Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University
Published Online ***Vol 5 Issue 1, January 2023***
(https://fcihib.journals.ekb.eg)

Re-Talk: Automated Speech Assistance for People with Dysarthria

Salma Khaled Ali[1,*], Aya Mohamed Hassan[1], Ahmed Mohamed Salah[1], Donia Wael Mohamed[1], Muhamed Mustafa[1], and Ensaf Hussein Mohamed[1]
[1]*Department of computer science, College of computer science and Artificial intelligence, Helwan University, Egypt*
salmakhaled708@gmail.com , ayhassan@nu.edu.eg , ahmed.mohamadsalah@gmail.com , doniawaelsamir@gmail.com , muhammedmuustafa@gmail.com , ensaf_hussein@fci.helwan.edu.eg

***Abstract*— Dysarthria is a speech motor disorder where the muscles responsible for speech production, such as in the face, mouth, or respiratory system, have trouble coordinating and controlling themselves. Our research goal is to help individuals with dysarthria communicate effectively. Often, physical conditions make it challenging for them to express their thoughts through writing. Our research introduces an automatic speech assistant solution, consisting of two main parts: speech recognition and auto-correct. The speech recognition component takes the person's distorted speech as input, converts it to text, and then sends it to the auto-correct module to fix any mistakes or unclear words. We tested our model on both English and Arabic datasets. The English dataset showed a 50% Word Error Rate (WER) which was reduced to 40% after using the auto-correct module. Our results outperformed previous studies by 4.5%. However, the WER on the Arabic dataset was 80% which is not a satisfactory result, due to the limited size of the Egyptian Dialect Dysarthric Speech (EDDS) database.**

***Index Terms*— Speech Disorder, Dysarthric Speech Recognition, Bi-directional LSTM, CNN-LSTM, TORGO Database, UASpeech Database, Auto Correction, Noisy Channel, EDDS Database, Arabic ASR, Arabic Auto Correction.**

I. INTRODUCTION

Dysarthria is a speech motor disorder where the muscles used for speech production are damaged, paralyzed, or weakened. As a result, individuals with dysarthria struggle to control their tongue or voice box, leading to slurred speech [1].

Research on dysarthria has been ongoing for over two decades, with most of the literature focusing on addressing the irregularity of the acoustic signals produced by patients with the disorder. Recently, deep learning techniques such as convolutional neural networks, recurrent networks, and long short-term memory have been applied to the field of dysarthric speech recognition. The use of transformers, a type of neural network architecture developed for sequence transduction tasks such as speech recognition and text-to-speech transformation, has also become more widespread. The idea behind transformers is to handle the dependencies between inputs and outputs with attention and recurrence, as well as self-attention and feed-forward layers. Decoders in transformers have an extra layer of encoder-decoder attention to help the decoder focus on relevant parts of the input sequence. [2].

However, most of these studies have focused on the English language, with only a few addressing Arabic. Collecting data for Arabic language speech recognition has been a challenge due to the lack of available datasets. To overcome this, data was collected from hospitals and specialized clinics. Arabic speech recognition faces additional difficulties such as a high number of errors, such as the confusion of Arabic letters (e.g. like baa (ب) is uttered hamza - aaa(ء), faa(ف) is uttered hamza(ء), raa(ر) is uttered lam(ل), There are also instances where most letters are pronounced as hamza, making it difficult to identify the spoken word.

In our study, we worked with both Arabic and English datasets, using the same models for both. The only variation between the two datasets was in the preprocessing phase. Both datasets were trained on an Automatic Speech Recognition (ASR) model, followed by an Autocorrect phase. The Autocorrect phase in the English dataset used the Noisy Channel model at the word level, while in the Arabic dataset it was used at the contextual level (Bigram).

Our contribution in this paper is the improvement of the Bi-directional LSTM (BLSTM) in dysarthric speech recognition (DSR) from 44.5% [3]. This was achieved by applying an Autocorrect model to the DSR output, which improved the word error rate (WER) of the Automatic Speech Recognition (ASR) prediction. The result was promising, with a WER of 40%. We also collected the Egyptian Dialect Dysarthric Speech (EDDS) dataset, but the results were not as promising, with a WER of only 80%.

The rest of the paper is organized as follows: the second section will review related research and their findings. The third section will cover the datasets used in this paper, provide brief information about them, and outline the proposed models, including the system's process flow. The fourth section will display the results and performance of each method we tested, and determine the best approach. Finally, the last section summarizes and concludes the paper, and outlines future work.

II. RELATED WORK

This section explains previous research related to dysarthria and their outcomes. In the past, many studies utilized Gaussian Mixture Models (GMMs) to handle the distribution of speech waveform's spectral representation and Hidden Markov Models (HMMs) to manage the speech signal's sequential structure. The Support Vector Machine (SVM) approach was discovered to be resilient against missing consonants [4]. Nevertheless, this approach requires a large corpus for training, which is not obtainable for dysarthric speech [5].

In 2004, Alexander et al. [6] proposed a transformation system that is based on the principle of extracting from the input speech signal acoustic parameters, that are particularly important to speech intelligibility, modifying those parameters, and then synthesizing a new speech signal from them. Due to the vastly varied types of the illness, it is likely that any variant of the suggested system will only be effective for a certain subgroup of people with dysarthria.

In 2019, Feifei et al. [7] explored a method to non-linearly alter speech pace. They used an Automatic Speech Recognition (ASR) system that analyzed speech tempo at the phonetic level using a forced-alignment method from the conventional Gaussian Mixture Model Hidden Markov Model (GMMHMM). Instead of using time-domain signals, the estimated tempo modifications were applied directly to the acoustic properties. The experiments showed that adjusting typical speech towards dysarthric speech was more effective for data augmentation in personalizing dysarthric ASR training. This resulted in nearly a 7% improvement over the baseline speaker-dependent system evaluated using the UASpeech corpus. In recent years, research has shifted towards using Deep Neural Networks (DNNs).

In 2019, Jeremy [3] used a bi-directional LSTM encoder and Connectionist Temporal Classification (CTC) decoder to train the UA speech database and achieved a Phoneme Error Rate (PER) of 44.5% on the test set.

In 2020, Hussain and Alaa [8] found that plain Deep Neural Networks (DNNs) were not effective for dysarthric speech recognition, so they created a hybrid model (CRNN) by combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) and trained it on the samples from the TORGO database [9]. The results showed that adding a convolution layer to the standard RNN improved performance, outperforming the standard CNN, and had the potential to improve dysarthric speech recognition accuracy to 40.6%, compared to 31.4% for CNN.

In 2020, Mohammed et al. [10] proposed a new approach to enhance dysarthric speech recognition (DSR). As a preprocessing step, they utilized empirical mode decomposition and Hurst-based mode selection (EMDH) to improve the speech quality. The system was designed by combining the EMDH-based enhancement process with a convolutional neural network, resulting in improved performance compared to the baseline HMM-GMM and CNN systems.

In 2020, Sidharth [11] found that the encoding method for audio signals used as input for a deep learning model during training affects its performance. It was also determined that Mel Spectrogram is typically the best option for classifying Dysarthria.

In 2015, Stacey et al. [12] investigated the ability to detect dysarthric versus non-dysarthric speech and the impact of dimensionality in Mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs). Their results showed that repetition stuttering in dysarthric speech was correctly diagnosed at around 86% and 84% for non-dysarthric speech using MFCC and LPCC features, respectively. Non-speech sounds in dysarthric speech were also recognized with an accuracy of about 75%. Convolutional neural networks (CNNs) can extract useful local features from speech and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) can model the temporal dependencies of those features, allowing Convolutional LSTM Recurrent Neural Networks (CLSTM-RNNs) to capture the unique characteristics of dysarthric speech.

In 2018, Myung Jong et al [13]. evaluated the effectiveness of CLSTM Recurrent Neural Networks (RNNs) in recognizing dysarthric speech. They considered four types of CLSTM-RNN: Frequency domain (F-CLSTM-RNN), Time domain (T-LSTM-RNN), TFLSTM-RNN, and Parallel Time Frequency (PTF-LSTM-RNN). They conducted multiple studies on 18 speech data sessions from 9 Amyotrophic Lateral Sclerosis (ALS) patients, evaluating the Phoneme Error Rate (PER). The results showed that CLSTM-RNN significantly outperformed CNN and LSTM-RNN. Among the CLSTM-RNN variants, TFCLSTM-RNN performed the best, achieving a PER of 30.6% for the average of all test sessions and 35.4% for the average of three speech intelligibility groups.

In 2017, Jun and Mingzhe [14] used Deep Belief Neural Networks (DBNs) to predict the distribution of dysarthric speech. They estimated the posterior probability of states in Hidden Markov Models (HMMs) using DBNs and developed the speech decoder in a continuous dysarthric speech recognition system using the Weighted Finite State Transducers framework. According to their findings, the suggested approach improved the accuracy of predicting the probability distribution of dysarthric speech's spectral representation.

In 2021, S R Mani et al. [15] proposed a transfer learning-based convolutional neural network model (TL-CNN) and converted audio samples to Mel-spectrograms to improve accuracy on the TORGO dataset. When compared to other machine learning models, the proposed TL-CNN achieved improved accuracy.

In 2021, Alim et al. [16] used time delay deep neural networks to evaluate the performance of speech recognition models for dysarthric speakers. They also studied the impact of combining a corpus of normal and dysarthric speech on the model's performance. The results showed that deep neural network structures with properly tuned hyperparameters produced excellent outcomes for dysarthria speech in both Mandarin and English.

In 2021, Brahim et al. [17] found that the CNN-based system using perceptual linear prediction features achieved an impressive 82% recognition rate, which represents an improvement of 11% and 32% over the LSTM- and GMM-HMM-based systems, respectively, compared to the widely used MFCC.

In 2014, Toru et al. proposed a feature extraction method using a Convolutive Bottleneck Network (CBN) [18]. The CBN creates a deep network by combining various layers, including a convolution layer, a subsampling layer, and a bottleneck layer. They believed that using the CBN for dysarthric speech feature extraction would mitigate the impact of unstable speaking styles caused by athetoid symptoms. The CBN-based method showed better results compared to the traditional feature extraction method.

In summary, previous work in the field of Dysarthric speech recognition (DSR) has been divided into two areas. The first area focused on preprocessing speech signals by analyzing their sequential structure and phonetic level tempo using GMMs and HMMs. Researchers also used techniques such as empirical mode decomposition and Hurst-based mode selection to improve Dysarthric speech. The field then shifted to using deep learning models, with a focus on hybrid models such as CNNs-RNNs and CLSTM-RNN. These models achieved promising results, with a WER of 58% and a PER of 30.6%. Bidirectional LSTMs performed well with a 44.5% PER. However, the limited size of existing Dysarthric databases has made it difficult to develop high-performing models without overfitting. As a result, there is a need for more research in DSR, particularly in the Arabic language, to take advantage of new NLP techniques and build a larger and better database.

III. METHODS AND MATERIALS

This section explains the dataset and the proposed model.

A. Dataset Used

In this paper, we examined both Arabic and English datasets, the details of each will be explained below.

1)English Dataset:

In this paper, we used two datasets, the UASpeech and the TORGO. The UASpeech dataset is used as the training set and consists of 90,000 words from 19 speakers with cerebral palsy. The speech materials include 765 isolated words per speaker, including 300 uncommon words and 3 repetitions of digits, computer commands, radio alphabet, and common words. Data was recorded using an 8-microphone array and one digital video camera. We classified the dataset into three classes based on the word error rate of each record and trained our auto-correction model on the references in the UASpeech database. The TORGO dataset is used as the validation set and was produced by seven dysarthric subjects, including 4 males, 3 females, and 1 with ALS, between the ages of 16 and 50 with dysarthria resulting from cerebral palsy. The dysarthric and non-dysarthric subjects were matched according to age and gender for comparison of acoustic and articulatory differences.

2)Egyptian Dialect Dysarthric Speech (EDDS)Dataset:

In this paper, we encountered a challenge in obtaining an Arabic dataset for dysarthric speech. However, with the assistance of the Egyptian Charity Organization "Resala," we were able to gather data from Egyptian patients diagnosed with cerebral palsy, multiple sclerosis, muscular dystrophy, ALS, and Down's syndrome. The dataset was collected from 4 female and 8 male patients with ages ranging from 8 to 35 years old. Additionally, normal speech data was gathered from 3 female and 2 male individuals as a reference for the model. All data was recorded with a text reference, and the dataset currently consists of 1052 records of dysarthric and normal speech, and the collection is still ongoing.

B. Proposed Model

The proposed model has four main sequential phases as depicted in Fig. 1: feature extraction, automatic speech recognition, auto correction, and text-to-speech. First, we apply noise reduction to the speech, then process it through our proposed pipeline to enhance the spoken words by converting speech to text with the ASR, then correcting the text through an auto correction noisy channel model. Finally, the text is converted back to speech.

1) Features Extraction Techniques:

We used Mel frequency cepstral coefficient (MFCC) to extract acoustic features from the audio clips, with a sample rate of 16000 and clip duration of 1000 milliseconds.

2) Automatic Speech Recognition (ASR):

In this paper, we utilized automatic speech recognition (ASR) to convert speech into text. ASR is a technology that allows computers to detect spoken language or utterances and take appropriate actions, with speech-to-text conversion being a common use case. STT applications are valuable for individuals with physical or neuromotor impairments, as they eliminate or reduce the need for manual input methods.

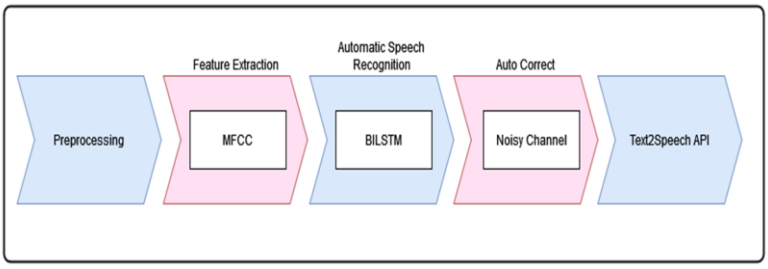

Fig. 1: Proposed Model Architecture

We conducted three experiments using a combination of Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory (BLSTM), and a pre-trained model, Wave2vec. The ***CNN*** consisted of four layers: convolution, max pooling, fully connected, and softmax. The convolution and pooling layers extracted features while the fully connected layer mapped the final output. On the other hand, ***LSTMs*** are a type of Artificial Recurrent Neural Network (RNN) used in deep learning that were designed to address the vanishing gradient problem in training conventional ***RNNs*** and perform well in classifying, processing, and making predictions on time series data. ***BLSTMs*** extend the conventional LSTM by using forward and backward cells to run inputs in two different directions. ***Wave2vec*** is a fully convolutional network that takes raw audio as input and calculates a general representation suitable for feeding into a speech recognition system.

To train the model, we used a learning rate of 3e-4 and varied the number of BLSTM units, batch size, and number of epochs. To prevent overfitting, we employed batch normalization and dropout for each activation layer.

3) Auto Correction:

This phase of the research addresses the problem of misspelled words generated from the ASR module or wrongly pronounced by the speaker. We need an auto-correct model to correct these misspelled words. There are many ways to implement auto-correct, such as using deep learning models, jamspell, or other libraries, but we chose the Noisy Channel model by Jurafsky and H. Martin [21] due to its effectiveness in correcting words. We improved the model to also correct words within context in both English and Arabic languages. Noisy Channel uses Bayesian Inference, applying Bayes' rule, as shown in equations (1) and (2).

$$W^* = argmax\ P\ (W|X) \quad (1)$$

$$W^* = argmax\ P\ (P(X|W)\ P(W))\ P(X) \quad (2)$$

$$W^* = argmax\ P\ (W)P(X|W) \quad (3)$$

The Noisy Channel model [22] operates on the assumption that a correctly spelled word has been "distorted" and misspelled when passed through a noisy communication channel. It uses Bayes' rule for its inference and simplifies the calculations by dropping out the denominator P(X) in equation (3).

The language model, P(W), represents the probability that W appears as a word, while the channel model, P(X|W), represents the probability that X would be typed in a text when the author intended W. The channel introduces errors in the form of substitutions, deletions, insertions, and other changes to the letters, making it difficult to recognize the true word.

To correct these misspelled words, we will run every word in the UASpeech dataset through the Noisy Channel model to identify the closest candidate words. We will create a dictionary for the model after feeding it with all the words in the dataset. Then, the misspelled word is passed to the model to correct it by making an edit distance of one or two steps at most through deletions, insertions, replacements, or transpositions of one or two letters. The list of words is then passed through a function to select only the true words, which are candidates for correction. Finally, the model chooses the most probable correction by calculating the probability of all the candidates and selecting the one with the highest probability.

The implemented model is divided into four parts for simplicity:
(a) Selection: The model selects the most likely correction by finding the candidate with the highest combined probability.
(b) Candidate Model: The candidate model generates correction suggestions by calculating the edit distance (number of letter changes needed to correct the word), which can include deletion, insertion, replacement, or transposition of one or two letters.
(c) Language Model: The language model calculates the probability of each word appearing by counting how often it appears in the dataset.
(d) Channel Model: The channel model determines the most likely correction by considering the edit distance of each candidate word. Words with an edit distance of 1 are considered more probable than those with an edit distance of 2, and words with an edit distance of 0 (already correct words) are returned as the most likely correction.

IV. PERFORMANCE EVALUATION AND EXPERIMENTS RESULTS

The results and performance of the model are explained in this section.

A. Performance Evaluation Metrics

The metric used to evaluate performance is Word Error Rate (WER), which calculates the ratio of errors in a transcript to the total number of words spoken.

$$WER = \frac{Substitution + Deletion + Insertion}{No.of\ words\ in\ the\ reference} \quad (4)$$

B. Experiments and Results

Three different models were experimented in the ASR module: Convolutional BLSTM, Bidirectional LSTM, and wav2vec. We described their structure in the previous section and attempted to optimize the parameters to achieve the best results. Our auto correct model was then applied to improve performance. We compared the accuracy of our model to the Bidirectional LSTM model of Jeremy [3], which uses BLSTM with specific parameters and the CTC loss function. Additionally, we compared our model to the models of Hussain and Alaa [8], which use a vanilla CNN model and compare it to a Hybrid RNN and CNN.

1) CNN - BLSTM model:

This model is a hybrid of two separate models, the CNN and BLSTM. It takes advantage of the local feature extraction capabilities of CNNs and the temporal modeling capabilities of LSTM-RNNs. The BLSTM component helps capture long-range temporal dependencies and overcome the vanishing gradient problem in traditional RNNs. This specific CBLSTM model used two BLSTM layers, each with 320 LSTM units, instead of a fully connected layer on top of two convolutional layers, two max pooling layers with a 0.2 dropout, and one softmax output layer. The result was a WER of 78%.

Model	WER
CNN – BLSTM without the noisy channel	78%
CNN – BLSTM + Noisy channel	68%
CRNN (Hussain and Alaa, 2020[8])	50.4%
CNN (Hussain and Alaa, 2020[8])	68.6%

Table 1: The baseline results are from Hussain and Alaa [8]. We adapted the same approach by experimenting a hybrid model with UAspeech database.

The hybrid model we proposed with a WER of 78% did not perform better than Hussain and Alaa's hybrid model, which had a WER of 50.4%. However, with the addition of auto correction, our results outperformed previous literature results for CNN models by 0.6% WER.

2) Wav2vec pre-trained model:

Wave2vec is a sequence-to-sequence (seq2seq) architecture model that consists of an encoder and a decoder. The encoder section of the model takes the input data and fine-tunes it for training before passing it on to the decoder section, which generates predictions and training results [20]. This model achieved a WER of 1.1%. We attempted to use this approach to enhance the accuracy and WER of the automatic speech recognition system. We believe that with a high-quality and extensive dataset, the results would be very promising.

3) BLSTM model:

In this study, we experimented with five BLSTM hidden layers, adjusting them with different batch sizes and hidden units. Our initial model had 50 epochs, 256 hidden units, and a batch size of 256, resulting in a WER of 68%. Our second model, which produced a WER of 58%, had 512 hidden units and a batch size of 512, and 80 epochs. The third model had 1024 hidden units, a batch size of 512, and 80 epochs, and produced the best results with a WER of 51%.

Comparing the performance of these models, we can conclude that the last BLSTM model was the most accurate with a WER of 51%. Our auto-correction model achieved a WER of 0.11, which improved the WER of the ASR models by 10%. There is a direct relationship between the accuracy of the ASR model and the auto-correction model, as the accuracy of the ASR improves, so does the performance of the auto-correction.

Finally, we applied our auto-correction model to the outputs of each of the models, resulting in improved performance. As the accuracy of the speech-to-text model increased, so did the accuracy of the auto-correction model.

In conclusion, the best performance for the last BLSTM model was a WER of 51%, which was improved to 40% WER with the addition of the auto-correction model. When comparing our model's performance (WER 40%) to that of Jeremy [4] (WER 44.5%), our model outperformed previous work by 4.5%.

Model	WER
BLSTM model + Auto correction (Re-talk)	40%
BLSTM model + CTC Loss function (Jeremy, 2019[3])	44.5%

Table 2: The baseline results are from Jeremy, 2019[3]. we used the same architecture and the UAspeech database but added the proposed auto correction model to enhance the performance.

4) Arabic ASR:

The top-performing technique from the English dataset was utilized to train the Arabic data. But, due to the insufficient size of the EDDS database, the model did not train effectively and achieved a WER of 80%.

V. CONCLUSION

We created an automated assistant for individuals with dysarthria to communicate with others by correcting their disordered speech and playing back the corrected audio. Although there were limitations with the ASR, it can be improved with more data and stronger training machines. We tested our BLSTM model on the English dataset and achieved a WER of 51%. With the addition of a noisy channel model to correct the faulty text, the final ASR WER was 40%.

However, when tested on the smaller Arabic EDDS dataset, the results were not as favorable and require more training data.

In the future, we plan to gather more Arabic data to train the model, and to explore the results of training larger datasets with Wave2vec. We aim to integrate the proposed model with Google Assistant to make it easier for patients to browse the web. Additionally, the application will have a therapist account for supervising patients and providing weekly progress reports. The model will also be connected to phone contacts so that when a patient calls, the recipient can understand them. This application will be accessible in emergency locations such as hospitals and police stations to aid patients in communication during emergencies.

ACKNOWLEDGMENT

We would like to extend our heartfelt gratitude to Resala Charity Organization for permitting us to gather the EDDS dataset and to our supervisor Dr. Ensaf Hussein for her constant support. Finally, we are grateful to Mr. Mark Allen for giving us access to the UASpeech dataset and allowing us to work on it.

REFERENCES

[1] Cleveland Clinic. (no date). Dysarthria & Speech: Symptoms, causes, treatments. Available at: https://my.clevelandclinic.org/health/diseases/17653-dysarthria#:~:text=Dysarthria%20is%20a%20motor%20speech,are%20strategies%20to%20improve%20communication. (Accessed: January 29, 2023).

[2] Chromiak, M. (2021). Exploring Recent Advancements of Transformer Based Architectures in Computer Vision.

[3] Cs230.stanford.edu. (2021). Dysarthric speech recognition using bi-directional LSTM. Available at: http://cs230.stanford.edu/projects_spring_2019/reports/18677882.pdf (Accessed 20 August 2021).

[4] Ieeexplore.ieee.org. (2021). Hmm-Based and SVM-Based Recognition of the Speech of Talkers with Spastic Dysarthria. Available at: https://ieeexplore.ieee.org/abstract/document/1660840 (Accessed 20 August 2021).

[5] Aclanthology.org. (2021). Automatic dysfluency detection in dysarthric speech using deep belief networks. Available at: https://aclanthology.org/W15-5111.pdf (Accessed 20 August 2021).

[6] Kain, A., & Niu, X. (2004). Formant Re-Synthesis of Dysarthric Speech. Available at: https://www.researchgate.net/publication/2905808_Formant_Re-Synthesis_Of_Dysarthric_Speech (Accessed 20 September 2022).

[7] Xiong, F., Barker, J., & Christensen, H. (2019). Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalized Dysarthric Speech Recognition. 10.1109/ICASSP.2019.8683091.

[8] Oaji.net. (2021). Dysarthric Speech Recognition using Convolutional Recurrent Neural Networks. Available at: http://oaji.net/articles/2020/3603-1603768199.pdf (Accessed 20 August 2021).

[9] Researchgate. (2021). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Available at: https://www.researchgate.net/publication/225446742_The_TORGO_database_of_acoustic_and_articulatory_speech_from_speakers_with_dysarthria (Accessed 20 August 2021).

[10]Asmp-eurasipjournals.springeropen.com. (2021). Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. Available at: https://asmpeurasipjournals.springeropen.com/track/pdf/10.1186/s13636-019-01695.pdf (Accessed 20 August 2021).

[11]Cs230.stanford.edu. (2021). Deep learning-based detection of dysarthric speech disability. Available at: http://cs230.stanford.edu/projects_winter_2020/reports/32533441.pdf

[12] S. Oue, R. Marxer, and F. Rudzicz, "Automatic dysfluency detection in dysarthric speech using deep belief networks," available online: https://aclanthology.org/W15-5111.pdf, accessed Sep. 20, 2022.

[13] Myungjong Kim, Beiming Cao, Kwanghoon An, and Jun Wang, "Dysarthric Speech Recognition Using Convolutional LSTM Neural Network," in Interspeech, 2018, pp. 1–4, doi: 10.21437/interspeech.2018-2250.

[14] J. Ren and M. Liu, "An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks," available online: https://thesai.org/Downloads/Volume8No12/Paper_7An_Automatic_Dysarthric_Speech_Recognition_Approach.pdf, accessed Sep. 20, 2022.

[15] S.R. Mani Sekhar, G. Kashyap, A. Bhansali, A.A. Abishek, K. Singh, "Dysarthric-speech detection using transfer learning with convolutional neural networks," ICT Express, vol. 8, no. 1, pp. 61–64, 2022, doi: 10.1016/j.icte.2021.07.004.

[16] A. Misbullah, H.-H. Lin, C.-Y. Chang, H.-W. Yeh, and K.-C. Weng, "Improving Acoustic Models for Dysarthric Speech Recognition using Time Delay Neural Networks," in 2020 International Conference on Electrical Engineering and Informatics (ICELTICs), 2020, pp. 1–4, doi: 10.1109/ICELTICs50595.2020.9315506.

[17] B.F. Zaidi, S.A. Selouani, M. Boudraa, et al., "Deep neural network architectures for dysarthric speech analysis and recognition," Neural Comput & Applic, vol. 33, pp. 9089–9108, 2021, doi: 10.1007/s00521020-05672-2.

[18] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric Speech Recognition Using a Convolutive Bottleneck Network," available online: http://www.me.cs.scitec.kobe-u.ac.jp/takigu/pdf/2014/0099.pdf, accessed Sep. 20, 2022.

[19] "Dysarthric speech database for universal access research," available online: https://www.researchgate.net/publication/221481038_Dysarthric_speech_database_for_universal_access_research, accessed Aug. 20, 2021.

[20] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Proceedings of the Conference on Neural Information Processing Systems, 2020. Available online: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07Paper.pdf. [Accessed 25 September 2022].

[21] D. Jurafsky and J. H. Martin, "Spelling Correction and Noisy Channel," Web.stanford.edu, 2021. Available online: https://web.stanford.edu/jurafsky/slp3/B.pdf. [Accessed 19 October 2022].

[22] "How to Write a Spelling Corrector," Norvig.com, 2021. Available online: http://norvig.com/spell-correct.html. [Accessed: 03-Jul-2021].