Study of the performance of automatic speech recognition systems in speakers with Parkinson's Disease

Moro-Velazquez, Laureano; Cho, JaeJin ; Watanabe, Shinji; Hasegawa-Johnson, Mark A.; Scharenborg, Odette; Kim, Heejin; Dehak, Najim

DOI
10.21437/Interspeech.2019-2993

Publication date
2019

Document Version
Final published version

Published in
Proceedings of Interspeech 2019

INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Study of the performance of automatic speech recognition systems in speakers with Parkinson's Disease

Laureano Moro-Velazquez[1,*] *, JaeJin Cho*[1,*]*, Shinji Watanabe*[1]*, Mark A. Hasegawa-Johnson*[2]*, Odette Scharenborg*[3]*, Heejin Kim*[4]*, Najim Dehak*[1]

[1]Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, USA
[2]Beckman Institute and ECE Department, University of Illinois at Urbana-Champaign, IL, USA
[3]Multimedia Computing Group, Delft University of Technology, the Netherlands
[4]Department of Linguistics, University of Illinois at Urbana-Champaign, IL, USA

laureano@jhu.edu, jcho52@jhu.edu

Abstract

Parkinson's Disease (PD) affects motor capabilities of patients, who in some cases need to use human-computer assistive technologies to regain independence. The objective of this work is to study in detail the differences in error patterns from state-of-the-art Automatic Speech Recognition (ASR) systems on speech from people with and without PD. Two different speech recognizers (attention-based end-to-end and Deep Neural Network - Hidden Markov Models hybrid systems) were trained on a Spanish language corpus and subsequently tested on speech from 43 speakers with PD and 46 without PD. The differences related to error rates, substitutions, insertions and deletions of characters and phonetic units between the two groups were analyzed, showing that the word error rate is 27% higher in speakers with PD than in control speakers, with a moderated correlation between that rate and the developmental stage of the disease. The errors were related to all manner classes, and were more pronounced in the vowel /u/. This study is the first to evaluate ASR systems' responses to speech from patients at different stages of PD in Spanish. The analyses showed general trends but individual speech deficits must be studied in the future when designing new ASR systems for this population.
Index Terms: automatic speech recognition, Parkinson's disease, dysarthria, word error rate, deep neural networks

1. Introduction

Parkinson's Disease (PD) is a chronic condition caused by the gradual death of brain cells implicated in the production of dopamine neurotransmitters. Dopamine plays an important role in motor tasks, and its absence or decrease affects the coordination, velocity, and acceleration of movements. Speech production involves the movement and coordination of multiple articulators and, consequently, it is affected by PD, that causes dysphonia and dysarthria (in particular *hypokinetic dysarthria*) in patients [1, 2, 3]. Dysphonia can be defined as the incapacity of the subject to produce a normal voiced sound, while dysarthria is more related to problems in articulation during the pronunciation of words. More specifically, *hypokinetic dysarthria* is characterized by a reduction of speech sound pressure level and articulation amplitude, slow speech rates combined with occasional rushes of fast speech, and a decrease in intelligibility. Some studys suggest that 90% of PD patients suffer from dysarthria after 7 years since diagnosis [4]. However, although the influence of PD on the speech produced by patients with PD is not always perceivable by human listeners, research using machine learning approaches has found PD-related cues in the speech of most of the studied patients, even those in the early developmental stages [5, 6, 7]. We therefore hypothesize that given that parkinsonian speech includes some specific traits even when no dysarthria or dysphonia is perceived by human listeners, these PD-related cues may influence the performance of Automatic Speech Recognition (ASR) systems.

We investigate this hypothesis by analyzing the differences in recognition performance of two different state-of-the art ASR systems (an end-to-end ASR and a conventional Hidden Markov Model (HMM)/Deep Neural Networks (DNN) hybrid ASR) on the speech of speakers with different developmental stages of PD and the speech of healthy controls (HC). Since people with motor-related diseases might need to use human-computer interaction systems to increase or regain their independence [8], the analysis of the recognition performance of state-of-the-art ASRs on parkinsonian speech can guide to design more robust ASR systems for speakers affected by PD.

2. Related work

Early research investigating the performance of ASR systems on speech from speakers with different degrees of dysarthria, e.g., suffering from Friedrichs Ataxia [9, 10], traumatic brain injury [11, 10] and cerebral palsy [11, 10], reported in all cases a higher Word Error Rate (WER) on speech from patients than on that of controls. The number of speakers in those studies, however, was rather low, i.e., below 7 while the present study scales up the number of speakers with PD to 43.

More recently, Tu et al. [12] performed an evaluation of the performance of the Google ASR engine [13] on the speech of 32 dysarthric speakers (the studied diseases are not specified) with a mean dysarthria severity of 5, which was perceptually rated on a scale ranging from 1 to 7. They found that WER was correlated with several different types of perceptual evaluation ratings, with the strongest correlation with articulatory ratings.

Despite the pervasiveness of PD in society, i.e., PD affects more than 1% of the population older than 60 and 3% older than 80 years of age [14], no research has been carried out to analyze the performance of state-of-the-art ASR systems with speakers with PD. Only some attempts to quantify the patient's intelligibility have been carried out using speech-to-text tools [15]. Some related applications, however, have been reported: Utilizing WER obtained with a cloud-based ASR can automatically detect PD with precision over 90% discriminating patient and control speakers [16].

*Equal contribution

http://dx.doi.org/10.21437/Interspeech.2019-2993

Finally, other studies proposed new ASR schemes for speakers suffering from severe dysarthria [17, 18, 19, 20], improving the performance of the conventional ASR in limited or wide-range vocabulary, still existing room for improvement.

3. Methodology

In this study we compare the performance of two ASR systems, i.e., an end-to-end and a hybrid HMM/DNN system. Both were trained using a Spanish speech corpus and tested on a different corpus containing speakers with and without PD in order to evaluate the differences in the ASR performance respect to the speech of the two groups. Various analyses were carried out:

- **Performance.** The performance was measured in terms of WER, Word Substitution Rate (WSR), Word Deletion Rate (WDR) and Word Insertion Rate (WIR). Spearman's correlations between the WER and ratings for the severity and stage of the disease in speakers with PD were calculated to evaluate the impact of these factors on the WER obtained with both ASR systems.
- **Character/phonetic unit analysis.** The substitution, insertion and deletion rates per character (with the end-to-end system), and per phonetic unit (with the hybrid system) were assessed in order to identify trends in the behavior of parkinsonian speech. These rates were calculated by normalizing the occurrences of each class (substitution, insertion and deletion) for a character/phoneme with the frequency of the character/phoneme in the reference text. For example, if 3 <a> were substituted by other characters, considering that the reference speech contained 10 <a>, its substitution rate is 0.3.
- **Manner classes analysis.** Finally, the differences in the ASRs' recognition performances among 5 consonant manner classes (i. e., manner of articulation) and vowels of both groups of speakers were analyzed by categorizing the Spanish phonemes into affricates, fricatives, liquid, nasals, plosives and vowels, according to [21].

3.1. Corpora

Two main speech corpora are employed in this study: Fisher Spanish (FisherSP) [22] was used to train two different ASR systems, and Neurovoz [7] to evaluate their performance on parkinsonian speech.

3.1.1. Fisher Spanish

The FisherSP corpus, created by the Linguistic Data Consortium to develop ASR systems in Spanish, was sampled at 8 kHz and 16 bits. It contains 163 h of telephone conversations from native Spanish speakers from at least 20 countries, along with their transcriptions. This corpus is split into subsets to train, evaluate, and test models. The test subset, which will be called *in-domain test set* to avoid confusion, was used to evaluate the performances of the two systems to be trained in this paper.

3.1.2. Neurovoz

The subset from the Neurovoz corpus employed in this work contains 43 speakers with PD and 46 control speakers whose mother tongue is Castillian Spanish. The speech material employed in this study consisted of 15 fixed sentences or Text-Dependent Utterances (TDU). Fig. 1 shows the frequency of each character over all speakers in the corpus for that material.

The demographic statistics of the subset used in this study are described in Table 1, including Unified Parkinson's Disease

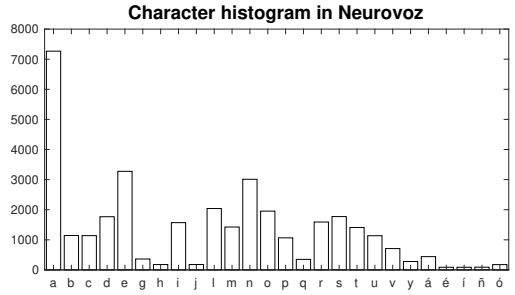

Figure 1: *Character histogram of Neurovoz, considering all speakers.*

Rating Scale (UPDRS) III [23] and Hoehn & Yahr (H-Y) [24] mean ratings. UPDRS III is a common scale used by clinicians to score the motor part of PD of patients based on clinical observations and questionnaires. UPDRS part III ranges from 0 (no motor problems) to 56 (motor functions seriously affected). H-Y is a more general scale ranging from 1 to 5 where 1 means mild symptoms or initial stage, and 5, patient totally dependent.

Table 1: *Neurovoz demographic statistics including average values with standard deviation in parenthesis in all rows except for age range. YSD stands for years since diagnosis.*

	Female		**Male**		**Total**	
	PD	**HC**	**PD**	**HC**	**PD**	**HC**
#Subjects	20	23	23	23	43	46
Age	70.7 (8.0)	69.5 (7.2)	67.0 (9.4)	61.3 (7.3)	68.7 (8.1)	65.4 (8.5)
Age range	59-86	58-86	41-80	53-77	41-86	53-86
UPDRS III	16.1(11.8)	-	14.8 (6.8)	-	15.4 (9.5)	-
H-Y	2.3 (0.8)	-	2.0 (0.5)	-	2.1 (0.7)	-
YSD	6.3 (6.1)	-	6.2 (4.8)	-	6.3 (5.1)	-

3.2. ASR systems

The end-to-end and HMM/DNN ASR systems have been selected to be tested with parkinsonian speech due to their popularity. Both models were trained and tested using the open source ASR toolkits: ESPnet [25] and Kaldi [26].

3.2.1. ASR 1: end-to-end approach

The end-to-end system was trained using sequences of acoustic frames as input and Spanish character sequences as output in one big neural network. The 83 dimensional feature including 80 dimensional filter bank and 3 dimensional pitch was extracted every 10ms for input acoustic frames. The model combined connectionist temporal classification (CTC) loss with cross entropy in the attention module to help with learning monotonic attention as in [27].

Since applying an explicit language model (LM) has been shown to improve WER performance in many languages [28], a word-level Recurrent Neural Network (RNN) LM [29] was included during word decoding. For the analyses of character deletions, insertions, and substitutions, the LM was not used as it degraded Character Error Rate (CER).

The model parameters were initialized with the values of pre-trained model's parameters, which were learned from the mixture of 10 languages [28]. An open source ASR toolkit, ESPnet, was employed for this end-to-end model training.

3.2.2. ASR 2: HMM/DNN hybrid approach

The HMM/DNN hybrid ASR system was trained by optimizing multiple modules separately: an acoustic model, pronunciation lexicon, and language model. The system was trained with phonetic units that are usually related to one allophone (a list of the Kaldi phonetic units and the correspondent allophones can be found in [7]). Thus, this system is employed to analyze phone-wise prediction between speakers with and without PD.

The acoustic model was trained based on Resnet-style chain Time delay neural network (TDNN) architecture allowing skip connections. It consists of 13 factorized TDNN layers with 128 dimensional bottle-neck, 1024 dimensional output layers, and L2 regularization in each. 40-dimensional high resolution MFCC features were used as input features and output predicts the distribution over senones. Additionally, i-vectors [30] were used as auxiliary features for speaker adaptation.

4. Results and discussion

4.1. Performance

Table 2 presents the results of the first type of analysis considering all the TDU (the same 15 per each speaker) from Neurovoz. On average, speakers with PD have at least a 27.4% (relative) WER higher than controls in both ASR systems, which suggests that none of the two systems presents any advantage respect to the other when used by speakers with PD, since average WER for speakers with PD was similar in both systems.

Average WER of control speakers was lower in the end-to-end system than in the HMM/DNN one. However, these values were higher than the WERs obtained with the *in-domain test set* of FisherSP (23.5% and 22.6% for ASR1 and ASR 2 respectively). This deviation can be attributed to differences between FisherSP and Neurovoz regarding type of speech (the first contains conversational speech and the second, TDU), recording environment, age or accent. It has been considered that these factors influence speakers equally both with and without PD.

Table 2: *Average error rates produced in the two ASR schemes analyzed with the two groups of speakers from Neurovoz. Best results are marked in bold.*

	ASR 1		**ASR 2**	
	PD	**HC**	**PD**	**HC**
WER	47.0%	**36.7%**	47.3%	**39.4%**
WSR	34.9%	**27.3%**	35.1%	**29.3%**
WDR	7.1%	**5.5%**	8.7%	**6.8%**
WIR	5.0%	**3.8%**	3.6%	**3.4%**

Fig. 2 shows the probability density functions of the average WER from the two groups of speakers. The WER tends to be higher in most of the speakers with PD compared to controls. The dissimilarity between the two groups is reduced in ASR 2, because fewer sentences from speakers with PD have very high WER ($> 70\%$), while, simultaneously, more control speakers have moderately high WER ($> 40\%$). A two-sample t-test [31] was applied to the distributions observed in the two ASR and in both cases the null hypothesis was rejected at the 5% significance level with p-value< 0.05 and confidence intervals $[-16.5, -5.0]$ for ASR 1 and $[-14.0, -2.6]$ for ASR 2. The differences between the distributions of the two schemes suggest that the end-to-end system tends to provide a lower WER in absence of dysarthria or pronunciation deviations.

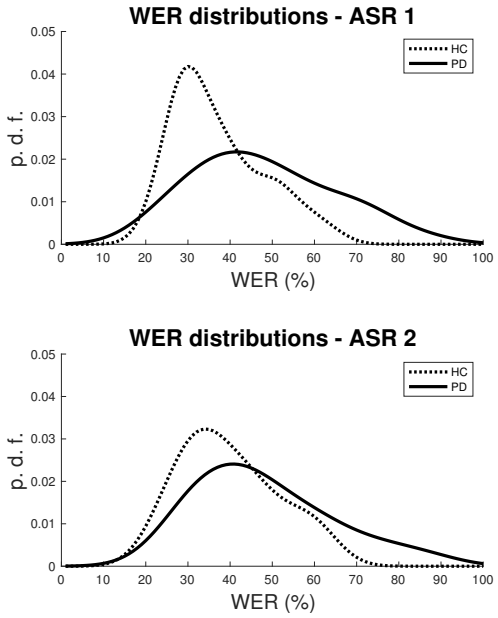

Figure 2: *Probability density functions (denoted as p.d.f.) of the average WER per speaker from the two groups of speakers employing ASR 1 (top) and ASR 2 (bottom) schemes. All speech material from both groups has been used to obtain the curves, which were estimated using a gaussian kernel.*

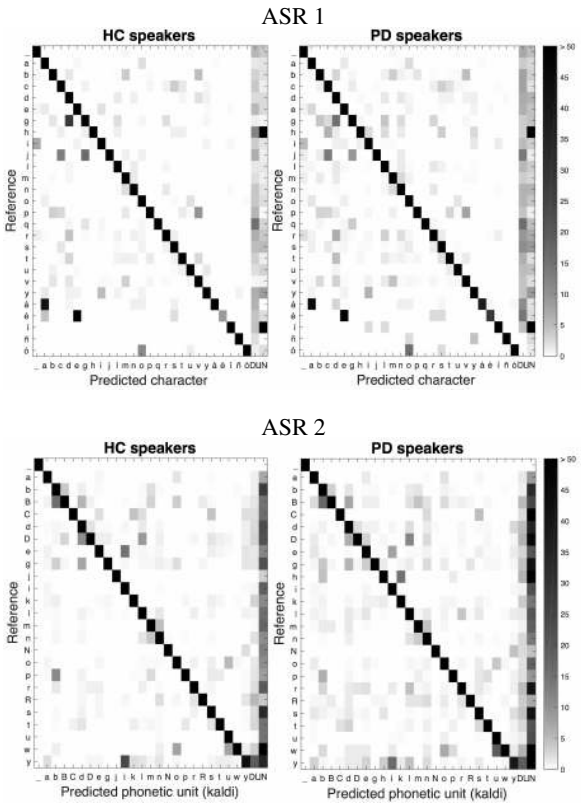

Figure 3: *Character substitution, insertion and deletion rates (%) per character in ASR 1 (top) and phonetic unit in ASR 2 (bottom). The two last columns labeled as DL and IN are deletion and insertion error rates, respectively. <_> represents the space character.*

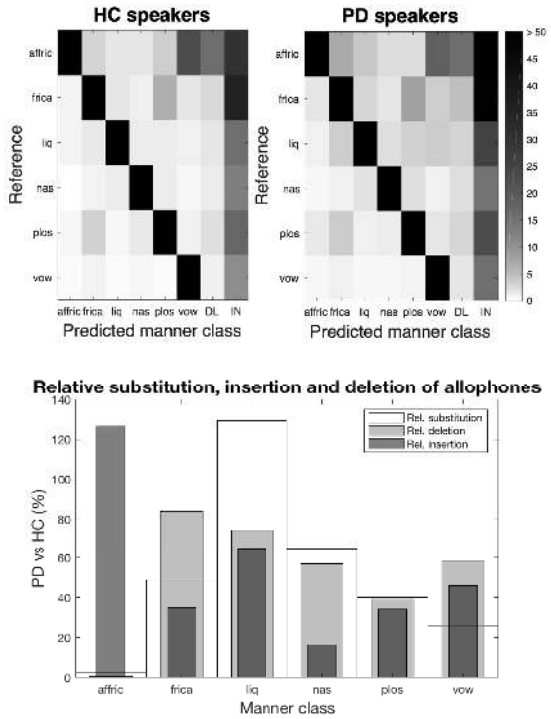

Figure 4: *Top: Manner classes substitution, insertion and deletion error rates in ASR 2. Bottom: relative substitution, insertion and deletion rates of speakers with PD respect to HC.*

Table 3 shows the Spearman's correlation between WER and the UPDRS III scores, H&Y scores and Years Since Diagnosis (YSD). These results suggest that the influence of PD in speech is not highly dependant on the stage of the disease as the WER is scarcely correlated with the H-Y rating and moderately correlated with the UPDRS III and YSD. This also supports the findings of previous studies indicating that PD affects speech even in early stages [5, 7].

4.2. Character/phonetic unit analysis.

Fig. 3 depicts the substitution, deletion and insertion rates per character for ASR 1 and per phonetic unit for ASR 2 both without the LM. For both ASR systems, the deletion and insertion rates tend to be higher in patients than in controls and affect more characters/phonetic units. This effect is related with the changes associated with the patient's motor functions that, for instance, can affect the onset and offset of the glottal source, the breathing control, and longer vowel duration (slow speech) combined with occasional rushes of fast speech [2], deriving in insertions and deletions. Similarly, Fig. 3 shows that the substitution rates are higher and more diverse in the case of patients, which is related to the reduction of the articulatory amplitude of tongue, lips and jaw, that produce a change in the frequency values of the formants, affecting vowels, and imprecise articulatory movements, affecting consonants. More specifically, it is associated to effects such as spirantization in which some types of consonants are substituted by fricatives during pronunciation, an effect that has been observed in speakers with PD [32]. On the other hand, substitution affects specially to phoneme /u/ (labels 'u' and 'w') in ASR 2, in which the substitution rate is 12 times higher in patients than in controls, supporting previous findings pointing to a higher degradation of this vowel respect to others [5, 7, 33]. Additionally, the analysis relative to ASR 1 suggests that the substitution rate of accented vowels (specially in the case of <í>) tends to be higher in patients, which is related to disprosody, commonly present in speakers with PD. Other characters that have a clearly higher substitution rate in patients compared to controls are <g, r, p, v> and <y>, findings which are in line with previous studies [34, 32, 35, 6].

Table 3: *Spearman's correlation (ρ) between WER and UPDRS III / H-Y scales / YSD in speakers with PD. ‡: p-value < 0.05*

	ASR 1	**ASR 2**
ρ_{UPDRS}	$0.32^{\ddagger}$	0.23
$\rho_{H\&Y}$	0.19	0.16
ρ_{YSD}	$0.47^{\ddagger}$	$0.38^{\ddagger}$

4.3. Manner classes analysis

Fig. 4 shows a comparison between the substitution, deletion and insertion rates in terms of manner classes for ASR 2, where affric. refers to affricate, frica. is fricative, liq. is liquid, nas. is nasal, plos. is plosive and vow. is vowel. The analysis shows that the substitution, deletion and insertion rates are always higher in the PD group, ranging from a 2.3% higher deletion rate of affricates to a 129.3% higher substitution rate in liquids. It is also observed that while substitution of plosives with fricatives is common in both speaker groups, the rate is higher in patients, supporting the prevalence of spirantization effect in PD-associated dysarthria.

For the future work, an analysis of the error rates depending on the points of articulation must be performed, along with a differentiation between the general trends and the individual deviations found in the parkinsonian speech, since that can have a high relevance in the proposal of new ASR systems for speakers with PD, which is the final goal. The information about these trends can help creating new more general ASR for speakers with PD while a system able to extract individual deviations will provide the information required to adapt the general ASR to each speaker.

5. Concluding remarks

We compared an end-to-end and hybrid speech recognizer on the task of the recognition of speech from speakers with Parkinson's Disease. Speech from individuals with PD suffered higher WER than speech from controls but the two distributions overlapped considerably. Speech from speakers with PD generally suffered greater de-accentuation (for instance, <í> to <i>) in the end-to-end scheme and greater mean deletion, insertion and substitution rates than speech from control speakers in all manner classes and almost all character/phonetic units studied, possibly correlated with the early onset of dysphonia and/or hypokinetic dysarthria. Future research will further investigate these differences and individual deficits to better understand the performance of ASR systems with speakers with PD in order to propose more adequate ASR schemes to be used by this population.

6. Acknowledgements

Authors want to thank Juan I. Godino-Llorente and Jorge A. Gomez-Garcia from Universidad Politecnica de Madrid for sharing their invaluable corpus Neurovoz.

7. References

[1] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential Diagnostic Patterns of Dysarthria," *Journal of Speech Language and Hearing Research*, vol. 12, no. 2, p. 246, 1969.

[2] J. Kegl, H. Cohen, and H. Poizner, "Articulatory consequences of Parkinson's disease: perspectives from two modalities." *Brain and Cognition*, vol. 40, no. 2, pp. 355–86, 1999.

[3] P. Blanchet and G. Snyder, "Speech Rate Deficits in Individuals with Parkinson's Disease: A Review of the Literature," *Journal of Medical Speech - Language Pathology*, vol. 17, no. 1, pp. 1–7, 2009.

[4] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[5] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of parkinson's disease: Effect of speaking task," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171–2181, 2013.

[6] L. Moro-Velazquez, J. Gomez-Garcia, J. Godino-Llorente, J. Rusz, S. Skodda, F. Grandas, J. Velazquez, J. Orozco-Arroyave, E. Noth, and N. Dehak, "Study of the automatic detection of parkisons disease based on speaker recognition technologies and allophonic distillation," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 1404–1407.

[7] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, and N. Dehak, "A forced gaussians based methodology for the differential evaluation of parkinson's disease by means of speech processing," *Biomedical Signal Processing and Control*, vol. 48, pp. 205–220, 2019.

[8] R. K. Megalingam, R. N. Nair, and S. M. Prakhya, "Automated voice based home navigation system for the elderly and the physically challenged," in *2011 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE)*. IEEE, 2011, pp. 1–5.

[9] J. Wilson, Bronagh Blaney, "Acoustic variability in dysarthria and computer speech recognition," *Clinical Linguistics & Phonetics*, vol. 14, no. 4, pp. 307–327, 2000.

[10] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[11] N. Thomas-Stonell, A.-L. Kotler, H. Leeper, and P. Doyle, "Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy," *Augmentative and Alternative Communication*, vol. 14, no. 1, pp. 51–56, 1998.

[12] M. Tu, A. Wisler, V. Berisha, and J. M. Liss, "The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. EL416–EL422, 2016.

[13] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "your word is my command: google search by voice: A case study," in *Advances in speech recognition*. Springer, 2010, pp. 61–90.

[14] A. Lee and R. M. Gilbert, "Epidemiology of parkinson disease," *Neurologic clinics*, vol. 34, no. 4, pp. 955–965, 2016.

[15] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in parkinsons disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22 199–22 208, 2017.

[16] J. Vasquez-Correa, J. Orozco-Arroyave, and E. Nöth, "Word accuracy and dynamic time warping to assess intelligibility deficits in patients with parkinsons disease," in *2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*. IEEE, 2016, pp. 1–5.

[17] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[18] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4924–4927.

[19] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech & Language*, vol. 27, no. 6, pp. 1147–1162, 2013.

[20] E. Yilmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, "Multi-stage dnn training for automatic recognition of dysarthric speech," 2017.

[21] A. Quilis, *Tratado de fonología y fonética españolas*. Editorial Gredos, 1993.

[22] D. Graff, S. Huang, I. Cartagena, K. Walker, and C. Cieri, "Fisher spanish transcripts ldc2010t04," *Web Download, Philadelphia, USA*, 2010.

[23] S. Fahn, *Recent developments in Parkinson's Disease*. Raven Pr, 1986.

[24] M. M. Hoehn and M. D. Yahr, "Parkinsonism onset, progression, and mortality," *Neurology*, vol. 17, no. 5, pp. 427–427, 1967.

[25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[27] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[28] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiat, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," in *Proc. of SLT*, 2018.

[29] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based rnn language models," *arXiv preprint arXiv:1808.02608*, 2018.

[30] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[31] N. Cressie and H. Whitford, "How to use the two sample t-test," *Biometrical Journal*, vol. 28, no. 2, pp. 131–148, 1986.

[32] J. A. Logemann and H. B. Fisher, "Vocal Tract Control in Parkinson's Disease," *Journal of Speech and Hearing Disorders*, vol. 46, no. 4, p. 348, 1981.

[33] K. Tjaden, J. Lam, and G. Wilding, "Vowel acoustics in Parkinson's Disease and multiple sclerosis: comparison of clear, loud, and slow speaking conditions." *Journal of speech, language, and hearing research : JSLHR*, vol. 56, no. 5, pp. 1485–502, 2013.

[34] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *Journal of Speech and hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[35] G. Weismer and M. McNeil, "Articulatory characteristics of parkinsonian dysarthria: Segmental and phrase-level timing, spirantization, and glottal-supraglottal coordination," *The dysarthrias: Physiology, acoustics, perception, management*, pp. 101–130, 1984.