-adaptive ASR delivered the best results for moderate dysarthria (speakers with speech intelligibility 43% to 62%).

A comparison of the average WRAs across each speech intelligibility level is also provided by Table IV. Instead of mild intelligibility, SV provided the best average WRAs for the rest of the severity levels with an average 6.12%, 6.26%, 2.67% higher average accuracy for very low, low, and high intelligibility, respectively compared to the next best version of the baseline systems. With respect to the absolute average accuracy improvements, SV trained with the synthetic speech delivered the best performance than all baselines versions. SV performed better for mild intelligibility than baseline 2 but did not offer better results over baseline 1.

The better results delivered by SV are because SV does not perform phoneme recognition to recognize the words. Instead, the visual mapping of the speech signals to voicegrams and then to words enables the system to leverage data augmentation. Moreover, the differences between synthetic and original speech are masked by converting the data into voicegrams, which may further increase the synthetic data's effectiveness.

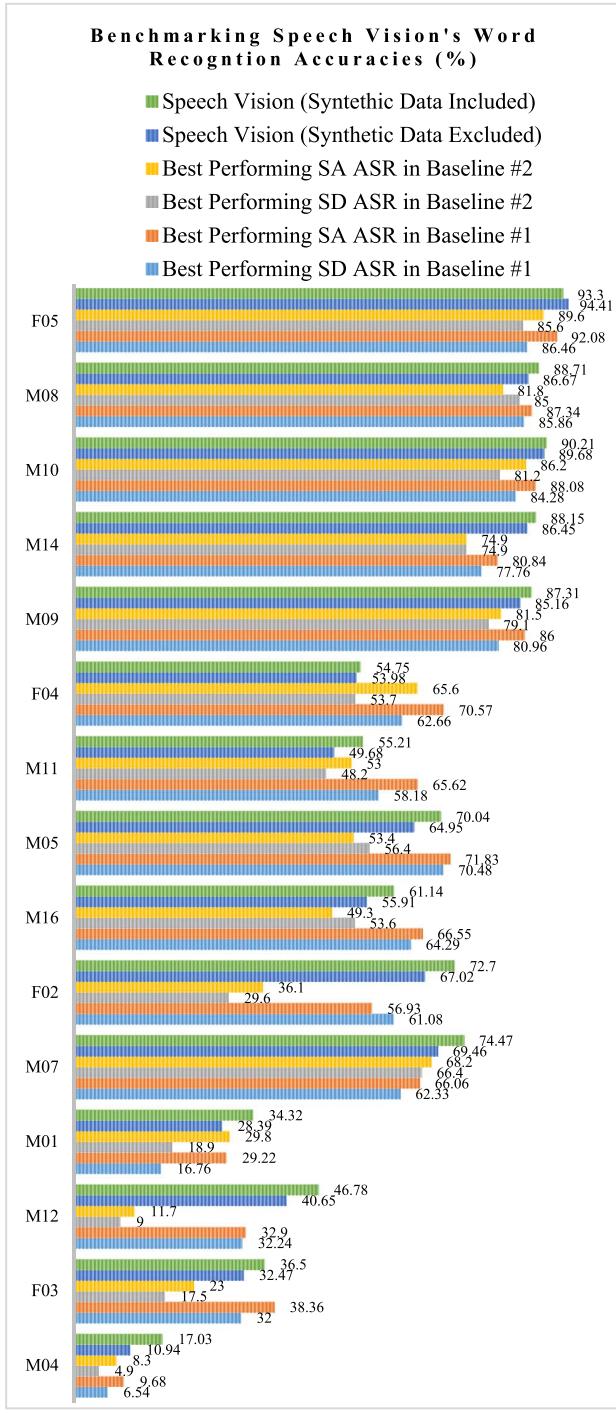


Fig. 6. Speech vision's WRAs versus baseline #1 and #2.

Likewise, using a set of convolutional neurons has been instrumental in the efficacy improvements offered by SV since convnets learn translation-invariant patterns in the input feature space in contrast to dense neurons that learn global patterns. In other words, after convnets learn a certain pattern in the input data, they can recognize it anywhere in the feature space, whereas a dense neuron will have to learn the pattern again if it appears at a new location. This feature of convnets makes them more data efficient, which is a desirable attribute given the limited availability of dysarthric acoustic samples. Another benefit of convnets is the facilitation of visual data augmentation built-in with most DL platforms; not only it

TABLE IV
BENCHMARKING AVERAGE WRAs

	Very Low Intelligibility Average WRA (%)	Low Intelligibility Average WRA (%)	Mild Intelligibility Average WRA (%)	High Intelligibility Average WRA (%)	Absolute Average WRA (%)
Best Performing SD ASR in Baseline #1	21.89	62.57	63.77	83.06	58.79
Best Performing SA ASR in Baseline #1	27.54	63.18	69.34	86.87	62.80
Best Performing SD ASR in Baseline #2	12.57	49.87	52.77	81.16	50.93
Best Performing SA ASR in Baseline #2	18.2	51.2	57.33	82.80	54.16
Speech Vision (Synthetic Data Excluded)	28.11	64.13	56.20	88.47	61.11
Speech Vision (Synthetic Data Included)	33.66	69.44	60.00	89.54	64.71

helps to achieve better learning dynamics when the training data is scarce, but it is also highly effective in maximizing the impact of learning translation-invariant patterns.

Additionally, since SV does not require labeled phoneme data to operate, it is more robust to the effects of the inaccuracies of dysarthric phonemes and difficulties in labeling them, making it easier for future studies to collect dysarthric data to increase the vocabulary and training size. In contrast, conventional dysarthric ASR may need to label the phonemes, which can be tedious yet insufficient.

VI. CONCLUSION

In this study, we identified three challenges in developing dysarthric ASR systems and proposed a system called Speech Vision that attempts to address them. The first challenge was due to the alternation and inaccuracy of phonemes in dysarthric speech, making conventional ASR systems less effective to recognize dysarthric speech. SV addresses this issue by converting word utterances into visual-feature representations and attempting to recognize the shape of the words instead of recognizing phonemes. The next issue was related to the unavailability of dysarthric speech samples, where SV facilitates by considering three measures: 1) applying visual data augmentation approaches to benefit from the translation-invariant learning feature of convnets; 2) generating synthetic dysarthric speech using state-of-the-art text-to-speech technologies to have extra training samples; and 3) utilizing transfer learning and neuron freezing to learn the basic word shapes from normal speech. The final issue we identified was the inaccuracy and difficulties with labeling dysarthric phonemes, which SV was immune to as it did not rely on phoneme data to recognize dysarthric speech.

Speech Vision was evaluated in two steps: initially, it was trained and tested by only considering the original data supplied by UA-Speech dataset, and then generating extra synthetic voicegrams for all vocabulary words and speakers and use them in addition to the original data for training. SV delivered absolute average WRAs of 61.11% and 64.71% for the two steps experiments, respectively. Moreover, in a

detailed comparison with other dysarthric speech recognizers verified with the same dysarthric speakers' data, SV outperformed them in recognizing mild and severe dysarthric speech achieving state-of-the-art results.

Nevertheless, the following limitations are identified, and future studies could investigate them to further improve SV's performance:

- Synthetic data generation produced the same output for the same prompt - we could only generate one additional sample per word for each speaker.
- SV did not deliver the best average WRA for moderate dysarthria.
- The S-CNN architecture could be improved by the inclusion of measures to minimize vanishing gradients and representational bottlenecks.

Speech Vision's source code is available from [34].

ACKNOWLEDGMENT

The author would like to thank Andrew Hu and Dhruv Phadnis for their help with respect to the synthetic data generation section of this project.

REFERENCES

- [1] M. Fernández-Díaz and A. Gallardo-Antolín, "An attention long short-term memory based system for automatic classification of speech intelligibility," *Eng. Appl. Artif. Intell.*, vol. 96, Nov. 2020, Art. no. 103976, doi: [10.1016/j.engappai.2020.103976](https://doi.org/10.1016/j.engappai.2020.103976).
- [2] C. Whillans, M. Lawrie, E. A. Cardell, C. Kelly, and R. Wenke, "A systematic review of group intervention for acquired dysarthria in adults," *Disability Rehabil.*, pp. 1–17, Dec. 2020, doi: [10.1080/09638288.2020.1859629](https://doi.org/10.1080/09638288.2020.1859629).
- [3] N. P. Narendra and P. Alku, "Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features," *Comput. Speech Lang.*, vol. 65, Jan. 2021, Art. no. 101117, doi: [10.1016/j.csl.2020.101117](https://doi.org/10.1016/j.csl.2020.101117).
- [4] Y. Zhao, M. Kuruvilla-Dugdale, and M. Song, "Voice conversion for persons with amyotrophic lateral sclerosis," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2942–2949, Oct. 2020, doi: [10.1109/JBHI.2019.2961844](https://doi.org/10.1109/JBHI.2019.2961844).
- [5] I. Calvo *et al.*, "Evaluation of an automatic speech recognition platform for dysarthric speech," *Folia Phoniatrica et Logopaedica*, pp. 1–10, Nov. 2020, doi: [10.1159/000511042](https://doi.org/10.1159/000511042).
- [6] M. Tu, A. Wisler, V. Berisha, and J. M. Liss, "The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance," *J. Acoust. Soc. Amer.*, vol. 140, no. 5, pp. EL416–EL422, Nov. 2016.
- [7] S. R. Shahamiri and S. S. B. Salim, "A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 5, pp. 1053–1063, Sep. 2014, doi: [10.1109/TNSRE.2014.2309336](https://doi.org/10.1109/TNSRE.2014.2309336).
- [8] S. R. Shahamiri and S. S. B. Salim, "Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach," *Adv. Eng. Informat.*, vol. 28, no. 1, pp. 102–110, Jan. 2014, doi: [10.1016/j.aei.2014.01.001](https://doi.org/10.1016/j.aei.2014.01.001).
- [9] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2008, pp. 1741–1744.
- [10] D. Ellis and N. Morgan, "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Mar. 1999, pp. 1013–1016, doi: [10.1109/icassp.1999.759875](https://doi.org/10.1109/icassp.1999.759875).
- [11] G. Saon and J.-T. Chien, "Large-vocabulary continuous speech recognition systems: A look at some recent advances," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 18–33, Nov. 2012, doi: [10.1109/MSP.2012.2197156](https://doi.org/10.1109/MSP.2012.2197156).
- [12] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *Proc. SLPAT, 6th Workshop Speech Lang. Process. Assistive Technol.*, 2015, pp. 65–71, doi: [10.18653/v1/W15-5112](https://doi.org/10.18653/v1/W15-5112).
- [13] N. Rajeswari and S. Chandrakala, "Generative model-driven feature learning for dysarthric speech recognition," *Biocybernetics Bio-med. Eng.*, vol. 36, no. 4, pp. 553–561, 2016, doi: [10.1016/j.bb.2016.05.003](https://doi.org/10.1016/j.bb.2016.05.003).
- [14] B. Vachhani, C. Bhat, B. Das, and S. K. Kopparapu, "Deep autoencoder based speech features for improved dysarthric speech recognition," in *Proc. Interspeech*, Aug. 2017, pp. 1854–1858, doi: [10.21437/Interspeech.2017-1318](https://doi.org/10.21437/Interspeech.2017-1318).
- [15] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. Interspeech*, Sep. 2018, pp. 471–475, doi: [10.21437/Interspeech.2018-1751](https://doi.org/10.21437/Interspeech.2018-1751).
- [16] K. Gurugubelli, A. K. Vuppala, N. P. Narendra, and P. Alku, "Duration of the rhotic approximant /r/ in spastic dysarthria of different severity levels," *Speech Commun.*, vol. 125, pp. 61–68, Dec. 2020, doi: [10.1016/j.specom.2020.09.006](https://doi.org/10.1016/j.specom.2020.09.006).
- [17] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORG database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, Dec. 2012.
- [18] C. España-Bonet and J. A. R. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *Advances in Speech and Language Technologies for Iberian Languages*. Cham, Switzerland: Springer, 2016, pp. 97–107.
- [19] N. M. Joy and S. Umesh, "Improving acoustic models in TORG dysarthric speech database," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 637–645, Mar. 2018, doi: [10.1109/TNSRE.2018.2802914](https://doi.org/10.1109/TNSRE.2018.2802914).
- [20] N. M. Joy, S. Umesh, and B. Abraham, "On improving acoustic models for TORG dysarthric speech database," in *Proc. Interspeech*, Aug. 2017, pp. 2695–2699, doi: [10.21437/Interspeech.2017-878](https://doi.org/10.21437/Interspeech.2017-878).
- [21] Y. Takashima, T. Nakashika, T. Takiguchi, and Y. Ariki, "Feature extraction using pre-trained convolutional bottleneck nets for dysarthric speech recognition," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 1411–1415, doi: [10.1109/EUSIPCO.2015.7362616](https://doi.org/10.1109/EUSIPCO.2015.7362616).
- [22] T. A. M. Celin, G. A. Rachel, T. Nagarajan, and P. Vijayalakshmi, "A weighted speaker-specific confusion transducer-based augmentative and alternative speech communication aid for dysarthric speakers," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 2, pp. 187–197, Feb. 2019, doi: [10.1109/TNSRE.2018.2887089](https://doi.org/10.1109/TNSRE.2018.2887089).
- [23] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, Oct. 1996, pp. 1962–1965.
- [24] B. F. Zaidi, S. A. Selouani, M. Boudraa, and M. S. Yakoub, "Deep neural network architectures for dysarthric speech analysis and recognition," *Neural Comput. Appl.*, pp. 1–20, Jan. 2021, doi: [10.1007/s00521-020-05672-2](https://doi.org/10.1007/s00521-020-05672-2).
- [25] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, [arXiv:1212.5701](https://arxiv.org/abs/1212.5701). [Online]. Available: <http://arxiv.org/abs/1212.5701>
- [26] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, [arXiv:1712.04621](https://arxiv.org/abs/1712.04621). [Online]. Available: <http://arxiv.org/abs/1712.04621>
- [27] K. Park and T. Mulc, (2018). *A TensorFlow Implementation of Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model*. [Online]. Available: <https://github.com/Kyubyong/tacotron>
- [28] R. Yamamoto. (2018). *Deepvoice3_Pytorch*. [Online]. Available: https://github.com/r9y9/deepvoice3_pytorch
- [29] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4784–4788, doi: [10.1109/ICASSP.2018.8461829](https://doi.org/10.1109/ICASSP.2018.8461829).
- [30] A. Van Opbroek, H. C. Achterberg, M. W. Vernooij, and M. D. Bruylants, "Transfer learning for image segmentation by combining image weighting and kernel learning," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 213–224, Jan. 2019, doi: [10.1109/TMI.2018.2859478](https://doi.org/10.1109/TMI.2018.2859478).
- [31] H. M. Chandrashekhar, V. Karjigi, and N. Sreedevi, "Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2880–2889, Dec. 2020, doi: [10.1109/tnsre.2020.3035392](https://doi.org/10.1109/tnsre.2020.3035392).
- [32] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6009–6013, doi: [10.1109/ICASSP.2018.8462290](https://doi.org/10.1109/ICASSP.2018.8462290).
- [33] H. Christensen *et al.*, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2012, pp. 1776–1779.
- [34] S. R. Shahamiri. (2021). *Speech Vision*. [Online]. Available: <https://github.com/rshahamiri/SpeechVision>