Dysarthric Speech Recognition: From Impaired to Understandable

Karima Kadaoui
karima.kadaoui@mbzuai.ac.ae

1 Introduction

Speech impairment is a life-altering disorder that prevents patients from communicating with their environment. In 2012, around 10% of the adult US population reported suffering from one [49]. Speech disorders impact a person's employment, self-esteem and quality of life, and can come hand in hand with other conditions such as ALS, strokes and Parkinson's disease [3].

Dysarthria is the technical term to refer to disorders that negatively impact the physical production of speech in the form of weakened speech muscles (lips, tongue, jaw etc.) [56] Dysarthric speech has different characteristics that make it hard to understand. It is "between 10 and 17 times slower than regular speech" [56] which increases inter-speaker variability [62]. It includes involuntary repetitions, pauses [38] and is either monotonic or has high pitch and loudness variation [16] [39]. People with dysarthria also find difficulties with consonants and consonant clusters, especially the ones occurring in word starts [61].

Unfortunately, impaired speech is usually accompanied by additional motor disabilities [65]. It is thus more crucial for patients to be able to communicate with their surroundings to get their needs met. Augmentative and Alternative Communication (AAC) [2], which refers to ways to communicate that are not reliant on talking, such as using images, computers, writing or hand gestures, is used to bridge this communication gap. An AAC's capacity, however, can be limited in terms of the number of words or messages that can be expressed and is also much slower compared to speech (around 8 to 10 words/minute) [26]. Additionally, in [23], dysarthric speakers have expressed that they find making voice commands less tiring that using a scanning interface (where they can scroll and select their command), even when the scanning interface has a higher accuracy. Certain disabilities like Parkinson's disease might even prevent the correct use of AACs altogether. An Automatic Speech Recognition (ASR) solution is therefore a better alternative.

ASR for dysarthric speech still suffers from lower performance compared to healthy speech systems due to variability at both inter-speaker and intra-speaker levels, in addition to a lack of impaired speech data [51] [60].

In this report, we will explore the performance of state-of-the-art (SOTA) ASR models when applied to dysarthric speech.

2 Literature Review

According to [11], there are 2 techniques to improve the recognition of impaired speech: *dysarthric speech enhancement* and *speaker adaptation-based ASR*.

2.1 Dysarthric Speech Enhancement

Dysarthric speech enhancement refers to acoustically transforming dysarthric speech to be closer to normal speech through altering formants, tempo, F_0, vowel space among other aspects: [54] found that a reduction in variation in F_0 is correlated with a reduction in intelligibility, while [28] studied alterations to spectral content, formants and prosody, applying transformations on both spectral and temporal domains to match typical speech characteristics. In [57], phase vocoding was used to adjust dysarthric speech tempo in addition to transformations to the spectral domain and pronunciation correction at the phoneme level. Vowel duration manipulation has also been explored in [36].

[55] identified some differences between dysartric and typical speech from the TORGO [58] dataset and employed some techniques to correct them:

- *Voicing unvoiced consonants:* Consonants can be characterized as either voiced or unvoiced. Voiced consonants are characterized by the vibration of the vocal cords during the articulation. The unvoiced ones only involve passing air and restricting it. Most of the consonants in English are organized in pairs of voice and unvoiced, having the same tongue and mouth position with voicing being the only difference (e.g. unvoiced: /t/, voiced: /d/).
  Dysarthric speech can have up to 27.2% of the unvoiced consonants replaced by their voiced counterparts [55]. The vibrations of the vocal cords when pronouncing voiced consonants result in an energy concentration under 150 Hz in the speech signal, called the voice bar. Removing the voice bar with a high-pass Butterworth filter [13] can correct the wrong pronunciation.
- *Inserted phonemes:* When phonemes are involuntarily added to speech (e.g. repetitions of the same syllable), the segment of that phoneme is removed completely when it's surrounded by silence. When it's not the case, phonemes that are supposed to be adjacent are merged together using a technique called Pitch Synchronous OverLap and Add (PSOLA) [14].
- *Deleted phonemes:* When the speech is missing some phonemes, they are added to it after being extracted from an associated synthesized speech (generated with a text-to-speech algorithm using the annotation of the dysarthric one).
- *Time differences:* Speakers with dysarthria speak slower, making the phoneme durations longer. [55] thus shorten those phonemes using a digital short-time Fourier analysis-base phase vocoder [52], since it can contract the needed segments without altering the frequency.
- *Frequency differences:* Atypical speech is also characterized by a different vowel space due to different formant trajectories (formants are the resonance frequencies of the vocal tract. They are relevant to speech and vowel differentiation). To make the vowels more typical, a Gaussian mixture mapping conversion function is learned to substitute the original formants with ones that are known to be associated with the target vowel.

In order to evaluate the impact of these transformations on intelligibility, listeners were given 20 randomly selected sentences ensuring that there is a minimum of two sentences per transformation. They were then asked to write transcripts for the given audios. Results showed that missing phoneme insertion and the deletion of involuntary extra phonemes benefited intelligibility the most with +7.8% of correctly identified phonemes. The time and frequency transformations, however, did not show improvement.

2.2 Speaker Adaptation-Based ASR

Speaker adaptation-based ASR, on the other hand, relies on adapting a model to a specific speaker:

- *Speaker Independent models (SI)* are models trained on a big number of speakers. They have better accuracy for unseen speakers compared to adapted models, but generally have a worse performance overall.

- *Speaker Dependent models (SD)* are models trained on a single speaker. They have the best performance if tested on the same speaker, but are not very useful since their generalization ability is extremely limited.

- *Speaker Adapted models (SA)* are SI models that were then fine-tuned on a particular speaker. They need relatively little fine-tuning data and have important accuracy improvements over SI models. In addition to that, speaker adaptation has a low cost.

Speaker adaptation methods include Maximum a posteriori (MAP) Adaptation [31], Cluster Adaptive Training (CAT) [18] and Maximum Likelihood Linear Regression (MLLR) [44]. Selecting an adequate initial model to adapt plays a huge role. A technique introduced by [32] involves interpolating an SD background model, that represents the characteristics of a certain dysarthric speaker, with regular SI models before applying MAP adaptation. [59] used both MLLR and MAP in a hybrid speaker adaptation, while [12] employ a multi-taper spectral estimation along with feature space MLLR (fMLLR) while exploring the usefulness of jitter (F_0 variation between cycles) and shimmer (amplitude variation).

Vocal Tract Length Normalization (VTLN) [34] is another speaker adaptation method. Its premise is learning a warping function whose role is speech features normalization. Although it is robust and efficient (a single parameter has to be estimated for each speaker), the function needs to be manually designed.

[8] introduced a technique to adapt hybrid Neural Network - Hidden Markov Models (NN-HMMs). They insert an adaptation network (made of fully connected layers) under the model to learn feature vectors called speaker codes. It learns one for each speaker using only their specific data. The speaker code for the target speaker is then fed to the model and is responsible for controlling the way the speaker's data is transformed to a general speaker-independent feature space. The size of the speaker codes can also be adjusted depending on how much adaptation data is available.

This technique however, has poor performance with CNNs. [9] theorized that it could be due to the CNN's max-pooling making it harder to learn good adaptation weights or because of the CNN managing to normalize the differences between the speakers.

They thus placed the adaptation network above the pooling layers and adapted their outputs instead of adapting input speech features. In addition to that, they learned each speaker's nodes output weights solely from that specific speaker's adaptation speech, considerably reducing the amount of weights to be learned for each speaker.

As adaptation is performed on a model, the weights learned originally change and the model might end up not performing well anymore on the initial data it has been trained on. In order to mitigate the issue, [67] added a regularization term to the cost function. Since the goal is to keep the adapted model's distribution close to the one of the originally trained model, adding the Kullback-Leibler (KL) divergence [42] as a regularizer prevents said model from overfitting on the new data.

2.3 Data Augmentation

Despite the inherent difficulty of the task, the lack of dysarthric data only complicates the problem further. Due to the physical strain that comes with speaking for dysarthric patients, it is not easy to gather a dataset that is comparable in size to the standard speech ones.

Data augmentation thus comes handy. In addition to the conventional approaches mentioned, [63] introduced for the first time, a data augmentation technique specific to atypical speech. They used a fixed factor to alter the tempo and timing of standard speech to make it sound like dysarthric speech. [66] built on this work by making the fixed factor adjustable and tailored to each speaker.

[33] used Deep Convolutional Generative Adversarial Networks (DCGAN) [53] to explore adversarial data augmentation. A DCGAN learns features specific to a dysarthric speaker and perturbs standard speech to apply changes in volume and speech rate, in addition to other fine-grained aspects of

atypical speech. A comparison with 7 other augmentation methods showed that when using DCGAN augmentation, WER reaches 25.89% with a relative decrease that ranges between 11.61% and 0.48%.

The DCGAN, however, does not affect the length of the generated speech and speech rate has to be modified separately. To avoid that, another study [29] on dysarthric data augmentation used a Voice Transformer Network (VTN) [30] to transform standard speech, along with a Transformer-based Voice Conversion (VC) model to reistablish the identity of the dysarthric speaker that was used in the transformation. The evaluation in [29] did not involve augmenting data that was then used to train a model and thus did not supply WER scores. They argued that a WER score does not accurately evaluate whether dysarthric aspects are correctly introduced to the standard speech. Hence, their evaluation consisted of asking human listeners to determine the nature of a mix of generated and real dysarthric audios. They have found, however, that the higher the severity of the dysarthria, the more listeners consider the speech to be generated, even when it's not the case. There is therefore a difficulty in differentiating between naturalness and intelligibility.

2.4 Intelligibility Classification

One helpful task in the field of dysarthric ASR is intelligibility classification. Other than its use for diagnosis and disease progression monitoring, learning to categorize the severity level of the dysarthric speaker can help to build tailored models for each one of those categories.

[17] experimented with Support Vector Machines (SVMs) [15] and Attention-enabled Long Short Term Memory networks (LSTMs) [27]. For the SVM, hand-crafted acoustic features had to be extracted from the data, while the LSTM was directly fed spectrograms. Thanks to the attention mechanism, the LSTM had the ability of modeling each speech frame's contribution to the classification output. The SVM-based system showed a maximum accuracy of 61.05% while the LSTM one achieved 76.97%.

One issue with this method, however, is that attention makes the model very complex and prevents it from learning correctly due to the very limited dysarthric data available. [19] thus investigated to find a way to compute the weights of the heavy Weight Pooling (WP) mechanism before the training stage. They managed to get the weights through saliency pooling using the saliency model from [37]. The role of this model is to determine what parts of the audio signal contribute the most to the comprehension of the message. It thus constitutes a replacement for attention given that saliency is essentially attention but deployed by noticeable stimuli, in a bottom-up approach. The system achieved an error reduction of 57.56% compared to the SVM and 20.34% compared to the attention-based LSTM [19].

3 Data Exploration and Pre-processing

3.1 Datasets

3.1.1 The Nemours Database of Dysarthric Speech

Created by the Applied Sience Engineering Laboratories (ASEL) [1] at the University of Delaware, the Nemours database [48] comprises 814 short nonsense sentences in total. 74 sentences were spoken by each of 11 male speakers with impaired speech and 1 healthy control speaker. In addition to those sentences, the participants provided 2 paragraphs of connected speech. The audio was sampled at 16 kHz.

We contacted the authors of the database, who gave me an email address to contact. We have not gotten a reply.

3.1.2 The TORGO Databse

The TORGO database [58] was collected by the Department of Computer Science [6] of the University of Toronto. It contains speech data sampled at 16kHz from 4 male and 3 female dysarthric patients in the 16-50 age range. The same data was also collected from control speakers that match the dysarthric ones in number, gender and age. Each participant recorded around 500 utterances resulting in 3 hours of data. The data consists of non-words (e.g. /p-ah-t-ah-k-ah/ or "maintain 'eee' for 5 seconds"), digits, short words ("yes", "no", "up", "down" etc.), international radio, written sentences and unrestricted sentences where they were told to describe 30 images in their own words.

In addition to the audio data, the database includes 3D videos showing the articulatory movements that happen inside the vocal tract and out of it. Although these articulatory features could be helpful to train a speech recognizer, that kind of data will not be easily obtainable "in the wild" and will thus not be exploited. The data is published publicly.

3.1.3 The UASpeech Database

The UASpeech database [40] was put together at the University of Illinois. It includes 9.4 hours of speech from 15 dysarthric speakers (11 of which are male) and 4.8 hours from 13 (age-matched) control speakers sampled at 16 kHz. There are 765 isolated words, namely digits, commands, radio alphabet, 155 common short words and 300 uncommon long ones. Each speaker's data is split into 3 blocks that contain 155 common and a 100 uncommon words in total. Typically, the first and third blocks are used as for training and the second block serves as a testing split. Patients are classified across 5 intelligibility levels (very low, low, mid, high, control) which can come in handy because adapting models on a specific severity level can improve performance. We requested access to the database and were given credentials to download the data via ftp.

4 Methodology

4.1 Data Centric

ASR on dysarthric speech is already a tricky task on its own. In addition to that, data scarcity further hinders its development. The size of the exisiting dysarthric datasets is very far from the regular ASR datasets (such as LibriSpeech [50] with 960h of annotated speech). Furthermore, there is even less data in languages other than English. It is thus as imperative to work on gathering large-enough datasets in multiple languages as it is to develop new approaches to solve the task.

One of the main reasons for which collecting a dysarhtric dataset is hard, is the difficulty of the task for the patients. People with speech impairments find it very hard to talk and tire easily. They need to rest a lot during the recording sessions and they cannot go beyond a couple of hours at a time.

However, and contrary to a lot of other AI fields, the collection of speech data can be much easier to achieve. Sound doesn't need costly procedures to be recorded (the way medical images would be for example) and it can be done by anyone owning a device with a microphone. We thus decided to create a mobile application that gives a user with a speech impediment sentences to read and record before saving them to the database. This will allow for more data to be collected because users can record their sentences at their own pace. The goal is to have a dynamic continuously growing dataset as new users join and sentences get added. It is also important to record sentences that make sense (Nemours has non-sense sentences and UASpeech only has single words) because context can tremendously help in speech recognition, and it's important to train a model to use that.

4.2 Model Centric

4.2.1 Models

QuartzNet

The QuartzNet [41] model is made of blocks that contain one or multiple module of time-channel separable 1-D Convolutions, ReLU and batch normalization layer. It uses CTC loss [21] for training. It is similar to the Jasper model [45] but thanks to its 1-D Convolutions being time-channel separable, it is able to reduce the number of weights all while maintaining the same accuracy (Figure 1). Results on test-clean show 3.9% WER while test-other is of 11.28%.

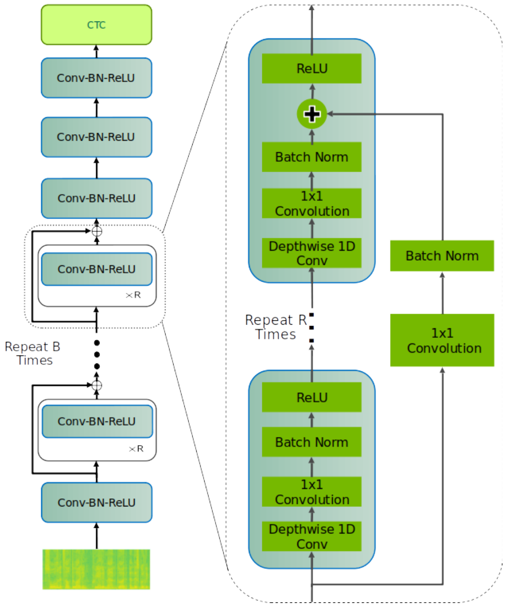


Figure 1: QuartzNet Architecture

Conformer

Although transformers have shown good results at capturing global features, they are not as successful with local ones. At the same time, Convolutional Neural Networks (CNNs) [43] are good at capturing those fine-grained features but require many layers to capture more global information. [22] combined the two to model both local and global representations simultaneously. Each Conformer block contains 2 half-step feed forward modules (in Macaron-Net [47] style), between which are a multi-headed self attention module and a Convolution module, followed by layer normalization (Figure 2).

[69] uses wave2vec 2.0 pre-training [10] with the Conformer architecture on Libri-Light [35] and Librispeech [50] for the unlabeled and labeled datasets respectively. Results show a 1.4% WER on the test-clean set and 2.6% on the test-other set.

Wav2Vec2

The Wav2Vec2 [10] model consists of a multi-layer convolutional feature encoder followed by a Transformer [64] that uses a convolutional layer that serves as relative positional embedding (instead of the usual absolute one) and is responsible for creating contextualized representations. Wav2Vec2 uses contrastive learning, where the same input goes through 2 different transformations and the model has to recognize pairs of transformations that resulted from the same input. The outputs of the feature encoder are thus also fed to a quantization module, and its outputs are compared against the output of the transformer to solve the contrastive task (Figure 3).

4.2.2 Experiments

In a first set of experiments, we used the Nvidia Neural Modules Toolkit [5] to fine-tune a QuartzNet 15x5 model and a Conformer CTC on the UASpeech corpus.

We then used the Huggingface [4] toolkit to fine-tune a wav2vec 2.0 model [10] (Figure 3) pre-trained on 960 hours of unlabeled speech audio from the Librispeech corpus sampled at 16kHZ.

Fine-tuning was performed using 3 different datasets. In one of the Wav2Vec2 experiments, the UASpeech corpus was used directly to train the model to decode the acoustic features into their corresponding characters (since it was only pre-trained on unlabeled data). All the utterances of the

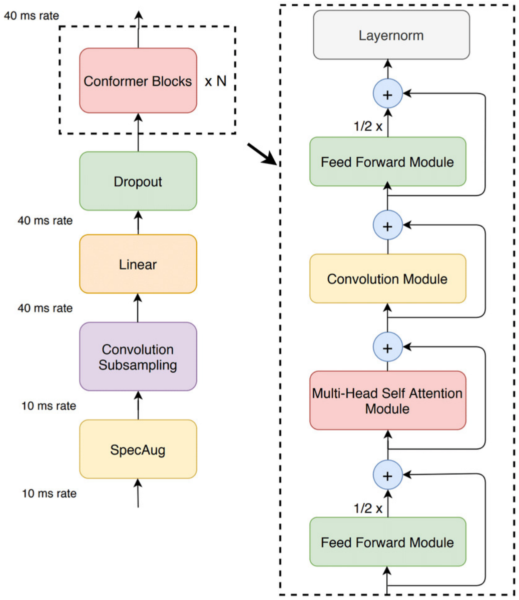


Figure 2: Conformer Architecture

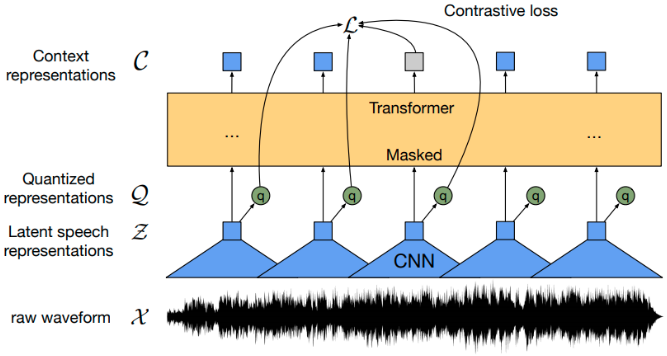


Figure 3: Wav2Vec2 Architecture

15 dysarthric speakers were used during training. No data from the healthy control speakers was used. Similar to other papers [68] [24] [46], blocks 1 and 3 of each speaker were used as the training set, while the second block served as the test set. Another experiment involved the TIMIT dataset [20] (healthy speech) and was first used before fine-tuning on The UASpeech corpus. Finally, data from an accented English dataset called VCTK was used between the TIMIT and UASpeech fine-tunings.

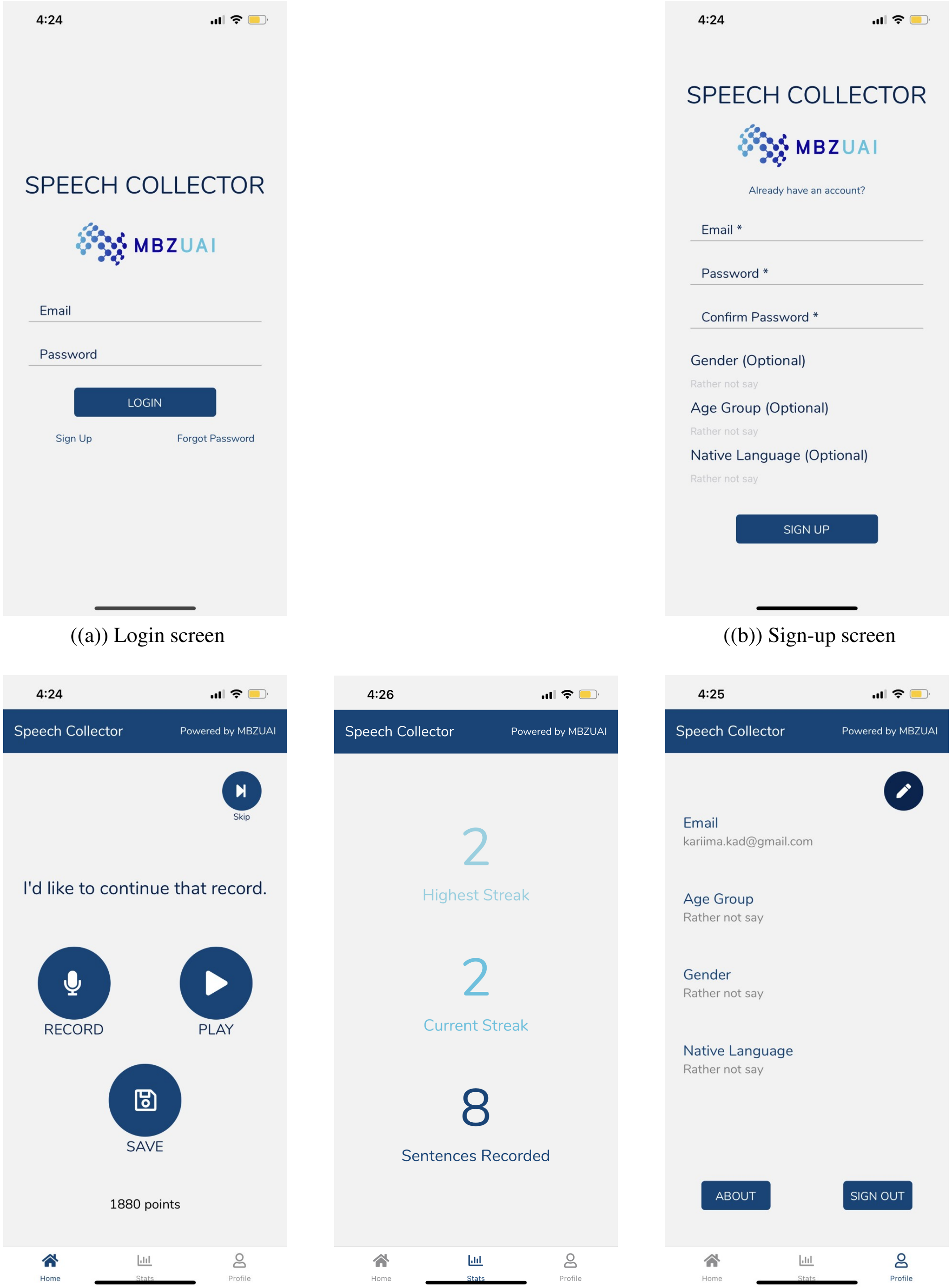
Previous literature trimmed silences in the beginning and ends of the UASpeech utterances but we have left them to mimic "in-the-wild" conditions, where people with speech impairments tend to insert long silences in their speech or start later than healthy subjects when prompted.

Due to the limited resources, we were only able to run the smallest versions of the models with a maximum batch size of 16.

5 Discussion of Experiments and Results

5.1 Data Centric

An APK was developed for the app and it is now being alpha tested before being deployed (Figure 4).



Gender (Optional)
Rather not say

Age Group (Optional)
Rather not say

Native Language (Optional)
Rather not say

I'd like to continue that record.

1880 points

2
Highest Streak

2
Current Streak

8
Sentences Recorded

Email
kariima.kad@gmail.com

Age Group
Rather not say

Gender
Rather not say

Native Language
Rather not say

((a)) Login screen

((b)) Sign-up screen

((c)) Home screen

((d)) Stats screen

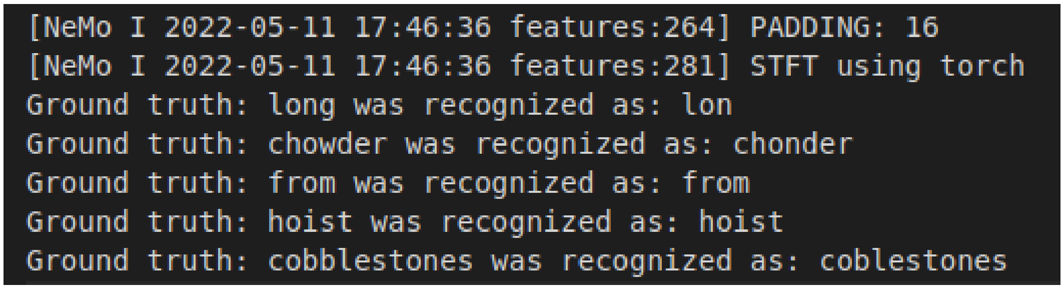
((e)) Profile screen

Figure 4: Screenshots of the Speech Collector app

5.2 Model Centric

5.2.1 Quartznet

Training QuartzNet on UASpeech for around 150 epochs showed a decrease in WER (from 200 to 99%) (Table 1 but it's nowhere near a satisfactory result. After doing inference however, it does seem that the model is in the right track and might just need further training (Figure 5).



```
[NeMo I 2022-05-11 17:46:36 features:264] PADDING: 16
[NeMo I 2022-05-11 17:46:36 features:281] STFT using torch
Ground truth: long was recognized as: lon
Ground truth: chowder was recognized as: chonder
Ground truth: from was recognized as: from
Ground truth: hoist was recognized as: hoist
Ground truth: cobblestones was recognized as: coblestones
```

Figure 5: Inference of samples from UASpeech speaker M05 using QuartzNet15x5

5.2.2 Conformer

Although the Conformer model showed a decrease in loss when being trained from scratch, it did not learn when it was fine-tuned. We thus attempted with a normal speech dataset, VCTK [7], which is characterized by speech with different accents. This would thus represent a task that is between normal unaccented speech ASR and impaired speech ASR in terms of difficulty. The model, however, was still not learning. During the initial validation sanity check, the model would make correct predictions thanks to its weights from pre-training, but the loss would start to exponentially increase as the training advances. The predictions would then be random words repeated several time (e.g. "brick brick brick") and the loss would reach the order of 10^6 before displaying "nan". We attempted to use gradual clipping, but although it somewhat mitigated the issue, it still didn't manage to prevent the loss from exploding.

5.2.3 Wav2Vec2CTC

We initially performed evaluation directly on the UASpeech dataset after fine-tuning the model on TIMIT and obtained a WER of 193.95%, which was not surprising given that even human beings find it hard to understand dysarthric speech.

Fine-tuning on VCTK before UASpeech has shown an improvement and is likely due to the difficulty of the accented speech task compared to healthy and dysarthric.

Although it is typical to freeze the feature encoder when fine-tuning the Wav2Vec2 model (the model learns to do the extraction sufficiently during pre-training), we have run experiments without freezing the feature encoder. The motivation behind doing that was that dysarthric data can be different than normal data acoustically as well. Therefore, the pre-training of the feature encoder on normal data alone might not be enough. Results did not show, however, a notable change in accuracy.

While a 50% WER is not low enough for the model to be ready for use, it is an important improvement on the initial WER before fine-tuning on the UASpeech dataset.

Table 2 shows the WER scores for each speaker resulting from fine-tuning the model on all 3 datasets.

5.2.4 Future Work

It seems to be a common practice to use the first and third blocks of the UASpeech dataset as a training set while the second block is used as a test set. There could be a problem with this approach however because all speakers appear in each block. In order for the model to learn how to generalize

Table 1: Results from trained architectures. All WER scores are computed on the UASpeech test set. FE: Feature Encoder

Experiment	**Pre-trained on**	**Fine-tuned/ Trained on**	**WER**
Quartznet	LibriSpeech	-	201.24%
Quartznet	LibriSpeech	UASpeech	99.68%
Conformer-CTC	LibriSpeech	UASpeech	-
Wav2Vec2 CTC	LibriSpeech (Unlabeled)	TIMIT	193.95%
Wav2Vec2 CTC	LibriSpeech (Unlabeled)	TIMIT + UASpeech	54.26%
Wav2Vec2 CTC	LibriSpeech (Unlabeled)	TIMIT + UASpeech (Unfrozen FE)	58.22%
Wav2Vec2 CTC	LibriSpeech (Unlabeled)	TIMIT + VCTK + UASpeech	**50.37**%

Table 2: Individual results from the Wav2Vec2 CTC experiment with the lowest WER score.

Speaker	**Correct**	**Total**	**Individual WER**	**Intelligibility**	**Dysarthria Type**
M01	311	1020	69.51%	Very Low	Spastic
M04	103	1275	91.92%	Very Low	Spastic
M05	934	1785	47.68%	Mid	Spastic
M07	973	1785	45.49%	Low	Spastic
M08	1049	1785	41.23%	-	Spastic
M09	1028	1785	42.41%	High	Spastic
M10	1074	1785	39.83%	-	Mixed
M11	824	1530	46.14%	Mid	Athetoid
M12	651	1530	57.45%	-	Mixed
M14	989	1785	44.59%	-	Spastic
M16	753	1530	50.78%	-	-
F02	876	1785	50.92%	Low	Spastic
F03	680	1783	61.86%	Very Low	Spastic
F04	920	1759	47.70%	Mid	Athetoid
F05	1089	1785	38.99%	High	Spastic

well and be useful "in the wild", we believe that it is imperative that some of the speakers are not seen at all during training. In the future, we plan to run experiments where we change the training and testing split to prevent "speaker contamination". We are expecting the "contaminated" model to perform better on the datasets speakers but to perform worse on entirely new speakers.

We also intend to explore using multiple languages as was attempted in [25] and to look at the problem as a Speech-to-Speech problem rather than a Speech-to-Text one.

6 Conclusion

Dysarthria is a speech disorder that poses a barrier between patients and communicating with their environments. Accurate ASR on dysarthric speech could change the harsh situation of dysarthric patients and offer them a way to connect with other people. Although some work has already been done in the field, it has not yielded satisfactory results yet. This project explored ways to improve ASR on dysarthric speech.

References

[1] Applied science and engineering laboratories. `https://www.asel.udel.edu/`. Accessed: 2021-04-15.

[2] Dysarthria.

[3] Dysarthria amp; speech: Symptoms, causes, treatments.

[4] Hugging face – the ai community building the future. `https://huggingface.co/`. Accessed: 2021-04-15.

[5] Nvidia nemo. `https://developer.nvidia.com/nvidia-nemo`. Accessed: 2021-04-15.

[6] University of toronto, department of computer science. `https://web.cs.toronto.edu/`. Accessed: 2021-04-15.

[7] Vctk. `https://datashare.ed.ac.uk/handle/10283/2950`. Accessed: 2021-04-15.

[8] Ossama Abdel-Hamid and Hui Jiang. Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code. 2013.

[9] Ossama Abdel-Hamid and Hui Jiang. Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition. 2013.

[10] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. volume 2020-December, 2020.

[11] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kopparapu. Improving recognition of dysarthric speech using severity based tempo adaptation. In *International Conference on Speech and Computer*, pages 370–377. Springer, 2016.

[12] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kumar Kopparapu. Recognition of dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation. In *Interspeech*, pages 228–232, 2016.

[13] Stephen Butterworth. On the theory of filter amplifiers, 1930.

[14] F. J. Charpentier and M. G. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. 1986.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.

[16] Joseph R Duffy. *Motor speech disorders e-book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2019.

[17] Miguel Fernández-Díaz and Ascensión Gallardo-Antolín. An attention long short-term memory based system for automatic classification of speech intelligibility. *Engineering Applications of Artificial Intelligence*, 96, 2020.

[18] Mark JF Gales. Cluster adaptive training for speech recognition. In *Fifth International Conference on Spoken Language Processing*, 1998.

[19] Ascensión Gallardo-Antolín and Juan M. Montero. An auditory saliency pooling-based lstm model for speech intelligibility classification. *Symmetry*, 13, 2021.

[20] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403, 1993.

[21] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. volume 148, 2006.

[22] Anmol Gulati, James Qin, Chung Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. volume 2020-October, 2020.

[23] Mark S Hawley, Pam Enderby, Phil Green, Stuart Cunningham, Simon Brownsell, James Carmichael, Mark Parker, Athanassios Hatzis, Peter O'Neill, and Rebecca Palmer. A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics*, 29(5):586–593, 2007.

[24] Enno Hermann and Mathew Magimai-Doss. Handling acoustic variation in dysarthric speech recognition systems through model combination. *Proc. Interspeech 2021*, pages 4788–4792, 2021.

[25] Abner Hernandez, Paula Andrea Pérez-Toro, Elmar Nöth, Juan Rafael Orozco-Arroyave, Andreas Maier, and Seung Hee Yang. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition, 2022.

[26] D Jeffery Higginbotham, Howard Shane, Susanne Russell, and Kevin Caves. Access to aac: Present, past, and future. *Augmentative and alternative communication*, 23(3):243–257, 2007.

[27] Sepp Hochreiter and Jurgen Schmidhuber. Long short term memory. neural computation. *Neural Computation*, 9, 1997.

[28] J-P Hosom, Alexander B Kain, Taniya Mishra, Jan PH Van Santen, Melanie Fried-Oken, and Janice Staehely. Intelligibility of modifications to dysarthric speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE, 2003.

[29] Wen-Chin Huang, Bence Mark Halpern, Lester Phillip Violeta, Odette Scharenborg, and Tomoki Toda. Towards identity preserving normal to dysarthric voice conversion. *CoRR*, abs/2110.08213, 2021.

[30] Wen Chin Huang, Tomoki Hayashi, Yi Chiao Wu, Hirokazu Kameoka, and Tomoki Toda. Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. volume 2020-October, 2020.

[31] Zhen Huang, Sabato Marco Siniscalchi, I-Fan Chen, Jiadong Wu, and Chin-Hui Lee. Maximum a posteriori adaptation of network parameters in deep models. *arXiv preprint arXiv:1503.02108*, 2015.

[32] Takaaki Ishii, Hiroki Komiyama, Takahiro Shinozaki, Yasuo Horiuchi, and Shingo Kuroiwa. Reverberant speech recognition based on denoising autoencoder. In *Interspeech*, pages 3512–3516, 2013.

[33] Zengrui Jin, Mengzhe Geng, Xurong Xie, Jianwei Yu, Shansong Liu, Xunying Liu, and Helen Meng. Adversarial data augmentation for disordered speech recognition. volume 4, 2021.

[34] Keith Johnson. Vocal tract length normalization. *UC Berkeley Phonology Lab Annual Reports*, 14, 2018.

[35] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. volume 2020-May, 2020.

[36] Alexander B Kain, John-Paul Hosom, Xiaochuan Niu, Jan PH Van Santen, Melanie Fried-Oken, and Janice Staehely. Improving the intelligibility of dysarthric speech. *Speech communication*, 49(9):743–759, 2007.

[37] Ozlem Kalinli and Shrikanth Narayanan. A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. volume 4, 2007.

[38] Ray D Kent. Research on speech motor control and its disorders. *Journal of Communication Disorders*, 33(5):391–428, 2000.

[39] Ray D Kent, Gary Weismer, Jane F Kent, Houri K Vorperian, and Joseph R Duffy. Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of communication disorders*, 32(3):141–186, 1999.

[40] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, and Simone Frame. Dysarthric speech database for universal access research. 2008.

[41] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. volume 2020-May, 2020.

[42] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 1951.

[43] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 1998.

[44] Christopher J Leggetter and Philip C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer speech & language*, 9(2):171–185, 1995.

[45] Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M. Cohen, Huyen Nguyen, and Ravi Teja Gadde. Jasper: An end-to-end convolutional neural acoustic model. volume 2019-September, 2019.

[46] Shansong Liu, Shoukang Hu, Xunying Liu, and Helen Meng. On the use of pitch features for disordered speech recognition. In *INTERSPEECH*, pages 4130–4134, 2019.

[47] Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *CoRR*, abs/1906.02762, 2019.

[48] Xavier Menendez-Pidal, James B. Polikoff, Shirley M. Peters, Jennie E. Leonzio, and H. T. Bunnell. Nemours database of dysarthric speech. volume 3, 1996.

[49] Megan A. Morris, Sarah K. Meier, Joan M. Griffin, Megan E. Branda, and Sean M. Phelan. Prevalence and etiologies of adult communication disabilities in the united states: Results from the 2012 national health interview survey. *Disability and Health Journal*, 9(1):140–144, 2016.

[50] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. volume 2015-August, 2015.

[51] Prasad D Polur and Gerald E Miller. Effect of high-frequency spectral components in computer recognition of dysarthric speech based on a mel-cepstral stochastic model. *Journal of Rehabilitation Research & Development*, 42(3), 2005.

[52] Michael R. Portnoff. Time-frequency representation of digital signals and systems based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28, 1980.

[53] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. 2016.

[54] Emily Elizabeth Redd. *The Effect of an Artificially Flattened Fundamental Frequency Contour on Intelligibility in Speakers with Dysarthria*. Brigham Young University, 2012.

[55] Frank Rudzicz. Acoustic transformations to improve the intelligibility of dysarthric speech. *SLPAT*, 2011.

[56] Frank Rudzicz. *Production knowledge in the recognition of dysarthric speech*. PhD thesis, 2011.

[57] Frank Rudzicz. Adjusting dysarthric speech signals to be more intelligible. *Computer Speech & Language*, 27(6):1163–1177, 2013.

[58] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46, 2012.

[59] Siddharth Sehgal and Stuart Cunningham. Model adaptation and adaptive training for the recognition of dysarthric speech. In *proceedings of SLPAT 2015: 6th workshop on speech and language processing for assistive technologies*, pages 65–71, 2015.

[60] Sid-Ahmed Selouani, Mohammed Sidi Yakoub, and Douglas O'Shaughnessy. Alternative speech communication system for persons with severe speech disorders. *EURASIP Journal on Advances in Signal Processing*, 2009:1–12, 2009.

[61] N Thubtong, P Kayasith, S Manochiopinig, W Leelasiriwong, and O Rukkharangsarit. Articulation analysis of thai cerebral palsy children with dysarthric speech. *Proceeding of the 6th Symposium on Natural Language Processing*, page 267–272, 2005.

[62] Ying-Chiao Tsao, Gary Weismer, and Kamran Iqbal. The effect of intertalker speech rate variation on acoustic vowel space. *The Journal of the Acoustical Society of America*, 119(2):1074, 2006.

[63] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. Data augmentation using healthy speech for dysarthric speech recognition. volume 2018-September, 2018.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. volume 2017-December, 2017.

[65] Ka Ho Wong, Yu Ting Yeung, Edwin H. Chan, Patrick C. Wong, Gina-Anne Levow, and Helen Meng. Development of a cantonese dysarthric speech corpus. *Interspeech 2015*, 2015.

[66] Feifei Xiong, Jon Barker, and Heidi Christensen. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. volume 2019-May, 2019.

[67] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide. Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. 2013.

[68] Jianwei Yu, Xurong Xie, Shansong Liu, Shoukang Hu, Max WY Lam, Xixin Wu, Ka Ho Wong, Xunying Liu, and Helen Meng. Development of the cuhk dysarthric speech recognition system for the ua speech corpus. In *Interspeech*, pages 2938–2942, 2018.

[69] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition, 2020.