Automatic Speech Recognition System for Dysarthria Patients

Ishtiaque Ahmed
Guided by
Dr. Waquar Ahmad

M.Tech Signal Processing
National Institute Of Technology, Calicut

June 2, 2023

Table of content

1. Introduction
2. Motivation
3. Literature Survey
4. Problem Definition
5. Methodology
6. Dataset
7. Results
8. Work done
9. References

Introduction

Automatic speech recognition (ASR) systems: Accurately translate spoken utterances into text (words, syllables, etc.)

- Well-known examples: YouTube closed captioning, Voicemail transcription, Dictation systems, Siri, Google Assistant, Cortana, etc.

Why is ASR desirable for all languages

- Speech is the primary means of human communication
- Develop natural interfaces for both literate & illiterate users
- Contribute to the preservation of endangered languages
- An Improved ASR system is needed in the case of pathological speech

[0] https://www.microsoft.com/en-us/research/video/automatic-speech-recognition-overview/

What makes ASR a difficult problem?

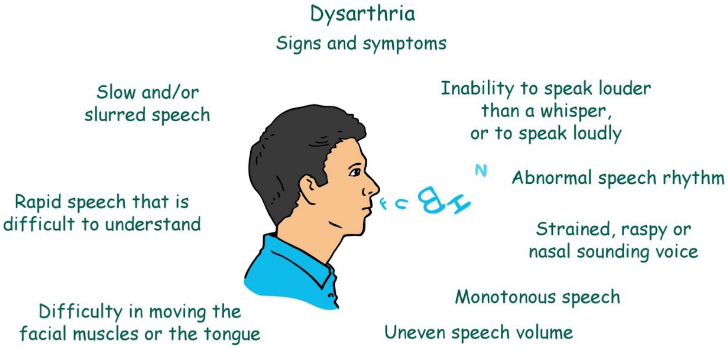
Several sources of variability!

- Style: Conversational (or casual) speech or read the speech? Continuous speech or isolated words?
- Environment: Background noise, channel conditions, room acoustics, etc.
- Speaker characteristics: Rate of speech, accent, etc.
- Task specifics: Number of words in the vocabulary, language constraints, etc.

[0] https://www.microsoft.com/en-us/research/video/automatic-speech-recognition-overview/

What is Dysarthria

Dysarthria is a speech disorder caused by muscle weakness. It can make it hard for one to talk. People may have trouble understanding what a person with dysarthria says


Dysarthria
Signs and symptoms
Slow and/or slurred speech
Rapid speech that is difficult to understand
Difficulty in moving the facial muscles or the tongue
Inability to speak louder than a whisper, or to speak loudly
Abnormal speech rhythm
Strained, raspy or nasal sounding voice
Monotonous speech
Uneven speech volume

[0] https://www.microsoft.com/en-us/research/video/automatic-speech-recognition-overview/

Causes of Dysarthria

- Stroke
- Brain injury
- Tumours
- Parkinson’s disease
- Amyotrophic lateral sclerosis, or ALS
- Huntington’s disease
- Multiple sclerosis Cerebral palsy
- Muscular dystrophy

[0] https://my.clevelandclinic.org/health/diseases/17653-dysarthria

Motivation

- **Inaccurate Word Recognition:** Existing assistive technologies for dysarthria patients often fall short of providing accurate and efficient communication solutions
- **Enhanced Communication:** Dysarthria patients often struggle to be understood by others, leading to frustration, isolation, and limited participation in social, educational, and professional settings
- **Inclusivity and Accessibility:** Accessibility is a fundamental right, and it is crucial to ensure that individuals with dysarthria have equal opportunities to communicate and be understood.
- **Advancements in Technology:** By leveraging the latest advancements in speech recognition algorithms, machine learning, and signal processing techniques we can build an ASR system.
- **Impact and Social Significance:** Improved communication can enhance their self-confidence, mental well-being, and the overall quality of life.

Literature Survey

- Researchers have noted the importance of understanding the characteristics of users' speech and how they affect interactions.**[1]**
- The accuracy results improved somewhat; however, the performance of these systems when used by people without dysarthria was notably different.**[1]**
- Some scholars have focused on improving systems to enhance interactions for people with dysarthria showed that individualized systems are more suitable.**[2]**
- Most studies evaluated systems using publicly available databases rather than real users and most recordings are produced in a controlled environment rather than natural environments.**[1]**

Literature Survey

- Commercial systems are mostly speaker-independent systems sold in the market and used by any user, that may be either built into computers. **[4][6]**
- Overall researchers have agreed that these devices require improvements to their recognition and robustness to fully understand dysarthric speech or to some extent. **[4]**
- Obtaining training data from users with dysarthria may be challenged by muscle fatigue and frustration, they experience.**[7]**

Problem Definition

The aim of this research work is to develop an effective automatic speech recognition system specifically designed to assist dysarthria patients in improving their communication abilities. The system should accurately recognize and transcribe the speech of dysarthria patients, accounting for their unique speech characteristics and challenges.

Dataset

We are using the UASpeech dataset for Test data and using the UASpeech Controlled dataset for Train data.[8] Subjects include 16 talkers with Cerebral Palsy and 13 age-matched healthy controls. Subjects were recruited based primarily on personal contact facilitated by disability support organizations. Subjects were selected based on self-report of either speech pathology or cerebral palsy. Before data were included in the UA-Speech distribution, the diagnosis of spastic dysarthria (sometimes mixed with other forms of dysarthria) was informally confirmed by a certified speech-language pathologist listening to these recordings. Subjects were asked to explicitly grant permission for the dissemination of their data; subjects who refused permission are not represented in the distribution.

Methodology

Need for Data Augmentation

- Data augmentation generates new training data by applying transformations or modifications to existing datasets, enhancing the system's accuracy.
- Data augmentation is particularly important to improve the accuracy and robustness of ASR models by introducing variations in speech duration and tempo.
- Collecting sufficient speech data, especially from individuals with dysarthria, can be challenging due to limitations in speaking duration caused by muscular weakness and fatigue.
- Limited speech data from a small number of dysarthric speakers hinders the development of accurate ASR models.

Need for Data Augmentation

- A proposed data augmentation approach focuses on modifying the duration of speech segments.
- Our study aims to develop an ASR system for dysarthric speakers using x-vector-based speaker embedding[**9**].
- Limited availability and diversity of dysarthric speech data pose challenges in training an x-vector-based system, risking under-fitting.
- Incorporating data augmentation helps overcome the challenges of limited data availability and diversity.
- Data augmentation techniques are needed to increase the available training data and improve the performance of ASR systems for dysarthric speech.

Methodology

Motivation for Duration Modification based Data Augmentation

- Dysarthric speech is characterized by longer durations of phonemes, inter-word delays, frequent pauses, and non-speech sounds.

- Duration modification-based data augmentation is proposed to address the missing acoustic attributes in dysarthric speech for training ASR systems.

- By extending the duration of training data from dysarthric speakers, ASR systems can learn larger phoneme durations and become more robust for dysarthric patients.

- Our experimental validation shows that duration-modification-based data augmentation significantly improves the performance of ASR systems for dysarthric speech.

Dysarthric Speech vs Controlled Speech

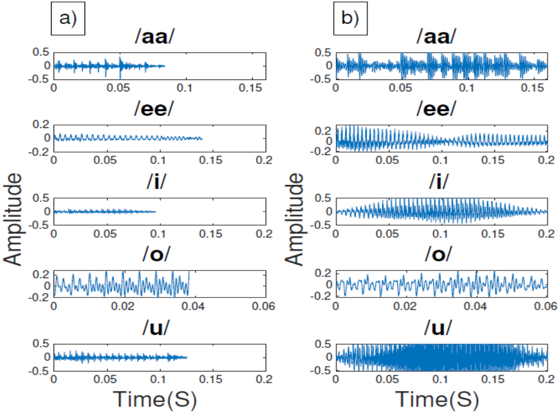


Figure 1: Waveform for vowel sounds /aa/, /ee/, /i/, /o/, /u/ spoken by (a) control subject, and (b) dysarthric subject

Methodology

Employed ASR System Architecture

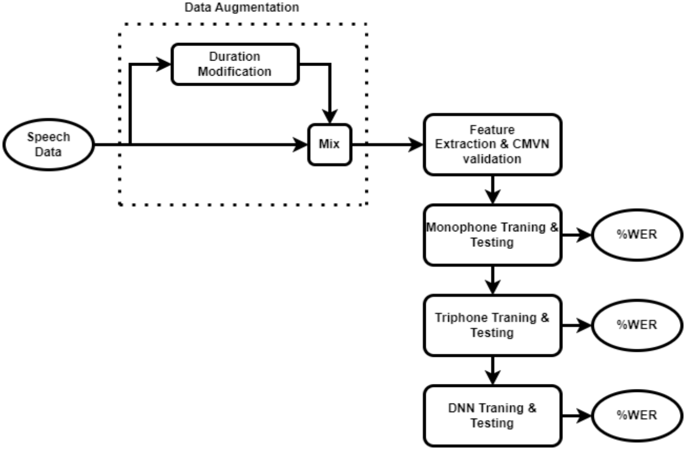

Data Augmentation
Duration Modification
Speech Data
Mix
Feature Extraction & CMVN validation
Monophone Traning & Testing
%WER
Triphone Traning & Testing
%WER
DNN Traning & Testing
%WER

Figure 2: Automatic Speech Recognition Architecture

ASR Architecture

- The ASR system architecture includes a data augmentation module based on duration modification to enhance acoustic diversity and incorporate targeted attributes.
- The duration modification technique involves extending the duration of all phonemes in speech utterances using the glottal closure (GC) and glottal opening (GO) instants identified through the zero frequency filtering (ZFF) method[**11**].
- The ZFF method[**11**] involves computing the differential input signal, passing it through ideal digital filters at zero frequency, and removing the trend in the filtered signal to obtain the ZFF signal.
- GC and GO instants are detected as positive zero crossings of the ZFF signal, which are then used for duration modification.
- Duration modification is achieved by creating a new epoch sequence with desired duration changes, modifying the duration according to a modification rate α, and reconstructing the duration-modified speech.

ASR Architecture

- In the training phase, speech data from dysarthric speakers and their duration-modified versions are combined with healthy speech data, then combined training speech is processed through VAD.
- The features extracted from the front end are normalized, and in the testing phase, dysarthric speech data undergo feature extraction, monophone training and testing, triphone training and testing, and DNN training and testing.
- The extracted features are stored in the form of x-vectors[**9**], which are then scored for verification, resulting in WER measurements for each model.
- The x-vector extractor used in this study is based on the system proposed which converts variable-length utterances into fixed-dimensional embeddings called x-vectors[**9**].

Experimental Evaluation

Experimental Setup

- (UA-Speech) corpus, along with an augmented dataset created using scaling is used for this project.
- The UA-Speech dataset includes recordings from 16 talkers with Cerebral Palsy and 13 age-matched healthy controls, selected based on self-report of speech pathology or cerebral palsy.
- The UA-Speech corpus consists of data from healthy speakers, dysarthric speakers, and an augmented dataset with scaled-up data of dysarthric speakers, totaling over 30,000 utterances from 16 speakers.
- The Kaldi speech recognition toolkit was utilized **[13]**for the development and evaluation of the ASR system.
- The evaluation involved 52,670 trials, including 2070 genuine trials and 50,600 impostor trials, to assess the performance and accuracy of the developed ASR system.

Results

- The ASR system is evaluated using speech data from the Torgo and UA-Speech databases, with Torgo serving as the baseline and UA-Speech as the proposed results.
- Monophone training and triphone modeling techniques are used for acoustic modeling in ASR, with triphones capturing context-dependent phonetic units and improving accuracy.
- Tri-1, Tri-2, and Tri-3 represent stages of ASR training, incorporating techniques to enhance performance and adapt to speaker characteristics.
- Deep Neural Networks (DNN) have shown significant improvements over traditional models like GMM-HMM in ASR, utilizing multiple hidden layers to learn acoustic-to-text mappings.
- Data augmentation through duration modification is performed to address the acoustic mismatch between training and test data, with representing the modification rate.

Work Done and Results

We did our experiments with the following modification factors.

Table 1: Overall %WER

Scaling 1.1			Scaling 1.2			Scaling 1.3			Scaling 1.4	
Method Used	%WER		Method Used	%WER		Method Used	%WER		Method Used	%WER
DNN	25.63		DNN	25.63		DNN	25.63		DNN	25.63
Scaling 1.5			Scaling 1.6			Scaling 1.7			Scaling 1.8	
Method Used	%WER		Method Used	%WER		Method Used	%WER		Method Used	%WER
DNN	25.63		DNN	25.52		DNN	25.45		DNN	24.02

Work Done and Results

- The WER is evaluated for dysarthric speakers of different severity levels, and the proposed system outperforms the baseline, particularly for high-severity dysarthric speech.
- Optimal modification factors $\alpha = 1.8$ is chosen based on performance, and merged training sets are used for the final ASR system.
- The proposed duration modification-based data augmentation approach effectively improves performance for healthy and dysarthric speakers.
- WER values are computed for each severity level, indicating significant improvement over the baseline, with higher improvement observed for higher severity levels.
- The proposed ASR system demonstrates the ability to effectively recognize speech from normal and dysarthric speakers of varying intelligibility by appropriately modifying speech duration.

Results

With our research work, there is a significant amount of improvement in the results across all models as compared to the baseline **[10]** model.

Table 4: Overall %WER

ASR System	Baseline	Proposed
Monophone	71.63 %	50.61 %
Triphone-1	71.41 %	39.75 %
Triphone-2	80.85 %	34.09 %
Triphone-3	55.04 %	28.14 %
DNN	47.11 %	19.03 %

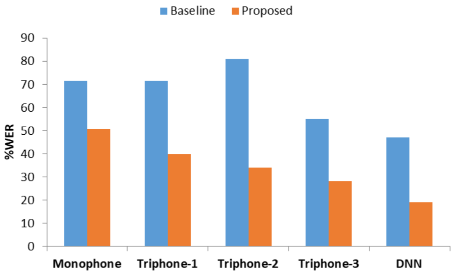


Figure 3: Comparison from base

Results

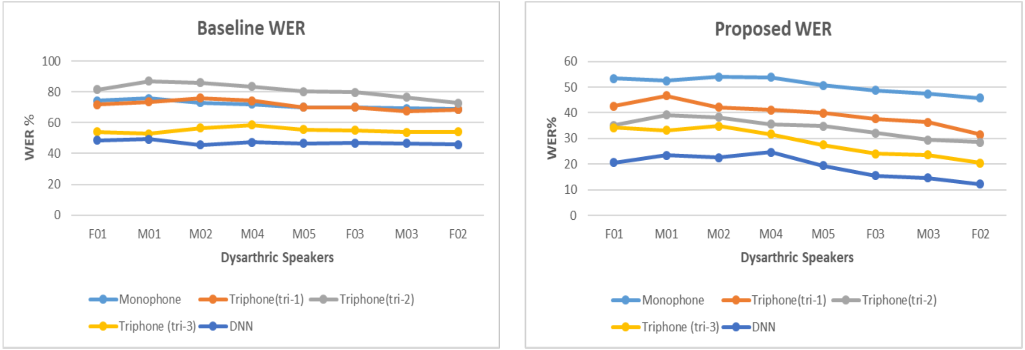
Table 2: WER of dysarthria speakers across several different models for baseline models

Baseline	Severely				Moderate to Severely	Moderately	Very Mild	
Speaker ID	F01	M01	M02	M04	M05	F03	M03	F02
Monophone	74.28	75.64	72.91	71.98	70.12	70.08	69.33	68.76
Triphone(tri-1)	71.76	73.45	76.02	74.12	70.04	69.93	67.46	68.53
Triphone(tri-2)	81.37	86.9	85.9	83.46	80.28	79.82	76.45	72.64
Triphone(tri-3)	54.14	52.9	56.62	58.43	55.63	54.98	53.72	53.96
DNN	48.56	49.32	45.59	47.36	46.68	46.92	46.56	45.90

Table 3: WER of dysarthria speakers across several different models for proposed models

Proposed	Severely				Moderate to Severely	Moderately	Very Mild	
Speaker ID	F01	M01	M02	M04	M05	F03	M03	F02
Monophone	53.45	52.46	53.98	53.89	50.7	48.68	47.44	45.8
Triphone(tri-1)	42.68	46.6	42.14	41.13	39.9	37.68	36.4	31.5
Triphone(tri-2)	35.12	39.16	38.27	35.55	34.79	32.12	29.43	28.54
Triphone(tri-3)	34.2	33.21	34.9	31.67	27.54	24.1	23.6	20.4
DNN	20.6	23.5	22.5	24.6	19.4	15.6	14.63	12.3

Results



(A)

(B)

Figure 4: Experimental plots (A) Baseline WER Plot, (B) Proposed WER Plot

Conclusion

- The study introduces an automatic speech recognition (ASR) system tailored for dysarthric speakers with varying speech intelligibility.
- A duration-modification-based data augmentation module is incorporated in the ASR system's front end to address the challenges.
- The proposed method involves applying duration modification with different scaling factors to dysarthric training data and augmenting it with the original version.
- Experimental results demonstrate the effectiveness of the proposed method, showing a 34% relative improvement in the performance of the baseline ASR system for dysarthric speakers.
- The method also improves performance for individual severity levels, with a relative improvement of approximately 29% observed for the high severity level.

References I

[1] Victoria Young and Alex Mihailidis. “Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review”. In: *Assistive Technology* 22.2 (2010), pp. 99–112.

[2] Mumtaz Begum Mustafa et al. “Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker”. In: *Expert Systems with Applications* 42.8 (2015), pp. 3924–3932.

[3] Kristin Rosen and Sasha Yampolsky. “Automatic speech recognition and a review of its functioning with dysarthric speech”. In: *Augmentative and Alternative Communication* 16.1 (2000), pp. 48–60.

[4] Neethu Mariam Joy and S Umesh. “Improving acoustic models in TORGO dysarthric speech database”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.3 (2018), pp. 637–645.

[5] Aisha Jaddoh, Fernando Loizides, and Omer Rana. “Interaction between people with dysarthria and speech recognition systems: A review”. In: *Assistive Technology* (2022), pp. 1–9.

[6] Luigi De Russis and Fulvio Corno. “On the impact of dysarthric speech on contemporary ASR cloud platforms”. In: *Journal of Reliable Intelligent Environments* 5.3 (2019), pp. 163–172.

[7] Fabio Ballati, Fulvio Corno, and Luigi De Russis. ““Hey Siri, do you understand me?”: Virtual Assistants and Dysarthria”. In: *Intelligent Environments 2018*. IOS Press, 2018, pp. 557–566.

[8] Heejin Kim et al. “Dysarthric speech database for universal access research”. In: *Proc. Interspeech 2008*. 2008, pp. 1741–1744. DOI: `10.21437/Interspeech.2008-480`.

References II

[9] Meet Soni et al. “Acoustic Model Adaption Using x-vectors for Improved Automatic Speech Recognition”. In: Nov. 2022. DOI: 10.23919/APSIPAASC55919.2022.9979886.

[10] Jun Ren and Mingzhe Liu. “An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks”. In: *International Journal of Advanced Computer Science and Applications* 8 (Jan. 2017). DOI: 10.14569/IJACSA.2017.081207.

[11] SRM Prasanna et al. “Fast prosody modification using instants of significant excitation”. In: *Speech Prosody 2010-Fifth International Conference*. 2010.

[12] Shinimol Salim, Syed Shahnawazuddin, and Waquar Ahmad. “Automatic Speaker Verification System for Dysarthria Patients”. In: *Proc. Interspeech 2022* (2022), pp. 5070–5074.

[13] Daniel Povey et al. “The Kaldi speech recognition toolkit”. In: *IEEE 2011 workshop on automatic speech recognition and understanding*. CONF. IEEE Signal Processing Society. 2011.