DOI: 10.61782/fa.2023.0766

forum acusticum 2023

DEVELOPING A MACHINE-LEARNING MODEL FOR DETECTING INTELLIGIBILITY DIFFERENCES IN INDIVIDUALS WITH VOICE DISORDERS: A FEASIBILITY STUDY

Mary Pietrowicz[1*] **Diana Orbelo**[2] **Amrit Kamboj**[3]
Manoj Krishna Yarlagadda [3] **Kevin Buller**[3] **Sara Charney**[2] **Cadman Leggett**[3]
Keiko Ishikawa[3]
[1] Applied Research Institute, University of Illinois, USA
[2] Department of Otolaryngology, Mayo Clinic, USA
[3] Department of Gastroenterology, Mayo Clinic, USA
[4] Department of Communication Sciences and Disorders, University of Kentucky, USA

ABSTRACT

Voice disorders can reduce an individual's ability to produce intelligible speech; however, intelligibility in dysphonia has limited study. Current methods of intelligibility assessment are subjective and time-consuming, making reliable, efficient monitoring of patient progress difficult for clinicians. Machine-learning techniques, however, may provide novel, automated assessment solutions. This study aims to discover machine-learning models that differentiate habitual speech (HS) from hyperarticulated or "clear speech" (CS). Two corpora with same-subject recordings of HS and CS were used. The corpus consisted of 115 speakers, 65 healthy and 50 with mild-to-moderate voice disorders, saying six sentences from the Consensus of Auditory-Perceptual Evaluation. Acoustic analyses revealed significant differences between HS and CS in speech rate and CPP for female speakers. Various machine modeling techniques are explored for their ability to differentiate HS and CS, and the results are reported.

Keywords: *voice disorders, intelligibility, clear speech, AI, machine learning*

[*] ***Corresponding author***: *marybp@illinois.edu.*

1. INTRODUCTION

Voice disorders, also known as dysphonia, affect a significant proportion of the global population, with a lifetime prevalence of an estimated 30% in the general United States population [1]. These disorders lead to reduced speech intelligibility, particularly in noisy environments [2], and can considerably impact affected individuals' quality of life and social interactions [3]. Although various voice therapy techniques have been developed to improve speech intelligibility, assessing the effectiveness of these interventions, and estimating the intelligibility gain can be challenging. The challenge is that intelligibility is typically judged subjectively and perceptually, with assessments influenced by listener experience, knowledge, and familiarity with the speaker [4]. Therefore, an objective, software tool that can accurately estimate intelligibility gains resulting from the implementation of voice therapy techniques is needed.

Clear speech, a speaking style characterized by hyper-articulation of speech sound, is known to yield greater intelligibility than casual speech. It has been widely studied for its potential to improve communication for individuals with hearing and speech disorders [5-7]. The acoustic-phonetic properties of clear speech are distinct from those of casual speech and include features, such as a slower speaking rate, increased intensity, articulatory precision, and expanded vowel space [8]. Due to its intelligibility benefits, clear speech was incorporated into Conversation Training Therapy, a voice therapy program [9].

Automated detection of clear speech could therefore support multiple ends, including providing 1) an objective, computer-aided assessment of the clear speech condition, 2) an objective, computer-aided measurement contributing to intelligibility assessment, and 3) input features useful in machine modeling of a variety of expressive speech states, speech disorders, or neuropsychiatric conditions. To our knowledge, prior work has yet to produce an AI system that can automatically detect or assess clear speech. Recent work in voice-enabled AI, however, has demonstrated techniques for automated detection of voice disorders [10-11], emotional states [12-14], neurological conditions [15-20], psychiatric conditions [21-25], and more. In this study, we aim to 1) assess acoustical differences between clear and conversational speech in dysphonic and healthy voices and 2) produce a preliminary machine model that can support the automated detection of clear speech using similar voice-enabled AI techniques. Results are reported.

2. METHODS

2.1 Description of the speech database

This study used a corpus of HS and CS audio recordings, including six sentences from the Consensus of Auditory-Perceptual Evaluation of Voice (CAPE-V) [26]. These recordings were collected as part of two clinical studies: one aimed to evaluate the effectiveness of gargle phonation therapy [27], and another sought to examine the feasibility of using speech-based biomarkers for automatic detection of gastroesophageal reflux disease [28]. The corpus included 37 females, with normal voice and speech with an average age of 55.95 years (SD = 15.06), and 39 females with mild-to-moderate dysphonia, with an average age of 59.21 years (SD = 16.20). Among male participants, 28 had normal voice and speech, with an average age of 60.43 years (SD = 14.36), and 11 had mild-to-moderate dysphonia with an average age of 64.45 years (SD = 13.60). Four speech-language pathologists determined dysphonia severity via auditory-perceptual rating on a 0-100 scale. Speakers were native speakers of American English.

2.2 Speech Recording Procedures

Speech samples were recorded on a digital recorder (TASCAM-DR-40X) with a headset microphone (AKG C555L), using a consistent 5 cm distance from the corner of the participant's mouth. The sampling rate of the recordings was 44.1 kHz with a depth of 16 bits. The recordings were collected in a quiet office room.

2.3 Acoustic analyses

The samples were acoustically analyzed with PRAAT to obtain speech rate, intensity, and cepstral peak prominence (CPP), a measure sensitive to dysphonic voice quality [29]. Speech rate was calculated using a script by de Jong and Wempe [30]. CPP was obtained with a PRAAT plug-in software with a voice detection function [31].

2.4 Statistical analyses for the acoustic measures

A repeated measures ANOVA was conducted to examine the effect of diagnosis, sex, and speech production style on acoustic measurements, specifically speech rate, intensity, and CPP (Cepstral Peak Prominence). A pairwise t-test with Bonferroni correction was utilized to identify pairs with statistically significant differences.

2.5 Machine Model Development

To develop a preliminary machine model capable of detecting clear speech, we first prepared the CAPE-V data for analysis by ensuring samples were single-channel, 44.1K sampling rate, 16-bit recordings, and rescaling the recording amplitudes for each speaker's recording. Next, the recordings were segmented by sentence. Afterwards, 130-frame-level features (low-level descriptors, or LLDs) and 6369 summary features were extracted using the OpenSMILE [32] ComParE 2016 data set [33] configured to use 60 msec frames advancing at a 10 msec rate. The feature set was selected because the range of features was suitable for detecting qualitative changes in voice and had been successfully used in practice and in paralinguistic challenges [33] to detect a variety of vocal expression modes and health states.

Next, random forest classifiers were explored for their ability to distinguish between clear and habitual speaking styles, including both dysphonic and normal speakers. Separate models were developed for males and females due to the fundamental gender differences and the small size of the dataset. Both frame-level features (the instantaneous LLD measurements) and summary features (statistical measures on the frame-level measurements across the CAPE-V utterances) were explored during model building. Models were validated using a nested 3-fold cross-validation, and conditions were randomly balanced so that each condition had equal representation within a given model and fold. First, low-variance features were removed from consideration. Next, features were ranked within fold using the ANOVA F-value, and models were trained, also

within fold, using a successively smaller number of features based on the within-fold feature ranking (a recursive feature elimination, or "RFE" approach). The best and average model performance measurements were reported (F1 scores). Then, the ability to distinguish between clear and habitual speaking styles was explored for dysphonic and normal speakers separately, using similar machine models, and compared with models that included both dysphonic and normal speakers. Finally, to measure the relationship between clear/habitual speech and intelligibility, similar random forest classifiers were developed to measure the ability to discern normal and dysphonic speech in the clear and habitual speaking styles.

3. RESULTS

3.1 Acoustic results

3.1.1 Speech rate

The results of a repeated measures ANOVA indicated no significant main effect of diagnosis, $F(1, 112) = 0.832$, $p = 0.3638$, $\eta^2 = 0.38$. However, there was a marginally significant main effect of sex, $F(1, 112) = 3.770$, $p = 0.0547$, $\eta^2 = 1.70$. For the within-subjects factor, a significant main effect of speech production style was observed, $F(1, 114) = 10.4$, $p = 0.002$, $\eta^2 = 2.246$.

A pairwise t-test with Bonferroni correction indicated a significant difference in speech rate between clear and habitual speech styles was observed ($t(75) = -4.68$, $p_{adj} = 0.0000123$) for females. For males, no significant difference in speech rate between clear and habitual speech styles was found ($t(38) = 0.262$, $p_{adj} = 0.794$): see Fig. 1.

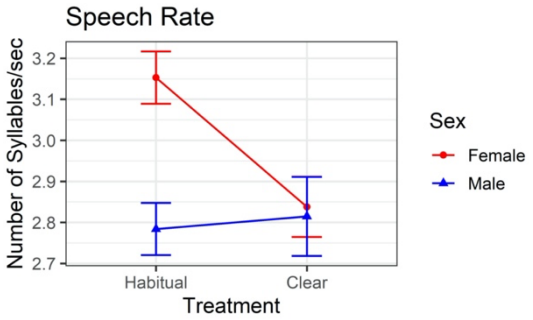

Figure 1. Line plot showing speech rate for habitual and clear speech in female and male speakers. The dot indicates mean and error bar indicates standard error.

3.1.2 Intensity

A repeated measures ANOVA revealed no significant main effect of diagnosis $F(1, 112) = 0.127$, $p = .722$, $\eta^2 = 4.00$. However, there was a marginally significant main effect of sex, $F(1, 112) = 3.813$, $p = .053$, $\eta^2 = 120.33$. A significant main effect of speech production style was observed for the within-subject factor: $F(1, 114) = 103.7$, $p < .001$, $\eta^2 = 194.32$: see Fig. 2.

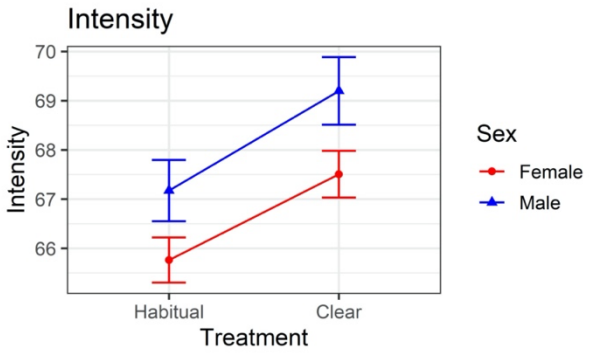

Figure 2. Line plot showing intensity for habitual and clear speech in normal and dysphonic speakers. The dot indicates mean and error bar indicates standard error.

3.1.3 CPP

A repeated measures ANOVA revealed a significant main effect of diagnosis, $F(1, 112) = 8.447$, $p = 0.004$, $\eta^2 = 19.06$, as well as a significant main effect of sex, $F(1, 112) = 6.276$, $p = 0.014$, $\eta^2 = 14.16$. For the within-subjects factor, a significant main effect of speech production style was observed, $F(1, 114) = 5.912$, $p = 0.0166$, $\eta^2 = 10.51$.

A pairwise t-test with Bonferroni correction was conducted to compare the CPP between clear and habitual speech production styles across different groups based on diagnosis and sex. In normal females, a significant difference in CPP between the two speech styles was also found ($t(36) = -2.06$, $p_{adj} = 0.047$). For dysphonic females, a significant difference in CPP between habitual and clear speech styles was observed ($t(38) = -5.06$, $p_{adj} = 0.00001$). For dysphonic males, no significant difference in CPP between clear and habitual speech styles was detected ($t(10) = -1.11$, $p_{adj} = 0.294$). Similarly, no significant difference in CPP between the two speech styles was observed in normal males ($t(27) = 0.863$, $p_{adj} = 0.396$): see Fig. 3

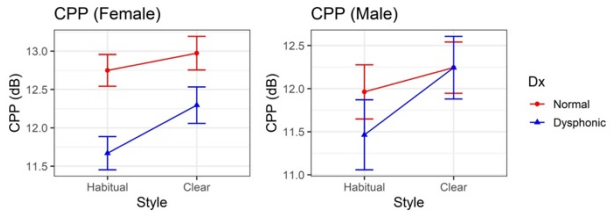

Figure 3. Line plot showing CPP rate for habitual and clear speech in normal and dysphonic speakers. The dot indicates mean and error bar indicates standard error.

3.1.4 Correlation between intensity and CPP

Because CPP is known to be sensitive to vocal intensity [34], a Pearson's product-moment correlation was conducted to analyze their relationship. The results indicated no significant correlation between intensity and CPP (r = 0.0779, $t(228)$ = 1.18, p = 0.239). The 95% confidence interval for the correlation ranged from -0.052 to 0.205.

3.1.5 Alpha Ratio

A repeated measures ANOVA indicated a significant main effect of diagnosis on the dependent variable, $F(1, 112)$ = 1.493, p = .224, η^2 = 0.011, and no significant main effect of sex, $F(1, 112)$ = 2.553, p = .113, η^2 = 0.019. However, style had a significant main effect on the dependent variable, $F(1, 114)$ = 79.53, p < .001, η^2 = 0.031.

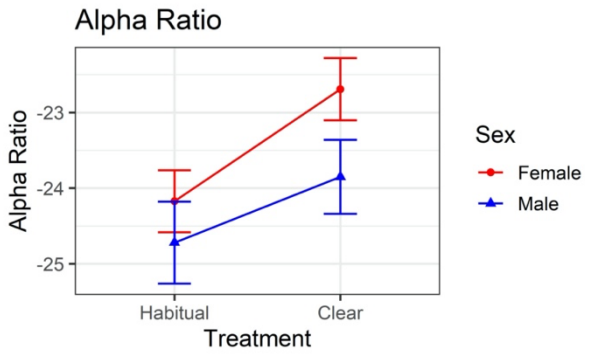

Figure 4. Line plot showing *Alpha ratio* for habitual and clear speech in normal and dysphonic speakers. The dot indicates mean and error bar indicates standard error.

3.2 Automated detection of the clear speech condition

The resulting models demonstrate that the clear speech condition is discernable in CAPE-V speech in both males and females. See Figure 5 below. We report the best and average F1 scores obtained in modeling for discernment of clear speech in both males and females (see Figure 5). The summary (SUM) features far outperformed frame-level (LLD) features in modeling, resulting in best F1 scores for models based on SUM features at 0.71 for males and 0.83 for females. The instantaneous, frame-level features, when extended across the entirety of the utterances, did not result in a random forest classifier that could easily discern between conditions. The models based on female speech also outperformed the models based on male speech by 10% or more. Figure 6 shows the receiver operating characteristic (ROC) curve for clear vs. habitual speech in both males and females.

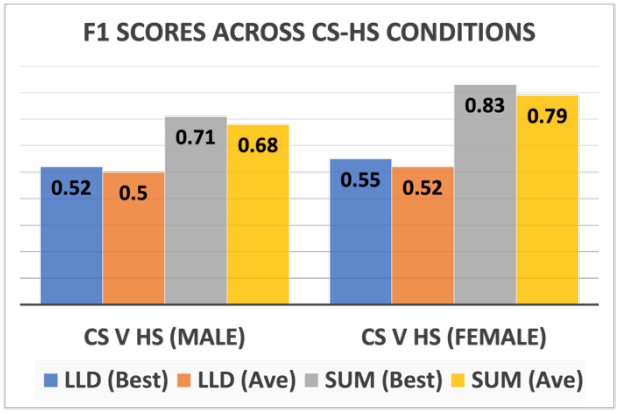

Figure 5. Results show the performance of Random Forest classifiers trained to discern clear from habitual conditions using both OpenSMILE frame-level (LLD) and Summary (SUM) features. Conditions were balanced and best and average F1 scores were reported.

Figures 7 highlights differences between the clear and habitual conditions for highly ranked features in men. Many of these features applied to the minSegLen, or minimum segment length condition, in which segment boundaries are defined by a signal changing more than a designated threshold when a current frame is compared to a running average computed over prior frames. The minSegLen condition differences are likely due to combined differences in articulation, vowel duration, and separation between words present in clear vs. habitual speech. Spectral skewness measures the symmetry of a spectrum around its arithmetic mean; therefore, signals that have relatively high energy around the fundamental frequency compared to the energy distributed to the rest

of the spectrum will have a higher skewness value. Examples of speech with more energy in the higher frequencies includes noisy and resonant speech. Since the minSegLength functional is also applied to spectral skewness, the model may be tracking articulation differences between clear and habitual speech within short segments (emphasized consonants).

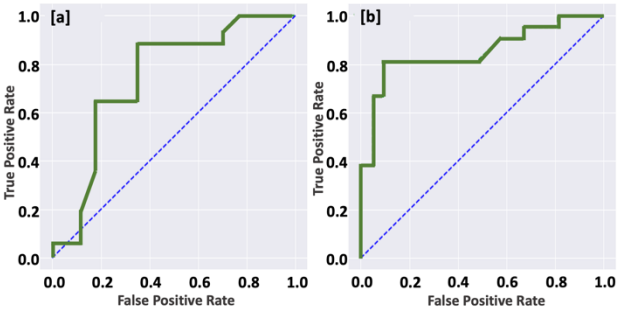

Figure 6. Receiver Operating Characteristic (ROC) curve for Clear vs. Habitual speech for a) males (F1= 0.71) and b) females (F1=0.83) speech.

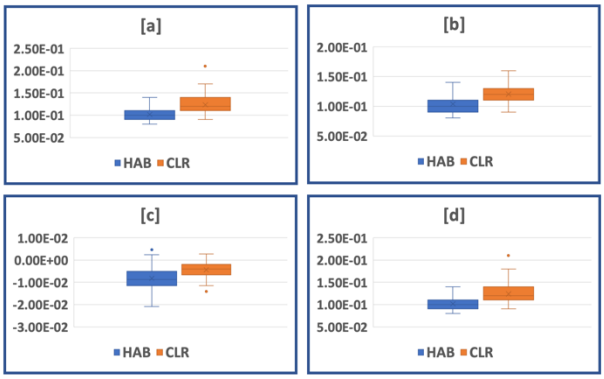

Figure 7. Differences between the clear (CLR) and habitual (HAB) conditions in highly-ranked features in males: a) mfcc_sma_de[7]_minSegLen,
b) mfcc_sma[11]_minSegLen,
c) pcm_fftMag_fband250-650_sma_linregc1, and
d) pcm_fftMag_spectralskewness_sma_minSegLen.

Figure 8 highlights differences between the clear and habitual conditions for highly ranked features in women. The harmonic-to-noise ratio (HNR) likely shows differences in consonant articulation between the clear and habitual conditions (logHNR_sma_centroid feature). In the pcm_zcr_sma_de_minSegLen features, the zero crossing rate (ZCR) mirrors differences in high frequency components between conditions. This is specifically looking at how ZCR changes during speech in the context of the minimum segment length condition. Differences in articulation of consonants, separation of words, and durations are probably reflected in both the HNR and ZCR features. The 1000-4000 Hz band in women includes higher harmonics, formants (most often F2, F3), and noise. Therefore, the pcm_fftMag_fband1000-4000_sma_linregc1 is likely reflecting formant differences, differences in articulations, differences in vowel/consonant durations, and differences in how these values change during an utterance between the two speaking styles. The "lengthL1norm" is the sum of the magnitudes of the frequency vectors in the space. This feature, too, is likely to reflect differences between signals containing many higher frequency components and those that do not. Note that the spectra of noisy, obstruent consonants typically have high-frequency components that sonorant consonants and vowels do not.

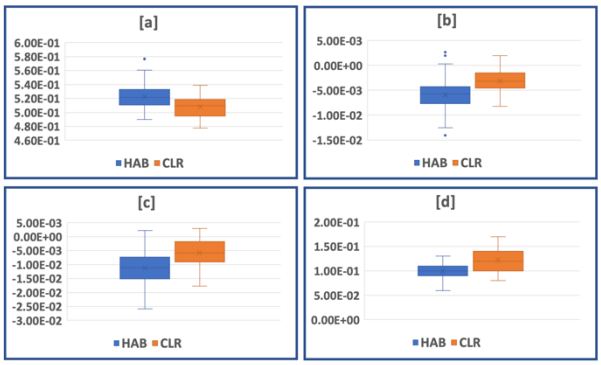

Figure 8. Differences between the clear (CLR) and habitual (HAB) conditions in highly-ranked features in females: a) logHNR_sma_centroid,
b) pcm_fftMag_fband1000-4000_sma_linregc1,
c) audspec_lengthL1norm_sma_linregc1, and
d) pcm_zcr_sma_de_minSegLen.

Figure 9 shows the differences in discerning CS and HS for dysphonic and normal speakers. This distinction is more challenging to make using OpenSMILE features when speakers are dysphonic, which is reflected by higher F1 scores for normal speakers.

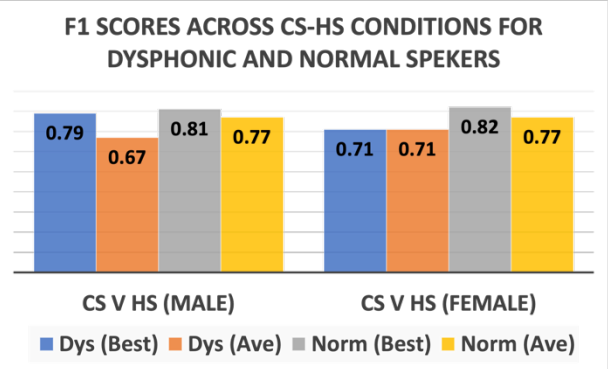

Figure 9. Results show the performance of Random Forest classifiers trained to discern clear from habitual speech for both dysphonic (Dys) and normal (Norm) speakers separately using OpenSMILE summary (SUM) features. Conditions were balanced and best and average F1 scores were reported.

3.3 Automated detection of the dysphonic condition in clear and habitual speech

Clear speech increased the difficulty of recognizing the dysphonic condition in the female voice samples by about 6%, potentially reflecting improved voice quality. Similar results were not seen in the male voice samples, probably because the signal differentiating clear speech in males was not as strong in our sample. See Table 1.

Table 1. Detection of dysphonic vs. normal voices in clear/habitual speaking styles for males and females.

Clear/ Habitual	Gender (M/F)	Best F1	Average F1 (μ/σ)
Clear	F	0.77	0.70/0.058
Habitual	F	0.83	0.76/0.067
Clear	M	0.78	0.72/0.074
Habitual	M	0.78	0.72/0.094

4. DISCUSSION & CONCLUSIONS

The results of acoustic analyses showed that intensity was greater for clear speech than habitual speech in both female and male speakers. Furthermore, Alpha ratio was greater for clear speech, indicating that clear speech signals contained more high-frequency energy than habitual speech [8]. These results corroborate previous acoustic studies of clear speech. However, speech rate and CPP were significantly greater for clear speech only in female speakers. The lack of difference in male speakers is likely because most female speakers came from the gargle phonation study, whereas male speakers came from the GERD study. Although the same instructions for eliciting clear speech were used in both studies, participants in the gargle phonation study were instructed by speech-language pathologists (SLPs). In contrast, participants in the GERD study were instructed by research staff without clinical training in speech-language pathology. Previous studies have shown that the way clear speech is produced can be affected by how the instructions are given to speakers [35]. The lack of correlation between intensity and CPP indicates that the increase in CPP was not due to an increase in intensity. Instead, the increase in CPP observed in clear speech is likely associated with increased periodicity in the signal.

The machine models successfully separated men's and women's clear and habitual speech conditions. The higher F1 scores for females most likely reflected differences in instructions given to the participants (most of the female participants were instructed by SLPs). The highly-ranked features in the models for both males and females likely mirrored differences in articulation, duration, and separation of words between the two conditions.

While this study discerns differences in habitual and clear speech and demonstrates the feasibility of the approach, it remains a preliminary study due to the limited set of speech data with imbalanced populations and simple machine modeling techniques that limit the generalizability of the results. Future work will address these limitations, incorporate a more diverse set of speakers from different cultural backgrounds and geographical regions, and include a broader range of dysphonic speech types and dysphonia severity levels. The data sample was also biased across age and demographics and imbalanced across conditions. We addressed the imbalanced data for machine modeling via random undersampling of the majority class. While the machine models used in this study were classic machine learning models, more advanced deep learning techniques will improve model performance in the future. Finally, more work is needed to quantify the relationship between clear speech and intelligibility.

5. ACKNOWLEDGMENTS

This work was funded in part by Mayo Clinic's Departments of Otolaryngology (PI: Orbelo) and Gastroenterology (PI: Leggett) Small Grant Programs and a Mayo Max Innovation Award (PI: Leggett).

6. REFERENCES

[1] N. Roy, R. M. Merrill, S. Thibeault, R. A. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of voice disorders in teachers and the general population," *Journal of Speech, Language, and Hearing Research*, vol. 47, no. 2, pp. 281-293, 2004.

[2] K. Ishikawa, S. Boyce, L. Kelchner, M. G. Powell, H. Schieve, A. de Alarcon, and S. Khosla, "The effect of background noise on intelligibility of dysphonic speech," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 7, pp. 1919-1929, 2017.

[3] C. A. Rosen, A. S. Lee, J. Osborne, T. Zullo, and T. Murry, "Development and validation of the voice handicap index-10," *The Laryngoscope*, vol. 114, no. 9, pp. 1549-1556, 2004.

[4] G. Weismer, "Speech intelligibility," in *The Handbook of Clinical Linguistics*, M. J. Ball, M. R. Perkins, N. Müller, and S. Howard, Eds. Oxford, UK: Blackwell Publishing, 2008, pp. 568-582.

[5] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," Journal of Speech, Language, and Hearing Research, vol. 28, no. 1, pp. 96-103, 1985.

[6] J. A. Whitfield and A. M. Goberman, "Articulatory–acoustic vowel space: Application to clear speech in individuals with Parkinson's disease," *Journal of Communication Disorders,* vol. 51, pp. 19-28, 2014.

[7] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," The Journal of the Acoustical Society of America, vol. 112, no. 1, pp. 272-284, 2002.

[8] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *Journal of Speech, Language, and Hearing Research*, vol. 29, no. 4, pp. 434-446, 1986.

[9] J. Gartner-Schmidt, S. Gherson, E. R. Hapner, J. Muckala, D. Roth, S. Schneider and A. I. Gillespie, "The development of conversation training therapy: A concept paper," in Journal of Voice, vol. 30, no. 5, pp. 563-573, 2016.

[10] M. Huckvale and C. Buciuleac, "Automated detection of voice disorder in the Saarbrucken voice database: Effects of pathology subset and audio materials," *INTERSPEECH*, 2021, pp. 4850-4854.

[11] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A Survey on Machine Learning Approaches for Automatic Detection of Voice disorders," in Journal of Voice, vol. 33, no. 6, pp. 947.e11-947.e33, 2019.

[12] A. Baird, S. Amiriparian, N. Cummins, S. Sturbauer, J. Janson, E.-M. Messner, H. Baumeister, N. Rohleder, and B. Schuller, "Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test," in *INTERSPEECH*, pp. 534-538, 2019.

[13] D. Bone, J. Mertens, E. Zane, S. Lee, S. Narayanan, and R. Grossman, "Acoustic-Prosodic and Physiologic Response to Stressful Interactions in Children with Autism Spectrum Disorder," in *INTERSPEECH*, 2017, pp. 147-151.

[14] T. L. New, Q. Xu, C. Guan, and B. Ma, "Stress Level Detection Using Double-Layer Subband Filter," in *INTERSPEECH*, 2015, pp. 3695-3699

[15] C. Agurto, O. Ahmad, G.A. Cecchi, R. Norel, M. Pietrowicz, E. Eyigoz, E. Mosmiller, E. Baxi, J.D. Rothstein, P. Roy, J. Berry, and N. Maragakis, "Analyzing Progression of Motor and Speech Impairment in ALS," in *Proc. IEEE Eng. Med. Biol. Soc. (EMBC)*, pp. 6097-6102, 2019.

[16] G. An, D.G. Brizan, M. Ma, M. Morales, A.R. Syed, and A. Rosenberg, "Automatic Recognition of Unified Parkinson's Disease Rating from Speech with Acoustic, i-Vector and Phonotactic Features," in *Proc. INTERSPEECH*, pp. 508-512, 2015.

[17] R. Norel, M. Pietrowicz, C. Agurto, S. Rishoni, and G.A. Cecchi, "Detection of Amyotrophic Lateral Sclerosis (ALS) via Acoustic Analysis," in *Proc. INTERSPEECH*, pp. 377-381, 2018.

[18] J.R. Orozco-Arroyave, F. Honig, J.D. Arias-Londono, J.F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Noth, "Automatic Detection of Parkinson's Disease in Running Speech Spoken in Three Different Languages," *J. Acoust. Soc. Am.*, vol. 139, no. 1, pp. 481-500, 2016.

[19] M. Perez, W. Jin, D. Le, N. Carlozzi, P. Dayalu, A.M. Roberts, and E. Mower Provost, "Classification of Huntington Disease Using Acoustic and Lexical Features," in *Proc. INTERSPEECH,* pp. 1898-1902, 2018.

[20] Y.A. Qadri and V. Kumar, "Early Detection of Epilepsy using Automatic Speech Recognition," *Indian J. Sci. Technol*., vol. 9, no. 47, 10.17485/ijst/2015/v8i1/106440, 2016.

[21] F. Carrillo, M. Sigman, D. Fernandez Slezak, P. Ashton, L. Fitzgerald, J. Stroud, D. J. Nutt, and R. L. Carhart-Harris, "Natural speech algorithm applied to baseline interview data can predict which patients will respond to psilocybin for treatment-resistent depression," *J Affect Disord*, vol. 230, pp. 84-86, 2018.

[22] C. Agurto, M. Pietrowicz, R. Norel, E. K. Eyigoz, E. Stanislawski, G. Cecchi, and C. Corcoran, "Analyzing acoustic and prosodic fluctuations in free speech to predict psychosis onset in high-risk youths," in *42nd Annual International conference of the IEEE Engineering in Medicine & Biology Society (EMBC),* 2020, pp. 5575-5579.

[23] C. Agurto, R. Norel, M. Pietrowicz, M. Parvaz, S. Kinreich, K. Bachi, G.A. Cecchi, R.Z. Goldstein, "Speech Markers for Clinical Assessment of Cocaine Users," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6391-6394, 2019.

[24] N. Cummins, S. Scherer, J. Krajewski, S. Schnider, J. Epps, and T.F. Quatieri, "A Review of Depression and Suicide Risk Assessment Using Speech Analysis," *Speech Communication*, vol. 71, pp. 10-49, 2015.

[25] R. Gupta, S. Sahu, C. Espy-Wilson, and S. Narayanan, "Affect Prediction Approach Through Depression Severity Parameter Incorporation in Neural Networks," in *Proc. INTERSPEECH*, pp. 3122-3126, 2017.

[26] R. I. Zraick, G. B. Kempster, N. P. Connor, S. Thibeault, B. K. Klaben, Z. Bursac et al., "Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)," in *American Journal of Speech-language Pathology*, vol. 20, no. 1, pp. 14-22, 2010.

[27] D. Orbelo, S. Charney, M. Pietrowicz, D. Aka, S. Bayan, and K. Ishikawa, "Vocal Effort and Acoustic Analysis of Gargle Phonation versus Water Swallow in Patients with Mild Muscle Tension Dysphonia: A Clinical Trial," Journal of Voice, in review.

[28] K. Kamboj, M. K. Yarlagadda, M. Pietrowicz, K. Buller, P. G. Iyer, D. A. Katzka, et al., "S402 Voice Enabled Artificial Intelligence for Detection of Pathologic Gastroesophageal Reflux Disease and Barrett's Esophagus," in Official Journal of the American College of Gastroenterology| ACG, vol. 117, no. 10S, 2022, pp. e281-e282.

[29] Y. D. Heman-Ackah et al., "Cepstral peak prominence: a more reliable measure of dysphonia," *Annals of Otology, Rhinology & Laryngology*, vol. 112, no. 4, pp. 324-333, Apr. 2003.

[30] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," in *Behavior Research Methods*, vol. 41, no. 2, pp. 385-390, 2009.

[31] E. S. H. Murray, A. Chao, and L. Colletti, "A Practical Guide to Calculating Cepstral Peak Prominence in Praat," in *Journal of Voice*, 2022.

[32] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM)*, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 2010.

[33] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," *Proc. INTERSPEECH*, pp. 2001-2005, 2106. doi: 10.21437/Interspeech.2016-129.

[34] M. Brockmann-Bauser, J. H. Van Stan, M. C. Sampaio, J. E. Bohlender, R. E. Hillman, and D. D. Mehta, "Effects of vocal intensity and fundamental frequency on cepstral peak prominence in patients with voice disorders and vocally healthy controls," *J. Voice,* vol. 35, no. 3, pp. 411-417, May 2021.

[35] J. Lam, K. Tjaden and G. Wilding, "Acoustics of clear speech: Effect of instruction," in Journal of Speech, Language, and Hearing Research, vol. 55, no. 1, pp. 30-45, 2012, doi: 10.1044/1092-4388(2011/10-0245).