Designing Voice-Assistive Technologies: Enhancing the Quality and Intelligibility of Pathological Speech

Meredith Moore

Arizona State University

mkmoore7@asu.edu

Abstract

The motivation behind my dissertation can be summed up in four words: help people be understood. I have focused on working with individuals with voice disorders to understand the challenges they face on a day-to-day basis and develop new assistive technologies to help them be better understood. I have assessed the needs of individuals with voice disorders both quantitatively and qualitatively and have developed a set of design considerations that should be followed in order for a voice-assistive technology to be as positively impactful as possible. I am currently in the process of building a dataset in a distributed manner in the hopes of publishing a large, publicly available speech disorder dataset for speech researchers to work with. Through this dataset, I expect there to be better representation of different voices in tomorrow's speech-based technology. Using this dataset, I am working on building a system that improves the quality and the intelligibility of pathological speech such that individuals with speech disorders can be better understood by both humans and machines in real-time.

Index Terms: speech enhancement, human-computer interaction, voice disorders, assistive technology

1. Introduction

In the United States, 9.4 million adults have trouble using their voices [1]. Speech is a complicated process with many potential breakpoints. A voice disorder occurs when voice quality, pitch, and loudness differ or are inappropriate for an individual's age, gender, cultural background, or geographic location [2, 3].

The majority of the work presented in this paper has been collected from a sample of participants who have Spasmodic Dysphonia (SD). Also known as laryngeal dystonia, SD is a voice disorder characterized by improper functioning of the muscles that generate a person's voice [4]. These muscles spasm, in what is referred to as a laryngospasm, which makes it difficult to speak or breath.

1.1. Voice-Assistive Technologies

Much of the literature dealing with pathological speech deals with the efficacy of speech therapies or building robust speech recognition systems that will recognize pathological speech. Very few voice-assistive technologies exist. The main voice-assistive technology that is available to individuals with voice disorders is a voice amplification system that focuses on removing stress from the voice. Research suggests that voice amplification may be an effective intervention to decrease vocal cord damage due to overuse [**?**, **?**, **?**]. 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.

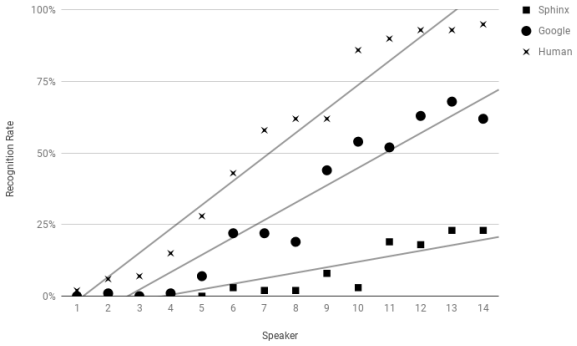

Figure 1: *A comparison of the recognition rate of the three different models of intelligibility: Human, Sphinx, and Google. Human recognition rates are denoted with the cross, Google with the circle, and Sphinx with the square.*

2. 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 wellbeing as characterized by isolation, frustration, stress, anxiety, and depression. The motivation behind my dissertation is to build a system that helps people with pathological speech be better understood. Rather than trying to build systems that recognize pathological voices, I am focusing on making the quality and intelligibility of the individual's voice better so that both humans and machines are better able to understand the speaker.

3. Needs Assessments

We have assessed the needs of individuals with voice disorders in two separate studies. The first is an analysis of the state-of-the-art off-the-shelf automatic speech recognition systems and their performance on pathological speech, and the second is an in-depth survey of the experience of individuals with voice disorders. Current off-the-shelf speech recognition packages do not recognize pathological speech as well as they recognize 'normal' speech. In [**?**], I tested the efficacy of two off-the-shelf speech recognition systems on both control and pathological speech (using the dysarthric speech datasets UASPEECH [5], and TORGO [6]), and the control speech was recognized 59% more often than pathological speech. In Figure 1, the recognition rates of the two different systems that were tested are shown against the human recognition rate for each speaker. It is clear that the available off-the-shelf speech recognition systems do not sufficiently recognize pathological speech.

Table 1: *This table demonstrates what participants reported as the primary effects of living with a voice disorder. The open-ended responses were coded into several different categories, the most prominent categories are shown below.*

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%

We conducted a need-finding survey to learn about the experience of individuals with voice disorders. In this survey, 458 participants responded to both open and close-ended questions relating to their experience with a voice disorder. The survey was conducted primarily on individuals with Spasmodic Dysphonia, but all individuals with voice disorders were welcome to participate. The primary effects of living with a voice disorder as reported by the survey participants are decreased social interaction, decreased emotional well-being, and a negative impact on the individual's career, the response rates for these categories are shown in Table 1. Respondents also reported significant difficulty talking on the phone, as well as a general decrease in self-confidence.

4. Design Considerations

In the same need-finding survey, the participants were asked to describe technologies that they would like to be developed. While many of the respondents did not have a specific technology in mind, they offered general design principles that should be followed when developing voice-assistive technologies. The reported design principles include the need for a voice-assistive technology to be unobtrusive, affordable, and for it to help them be better understood. As the average age of individuals with SD is 62 years old, and many of the respondents reported having minimal technical capabilities, the technology needs to be very user-friendly.

5. Dataset Collection

We are currently working on creating a database that will be collected in a distributed manner via a web application. Most existing pathological speech databases consist of very few speakers and do not have very many hours of speech. We believe that part of this is due to the difficulty in getting people to come into the lab in person to collect speech samples. To lower this barrier, we are deploying our speech collection application via the web. We are hoping to collect data from 200 individuals with voice disorders. This dataset will be made publicly available for researchers to work with. Making a large body of pathological speech samples will hopefully improve the representation of different voices in state-of-the-art speech systems.

6. Generative-Audio Systems

Using the discussed dataset, we are planning to build a system that generates speech that is more intelligible than the input speech. We plan to test several different machine learning models to accomplish this task including adversarial learning, as shown in Figure 2, and reinforcement learning, as shown in Figure 3. In the adversarial learning paradigm, pathological speech and control speech are the inputs, and the goal of the intelligible speech generator is to fool the discriminator into thinking that it is control speech rather than pathological speech. This system follows the paradigm of a Generative Adversarial Network as proposed in [**?**]. In the reinforcement learning system, a speech enhancement network is trained using the output of an automatic speech recognition (ASR) system. The ASR system takes the output of the speech enhancement network and returns a hypothesis of what it said. The difference between the hypothesis and the ground truth label is then calculated as a word error rate, and this the speech enhancement network is trained to minimize this difference.

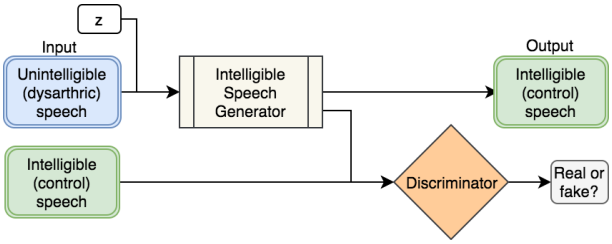

Figure 2: *An outline of a machine learning paradigm that will generate intelligible speech from pathological speech using the Generative Adversarial Network framework.*

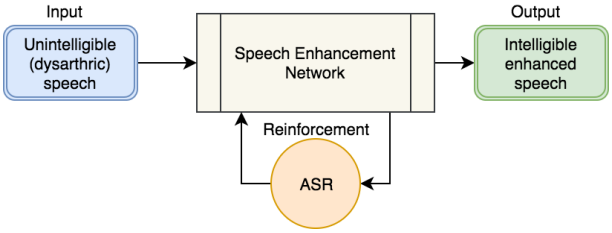

Figure 3: *An outline of a machine learning system that will generate intelligible speech from pathological speech using the output of an automatic speech recognition (ASR) system as the reward signal.*

7. Contributions

I have focused my dissertation on helping people with voice disorders be understood. I have made an effort to understand and clearly communicate the needs of individuals with voice disorders, and designed a framework of design considerations to guide the development of voice-assistive technologies. I am working on collecting a large dataset of pathological speech to help improve the representation of different voices in state-of-the-art systems, and to fuel the research on pathological speech, as well as developing a voice-assistive technology, based on novel machine learning paradigms, that improves the quality and intelligibility of an individual's voice to help them be better understood.

8. Acknowledgements

I would like to thank the National Spasmodic Dysphonia Association for their support as well as my advisors Dr. Sethuraman Panchanathan, Dr. Troy McDaniel, and Dr. Hemanth Venkateswara.

9. References

[1] N. Bhattacharyya, "The prevalence of voice problems among adults in the united states," *The Laryngoscope*, vol. 124, no. 10, pp. 2359–2362, 2014.

[2] L. Lee, 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*, vol. 35, no. 4, pp. 308–319, 2004.

[3] A. Aronson and D. Bless, *Clinical Voice Disorders*, ser. Thieme Publishers Series. Thieme, 2009. [Online]. Available: https://books.google.com/books?id=wOhkGWBzG2UC

[4] M. J. Aminoff, H. H. Dedo, and K. Izdebski, "Clinical aspects of spasmodic dysphonia." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 41, no. 4, pp. 361–365, 1978. [Online]. Available: http://jnnp.bmj.com/content/41/4/361

[5] H. Kim, 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, 2008, pp. 1741–1744.

[6] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, Dec 2012.