Interspeech 2018
2-6 September 2018, Hyderabad

Whistle-blowing ASRs: evaluating the need for more inclusive automatic speech recognition systems

Meredith Moore[1], *Hemanth Venkateswara*[1], *Sethuraman Panchanathan*[1]

[1]Arizona State University's Center for Cognitive Ubiquitous Computing

mkmoore7@asu.edu, hemanthv@asu.edu, panch@asu.edu

Abstract

Speech is a complex process that can break in many different ways and lead to a variety of voice disorders. Dysarthria is a voice disorder where individuals are unable to control one or more of the aspects of speech—the articulation, breathing, voicing, or prosody—leading to less intelligible speech. In this paper, we evaluate the accuracy of state-of-the-art automatic speech recognition systems (ASRs) on two dysarthric speech datasets and compare the results to ASR performance on control speech. The limits of ASR performance using different voices have not been explored since the field has shifted from generative models of speech recognition to deep neural network architectures. To test how far the field has come in recognizing disordered speech, we test two different ASR systems: (1) Carnegie Mellon University's Sphinx Open Source Recognition, and (2) Google®Speech Recognition. While (1) uses generative models of speech recognition, (2) uses deep neural networks. As expected, while (2) achieved lower word error rates (WER) on dysarthric speech than (1), control speech had a WER 59% lower than dysarthric speech. Future studies should be focused not only on making ASRs robust to environmental noise, but also more robust to different voices.

Index Terms: speech recognition, voice disorders, dysarthric speech,

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]. 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 [4]. This creates a social barrier which prevents some individuals with voice disorders from fully participating in their community [5].

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 [6, 7]. 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.

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 [8, 9]. Results from systems that use larger vocabularies are extremely varied from 30.84% [10] to 97% recognition rate [11]. The highest reported accuracy on the biggest vocabulary using the least intelligible subjects was 85.05% from [12] 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, [13] 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 [6, 7]. 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) [14, 15, 16] as well as distance measures such as the connectionist temporal classification (CTC) [17]. 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.

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 [18, 19, 20]. However, we suggest that there is a need for

10.21437/Interspeech.2018-2391

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.

1.3. Domain Adaptation

The idea of adapting a model that is trained on one dataset on a different, but similar dataset and optimizing the model to perform well across domains is referred to as domain adaptation [21, 22]. We can formulate this study in terms of a domain adaptation problem and can use recent advances in domain adaptation techniques to improve the robustness of ASR systems to dysarthric speech.

1.4. Contributions

This paper presents an evaluation of how robust state-of-the-art ASR systems are to dysarthric speech. We test a model that uses generative methods (Gaussian Mixture Models and Hidden Markov Models), and a system that uses deep learning techniques. We compare the performance of these two systems on dysarthric speech and normative speech to obtain a baseline of how well state-of-the-art systems perform on differently-abled voices. We then make a case that there is a significant opportunity for improvement in state-of-the-art systems when it comes to being robust to correlated noise. In general, the contributions of this paper are to:

- Evaluate the performance of ASR systems on dysarthric speech
- Bring the attention of the speech community to the need for more inclusive ASR systems

2. Methods

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 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.

2.2. Datasets

2.2.1. UASPEECH

The Universal Access Speech (UASPEECH) dataset from the University of Illinois [23] 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.

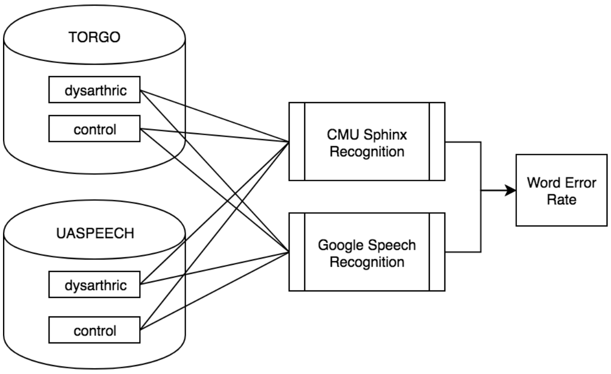

Figure 1: *The experimental set up used to test the performance of two ASR systems (Sphinx and Google) on dysarthric and control speech from two different dysarthric speech datasets (TORGO and UASPEECH)*

2.2.2. TORGO

The University of Toronto's TORGO database is a database of acoustic and articulatory speech from speakers with dysarthria [24] 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 [25], words from the word intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthria [26], the 10 most common words in the British National Corpus, and all of the phonetically contrasting pairs of words from [27]. 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.

2.2.3. Performance Measures

Word Error Rate (WER) is used to measure the performance of the ASR systems [28]. 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%.

$$WER = \frac{S + D + I}{N} \tag{1}$$

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,

Table 1: *Comparison between the combined performance of the ASR systems on dysarthric and control speech.*

	Dysarthric	Control	% Diff
WER	136%	74%	59%

Table 2: *Average Word Error Rate for each ASR system's performance on dysarthric and control speech*

Category	CMU Sphinx	Google	% Diff
Dysarthric	126%	43%	**84%**
Control	63%	20%	**74%**
%Diff	**55%**	**44%**	

divided by the total number of words.

$$RR = \frac{R}{N} \tag{2}$$

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.

3. Results

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 1. Table 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 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.

3.2. ASRs as a Model of Human Intelligibility

Figure 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 [23]. 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.

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 outside the range of 'normal' voices.

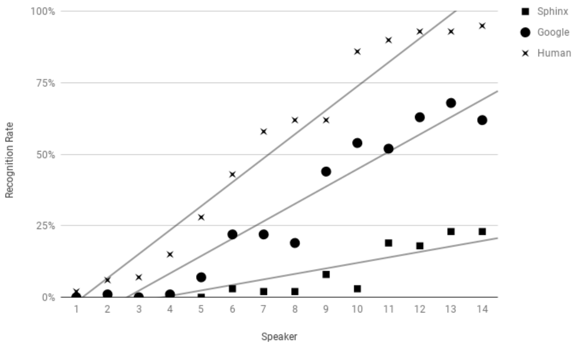

Figure 2: *A comparison of the recognition rate of the three different models of intelligibility: Human, Sphinx, and Google. Human recognition rates are denoted with the cross, Google with the circle, and Sphinx with the square.*

Table 3: *The Recognition Rate of Dysarthric Speech by Humans, Sphinx, and Google*

Speaker	Sphinx	Google	Human
1	0%	0%	2%
2	0%	1%	6%
3	0%	0%	7%
4	0%	1%	15%
5	0%	7%	28%
6	3%	22%	43%
7	2%	22%	58%
8	2%	19%	62%
9	8%	44%	62%
10	3%	54%	86%
11	19%	52%	90%
12	18%	63%	93%
13	23%	68%	93%
14	23%	62%	95%

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. The difference between control and one individual's dysarthric speech is shown in Figure 3, in this comparison it is clear that 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.

5. Proposed Directions

We propose that more research should be focused on creating ASR systems that are robust to both correlated and uncorrelated noises in order to make voice recognition systems more inclusive of different voices. 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

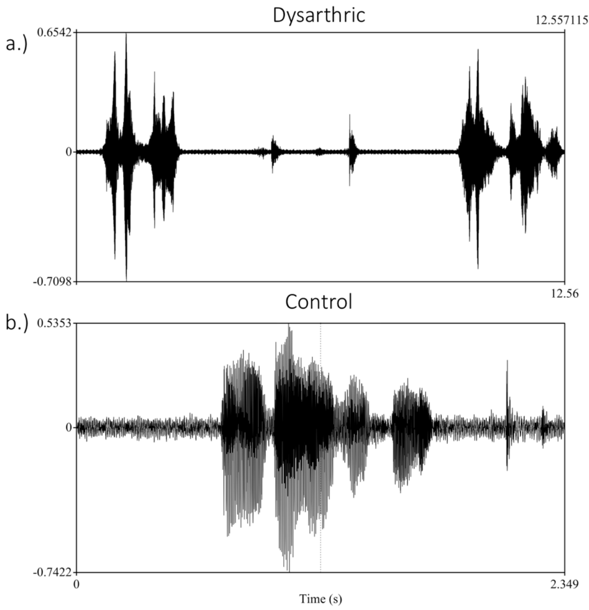

Figure 3: *A comparison of the waveform of two speech samples from UASPEECH. In **a**, the speech is from an individual with a dysarthria. The speech from **b** is from a control subject. The word that is spoken is 'autobiography'.*

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 [29]. We believe that the following areas will be essential in building these robust systems.

5.1. Datasets

The datasets used to train ASR systems need to be more inclusive of different voices than the current datasets. Currently, as shown in Table 4, 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 [30] 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.

5.2. 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.

Table 4: *An overview and comparison of the available datasets for dysarthric speech recognition—[31], [24], and [23], and one common normative speech dataset [32], shown in bold text.*

Dataset	Sub	Data Type	Utterances	Hours
[31]	11	Audio	Sentences	17.5
[23]	19	Audio, Visual	Isolated words	18
[24]	7	Audio, Visual, Articulatory	Non-words, Isolated words, Sentences	23
[32]	**543**	**Audio**	**Conversations**	**260+**

5.3. 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. Robust Models

With the collection and creation of more data and the application of domain adaptation techniques between normative speech and disordered speech, we expect to create significantly more robust models. 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.

6. Conclusions

In the last ten years, the performance of ASR system has significantly improved. Because of this increase in ASR system accuracy, the performance of ASR systems on dysarthric speech needed to be reevaluated. The re-evaluation demonstrated the poor performance of ASRs on dysarthric speech which led us to conclude that there is a need for systems that are not only robust to uncorrelated noise, but also for systems that are robust to correlated noise. If these ASR systems could be more robust to correlated noise, it would make them more usable by a population who have previously had a barrier to access for ASR systems, and therefore make ASR systems more inclusive.

7. Acknowledgements

We offer our sincere gratitude to the National Science Foundation's Alliance for Person-Centered Accessible Technologies Interdisciplinary Graduate Education Traineeship (APAcT IGERT), as well as the National Science Foundation's Graduate Research Fellowship, without which, this research would not have been possible.

8. 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] S. A. Borrie, M. J. McAuliffe, and J. M. Liss, "Perceptual learning of dysarthric speech: A review of experimental studies," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 1, pp. 290–305, 2012.

[5] L. Cooper, S. Balandin, and D. Trembath, "The loneliness experiences of young adults with cerebral palsy who use alternative and augmentative communication," *Augmentative and Alternative Communication*, vol. 25, no. 3, pp. 154–164, 2009.

[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. Hasegawa-Johnson, 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. IEEE, 2006, pp. III–III.

[9] P. D. Green, 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.

[10] P. D. Polur 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*, vol. 28, no. 8, pp. 741 – 748, 2006.

[11] H. V. Sharma 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*, ser. SLPAT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 72–79.

[12] S. Selva Nidhyananthan, 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*, vol. 19, no. 3, pp. 577–583, Sep 2016.

[13] J. R. Orozco-Arroyave, J. C. Vdsquez-Correa, F. Hnig, J. D. Arias-Londoo, 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)*, March 2016, pp. 6490–6494.

[14] G. Hinton, 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*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[15] J. M. Baker, 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*, vol. 26, no. 3, pp. 75–80, May 2009.

[16] L. Deng, 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*, May 2013, pp. 8599–8603.

[17] A. Graves, 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*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 369–376.

[18] X. Wang, 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*, vol. abs/1509.06103, 2015.

[19] Z. Pang 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*, vol. abs/1509.07211, 2015.

[20] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," *CoRR*, vol. abs/1711.05747, 2017.

[21] J. S. Bridle and S. J. Cox, "Recnorm: Simultaneous normalisation and classification applied to speech recognition," in *Advances in Neural Information Processing Systems 3*, R. P. Lippmann, J. E. Moody, and D. S. Touretzky, Eds. Morgan-Kaufmann, 1991, pp. 234–240.

[22] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1, pp. 151–175, May 2010.

[23] 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.

[24] 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.

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

[26] M. Walshe, N. Miller, M. Leahy, and A. Murray, "Intelligibility of dysarthric speech: perceptions of speakers and listeners," *International Journal of Language & Communication Disorders*, vol. 43, no. 6, pp. 633–648.

[27] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, pp. 482–499, 1989.

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

[29] S. Panchanathan, 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*, vol. 10, no. 5, pp. 942–951, 2016.

[30] I. Goodfellow, 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*, 2014, pp. 2672–2680.

[31] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, Oct 1996, pp. 1962–1965 vol.3.

[32] J. J. Godfrey, 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*, ser. ICASSP'92. Washington, DC, USA: IEEE Computer Society, 1992, pp. 517–520.