INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

UncommonVoice: A Crowdsourced Dataset of Dysphonic Speech

Meredith Moore[1*†], *Piyush Papreja*[2*†], *Michael Saxon*[3†], *Visar Berisha*[2], *Sethuraman Panchanathan*[2]

[1]Drake University, United States
[2]Arizona State University, United States
[3]University of California, Santa Barbara, United States

meredith.moore@drake.edu, {ppapreja, mssaxon, visar, panch}@asu.edu

Abstract

To facilitate more accessible spoken language technologies and advance the study of dysphonic speech this paper presents UncommonVoice, a freely-available, crowd-sourced speech corpus consisting of 8.5 hours of speech from 57 individuals, 48 of whom have spasmodic dysphonia. The speech material consists of non-words (prolonged vowels, and the prompt for diadochokinetic rate), sentences (randomly selected from TIMIT prompts and the CAPE-V intelligibility analysis), and spontaneous image descriptions. The data was recorded in a crowdsourced manner using a web-based application. This dataset is a fundamental resource for the development of voice-assistive technologies for individuals with dysphonia as well as the enhancement of the accessibility of voice-based technologies (automatic speech recognition, virtual assistants, etc). Research on articulation differences as well as how best to model and represent dysphonic speech will greatly benefit from a free and publicly available dataset of dysphonic speech. The dataset will be made freely and publicly available at www.uncommonvoice.org. In the following sections, we detail the data collection process as well as provide an initial analysis of the speech corpus.

Index Terms: voice disorder, spasmodic dysphonia, dataset human-computer interaction,

1. Introduction

There is much interest in mitigating "algorithmic unfairness" in machine learning-based production systems across a variety of domains [1]. In spoken dialogue systems, a dearth of disordered speech training data drives their inaccessibility to users exhibiting said disorders. In light of this problem, we present the UncommonVoice dataset to both better represent individuals with voice disorders in current voice-based technologies and to enable the development of future voice-assistive technologies. UncommonVoice was inspired by the work done at Mozilla on Common Voice [2], a large, freely-available, crowd-sourced 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. 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 and does not provide insight into how individuals with voice disorders speak.

[*]Denotes equal contribution
[†]Work done at Arizona State University

1.1. Accessibility of Voice-Based Systems

There exists a body of previous work demonstrating the lack of accessibility of voice-based technologies for individuals with voice disorders [3, 4, 5]. 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 [6, 7]. Voice Assistants (VAs) are becoming increasingly popular as a form of human-computer interaction leading to voice-based control of many systems around the house. It has been shown, however, that these systems recognize speech from individuals with voice disorders significantly less often than speech from individuals without voice disorders. This difference indicates that there is a barrier to access of VAs for individuals with voice disorders.

1.2. Spasmodic Dysphonia

One particular group for which improving the accessibility of voice-based systems could be beneficial to is individuals with 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 [8]. These muscles spasm, in what is referred to as a laryngospasm, which makes it difficult to speak or breathe. Depending on which muscles are affected these spasms can lead to either breathy and/or creaky speech.

Botulinum toxin A, (known commercially as Botox and referred to clinically as BTX) therapy has proven to be an effective treatment for SD [9], however, BTX often causes an individual's voice to change in a cyclic pattern. Over a period of several weeks, an individual's voice will go from creaky before injection, to breathy after an injection. Once the BTX injection starts to wear off, the individual's voice will sound 'modal' (normal) for a little while before becoming creaky again [10].

In a series of surveys, individuals with SD have described their difficulties using voice-based technologies such as smart speakers or speech to text as these systems rarely understand their speech. Voice disorders significantly affect an individual's social life, emotional wellbeing, and career.

Compared to neurologically dysarthric speech, SD is understudied. There are no large-scale publicly available datasets of speech from SD patients. The SD datasets that do exist are difficult to access as they often include sensitive health data.

1.3. Motivation and Contribution

Towards the goal of improving the representation of individuals with voice disorders in the vast corpora of speech, we present UncommonVoice, a crowdsourced, publicly available dataset of

http://dx.doi.org/10.21437/Interspeech.2020-3093

Table 1: *Pre-Collection Survey Questions*

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 for your voice?	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

speech from individuals with voice disorders. While datasets like TORGO [11] and UASPEECH [12] focus on freely providing speech data from individuals with dysarthria, UncommonVoice focuses on providing data from individuals with dysphonia.

We believe that UncommonVoice posits a significant contribution to the field and will enable advancement in improving the accessibility of voice-based technologies as well as the development of voice-assistive technologies.

2. Uncommonvoice Collection Process

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

2.1. Data Collection System

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.

2.2. 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 1. Only participants exhibiting Spasmodic Dysphonia were asked the last six questions. In the final two, participants 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 a scale from 'Very difficult' to 'Effortless'.

2.3. Data Collection Tasks

The UncommonVoice data collection process consists of 4 tasks. The design decision to keep the order of the tasks the same between users, but to randomize 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 non-word content so that the data exists to measure any change in vocal quality throughout the data collection process.

2.3.1. 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 [13].

2.3.2. Task 2: Read Sentences

In the second task, we asked users to read sentences that were randomly selected from TIMIT [14]. We asked the user to read 84 different sentences from the TIMIT dataset. These sentences were randomly presented to avoid any ordering effect. To calculate a speaker's CAPE-V as in [15], speakers were also asked to read the sentences involved in the calculation of the CAPE-V score.

2.3.3. 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 spontaneous speech that would have a more natural cadence than read speech. The images were chosen from the Microsoft Common Objects in Context (MSCOCO) [16].

2.3.4. 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 be able to measure any change in vocal quality over the duration of the tasks.

2.4. Participant Recruitment

Both dysphonic and control speakers were recruited primarily through email list solicitation. The National Spasmodic Dysphonia Association (NSDA) shared the data collection link with their network of individuals exhibiting SD, and healthy control speakers were recruited through university mailing lists. Data collection began in February 2020 and will remain active going forward to facilitate its further growth.

3. UncommonVoice Results

We present statistics of the 1.0 release of the UncommonVoice dataset, consisting of all recordings collected as of May 2020.

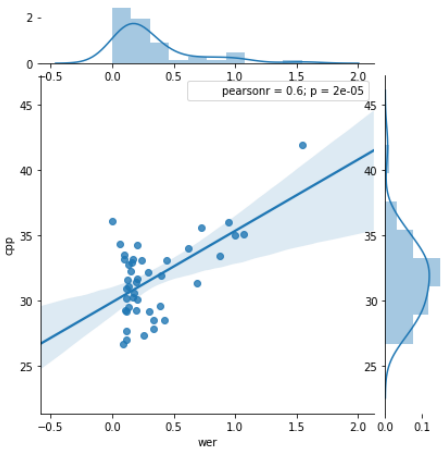

Figure 1: *Correlation Between Average WER Per Speaker and Average CPP.*

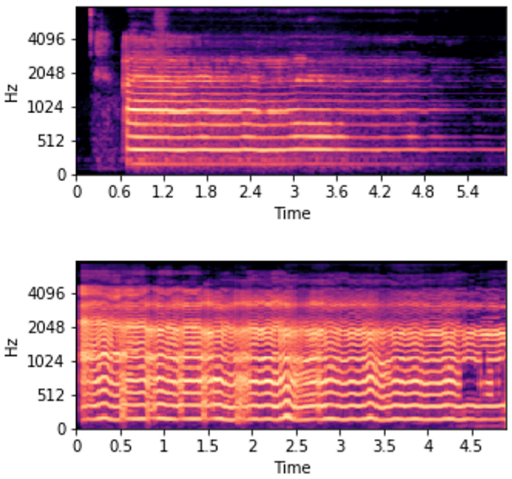

Figure 2: *Mel spectrograms of the Vowel /ae/ for Control (top), and Dysphonic (bottom).*

Additionally, we provide acoustic and ASR model-based analyses for validation and motivation for future work.

3.1. UncommonVoice Demographics

Currently, UncommonVoice consists of 4,683 speech recordings from 57 individuals–approximately 8.5 hours of data. Of those individuals, 44 (77%) of the individuals who recorded speech are female, while the other 13 (23%) are male. Of the individuals who contributed speech samples, 48 (84%) of them have a voice disorder, while the other 9 (16%) do not. Of the individuals who have a voice disorder, 18 (37.5%) of the individuals who provided speech samples regularly receive BTX injections as a treatment for their voice disorder, while the other 30 individuals with voice disorders (62.5%) 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, 49 (86%) indicated that they are native English speakers while the other 8 (14%) were not.

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 2.44 $\pm$ 1.13. 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 2.34 $\pm$ 1.10.

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). In response to this question, 43% of the participants with voice disorders answered 'tight/creaky', 31% 'combination', 10% responded as 'breathy', and 10% responded 'modal/normal'.

3.2. Studying the Acoustics of SD

As this is the first large-scale publicly available dataset of SD speech, there are many ways that this dataset can be used to demonstrate properties of SD speech. For example, Cepstral Peak Prominence (CPP) has been shown to be a reliable measure of dysphonia, more than the traditional acoustic metrics of jitter, shimmer, and the fundamental frequency [17].In [18], signal periodicity is shown to be highly correlated with the breathiness quality of speech. Calculating these acoustic properties and using them to predict dysphonia is one way that this dataset could be utilized. The dataset was collected in such a way that the Vowel Space Area for each speaker can be calculated at the beginning of the recording process and the end of the recording process [19, 20]. The change in Vowel Space Area over the duration of the data collection could provide insight into how SD voices are affected by heavy voice usage. In Figure 2, the difference between control and dysphonic speech when producing /ae/ is shown. The waves that are evident in the bottom mel-spectrogram are indicative of the 'choppier' glottal pulse, and lack of control that characterizes dysphonia.

3.3. Evaluating Intelligibility

3.3.1. ASR Performance

Given previous work on ASR system accessibility discussed in Section 1.1, we expected the dysphonic speech transcriptions to have a higher word error rate (WER) than the control speech. 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 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.

3.3.2. Acoustic Features and Intelligibility

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

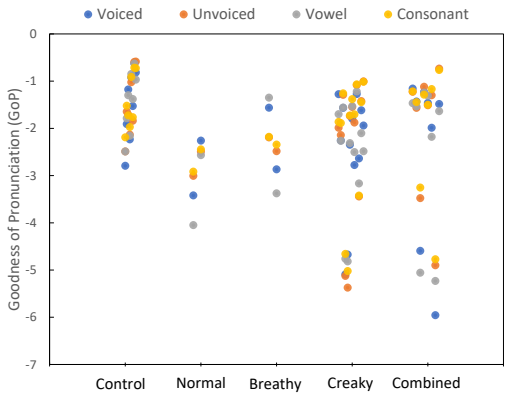

Figure 3: *Goodness of Pronunciation (GoP) grouped by the self-reported speech quality and phoneme type.*

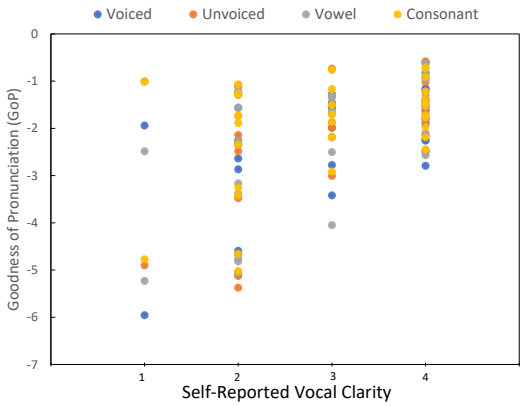

Figure 4: *Goodness of Pronunciation (GoP) for each level of self-reported vocal clarity (4 point opinion scale where 1 is not clear at all and 4 is very clear).*

Table 2: *Analysis of the intelligibility of control and dysphonic speech in UncommonVoice where Correct, S, I, D are the number of Correct, Substitutions, Insertions, and Deletions respectively, and WER is the Word Error Rate.*

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

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 [3]. The Pearson Correlation Coefficient between the CPP and WER is 0.75 and is shown in Figure 1.

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 [17, 21]. The Pearson Correlation Coefficient between the CPP and WER is 0.6.

The other features that were evaluated–jitter, shimmer, Harmonic Noise Ratio (HNR), and the fundamental frequency (f0)–all showed relatively low correlation with the WER for a given utterance.

3.3.3. Goodness of Pronunciation

We extract Goodness of Pronunciation (GoP) features [22] from each Task 2 sentence. Using a Kaldi GMM-HMM ASR model [23] each utterance is force aligned to its corresponding ground truth datasets. Each phoneme is then assessed with a log-likelihood ratio comparing the ASR model-assessed likelihood of the most probable phoneme in the language to the true phoneme as indicated by the transcript. A lower GoP score for a given phoneme corresponds to a less well-realized phoneme that "sounds" more like a different phoneme to the ASR model. As in [24] the GoP scores averaged by-phoneme across all utterances to generate a vector of 40 phoneme GoP scores for each speaker.

We average the scores of all voiced phonemes, unvoiced phonemes, vowels, and consonants to generate four composite GoP scores for each speaker, and plot the distributions of these composite scores for the control group and the four spasmodic dysphonia self-reported quality groups: modal (Normal), breathy, creaky, and combined. Figure 3 shows these distributions. Of them, the combined group has the most bimodal distribution, with a cluster of low GoP scores and a cluster of high GoP scores. The SD speakers, across all vocal quality types, have a higher cluster variance and lower mean GoP than the healthy control group, as expected.

Figure 4 depicts the same four composite GoP scores plotted against the speaker's self-reported vocal clarity rating, on a four point opinion scale. This figure lends credibility to the accuracy of these self-reported ratings, as mean and minimum composite GoP strictly increases with increasing self-reported clarity as composite GoP standard deviation decreases. This means that collectively, the 4-clarity speakers more consistently achieve better GoP than the lower self-rated clarity speakers, and so on.

4. Conclusion

We have described a database of dysphonic speech produced by 48 individuals with dysphonia. We currently have 9 control speakers, however, these control speakers are not age and gender-matched. We will strive to collect age and gender-matched control speakers to compare the two populations. We continue to collect speech samples from both individuals with and without voice disorders. For details on how to access the UncommonVoice dataset, visit `www.uncommonvoice.org`.

We believe this speech database is an impactful resource for the development of voice assistive technologies for individuals with voice disorders, as well as for improving the accessibility of state-of-the-art voice-based technologies. Analysis of this database will offer a deeper understanding of how to robustly model dysphonic speech.

5. Acknowledgements

The authors would like to acknowledge the National Spasmodic Dysphonia Association for their support throughout the development of UncommonVoice, particularly for their effort in recruiting speakers for UncommonVoice. Also a special thank you to the National Science Foundation Graduate Research Fellowship.

6. References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[2] "Mozilla Common Voice," 2017. [Online]. Available: https://voice.mozilla.org/

[3] M. Moore, 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*, 2019, pp. 2528–2532. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3096

[4] M. Moore, H. Demakethepalli Venkateswara, and S. Panchanathan, "Whistle-blowing ASRs: Evaluating the need for more inclusive automatic speech recognition systems," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-September, pp. 466–470, 1 2018.

[5] M. Jefferson, "Usability of automatic speech recognition systems for individuals with speech disorders: Past, present, future, and a proposed model," 2019.

[6] V. Young 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*, vol. 22, no. 2, pp. 99–112, 2010.

[7] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, no. 1, pp. 48–60, 2000.

[8] 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.

[9] J. Jankovic, K. Schwartz, and D. T. Donovan, "Botulinum toxin treatment of cranial-cervical dystonia, spasmodic dysphonia, other focal dystonias and hemifacial spasm." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 53, no. 8, pp. 633–639, 1990.

[10] A. Blitzer, "Spasmodic dysphonia and botulinum toxin: experience from the largest treatment series," *European Journal of Neurology*, vol. 17, no. s1, pp. 28–30, 2010.

[11] 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.

[12] 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.

[13] R. A. Portnoy and A. E. Aronson, "Diadochokinetic syllable rate and regularity in normal and in spastic and ataxic dysarthric subjects," *Journal of Speech and Hearing Disorders*, vol. 47, no. 3, pp. 324–328, 1982. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/jshd.4703.324

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[15] G. B. Kempster, 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*, vol. 18, no. 2, pp. 124–132, 2009.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[17] Y. D. Heman-Ackah, 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*, vol. 112, no. 4, pp. 324–333, 2003, pMID: 12731627. [Online]. Available: https://doi.org/10.1177/000348940311200406

[18] J. Hillenbrand and R. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of speech and hearing research*, vol. 39, pp. 311–21, 04 1996.

[19] S. Sandoval, 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*, vol. 134, no. 5, pp. EL477–EL483, 2013.

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

[21] R. A. Samlan, 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.

[22] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2-3, pp. 95–108, Feb. 2000. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(99)00044-8

[23] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-September, pp. 1636–1640, 1 2018.

[24] M. Saxon, J. Liss, and V. Berisha, "Objective measures of plosive nasalization in hypernasal speech," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6520–6524.