DOI: 10.61782/fa.2023.0922

forum acusticum 2023

EXPLORING AUTOMATED DETECTION OF BARRETT'S ESOPHAGUS VIA MACHINE MODELING AND ACOUSTIC ANALYSIS

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

ABSTRACT

Gastroesophageal reflux disease (GERD) affects approximately 18-27% of adults in North America; and chronic GERD is associated with Barrett's esophagus (BE), a precursor to esophageal adenocarcinoma. Current screening and diagnostic procedures for GERD/BE are invasive, expensive, and uncomfortable for the patient. Automated screening tools for GERD/BE based on voice analysis and modern machine learning techniques could, however, potentially enable early detection of GERD/BE without invasive procedures. In this study, standardized, scripted speech is collected, analyzed, and compared across three groups, including a) patients with BE (BE+), b) patients without endoscopic evidence of BE (BE-), and c) patients without GERD and without voice symptoms (normal). Acoustic differences across groups are reported. In addition, multiple machine learning techniques are explored, and machine models are trained to detect the BE+ condition. The ability of selected machine learning models to discern across BE+, BE-, and normal conditions is reported.

Keywords: *barrett's esophagus, gerd, machine learning, AI, computer-aided diagnosis (CAD)*

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

1. INTRODUCTION

Gastroesophageal reflux disease (GERD) is a common disorder in Western society affecting 18-27% of the population [1]. Undiagnosed chronic GERD can lead to development of Barrett's esophagus (BE) which is the only known precursor to esophageal adenocarcinoma (EAC). BE affects approximately 5% of the general population in the United States, and approximately 3-5% of patients with BE will develop EAC over their lifetime [2]. An esophagogastroduodenoscopy (EGD) with biopsies is required for diagnosis of BE, established when salmon-colored mucosa is visualized in the distal esophagus and esophageal biopsies demonstrate specialized intestinal metaplasia. At present, there is no non-invasive and cost-effective test to identify patients with pathological GERD, including BE.

The association between voice characteristics and GERD has been studied in patients with self-reported voice symptoms. However, pathologic GERD, even chronically present as in cases of BE, may lead to previously unrecognized voice changes in the absence of hoarseness or reflux laryngitis. Hoarseness attributed to GERD is present in approximately 10% of patients evaluated by otolaryngologists [3]. However, the exact mechanism as to how GERD causes laryngopharyngeal disease remains unclear. Animal studies suggest that exposure of gastric acid and pepsin may result in laryngeal damage [3]. Data are, however, lacking to support a causal association between GERD/BE and voice complaints [4].

Recent work in voice-enabled AI has demonstrated that embodied signals, particularly speech and language, can

successfully be used to model and detect a range of mental and emotional states [5-7] and health conditions, particularly in psychiatry [8-12], neurology [13-18], pulmonology [19-22], and cardiology [23]. The prior work explores a wide variety of machine/deep learning-signal processing applied to acoustic voice signals. We hypothesize that BE causes changes to the functioning of the upper GI system and vocal tract that are discernable via machine listening techniques, and present the results of an exploratory investigation into the following research question: "***Can the presence of Barrett's esophagus be detected via machine listening techniques applied to speech?***"

The primary contribution of this paper includes the development of a preliminary machine model capable of detecting the presence of BE via voice analysis. To our knowledge, no prior works have produced a voice biomarker or machine model for BE based on voice.

2. DATASET

The dataset consists of voice recordings, survey instrument data, and clinical assessment information from *114 participants* recruited from the Division of Gastroenterology and Hepatology at Mayo Clinic in Rochester, MN. Each subject completed the Gastroesophageal Reflux Disease Questionnaire (GERDQ) [24] and the Voice Handicap Index (VHI-10) [25]. Subjects were also asked whether they were currently taking proton pump inhibitors (PPIs), a common group of medications used to treat GERD. Their BE history was also recorded. For subjects with a history of BE, the study team further noted whether a subject had **BE Present** (BE-P) or treated **BE No Longer Present** (BE-N). Voice recordings of all subjects were collected in a quiet clinical office at the Mayo Clinic, Rochester MN, using professional recording equipment (TASCAM system).

Each participant read the first six sentences from the Rainbow Passage [26], a text commonly used in the assessment of voice disorders by speech language pathologists. This reading provided about 30-60 seconds of speech per subject (about an hour of speech). The recruiting efforts yielded *33 participants with BE Present (BE-P), 18 participants BE No Longer Present (BE-N), 14 GERD negative (GERD-) participants, and 47 Control subjects.* The Control group had GERDQ scores below 9 and VHI-10 scores below 11 (scores in the normal or healthy range); and they did not have a BE diagnosis. The Control group, however, also did not have an endoscopy to confirm the presence or absence of BE because an EGD was not medically indicated. While it is unlikely that the Control subjects had BE (especially given lack of GERD symptoms), it is possible, given that an estimated 5% of the US population has BE. Given these statistics, up to 0-3 of the Control subjects could possibly have asymptomatic BE. The GERD- group did not have GERD or voice symptoms as indicated by the GERDQ and the VHI-10 scores, and they also had endoscopic confirmation that BE was absent. Subjects with reported or confirmed voice disorders or with neurological or psychiatric disorders known to affect the voice (e.g., Parkinson's, dementia, major depressive disorder, etc.) were excluded from the study. All subjects with a history of BE were on PPIs as per clinical guidelines, while none in the Control group were on PPIs. Three participants in the GERD- group were on PPIs. All participants were over age 50 years (typical for BE patients) and were similar in age across the BE-P, BE-N, GERD-, and Control conditions.

All subjects in this initial study were male because of the demographics of the disease. Males with BE outnumber females at a ratio of at least 2:1, and the pool of available females with BE in the clinic was smaller than this. Further, since male and female voices have different spectral characteristics, modeling per gender separately is helpful for discovery of a BE voice biomarker, especially in the exploratory stages. Ongoing work is expanding the participant pool to include a more diverse group.

Table 1. The BE Study Dataset includes 33 with **BE Present (BE-P)** 18 with treated **BE No Longer Present (BE-N),** 14 **GERD-** (no evidence of GERD or voice difficulty, and confirmed BE negative via endoscopy), and 47 **Controls** (no evidence of GERD or voice difficulties, but did not undergo endoscopy to definitively confirm presence or absence of BE).

BE-P (# on PPI)	**BE-N** (# on PPI)	**GERD-** (# on PPI)	**Control** (# on PPI)
33 (33)	**18** (18)	**14** (3)	**47** (0)

3. ANALYSIS AND EXPERIMENTS

To address the research question, we first prepared the audio data for analysis. All data recordings were resampled to a single-channel, 44.1K, 16-bit format as needed. Excess leading and trailing silences were trimmed from all utterances. Any experimenter speech was removed, and

recording amplitudes were rescaled. Next, 130 frame-level features (low-level descriptors, or LLDs) and 6369 summary features were extracted using the OpenSMILE [27] ComParE 2016 data set [28] (configured for 60 msec frames with a 10 msec advance). This feature set was selected for its feature coverage and because it had been successfully used both in the detection of a variety vocal expression modes and health states and in prior paralinguistic challenges [28].

Next, support vector machine (SVM), random forest (RF), and nearest neighbor classifiers were explored for their ability to distinguish across the following conditions: 1) BE-P vs. Controls, 2) BE-N vs. Controls, 3) BE-P vs. GERD-, and 4) BE-N vs. GERD-. Both frame-level features (instantaneous measurements) and summary-level features (statistics on frame-level measurements across the utterances) were explored. Models were validated using 3-fold nested cross validation techniques. Conditions were randomly balanced so that each condition had equal representation within a given model. Features were ranked and selected within fold, and low-variance features were removed from consideration. Recursive feature elimination was used, and best and average model performance measurements (accuracy, F1, precision and recall) are reported for the RF classifier model, since RF generally outperformed the other models.

4. RESULTS

The resulting machine models demonstrate that both BE-P and BE-N conditions are discernable from both the Control and GERD- conditions. See Figure 1 below. Model results using the "Control" condition as opposed to the "GERD-" condition were generally better; however, this result may reflect the small numbers of GERD- samples. We report both the best F1 scores obtained in modeling for each classification task and the average classification scores across folds. Summary features far outperformed frame-level (LLD) features in modeling; models using only LLD features could not perform most of the tasks. Best results across all tasks using summary features resulted in F1 scores between 0.68 and 0.91.

Figure 2 compares the best performing models for each classification task via Receiver Operating Characteristic curves. These high-performing models were all trained with summary features.

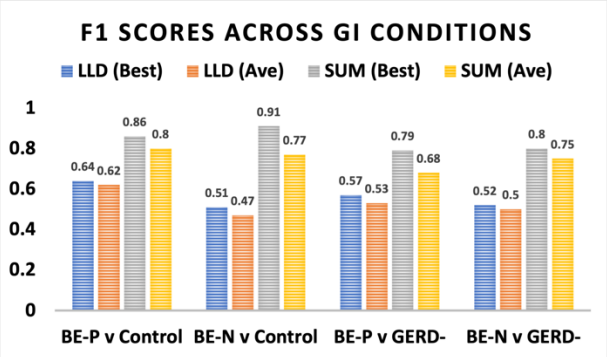

Figure 1. Machine modeling results show the performance of Random Forest classifiers trained to discern BE-P and BE-N from Control and GERD- conditions using both OpenSMILE LLD and Summary (SUM) features. Conditions were balanced and best and average F1 scores are reported.

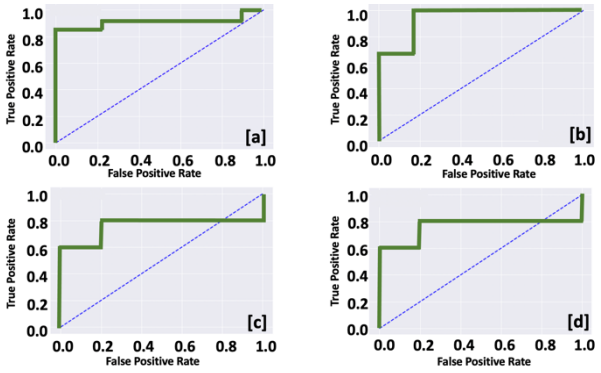

Figure 2. Receiver Operating Characteristic (ROC) curve for a) BE-P vs Control (F1=0.86), b) BE-N vs Control (F1=0.91), c) BE-P vs GERD- (F1=0.79), and d) BE-N vs Control (F1=0.80).

Figure 3 examines four highly-ranked features for differences between the BE-P and Control conditions: a) pcm_fftMagSpectralSkewness_sma_lpc0, b) pcm_fftMag_psySharpness_sma_percentile1.0, c) pcm_fftMag_spectralEntropy_sma_percentile1.0, and d) pcm_fftMag_spectralCentroid_sma_percentile1.0. Spectral skewness is a measure of how symmetric a spectrum is around its arithmetic mean. In general, spectral skewness will be high in signals that have relatively high energy around the fundamental frequency in comparison with the energy distributed in the rest of

the spectrum (at higher frequencies). Psychoacoustic sharpness (psysharpness) is a numeric measure of sound perception that is based on the amount of high-frequency components in a sound. Spectral entropy is a measure of randomness/uniformity. Sounds with primarily integer multiples of the fundamental will have lower spectral entropy than sounds with many random, high-frequency components (noise). The spectral centroid is the center of mass of the spectrum and indicates where most of the energy is concentrated; a sound with a higher spectral centroid may be higher in fundamental frequency, or it could have more/stronger high-energy components. All of these features point to differences in the high frequency and/or noise content in the voices. In addition, three of these features mark the 1st percentile, and highlight the number of values that are the less than or equal to this value. Voices that are hoarse, strained, breathy, or airy will contain more random, higher-frequency components that are not integer multiples of the fundamental. Voices that are tense are also generally higher in pitch than those that are not tense. The model appears to be detecting subtle differences in the high frequency content between the voices of healthy people and people with the BE-P condition. While these features each show distinct differences between conditions, no single feature clearly separates conditions in our model; multiple features work together to provide separation.

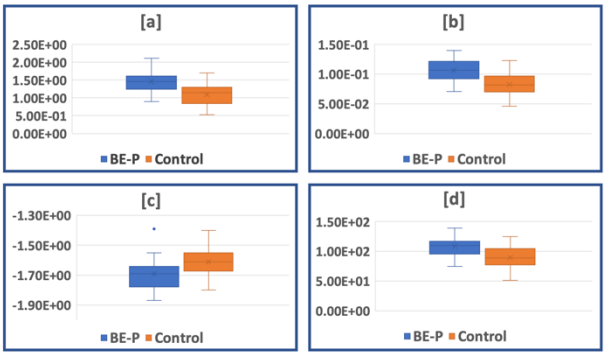

Figure 3. Differences between the BE-P and Control conditions in highly-ranked features: a) pcm_fftMag_spectralEntropy_sma_percentile1.0, b) pcm_fftMag_psySharpness_sma_percentile1.0, c) pcm_fftMag_spectralSkewness_sma_lpc0, d) pcm_fftMag_spectralCentroid_sma_percentile1.0.

DISCUSSION & CONCLUSIONS

This exploratory study demonstrates that the presence of BE, present or no longer present, can be detected via voice analysis. Furthermore, vocal characteristics across time, captured in summary features, are superior mirrors of BE when compared to instantaneous frame-level features. While this analysis demonstrates the feasibility of the approach, it is a preliminary study based on a very limited dataset of read speech and very simple machine modeling techniques. The study was also limited to male subjects from the Rochester, MN area. Future work will expand the dataset to include a variety of speech utterances, video, and a more diverse group of participants from different regions, cultures, and gender. A range of more powerful analytic and deep learning techniques will also be explored, particularly convolutional neural networks (CNN), long short-term memory networks (LSTM), pretrained embeddings, and transformer models. With respect to participants, comorbidity of conditions affecting the voice is quite common. In order to begin to understand the effects of BE on the voice, we attempted to isolate the BE condition and excluded several potential confounders. A clinically capable tool and model set, however, will need to be trained on a range of speakers with a variety of conditions affecting the voice, using techniques such that the models can learn to infer the presence of BE in the context of real-world conditions that include these comorbidities. Finally, further exploration into the differences between the BE-P and BE-N conditions is needed so that a clinical tool can make this distinction as well.

5. ACKNOWLEDGMENTS

This work was supported by a Mayo Clinic Jerry A. Wenger Career Development Award and a Mayo Clinic Division of Gastroenterology and Hepatology MAX Innovation Award.

6. REFERENCES

[1] H. B. El-Serag, S. Sweet, C. C. Winchester, and J. Dent, "Update on the epidemiology of gastro-oesophageal reflux disease: a systematic review," *Gut*, vol. 63, no. 6, pp. 871-880, Jun. 2014.

[2] P. Sharma, "Barrett's Esophagus A Review," *Jama-Journal of the American Medical Association*, vol. 328, no. 7, pp. 663-671, 2022.

[3] M. F. Vaezi, D. M. Hicks, T. I. Abelson, and J. E. Richter, "Laryngeal signs and symptoms and

gastroesophageal reflux disease (GERD): A critical assessment of cause and effect association," *Clin Gastroenterol Hepatol*, vol. 1, no. 5, pp. 333-344, 2003.

[4] G. T. Schneider, M. F. Vaezi, and D. O. Francis, "Reflux and Voice Disorders: Have We Established Causality?," *Curr Otorhinolaryngol Rep*, vol. 4, no. 3, pp. 157-167, 2016.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] G. Deshpande and B. W. Schuller, "An Overview on Audio, Signal, Speech, & Language Processing for COVID-19," *arXiv* preprint arXiv:2005.08579, 2020.

[20] X. Ding, D. Nassehi, and E. C. Larson, "Measuring Oxygen Saturation With Smartphone Cameras Using Convolutional Neural Networks," *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2603-2610, 2018.

[21] J. Han, K. Qian, M. Song, et al., "An Early Study on the Intelligent Analysis of Speech under COVID-19: Severity, Sleep, Quality, Fatigue, and Anxiety," *arXiv*:2005.00096v2, 2020.

[22] E. C. Larson, M. Goel, G. Boriello, S. Heltshe, M. Rosenfeld, and S. N. Patel, "SpiroSmart: using a microphone to measure lung function on a mobile phone," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 280-289, 2012.

[23] E. Maor, D. Perry, D. Mevorach, N. Taiblum, Y. Luz, I. Mazin, A. Lerman, G. Koren, and V. Shalev, "Vocal Biomarker Is Associated with Hospitalization and Mortality Among Heart Failure Patients," *Journal of the American Heart Association*, vol. 9, no. 7, pp. e013359, 2020.

[24] M. Tielemans and M. Oijen, "Online follow-up of individuals with gastroesophageal reflux disease using a patient-reported outcomes instrument: Results of an observational study," *BMC gastroenterology*, vol. 13, no. 1, pp. 144, 2013.

[25] C. A. Rosen, A. S. Lee, J. Osborne, T. Zullo, and T. Murry, "Development and validation of the voice handicap index-10," *Laryngoscope*, vol. 114, no. 9, pp. 1549-56, 2004.

[26] G. Fairbanks, *Voice and Articulation Drillbook*, 2nd ed. New York: Harper & Row, 1960, pp. 124-139.

[27] 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.

[28] 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.