"I'm Having Trouble Understanding You Right Now": A Multi-Dimensional Evaluation of the Intelligibility of Dysphonic Speech

by

Meredith Moore

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2020 by the
Graduate Supervisory Committee:

Sethuraman Panchanathan, Chair
Visar Berisha
Troy McDaniel
Hemanth Venkateswara

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

Individuals with voice disorders experience challenges communicating daily. These challenges lead to a significant decrease in the quality of life for individuals with dysphonia. While voice amplification systems are often employed as a voice-assistive technology, individuals with voice disorders generally still experience difficulties being understood while using voice amplification systems. With the goal of developing systems that help improve the quality of life of individuals with dysphonia, this work outlines the landscape of voice-assistive technology, the inaccessibility of state-of-the-art voice-based technology and the need for the development of intelligibility improving voice-assistive technologies designed both with and for individuals with voice disorders. With the rise of voice-based technologies in society, in order for everyone to participate in the use of voice-based technologies individuals with voice disorders must be included in both the data that is used to train these systems and the design process. An important and necessary step towards the development of better voice assistive technology as well as more inclusive voice-based systems is the creation of a large, publicly available dataset of dysphonic speech. To this end, a web-based platform to crowdsource voice disorder speech was developed to create such a dataset. This dataset will be released so that it is freely and publicly available to stimulate research in the field of voice-assistive technologies. Future work includes building a robust intelligibility estimation model, as well as employing that model to measure, and therefore enhance, the intelligibility of a given utterance. The hope is that this model will lead to the development of voice-assistive technology using state-of-the-art machine learning models to help individuals with voice disorders be better understood.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my committee for their guidance and dedication. Your thoughtful feedback and encouragement meant the world to me and has helped shape me into the researcher that I am today.

I would also like to extend my deepest gratitude for the support of the National Science Foundation's Graduate Research Fellowship which provided me with the both the funding and freedom of mind to explore the corners of computer science that most excited me.

I am also incredible grateful for the support and training offered from the National Science Foundation through the Interdisciplinary Graduate Education Research Traineeship (IGERT) on the Alliance for Person-Centered Accessible Technologies (APAcT), and specifically Jay Klein and Troy McDaniel for their effort and support in making APAcT the wonderful interdisciplinary learning experience that it was.

The completion of this dissertation would not have been possible without the support and nuturing of Kim Kuman, the Executive Director of the National Spasmodic Dysphonia Association. Thank you, Kim, for your effortless collaboration, ideas and the promotion of this research.

I'm extremely grateful to Piyush Papreja for his assistance and dedication to building the web experience for UncommonVoice.

To all the Cubites–a term of endearment for those who are part of the Center for Cognitive Ubiquitous Computing–thank you for the impromptu ping-pong matches, thought-provoking discussions, and general support over the past several years. The culture of CUbiC is something special, and I count myself as very lucky to have experienced it.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

The driving force of my dissertation is to help people with voice disorders be better understood. A voice disorder is characterized by any deviation in voice quality, pitch or loudness, inappropriate for an individual's age, gender or cultural backgroundAronson and Bless (2009). Communication plays a vital role in a person's participation in society Tiwari and Tiwari (2012). Voice disorders often make a significantly negative impact on an individual's quality of life Roy *et al.* (2005). The main driver of this negative impact is decreased intelligibility—not being able to be fully understood by communication partners. This decrease in intelligibility significantly impacts an individual's ability to communicate their thoughts, ideas, opinions, emotions, and general personality. Voice is very easy to take for granted–you don't realize how important it is until you are unable to use your voice to communicate. In fact, so much so that the American's with Disability Act (ADA) of 1990 qualified communication as a major life activity. Major life activities and an individual's ability to partake in major life activities qualify what defines an individual as having a disability under the ADA. As such, individuals with voice disorders that significantly affect their ability to communicate are also individuals with disabilities and are therefore protected by the ADA.

In 1988 the Technology-Related Assistance for Individuals with Disabilities Act (Public Law 100-407) defining an assistive technology device as 'any item, piece of equipment, or product system, whether acquired commercially off the shelf, modified, or customized, that is used to increase, maintain, or improve the functional capabilities of individuals with disabilities.' In the last several decades, there has been a

significant push to build technologies that help individuals with disabilities, resulting in the interdisciplinary field of Assistive Technologies. Assistive technologies span a wide variety of disabilities, from helping provide visual information to individuals with visual impairments, to a crutch that helps someone with a broken leg move around more freely. However, assistive technologies for individuals with voice disorders have been historically overlooked. The most relevant work in the field of assistive technologies to helping individuals with voice disorders the field of communication aids. Communication aids cover a large body of literature and devices, from low-tech augmentative, alternative communication (AAC) boards, to speech-generating devices and automatic recognition of sign language. The general dogma of the field has trended away from using speech as an input, instead, relying on text/images as the input into these systems, or gestures such as American Sign Language. While this is great for individuals who cannot communicate via speech, it overlooks individuals who still can use their voice to communicate, but might not be as intelligible as individuals with 'healthy speech'.

1.1 Motivation

Not being able to be understood has far-reaching effects on an individual's life. Having a voice disorder often causes individuals to withdraw socially, experience difficulties in their career, and experience a general decrease in emotional well-being as characterized by isolation, frustration, stress, anxiety, and depression. In this dissertation, I make significant strides towards understanding the impact that voice disorders have on the lives of individuals with dysphonia and identify areas of opportunity where voice-assistive technologies could be beneficial. I also take a closer look at what it means to be intelligible, and the many factors that impact the intelligibility of speech. From there, I evaluate the intelligibility of a wide variety of speech from

different angles to gain a better understanding of what it means to be understood– whether by humans or machines. The eventual goal of this work is to build a system that helps people with dysphonic speech be better understood.

1.2 Contributions

The contributions of this dissertation are as follows:

- A broad survey of the needs of individuals with dysphonia, including areas of opportunity for voice-assistive technologies to improve the quality of life of individuals with dysphonia.
- An evaluation of the accessibility and inclusivity of state-of-the-art voice-based technology
- A more nuanced and in-depth analysis of what it means for speech to be intelligible
- UncommonVoice: the largest publicly-available dataset of dysphonic speech, as well as accompanying metadataset.
- Design considerations for the development of voice-assistive technologies

Chapter 2

BACKGROUND AND PREVIOUS WORK

2.1 Voice Disorders

In the United States, it is estimated that 9.4 million adults have trouble using their voices Bhattacharyya (2014). Of those 9.4 million individuals, it is estimated that 2 million have a voice disability severe enough to affect their ability to be understood Beukelman and Mirenda (2005). Speech communication can be fundamental to an individual's participation in society. Dysphonia—the medical term for disorders of the voice—can result from alterations in respiratory, laryngeal, or vocal tract mechanisms, improper or inefficient use of the vocal mechanism, psychological distress, or a combination of these factors Lee *et al.* (2004). Voice disorders profoundly impact the quality of life and overall health of individuals with voice disorders often leading to anxiety, depression, and social isolation Merrill *et al.* (2011). While treatments and tools have been developed to help mitigate the symptoms of voice disorders—such as voice therapy, Botulinum Toxin injections, and voice amplification devices— many still experience symptoms that lead to a decrease in intelligibility.

2.1.1 A Note On Dysarthria v. Dysphonia

Dysarthria and Dysphonia are often confused. Dysarthria refers to difficulty speaking that may occur secondary to an injury or neurological disease. Damage to the musculature responsible for speech can also cause dysarthric speech. Dysarthria can present from mild to severe depending on the location and severity of the brain damage. There are several different types of dysarthria. Ataxic dysarthria stems from

poor coordination of the speech muscles making speech and volume slow, erratic, and irregular. In Flaccid dysarthria, the voice is often breathy and has a nasal quality due to poor control of the soft palate. Spastic dysarthria can sound slow, indistinct, and monotone while seeming like it is difficult for the speaker to articulate. Individuals with hyperkinetic dysarthria sound harsh and strained, while hypokinetic dysarthria presents with a hoarse voice and low volume. It is possible, and common, for individuals to have multiple types of dysarthrias, something that is referred to as Mixed dysarthria Lee *et al.* (2004).

Dysphonia, rather, is characterized by weakness or loss of voice. Dysphonia can present following a stroke, disease or trauma to the larynx. Dysphonia has more to do with the functioning of the vocal cords as well as dysfunctioning muscles related to phonation. The two main types of dysphonia are Organic and Functional. Organic dysphonia occurs when there is a physical problem with the vocal apparatus such as laryngitis, or a structural abnormality such as a nodule, tumor, or trauma to the larynx. Functional dysphonia occurs when the problem isn't structural, but there is still a voice problem.

2.1.2 Epidemiology of Dysphonia

Voice disorders affect elderly females far more than any other demographic Adler *et al.* (1997); Merrill *et al.* (2011); Cohen *et al.* (2012); Patel *et al.* (2015); Roy *et al.* (2005). The prevalence of voice disorders is also highly correlated with certain risk factors such as the individual's vocation Roy *et al.* (2004); Aminoff *et al.* (1978); Thibeault *et al.* (2004); Roy *et al.* (2005). For example, jobs that ask the individual to overuse their voice such as teachers, singers, construction workers, sales representatives, and clergy-members, all have an increased risk of developing a voice disorder.

There is a relatively large body of literature indicating that there is a strong correlation between depression and anxiety and voice disorders, however, the directionality of this relationship is relatively unexplored Willinger *et al.* (2005); Elena Nerriere (2009); Elam *et al.* (2010). For example, researchers are unsure as to whether the anxiety/depression increases the propensity for having a voice disorder, or if having a voice disorder increases the probability of experiencing anxiety and depressive symptoms.

2.1.3 Spasmodic Dysphonia

Spasmodic dysphonia (SD), also known as laryngeal dystonia, is a voice disorder that is characterized by the improper functioning of the muscles that generate a person's voice Aminoff *et al.* (1978). These muscles spasm, in what is referred to as a laryngospasm, which makes it difficult to speak or breathe. When the spasms cause the vocal cords to be too tight and overlap, it is referred to adductor spasmodic dysphonia (ADSD), while if the vocal cords are too loose, and open during the spasms, it is referred to as abductor spasmodic dysphonia (ABSD).

In ADSD, the voice is often strained, harsh, tight, and tremulous, while being low in volume and pitch Aminoff *et al.* (1978). The speech is often interrupted by irregular breaks and stoppages. In the less common form of SD, ABSD, the voice often sounds breathy and has a very low volume. This speech is often described as being a whisper Ludlow *et al.* (1991). These two types of SD are not mutually exclusive, some individuals have symptoms of both ADSD and ABSD, which is referred to as mixed SD.

While there is no known cure for Spasmodic Dysphonia, there are several different treatment paths for individuals with this voice disorder that have shown to alleviate or control the symptoms of the vocal spasms on a temporary or long-lasting basis.

While these do not address the underlying neurological dysfunction, they usually give enough symptom relief to enable a person to regain control of and improve the quality of their voice.

2.1.4 Voice Disorder Treatments

The purpose of voice therapy is the improvement of the vocal quality by teaching the patient to use his/her vocal mechanism more efficientlySpeyer (2008). In a systematic review of the filed, Speyer (2008) found that while many papers have methodological challenges, and subjective measures are often used, there is a tendency towards a modest positive effect of voice therapy on dysphonia. While voice therapy can teach individuals with voice disorders how to use their voice optimally—therefore somewhat improving the quality of the voice, it generally doesn't help the individual fully recover voice function. Voice therapy can also be costly and time-consuming.

Botulinum toxin (Botox ™, also referred to as BTX) injections have proven to be one of the most popular and effective treatments for individuals with voice disorders—specifically spasmodic dysphonia Ludlow (1990). BTX is a neurotoxin that acts as a 'blocker', inhibiting the contraction of muscles. Because SD is caused by involuntary spasms of the larynx (see 2.1.3), BTX injections into the muscles that spasm has shown to be an effective way to reduce the number of laryngospasms, which in turn lowers the acoustic measurement of the fundamental frequency, change in fundamental frequency and voice-break factor, therefore positively affecting an individual's voice quality Zwirner *et al.* (1991).

Several factors affect the efficacy of BTX injections as a treatment for SD. BTX can be injected unilaterally or bilaterally (either on the muscle that controls one vocal chord or the muscles that control both of the vocal cords). For ADSD, generally, the

[1]Image Source: https://www.dysphonia.org/anatomy.php

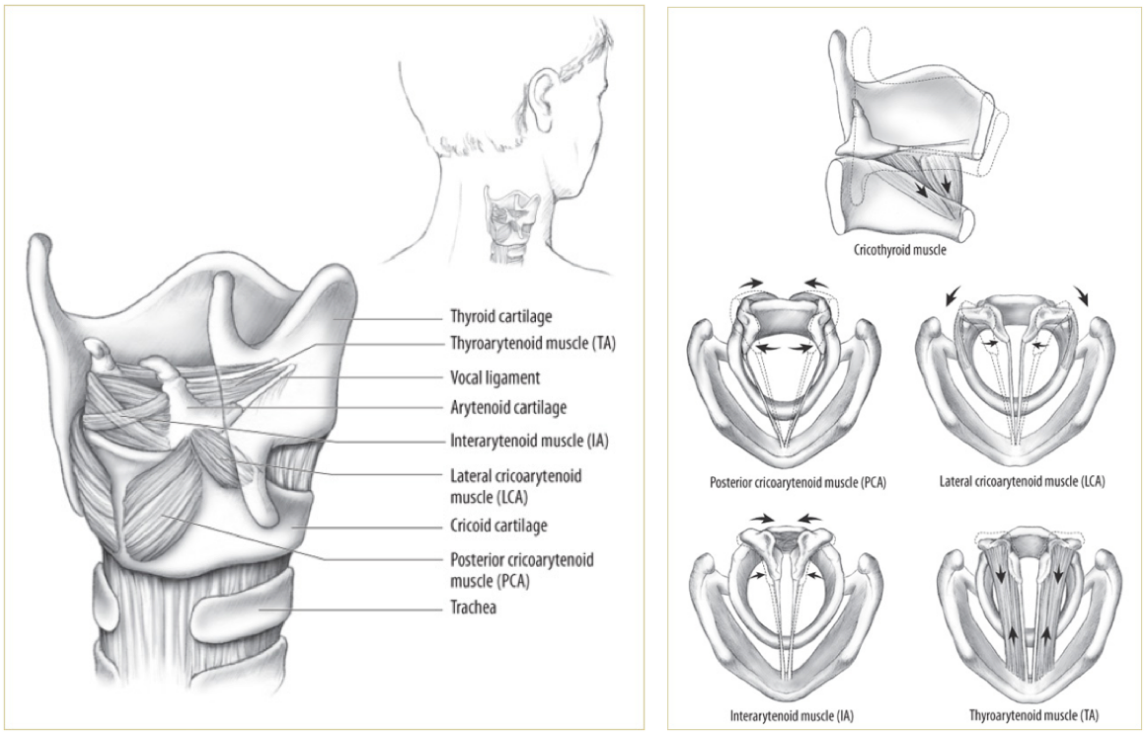


Figure 2.1: Illustration of the Anatomy of the Larynx[1]

thyroarytenoid muscle–the muscle responsible for closing the larynx–is targeted, while for ABSD, the posterior cricoarytenoid muscle–the muscle responsible for opening the larynx–is targeted Benninger *et al.* (2001). See Figure 2.1 for the location of these muscles and the general anatomy of the larynx.

The amount of BTX injected is up to the physician and generally involves a process of trial and error over time, but ranges from 15 to 30 units depending on the severity of the symptoms. The voice quality improvement afforded by BTX injections lasts anywhere from 6-12 weeks. Physicians often 'overshoot' the amount of BTX they inject for individuals with ADSD, causing the patient to experience a period where their normally creaky voice sounds breathy. This variation in voice quality over time has been deemed the colloquially deemed the 'botox rollercoaster', the process is shown in Figure 2.2.

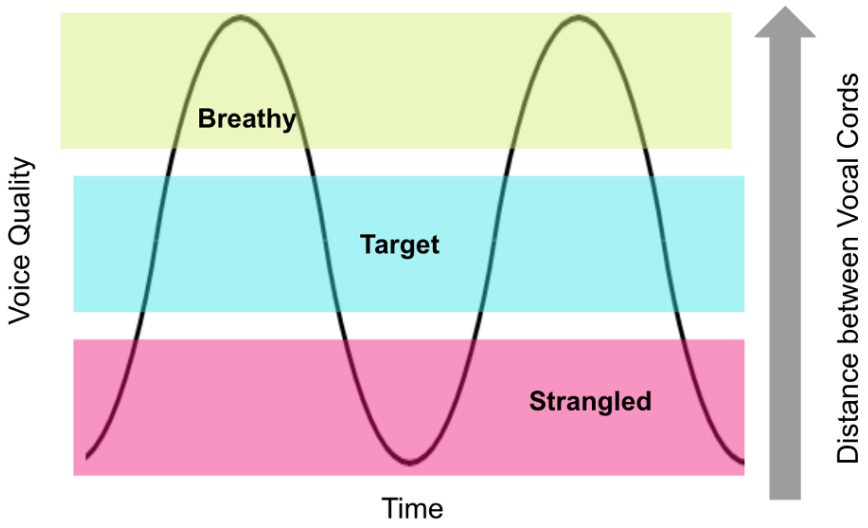


Figure 2.2: The BTX Treatment 'Rollercoaster'.

While BTX injections, voice therapy, or a combination of the two have shown to be a good way to manage the symptoms associated with common voice disorders, specifically SD, there are still several limitations associated with these treatment methodologies. For example, BTX injections primarily work for individuals with ADSD and show much less success in the ABSD population. These injections can also be costly if the individual's medical insurance doesn't cover them, and often patients need to travel hours to the nearest trained otolaryngologist to receive their injection. BTX injections also involve a significant amount of guesswork as to the dosage and the frequency of the injections. This can lead to varying degrees of success and potentially extended 'breathy' periods. Once the optimal dosage/frequency is achieved, there is still generally a week or so period of breathiness after a BTX injection. Despite the increase in voice quality after a BTX injection, the cost, travel time, and period of decreased intelligibility due to breathiness just after the BTX injection make BTX injections a significant inconvenience for some.

2.1.5 Relevant Acoustic Measures for Dysphonic Speech

Acoustic features such as the fundamental frequency, the standard deviation of the fundamental frequency, jitter, shimmer, signal-to-noise-ratio have been shown to change for individuals with ADSD after Botox injections. The fundamental frequency (F_0) of a voice reflects how high or low the pitch of a voice sounds, and is correlated with the changes in vocal fold tension. Typically adult males have a F_0 of from 85 to 155 Hz, while adult females range from 165 to 255 HZ. The standard deviation of the fundamental frequency (SDF_0) is the square root of the variance around the man fundamental frequency and reflects the variability of F_0, and can be used as a measure of instability in a voice. Jitter is the cycle-to-cycle variation in frequency. Shimmer is the cycle-to-cycle variation in amplitude. Signal to Noise Ratio (SNR) is the ratio of energy in the signal versus the noise components also contained in the acoustic spectrum. These features have been tested in both healthy and dysphonic speakers and pre-botox dysphonic speakers had significantly higher values for all while after BTX the SNR and SDF_0 were significantly less than before BTX injections for individuals with ADSD Zwirner *et al.* (1991).

There have been several acoustic parameters of speech that have been correlated with a perceived dysphonia severity score, more specifically, the cepstral peak prominence (CPP), the mean ratio of low-to-high frequency spectral energy, and the standard deviation of the ratio of low-to-high frequency spectral energy Awan *et al.* (2009). The CPP is a more reliable measure of dysphonia than other acoustic parameters such as jitter, shimmer and noise-to-nonharmonic ratio Heman-Ackah *et al.* (2003).

2.2 Communication Aids

Communication is a complex process involving the transfer of information between two participants. There are many points in the communication process where information could be lost. For example, if the information is being communicated verbally, the information could be lost in the creation/production of the signal (speaking), or the perception of the signal (listening).

There is a range of assistive technologies in the field of communication aids to enable people with disabilities to have full access to communication: including alternative augmentative communication (AAC) devices, speech-generating devices, automatic sign language translation systems, and hearing aids Beukelman and Mirenda (2005). To help individuals communicate AAC devices provide an interface that displays a combination of images, words, or phrases that are generally customizable by the user. The user can select from the vocabulary what they want to say, and the message will either be displayed on the device or if it's a speech-generating AAC device, it will speak the message out loud. This technology is great for individuals who are unable to speak, however, this may not be ideal for individuals who still can speak, but who may have problems being understood. These devices also pose difficulties when the user is trying to communicate in real-time. The process of creating a message can be time-consuming, therefore interrupting the natural flow of conversation.

The average human speaking rate is 130-190 words per minute (wpm), while the average speaking rate of an individual using an AAC device ranges from < 1 wpm to a max of about 35 wpm Trnka *et al.* (2009). It has been shown that this speaking rate can be marginally improved using natural language processing (NLP) techniques to predict word choice for the user Copestake (1997), however, this increase in speaking

rate comes at a cost of an increase in cognitive load as individuals have to scan the list of predictions to select a word that fits their use-case Trnka *et al.* (2008).

Because the speed of input is an important part of holding a conversation Yuan *et al.* (2006), it follows that if a user can use their voice, this ability should be leveraged in the design of communication aids. For this dissertation, communication aids that use voice as the input modality are referred to as voice-assistive technologies.

2.2.1 Voice-Assistive Technologies

While the majority of the literature in communication aids focuses on input modalities other than voice, there are several pockets of literature on voice-input devices as communication aids. Dysarthric speech recognition is one of these areas that attempts to solve the problem of helping people with speech disorders—primarily dysarthrias—be better understood. The general methodology here has been to train an automatic speech recognition (ASR) system on a body of speech from individuals with speech disorders, using the text output of the ASR system (the transcript of what was said) as an input into a speech synthesis system which 'speaks' the transcript in an intelligible voice.

A potential solution to recognizing significantly different voices is to build personalized ASR systems that fit individual voices. This methodology has been attempted for the last 30 years, and there has not been significant progress. Of the dysarthric speech recognizers created, those that use an extremely limited vocabulary (10 digits) achieve around 94% accuracy Hasegawa-Johnson *et al.* (2006); Green *et al.* (2003). Results from systems that use larger vocabularies are extremely varied from 30.84% Polur and Miller (2006) to 97% recognition rate Sharma and Hasegawa-Johnson (2010). The highest reported accuracy on the biggest vocabulary using the least intelligible subjects was 85.05% from Selva Nidhyananthan *et al.* (2016) using

recurrent models with Elman backpropagation networks.

However, due to the large variability in testing conditions—the intelligibility of subjects, the number of subjects, the complexity of the vocabulary, and the different evaluation metrics—it is very difficult to objectively compare the efficacy of different algorithms. Very few systems using dysarthric speech recognition have been robust enough to make it to the commercial market. VoiceItt [2] is one company focusing on non-standard speech recognition, and while they have several videos showing their product in action, they have not launched a product beyond the Beta stage yet.

Another potential solution to improving the intelligibility of dysarthric speech is to adjust certain temporal and acoustic features of the speech. In dysarthric individuals, it is common for the length of a given utterance to be significantly longer than that of a healthy speaker. In Bhat *et al.* (2018), they use a time-delay neural-network-based denoising autoencoder to adjust dysarthric speech to be recognized by ASR systems. This paper focused on temporal adaptations using the phase vocoder from Rudzicz (2013). In Rudzicz (2013), they take a very manual approach to the improvement of intelligibility by removing repeated sounds, inserting deleted sounds, devoicing the unvoiced phonemes, and adjusting the tempo and frequency of the speech to improve the intelligibility of the speech.

In Biadsy *et al.* (2019), they propose an end-to-end speech conversion model that normalizes different voices to an intelligible output voice. To achieve this result, for every input utterance they synthesized speech and set that as the target speech. In this model, they map an input spectrogram directly to another spectrogram, without utilizing any other intermediate discrete representation. The network that achieves this consists of an encoder, spectrogram and phoneme decoders, followed by a vocoder to synthesize a time-domain waveform. This model can be trained to normalize speech

[2] http://www.voiceitt.com/

from speakers with accents, prosodic differences, and speech signals with background noise into a single canonical speaker–including speech from individuals with speech disorders.

Empirically, the most commonly used voice-assistive technology is the voice amplifier. There have been several studies that have shown that voice amplification systems can be effective in decreasing the amount of vocal strain an individual experiences Morrow and Connor (2011); Roy *et al.* (2003). This is a good solution for individuals who primarily have trouble projecting their voice as it allows them to speak softly and still be heard by their communication partner. However, many individuals with voice disorders are unhappy with voice amplification systems as the root of the problem is in the lack of intelligibility of their voice. Amplifying an unintelligible voice still leads to having difficulties understanding what the speaker says. Another limitation of voice amplification systems is their design. Many amplification systems require the use of bulky microphones and speaker systems as shown in figure 2.3. These designs draw attention away from what the individual is saying and towards the fact that they are using an assistive technology.

Figure 2.3: An Example of the Voice Amplification System, ChatterVox[3]

[3]www.chattervox.com

2.3 Relevant Speech Processing Techniques

Improving the intelligibility of speech sits at the intersection of several other speech processing tasks such as speech recognition (speech to text), speaker enhancement (improving the quality and intelligibility of speech), speech synthesis (text to speech), and voice conversion (source speech to target speech) Purwins *et al.* (2019). In working towards building an intelligibility enhancing system, I will utilize techniques that have been successful in these other disciplines combined with task-specific domain knowledge to enhance intelligibility. But before we can work towards enhancing a metric like intelligibility, we need to be able to quantify that measure whether through subjective or preferably objective metrics.

2.3.1 Intelligibility and Voice Quality Metrics

Intelligibility is the ability of a speech signal to convey meaning and be understood by the listener. One way to measure intelligibility is to have a human orthographically transcribe what they hear Allen (1994). From the ground truth transcript and the predicted orthographic transcription, the word error rate (WER) can be calculated. WER takes the sum of substitutions S, insertions I, and deletions D from the hypothesized word divided by the number of words in the ground truth label N as shown in 2.1.

$$WER = \frac{S + I + D}{N} \tag{2.1}$$

Word Error Rate (WER) is used to measure the performance of the ASR systems Morris *et al.* (2004). While it may seem counter-intuitive, because of this formulation, it is possible to obtain a WER that is more than 100%. Although it is not the standard of the field, sometimes the intelligibility results are shown via the recognition rate (RR). The RR is calculated by the number of correctly recognized words R divided

by the total number of words in an utterance N.

$$RR = \frac{R}{N} \tag{2.2}$$

In the field of speech enhancement, there are many measures of intelligibility and speech quality, for example, the Speech Intelligibility Index Lee *et al.* (2019), Short-Time Objective Intelligibility Taal *et al.* (2010), and Perceptual Evaluation of Speech Quality (PESQ) Rix *et al.* (2001). These metrics were created to measure the effect of additive noise, and are intrusive—they require both a clean speech sample and a noisy speech sample that are time-aligned. Intrusive metrics don't easily translate to the task of improving the intelligibility of dysphonic as we don't have easy access to the 'clean' speech let alone time-aligned 'clean' speech.

2.3.2 Voice Conversion

While it might seem like speech enhancement would be the most relevant field to look to guidance on improving the intelligibility of dysphonic speech as the goal of speech enhancement is to enhance the quality and intelligibility of speech, the field's reliance on intrusive metrics—metrics that require a clean and noisy speech sample of the same length Benesty *et al.* (2005)—make it minimally relevant to the problem of improving the intelligibility of voice disorder speech. The problem of improving the intelligibility of dysphonic speech can be better framed as a voice conversion problem. In the field of voice conversion, the goal is to convert one voice, the source voice into another, the target voice Desai *et al.* (2009); Narendranath *et al.* (1995). The most simple form of voice conversion requires parallel corpora of one speaker saying the same utterances as the other speaker. However, recent techniques have shown that it is possible to convert speech from one speaker into speech from another speaker without parallel speech corpora Hsu *et al.* (2016); Kaneko *et al.* (2019), and without

needing to match the lengths of the input and output speech Zhang *et al.* (2019b).

2.3.3 Latent Representations of Speech

The deep learning revolution has led to neural networks being able to learn useful data representations in both supervised and unsupervised manners. In computer vision, it has been shown that neural networks process information in a hierarchical way—each layer can be interpreted as a feature extractor passing the output on to the next layer Zeiler and Fergus (2013). These feature extractors have also been shown to match some properties of the visual cortex Lee *et al.* (2008). Similarly, when applied to audio, neural networks have been shown to learn auditory frequency decompositions in both music Dieleman and Schrauwen (2014) and speech in the first several layers Jaitly and Hinton (2011). The use of autoencoders, specifically variational autoencoders (VAEs) to extract a latent representation has become a popular and successful method of unsupervised feature learning.

Autoencoders work by imposing an information bottleneck, which effectively compresses the relevant information into a latent representation from which the original signal can be reconstructed. The use of variational autoencoders (VAEs) has been shown to model the generative process of natural speech Chorowski *et al.* (2019). Through disentangling the learned latent representations of speech, Hsu *et al.* (2017b) demonstrates the ability to modify the phonetic content or the speaker identity for speech segments without the need for parallel data.

2.4 Remaining Challenges

While there are several relevant treatments of voice disorders, many individuals still experience some degree of symptoms that harm their ability to communicate and their general quality of life. Voice-assistive technologies have either focused on

recognizing dysarthric speech (a range of speech disorders that manifest differently than voice disorders) or simply amplifying the voice. There is a need for voice-assistive technologies that improve the voice quality and intelligibility of speech from individuals with voice disorders. While speech processing has been able to accomplish a lot with the deep learning revolution, the problem of improving the intelligibility of voice disorder speech has not received much research attention. While there has been a lot of work in the speech enhancement field, this field relies heavily on the ability of the data to be split into time-aligned noisy and clean speech. Unfortunately, in the problem-space of voice disorder speech, we do not have the luxury of working with paired clean and noisy speech samples that are time-aligned. This means that we'll need to employ techniques such as sequence to sequence models, generative adversarial networks, and latent representations of speech to solve this problem.

Chapter 3

INTELLIGIBILITY

3.1 Defining Intelligibility

In this chapter, I will evaluate the existing definitions of intelligibility, and propose a model of intelligibility that is inclusive of the many factors that play a role in whether or not speech is understood by a listener.

In many studies, intelligibility is simply defined as the ability of speech to be understood either by a human or by a computer. While this is a practical working definition, it fails to capture many of the intricacies and complexities involved in the process of communication. Intelligibility is affected by *many* other variables: the length of an utterance, context, language, accent, the speaker's voice quality, nasality, articulation, prosody, age, gender, vocal identity, speaking rate, vowel space area, and a lot more acoustic features like $f0$ variability, formant slopes, modulation energies, residual signal distributions, and cepstral coefficients–all of these are correlated with intelligibility.

Humans are very good at adapting to different accents and speaking styles, making the concept of intelligibility not only relative but also dynamic. The ability of the receiver also plays a role in intelligibility whether the receiver is a human or a computer. This ability comes down to what types of speech the receiver is most 'familiar' with, and for both humans and machines, this ability is a dynamic thing that changes based on the amount of data that the system has been exposed to. Even with all of these complexities and confounding variables, intelligibility is still used as a general indicator of communication ability, and as a metric for the diagnosis and

treatment of speech and voice disorders in a clinical setting. While several models of intelligibility have been proposed from different fields–acoustics, signal processing, speech-language pathology, and performance of voice-based systems–there has yet to be a unifying and holistic model of intelligibility.

In the clinical setting, the most commonly used intelligibility assessment is the clinician's information perceptual estimation of the patient's speech. It is relatively easy to see how there has been a notoriously bad inter-rater agreement between clinicians who rate the intelligibility of individuals with speech and voice disorders Lu and Matteson (2014). There is a lot of evidence that suggests that auditory-perceptual judgments are inherently biased, especially for the clinician whose perceptual systems have adapted to the patient's speech patterns.

3.2 Acoustic Measures of Intelligibility

Machine learning-based speech processing techniques often rely on the extraction of standard speech feature sets. In this section, I'll go through some of the more common speech feature sets (MFCCs, DWTs, and LPC) and describe the advantages and disadvantages as well as known sensitivities for each of the most standard speech feature sets. One of the most common speech input features is mel-frequency cepstral coefficients (MFCCs). MFCCs try to mimic the human ear where frequencies are non linearly resolved across the audio spectrum. To accomplish this, the mel filters are used to symbolize the spatial relationship of the hair cell distribution of the human ear. The mel frequency scale corresponds to a linear scale below 1 kHz and a logarithmic scale above 1 kHz. While MFCCs provide good discrimination, are non-linear, and can capture important phonetic characteristics they are not very robust to noise, and only consider the power spectrum ignoring the phase spectrum of the speech signal Cutajar *et al.* (2013).

Discrete Wavelet Transforms (DWTs) separate the temporal and frequency information in speech signals, analyzing different frequencies with different resolutions. DWTs are capable of compressing a signal without major degradation but are not flexible since the same basic wavelets have to be used for all speech signals Anusuya and Katti (2011).

Linear predictive coding (LPC) is a relatively simple to implement and mathematically precise time-domain approach that attempts to mimic the resonance structure of the human vocal tract when a sound is pronounced, obtaining a good source-to-vocal tract separation. However, LPC is a linear scale that is not necessarily adequate for representing speech production and perception, and the feature components are highly correlated with each other Cutajar *et al.* (2013).

3.3 Speech Intelligibility v Speech Quality v Vocal Quality

The decision to focus on predicting intelligibility over vocal quality came from my understanding of speech quality from a communication network point of view, not a clinical point of view. The research that I had been looking into dealt with speech quality as a measure of the degradation of the speech signal due to lossy compression or noise from the phone system. The vast majority of these metrics are intrusive, and the constraints of the problem that I am interested in solving do not lend themselves to being formulated in a way that we can utilize an intrusive quality metric. There seems to be overlap in the field of the definitions of *quality* when relating to a speech signal. This overlap can be decomposed into *speech quality* and *vocal quality*, *Voice quality* referring to the perceptual construct having to do with voice disorders while *speech quality* focuses on the degradation of a speech signal due noise introduced throughout the telecommunication system.

When deciding what quantity to estimate in order to improve the intelligibility, my understanding of the concept of *speech quality* as a perceptual measure using intrusive metrics in telecommunication systems pushed me towards intelligibility metrics instead as the essence of intelligibility and lack thereof is the main root cause of the decrease of quality of life for individuals with voice disorders. Now understanding the distinction between *speech quality* and *vocal quality*, I would consider including the use of common multidimensional perceptual scales of vocal quality in the values that we will predict about a given speech utterance.

3.3.1 Clinician-Based Voice Quality Assessments

In 1981, Hirano published the Grade, Roughness, Breathiness, Asthenia, and Strain (GRBAS) scale to evaluate the auditory-perceptual judgments of vocal quality. To use GRBAS, a speaker's voice is evaluated by a clinician on a scale of 0-3 where 0: normal, 1: mild degree, 2: moderate degree, and 3: high degree Hirano and McCormick (1986). In 2009, an ASHA special interest group met to promote a standardized approach to evaluating and documenting the auditory-perceptual judgments of vocal quality and the Consensus Auditory Perceptual Evaluation-Voice (CAPE-V) assessment was created Kempster *et al.* (2009). CAPE-V was developed as a tool for clinical auditory-perceptual assessment of voice with the primary purpose of describing the severity of auditory-perceptual attributes of a voice problem in a way that can be communicated among clinicians Kempster *et al.* (2009). This tool was developed to promote a standardized approach to evaluating and documenting auditory-perceptual judgments of vocal quality. In the description of the CAPE-V procedure, the characteristics of the voice are defined as follows Kempster *et al.* (2009):

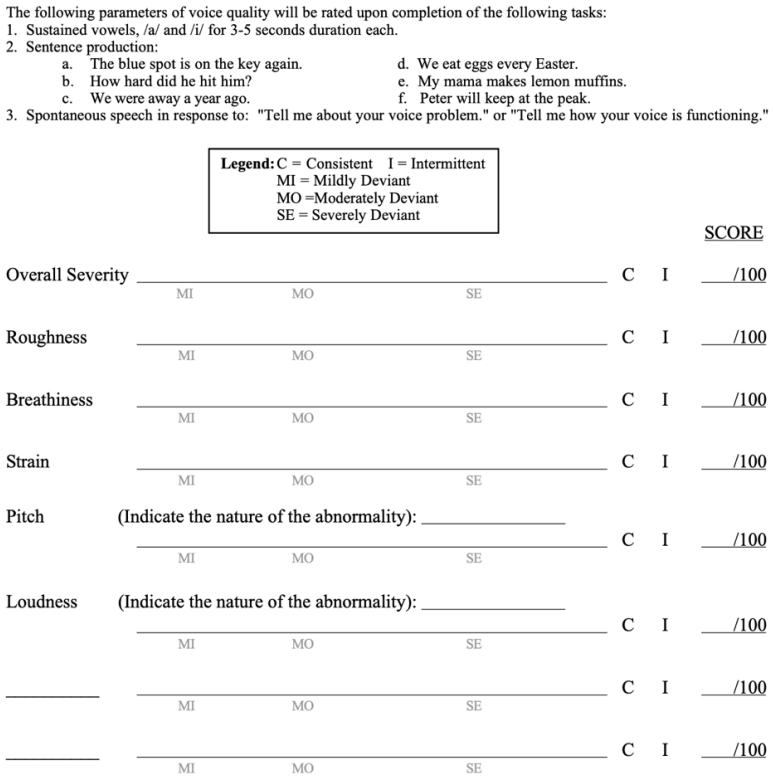
- **OVERALL SEVERITY:** Global, integrated impression of voice deviance.

- **Roughness:** Perceived irregularity in the voicing source.
- **Breathiness:** Audible air escape in the voice
- **Strain:** Perception of excessive vocal effort (hyperfunction)
- **Pitch:** Perceptual correlate of fundamental frequency. This scale rates whether the individual's pitch deviates from normal for that person's gender, age, and referent culture.
- **Loudness:** Perceptual correlate of sound intensity. This scale rates whether the individual's loudness deviates from normal for that person's gender, age, and referent culture.

In CAPE-V, speakers are asked to sustain the vowels /*a*/ and /*i*/ for 3-5 seconds, speak 6 short sentences, and spontaneously describe their voice problem, as shown at the top in Figure 3.1. After the completion of all of these utterances, the above attributes are evaluated by a clinician. The clinician is asked to place a mark somewhere on the scale between MI (mildly deviant), MO (moderately deviant), SE (severely deviant). After they have completed their ratings, the clinician is then asked to physically measure the distance of the line, and the distance from the start to where they made their mark to get a score out of 100 for each attribute. While this process could be automated so that the clinician wouldn't have to pull out their ruler and make 5-7 measurements for each assessment, the process of obtaining a CAPE-V score is a relatively involved one.

3.3.2 Relationship between Quality and Intelligibility

In speech enhancement systems, it is widely recognized that speech enhancement algorithms can harm speech intelligibility—while speech enhancement algorithms are


The following parameters of voice quality will be rated upon completion of the following tasks:
1. Sustained vowels, /a/ and /i/ for 3-5 seconds duration each.
2. Sentence production:
   a. The blue spot is on the key again.
   b. How hard did he hit him?
   c. We were away a year ago.
   d. We eat eggs every Easter.
   e. My mama makes lemon muffins.
   f. Peter will keep at the peak.
3. Spontaneous speech in response to: "Tell me about your voice problem." or "Tell me how your voice is functioning."

Legend: C = Consistent I = Intermittent
MI = Mildly Deviant
MO =Moderately Deviant
SE = Severely Deviant

SCORE

Overall Severity _____ C I /100
MI MO SE

Roughness _____ C I /100
MI MO SE

Breathiness _____ C I /100
MI MO SE

Strain _____ C I /100
MI MO SE

Pitch (Indicate the nature of the abnormality): _____
_____ C I /100
MI MO SE

Loudness (Indicate the nature of the abnormality): _____
_____ C I /100
MI MO SE

_____ _____ C I /100
MI MO SE

_____ _____ C I /100
MI MO SE

Figure 3.1: Screenshot of the CAPE-V scale.

proficient at removing background noise and improving the speech quality of an input signal, the reconstructed signal is often less intelligible than the input signal Xu *et al.* (2017); Kim and Loizou (2010). This indicates that the relationship between speech quality and intelligibility is nonlinear–when speech intelligibility is high, speech quality can vary across the entire spectrum, while when speech intelligibility is low, the speech quality is mainly determined by intelligibility Schiffner *et al.* (2014).

However, the relationship between vocal quality and intelligibility seems to be more clear, when vocal quality is affected by pathology, intelligibility is negatively affected. For adductor spasmodic dysphonia, in particular, speech from individuals with SD both pre- and post- BTX injections–speech that has a decreased vocal quality–is less intelligible than control subjects Bender *et al.* (2004). It has been shown that listeners take longer to transcribe speech from individuals with voice disorders, and

they also are more likely to make mistakes while transcribing voice disorders speech in comparison to control speech Evitts *et al.* (2016).

3.3.3 Advantages and Disadvantages of Multidimensional Vocal Quality Scale

One significant drawback of using a multidimensional vocal quality scale like GRBAS or CAPE-V is the level of human involvement in collecting quality data. A skilled clinician is necessary to rate the speech samples. Collecting auditory-perceptual judgments is a costly and time-consuming endeavor that is wrought with inherent bias, especially when compared to using ASR systems to obtain an estimation of speech intelligibility. In comparison, this method of intelligibility assessment does not involve any human annotation and is fast, cheap and objective. However, a more holistic, standard view of the individual's voice is provided by auditory-perceptual voice quality scales like CAPE-V. Having the auditory-perceptual rating from trained clinicians would be advantageous to have a metric that takes voice quality into account.

3.3.4 Standard Speech Features

Standard speech features for the estimation of vocal quality include jitter, shimmer, cepstral peak prominence, signal periodicity, and harmonic noise ratio. These speech features have shown varying degrees of success in aiding the analysis of pathological speech. Jitter and shimmer are perturbation measures commonly used in the acoustic analysis of pathological speech. Jitter is a measure of the frequency instability in a voice, while shimmer is a measure of the amplitude instability. Jitter and shimmer are shown on a speech signal in Figure 3.2. The ratio of the harmonic component to noise component (the Harmonic to Noise Ratio) yields information on the ability of the individual to coordinate source and filter acoustics Teixeira and Fernandes (2015).

Both jitter and shimmer can be calculated in several ways, the absolute jitter is the cycle-to-cycle variation of the fundamental frequency as shown by Equation 3.1. The relative jitter $J_{relative}$ can be calculated as the average absolute difference between consecutive periods, divided by the average period. Jitter can also be calculated as the Relative Average Perturbation (J_{RAP}), the average absolute difference between a period and the average of it and its two neighbors, divided by the average period. The last common way to calculate jitter is the Five-Point Period Perturbation Quotient (J_{PPQ5}), which is computed as the average of it and its four closest neighbors divided by the average period. Both J_{RAP} and J_{PPQ5} are expressed as percentages Teixeira and Fernandes (2015). The J_{RAP} is useful when we want to ignore the physiological difference in jitter between males and females–males generally have a longer glottal period than females and have a higher absolute variation. The difference between males and females is no longer relevant when using J_{RAP} and J_{PPQ5} as these are more measures of the local variation.

$$J_{abs} = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i-1}| \tag{3.1}$$

Shimmer is expressed as the variability of the peak-to-peak amplitude in decibels as shown in Equation 3.2 where A_i is the peak-to-peak amplitude and N is the number of periods. Similarly to jitter, shimmer can also be expressed as relative Shimmer, three-point Amplitude perturbation quotient, or a five-point amplitude perturbation quotient Teixeira and Fernandes (2015).

$$S_{abs} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left| 20 * log \left(\frac{A_{i+1}}{A_i} \right) \right| \tag{3.2}$$

The Harmonic to Noise Ratio (HNR) indicates the overall periodicity of the voice signal by quantifying the ratio between the periodic (harmonic) and aperiodic (noise)

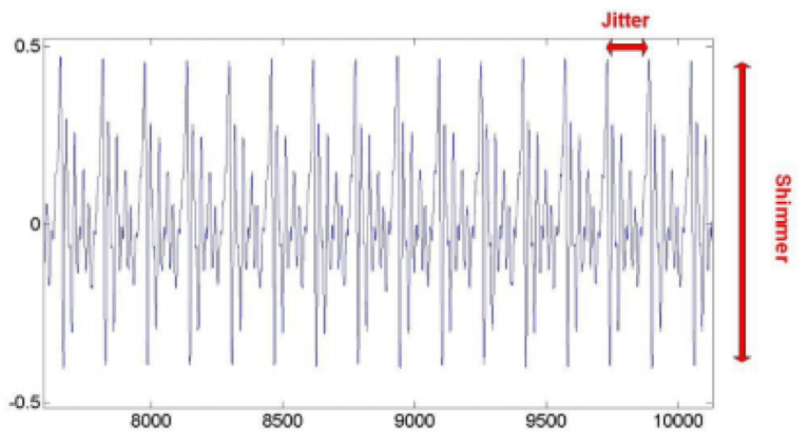


Figure 3.2: Jitter and Shimmer Perturbations in a Speech Signal.

components. HNR is given by Equation 3.3 where $AC_V(0)$ is the autocorrelation coefficient at the origin consisting of all the energy of the signal, $AC_V(T)$ is the component of the autocorrelation corresponding to the fundamental period. The difference between $AC_V(0)$ and $AC_V(T)$ is assumed to be the noise energy Boersma (2000). Harmonic-Noise Ratio (HNR) has been shown to effectively correlate with a hoarse vocal quality Ferrand (2002); Yumoto *et al.* (1982).

$$HNR = 10 * log_{10} \frac{AC_V(T)}{AC_V(0) - AC_V(T)} \tag{3.3}$$

Another acoustic feature that is mentioned a lot in regards to dysphonic speech is the Cepstral Peak Prominence (CPP). CPP has shown to be a more reliable measure of dysphonia than jitter, shimmer, and HNR Heman-Ackah *et al.* (2003). CPP is a measure of the relative amplitude of the cepstral peak prominence in relation to the expected amplitude as derived via linear regression. This measure reflects the degree of regularity or periodicity in the voice signal. Higher CPP values reflect greater periodicity. In Hillenbrand and Houde (1996), signal periodicity is shown to be highly correlated with the breathiness quality of speech.

3.3.5 Speech Feature Sensitivity in low SNR

Standard speech features are often analyzed by looking at the relationship between a speech feature and perturbation measures. In Fraile and Godino-Llorente (2014), the relationship between cepstral peak and perturbation measures is analyzed. There is an inverse relationship between measures of amplitude, frequency, and noise perturbations and the amplitude A_1 of the cepstral peak. The dependence between A_1 and jitter is more significant than the dependence between A_1 and shimmer and noise. There still exists an inverse relation between shimmer and A_1 and noise and A_1, so that with an increase in the standard deviation (σ) of the shimmer or noise, there will be a decrease in the intensity (dB) or the cepstral peak, however, this decrease will not be as significant for a similar increase in σ as it would be for jitter. These results are consistent with Hillenbrand and Houde (1996); Samlan *et al.* (2013) This means that should CPP be chosen to use as a speech feature for the estimation of vocal quality in low SNR, the slight decrease of the intensity of the cepstral peak with the σ of the noise will have to be accounted for.

In de Krom (1993), the authors demonstrate a major effect of both noise and jitter on HNR, in that HNR decreases almost linearly with increasing noise levels or increasing jitter. This indicates that using HNR in low SNR scenarios may not be the best choice for a speech feature.

In Kreiman *et al.* (2002), they conduct a clever study that evaluates how jitter, shimmer, and noise are perceived. In this study, they play a pathological voice for a listener and ask the listener to adjust different parameters (jitter, shimmer, and noise) of a synthesized voice to match the characteristics of the pathological voice. They found that there was no correspondence between jitter and shimmer and the perceived vocal quality. They argue that jitter and shimmer cannot be used as reliable or valid

measures of perceived vocal quality. However, the found noise to be a highly salient perceptual attribute of pathological voices. Listeners' noise responses varied much less than their jitter and shimmer responses. This study showed that listeners are highly insensitive to differences in the amounts of jitter and shimmer in a voice. This may be due to listeners not being able to differentiate between jitter and shimmer. The authors argue that jitter and shimmer do not perceptually distinguish mild from severe vocal pathology. This finding shows that jitter and shimmer are not intuitive perceptual features that distinguish pathological speech, however, other studies have successfully used jitter and shimmer to computationally recognize pathological speech Adnene *et al.* (2003); Dibazar and Narayanan (2002).

In summary, using standard speech features that are correlated in dysphonic speech in low signal to noise ratios will be tricky. However, it seems like jitter and shimmer are the least relevant acoustic features in determining whether and to what degree a voice is dysphonic. Using CPPs as a feature to help determine vocal quality seems like a relatively promising way to proceed if we need to use a standard speech feature to capture aspects of vocal quality.

3.4 Defining Intelligibility Multidimensionally

Intelligibility can be constructed as a binary thing: did the listener receive the information that they needed or not, or a scalar value: to what extent was the message received. There's a certain intelligibility threshold–a point where someone who is originally unintelligible becomes intelligible to a listener. Every listener's threshold is slightly different based on their previous experience: what languages are they familiar with, what kinds of accents do they understand, do they have experience communicating with individuals with voice disorders, etc. To help a speaker be more intelligible to communication partners, their intelligibility needs to cross the listener's

threshold. This increase in intelligibility can come from either our brain's natural tendency to adapt and learn the speech patterns of others, from the speaker's ability to adapt how they're speaking to be better understood, or it could also potentially be moved towards the intelligibility threshold by the aid of a voice-assistive technology. There are also instances where the listener might not understand very many of the words in a sentence, however, whether it was through the context, non-verbal cues, or other idiosyncrasies of human interaction that the message is still clearly understood.

Variables that affect intelligibility include local acoustic features, global acoustic features, linguistic data, the language it was spoken in, as well as the listener's physical and prior experience with similar speakers. The amount of energy that a listener must put in is also a significant

Chapter 4

A QUALITATIVE ANALYSIS OF THE NEEDS OF INDIVIDUALS WITH VOICE DISORDERS

4.1 Overview

One of the first steps towards the goal of helping people with voice disorders be better understood was to dig deep into the problem and learn more about how voice disorders affect the lives of people with dysphonia. To accomplish this, I conducted two surveys to learn more about the experience of individuals with dysphonia. The first survey focused on qualitative feedback about the overall effects of voice disorders, while the second survey focused on and 'double-clicked' on some of what was found in the first survey as well as dug deeper into people's experiences using voice-assistive technologies.

Individuals with voice disorders often find it difficult to be understood while speaking on the phone, conversing in a noisy environment (restaurants, parties, etc.), ordering at a drive-thru and meeting someone new for the first time. Engaging in social interactions and completing the tasks necessary to acquire, maintain, or advance in a career become particularly demanding. These trying situations often lead to low self-esteem and confidence, as well as feelings of isolation, anxiety, frustration, stress, and sometimes depression.

Individuals with voice disorders work hard to be understood in day-to-day interactions. This affects their social life, their career, and their emotional wellbeing. When it is difficult to socialize, it is easy to withdraw from social situations and become isolated. This lack of social interaction can lead to a decrease in self-esteem

and confidence, much like being consistently overlooked when it comes to obtaining a promotion in your job. These are all things that individuals with voice disorders deal with daily, largely because they are not easily understood.

While 88.83% of the respondents have experienced a limitation or a barrier because of their voice disorder, only 1.75% have used an assistive technology designed to help them be better understood. Of the respondents, 63.16% indicated that they would use voice-assistive technology. Very few voice-assistive technologies have been developed to help people with voice disorders be better understood, despite individuals with voice disorders being open and interested in using these kinds of technologies. This suggests that there are opportunities for innovation in creating voice-assistive technologies that help to more easily facilitate day-to-day interactions.

4.2 Methodology

Both surveys were administered using Google Forms and were distributed to the members of the National Spasmodic Dysphonia Association through email and social media. The inclusion criteria for both surveys was that participants had to be 18 years or older, and self-identify as having a voice disorder. Participants were allowed to skip any question they did not want to answer. In the first survey, we surveyed 471 participants (386 female, 76 male, and 9 who did not disclose), who have a voice disorder.

4.2.1 Types of Survey Questions

There were three main types of survey questions that we used to learn more about the experiences of individuals with voice disorders: open-ended, multiple select, single select, and Likert-scale questions. Open-ended questions are asked such that the respondent has an opportunity to write out their thoughts and opinions. In

multiple select questions, the respondent can select as many answers as they would like to answer the question. In single-select questions, the respondent must make a choice and only select one of the provided answers. Likert-scale questions have the format 'Please rate the extent to which you agree or disagree with the following statements', followed by the statement in question, and then the respondent is asked to choose between 'strongly disagree, 'disagree', 'neutral', 'agree', or 'strongly agree', or a number between 1 and 5 where 1 is 'Strongly disagree' and 5 is 'Strongly agree' as shown in Figure 4.1.

Figure 4.1: Example of a Likert Scale Question

4.3 Initial Voice Disorder Survey

There were three main parts of the first survey. The first section checks that the participants meet the criteria to participate. The second section asks for details about the respondent's voice, while the third section asks questions about what kinds of technology individuals with voice disorders use. The survey ends with a series of open-ended questions about how having a voice disorder has impacted their lives.

In the screener section of the survey, we ask participants to acknowledge that they meet our requirements to take the survey–that they're more than 18 years old and have a voice disorder. We also ask for some primary demographic information: their location (city/state), gender, and age.

In the first part of the survey, we begin to dive into learning more about the participant's voice. We ask questions about their voice disorder, including diagnosis, age of onset, voice description, voice treatment, and efficacy of treatment. We also

Table 4.1: Survey 1 Part 1: About Your Voice

Question	**Type**
What voice disorder do you have?	Multiple select
Do you have a more specific diagnosis for your voice disorder?	Open-ended
How long has it been since you've been diagnosed with a voice disorder?	Single select
At what age did your voice disorder first develop?	Number
How would you describe your voice?	Open-ended
What kind of treatment have you received for your voice?	Multiple select
"My voice is more easily understood after treatment"	Likert
"Most of the time people understand my voice."	Likert
"People are less likely to understand my voice when we first meet than after I've known them for a while"	Likert
Which situations do you find it difficult to be understood?	Multiple select

ask respondents to identify situations where they find it difficult to be understood. These questions are shown in more detail in Table 4.3.

In the second section of the survey, expanded on in Table 4.3, we ask individuals about the technology that they use. This involved non-voice related technology such as electronic devices (mobile devices, laptop, desktop, etc), and web browsers, as well as voice recognition systems. We ask about non-voice technology to obtain a baseline of what kinds of platforms we might develop systems for in the future. We use this information later in this dissertation to determine what browsers/devices we developed a data collection system for. Concerning voice recognition systems, we asked respondents what voice recognition systems they had used, as well as which ones had worked better than others, and which ones had performed worse than others.

Table 4.2: Survey 1 Part 2: Technology Use

Question	**Type**
Which electronic device(s) do you use on a daily basis?	Multiple select
Which internet browser (s) are you most comfortable or familiar with?	Multiple select
Which voice recognition systems have you used?	Multiple select
Are there any voice recognition systems that seem to recognize your voice better than others?	Multiple select
Are there any voice recognition systems that seem to recognize your voice worse than others?	Multiple select
"Voice Recognition systems recognize my voice most of the time:"	Likert
"I would use a voice assistive technology that helped me to be better understood"	Likert

We also asked respondents how they felt about using a voice assistive technology that helped them be better understood.

In the third section of the survey, expanded on in table 4.3, we asked a series of open-ended questions. These questions revolved around how voice disorders affected their lives. We asked them to identify any activities that they no longer could participate in due to their voice disorder, as well as how their voice disorder has affected their career, and social life. We also ask them again, to identify any situations that they've found it particularly difficult to be understood. We finish the survey by asking if they've ever faced any discrimination due to their voice disorder and allow them an opportunity to add anything that they'd like to share.

Table 4.3: Survey 1 Part 3: Open-Ended Responses

Question	**Type**
How does your voice disorder affect your life?	Open-ended
List any activities that you can no longer participate in due to your voice disorder	Open-ended
What kinds of technologies would you like to see developed for your voice disorder?	Open-ended
How does your voice disorder affect your social interactions?	Open-ended
How does your voice disorder affect your career?	Open-ended
Are there any situations in which is is particularly difficult to be understood?	Open-ended
Have you faced any kind of discrimination due to your voice disorder?	Open-ended
Is there anything else that you'd like to add?	Open-ended

The National Spasmodic Dysphonia Association was gracious enough to assist with the recruitment of participants. It is through their support that we were able to obtain such significant sample sizes, something that is a great feat, especially considering that Spasmodic Dysphonia is technically a rare disorder.

4.3.1 Initial Voice Disorder Survey Results

We surveyed the needs of individuals with voice disorders to get a better idea of the effect of voice disorders on people's lives to identify challenges that could be aided by the use of voice-assistive technologies. We present the experience of individuals with voice disorders and analyze the results concerning the challenges experienced daily.

The results of the first part of the survey mirrored what has been seen in other related research on the demographics of individuals with SD Adler *et al.* (1997); Patel *et al.* (2015); Roy *et al.* (2004): the respondents were primarily female (82.31%), the average age was 62 ± 11.84 years old, the onset of SD took place on average at the age of 45 ± 14 years, and 10.30% of the respondents were in the teaching profession.

When asked in an open-ended format *'How does your voice disorder affect your life?'*, the most prominent responses can be categorized into three distinct groups: social interactions (41.11%), career (29.33%), and emotional wellness (30.95%). The coded response rates are shown in Table 4.4. Following these three, other significant effects of having a voice disorder are reported as not being able to be understood when using the phone (18.71%), general communication (15.94%), and overall confidence (11.28%).

4.3.2 Diagnosis and Treatments

The majority of the respondents have a diagnosis of SD (95.6%), while 21.0% have Vocal Tremor, and 9.4% have Muscle Tension Dysphonia. When asked if they had a more specific diagnosis (such as ADSD or ABSD), 49 of the respondents reported that they were diagnosed with ADSD, while 31 individuals reported being diagnosed with ABSD, and 17 reported having mixed ADSD, and ABSD. These voice disorders are tightly coupled and exhibit very similar symptoms. As these disorders have similar causes, and symptoms, it makes sense that they would be treated similarly. The most common treatment method is Botox injections which 81.6% of the respondents had been treated with. The second most common treatment is speech therapy of which 65.4% of the surveyed population had participated in. These treatments seem to be relatively effective as 53.4% of the responses either agreed or strongly agreed with the statement *I would say that my voice is more easily understood after treatment*, while

Table 4.4: Primary Effects of Living with a Voice Disorder.

Response	**Response Rate**
Decreased Social Interations	41.11%
Decreased Emotional Wellness	30.95%
Negative Impact on Career	29.33%
Difficulty Using the Phone	18.71%
Decreased Communication	15.94%
Decreased Confidence	10.85%

when responding to the prompt *I would say that most of the time people understand my voice*, only 34.2% of the respondents would agree or strongly agree.

4.3.3 Difficult Situations and Limitations

The survey addressed the idea of difficult situations, barriers, and limitations through two questions, one close-ended 'choose all that apply' question, and another open-ended prompt that asked respondents to "List any activities that you can no longer participate in due to your voice disorder." From these two questions, we have identified several situations that are particularly difficult for individuals with voice disorders to participate in. These situations are shown in Table 4.5 in decreasing order of frequency. The most commonly reported difficult situation was speaking on the telephone (91.4% of respondents), followed by speaking in a noisy environment (87.5% of the respondents). Ordering food at a drive-thru was reported as one of the most difficult situations to be understood as 339 respondents indicated (74.3%). Speaking with a new person was also reported to be difficult by 60.5% of respondents. The least mentioned difficult situation to be understood is speaking with family and friends, however, 30.0% still reported having difficulty speaking with family and friends.

Table 4.5: Situations Identified as Particularly Difficult

Option	**Response Rate**
Speaking on the phone	90.95%
Speaking in a noisy environment	86.55%
Ordering at a drive thru	74.08%
Meeting someone new for the first time	60.88%
Talking with family or friends	28.85%
Stress*	2.69%

When asked to list which activities they no longer participate in because of their voice disorder, the results were very similar to the results from the close-ended question. In the open-ended question, respondents reported the loss of general communication abilities (32.91%), singing (27.00%), social interactions (23.31%), public speaking (22.42%), group conversations (21.19%), and teaching (9.21%). Only 11.17% of the participants responded that they have no limitations from their voice disorder. It follows that 88.83% of the respondents reported that they have experienced a limitation or barrier because of their voice disorder.

4.3.4 Social Effects

Not being able to be understood or heard has a significant impact on both the quantity and quality of social interactions. We asked participants '*How does your voice disorder affect your social interactions?*'. Of the respondents, 91.65% experienced a decline in the amount and quality of their social interactions. In general, the responses indicated that social interactions are stressful and difficult for individuals with voice disorders. So much so that 39.91% of the respondents indicated that they actively avoid participating in social gatherings.

In 14.62% of the responses, the reason for avoiding social interactions was reported to be that these gatherings usually take place in noisy settings. Environments with a lot of noise make it even more difficult for individuals with voice disorders to be understood. Gatherings in noisy environments are often frustrating for individuals with voice disorders because they feel like they are being ignored, overlooked, and generally disregarded. Another 9.28% of respondents mentioned being embarrassed by their voice, or fearful of how others would perceive them. Some respondents—10.44%—reported that having a voice disorder makes it much more difficult to meet new people. Withdrawing from social engagements can lead to isolation and depression, a scenario that 26 of the respondents (6.03%) acknowledged happening to them.

A few individuals decided to focus their response on how they cope with the fact that their voice disorder makes social interactions more difficult. The most common coping strategy focused on informing the communication partner of their voice disorder. Through explaining the condition, the communication partner reportedly is more likely to give the individual with a voice disorder more time to answer. Other respondents—6.49%—focused on the opportunity that their voice disorder gave them to build their listening skills, and be more self-reflective.

4.3.5 Career Effects

Having a voice disorder can have a significantly negative effect on a career. When asked to respond to the question '*How does your voice disorder affect your career?*,' respondents made it clear that it is difficult to acquire, maintain, and advance in a career when diagnosed with a voice disorder. Of the responses, 81.22% discussed the negative effect that their voice disorder has had on their career. Even further, 122 people shared that their careers had ended due to their voice disorder. As reported,

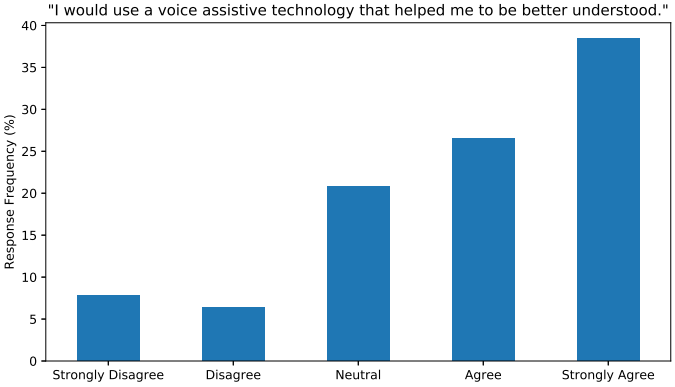


Figure 4.2: Response Rate to the Question *'I Would Use a Voice Assistive Technology that Helped me to Be Better Understood'.*

28.71% of the respondents have lost a job, or have to retire early because of their voice. Only eight respondents—1.75%—have used some sort of assistive technology, namely, an amplifier, to extend their careers.

4.3.6 Emotional Wellness

With fewer social interactions, difficulties finding, keeping and advancing in a career, and the prevalence of discrimination, it is not difficult to see the connection between having a voice disorder and a decline in emotional well-being. When asked to respond to the question '*How does your voice disorder affect your life?*' 30.95% of the respondents mentioned one or more of the following: stress, anxiety, frustration, isolation, or depression.

When asked if the participant had ever experienced discrimination due to their voice disorder, 59.66% indicated that they had experienced some form of discrimination because of their voice. Most of this discrimination manifests in people with voice disorders being overlooked for promotions because of their voice, or not getting jobs or interviews because of the way they sound.

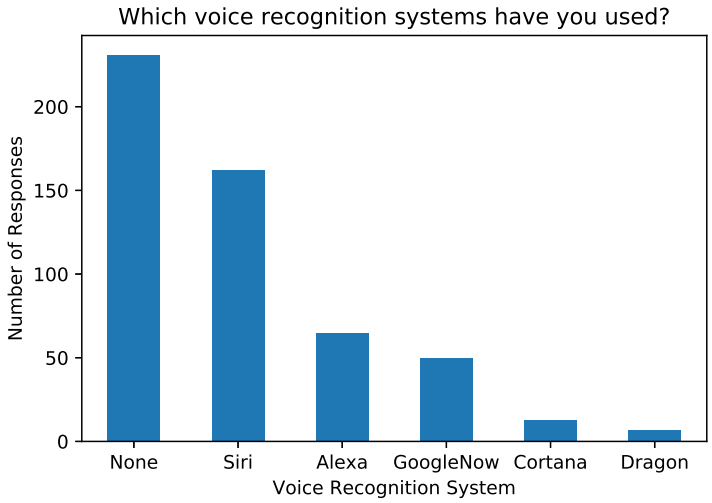


Figure 4.3: Responses for *'Which Voice Recognition Systems Have you Used'.*

4.3.7 Technology Usage

Two main pieces of technology usage were asked about in the initial voice disorder survey: voice recognition system usage, and general device/browser configuration. Questions were asked about what voice recognition systems respondents had used as well as if the respondents noticed any voice recognition systems working better than other voice recognition systems. Most respondents (50%) had not used any voice recognition systems while of those who had used voice recognition systems, Apple's Siri®was used the most (35.1%), followed by Amazon's Alexa®(14.1%), GoogleNow®(10.8%), Cortana®(2.8%), and Dragon Naturally Speaking®(1.5%).

4.4 Follow-Up Voice Disorder Survey

4.4.1 Limitations to Initial Voice Disorder Survey

There were a few themes that came up in the initial voice disorder survey that required further evaluation. There were specific things that came up that we didn't

know to ask about in the initial survey until we had the results to guide a follow-up survey. The initial survey asked relatively broad, qualitative questions, and the follow-up survey dug a little bit deeper to be able to obtain some quantitative data around the needs of individuals with dysphonia.

One main theme that was continually brought up in the results of the initial voice disorder survey was the impact of voice disorders on an individual's emotional wellbeing. However, we did not ask any specific questions to quantify how many individuals with SD experience anxiety, depression, isolation, or frustration because of their voice disorder. Another oversight in the initial survey was that we didn't ask them about what voice treatment they used. In the follow-up survey, we added a few questions about if they use Botox injections or not. We also found that many people were eager to share their coping mechanisms and pieces of advice as to how they've learned to deal with their voice disorder. After learning more about the population of individuals with SD in the first survey, we also learned that there's a big opportunity to develop voice-assistive technologies, but we didn't ask any questions about voice-assistive technologies in the original survey. In the penultimate question of the initial survey, we asked participants if they've ever been discriminated against. In this question, we were very surprised by how many individuals told stories of discrimination, particularly in the field of employment. We didn't specifically ask for employment status in the original survey, nor did we ask if the respondent had lost their job or retired early because of their voice. We used the follow-up survey as an opportunity to obtain more detailed information about these themes that appeared in the initial voice disorder survey.

In the first part of the follow-up survey, we check to make sure that the respondents met the criteria of being over the age of 18 and having a voice disorder. After that, we dive into general questions about their voice, in a very similar manner to how we

Table 4.6: Survey 2 Part 1: General Questions

Question	**Type**
What voice disorder do you have?	Multiple select
Do you regularly receive Botox injections?	Yes/No
If so, how often do you receive injections?	Open-ended
If you do not receive Botox, why not?	Open-ended
Are you able to sing when your voice is at it's worst?	Yes/No
Are you able to sing when your voice is at it's best?	Yes/No
What is your current employment status?	Multiple select
Have you ever lost a job because of your voice disorder?	Yes/No/Maybe
Did you retire early because of your voice disorder?	Yes/No/NA
If you were designing a voice-assistive device, what functionality and characteristics would it have?	Open-ended
Have you used an assistive technology for your voice?	Yes/No/Maybe

asked in the first survey, however adding questions about whether or not they receive Botox injections for their voice, and if not why they don't. The full questions in this section are shown in Table 4.4.1. We also learned in the initial survey that many individuals with voice disorders are unable to sing. We wanted to obtain a more quantitative measure of what percent of individuals can sing with a voice disorder, so we asked if the respondent can sing when their voice is 'bad' and 'good'. We also asked respondents to provide their employment status as well as if they have ever lost a job or retired early because of their voice disorder. We ended the first section of the survey by asking what functionality and characteristics they would like to see in a voice-assistive technology as well as if they had used a voice assistive technology before. If they had used a voice-assistive technology they moved on to the next section

Table 4.7: Survey 2 Part 2: Voice-Assistive Technology

Question	**Type**
What kind of assistive technology have you used for your voice?	Multiple select
Do you still use a voice assistive technology?	Single select
Describe how you use the voice assistive technology (times when it's useful, times when it is not).	Open-ended
What contributed to you continuing to use the assistive technology or stopping use of the technology?	Open-ended
"I am satisfied with my experience with assistive technologies"	Likert
What would you change about the technologies that you've used?	Open-ended
If you were designing a voice-assistive device, what functionality and characteristics would it have?	Open-ended
Is there anything else that you'd like to add about your experience with voice-assistive technologies?	Open-ended

about their experience using voice-assistive technologies, however, if they have not had that experience, they skipped this section and moved on to the effects of voice disorders section.

In the Voice-Assistive Technology Experience section of the survey, we were interested in learning more about the experience of respondents who had used voice-assistive technologies in the past. The questions used in this section are shown in Table 4.4.1. We asked respondents who made it to this section what kind of technology they have used for their voice, and whether or not they still used the technology. We asked them to describe how they used the technology–times when it was useful, and times when it was not. We asked them how satisfied they were with their voice-

assistive technology as well as what they would change about the technologies that they have used. We asked them what kind of voice-assistive technology they would like to see developed, and what characteristics/functionality it would have.

In the Effects of Voice Disorders section of the follow-up voice disorder survey, we asked many Likert scale questions. The questions asked in this section are shown in Table **??**. These questions focused a lot more on the emotional wellbeing of the respondent than the effects of voice disorder section in the initial voice disorder survey did. We also ask a couple of open-ended questions as follow-up questions to the Likert-scale questions about how their voice disorder has affected their career, social life, and emotional wellbeing.

Respondents were also asked to respond to a few questions about the mechanisms that they use to cope with their voice disorder. In the initial survey, there were several coping mechanisms brought up without specifically asking a question about coping mechanisms. The follow-up survey asked for more information on the coping mechanisms mentioned in the initial survey: using alcohol or cannabidiol (CBD), informing communication partners, or generally decreasing stress and anxiety. The specific questions asked are shown in Table 4.4.1.

4.5 Follow-Up Voice Disorder Survey Results

In the follow-up voice disorder survey, we obtained 453 responses from individuals with voice disorders.

4.5.1 General Questions

Voice Disorder Distribution

The distribution of voice disorders was very similar to the distribution from the initial survey, with the majority of respondents having Adductor Spasmodic Dysphonia

Table 4.8: Survey 2 Part 3: Effects of Voice Disorders

Questions	**Type**
"Before having a voice disorder, I would describe myself has having anxiety"	Likert
"After having a voice disorder, I would describe myself as having anxiety"	Likert
"Having a voice disorder increases my anxiety"	Likert
"Having a voice disorder increases my level of stress"	Likert
"Because of my voice disorder, I feel isolated"	Likert
"Most of the time, my voice is intelligible to communication partners"	Likert
"After treatment, my voice is intelligible to communication partners"	Likert
"Before having a voice disorder, I was confident in myself"	Likert
"After having a voice disorder, I am very confident in myself"	Likert
"My voice disorder has negatively impacted my career"	Likert
Please describe how having a voice disorder has affected your career	Open-ended
"My voice disorder has negatively affected my social life"	Likert
Please describe how having a voice disorder has affected your social life	Open-ended
"My voice disorder has negatively affected my emotional wellbeing"	Likert
Please describe how having a voice disorder has affected your emotional wellbeing	Open-ended

Table 4.9: Survey 2 Part 4: Coping Mechanisms

Questions	Type
What coping mechanisms have you found useful?	Open-ended
When your voice is 'good', do you tell people about your voice disorder?	Open-ended
When your voice is 'bad', do you tell people about your voice disorder?	Open-ended
How does stress affect your voice?	Open-ended
Have you used CBD oil as a treatment for your voice disorder? If so, how has it affected your voice?	Open-ended
How does alcohol affect the symptoms of your voice disorder?	Open-ended

(66.4%), followed by Abductor Spasmodic Dysphonia (34.3%), Vocal Tremor (21.1%), Essential Tremor (14.1%), and Muscle Tension Dysphonia (10.3%).

Botox Injections

We asked respondents if they regularly received Botox injections, and 48.1% of respondents responded 'Yes', while the other 51.9% responded 'No'. We asked those that received botox injections how frequently they received injections, and the average of the responses was every 3.8 months. For those who do not receive Botox injections, we asked them to explain why they don't use it as a treatment. The primary reasons cited were that it didn't sufficiently improve their intelligibility, it was too costly, the side effects (breathy voice for a while) were not worth it, the temporary nature of the treatment, or their voice disorder not being severe enough to warrant Botox injections.

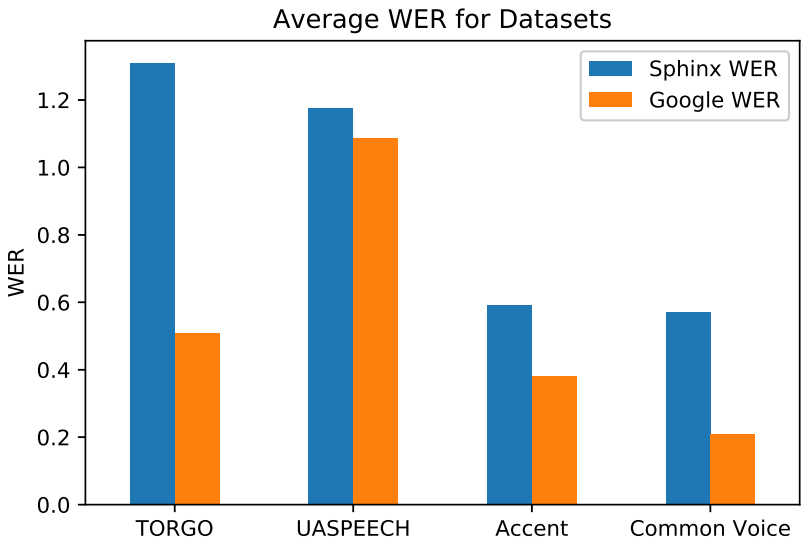


Figure 4.4: The Average WER Per Dataset Represented in Metadataset

Singing

In response to the question 'Are you able to sing when your voice is at its worst?', 72.3% responded with 'No'. When asking 'Are you able to sing when your voice is at its best?', the percent of individuals who responded with 'No' decreased to 45.3%.

Employment Status

When asked 'What is your current employment status?', 49.9% of respondents reported to be retired, 28.4% Employed full-time, 12.3% employed part-time, and 6.7% were unemployed. In response to 'Have you ever lost a job because of your voice disorder, 71.8% responded 'No', 14.4% responded 'Maybe', and 13.7% responded 'Yes'. In response to 'Did you retire early because of your voice disorder?' 44.9% responded 'No', 30.3% responded 'Not retired', and 24.7% responded 'Yes'.

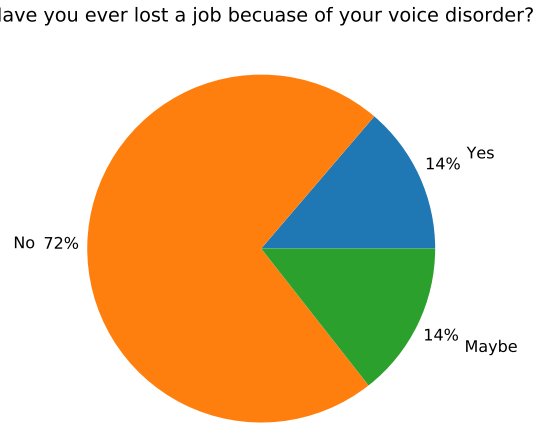


Figure 4.5: Responses to the Question *'Have you Ever Lost a Job Because of your Voice Disorder?'*.

4.5.2 Voice Assistive Technology Experience

Before splitting the survey up into individuals who have used a voice-assistive technology and individuals who have not, we asked everyone 'If you were designing a voice-assistive device, what functionality and characteristics would it have?'. The main theme of the responses to this question is that respondents are looking for technologies to help them be better understood, they want the technology to be inexpensive, unobtrusive, and flexible.

When asked 'Have you used assistive technology for your voice?', 20.5% of respondents responded with 'Yes', 3.4% responded with 'Maybe', and 76.1% responded with 'No'. The next question was 'What kind of assistive technology have you used for your voice', 85.4% of the respondents have used a voice amplification system, and 16.5% indicated that they had used text to speech, 1.9% had used Augmentative Alternative Communication devices. When asked 'Do you still use a voice assistive technology?', 56.2% responded 'No', 31.4% responded 'sometimes', and 12.4% responded with 'Yes'.

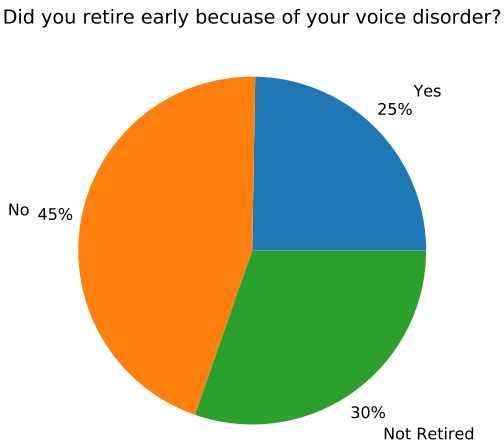


Figure 4.6: Responses to the question *'Did you Retire Early Because of your Voice?'*.

When asked to 'describe how you use the voice assistive technology (times when it's useful, and times when it is not)', the general themes of the responses were that they used amplification systems when their voice was particularly bad, but it only helps with volume, not clarity. Some individuals would use text to speech during times of complete voice rest. Most respondents agree that the voice-assistive technology that they've used has not been very useful and that communication partners still can't understand what they're saying.

Respondents were asked 'What contributed to you continuing to use the assistive technology or stopping use of the technology', and in the general themes of the answers are that either they used the technology when they were at work, but now are retired, they were put off by the time-delay in conversation, the system was too cumbersome/inconvenient. In general, there's a lot of frustration of only improving the volume, not the vocal quality/clarity.

When asked to rate the extent to which respondents agree or disagree with the statement 'I am satisfied with my experience with assistive technologies', 50.5% of

respondents either strongly disagreed or disagreed, while only 15.4% of respondents either strongly agreed or agreed with the statement and 34.0% of the respondents remained neutral.

The respondents were also asked 'What would you change about the technologies that you've used'. In general, respondents replied with suggestions that the devices be smaller, less obtrusive, and improve not only volume but also voice quality. Things like avoiding wires and feedback were also mentioned. Another thing that seems important is to minimize the delay in the system to be able to maintain the natural pace of a conversation.

4.5.3 Effects of Voice Disorders

Much like the initial survey, in the follow-up survey, our goal was to learn more about the effects of voice disorders on the lives of those affected by dysphonia. In this section, we asked questions that were more targeted than the initial survey and dug into more specific aspects of the lives of individuals with dysphonia.

Emotional Wellbeing

One aspect that came up in the initial survey, but that wasn't specifically asked about is the emotional wellbeing of individuals with dysphonia. In this survey, respondents were asked specifically about how their emotional wellbeing has changed between when they received their diagnosis and now.

Figure 4.7 shows that there is a very clear pattern of respondents strongly disagreeing or disagreeing (combined 61.7%) with the statement 'Before my voice disorder, I would describe myself as having anxiety', while strongly agreeing or agreeing (combined 64.6%) with the statement 'After my voice disorder, I would describe myself as having anxiety'. The weighted average of the responses to the before statement

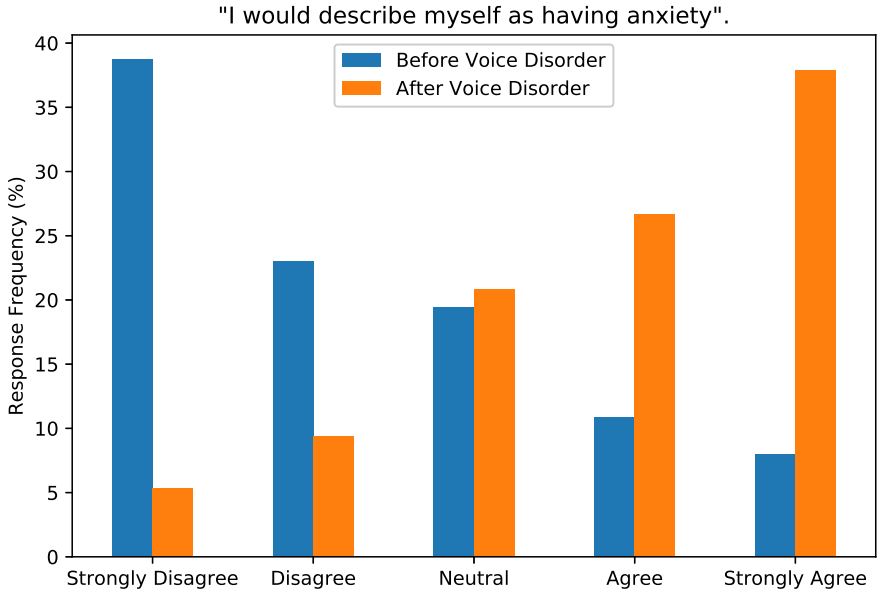


Figure 4.7: Responses to the Likert Question *'I Would Describe Myself as Having Anxiety'* Before and After their Voice Disorder Diagnosis.

was 2.26 (where 1 is Strongly Agree and 5 is Strongly Disagree), while the weighted average of the after statement was 3.82.

When respondents were asked to rate the extent to which they agree or disagree with the statement 'Having a voice disorder increases my anxiety, 75.5% of the respondents either strongly agreed or agreed and the weighted average was 4.08. The same pattern was found when the same question was asked referring to stress rather than anxiety–76.5% either strongly agreeing or agreeing with a weighted average of 4.16. These results are shown in Figure 4.8.

Respondents were asked to rate how much they agreed/disagreed with the statement 'Before having a voice disorder, I was very confident in myself.' and 'After having a voice disorder, I am very confident in myself'. The results are shown in Figure 4.9. Before diagnosis, 76.1% either agreed or strongly agreed with the statement

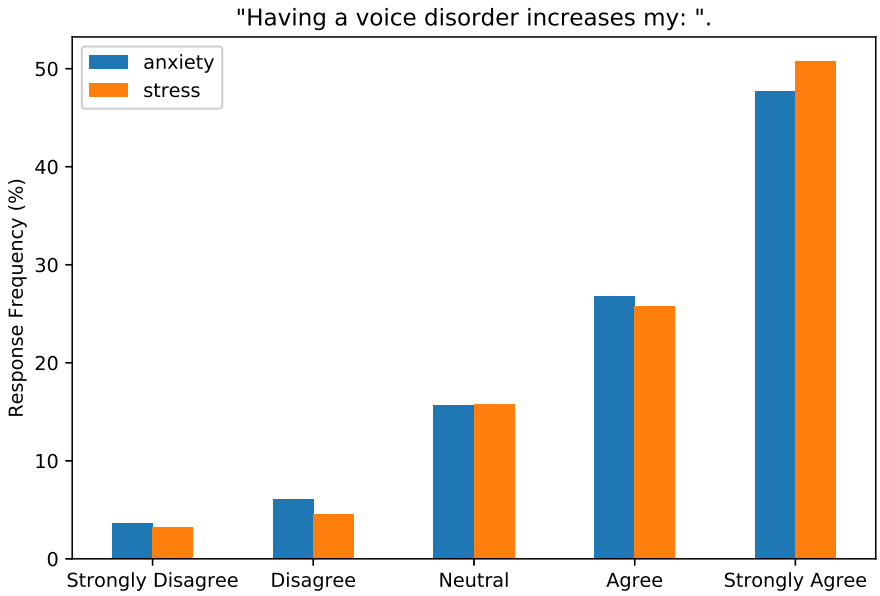


Figure 4.8: Responses to *'Having a Voice Disorder Increases my Level of Stress'* and *'Having a Voice Disorder Increases my Anxiety'*

with a weighted average of 4.15, while after diagnosis, only 22.4% of respondents agreed or strongly agreed with the statement with a weighted average of 2.60.

Another aspect that was brought up in the responses to the initial survey was a feeling of isolation. To quantify this feeling, the respondents were asked to rate the extent to which the agree/disagree with the statement 'Because of my voice disorder, I feel isolated' and 56.7% either strongly agreed or agreed with this statement while 21.7% either strongly disagreed or disagreed (weighted average 3.53).

4.6 Discussion

On a surface level, our investigation into the demographics of individuals with voice disorders—primarily Spasmodic Dysphonia—align very well with previous work in the area as Spasmodic Dysphonia occurs more often in women than men and often onsets in middle age. However, we go into more detail and explore other consequences

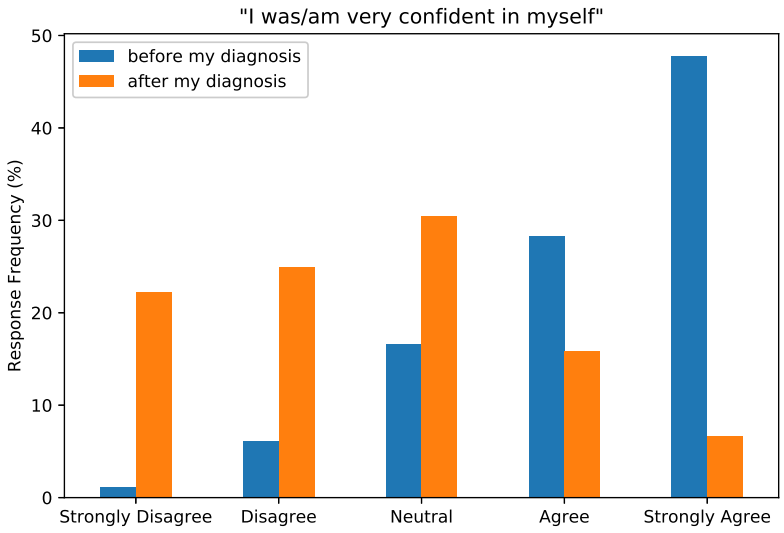


Figure 4.9: Responses to *'Before Having a Voice Disorder, I was Very Confident in Myself.'* and *'After Having a Voice Disorder, I am Very Confident in Myself'*

of having a voice disorder—providing insight into some of the day-to-day experiences of people with voice disorders in the hopes of making it clear what kinds of tasks voice-assistive technologies could assist.

4.6.1 Difficult Situations

One of the main outcomes of the initial voice disorder survey is a clear understanding of some of the most common situations that are particularly difficult for individuals with dysphonia to be understood. Individuals with voice disorders often find it difficult to be understood while speaking on the phone, speaking in a noisy environment, or meeting someone new for the first time. Because people with voice disorders find these situations difficult, it is more difficult for them to engage in social interactions, and complete the tasks necessary to acquire, maintain, or advance in a career. These difficulties often lead to low self-esteem and confidence, as well as feelings of isolation, anxiety, frustration, stress, and sometimes depression.

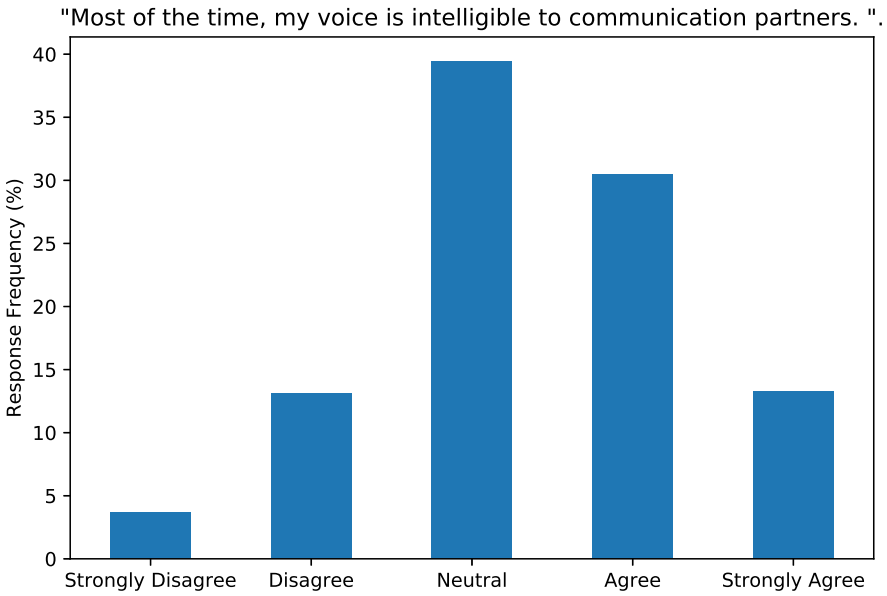


Figure 4.10: Responses to the Likert Question *'Because of my Voice Disorder, I Feel Isolated'*

4.6.2 Emotional Impact of Dysphonia

The emotional impact of having dysphonia is something that stood out as a theme in the initial survey, and that was further quantified in the follow-up survey, more specifically dealing with changes in levels of anxiety, confidence, and stress before and after diagnosis. As shown in Figure 4.7, before having a voice disorder, most respondents would strongly disagree or disagree with the statement 'I would describe myself as having anxiety' before being diagnosed with a voice disorder, however, after being diagnosed, the majority of respondents would either strongly agree or agree. This points towards the idea that individuals with voice disorders experience anxiety secondary to their voice disorder, rather than having anxiety before having a voice disorder, as was previously thought, however, is purely directional.

In Figure 4.8, the responses to the statement 'Having a voice disorder increases my stress' and 'Having a voice disorder increases my anxiety' were very similar. In

general, the surveyed population either strongly agreed or agreed with both statements. The increased levels of stress and anxiety have a significant impact on an individual's mental health and emotional wellbeing.

Figure 4.9 shows the responses to the statement 'I was/am very confident in myself' both before and after diagnosis with a voice disorder. Before diagnosis, most respondents strongly agree or agree with the statement, however, after diagnosis, the responses lean more towards strongly disagree/disagree. This change denotes a general decrease in confidence due to the development of a voice disorder.

Anxiety, stress, and confidence are all key parts of an individual's emotional and mental health. The development of a voice disorder has shown to have a strong negative impact on most individual's levels of stress, anxiety, and self-confidence.

4.6.3 Social Impact of Dysphonia

Decreased intelligibility makes being social very difficult. Many individuals with voice disorders report feeling isolated and frustrated by not being able to maintain the same social life that they were able to before being diagnosed with a voice disorder. When it is difficult to socialize, it is easy to withdraw from social situations and become isolated.

4.6.4 Career Effects of Dysphonia

Individuals with voice disorders often find it difficult to find, keep, and advance in a career. The results show that 24.7% of all respondents (40% of respondents who are retired) retired early due to their voice disorder, and 28.1% of respondents (125 people) report that they either have lost a job because of their voice, or they suspect that they lost a job because of their voice. These results demonstrate that being

able to find, maintain, and advance in a career with a voice disorder is difficult due primarily due to the decreased intelligibility of individuals with dysphonia.

4.6.5 Voice-Assistive Technologies

Despite 89.0% of respondents reporting that they experience some kind of barrier on a day-to-day basis due to their voice disorder, only 20.5% have used assistive technology for their voice, and only 2.9% (13 individuals) reported continued use of a voice-assistive technology. Very few voice-assistive technologies have been developed to help people with voice disorders be better understood, however, individuals with voice disorders are open to using voice-assistive technologies that help them be better understood. This suggests that there are opportunities for innovation in creating voice-assistive technologies that help to facilitate these day-to-day interactions and help mitigate the social, career, and emotional impact of having a voice disorder.

4.6.6 Treatments/Coping Mechanisms

While 48.1% of the respondents regularly receive BTX injections at an average frequency of every 3.8 months, many are still unsatisfied by the treatment options. BTX injections work very well for some individuals but others are frustrated by the period directly following an injection where their voice may be extra breathy or tight (as the dosage of BTX is tweaked). Some individuals find the process painful, and for others, it is too expensive and time consuming to continue.

Respondents shared many different ways that they cope with their voice disorder. These coping strategies range from adjusting social expectations and avoiding places with loud background noises, to having a glass of wine at dinner for the alcohol to relax their voice a little bit. There have been anecdotal discussion around the use of CBD oil to help treat voice conditions, however, there is currently no data to support

this claim, and when asked in the survey, many of those who had tried to use CBD for their voice disorder reported no difference in their voice.

4.7 Conclusions

We conducted a survey that evaluated the potential for assistive technologies in the domain of voice disorders. In two joint investigations, 471 and 453 individuals with voice disorders were surveyed respectively leading to the conclusion that there is a significant need for voice-assistive technologies. The experience of individuals with voice disorders concerning existing voice-based technologies and opportunities for the development of new assistive-technologies is presented. From these findings, new research directions are proposed that focus on creating voice-assistive technologies for individuals with dysphonia.

Here are a few particularly impactful quotes from the surveys:

> *I have amazing, thought-provoking, earth-changing things to say. I would love if someone could help me make it as easy as it is for everyone else to say them.*

> *Your voice is you. Your intelligence, emotions, and abilities all come through in your voice... without it you become invisible.*

> *People don't hear me even when I try. I miss being heard.*

> *I hate all the pressure to use voice-activated systems. They don't work for me, and it's just aggravating to be told by a computer that I need to repeat an answer.*

> *One of the biggest challenges with this disorder is hardly anyone knows what it is. It's not like walking with a cane or having a recognized disorder,*

like blindness where people will respond with kindness and helpfulness. For most people this disorder is off-putting. The worst is on the phone with people who don't know me, such as when I have to call an agency, make an appointment, get information, etc. I've had to put up with some hurtful comments. And voice response calls ("say 'placing an order') on the phone are next to impossible.

Coming at a crucial time in my development as an adult, (age 19) and being undiagnosed for 6 years, untreated for 12 and then poorly understood throughout my life, SD has been about as impactful as a serious spinal cord injury causing quadriplegia. Except quadriplegia is better understood.

Chapter 5

EVALUATING THE ACCESSIBILITY OF VOICE-BASED TECHNOLOGIES

5.1 Accessibility of ASR Systems

Speech that is less intelligible due to a neuromuscular disorder is referred to as dysarthric speech. The speech of individuals with dysarthria is highly variable—speech may be slurred; have nasal, strained, or hoarse vocal quality; and vary in tempo, rhythm, or volume of speech production. This wide breadth of articulatory differences makes recognizing and understanding dysarthric speech a challenging problem. People with voice disorders will often be able to communicate quite clearly with those who are close to them: family, friends, caregivers, however, they will be significantly less intelligible to unfamiliar communication partners Borrie *et al.* (2012). This creates a social barrier which prevents some individuals with voice disorders from fully participating in their community Cooper *et al.* (2009).

With the popularization of products like Amazon Alexa®, Google Home®, and Voice Assistants like Siri®, Cortana®, and Google Now®, speech is being used now, more than ever, as a means of digital interaction. Automatic speech recognition can be used for a variety of assistive contexts, such as computer interactions and phone-based interactions. However, individuals with voice disorders generally cannot obtain satisfactory performance with commercially available ASR systems Young and Mihailidis (2010); Rosen and Yampolsky (2000). To address this problem, many researchers have developed specific, robust, dysarthric speech recognition systems to varying degrees of success. Dysarthric speech recognition is a difficult problem to solve due to two main factors: the immense variability in the speech produced

by individuals with dysarthrias, and the relatively small datasets available to model dysarthric speech and train robust recognition models.

It is colloquially known that current off-the-shelf speech recognition packages do not recognize pathological speech as well as they recognize 'normal' speech. To investigate this hypothesis the accessibility of two off-the-shelf speech recognition systems was evaluated on both control and pathological speech (using the dysarthric speech datasets UASPEECH Kim *et al.* (2008), and TORGO Rudzicz *et al.* (2012)), and later a collection of datasets from speech with a wide range of intelligibility including data collected in noisy conditions, and speech from individual's with accents.

5.1.1 Previous Work

A potential solution to recognizing significantly different voices is to build personalized ASR systems that fit individual voices. This methodology has been attempted for the last 30 years, and there has not been significant progress. Of the dysarthric speech recognizers created, those that use an extremely limited vocabulary (10 digits) are able to achieve around 94% accuracy Hasegawa-Johnson *et al.* (2006); Green *et al.* (2003). Results from systems that use larger vocabularies are extremely varied from 30.84% Polur and Miller (2006) to 97% recognition rate Sharma and Hasegawa-Johnson (2010). The highest reported accuracy on the biggest vocabulary using the least intelligible subjects was 85.05% from Selva Nidhyananthan *et al.* (2016) using recurrent models with Elman backpropagation networks. However, due to the large variability in testing conditions—the intelligibility of subjects, the number of subjects, the complexity of the vocabulary, and the different evaluation metrics—it is very difficult to objectively compare the efficacy of different algorithms.

This is not the first paper to evaluate the efficacy of off-the-shelf ASR systems on non-normative voices. Most recently, Orozco-Arroyave *et al.* (2016) evaluated the

performance of Google's®cloud-based ASR system on speech from individuals with Parkinson's Disease in three different languages. However, speech from individuals with dysarthrias has not been tested since 2010 Young and Mihailidis (2010); Rosen and Yampolsky (2000). In the last eight years, there have been significant improvements in ASR systems largely from the application of different deep neural network models to the domain—namely long short-term memory systems (LSTMs) Hinton *et al.* (2012); Baker *et al.* (2009); Deng *et al.* (2013) as well as distance measures such as the connectionist temporal classification (CTC) Graves *et al.* (2006). We predict that when these off-the-shelf ASR systems are tested with dysarthric speech, the system that uses deep neural networks will outperform the system that uses generative models.

5.1.2 Robust Speech Recognition

Most of the robust speech recognition research has focused on making speech recognition systems robust to background noise such as bustling traffic, or a crying baby. These kinds of noise are what we refer to as uncorrelated noise—meaning that there is no correlation between the speech and the noise. The dogma of the field of robust speech recognition is to take a dataset, add noise to it, and then reconstruct the original utterance from the noisy data. This has led to many good results as can be seen in Wang *et al.* (2015); Pang and Zhu (2015); Donahue *et al.* (2017). However, we suggest that there is a need for a stronger focus on what we refer to as correlated noise—i.e., noise that comes from the voice itself. Much of the noise-robust ASR literature revolves around the central assumption that the noise is uncorrelated with the speech. In many cases, this is not a safe assumption, such as when dealing with accented speech or speech from individuals with voice disorders.

5.1.3 Motivation

In order to obtain a clearer picture of how well state-of-the-art voice-based technologies recognize speech from individuals with voice disorders, a series of experiments were conducted where the performance of two different ASR systems was compared between dysarthric and control speech.

5.2 Methods

5.2.1 Experiments

The performance of the two ASR systems was tested using the two datasets described above—TORGO and UASPEECH. Each dataset was fed to the ASR systems, and the word error rate (WER) was calculated from the resulting prediction, as shown in figure 5.1. Carnegie Mellon University's Sphinx Open Source Recognition (Sphinx), and Google Speech Recognition were used as the ASR systems to test. Sphinx uses a combination of HMMs and GMM models to recognize speech while Google reportedly uses an LSTM based network. Unfortunately, we must treat these two ASR systems as black boxes, and rather than directly compare their architectures, we will use them as benchmarks for how the field has progressed in the last ten years, as it has shifted from generative models to deep neural network models.

We predicted that the Google model would have a lower WER than the Sphinx model for both control and dysarthric speech and that the dysarthric speech would have a higher WER than the control speech.

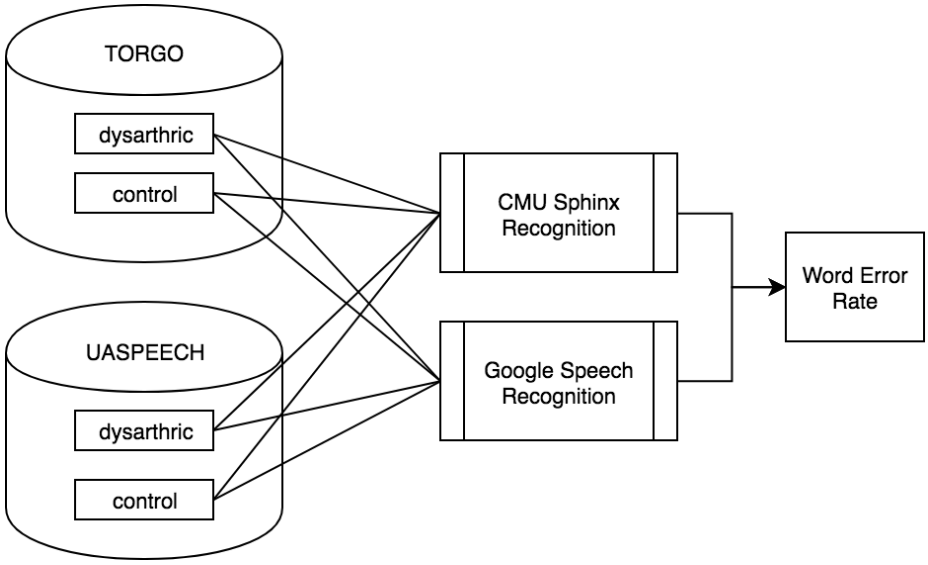


Figure 5.1: Experimental Setup to Test the Accessibility of Two ASR Systems

5.2.2 Datasets

UASPEECH

The Universal Access Speech (UASPEECH) dataset from the University of Illinois Kim *et al.* (2008) was published in 2008 and consists of speech samples from 15 individuals with dysarthrias, and 13 age and gender-matched control voices. The vocabulary used in UASPEECH consists of command words (up, left, down, right, etc.), common words (the, and, I, you, etc.), the phonetic alphabet (alpha, bravo, charlie, etc.), digits 1-10, and 300 uncommon words. There are a total of 765 words for each speaker, three repetitions of each of the commands, letters, digits, and common words, and only one instance of the 300 uncommon words per speaker. The speech from UASPEECH was collected using a 'beep' sound to segment the instances of speech, and because of this, there is a lot of silence in the dataset.

TORGO

The University of Toronto's TORGO database is a database of acoustic and articulatory speech from speakers with dysarthria Rudzicz *et al.* (2012) which was created in 2012. This dataset consists of speech samples from 8 individuals with dysarthria and 7 control voices. For our use case, we did not use the articulatory data, and just focused on the speech. The vocabulary of TORGO consists of non-words (vowel sounds, phoneme repetitions, etc), short words (computer command words, words from the Frenchay Dysarthria Assessment Enderby (1983), words from the word intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthria Walshe *et al.* (????), the 10 most common words in the British National Corpus, and all of the phonetically contrasting pairs of words from Kent *et al.* (1989). The dataset also contains both restricted sentences and unrestricted sentences. Unrestricted sentences are recorded from asking an individual to freely describe an image rather than reading from the screen.

Performance Measures

Word Error Rate (WER) is used to measure the performance of the ASR systems Morris *et al.* (2004). WER takes the sum of substitutions S, insertions I, and deletions D from the hypothesized word divided by the number of words in the ground truth label N. While it may seem counter-intuitive, because of this formulation, it is possible to obtain a WER that is more than 100%.

$$WER = \frac{S + D + I}{N} \tag{5.1}$$

A slightly less common performance metric used in the ASR literature is the Word Accuracy Rate ($WAcc$). This is defined in equation 5.2, where WER is as defined in

Table 5.1: Difference Between ASR Performance on Control and Dysarthric Speech

	Dysarthric	**Control**	**% Diff**
WER	136%	74%	59%

equation 5.1 and $R = N - (S + D)$, which refer to the number of correctly recognized words.

$$WAcc = 1 - WER = \frac{N - S - D - I}{N} = \frac{R - I}{N} \tag{5.2}$$

In creating the UASPEECH dataset, the authors tested how well the dataset could be understood by humans. To do this they calculated the recognition rate of each dysarthric speaker to correspond to the percent intelligibility. They calculated the recognition rate as the number of correctly recognized words R, divided by the total number of words.

$$RR = \frac{R}{N} \tag{5.3}$$

To compare the performance of both ASR systems to the human intelligibility baseline recognition rate, we calculated the recognition rate of both ASR systems. This recognition rate is used to assess how well these ASR systems model human intelligibility.

5.3 Results

5.3.1 ASR Performance

When the performance of the two chosen ASR systems was evaluated, as expected, Google ubiquitously achieved a lower WER than Sphinx. The WER of the control speech was lower than the dysarthric speech on all test cases as shown in Table 5.1. Table 5.2 shows that Sphinx had an 84% larger WER than Google when the dysarthric speech was evaluated, and 74% larger when control speech was tested. Sphinx had

Table 5.2: ASR System Performance on Control and Dysarthric Speech

	Sphinx	**Google**
Dysarthric	126%	43%
Control	63%	20%

a 55% larger WER in dysarthric data than control, and there was a 44% difference between the WER of the control and dysarthric speech when using Google.

5.3.2 ASRs as a Model of Human Intelligibility

Figure 5.2 demonstrates the correlation between human recognition rate and what the ASR systems were able to correctly recognize. Each speaker from the UASPEECH database was tested using human listeners to establish a level of intelligibility. These percent recognition rates for each speaker are compared to the human recognition rate reported in Kim *et al.* (2008). The numbers on the x-axis correspond to a speaker, and the y-axis is the recognition rate. Humans consistently perform better than both Google and Sphinx in recognizing dysarthric speech, and Google outperforms Sphinx. When a simple linear regression is performed, the correlation coefficient values for the trend lines show similar patterns: 0.958 for human, 0.920 for Google and 0.765 for Sphinx.

5.4 Discussion

In general, the results were as expected: models that employ deep neural networks (as Google does) perform better on both control and dysarthric speech compared to models that use generative strategies (like HMMs and GMMs). Dysarthric speech is recognized less often than control speech. Our analysis demonstrates that ASR systems do not provide robust speech recognition to individuals with voices that fall

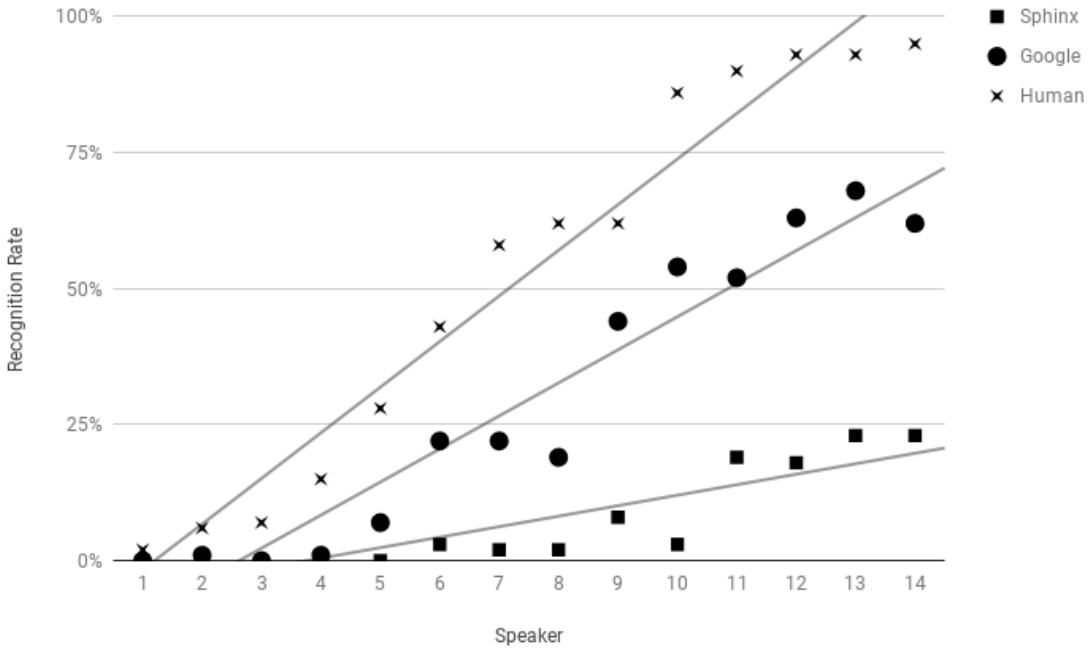


Figure 5.2: Three Different Models of Intelligibility of Dysarthric Speech: Sphinx, Google and Human

outside the range of 'normal' voices. Humans perform the best at recognizing speech from individuals with dysarthria, while the more advanced models of speech and langauge used in Google's speech recognition system perform better than the HMMs used in CMU's Sphinx.

Part of the reason that these error values are so large is that the average length of the utterances $N_\sigma = 1.56$ is very small. Often times, individuals with dysarthrias will speak slowly or add breaths between syllables. The models tested do not seem to be robust to this kind of noise. Often times, in dysarthric speech the speech is staccato and slow. These systems often interpret these pauses or changes in tempo as the beginning of new words, and thus the WER of the word is often greater than one. With N_σ being so small, any language model that the ASR systems have built are able to be used. This also could lead to an increase in WER.

ASR systems have room for improvement to be robust to both correlated and

uncorrelated noises. Should these improvements be made, ASRs would be a more inclusive an accessible tool for individuals with voice disorders. Through creating such a system, not only will individuals with speech disorders be able to be better understood by ASR systems but in general ASR systems will be more robust to complex noise. This is a great example of universal design—the explicit needs of individuals with disabilities become the implicit needs of the general population. Creating inclusive ASR systems for individuals with dysarthria will only help to make ASR systems more robust and widely applicable in real-world settings Panchanathan *et al.* (2016). The following areas will be essential in building these robust systems.

5.4.1 Datasets

The datasets used to train ASR systems need to be more inclusive of different voices than the current datasets. As shown in Table 5.3, there are three main dysarthric speech datasets that are used. The total number of hours of dysarthric data is around 58 hours of speech with very high variation. However, one dataset of normative speech, Switchboard, has 260+ hours of speech data. Comparatively, the three dysarthric speech datasets seem insignificant when compared to the size of normal speech corpora. The lack of sufficient training data for disordered speech is a bottleneck for the field. With the collection and publication of more data, we expect to create systems that are more robust to complex types of noise, both correlated and uncorrelated. One potential way to get more data is to create it. In the last three years, Generative Adversarial Networks Goodfellow *et al.* (2014) have shown that they have the power to generate lots of data from a distribution. In order to augment the existing dysarthric data that we have, we may need to collect more dysarthric data to get a better idea of the distributions.

Table 5.3: Overview and Comparison of Available Datasets

Dataset	**Sub**	**Data Type**	**Utterances**	**Hours**
Menendez-Pidal *et al.* (1996)	11	Audio	Sentences	17.5
Kim *et al.* (2008)	19	Audio, Visual	Isolated words	18
Rudzicz *et al.* (2012)	7	Audio, Visual, Articulatory	Non-words, Isolated words, Sentences	23
Godfrey *et al.* (1992)	**543**	**Audio**	**Conversations**	**260+**

5.4.2 Benchmarking Tests

In order to create systems that are fully robust, a standard benchmark test will need to be created. Ideally, a standard test of how robust a model is to different voices should be used to measure the performance of new ASR systems. One of the biggest problems with the field of dysarthric speech recognition is that there is not a consistent, objective way to compare the performance of different algorithms.

5.4.3 Domain Adaptation

There seems to be great potential for domain adaptation techniques to make ASR systems more robust to correlated noise. The goal of domain adaptation is to optimize a model that is trained on a source distribution D_s to also perform well for a target distribution D_t. In the case of making ASR systems more robust to different voices, D_s would be the normal speech corpora that ASR systems are trained on, and D_t would be the datasets that have data from individuals with speech disorders. Domain adaptation and transfer learning show a lot of promise in making ASR systems more inclusive of different voices.

5.4.4 Robust Models

With the collection and creation of more data and the application of domain adaptation techniques between normative speech and disordered speech, significantly more robust models will be created. These systems could also benefit from the application of a person-centered model. By fine-tuning the machine learning architectures to better understand the speaker's voice, the model can be made more robust. The application of other cutting-edge machine learning techniques, coupled with more data and benchmarking tests should lead to a system that is inclusive of all voices.

5.5 Say What? Intelligibility Metadataset

In an expansion of the work on evaluating the accessibility of voice-based technologies, I analyzed the performance of the same two ASR systems on three other datasets: Mozilla's Common Voice [1] , TIMIT Garofolo *et al.* (1992), and the Speech Accent Archive Weinberger (2013). The goal of adding the evaluation of more datasets is to obtain a better understanding of how intelligibility is encoded in speech. By only working with pathological speech, the data that resulted was biased towards poorly intelligible speech. By adding large datasets of speech from control speakers, especially speech from datasets like Common Voice, where the data is collected in a distributed, crowdsourced manner, causing some level of noise in the data, the expectation is to learn a more robust model of intelligibility. The process of collecting 'Say what?', the intelligibility dataset is shown in figure 5.3. The goal of this work was to have a dataset of speech intelligibility information that is representative of a wide range of types of speech.

[1] https://voice.mozilla.org

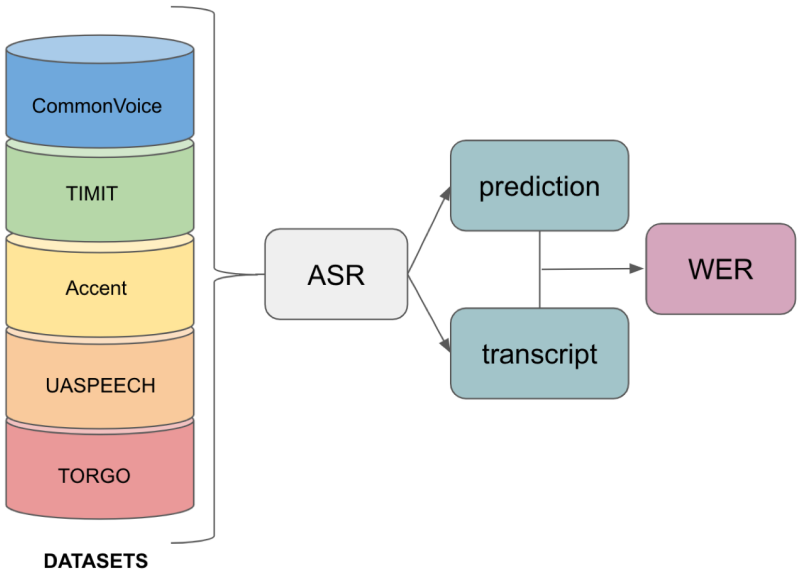


Figure 5.3: The Process of Collecting the 'Say What?' Intelligibility Metadataset.

5.5.1 Metadatset Collection

Two different ASR systems were used to obtain this metadata: Google Speech Recognition and CMUSphinx Open Source Speech Recognition. Google Speech Recognition uses deep neural networks while CMUSphinx uses hidden Markov models to achieve its speech recognition. Figure **??** shows the relatively straightforward process of obtaining the data. The speech files were fed into the two ASR systems to obtain the prediction of what was said. Then, using the transcript, the WER was calculated and recorded, along with the number of substitutions, insertions, deletions, and the time taken to obtain the results. The time is included to provide a comparison for alternative techniques such as estimation models that may operate faster.

5.5.2 Metadataset Analysis

While many utterances are understood by the ASR systems, there is enough variability in the metadataset to find some interesting patterns. In Table 5.4, the type

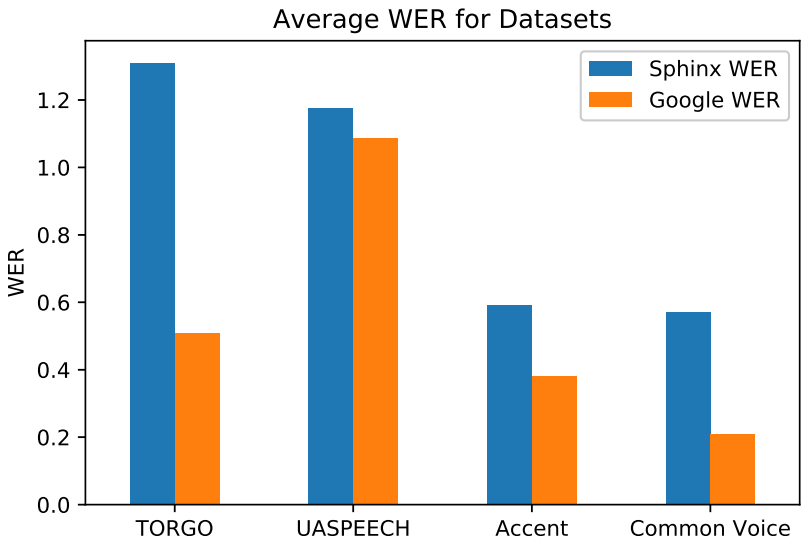


Figure 5.4: The Average WER Per Dataset Included in the Metadataset.

of data, and number of speakers per dataset are shown. As one would expect, the pathological datasets TORGO and UASpeech have the fewest speakers (27) and the highest average WER (1.16). In Figure 5.4 a bar chart shows the average WER per dataset. The dataset with the smallest average WER is TIMIT (0.09), while the Accent database (0.33) and Common Voice (0.20) sit somewhere in the middle. This distribution of WER is what we would expect. Obtaining the WER from these datasets was a slow process, it took on average 1.87 seconds per utterance, and 0.33 seconds per word for a total of 513533.93 seconds of continuous computation to obtain the WER data that is included in this dataset–that's almost 6 full days of continuous computation.

5.6 Conclusions

This work introduced a new metadataset of WER labels for several popular speech datasets spanning a wide intelligibility range. With the goal of stimulating research into modeling intelligibility, the resulting transcripts from two different ASR systems

Table 5.4: Metadetaset Metadata

Dataset	**Type**	**Speakers**
TORGO/UASPEECH	path	27
Accent	accented	2140
CommonVoice	average	33,541
TIMIT	average	630
Total		36338

and an analysis of the number of substitutions, insertions, deletions, and total words in the predicted transcript were collected.

Chapter 6

EXPERIMENTS WITH DYSPHONIC SPEECH

6.1 Intelligibility Estimation using SayWhat? Metadataset

In an effort to estimate the intelligibility of an utterance, we decided to use a state-of-the-art model for estimating speech quality– Quality-Net Fu *et al.* (2018). We created Intellinet, our implementation of Quality-Net to test its ability to predict intelligibility data in the form of estimating the WER. Quality-Net consists of a bidirectional LSTM layer with 100 nodes followed by two fully connected linear layers with 50 nodes each, one linear layer, and then a global averaging layer. The output of the last layer is the global average of the frame-level predictions which is the utterance-level WER prediction. As input into the network, we used spectrograms extracted from TIMIT to replicate Quality-Net. We selected a random subset of TIMIT (4200 utterances) to use as the training set and split the rest into a validation set (1049 utterances) and test set (1049 utterances).

6.1.1 Results

Much like the results seen in Quality-Net, our model quickly converges after around 500 iterations, however, unlike Quality-Net, our model converges to the global average of the labels rather than learning the mapping between the utterances and

Table 6.1: Performance of Intellinet in Comparison to Quality-Net

	MSE	**LCC**	**SRCC**
Quality-Net	0.1225	0.9054	0.9065
Intellinet	0.023	0.023	0.007

the WER. This difference in performance is demonstrated in figure 6.1. In Figure 6.1A, the results from Quality-Net show the predicted PESQ scores compared to the true PESQ scores, and Figure 6.1B shows the predicted WER compared to the true WER. The PESQ predictions line up relatively along the line y=x, while the predicted WER is only predicting the global average of the data, and the Linear Correlation Coefficient (LCC) for Quality-Net's predictions is 0.9054, indicating a very strong correlation between the true PESQ value and the predicted PESQ. The Spearman's Rank Correlation Coefficient (SRCC) is 0.9065 for Quality-Net reinforcing the idea that there is a strong correlation between the true and predicted values. However, when we plot the results from the intelligibility data, we see roughly a horizontal line right around where the global average of the WER is, an LCC of 0.023 and an SRCC of 0.007. Both the LCC and SRCC for the intelligibility data are very close to zero, which indicates that there is no correlation between the true WER and the predicted WER. The mean squared error of the intelligibility network is lower than Quality-Net's but this doesn't say much.

6.1.2 Discussion

Using the same network as Quality-Net and the same features from TIMIT, we were unable to replicate the success of Quality-Net with the task of predicting the WER instead of PESQ. While intelligibility and quality may seem like superficially similar concepts, a state of the art model for estimating quality is ill-suited for assessing intelligibility. There are a variety of reasons this might be the case.

Linguistic context is very important for understanding intelligibility. Realization difficulties become exhibited in varying phonemes for various reasons when considering disordered speech and accented speech. Word errors in ASR systems often involve the mischaracterization of one word to another as a result of the complicated

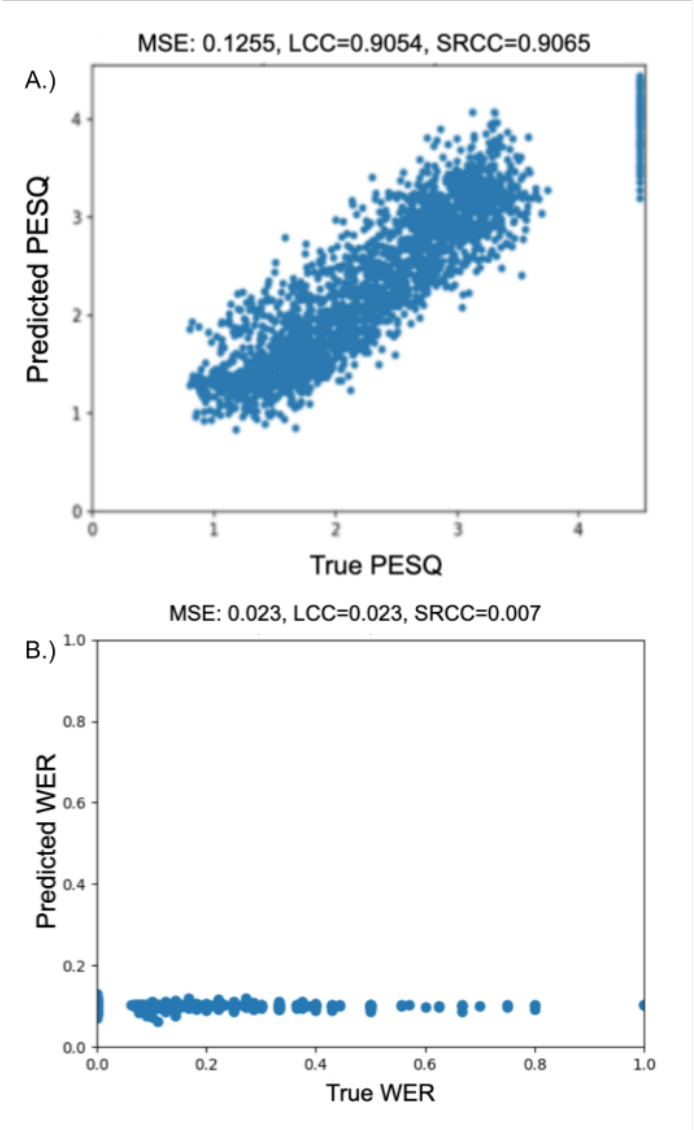


Figure 6.1: (A) Quality-Net's Performance. (B) Intellinet's Performance

interaction of internal language models and acoustic models. Out of domain acoustic patterns produced by disordered or accented speakers lead to incorrect classification calls made by the ASR model. This means that Quality-net's focus on frame-level details of the speech acoustics will miss broad patterns on a word-level scale – the sort of short-term spectral artifacts in low-quality compressed speech are in no way similar to the broader, more complicated patterns of difficult-to-understand natural speech.

The intelligibility model that was built trains and converges but learns no useful WER prediction capabilities. Integrating linguistic data acquired solely from a speech signal and learning broad patterns or acoustics across words and the propensity for an ASR system to make errors from only WER counts is a very tall order for a simple neural network. Much more work in this area is necessary.

6.1.3 Conclusions

After attempts to predict intelligibility using models that were successful in predicting speech quality failed, the conclusion was made that modeling intelligibility is a nontrivial task that requires novel approaches unrelated to quality assessment methods. From this data, we can conclude that intelligibility and quality are encoded significantly differently in speech. The complexity of intelligibility and the diversity of reasons intelligibility difficulties can arise mean that significantly more complicated models integrating local acoustic, global acoustic, and linguistic data are probably necessary to model it adequately. Once such a model is created many clinical and educational applications will be available. We hope that a direct speech intelligibility estimation system requiring no transcripts and no 'gold' examples will drive new applications in clinical, educational, and research settings.

6.2 Intelligibility Detection

As a simplification of the task discussed above in 6.1, the goal of building an intelligibility detection model is to predict whether or not an utterance would be recognized by an ASR system. The goal of building this experiment was to build an error detection model that quickly predicts whether or not speech is intelligible to an ASR system–the first step towards building a system that improves the intelligibility of speech.

Table 6.2: Performance of ASR Systems on Dysphonic Speech

Voice Type	S	I	D	WER
Control	0.56	0.05	0.15	0.09
Dysphonic	1.84	0.15	9.96	0.28

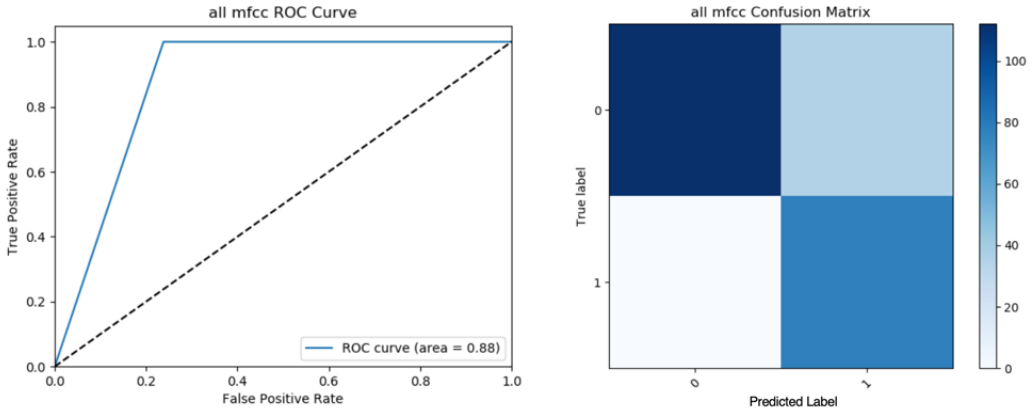


Figure 6.2: ROC Chart and Confusion Matrix for Error Detection Model.

To test the performance of state-of-the-art ASR systems on speech from individuals with spasmodic dysphonia, we collected a small dataset of speech samples from individuals with SD. For this experiment, the same dataset described in 6.3.1 was used to represent speech from individuals with SD. This dataset was fed through an ASR system and the resulting ASR performance is shown in Table 6.2. This measure of intelligibility was used to determine whether or not an ASR system would make an error on a given utterance or not.

The same model that was used to estimate the intelligibility was trained on the error detection data. The model achieved 85% accuracy after 100 epochs. The results of this set of experiments are shown in the confusion plot and ROC chart in Figure 6.2. As shown in the confusion plot, the model was more likely to predict false positives than false negatives.

6.2.1 Discussion

Upon closer evaluation of the output of the intelligibility detection model, while the model seems to perform well, with an accuracy of 85%, it seems like the model just learned to apply a threshold based on utterance length. The longer the utterance, the more likely it is that an error is made in the utterance. The model seems to find a cutoff point x for an utterance length where it predicts every utterance shorter than x as intelligible, and every utterance greater than x as unintelligible.

6.3 Voice Disorder Classification

6.3.1 Dysphonic Speech Dataset

I collected a small sample of data to work with from individuals with Spasmodic Dysphonia, to be able to better understand how SD speech differs from control speech and what makes it less intelligible. I collected this small dataset of 10 speakers in-person at the 2018 National Spasmodic Dysphonia Association's Annual Symposium. I set up a small recording station and had individuals who volunteered to record speech samples walk through a few prompts. Each individual was asked to read several sentences from TIMIT, as well as few paragraphs that are commonly used in the speech pathology field, they were also asked to describe images to obtain some spontaneous speech.

Currently, the voice disorder data that we have to work with consists of 10 speakers saying 24 utterances each. This is a small dataset that makes up less than 1.5 hours of speech. To guarantee a balanced dataset, I randomly selected 10 speakers from the Voice Conversion Tool Kit Veaux *et al.* (2017) and pulled 24 random utterances from each of the 10 selected speakers. This data serves as the control speech data. For the noisy control speech condition, I added random Gaussian noise to a different

10 random speakers and different 24 random utterances per speaker. The dataset for this question is comprised of 720 utterances, 240 from individuals with SD, 240 from control speech, and 240 from control speech with added noise. These are the three classes that the voice disorder detection model was trained with.

6.3.2 Voice Disorder Classification Experiment

For preprocessing, 80-dimensional log-melspectrograms were calculated for each utterance with 12.5 ms of overlap, and 50 ms frame length. These spectrograms were computed with a 1024 point STFT. The output of the model is encoded as a one-hot vector where the classes are '0: dysphonic', '1:control speech + noise' and '2:control speech'.

The model that I used to train this classification system is based on the model that we had used for Moore *et al.* (2019). It consists of a bidirectional LSTM layer with 100 nodes followed by two fully connected linear layers with 50 nodes each and a 3-dimensional output as there are three classes to predict. The predictions were made by taking the index of the max of the output. The model was trained using cross-entropy loss and stochastic gradient descent with a learning rate of 0.001 and a momentum of 0.9. Dropout was applied to the bidirectional LSTM layer with $p = 0.2$. ReLU activations were used for the main network followed by a Sigmoid activation for the final prediction.

6.3.3 Results and Discussion

As is shown in Figure 6.3, the model performed relatively well for this constrained task. Our model achieves 87.5% accuracy on this small test dataset. Most of the errors were made by mistaking control speech for dysphonic speech. This makes sense as there were a couple of speakers in the SD dataset who had just had a botox injec-

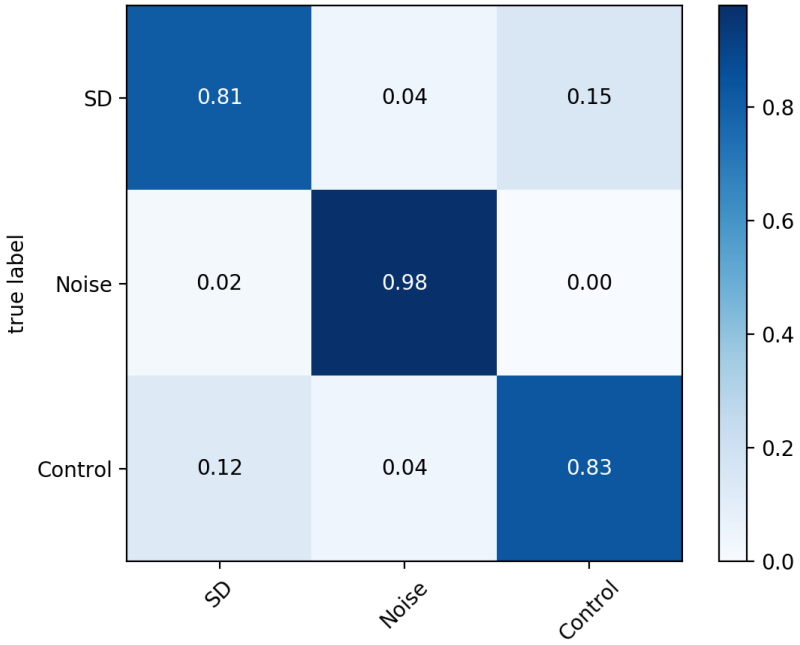


Figure 6.3: The Confusion Matrix For Dysphonic Speech Classification

tion and it was more difficult to perceive their dysphonia. I believe that I might have made this classification problem too much of a 'toy' example by not taking a sophisticated approach in adding noise to the noisy speech. The classification model achieved 97.9% accuracy in classifying the control+noise condition, while it was 83% and 81% accurate on classifying the control speech and SD speech respectively. Expanding this classification problem to dysphonic voice in comparison to other speech-based disorders could be an interesting follow-up study. For these experiments to be successful, it's important that datasets that represent individuals with voice disorders be not only collected but also made freely and publicly available.

Chapter 7

UNCOMMONVOICE DATASET

7.1 UncommonVoice Overview

UncommonVoice was created to build a dataset that helped better represent individuals with voice disorders in current voice-based technologies. Inspired by the work done at Mozilla in Common Voice [1] , a large, freely-available, crowdsourced dataset with speakers from all over the world. Common Voice was created as a high-quality, publicly-open dataset of voice data, with the goal of teaching machines how real people speak. Mozilla also built an open-sourced speech recognition system called Deep Speech, which is trained using Common Voice. While Common Voice has made great strides towards making large volumes of speech data readily available for hobbyists or researchers to jump in and start playing with the data, Common Voice still is made of up mostly healthy speakers. After evaluating the state-of-the-art automatic speech recognition systems in Section 5.1, we concluded that these voice-based technologies are not inclusive of different voices.

The creation of UncommonVoice seeks to freely and publicly provide speech samples of individuals with voice disorders to fuel the research of improving the accessibility of voice-based technologies as well as providing a platform for voice-assistive technologies to be built off of. There currently does not exist a publicly available dataset of voice disorder speech, and as such, UncommonVoice is a significant contribution to the field.

[1] https://voice.mozilla.org/

7.2 'In the Wild' Dataset Considerations

Crowdsourcing is defined as the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community. Crowdsourced datasets are notoriously noisy due to logistical and technical issues during data collection, as well as not being able to control variables that would be controlled if the dataset were collected on-site such as recording instrument type, background noise, distance from the microphone, etc. In the crowdsourced speech acquisition dataset that we propose, Uncommon Voice, we are relying on the participants to classify their voice and follow the speech recording instructions. We have proactively built-in functionality to help reduce the number of logistical and technical issues and have attempted to make the web interface as usable as possible.

UncommonVoice is unique in that the goal isn't necessarily to collect 'perfect' data. One of the advantages of collecting a crowdsourced dataset is that you obtain realistic data from the environment where the resulting machine learning models will be used. By training in the test environment, machine learning models can better anticipate and deal with noisy signals. Uncommon Voice is also expected to have more noise than other similarly sourced datasets as the voices that we are collecting have voice disorders and the signal is inherently less clear. The first step in reducing the amount of post-processing and cleaning necessary for creating an 'in the wild' dataset is building in functionalities to help reduce the number of unsuitable speech samples.

In McGraw *et al.* (2010), only 65% of the data that was collected in a crowdsourced manner was usable in comparison to 90% usable data when it was collected in a controlled manner. The authors directly attribute this decreased yield of high-quality recordings to the lack of training the crowdsourced speakers received. They

theorize that the more training an individual receives, the better the quality of the data that is collected. To this end, UncommonVoice has a video walkthrough of the data collection process as well as a document that provides a walkthrough of the data collection system. UncommonVoice also has a troubleshooting document to help users troubleshoot common issues like browser permissions, presence/absence of a camera/microphone, which browser they are using, etc. A support email address and phone number will also be provided so that should participants need support. The data collection tool also has built-in functionalities such as a 'Review' option—-a button that is available for users to press and hear the recorded speech to make sure that it is valid, and sounds how they want it to.

Before developing ways to automate the process of removing erroneous speech data, it is important to evaluate the potential types of errors in a crowdsourced speech acquisition dataset. When recording prompted speech there are three main types of issues that result in unsuitable data: garbage audio, low-quality recordings, and speaker errors Parent and Eskenazi (2011). Garbage audio consists of recordings that are empty, clipped, have insufficient power or are incorrectly segmented. Low-quality recordings have a low signal-to-noise ratio due to poor equipment or large background noise. Speaker errors are when users misspeak the prompts either maliciously or accidentally. In UncommonVoice, as participants are asked to record three different kinds of prompts (read speech, image descriptions, and non-words) one potential speaker error is reading the non-word prompt ('Please hold 'e' as in leap) rather than following the instructions. We also expect clipping to happen as well as the presence of background noise. These are errors that through the design and development of UncommonVoice, we hope to have done our best to mitigate, but will also do our best to recognize these errors when they are made.

7.3 Design and Development of UncommonVoice

7.3.1 UncommonVoice Data Collection System Features

The UncommonVoice data collection website was implemented with the goal of it to be as convenient as possible for users to provide speech samples. This included building out a feature that allows users to stop at any point in the collection process, should they need a break, etc, the data collection tool saves their spot. The next time the user logs in, the system will ask if they've received a Botox Injection (if they receive BTX therapy) since they last recorded speech samples, and if so get a date, but then it will launch them right back where they left off. This feature was implemented after the realization that it may not be convenient for everyone to collect the speech in one sitting. Throughout the entire dataset collection process, it was made clear to participants that participation was voluntary, and that they could skip any tasks at any time, except for the screener question asking if they were 18 years or older.

To help users get a better idea of the goal of UncommonVoice and how to operate the data collection system, a video was made detailing both the overall goal and the specific instructions of how to use the interface. These videos served as useful tools for the participants. Should the participants run into any issues, an FAQ page was available as well, and demonstrated how to troubleshoot common issues like denying the browser permission to record audio or video.

7.3.2 UncommonVoice Design Limitations

One significant limitation of the data collection tool is that there is a subset of browser/device combinations that the UncommonVoice data collection system is not compatible with. While the data collection system works for Desktops and Laptops

Table 7.1: Device/Operating System/Browser Configurations for UncommonVoice

Device	**Operating System**	**Browsers**
Desktop/Laptop	Windows	Firefox, Chrome, Opera
Desktop/Laptop	Linux/Ubuntu	Firefox, Chrome, Opera
Desktop/Laptop	Mac	Firefox, Chrome, Opera
Mobile	Android	Firefox, Chrome

that use Firefox, Chrome, or Opera, the data collection tool does not support Safari or any Apple mobile browsers. In Table 7.1, the configurations that were compatible with UncommonVoice are shown.

7.4 Dataset Collection Process

The process of contributing data to UncommonVoice includes 5 main steps: the pre-collection survey, and then four main speaking tasks. These tasks are outlined in more detail in the following sections.

7.4.1 Pre-Collection Survey

Before the voice sample recordings, users were asked to provide some demographic information about themselves, as well as provide more information about their voices. The exact questions asked to participants are shown in 7.4.1. There was some conditional logic involved in this survey such that individuals who answered that they were younger than 18 years old were redirected to a disqualification page, and individuals who indicated that they did not have a voice disorder were routed directly to begin the data collection process as the rest of the questions were not relevant to them. The last two questions asked the respondent to rate how clear their voice is on a scale from 'Not clear at all' to 'Very clear', and to rate how easy it is for them to speak on

Table 7.2: UncommonVoice Pre-Collection Survey

Question	**Answer Type**
Are you 18 years or older?	Yes/No
Are you a native English speaker?	Yes/No
Do you have a voice disorder?	Yes/No
What voice disorder do you have?	Multiple Select
Do you regularly receive Botox injections?	Yes/No
When was your last injection?	Date
How often do you normally receive injections?	Number
How would you describe your voice today?	Multiple Choice
How would you rate your voice quality in terms of clarity?	Rating Scale
How easy is it for you to speak?	Rating Scale

a scale from 'Very difficult' to 'Effortless'.

7.4.2 Multimodality

Before beginning the recording process, users were given the choice to either provide audio-only or both audio and visual (video) data. It was made clear that should they choose to provide video data, the visual portion would not be shared publicly, however, the audio would. Upon making a selection between providing audio-only or audio-visual speech samples, the interface for which is shown in Figure 7.1, the speakers were then asked to begin the recording process.

7.4.3 Tasks

The UncommonVoice data collection process consists of 4 different tasks. The design decision to keep the order of the tasks the same between users, but to randomize

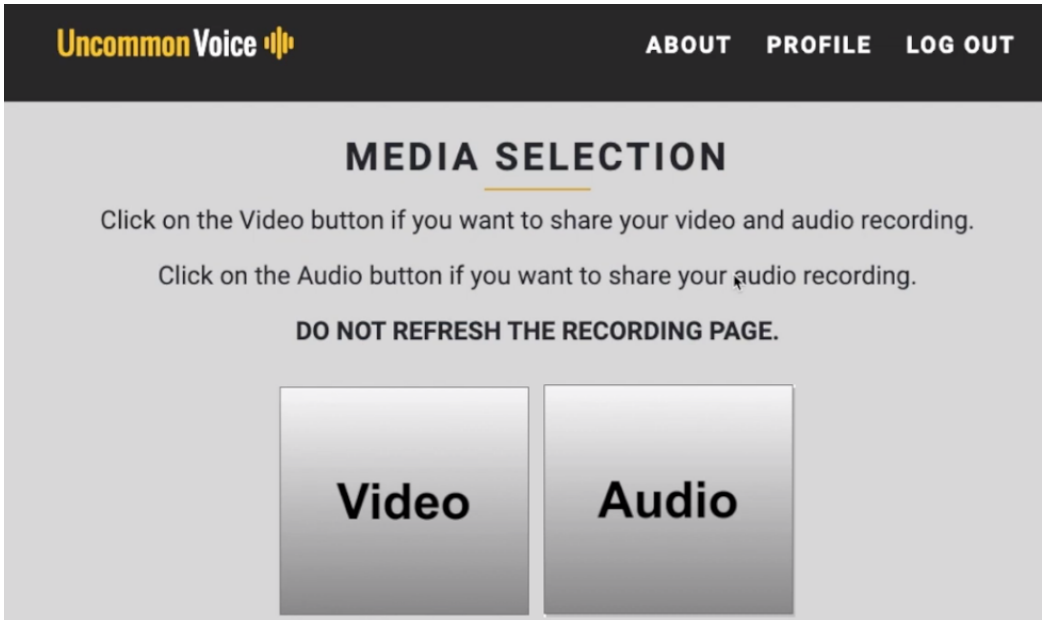


Figure 7.1: Screenshot of the Media Selection Section of UncommonVoice.

the presentation of stimuli within each task was made to obtain the highest value data first as there was an expectation for some of the participants to drop-off mid data collection. To control for–or at least be able to measure– any ordering effects due to this decision, Tasks 1 and 4 contain the same tasks so that the data exists to measure any change in vocal quality throughout the data collection process.

Task 1: Non-words Round 1

The first task that users were asked to complete is holding vowels for 5 seconds. The respondents were asked to hold the corner vowels, so /*a*/, /*u*/, /*ae*/, and /*i*/. To make sure the task was clear, a target word was provided so that the speaker knew what sound they should be holding–for example for /*ae*/, we asked them to hold /*ae*/ as in 'nap'. The goal behind this task was to be able to calculate vocal quality measures. The participants were also asked to repeat 'puh-tuh-kuh' as many times as possible in 5 seconds to obtain the speaker's diadochokinetic rate as described in Portnoy and Aronson (1982).

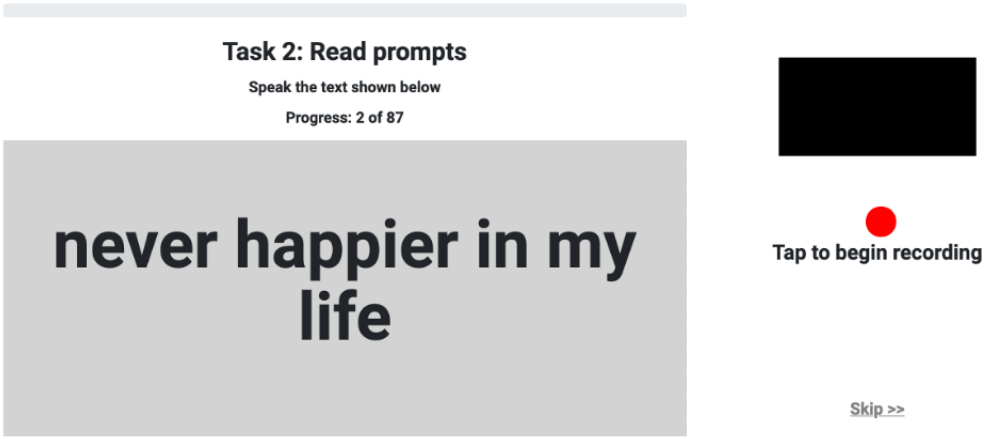


Figure 7.2: Screenshot of UncommonVoice Task 2, Read Prompts from TIMIT

Figure 7.3: Screenshot of UncommonVoice Task 3, Image Description

Task 2: Read Sentences

In the second task, we asked users to read sentences that were randomly selected from TIMIT Garofolo *et al.* (1992). We asked the user to read 84 different TIMIT sentences. These sentences were randomly presented to avoid any ordering effect. To be able to calculate a speaker's CAPE-V as in Kempster *et al.* (2009), speakers were also asked to read the sentences involved in the calculation of the CAPE-V score.

Task 3: Image Descriptions

In the third task, we asked users to describe three different images in their natural way of speaking. We chose to include an image description task to have some speech that wasn't directly read.

Task 4: Non-words Round 2

In the final task, we asked users to repeat the non-words tasks that they completed in Task 1 again. The purpose of this is to measure any change in vocal quality over the duration of the tasks.

7.5 UncommonVoice Results

While UncommonVoice is still open for collecting speech samples from individuals with or without voice disorders, the majority of the data collection has taken place.

7.5.1 UncommonVoice Demographics

Currently, UncommonVoice consists of 4,184 speech recordings from 52 individuals– approximately 7.5 hours of data. Of those individuals, 39 (75%) of the individuals who recorded speech are female, while the other 13 (25%) are male. Of the individuals who contributed speech samples, 35 (67%) of them have a voice disorder, while the other 9 (17.3%) do not. Of the individuals who have a voice disorder, 15 (35%) of the individuals who provided speech samples regularly receive BTX injections as a treatment for their voice disorder, while the other 28 individuals with voice disorders (65%) do not regularly receive BTX injections as a treatment. The respondents were also asked to disclose whether or not they were native English speakers. In response to this question, 44 (84.6%) indicated that they are native English speakers while the other 8 (15.4%) were not.

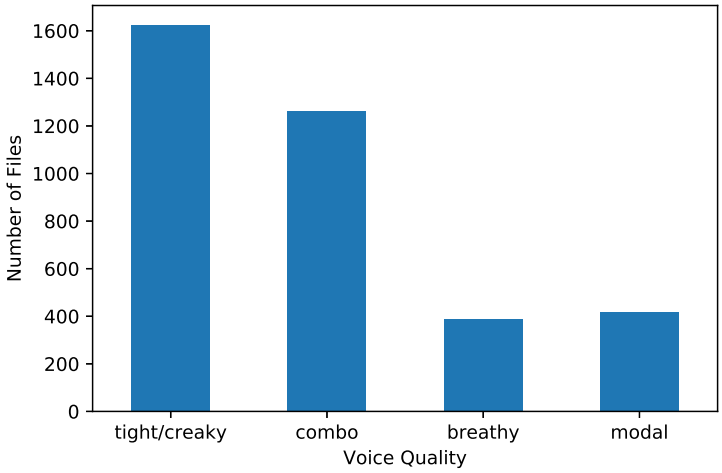


Figure 7.4: Distribution of UncommonVoice Speech Data with Regard to Vocal Quality.

In the pre-voice-recording survey, participants who acknowledged having a voice disorder were asked to rate 'How would you rate your voice quality in terms of clarity', on a scale from 'Not at all clear' (1) to 'Very clear' (4), and the average rating was a 2.34 ± 1.12. Participants were also asked to rate 'How easy is it for you to speak' on a scale from 'Very difficult' (1) to 'Very easy' (4). The average rating for the speaking effort was 2.44 ± 1.16.

Respondents with voice disorders were asked to classify their voice into one of the following categories: tight/creaky, breathy, modal (normal), or combination (breathy and tight). Figure 7.4 shows the distribution of the files in UncommonVoice with regard to the participant's self-reported vocal quality.

7.5.2 UncommonVoice Acoustic Analysis

In Figure 7.5, the difference between control and dysphonic speech when producing */ae/* is shown. The waves that are evident in the bottom melspectrogram are indicative of the 'choppier' glottal pulse, and lack of control that characterizes

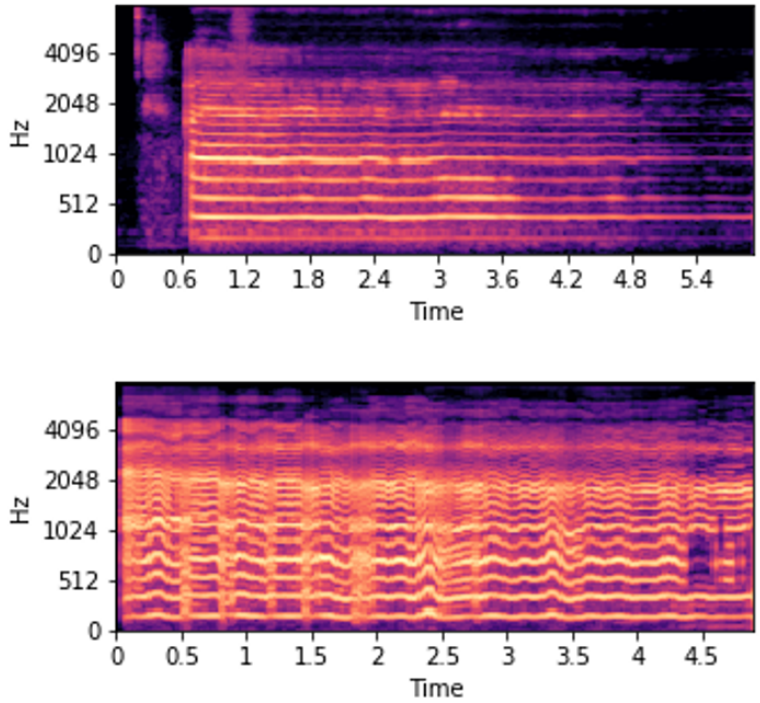


Figure 7.5: Melspectrograms of the Vowel /*ae*/ for Control (top), and Dysphonic (bottom).

dysphonia.

To learn more about the acoustics behind dysphonic speech, several different acoustic features were calculated for the speech recordings in UncommonVoice. Based on previous literature, the features most often correlated with dysphonia were the Cepstral Peak Prominence, jitter, shimmer, harmonic noise ratio, and the variability of the fundamental frequency. For all utterances, these acoustic features were collected, and the results were analyzed with regard to whether or not an individual had a voice disorder as well as the intelligibility of the speaker.

Vowel Space Area

Another common acoustic measure focused more on the articulatory precision of an individual with the Vowel Space Area (VSA) Sandoval *et al.* (2013); Jacewicz

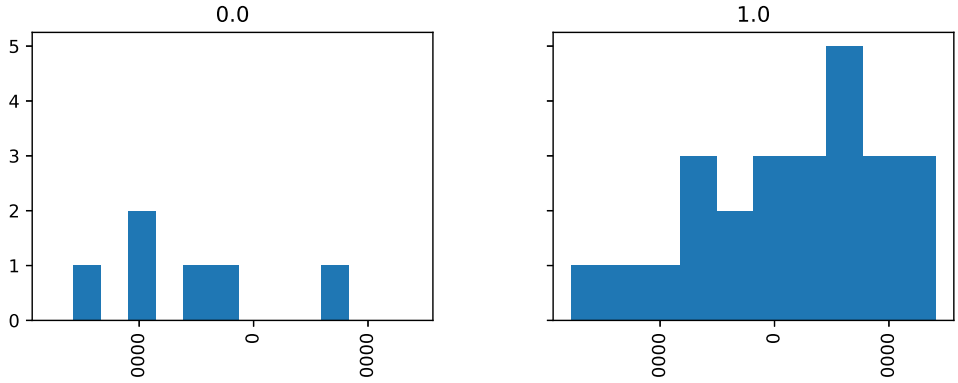


Figure 7.6: Histogram of Change in the VSA Over Recording Process. Control is Shown by '0.0', Dysphonia is shown by '1.0'.

et al. (2007). Due to how UncommonVoice was designed, the ability to measure the articulatory precision at the beginning of the recording process as well as the articulatory precision at the end of the recording process was afforded should the speaker make it through all of the prompts and record all four vowels in the first round and last round. In Figure 7.6, the results of the change in VSA from the beginning of the recording process to the end of the recording process is shown. On the left, the histogram of the change in VSA is shown for speakers without a voice disorder, while on the left the histogram of the change in VSA is shown for speakers with a voice disorder.

7.5.3 UncommonVoice Intelligibility Analysis

It was expected that as is consistent with previous results such as those in Section 5.1, the dysphonic speech would have a larger WER than the control speech. The results of the intelligibility analysis are shown in Table 7.3. On average, when fed into an ASR system, the ASR system recognized more words correctly in the control speech (7.46) than the dysphonic speech (6.35). There were more substitutions in the dysphonic speech (1.35) compared to the control speech (1.02). There were on average 0.45 insertions per utterance for dysphonic speech, while only 0.07 insertions

Table 7.3: Analysis of the Intelligibility of Control and Dysphonic Speech in UncommonVoice

Voice Type	Correct	S	I	D	WER
Control	7.46	1.02	0.07	0.29	0.15
Dysphonic	6.35	1.35	0.45	1.07	0.32

per utterance in control speech. The deletions showed a similar pattern with 0.29 average deletions per utterance for control speech and 1.07 average deletions per utterance for dysphonic speech. Overall, the WER for the control speech was 0.15, while the WER for the dysphonic speech was more than double that at 0.32. It is worth noticing that dysphonic speech seems to be recognized more successfully than dysarthric speech. The most common error that the ASR system made when transcribing dysphonic speech was substituting words, followed by deleting words.

To better understand what acoustic features might be correlated with the intelligibility– or in this case, the proxy for intelligibility that is the WER–the extent to which each acoustic feature is correlated with WER was investigated.

The most highly correlated feature with WER was the duration of the speech sample. This result is very similar to the result observed in 6.2. The Pearson Correlation Coefficient between the CPP and WER is 0.75, and the distribution of the duration of the utterance and the wer is shown in Figure 7.7.

The second most highly correlated acoustic feature was the cepstral peak prominence (CPP). This result was what we expected to find, as the CPP has been demonstrated to be a viable predictor of dysphonia in previous work Heman-Ackah *et al.* (2003); Samlan *et al.* (2013). The Pearson Correlation Coefficient between the CPP and WER is 0.6, and the distribution of the two features is shown in more detail in Figure 7.8.

The other features that were evaluated–jitter, shimmer, hnr, and f0–all showed

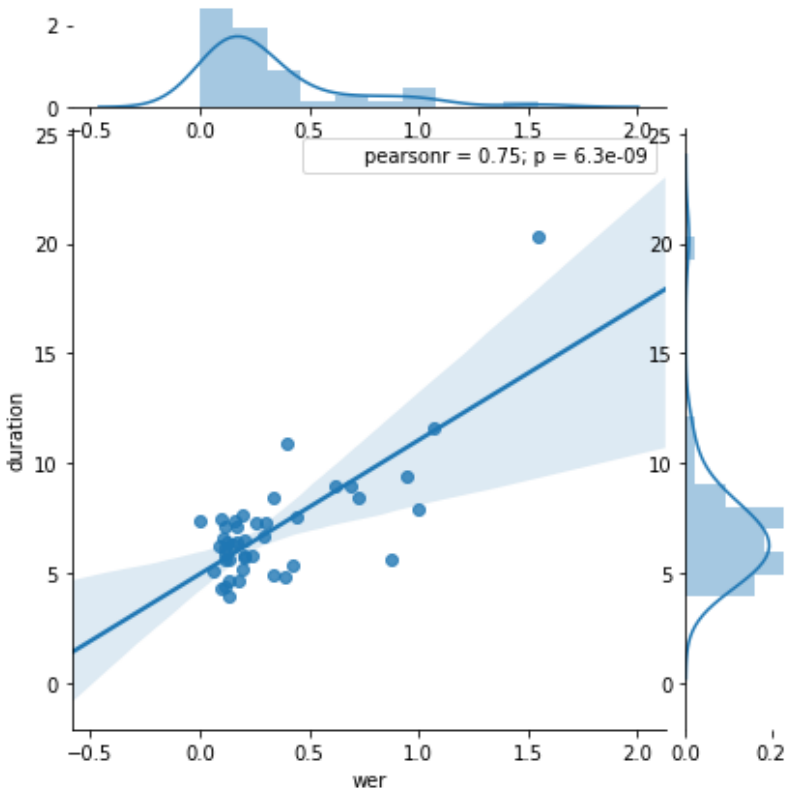


Figure 7.7: Correlation Between Average WER Per Speaker and Average Duration.

relatively low correlation with the WER for a given utterance.

7.6 Discussion

UncommonVoice is the first dysphonia dataset made freely and publicly available. As such, it is a significant contribution to the field and hopefully will fuel future research into improving the intelligibility of voice-based technology as well as the development of voice-assistive technologies. While still growing, UncommonVoice has 52 speakers–a volume that is unheard of in publicly available speech disorder datasets such as TORGO and UASPEECH which have a combined 27 speakers. However, where UncommonVoice falls a bit short is in the depth of the data for each speaker. As this dataset was collected in a distributed manner, the decision was made to keep the task of recording speech to be as quick and easy as possible. While this constrained

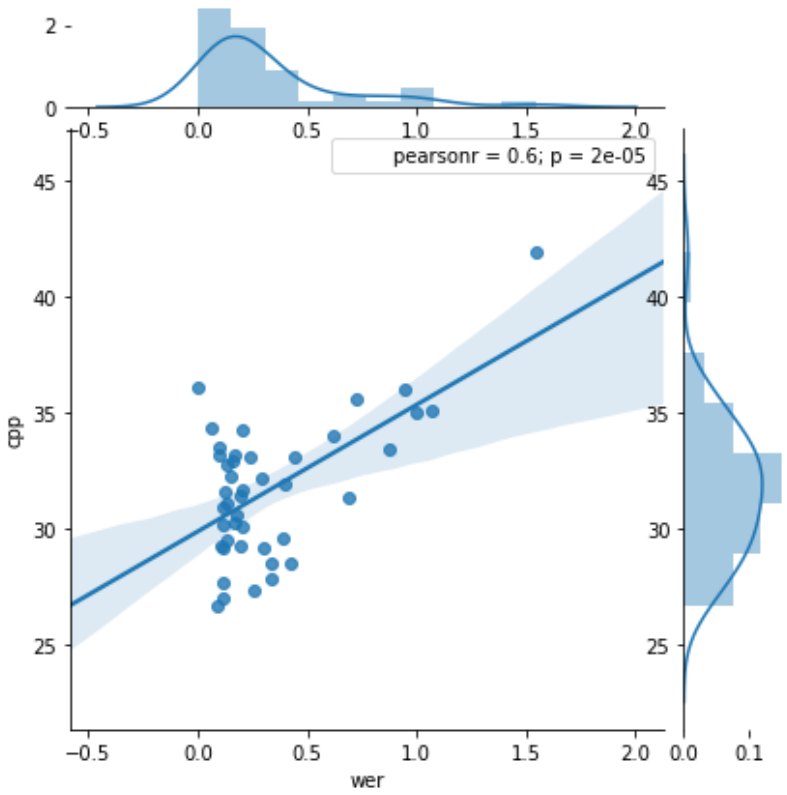


Figure 7.8: Correlation Between Average WER Per Speaker and Average CPP.

the total hours of data that were recorded in UncommonVoice to currently around 7.5 hours of data, this decision is also part of what led UncommonVoice to have as many speakers as it does.

The VSA for speakers in UncommonVoice was calculated and interestingly, for individuals with voice disorders, there seems to be an increase in VSA from the beginning of the recording session to the end. This is the opposite outcome from what we expected, as we expected the VSA to decrease due to vocal fatigue. This increased VSA could indicate that a 'warm-up' period for individuals with dysphonia could lead to articulatory gains.

When evaluating acoustic features that correlate with intelligibility (as measured by the WER of an utterance determined by an ASR system), the duration of the utterance was shown to be the most highly correlated feature that was explored.

While the idea that the duration of an utterance is important as the WER is inversely proportional to the total number of words in an utterance–as the total number of words increases, there are more opportunities for mistakes (insertions, deletions, or substitutions) to be made. A longer duration utterance doesn't necessarily equate more words in the utterance–it could also be caused by the speaker using a slower speaking rate. While the duration of the utterance may be indicative of a higher word error rate, this kind of correlation isn't particularly useful when attempting to obtain a more robust model of intelligibility that takes into account speech differences seen in conditions such as dysphonia.

Chapter 8

DESIGN CONSIDERATIONS FOR VOICE-ASSISTIVE TECHNOLOGY

8.1 Person-Centered Design

Person-centered multimedia computing (PCMC) is a paradigm that is sensitive to a specific user, task, and environment, leading to a shift in perspective from the population's needs to the individual's needs. This design paradigm starts with an individual, focusing on the individual's needs and specific problems that an individual encounters daily, and then adapts to a broader population through integration and adaptation.

Individuals with voice disorders often have specific requirements that necessitate a personalized, adaptive approach to multimedia computing. To address this challenge, our proposed approach places emphasis on understanding the individual user's needs, expectations, and adaptations toward designing, developing, and deploying effective multimedia solutions. The first step towards making a person-centric intelligibility support tool is gathering user input. While I have completed two online surveys of the experience of individuals with voice disorders, I want to conduct a focus group to discuss the needs and problems encountered in daily life of individuals with voice disorders. From these discussions, I hope to create a set of requirements for developing technologies for individuals with voice disorders.

In an ideal system, a user would speak into the input device of their choice, and the output would say exactly what they said in their voice, but a more intelligible version. To achieve such a system, it seems like we would need to have paired data, a less intelligible speech sample and a more intelligible speech sample from the same

voice. While usually in the dysphonia space this ask is unrealistic, for individuals with spasmodic dysphonia, it isn't necessarily that hard to imagine collecting a dataset that has paired samples from the same speaker where one sample is significantly more intelligible than the other because of the botox cycle. Botulinum toxin injections help decrease the number of larynogospams in the voice, therefore making the speaker's voice more intelligible for some time. Collecting a longitudinal dataset of speech from throughout the speakers' botox cycle could lead to the collection of a dataset that has paired samples between an individual's 'bad' voice and 'good' voice. However, if we are unable to achieve enough data to make this sort of ideal system work, a person-centered design consideration to include would be enabling the individual to choose their output 'voice'—the gender, pitch, accent, etc. Allowing an individual flexibility and options in the output voice makes this technology adaptable and flexible.

In person-centered multimedia computing, accessibility is woven into the design process through evaluating the person, the task, and the environment, and how the constraints of having a disability affect all of these factors. A person-centered solution uses tools like flexible inputs, multimodal outputs, user awareness, automated guidance, customizability, morphable interfaces, and content sensitivity to enable the user to adapt and integrate the technology to obtain a solution Panchanathan *et al.* (2012). There are two important processes from which a solution can be applied to multiple contexts: the processes of adaptation and integration. In taking a person-centered approach to the design of an intelligibility support tool, we will incorporate flexible inputs, customizability, modularity and morphable inputs and outputs. Given these attributes, this system will be able to adapt to many contexts and users will be able to integrate this process into their daily lives to help improve their quality of life.

Many different types of microphones could be used to capture the user's speech.

Each microphone has it's own set of affordances that lead to different user experiences. Figure 8.1 shows a few different microphone input options. A Bluetooth headset that includes a microphone could be used as shown in Figure 1a, or a microphone that plugs into the phone and can be directed towards where the speech is coming from as shown in 1b. Another option is a Bluetooth lapel microphone as shown in 1c. Depending on the situation, and the individual's personal preference, they will have the ability to select the input modality that best fits their preference, environment, and task. The user will also have the ability to input text if they prefer to type rather than speak. Enabling the user to choose the input modality and tool makes the system highly adaptable and able to be integrated into the user's daily life.

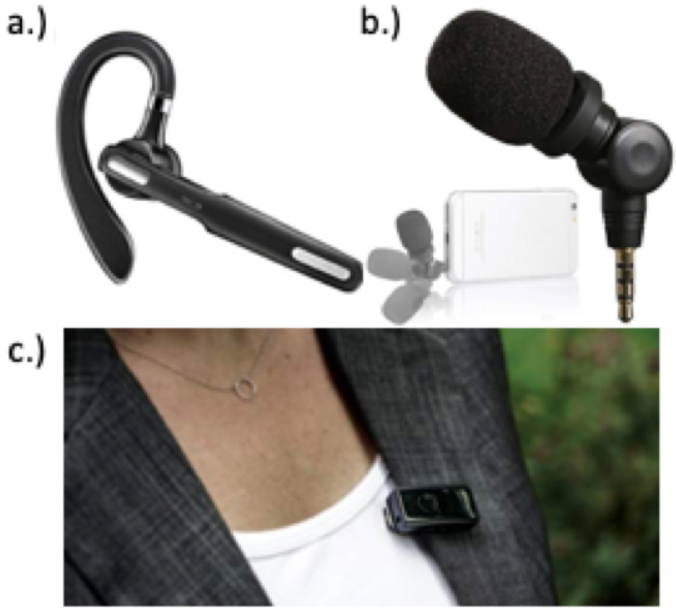


Figure 8.1: Different Audio-Input Options for Voice Assistive Technologies

Another way that this system is person-centric is in its morphable and multimodal output. The user will have the ability to choose what hardware they would like the output to be played on as well as what modality (either speech or text). For example, take one of the scenarios that have been reported in our initial survey: Susan has

SD and her husband is hard of hearing, and they would like to go out to eat at a restaurant. However, with the additional background noise, it is very difficult to understand Susan for the waiter, and even harder for her husband to understand her. In this scenario, Susan could use the speech intelligibility support tool by selecting her input, and output modality. In this case, it might work well for Susan to use a lapel microphone and have the output speech played to a set of headphones that her husband could use to cancel out the background noise and turn up the volume of Susan's output speech. When the waiter comes around to take their order, Susan could switch the output modality to either a portable Bluetooth speaker, the phone's internal speaker, or even use text to communicate her order to the waiter. Eventually, a feature that could be useful in this scenario, as well as many others, is using the background noise to select a reasonable output volume given the situation.

In another scenario gathered from speaking with individuals with voice disorders, a different individual with a voice disorder described her experience to me at an SD support group meeting. In her experience, she will completely lose her voice for months on end occasionally with no warning. For this individual, the option to use text input would be a very helpful feature.

A non-person-centered approach to improving the intelligibility of voice disorder speech would most likely result in far less flexible technology that would not help individuals with voice disorders be understood to the best of their abilities. Without involving the user in the development of this technology, I would not have known about the two situations above where users need intelligibility support. A non-person-centered system would likely result in a one-input, one-output speech-to-speech system. This system would fail to be useful in the above scenarios, as well as many other scenarios–for example if a user was in a quiet environment or required a little bit of privacy, having the option to output the speech as text would also be

quite advantageous for situations when it might be more convenient for the user to not user their own voice as an input.

Since person-centered technologies are inherently designed with an individual user in mind, they can sometimes be a bit rigid and inflexible. Technology developed using PCMC guarantees accessibility, usability, and often optimality for an individual. This granular level of design comes at the price of rendering the technology inflexible toward the broader audience. However, the Social Interaction Assistant Panchanathan *et al.* (2016) and the Autonomous Training Assistant Tadayon *et al.* (2016) provide examples of person-centric technologies that while inspired and designed to meet the explicit needs of an individual, also meet the implicit needs of a much broader audience through methods of adaptation and integration.

8.2 Designing Voice-Assistive Technology From Survey Results

8.2.1 Use and Design of Voice-Assistive Technologies

As part of the surveys described in 4, when asked to respond to the statement '*I would use a voice assistive technology that helped me to be better understood*' on a five-point scale between strongly disagree, and strongly agree, 63.16% of the respondents indicated that they would be willing to use an assistive technology that helped them to be better understood, the results from this question are shown in Figure 4.3.

To get a better idea of what kinds of technologies individuals with voice disorders are interested in having developed, we asked the question *'What kinds of technologies would you like to see developed for your voice disorder?*'. In general, the respondents had a strong preference for speech-based technologies (67.42%) rather than text-based (9.63%). As mentioned above, the only voice-assistive technology that exists today for individuals with voice disorders is an amplifier. As this is what people are most

Table 8.1: Responses to the Question *"What Kind of Technologies Would you Like to See Developed for Your Voice Disorder?"*

Response	**Response Rate**
Better amplification system	19.55%
Improve phone-based interactions	15.86%
Better automatic speech recognition systems	9.63%
Better text-to-speech systems	9.63%
Speech smoothing device	9.35%

familiar with, it makes sense that the response 'a better amplifier' was a frequent answer (19.55%), followed by the answer 'I'm not sure' at 16.43% of the respondents. The next most prevalent answer was that they wanted technology to be developed to help them be better understood on the phone with 15.86% of the respondents. This answer is followed by 'better automatic speech recognition systems' (9.63%), 'better text to speech systems' (9.63%), and 'speech smoothing device' (9.35%). Of course, many people just wished to see a cure developed (7.65%), and others hoped to see the development of better and less invasive treatments (7.37%). One thing that was consistent across most answers was that these individuals just want their 'normal' voice back. They want to be able to be understood without the stress of worrying about what their voice will sound like, and if they will be understood or judged. On top of these ideas, there was an emphasis that whatever technology will be developed should be both unobtrusive and affordable.

Voice-assistive technologies should take into account the user's preferences as described in Section 8.2.1. Technologies should be focused on helping individuals with voice disorders be better understood while being unobtrusive and affordable. It was made clear through this survey that individuals with voice disorders would much rather communicate through speech-based systems than text-based systems, and as

such future, voice-assistive technologies should focus on using speech as the input rather than text. As the average age of individuals with SD is 62 years old, any voice-assistive technologies that are developed must be inherently user-friendly.

By improving the intelligibility of individuals with voice disorders we can help them be better understood and help put them in a better position to fully participate in social interactions, acquire, maintain, and advance their careers, and in general, maintain a better quality of life.

Chapter 9

CONCLUSIONS

9.1 Contributions

The contributions of this dissertation are all centered around the development of more accessible technologies for individuals with voice disorders. As outlined in Section 1.2 the contributions of this dissertation are as follows:

- A broad survey of the needs of individuals with dysphonia, including areas of opportunity for voice-assistive technologies to improve the quality of life of individuals with dysphonia.
- An evaluation of the accessibility and inclusivity of state-of-the-art voice-based technology
- A more nuanced and in-depth analysis of what it means for speech to be intelligible
- UncommonVoice: the largest publicly-available dataset of dysphonic speech, as well as accompanying metadataset.
- Design considerations for the development of voice-assistive technologies

9.2 A Broader Definition of Intelligibility

While on the surface, the concept of intelligibility seems deceivingly simple–the ability of speech to be understood–it is inherently more complicated than this definition makes it seem. Intelligibility is both a relative and dynamic concept, continually

changing based on the ability of the listener and the speaker. Intelligibility can be represented as either a scalar concept–the extent to which the message was received, or a binary concept–whether or not the message was received. Many external factors affect the construct of intelligibility from the native language of the speaker/listener, to the non-verbal cues and context provided by the surroundings. Communication is a multimodal and messy interaction, and as a measure of communication, intelligibility should also reflect these intricacies.

9.2.1 A Qualitative Evaluation of Dysphonic Needs

The first contribution is a deep understanding of the experience of individuals with voice disorders through two surveys. In the first survey, particularly difficult situations for individuals with voice disorders including speaking on the phone, ordering at a drive-thru, communicating in a noisy environment, and meeting someone new for the first time. This survey also provided some directional guidance as to areas that required further questions to get a broader view of the experience of individuals with voice disorders. In the follow-up survey, questions developed from the initial survey such as more details about the emotional impact of having a voice disorder as well as the impact of having a voice disorder on employment and careers. In general, these surveys found that having a voice disorder significantly impacts most individual's social life, emotional wellbeing, and career. Often individuals with voice disorders experience isolation, frustration, stress, and anxiety.

The second survey also helped to get a better idea of what experience individuals had with assistive technologies as well as what kinds of characteristics and functionality they would be looking for in assistive technology. While 89% of our respondents recognize having some sort of limitation or barrier due to their voice disorder, and 20% of these individuals have used assistive technology before, only 3% of individuals

continue to use assistive technology to mitigate these barriers. From this data, we can conclude that the existing voice-assistive technologies are not meeting the needs of the population of individuals with dysphonia.

9.2.2 Accessibility and Inclusion of ASR Systems

Currently, state-of-the-art voice-based technologies do not recognize voice disorder speech with the same accuracy that it does control speech. This performance difference creates barriers for individuals with disabilities and implies that today's voice-based technologies are not inclusive or accessible for individuals with voice disorders. From the data used to assess the performance of these ASR systems, we created SayWhat? a metadataset that consists of intelligibility data for a wide range of speech–from control speech, accented speech, to voice disorder speech. This dataset was published to provide more data to help build more robust models of intelligibility.

9.2.3 UncommonVoice

The crowdsourced collection of speech from 52 individuals, 35 of whom have a voice disorder for a combined 7.5 hours of speech data is a significant contribution to the field. Dysphonic speech datasets exist, however, most of them sit behind the barrier of a clinic and are not readily accessible for research purposes. By making UncommonVoice available to researchers, we expect to fuel research in the field of voice-assistive technologies. UncommonVoice is still accepting speech contributions, and we hope to continue to collect speech from individuals with voice disorders.

9.2.4 Voice-Assistive Technology Design Considerations

From the need-finding surveys discussed above, this dissertation builds out a few guidelines for the development of voice-assistive technology such that these devices

meet the needs of individuals with dysphonic speech. The main guidelines are: if a user can speak, they would rather speak than write out what they are saying, devices need to be affordable, and devices need to be as unobtrusive and flexible as possible.

Chapter 10

FUTURE WORK

10.1 Uncommon Voice Extensions

10.1.1 Other Intelligibility Measures

There are still a few steps that need to be completed for UncommonVoice to be as useful as possible to the research community. The first of these steps is obtaining human intelligibility data–by using Amazon Mechanical Turk to get orthographic transcriptions of the data as well as a rating of how difficult it was for the listener to understand what was being said, and the time taken for the listener to complete the transcription. UncommonVoice also contains the sentences necessary for clinicians to provide a CAPE-V score. CAPE-V is a voice quality measurement that is commonly used in the realm of speech-language pathology and is described in more detail in 3.3.1. Having more measures of vocal quality and intelligibility will lead to a more well-rounded and accurate picture of the intelligibility of dysphonic speech.

10.1.2 Automatic Dataset Cleaning

There are a few different possible approaches to automate cleaning the dataset. In previous work McGraw *et al.* (2010) proposes using an automatic speech recognition system in the loop to evaluate the utterance as it is collected to recognize speaker errors. While this is a good idea, in theory, automatic speech recognition systems (ASRs) do not perform very well of speech from individuals with voice or speech disorders Moore *et al.* (2018). Instead, we could set a threshold for which if the WER is higher than the average WER from the given ASR system's performance

on dysphonic speech, then we could flag these samples as potentially unusable and manually review flagged samples. If we implement this into the existing system, we could also provide some sort of live feedback—-for example, asking a participant to move closer to the microphone—to help get the participant back on track.

To identify speech samples that have been clipped, we could write a script to check for silence on either end of the speech. If there is not a small amount of silence before the speech and after the speech, clipping has likely occurred. To identify and flag low-quality signals, we could employ one of several unintrusive speech quality metrics Falk *et al.* (2005); Santos *et al.* (2014); Gray (2000) and set a threshold for which if a given speech sample falls above or below (depending on the directionality of the metric) a given threshold it will be flagged for further review. Integrating these results could lead to a ranking of samples that need to be manually reviewed. For example, if a speech sample was flagged as having an above-average WER, low-quality and potentially clipped, it would be on the top of the list of samples to manually review. In doing so, we prioritize severely unusable speech samples for manual review and hopefully remove the least usable data from the dataset.

Another approach to automating the process of quality control for Uncommon Voice is to use Amazon's Mechanical Turk (AMT) (an online crowdsourcing tool where workers are paired with Human Intelligence Tasks and are paid in micro currency as low as $0.005 per task). In this case, we could present the speech samples to an AMT worker and ask them to transcribe the speech, as well as provide a Mean Opinion Score (MOS) rating the quality of the speech, and their confidence in their transcription. Speech samples that receive a low-quality rating or a transcription that is vastly different than the expected prompt would be flagged for manual review and removal from the dataset. Having multiple workers review each speech sample would also enable us to employ standard merging/voting algorithms like ROVER to

identify inter-rater reliability and be even more confident about flagging particular speech samples for manual review.

The main advantage of using the human-in-the-loop feedback such as the feedback obtained by AMT is that you are collecting actual human perceptual data. However, sometimes AMT transcriptions can be difficult to interpret or score because of human tendencies to misspell words, use truncated abbreviations for words, and/or insert punctuation. These human errors may or may not be related to the intelligibility or quality of the speech, and that distinction is difficult to tease out. AMT also has a cost associated with it, while the objective methods mentioned above are all free (except the use of an external ASR system, however, this cost is relatively negligible).

10.1.3 Machine Learning Experiments with UncommonVoice

There are a few models that would be particularly interesting to use UncommonVoice to build. There are so many different tasks that could be completed using machine learning and UncommonVoice. For example, fine-tuning an automatic speech recognition system with UncommonVoice is expected to lead to a speech recognition system that is more tolerant of speech from individuals with voice disorders. Predicting intelligibility and severity of the voice disorder is also a potential task that UncommonVoice could be used for. Being able to automatically obtain a rating of severity of a voice disorder is a potential application of machine learning using UncommonVoice. Voice disorder classification is also a potential task that UncommonVoice could be used for.

10.2 Machine Learning to Improve Intelligibility

To meet the needs of users with voice disorders, I am proposing the implementation of an intelligibility improving speech support tool. This system will take in

speech from an individual with a voice disorder and with a reasonably low-latency output speech that is more intelligible than the input speech. In my proposal, I laid out a few different road maps to potentially building a system like this based on current literature in the voice-conversion field as well as closely related fields. In voice conversion, there are several state-of-the-art methodologies of achieving an effective voice conversion system. Several different parameters define most voice conversion systems, mainly dealing with the number of input and output voices, as well as the structure of the data that was used to train the model. For example, there are many-to-many voice conversion systems with either parallel or non-parallel speech corpus inputs. The most simplistic voice conversion system consists of a parallel one-to-one corpus of speech. A voice conversion system trained on a parallel one-to-one corpus of speech would be very limited. It would only be able to convert from the source voice to the target voice.

10.2.1 Voice Conversion Techniques

Within the last few years, there have been several significant developments in the field of voice conversion. Voice conversion models that use non-parallel data for many-to-many conversions are now state-of-the-art, and several different generative modeling techniques are used. Voice conversion is a somewhat more difficult and nuanced task than speech enhancement in that voice conversion must function on a sequence-to-sequence level. Voice conversion data do not have the luxury of having equal, time-aligned lengths for the source and target speakers. A few years ago, it wasn't uncommon to see a voice conversion system using Dynamic Time Warping followed by a Gaussian Mixture Model, however, now advanced generative techniques are utilized. Generative Adversarial Networks are a popular choice among researchers working on voice conversion tasks Kameoka *et al.* (2018); Hsu *et al.* (2017a); Kaneko

et al. (2019). While Kameoka *et al.* (2018) uses a relatively vanilla GAN formation, Hsu *et al.* (2017a) uses Variational Autoencoder as the generative model within the GAN, and Kaneko *et al.* (2019) uses a Cycle-GAN, employing the cycle-consistency loss to make sure that it is possible to recreate not only the target from the source but also the source form the target. Autoencoders are another popular choice when it comes to implementing a voice conversion system partially due to the ability of an autoencoder to come up with latent representations of a speaker's identity Hsu *et al.* (2017a). Pascual *et al.* (2018) Serrà *et al.* (2019)

The minimum criteria for creating a voice conversion system to use as a speech support tool is a many-to-one voice conversion system similar to the one demonstrated in Biadsy *et al.* (2019). In Biadsy *et al.* (2019), they take a relatively unique 'voice normalization' approach and create a many-to-one parallel corpus by synthesizing the speech from the target speaker. Using a 'single canonical speaker's they call it– inherently improves the intelligibility of the speaker's voice by converting the speech into the target speech. This is the only other current paper that is attempting to do something similar to what I am proposing in this work. In another similar work, the authors of Pascual *et al.* (2017) reimplement the speech enhancement generative adversarial network (SEGAN) as a voice conversion system that takes alaryngeal– whispered–speech, and converts it into voiced speech Pascual *et al.* (2018) through only a few small changes to the SEGAN network.

10.2.2 Intelligibility Optimized Model

In this approach, a WER estimation model is utilized to enhance the intelligibility of speech. After training a successful WER estimation model, we would use this network as a component of the loss function of an end-to-end voice conversion model that takes in voice disorder speech and outputs more intelligible speech. In this model,

the 'mentor' network being incorporated into the loss function of the generative model could be one or a combination of a few different models learned from the intelligibility data as discussed above, or it could be a multitask learning system that makes several predictions about the input utterance.

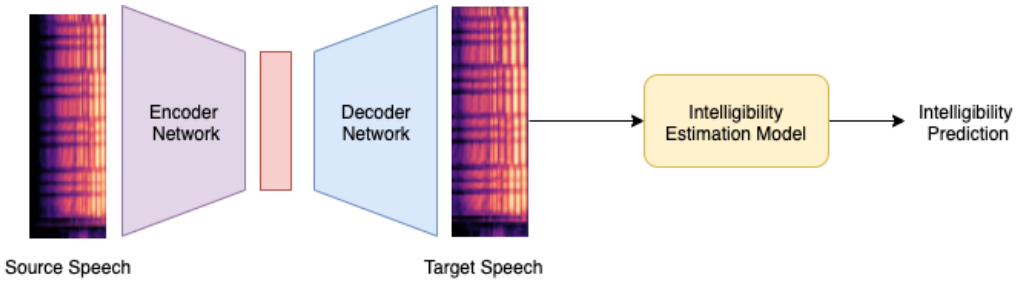


Figure 10.1: Overview of the Intelligibility Optimized Model.

The first step of building this intelligibility optimized model is building an intelligibility estimation model. This estimation model is trained on spectrogram inputs paired with an automatic speech recognition system's word error rate of the given utterance. The goal of this model is to quickly and efficiently estimate the intelligibility of a given speech sample. Using an automatic speech recognition system to obtain the actual WER of a given speech sample is too time-intensive, taking on average 1.87 seconds per utterance, while the inference step in the intelligibility estimation model takes 0.000129 seconds per utterance. While estimating the WER of a given speech utterance may compromise the accuracy, the decreased computation time makes integrating a prediction about the intelligibility of a speech sample into the optimization of a neural network possible.

Once we have a reliable intelligibility estimation model trained, we will integrate this estimation model into an intelligibility enhancement model which is largely based on sequence-to-sequence voice conversion models.

Voice conversion models consist of three main parts, the data pre-processing, acoustic modeling, and speech synthesis from the output of the acoustic model. For

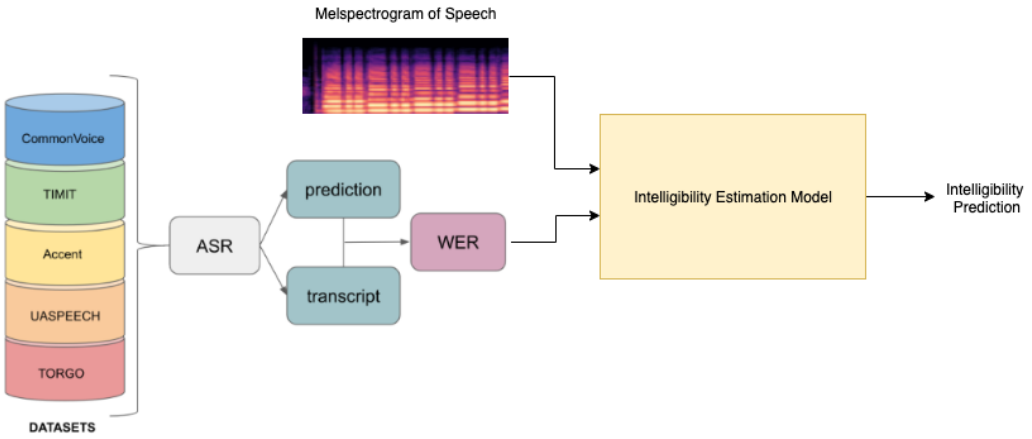


Figure 10.2: Overview of the Intelligibility Estimation Model Used in Moore *et al.* (2019)

each of these sections, there are several different options to build out an intelligibility enhancement model.

As far as pre-processing, the most common input into voice conversion systems is a log-melspectrogram. In Biadsy *et al.* (2019); Wang *et al.* (2017); Zhang *et al.* (2019b), log-melspectrograms are used as inputs into model. In the current implementation, I chose to use 80-dimensional log-melspectrograms from speech sampled at 16000 Hz with 12.5 ms overlap, a Hann window, and a 1024 point SFTP. While there are advantages to using raw-audio like avoiding confounding variables that standard speech features are sensitive to, see description from Chapter 3.3, and recently there have been models that have gotten close to state-of-the-art that are trained on raw-audio Serrà *et al.* (2019), using raw-audio does not seem like the smartest input to use for a system that requires as real-time of predictions as possible.

Adversarial Learning Based Enhancement

We propose an adversarial learning framework to enhance the intelligibility of dysphonic speech. The standard Generative Adversarial Network (GAN) model Goodfellow *et al.* (2014), generates data from noise inputs z sampled from a distribution $P_z(z)$. A

vanilla GAN consists of a generator neural network G that takes $z \sim P_z(z)$ as input and generates $G(z)$ which tries to imitate real data $x \sim P_{data}(x)$. The GAN also has a discriminator neural network D which tries to distinguish between real data x and generated fake data $G(z)$. The adversarial framework with G and D competing against each other aligns the distributions of $G(z)$ and x.

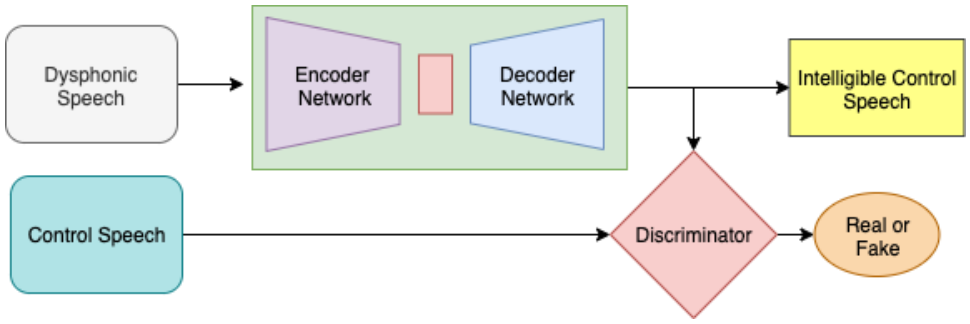


Figure 10.3: Proposed Adversarial Learning System for Generating Intelligible Speech.

We propose to apply an adversarial framework to generate control/target speech from dysphonic speech. Computer vision literature outlines related frameworks for applications in domain adaptation and image translation Isola *et al.* (2016). In our model shown in figure 10.3, the generator is the Intelligible Speech Generator network which is made up of an encoder network, a bottleneck layer, and a decoder network. This system will take melspectrograms from dysphonic speech and control speech as input.

$$\min_G \max_D = \mathbb{E}_{x \sim P_{\text{data}}(x)}[\log(D(x|y))] + \mathbb{E}_{z \sim P_z(z)}[\log(1 - D(G(z|y)))] \tag{10.1}$$

The discriminator learns to distinguish between 'real' data (the intelligible speech), and the generated data (enhanced speech). We also propose using a variation of the Conditional GAN (CGAN) Mirza and Osindero (2014) as a way to embed domain knowledge about vocal quality and intelligibility into the network, potentially in the

form of a perceptual quality assessment that has been conducted by a clinician. This conditional information is denoted in equation 10.1 as y.

10.2.3 Multimodal Intelligibility Improvement

The majority of the approaches to speech processing use only audio input despite the propensity for audio to be compromised by external noise. Alternative sources of information are becoming more readily available with the increasing use of multimedia data in everyday communication. The primary alternative source of information used in speech is vision, as it is naturally used in human speech processing.

Humans process the world in a multimodal way, often relying on visual information in their perception of speech. The perceptual integration of vision and hearing is demonstrated through the McGurk effect—an auditory illusion where the auditory component of one sound is paired with the visual component of another sound, leading to the perception of a third sound McGurk and MacDonald (1976). This effect demonstrates that visual information is an important part of how humans recognize and perceive speech. In noisy settings, humans routinely exploit the audio-visual nature of speech to selectively suppress the background noise to be better able to focus on the target speaker Zekveld *et al.* (2008).

It follows, then, that adding visual information to a speech processing model would add information to the system and help the model better perform speech processing tasks. There have been several papers that have focused on using the visual modality to better accomplish speech processing tasks such as source separation Wang *et al.* (2005), speech enhancement Afouras *et al.* (2018); Hou *et al.* (2017); Gabbay *et al.* (2017), and speech recognition Cooke *et al.* (2006). In general, for both humans and computers, bimodal perception leads to better speech understanding than auditory perception alone Tiippana *et al.* (2004); Rosenblum (2008); Mroueh *et al.* (2015). The

visual modality enhances the performance of speech processing when compared with auditory-only speech processing.

While currently, the proposed model of speech intelligibility enhancement focuses on using only the auditory modality, combining audio and visual modalities for the enhancement of the intelligibility of the speech seems like a promising endeavor. A multimodal approach to intelligibility enhancement of dysphonic speech would rely on both audio and visual information. To implement such a model, we would need to have a corpus of audio-visual speech from individuals with voice disorders. While not mandated, in the data collection process of UncommonVoice (the dataset that I am collecting as part of my dissertation), the user has the option of contributing video samples as well as audio samples. While we don't expect a large majority of the participants to choose to contribute video as well as speech, we hope to collect at least some audio-visual dysphonic speech samples in the coming months.

Audio-Visual Speech Models

The first step of any audio-visual intelligibility enhancement model would be to pre-process the data, on top of the normal audio pre-processing, this would also include using existing algorithms to recognize the bounding box of the lips and cropping the lips out of the images. The Viola-Jones algorithm is commonly used to locate the lips on a face Wang (2014). These cropped sections are generally resized to be of equal size and are generally recolored as well. Once the images are cropped, resized and color-corrected, they would then be used as input into a deep neural network.

One major challenge of pre-processing is making sure that the audio and visual elements are precisely aligned. Any misalignment can be extremely disruptive–just think about when a video lags and the audio and visual components are out of sync. The effect of mismatched visual features is evaluated in Hou *et al.* (2017). They find

that speech that was enhanced with speech signals that were paired with incorrect lip feature sequences lacked the detailed structures that were otherwise preserved incorrectly matched audio and visual features.

Integrating heterogeneous information streams is a challenging task. When, and how, to fuse modalities is still up for debate in the literature. In Hou *et al.* (2017), they use deep convolutional neural networks as an audio-visual encoder-decoder network in which audio and visual data are first processed using individuals convolutional neural networks, and then later fused into a joint network to generate enhanced speech. This is an example of late-fusion, as the two modalities are fused after being input into independent networks. In Papandreou *et al.* (2009), they recognize that sensory information is often fused dynamically. That is, humans adapt which multimedia channel they pay the most attention to based on environmental conditions. For example, having a conversation in an environment with a lot of background noise will lead a person to focus more on the visual modality of speech than normal. The authors propose an adaptive algorithm for multimodal fusion that takes into account the uncertainties of the different modalities.

In the case of audio-visual speech enhancement, the task of enhancing speech is simplified as in post-processing they can add noise and obtain a clean and noisy speech signal. In these speech enhancement paradigms, paired clean and noisy speech and visual samples are input into their respective networks based on the modality (e.g. an audio network and a visual network as shown in Figure **??**, and the output of these networks is merged in the fusion network. The output of the fusion network is enhanced speech and reconstructed mouth images. The loss function that they use to train this audio-visual speech enhancement network is just the mean-squared error (MSE) between the clean and noisy audio plus the reconstruction loss of the mouth images which is multiplied by a scaling factor μ. he number of image frames and

audio frames is forced to be equal and the number of frames is represented in 10.2 by K. In equation 10.2, Y represents the audio output, $\hat{Y}$ being the clean audio, and Y representing the predicted audio. In this equation, X and $\hat{X}$ represent the reconstructed and original visual inputs respectively.

$$\min_{\theta} \left(\frac{1}{K} \sum_{i=1}^{K} ||\hat{Y}_i - Y_i||_2^2 + \mu ||\hat{X}_i - X_i||_2^2 \right) \tag{10.2}$$

To adapt this model to train a multimodal intelligibility improving system, the first step would be to obtain intelligibility data. This data can be obtained from an objective measure of intelligibilitySantos *et al.* (2014); Gray (2000), from subjective measures of intelligibility Streijl *et al.* (2016), or a combination of the two. We would need to build a dataset that has pairs of clean and noisy speech. In this case, noisy speech would refer to pathological speech while clean speech can either be synthesized speech or if the individual's voice has a cyclic property (such as the BTX cycle in individuals with Spasmodic Dysphonia), it could be speech from the individual during a 'good voice' period.

There are a few adjustments that we would need to make to the model used in Hou *et al.* (2017) to change the task of the model from enhancing/denoising the speech to improving the intelligibility of the speech. The first step would be to turn the model into a sequence-to-sequence model as the length of the clean and noisy speech is not going to be the same. Sequence-to-sequence models have been employed in audio processing models, particularly voice conversion models quite often Zhang *et al.* (2019b,a); Narayanan *et al.* (2019); Tanaka *et al.* (2018); Weiss *et al.* (2017). Most of these seq2seq voice conversion models rely on an encoder-decoder system that takes in melspectrograms from the source speaker and output melspectrograms from the target speaker. In Sadeghi *et al.* (2019) they take a multimodal approach to speech enhancement, however, they use a conditional variational autoencoder to enhance the

speech. A model similar to this could potentially be implemented for the intelligibility improvement.

Usability of a Multimodal Intelligibility Enhancement System

If an audio-visual intelligibility model were to be implemented, the user experience would significantly change. Rather than just using a microphone as the input to the system, the user would also have to take a video of their face. While this has become a more common paradigm of interacting with technology (SnapChat, Instagram, FaceTime and other Video Chat/live streaming tools), there still exist some social boundaries that would make taking a video of oneself somewhat uncomfortable. Adding a visual element also adds another layer of interaction that could potentially complicate the application and make it less usable. Adding visual data will also likely increase the latency of acquiring the intelligibility enhanced speech. These usability considerations would have to be tested before deciding one way or another. The amount of intelligibility improvement from incorporating the visual modality would need to be evaluated with users to see if it would be worth the extra inconvenience of needing to have a video of their face to have their intelligibility improved.

I hypothesize that it all depends on whether the speech can be improved above the intelligibility threshold with audio alone. As this is the main goal of this project, if the speech can be improved by using either the native microphone, or a Bluetooth microphone (either attached at the ear, lapel, or phone), I would expect users to choose the simpler form of input. However, if audio didn't cut it as the only input source and failed to improve the speech to the point where it was understood, then I would expect users to be willing to take a video of their face while they are speaking to be better understood. I also would expect a system that required visual input to be used less often as intelligibility support than a system that requires only audio

input. These expectations would have to be tested in a usability test before making these conclusions, however.

In summary, while multimodal learning can produce potentially superior performance compared to audio-only intelligibility improvement, adding the visual modality to the model will increase the dimensionality of the input data, and make the model more complex. We expect that this complexity will increase the latency to obtain a prediction, as well as increase the social barrier of using an intelligibility support tool as it is more socially invasive. Using an audio-only system provides lower dimensionality and a less invasive input modality in practice, however, the performance of the system will be potentially inferior to a system that fuses both audio and visual data.

10.3 Next Steps to Build Person-Centered Voice-Assistive Technologies

As we are taking a person-centered approach to building an intelligibility support tool, involving the end-user in the development of the system is incredibly important. The first step of this has been completed through a series of online surveys evaluating the experience of individuals with voice disorders.

The next step in developing this technology through inclusive design is to run a focus group of individuals with voice disorders and talk about how an intelligibility support tool might be implemented in their daily lives. In this focus group, different forms of input and output will be presented to the group, and a discussion will be had about the preferences and requirements of an intelligibility support tool to make it optimally useful. In this focus group we will also discuss the importance of factors like the naturalness of the voice, whether or not the output voice sounds like them, the importance of latency and maximum amount of latency that is acceptable, and what kinds of wearables would be most acceptable. We will also ask the focus group participants to discuss the predetermined scenarios that are particularly difficult for

individuals with voice disorders and will ask them to rank these situations in order of increasing difficulty. Results from this focus group will drive the development of specific requirements of the intelligibility support tool. Depending on how the focus group goes, we may potentially ask users to run a diary study over a week and detail their experiences communicating. This would give a window into what it's like to live with a voice disorder and will provide insight into even more situations where we should prove our speech support device.

After these requirements are built out into a mobile application that improves the intelligibility of the user, we will run a series of usability tests to evaluate the usability of the device. In this study we will test each of the scenarios that we found to be particularly difficult for individuals with voice disorders: talking on the phone, speaking in environments with background noise, ordering something at a drive-thru, and meeting someone new. It turns out that all of these scenarios can be experienced at your local Starbucks, so to run these usability tests, I intend to partner with Starbucks. The following sections explain how I would test the usability of the intelligibility support tool in the given particularly difficult situations. As mentioned, the order of these usability tests would depend on the results of the prior focus group, and we would start with the task that was identified as the most important to individuals with voice disorders.

10.3.1 The Starbucks Intelligibility Challenge

For this usability test, we would recruit individuals with voice disorders from the Phoenix area and work with a local Starbucks location. The first phase of the usability study would be a short pre-interview, where we would ask the user to input what their Starbucks order was to use as the groundtruth, as well as ask them standard questions about their voice and how their voice affects their life, and their previous experiences

Table 10.1: The Voice Handicap Index 10 Rosen *et al.* (2004). The Frequency Scale is 0: Never, 1: Almost Never, 2: Sometimes, 3: Almost Always, 4: Always

Situation	**Frequency**
My voice makes it difficult for people to hear me	0 1 2 3 4
People have difficulty understanding me in a noisy room	0 1 2 3 4
My voice difficulties restrict my personal and social life	0 1 2 3 4
I feel left out because of my voice.	0 1 2 3 4
My voice problem causes me to lose income.	0 1 2 3 4
I feel as though I have to strain to produce voice.	0 1 2 3 4
The clarity of my voice is unpredictable.	0 1 2 3 4
My voice problem upsets me	0 1 2 3 4
My voice makes me feel handicapped	0 1 2 3 4
People ask, "What's wrong with your voice?"	0 1 2 3 4

ordering at drive-thrus. Here we will obtain a baseline for the questions that we will ask in the post-interview. I plan on using the Voice Handicap Index-10 Rosen *et al.* (2004), shown in Table 10.1. The VHI-10 is comprised of statements that many people with voice disorders have used to describe their voices and the effects of their voices on their lives. VHI-10 is often used to get a base assessment of the individual's quality of life and how their voice affects their quality of life. The main task of this usability study will be for the participant to place a drive-thru order at Starbucks in two conditions, one with the intelligibility support tool, and one without. The conditions would be counterbalanced to avoid any ordering effect (pun intended).

To complete the first task, we would ask the participant to drive their car, the researcher would either ride along or set up a recording device (both audio and visual) to capture the interactions between the driver (the individual with a voice disorder) and the cashier. To assess this interaction, there are several things that we would look for:

asking the driver to repeat themselves or generally asking for clarification, any errors in the order that they received (as depicted by the receipt and the difference between the ground truth and the order on the receipt). Between each condition, we would ask the participant a series of reflection questions in a post-interview. These questions would relate to their satisfaction with using the device, the difficulty of completing the task, and things they generally found good, bad, surprising or confusing.

For the second study, we will repeat the first task, however rather than going through the drive-thru, the individual will be asked to order coffee inside of the Starbucks, where there is lots of background noise–other conversations, music, coffee grinders, and steamers. This interaction will be video recorded and the interaction will be assessed similarly as the drive-thru study. We will look for any trouble being understood by the barista, either in the form of the barista asking for clarification or for the individual to repeat themselves. We will also check the accuracy of the order. We will also conduct a post-interview, where the participant will be asked to reflect on their interaction and ask them the same quantitative rating scale questions as we will ask after the drive-thru task.

To avoid any adaptation to the participant's voice by the barista, we will make sure that a different barista takes the individual's order, and we will make sure that the order is different between the two conditions. We will shoot to have 5-10 participants for each study, and will work with Starbucks so that they know what is going on, and so that we have the approval to record the interactions. Both of these studies will be completed in person, and I will recruit participants through the local Spasmodic Dysphonia support group.

This study covers three of the four situations that were identified in need-finding surveys: ordering at a drive-thru, speaking in noise, and meeting someone new. This just leaves speaking on the phone for evaluation.

10.3.2 Speaking on the Phone

From the initial need-finding survey, we have received two main complaints from individuals with voice disorders about speaking on the phone. The first is not being understood by auto-attendant systems, and the second is generally not being able to be understood by other people on the phone. In this study, we will test the intelligibility of individuals with voice disorders when speaking to an auto-attendant (machine), and when speaking to another person. The speaker will complete this study under two conditions, once without an intelligibility support tool, and once with an intelligibility support tool.

To test the auto-attendant feature, we will use a combination of Zoom and Amazon Connect. We will build an auto-attendant system using Amazon Connect. The participant will have a scheduled time to call into a Zoom session and from the Zoom session, we will have them call the auto-attendant. This means that we will be able to listen to their interaction with the auto-attendant. The first task will be for them to provide a fake account number (that we will provide via email before the study). We will use Amazon Connect's automatic speech recognition to transcribe the account number. From this, we will be able to calculate an error rate between the true account number and the transcription of the account number. We will also have the auto-attendant ask a couple of simple questions, for example, 'What is your favorite color?', 'How old are you?', etc. This will give us a baseline estimation of how auto-attendants understand dysphonic speech.

Integrating the speech support system into a phone call may prove difficult. If we can integrate the speech support system into a Voice Over Internet Protocol (VOIP) service, we will test the intelligibility of the participant in two different conditions, one with the speech support tool, and one without. The user will be given a series of

'messages' to provide over the phone. These will be practical prompts like 'Harvey has a dance recital at 9:30 am on the 15th of June, 2020.' or other scenarios that have enough information to be slightly complicated, however, are still very realistic situations to be communicating over the phone about. We will ask the participant to call a phone number that we provide, and we will have a standard listener on the other end of the line transcribing the message. We will try to keep the interaction between the speaker and the listener as realistic and natural as possible by allowing the listener to ask for clarification. The interactions will be recorded and analyzed similarly to the Starbucks intelligibility challenge with pre- and post-interviews.

The next step after any usability issues are identified in the above studies will be to have a group of Beta testers download the application and use it for a week while asking them to do a diary study detailing their interactions using the speech support tool.

REFERENCES

Adler, C. H., B. W. Edwards and S. F. Bansberg, "Female predominance in spasmodic dysphonia", Journal of Neurology, Neurosurgery & Psychiatry 63, 5, 688–688 (1997).

Adnene, C., B. Lamia and M. Mounir, "Analysis of pathological voices by speech processing", in "Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.", vol. 1, pp. 365–367 vol.1 (2003).

Afouras, T., J. S. Chung and A. Zisserman, "The conversation: Deep audio-visual speech enhancement", pp. 3244–3248 (2018).

Allen, J. B., "How do humans process and recognize speech?", IEEE Transactions on Speech and Audio Processing 2, 4, 567–577 (1994).

Aminoff, M. J., H. H. Dedo and K. Izdebski, "Clinical aspects of spasmodic dysphonia.", Journal of Neurology, Neurosurgery & Psychiatry 41, 4, 361–365 (1978).

Anusuya, M. and S. Katti, "Front end analysis of speech recognition: A review", International Journal of Speech Technology 14, 99–145 (2011).

Aronson, A. and D. Bless, *Clinical Voice Disorders*, Thieme Publishers Series (Thieme, 2009).

Awan, S. N., N. Roy and C. Dromey, "Estimating dysphonia severity in continuous speech: Application of a multi-parameter spectral/cepstral model", Clinical Linguistics & Phonetics 23, 11, 825–841, URL `https://doi.org/10.3109/02699200903242988`, pMID: 19891523 (2009).

Baker, J. M., L. Deng, J. Glass, S. Khudanpur, C. h. Lee, N. Morgan and D. O'Shaughnessy, "Developments and directions in speech recognition and understanding, part 1 [dsp education]", IEEE Signal Processing Magazine 26, 3, 75–80 (2009).

Bender, B. K., M. P. Cannito, T. Murry and G. E. Woodson, "Speech intelligibility in severe adductor spasmodic dysphonia", Journal of Speech, Language, and Hearing Research (2004).

Benesty, J., S. Makino and J. Chen, *Speech enhancement* (Springer Science & Business Media, 2005).

Benninger, M. S., G. Gardner and C. Grywalski, "Outcomes of Botulinum Toxin Treatment for Patients With Spasmodic Dysphonia", JAMA Otolaryngology Head & Neck Surgery 127, 9, 1083–1085, URL `https://doi.org/10.1001/archotol.127.9.1083` (2001).

Beukelman, D. and P. Mirenda, "Augmentative and alternative communication: Supporting children and adults with complex communication needs", Paul H. Brookes, Baltimore, MD (2005).

Bhat, C., B. Das, B. Vachhani and S. K. Kopparapu, "Dysarthric speech recognition using time-delay neural network based denoising autoencoder", in "Proc. Interspeech 2018", pp. 451–455 (2018), URL http://dx.doi.org/10.21437/Interspeech.2018-1754.

Bhattacharyya, N., "The prevalence of voice problems among adults in the united states", The Laryngoscope 124, 10, 2359–2362 (2014).

Biadsy, F., R. J. Weiss, P. J. Moreno, D. Kanvesky and Y. Jia, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation", in "Proc. Interspeech 2019", pp. 4115–4119 (2019), URL http://dx.doi.org/10.21437/Interspeech.2019-1789.

Boersma, P., "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", Proceedings of the Institute of Phonetic Sciences 17 (2000).

Borrie, S. A., M. J. McAuliffe and J. M. Liss, "Perceptual learning of dysarthric speech: A review of experimental studies", Journal of Speech, Language, and Hearing Research 55, 1, 290–305 (2012).

Chorowski, J., R. J. Weiss, S. Bengio and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders", CoRR abs/1901.08810, URL http://arxiv.org/abs/1901.08810 (2019).

Cohen, S. M., J. Kim, N. Roy, C. Asche and M. Courey, "Prevalence and causes of dysphonia in a large treatment-seeking population", The Laryngoscope 122, 2, 343–348, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/lary.22426 (2012).

Cooke, M., J. Barker, S. Cunningham and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition", The Journal of the Acoustical Society of America 120, 5, 2421–2424 (2006).

Cooper, L., S. Balandin and D. Trembath, "The loneliness experiences of young adults with cerebral palsy who use alternative and augmentative communication", Augmentative and Alternative Communication 25, 3, 154–164 (2009).

Copestake, A., "Augmented and alternative NLP techniques for augmentative and alternative communication", in "Natural Language Processing for Communication Aids", (1997), URL https://www.aclweb.org/anthology/W97-0506.

Cutajar, M., E. Gatt, I. Grech, O. Casha and J. Micallef, "Comparative study of automatic speech recognition techniques", Signal Processing, IET 7, 25–46 (2013).

de Krom, G., "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals", Journal of Speech, Language, and Hearing Research 36, 2, 254–266 (1993).

Deng, L., G. Hinton and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview", in "2013 IEEE International Conference on Acoustics, Speech and Signal Processing", pp. 8599–8603 (2013).

Desai, S., E. V. Raghavendra, B. Yegnanarayana, A. W. Black and K. Prahallad, "Voice conversion using artificial neural networks", in "2009 IEEE International Conference on Acoustics, Speech and Signal Processing", pp. 3893–3896 (IEEE, 2009).

Dibazar, A. A. and S. Narayanan, "A system for automatic detection of pathological speech", in "Conference Signals, Systems, and Computers, Asilomar, CA", (2002).

Dieleman, S. and B. Schrauwen, "End-to-end learning for music audio", in "2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", pp. 6964–6968 (2014).

Donahue, C., B. Li and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition", CoRR abs/1711.05747 (2017).

Elam, J. C., S. L. Ishman, K. B. Dunbar, J. O. Clarke and C. G. Gourin, "The relationship between depressive symptoms and voice handicap index scores in laryngopharyngeal reflux", The Laryngoscope 120, 9, 1900–1903, URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/lary.21012` (2010).

Elena Nerriere, F. G. V. K.-M., Marie-Noel Vercambre, "Voice disorders and mental health in teachers: a cross-sectional nationwide study", BMC Public Health 9, 370 (2009).

Enderby, P. M. P. M., *Frenchay dysarthria assessment* (San Diego, Calif. : College-Hill Press, 1983), includes index.

Evitts, P., H. Starmer, K. Teets, C. Montgomery, L. Calhoun, A. Schulze, J. MacKenzie and L. Adams, "The impact of dysphonic voices on healthy listeners: Listener reaction times, speech intelligibility, and listener comprehension", American journal of speech-language pathology 25, 1–15 (2016).

Falk, T. H., Qingfeng Xu and Wai-Yip Chan, "Non-intrusive gmm-based speech quality measurement", in "Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.", vol. 1, pp. I/125–I/128 Vol. 1 (2005).

Ferrand, C. T., "Harmonics-to-noise ratio: An index of vocal aging", Journal of Voice 16, 4, 480 – 487, URL `http://www.sciencedirect.com/science/article/pii/S0892199702001236` (2002).

Fraile, R. and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis", Biomedical Signal Processing and Control 14, 42 – 54, URL http://www.sciencedirect.com/science/article/pii/S1746809414000986 (2014).

Fu, S.-W., Y. Tsao, H.-T. Hwang and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm", in "INTERSPEECH", (2018).

Gabbay, A., A. Shamir and S. Peleg, "Visual speech enhancement using noise-invariant training", CoRR abs/1711.08789, URL http://arxiv.org/abs/1711.08789 (2017).

Garofolo, J., L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren and V. Zue, "Timit acoustic-phonetic continuous speech corpus", Linguistic Data Consortium (1992).

Godfrey, J. J., E. C. Holliman and J. McDaniel, "Switchboard: Telephone speech corpus for research and development", in "Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1", ICASSP'92, pp. 517–520 (IEEE Computer Society, Washington, DC, USA, 1992), URL http://dl.acm.org/citation.cfm?id=1895550.1895693.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets", in "Advances in neural information processing systems", pp. 2672–2680 (2014).

Graves, A., S. Fernández, F. Gomez and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks", in "Proceedings of the 23rd International Conference on Machine Learning", ICML '06, pp. 369–376 (ACM, New York, NY, USA, 2006).

Gray, P., "Non-intrusive speech-quality assessment using vocal-tract models", IEE Proceedings - Vision, Image and Signal Processing 147, 493–501(8) (2000).

Green, P. D., J. Carmichael, A. Hatzis, P. Enderby, M. S. Hawley and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers.", in "INTERSPEECH", (2003).

Hasegawa-Johnson, M., J. Gunderson, A. Perlman and T. Huang, "Hmm-based and svm-based recognition of the speech of talkers with spastic dysarthria", in "Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on", vol. 3, pp. III–III (IEEE, 2006).

Heman-Ackah, Y. D., D. D. Michael, M. M. Baroody, R. Ostrowski, J. Hillenbrand, R. J. Heuer, M. Horman and R. T. Sataloff, "Cepstral peak prominence: A more reliable measure of dysphonia", Annals of Otology, Rhinology & Laryngology 112, 4, 324–333, URL https://doi.org/10.1177/000348940311200406, pMID: 12731627 (2003).

Hillenbrand, J. and R. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech", Journal of speech and hearing research 39, 311–21 (1996).

Hinton, G., L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", IEEE Signal Processing Magazine 29, 6, 82–97 (2012).

Hirano, M. and K. R. McCormick, "Clinical examination of voice by minoru hirano", (1986).

Hou, J., S. Wang, Y. Lai, J. Lin, Y. Tsao, H. Chang and H. Wang, "Audio-visual speech enhancement based on multimodal deep convolutional neural network", CoRR abs/1703.10893, URL `http://arxiv.org/abs/1703.10893` (2017).

Hsu, C., H. Hwang, Y. Wu, Y. Tsao and H. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder", in "2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)", pp. 1–6 (2016).

Hsu, C., H. Hwang, Y. Wu, Y. Tsao and H. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks", CoRR abs/1704.00849, URL `http://arxiv.org/abs/1704.00849` (2017a).

Hsu, W.-N., Y. Zhang and J. Glass, "Learning latent representations for speech generation and transformation", pp. 1273–1277 (2017b).

Isola, P., J. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks", CoRR abs/1611.07004, URL `http://arxiv.org/abs/1611.07004` (2016).

Jacewicz, E., R. A. Fox and J. Salmons, "Vowel space areas across dialects and gender", in "16th International Congress of Phonetic Sciences, Saarbrücken, Germany", (2007).

Jaitly, N. and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines", in "2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", pp. 5884–5887 (2011).

Kameoka, H., T. Kaneko, K. Tanaka and N. Hojo, "Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks", 2018 IEEE Spoken Language Technology Workshop (SLT) pp. 266–273 (2018).

Kaneko, T., H. Kameoka, K. Tanaka and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion", in "ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", pp. 6820–6824 (2019).

Kempster, G. B., B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol", American Journal of Speech-Language Pathology 18, 2, 124–132 (2009).

Kent, R. D., G. Weismer, J. F. Kent and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria", Journal of Speech and Hearing Disorders 54, 4, 482–499 (1989).

Kim, G. and P. C. Loizou, "Why do speech-enhancement algorithms not improve speech intelligibility?", in "2010 IEEE International Conference on Acoustics, Speech and Signal Processing", pp. 4738–4741 (2010).

Kim, H., M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin and S. Frame, "Dysarthric speech database for universal access research.", in "Interspeech", vol. 2008, pp. 1741–1744 (2008).

Kreiman, J., B. Gerratt and B. Gabelman, "Jitter, shimmer, and noise in pathological voice quality perception", The Journal of the Acoustical Society of America 112, 2446 (2002).

Lee, H., C. Ekanadham and A. Y. Ng, "Sparse deep belief net model for visual area v2", in "Advances in Neural Information Processing Systems 20", edited by J. C. Platt, D. Koller, Y. Singer and S. T. Roweis, pp. 873–880 (Curran Associates, Inc., 2008).

Lee, L., J. C. Stemple, L. Glaze and L. N. Kelchner, "Quick screen for voice and supplementary documents for identifying pediatric voice disorders", Language, Speech, and Hearing Services in Schools 35, 4, 308–319 (2004).

Lee, S., L. L. Mendel and G. M. Bidelman, "Predicting speech recognition using the speech intelligibility index and other variables for cochlear implant users", Journal of Speech, Language, and Hearing Research 62, 5, 1517–1531, URL https://jshd.pubs.asha.org/doi/abs/10.1044/2018_JSLHR-H-18-0303(2019).

Lu, F.-L. and S. Matteson, "Speech tasks and interrater reliability in perceptual voice evaluation", Journal of voice : official journal of the Voice Foundation 28 (2014).

Ludlow, C. L., "Treatment of Speech and Voice Disorders With Botulinum Toxin", JAMA 264, 20, 2671–2675, URL https://doi.org/10.1001/jama.1990.03450200079035 (1990).

Ludlow, C. L., R. F. Naunton, S. Terada and B. J. Anderson, "Successful treatment of selected cases of abductor spasmodic dysphonia using botulinum toxin injection", Otolaryngology Head and Neck Surgery 104, 6, 849–855 (1991).

McGraw, I., C.-y. Lee, I. Hetherington, S. Seneff and J. Glass, "Collecting voices from the cloud", (2010).

McGurk, H. and J. MacDonald, "Hearing lips and seeing voices", Nature 264, 5588, 746 (1976).

Menendez-Pidal, X., J. B. Polikoff, S. M. Peters, J. E. Leonzio and H. T. Bunnell, "The nemours database of dysarthric speech", in "Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96", vol. 3, pp. 1962–1965 vol.3 (1996).

Merrill, R. M., A. E. Anderson and A. Sloan, "Quality of life indicators according to voice disorders and voice-related conditions", The Laryngoscope 121, 9, 2004–2010, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/lary.21895 (2011).

Mirza, M. and S. Osindero, "Conditional generative adversarial nets", CoRR abs/1411.1784, URL http://arxiv.org/abs/1411.1784 (2014).

Moore, M., M. Saxon, H. Venkateswara, V. Berisha and S. Panchanathan, "Say What? A Dataset for Exploring the Error Patterns That Two ASR Engines Make", in "Proc. Interspeech 2019", pp. 2528–2532 (2019), URL http://dx.doi.org/10.21437/Interspeech.2019-3096.

Moore, M., H. Venkateswara and S. Panchanathan, "Whistle-blowing asrs: evaluating the need for more inclusive automatic speech recognition systems", in "Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH", vol. 2018, pp. 466–470 (2018).

Morris, A. C., V. Maier and P. D. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition", in "INTERSPEECH", (2004).

Morrow, S. L. and N. P. Connor, "Voice amplification as a means of reducing vocal load for elementary music teachers", Journal of Voice 25, 4, 441 – 446 (2011).

Mroueh, Y., E. Marcheret and V. Goel, "Deep multimodal learning for audio-visual speech recognition", in "2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", pp. 2130–2134 (IEEE, 2015).

Narayanan, P., P. Chakravarty, F. Charette and G. Puskorius, "Hierarchical sequence to sequence voice conversion with limited data", (2019).

Narendranath, M., H. A. Murthy, S. Rajendran and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks", Speech Communication 16, 2, 207 – 216, URL http://www.sciencedirect.com/science/article/pii/016763939400058I, voice Conversion: State of the Art and Perspectives (1995).

Orozco-Arroyave, J. R., J. C. Vdsquez-Correa, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz and E. Noth, "Towards an automatic monitoring of the neurological state of parkinson's patients from speech", in "2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", pp. 6490–6494 (2016).

Panchanathan, S., S. Chakraborty and T. McDaniel, "Social interaction assistant: A person-centered approach to enrich social interactions for individuals with visual impairments", IEEE Journal of Selected Topics in Signal Processing 10, 5, 942–951 (2016).

Panchanathan, S., S. Chakraborty and T. McDaniel, "Social interaction assistant: A person-centered approach to enrich social interactions for individuals with visual impairments", IEEE Journal of Selected Topics in Signal Processing 10, 5, 942–951 (2016).

Panchanathan, S., T. McDaniel and V. Balasubramanian, "Person-centered accessible technologies: Improved usability and adaptation through inspirations from disability research", in "Proceedings of the 2012 ACM Workshop on User Experience in e-Learning and Augmented Technologies in Education", UXeLATE '12, pp. 1–6 (ACM, New York, NY, USA, 2012), URL http://doi.acm.org/10.1145/2390895.2390897.

Pang, Z. and F. Zhu, "Noise-robust ASR for the third 'chime' challenge exploiting time-frequency masking based multi-channel speech enhancement and recurrent neural network", CoRR abs/1509.07211 (2015).

Papandreou, G., A. Katsamanis, V. Pitsikalis and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition", IEEE Transactions on Audio, Speech, and Language Processing 17, 3, 423–435 (2009).

Parent, G. and M. Eskenazi, "Speaking to the crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges.", pp. 3037–3040 (2011).

Pascual, S., A. Bonafonte and J. Serrà, "SEGAN: speech enhancement generative adversarial network", CoRR abs/1703.09452, URL http://arxiv.org/abs/1703.09452 (2017).

Pascual, S., A. Bonafonte, J. Serrà and J. Gonzalez, "Whispered-to-voiced alaryngeal speech conversion with generative adversarial networks", (2018).

Patel, A. B., S. F. Bansberg, C. H. Adler, D. G. Lott and L. Crujido, "The mayo clinic arizona spasmodic dysphonia experience: A demographic analysis of 718 patients", Annals of Otology, Rhinology & Laryngology 124, 11, 859–863, pMID: 26024910 (2015).

Polur, P. D. and G. E. Miller, "Investigation of an hmm/ann hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals", Medical Engineering & Physics 28, 8, 741 – 748 (2006).

Portnoy, R. A. and A. E. Aronson, "Diadochokinetic syllable rate and regularity in normal and in spastic and ataxic dysarthric subjects", Journal of Speech and Hearing Disorders 47, 3, 324–328, URL https://pubs.asha.org/doi/abs/10.1044/jshd.4703.324 (1982).

Purwins, H., B. Li, T. Virtanen, J. Schlüter, S. Chang and T. N. Sainath, "Deep learning for audio signal processing", CoRR abs/1905.00078, URL http://arxiv.org/abs/1905.00078 (2019).

Rix, A. W., J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs", in "2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)", vol. 2, pp. 749–752 (IEEE, 2001).

Rosen, C. A., A. S. Lee, J. Osborne, T. Zullo and T. Murry, "Development and validation of the voice handicap index-10", The Laryngoscope 114, 9, 1549–1556 (2004).

Rosen, K. and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech", Augmentative and Alternative Communication 16, 1, 48–60 (2000).

Rosenblum, L. D., "Speech perception as a multimodal phenomenon", Current Directions in Psychological Science 17, 6, 405–409 (2008).

Roy, N., R. M. Merrill, S. D. Gray and E. M. Smith, "Voice disorders in the general population: Prevalence, risk factors, and occupational impact", The Laryngoscope 115, 11, 1988–1995 (2005).

Roy, N., R. M. Merrill, S. Thibeault, R. A. Parsa, S. D. Gray and E. M. Smith, "Prevalence of voice disorders in teachers and the general population", Journal of Speech, Language, and Hearing Research 47, 2, 281–293 (2004).

Roy, N., B. Weinrich, S. D. Gray, K. Tanner, J. C. Stemple and C. M. Sapienza, "Three treatments for teachers with voice disorders: a randomized clinical trial.", Journal of speech, language, and hearing research : JSLHR 46 3, 670–88 (2003).

Rudzicz, F., "Adjusting dysarthric speech signals to be more intelligible", Computer Speech and Language 27, 6, 1163 – 1177, URL http://www.sciencedirect.com/science/article/pii/S0885230812001003, special Issue on Speech and Language Processing for Assistive Technology (2013).

Rudzicz, F., A. K. Namasivayam and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria", Language Resources and Evaluation 46, 4, 523–541 (2012).

Sadeghi, M., S. Leglaive, X. Alameda-Pineda, L. Girin and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoder", ArXiv abs/1908.02590 (2019).

Samlan, R. A., B. H. Story and K. Bunton, "Relation of perceived breathiness to laryngeal kinematics and acoustic measures based on computational modeling", Journal of Speech, Language, and Hearing Research (2013).

Sandoval, S., V. Berisha, R. L. Utianski, J. M. Liss and A. Spanias, "Automatic assessment of vowel space area", The Journal of the Acoustical Society of America 134, 5, EL477–EL483 (2013).

Santos, J. F., M. Senoussaoui and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech", in "2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)", pp. 55–59 (2014).

Schiffner, F., J. Skowronek and A. Raake, "On the impact of speech intelligibility on speech quality in the context of voice over ip telephony", in "2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)", pp. 59–60 (2014).

Selva Nidhyananthan, S., R. Shantha Selva kumari and V. Shenbagalakshmi, "Assessment of dysarthric speech using elman back propagation network (recurrent network) for speech recognition", International Journal of Speech Technology 19, 3, 577–583 (2016).

Serrà, J., S. Pascual and C. Segura, "Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion", (2019).

Sharma, H. V. and M. Hasegawa-Johnson, "State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition", in "Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies", pp. 72–79 (Association for Computational Linguistics, Los Angeles, California, 2010), URL https://www.aclweb.org/anthology/W10-1310.

Speyer, R., "Effects of voice therapy: A systematic review", Journal of Voice 22, 5, 565 – 580, URL http://www.sciencedirect.com/science/article/pii/S0892199706001378 (2008).

Streijl, R. C., S. Winkler and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives", Multimedia Systems 22, 2, 213–227 (2016).

Taal, C. H., R. C. Hendriks, R. Heusdens and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech", in "2010 IEEE International Conference on Acoustics, Speech and Signal Processing", pp. 4214–4217 (2010).

Tadayon, R., T. McDaniel and S. Panchanathan, "Autonomous training assistant: A system and framework for guided at-home motor learning", pp. 293–294 (2016).

Tanaka, K., H. Kameoka, T. Kaneko and N. Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms", ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 6805–6809 (2018).

Teixeira, J. P. and P. O. Fernandes, "Acoustic analysis of vocal dysphonia", Procedia Computer Science 64, 466 – 473, URL http://www.sciencedirect.com/science/article/pii/S1877050915026794, conference on ENTERprise Information Systems/International Conference on Project MANagement/Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2015 October 7-9, 2015 (2015).

Thibeault, S. L., R. M. Merrill, N. Roy, S. D. Gray and E. M. Smith, "Occupational risk factors associated with voice disorders among teachers", Annals of Epidemiology 14, 10, 786 – 792, URL http://www.sciencedirect.com/science/article/pii/S1047279704000602 (2004).

Tiippana, K., T. Andersen and M. Sams, "Visual attention modulates audiovisual speech perception", European Journal of Cognitive Psychology 16, 3, 457–472 (2004).

Tiwari, M. and M. Tiwari, "Voice - How humans communicate?", Journal of Natural Science, Biology and Medicine 3, 1, 3–11 (2012).

Trnka, K., J. McCaw, D. Yarrington, K. Mccoy and C. Pennington, "User interaction with word prediction: The effects of prediction quality.", TACCESS 1 (2009).

Trnka, K., J. McCaw, D. Yarrington, K. F. McCoy and C. Pennington, "Word prediction and communication rate in aac", in "Proceedings of the IASTED International Conference on Telehealth/Assistive Technologies", Telehealth/AT '08, pp. 19–24 (ACTA Press, Anaheim, CA, USA, 2008), URL http://dl.acm.org/citation.cfm?id=1722763.1722768.

Veaux, C., J. Yamagishi and K. Macdonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit", (2017).

Walshe, M., N. Miller, M. Leahy and A. Murray, "Intelligibility of dysarthric speech: perceptions of speakers and listeners", International Journal of Language & Communication Disorders 43, 6, 633–648 (????).

Wang, W., D. Cosker, Y. Hicks, S. Saneit and J. Chambers, "Video assisted speech source separation", in "Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.", vol. 5, pp. v–425 (IEEE, 2005).

Wang, X., C. Wu, P. Zhang, Z. Wang, Y. Liu, X. Li, Q. Fu and Y. Yan, "Noise robust IOA/CAS speech separation and recognition system for the third 'chime' challenge", CoRR abs/1509.06103 (2015).

Wang, Y., R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark and R. Saurous, "Tacotron: Towards end-to-end speech synthesis", pp. 4006–4010 (2017).

Wang, Y.-Q., "An analysis of the viola-jones face detection algorithm", Image Processing On Line 4, 128–148 (2014).

Weinberger, S., "Speech accent archive. George Mason University.", (2013).

Weiss, R. J., J. Chorowski, N. Jaitly, Y. Wu and Z. Chen, "Sequence-to-sequence models can directly transcribe foreign speech", CoRR abs/1703.08581, URL http://arxiv.org/abs/1703.08581 (2017).

Willinger, U., S. Valkl-Kernstock and H. N. Aschauer, "Marked depression and anxiety in patients with functional dysphonia", Psychiatry Research 134, 1, 85 – 91, URL http://www.sciencedirect.com/science/article/pii/S016517810500020X (2005).

Xu, X., R. Flynn and M. Russell, "Speech intelligibility and quality: A comparative study of speech enhancement algorithms", in "2017 28th Irish Signals and Systems Conference (ISSC)", pp. 1–6 (2017).

Young, V. and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review", Assistive Technology 22, 2, 99–112 (2010).

Yuan, J., M. Liberman and C. Cieri, "Interspeech 2006 towards an integrated understanding of speaking rate in conversation", (2006).

Yumoto, E., W. J. Gould and T. Baer, "Harmonicsâtoânoise ratio as an index of the degree of hoarseness", The Journal of the Acoustical Society of America 71, 6, 1544–1550, URL https://doi.org/10.1121/1.387808 (1982).

Zeiler, M. D. and R. Fergus, "Visualizing and understanding convolutional networks", CoRR abs/1311.2901, URL http://arxiv.org/abs/1311.2901 (2013).

Zekveld, A., S. Kramer, M. Vlaming and T. Houtgast, "Audiovisual perception of speech in noise and masked written text", Ear and hearing 29, 99–111 (2008).

Zhang, J.-X., Z.-H. Ling and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations", arXiv preprint arXiv:1906.10508 (2019a).

Zhang, J.-X., Z.-H. Ling, L.-J. Liu, Y. Jiang and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion", IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 27, 3, 631–644 (2019b).

Zwirner, P., T. Murry, M. Swenson and G. E. Woodson, "Acoustic changes in spasmodic dysphonia after botulinum toxin injection", Journal of Voice 5, 1, 78 – 84, URL http://www.sciencedirect.com/science/article/pii/S0892199705801675 (1991).