Research Article ISSN 2835-6276

American Journal of Medical and Clinical Research & Reviews

# Artificial Intelligence for Screening Voice Disorders: Aspects of Risk Factors

**Pedersen M**

*Correspondence: *Pedersen M*

*Received:* **17 Jan 2025;** *Accepted:* **25 Jan 2025;** *Published:* **05 Feb 2025**

**Citation:** Pedersen M. Artificial Intelligence for Screening Voice Disorders: Aspects of Risk Factors. AJMCRR. 2025; 4(2): 1-8.

## Abstract

*Early detection of voice disorders significantly enhances diagnostic accuracy and treatment outcomes. The objective of this paper is to emphasize the existing lack of evidence regarding the clinical application of artificial intelligence (AI) in verbal communication disorders. A literature search conducted through the Royal Society of Medicine, UK, on AI and voice disorders identified 24 AI-related articles, with Parkinson's Disease being the most frequently studied condition. However, only a limited number of AI applications provided clinically useful results. The underlying challenges pertain to data measurement, data detection, software training and testing, and inadequate specificity, sensitivity, and accuracy. The necessity of clinically validated AI models is crucial, also in addressing neurological and genetic disorders, which affect 6% and 15% of the population, respectively, aside from primary laryngeal disorders. Transparent AI software is essential for future applications in foundational software models.*

## Introduction

Significant progress has been made in the diagnostics of voice disorders. The role of artificial intelligence (AI) encompasses endoscopic imaging as well as other measurement modalities. To date, AI has not been clinically implemented in any of these domains. The models have not been adapted to randomized, prospective, double-blinded clinical trials. A high-speed imaging setup revealed that only half of the images obtained in a clinical setting were suitable for AI analysis [1]. AI-assisted laryngeal endoscopy remains at the conceptual stage [2].

Voice evaluation is inherently complex. A recent presentation at a joint conference of the European Laryngological Society, the Union of European Phoniatricians, and the European Academy of Phoniatrics discussed key aspects that should be considered in AI-based voice evaluation [3]. Despite an updated consensus in 2023 [4], AI applications were not addressed. Established evaluation methods such as the Voice Handicap Index (VHI) [5], its modified short version, and the singer-specific VHI remain crucial [6]. While AI has the potential to analyze these test results, its implementation in clinical practice has not yet occurred but is currently under development. Airflow-related voice measurements, such as maximum phonation time

(MPT), are fundamental tools in voice pathology that require further AI development [7]. Additionally, expert evaluations of voice remain essential, although AI applications, such as those involving the GRBAS test, have yet to produce clinically viable results due to challenges related to dataset quality and accuracy [8].

To explore the application of AI in the analysis of acoustical parameters in verbal communication, a review of the past decade (2013-2023) was conducted using the Royal Society of Medicine (RSM) library, which identified 54 AI-related studies on Parkinson's Disease. A focused search on Parkinson's Disease and voice disorders revealed 98 relevant studies, 24 of which included AI applications, with 20 including reviews, published in the last five years, demonstrating a rapid increase in research activity[9]. However, these studies primarily focused on disease classification rather than treatment efficacy.

The aim of this study is to highlight the limitations of current AI studies to achieve more transparent and clinically applicable results in the future. These findings have broader implications for other voice-related disorders. The analysis presents critical perspectives on voice-related assessments, particularly in comparison to other biological parameters such as genetic regulation of voice function.

## Methods: Insufficiency of Studies

Table 1 provides an overview of various challenges associated with voice-related acoustical datasets, including measurement parameters, dataset size, and insufficient demographic information (e.g., age, gender, and socio-economic status).

**Table 1**. The most obvious problems in the referred voice-related acoustical datasets

| Category | Problems Identified |
|---|---|
| Usability | Sufficient dataset size, Sample duration |
| Precision | Measurement recordings |
| Content | Articulated vowels, Spoken sentences |
| Population | Age, gender, race, socio-economic status, and other disorders |
| Disorder Characteristics | Other characteristics of the disorder in question |

Issues related to data detection arise primarily from the lack of detailed software descriptions. Factors such as microphone placement, noise parameters, and feature extraction methods, which are the process of identifying and deriving meaningful acoustical measurements from raw audio signals, using AI, for analysis, classification, or further processing significantly impact measurement reliability. Which differs from traditional acoustical voice measurements without the use of AI. Table 2 outlines the primary challenges associated with data detection.

**Table 2**. Challenges Associated with Data Detection.

| Category | Challenges Identified |
|---|---|
| **Microphone Placement** | Distance of microphone |
| **Noise Factors** | Environmental noise, System noise, Background noise, Room acoustics |
| **Measurement Parameters** | Frequency area measurement |
| **Feature Extraction** | Signal processing techniques, Feature selection methods |

Key AI-related performance metrics, including sensitivity, specificity, and accuracy, are often inadequately reported. Furthermore, the rationale for selecting specific AI models is frequently absent. Table 3 summarizes these challenges.

**Table 3**. Software Description and Evaluation Metrics.

| Evaluation Criteria | Measurement Description |
|---|---|
| **Sensitivity (recall)** | Ability to correctly identify positive cases (True Positive Rate). |
| **Specificity** | Ability to correctly identify negative cases (True Negative Rate). |
| **Accuracy** | Overall correctness of the model. |
| **Cross-Validation** | Validation technique (e.g., k-fold, leave-one-out) to assess performance. |
| **Training Setup** | Dataset split ratio, preprocessing methods, feature selection. |
| **Testing Setup** | Evaluation metrics, unseen data performance, generalization ability. |
| **AI Model Choice & Description** | Justification of model selection, architecture, and application suitability. |

## Results of Analysis of the AI-Related Studies

Analysis of the 24 reviewed studies revealed significant deficiencies in dataset descriptions and methodological transparency, as exemplified in Tables 4, 5, and 6.

**Table 4**. Acoustical Datasets

| Category | Problems Identified | Articles That Provide Data | Articles That Do Not Provide Data | Reason for Missing Data |
|---|---|---|---|---|
| **Usability** | Insufficient dataset size was explicitly noted in 6 articles. Sample duration inconsistencies were mentioned in 4 articles, highlighting variability in recording lengths. | 6 articles | 12-13 articles | Articles may focus on model performance or feature analysis without discussing dataset size or duration inconsistencies. |
| **Precision** | Variability in measurement protocols was identified in 5 articles, focusing on inconsistent quality and lack of standardization. Subjective assessments integrated with objective measures in 3 articles. | 5 articles | 13-14 articles | Many articles assume standardized datasets or do not detail measurement protocols explicitly. |
| **Content** | Limited diversity in vocal tasks was reported in 7 articles; most datasets included only basic phonemes like /a/, /o/, /u/ or simple phrases. | 7 articles | 11-12 articles | Studies may focus on specific phonemes or a single type of vocal task, ignoring the diversity of speech content. |
| **Population** | Underrepresentation of demographic groups was noted in 4 articles. Insufficient age diversity was noted in 5 articles. Lack of consideration for co-occurring disorders in 3 articles. | 4–5 articles | 13–15 articles | Articles often do not address demographic diversity or co-occurring disorders, focusing on the primary disorder (PD). |
| **Disorder Characteristics** | Limited characterization of specific vocal impairments was mentioned in 6 articles, including tremor, monotone voice, and pitch irregularities. Lack of integration with neuropsychological assessments in 3 articles. | 6 articles | 12-13 articles | Some studies focus purely on classification accuracy without delving into disorder-specific vocal impairments. |

Table 5 shows that only some articles partially address a given problem or touch on it indirectly, making it unclear whether they should definitively count toward the total.

**Table 5**. Challenges in Acoustic Data Processing.

| Category | Challenges Identified | Articles That Provide Data | Articles That Do Not Provide Data | Reason for Missing Data |
|---|---|---|---|---|
| **Microphone Placement** | Distance of the microphone was identified as a challenge, impacting recording quality and consistency. | 5 articles | 14 articles | Many articles assume ideal recording conditions or focus on software processing without detailing placement issues. |
| **Noise Factors** | Environmental noise variability noted in 6 articles. System noise reported in 4 articles. Background noise and room acoustics noted in 5 articles. | 6 articles | 12-13 articles | Articles often assume noise-free environments or do not evaluate noise impact explicitly. |
| **Measurement Parameters** | Frequency area measurement inconsistencies discussed, focusing on frequency resolution and range limitations. | 4 articles | 14-15 articles | Many studies do not report detailed frequency analysis, focusing on simpler feature extraction methods. |
| **Feature Extraction** | Challenges in signal processing noted in 7 articles, particularly for non-linear or dynamic features. Feature selection challenges reported in 5 articles. | 7 articles | 11-12 articles | Some articles focus on algorithm testing or dataset creation without detailing feature extraction. |

Table 5 shows that only a few papers have well-defined features.

**Table 6**. Evaluation Metrics and Experimental Frameworks for Parkinson's Disease Detection Models.

| Evaluation Criteria | Articles That Provide Data | Articles That Do Not Provide Data | Measurement Description |
|---|---|---|---|
| **Sensitivity (recall)** | 7 articles | 11-12 articles | Sensitivity ranged from 73% to 95%. Specifically: 73–80%: 2 articles; 81–90%: 3 articles; 91–95%: 2 articles. Higher values were associated with well-defined datasets and robust feature engineering. |
| **Specificity** | 7 articles | 11-12 articles | Specificity ranged from 60% to 96%. Specifically: 60–70%: 2 articles; 71–85%: 3 articles; 86–96%: 2 articles. Variability was influenced by dataset imbalance and the inclusion of healthy controls. |
| **Accuracy** | 6 articles | 12-13 articles | Accuracy ranged between 84% and 96%. Specifically: 84–89%: 3 articles; 90–96%: 3 articles. Higher accuracies were often observed in ensemble models or those using optimized feature sets. |
| **Cross-Validation** | 8 articles | 10-11 articles | Common validation techniques included 10-fold cross-validation (used in 5 articles), and leave-one-out validation (used in 3 articles). The use of robust cross-validation methods mitigated overfitting risks. |
| **Training Setup** | 7 articles | 11-12 articles | An 80:20 split for training and testing was most common (reported in 4 articles), while feature selection methods such as PCA were employed in 3 articles. |
| **Testing Setup** | 6 articles | 12-13 articles | Evaluation metrics included F1-scores: 0.75–0.79: 2 articles; 0.80–0.89: 3 articles. AUC values ranged between 85–90% in well-tuned models, reported in 4 articles. |
| **AI Model Choice & Description** | 7 articles | 11-12 articles | Popular models included SVM (used in 4 articles), CNN (used in 3 articles), and AdaBoost (used in 2 articles). Novel architectures like p-CRNN were mentioned in 1 article. |

Table 6 shows that there is not an adequate number of papers that provide specific numbers for evaluation metrics clinics.

The lack of clinically relevant outcomes underscores the need for improved AI models tailored to voice-related disorders.

## Results of Voice-Related Measurements in Parkinson's Disease and Genetics

Table 7 presents frequency calculations of various voice-related disorders. Software applications have been used to assess neurological disorders; however, clinical utility remains limited. Similarly, only a small fraction of genetic disorder studies have incorporated AI methodologies in the past five years.

**Table 7**. The Calculation of the Frequency of Some Voice-Related Disorders [3].

| Individuals: |
| --- |
| 1 Dysphagia: 4% of the adult population |
| 2 Dysphonia: 3-9% of the adult population |

| Patients: |
| --- |
| 3 Parkinson's Disease: 80% |
| 4 Alzheimer's Disease: 84-93% |
| 5 Head &Neck oncology: +/- 40% |

| Country | Population | # Adults (25-65j) | Dysphagia (4%) | Dysphonia (3-9%) |
| --- | --- | --- | --- | --- |
| Belgium | 11,6M | 52% | 240K | 217K |
| The Netherlands | 17,5M | 52% | 364K | 328K |
| Germany | 83M | 53% | 1,80M | 1,5M |
| United-States | 332M | 65% | 8,6M | 6,2M |

| Region | Population | Parkinson's Disease | Alzheimer's Disease | H&N oncology |
| --- | --- | --- | --- | --- |
| Europe | 746M | 1,2M | 9,7M | 450K |
| United-States | 332M | 1M | 6,2M | 66K |

K = thousand, M = million people.

Software is used in tests of neurological disorders, but not with clinical consequences, meaning no observed clinical impact. This is also the case for genetic disorders where during the last 5 years in a search of RSM only 5 out of 61 voice-related studies found, had an AI-related implication. There is a good argument for the preliminary results of the AI studies. It is that measurement of voice-related parameters as such is a new area in many areas of disorders. In Table 8 the measured data are presented with the fundamental frequency as the paramount one. In many of the papers, it is noted that this was the first time voice-related parameters were used for the genetic syndrome.

**Table 8.** Frequency of Voice-Related Parameters in Papers on Genetics in the Last 5 Years.

| Assessment Method/ Feature | Number of Articles Reporting |
| --- | --- |
| VHI (Voice Handicap Index) | 6 |
| GRBAS (Listeners Test) | 10 |
| F0 (Fundamental Frequencies) | 20 |
| Jitter, shimmer | 8 |
| HNR/NHR (Harmonics to Nois Ratio/Noise to Harmonics Ratio) | 6 |
| MPT (Maximum Phonation Time) | 6 |
| ML (Machine Learning) | 5 |

Table 8 shows that machine learning is only used in 5 cases.

Regarding Parkinson's Disease, among the 98 studies conducted between 2013-2023, several focused on non-AI-based voice assessments, as summarized in Table 9. Common voice parameters such as fundamental frequency, jitter, shimmer, and harmonics-to-noise ratio were frequently used, along with subjective assessments such as the Voice Handicap Index and the GRBAS test. Although endoscopic evaluations were occasionally utilized, no studies integrated computerized image analysis with voice measurements.

**Table 9**. Non-AI papers [3].

| Parameters | Total |
|---|---|
| No Patient (cases) | 7561 (23 without no.) |
| Prospective | 25 |
| Randomized | 5 |
| (Case) Controls | 1513 |
| Retrospective | 6 |
| **HNR** | **23** |
| SNR (Signal to Noise Ratio) | 8 |
| **F0 (+stnd. dv.)** | **40** |
| **Intensity** | **24** |
| MPT | 14 |
| **JITTER APS/%** | **29** |
| **SHIMMER APS/%** | **23** |
| Spectrum, LTAS (Long Term Average Spectrum) | 9 |
| CEPSTRUM analysis | 5 |
| VRP (Voice Range Profile) | 4 |
| Telephone calls | 3 |
| Praat (Software) | 13 |
| **VHI** | **25** |
| GRBAS | 10 |
| Deep Brain Surgery | 7 |
| **AI** | **4** |
| Deep Learning | 9 |
| Laryngoscopy | 6 |

Table 9 shows the amount of non-AI paper measures.

## Discussion and Conclusion

This paper discusses the risk factors associated with artificial intelligence applications for various voice parameters. We have highlighted how these parameters are utilized in clinical settings, as, to date, voice parameters—referred to as features in artificial intelligence research—have yet to be adopted for clinical use.

The tables presented in this article provide a comprehensive overview of the challenges and advancements in applying artificial intelligence to voice-related disorders.

Table 1 highlights key issues in voice-related acoustic datasets, including insufficient dataset size, demographic representation, and content diversity. Table 2 explores the technical challenges of data detection, such as microphone placement, noise factors, and feature extraction techniques, which underscore the need for standardized data collection methodologies. Table 3 focuses on evaluation metrics and experimental frameworks, pointing out frequent inconsistencies in sensitivity, specificity, accuracy, and training/testing setups across studies.

Table 4 expands on the challenges in acoustic datasets, quantifying the number of articles addressing or neglecting specific issues, and providing valuable insight into the gaps in the literature. Table 5 continues this focus by detailing the challenges in acoustic data processing, including variability in noise and measurement parameters, as well as the lack of standardized feature extraction. Table 6 offers a quantitative breakdown of AI-related performance metrics, showing disparities in sensitivity, specificity, and accuracy reporting, and highlighting the limitations in clinical applicability.

Table 7 presents calculations of the frequency of some voice-related disorders across populations, offering a broader epidemiological context for the clinical significance of voice analysis. Table 8 examines the frequency of voice-related parameters in genetics-focused studies, emphasizing the limited integration of machine learning approaches. Lastly, Table 9 outlines non-AI-based voice assessment methods used in Parkinson's Disease studies, showcasing the reliance on traditional voice parameters like jitter, shimmer, and harmonics-to-noise ratio, alongside subjective evaluations like the Voice Handicap Index and GRBAS.

A Meta-analysis revealed that several voice parameters including jitter, shimmer, and fundamental frequency variation presented significant deviation from healthy controls. Significant variations of F0, MPT, HNR, were observed but with high heterogeneity between the studies [10].

AI holds substantial potential for the screening and assessment of voice disorders; however, significant challenges remain in terms of dataset quality, software transparency, and clinical validation. Future research should prioritize the establishment of standardized protocols to enhance the clinical applicability of AI in voice disorder diagnosis and treatment.

## Acknowledgment

I am grateful to Vitus Girelli Meiner, MSc in Computer Science.

## References

1. Pedersen M. (2021). Accuracy of Laryngoscopy for Quantitative Vocal Fold Analysis in Combination with AI, A Cohort Study of Manual Artefacts. Scholarly Journal of Otolaryngology, 6(3). https://doi.org/10.32474/sjo.2021.06.000237
2. Kim YEA, Holsinger C, Paderno A, Rau A, Chang M, Crowson M, Curtis J, Ahmad O, Cha DC, Donoho D, Dunham M, Enver N, Habib AR, Johnson A, Kiaer E, Li J, Bur A, Naunheim M, Ni XG, Patel R, Pedersen M, Srivastava R, Thamboo A, Rameau A. (2024). Establishing research priorities for AI implementations in Upper Aerodigestive Tract Endoscopy: a modified Delphi study. POSTER Weill Cornell Medicine Sean Parker Institute for the Voice.
3. Moerman M, Pedersen M. (2024). Voice-Related Biomarkers. Presentation from the commission on voice-related biomarkers at the 2nd joint meeting of the Union of European Phoniatrics (UEP)/European Academy of Phoniatrics (EAP) with the British Laryngology Association (BLA).
4. Lechien, J. R., Geneid, A., Bohlender, J. E., Cantarella, G., Avellaneda, J. C., Desuter, G., Sjogren, E. V., Finck, C., Hans, S., Hess, M., Oguz, H., Remacle, M. J., Schneider-Stickler, B., Tedla, M., Schindler, A., Vilaseca, I., Zabrodsky, M., Dikkers, F. G., & Crevier-Buchman, L. (2023). Consensus for voice quality assessment in clinical practice: guidelines of the European Laryngological Society and Union of the European Phoniatricians. European Archives of Oto-Rhino-Laryngology, 280(12), 5459–5473. https://doi.org/10.1007/s00405-023-08211-6
5. Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., Benninger, M. S., & Newman, C. W. (1997). The Voice Handicap Index (VHI). American Journal of Speech-Language Pathology, 6(3), 66–70. https://doi.org/10.1044/1058-0360.0603.66
6. Sobol, M., Sielska-Badurek, E. M., & Osuch-Wójcikiewicz, E. (2019). Normative values for singing voice handicap index – systematic review and meta-analysis. Brazilian Journal of Otorhinolaryngology, 86(4), 497–501. https://doi.org/10.1016/j.bjorl.2018.12.004
7. Sawaya, Y., Sato, M., Ishizaka, M., Shiba, T., Kubo, A., & Urano, T. (2022). Maximum Pho-

nation Time is a Useful Assessment for Older Adults Requiring Long-term Care/support. Physical Therapy Research, 25(1), 35–40. https://doi.org/10.1298/ptr.e10152

8. Hidaka, S., Lee, Y., Nakanishi, M., Wakamiya, K., Nakagawa, T., & Kaburagi, T. (2022). Automatic GRBAS Scoring of Pathological Voices using Deep Learning and a Small Set of Labeled Voice Data. Journal of Voice. https://doi.org/10.1016/j.jvoice.2022.10.020
9. Pedersen M, Meiner VG. (2023). Overview of Voice Parameters in Parkinson's Disease eventually usable as biomarkers. Webinar in the Biomarkers Committee, 6th of Sep. 2023.
10. Chiaramonte, R., & Bonfiglio, M. (2020). Acoustic analysis of voice in Parkinson's disease: a systematic review of voice disability and meta-analysis of studies, 70(11), 393. https://doi.org/10.33588/rn.7011.2019414