DOI: 10.71568/daga2026.037

Towards a General Framework for Pathologic Voice Rating Across Different Domains

Karolin Krüger[1], Marten J. Finck[2], Steffen Paschen[3], Gerhard Schmidt[1]

[1] *Digital Signal Processing and System Theory, Department of Electrical and Information Engingeering, Kiel University,*

[2] *Visual Computing und Artificial Intelligence, Department of Computer Science, Kiel University,*

[3] *Department of Neurology, Kiel University,*

Kiel, Germany, E-Mail: {kkru, gus}@tf.uni-kiel.de

Abstract

Perceptual evaluation of pathological voice and speech remains a central challenge in both clinical and technological domains, as subjective impressions often influence decisions in diagnostics and training. This work presents a domain-focused approach to design a questionnaire that systematically captures human perception of speech and voice quality. The survey enables the evaluation of various speech dimensions, such as phonation, articulation, and prosody, and is intended to support studies comparing subjective listener impressions with features extracted through automated speech analysis tools. Breaking down the overall voice impression into distinct domains is expected to reflect speakers abilities and specific areas of difficulty. The study investigates whether a concise set of standardized questions is sufficient to consistently and meaningfully capture listeners' impressions across different contexts. The resulting ratings will inform the refinement of the questionnaire and its application as training or reference data for machine learning models aimed at predicting human judgments from acoustic input. This approach lays the groundwork for bridging perceptual and computational assessments of pathological speech, contributing to the development of interpretable, human-aligned evaluation tools. It also provides a basis for examining whether a feature set extracted by automated analysis tools adequately captures all relevant aspects of pathological speaker performance.

Introduction

Perceptual ratings remain central to the evaluation of pathological speech, yet they are often difficult to reproduce consistently across listeners and contexts [1]. Meanwhile, automated speech analysis tools aim to approximate these human impressions using acoustic features. However, the relationship between perceptual dimensions and computational representations is not yet fully understood [2, 3]. This discrepancy highlights the need for a systematic, specialized framework that makes perceptual evaluation easier to understand and more comparable. Such a framework is essential for aligning human judgements with data-driven modelling approaches, and for developing interpretable, clinically meaningful assessment tools.

Therefore, our goal is to develop a structured perceptual questionnaire that systematically captures listeners impressions of speech across distinct dimensions, including phonation, articulation, and prosody. This questionnaire is designed to complement objective acoustic features extracted by our real-time speech analysis tool [4], providing a framework for directly comparing human perception with measurable speech characteristics. While the current work focuses on developing and validating the questionnaire, the resulting data will support machine learning approaches later on by acting as ground truth labels that align perceptual domains with objective features. This will enable more interpretable and clinically relevant automated assessments of impaired speech.

Background and Speech Study

Voice and speech disorders can manifest as a variety of perceptual and acoustic symptoms. To capture these characteristics systematically, speech is described in terms of three complementary domains: phonation, articulation, and prosody. Although respiratory support underlies and influences all three domains, it is not treated as a separate assessment target in this study, as its effects are reflected in, for example, tone duration and therefore phonation [5]. Phonation refers to the production of voiced sound, reflecting the voice's stability, intensity, and resonance. Impaired phonation can result in reduced loudness, an unstable or breathy voice, a rough timbre, and a limited pitch range, for example [5, 6]. Articulation involves the precise formation of speech sounds and affects intelligibility. Difficulties with articulation can result in slowed or irregular speech, as well as imprecise pronunciation. These issues can have a significant impact on how well a speaker is understood [5, 6]. Key elements of prosody include, for example, intonation, speech rate, and rhythm, all of which may be affected by voice and speech disorders [5]. Evaluating these domains separately makes it possible to describe and quantify the different aspects of impaired speech in a structured way. This supports both perceptual assessment and subsequent comparison with objective acoustic measures.

To perform the perceptual assessment, our study *Study on the automated examination of voice, speech, and associated motor skills* collects speech data in collaboration with the Neurology Department of the University Hospital Schleswig-Holstein (UKSH) in Kiel [7]. The dataset includes recordings from individuals with Essential Tremor (ET) and Parkinsons Disease (PD), performing a variety of speech tasks capturing controlled and spontaneous speech. These tasks include sustained vowels (all five vowels, repeated twice), basic articulation exercises such as *dadada*, *gagaga*, and *bababa*, three different rhyme sequences based on [8], reading of short pho-

netically balanced texts (twice) [9], and answering two questions in two to three sentences to obtain samples of natural, spontaneous speech. Additional movement data in different seating positions was recorded using the Delsys Trigno Avanti surface EMG sensor system, containing EMG and acceleration data [10].

All participants were German speakers (21 native, participants 1 and 6 non-native). The ongoing study includes 23 participants (8 women, 15 men; mean age 67.7 ± 12.3), comprising 15 ET (5 women, 10 men) and 9 PD (4 women, 5 men) patients (one with both). In addition to sensor recordings, we collect demographic information, disease onset and medication data, as well as perceptual measures such as a tremor index and the Voice Handicap Index (VHI) in German [11], which provides standardized information about the patients self-perceived voice-related limitations. Figure 2 presents the VHI results for 22 patients, as data from one patient (23) could not be included due to technical difficulties. Based on the VHI scores, experts evaluated a subset of speakers to identify suitable examples for the explanations used in the survey.

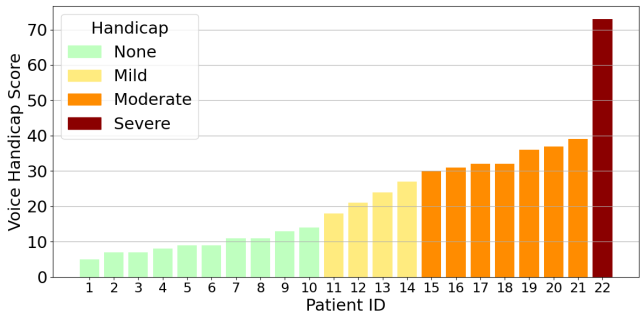

Figure 1: Voice handicap index results.

During the recordings, the patients were seated in front of a Microsoft Surface Pro 6 tablet while their voices were recorded using a Fox microphone by beyerdynamic [12, 13]. In addition, video recordings were obtained. The distance between the patient and the microphone was kept constant at 30 cm throughout the recording. The recording setup is illustrated in Figure 2.

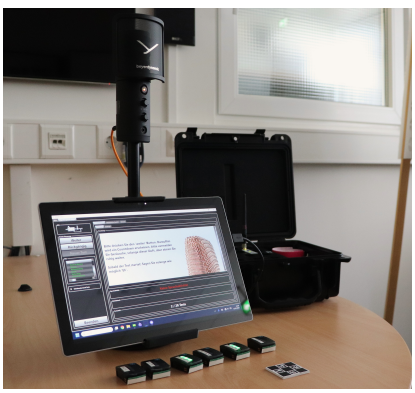

Figure 2: Measurement setup including the microphone, tablet and Delsys Trigno Avanti surface EMG sensor system.

Survey Design

As mentioned before, the evaluation of speech recordings is organized into three sections, each focusing on a specific speech domain: phonation, articulation, and prosody. Raters are not explicitly provided with definitions of these domains. Instead, each domain is assessed through two targeted questions designed to capture the relevant perceptual aspects, as described in Table 1. A supplementary question is also included to help identify the most informative speech task for each domain and patient. Each patient recorded 24 speech files, including

Table 1: Survey questions depending on the voice domain.

Domain	Questions
Phonation	1. How steady or shaky does the voice sound to you? (*Stability*)
	2. How consistently can the voice quality be maintained over time? Please consider the entire duration. (*Consistency*)
Articulation	3. How easy was it to understand the syllables or content? (*Intelligibility*)
	4. How clear or muffled/blurry was the pronunciation? (*Clarity*)
Prosody	5. How lively or monotonous does the voice sound to you? (*Dynamics*)
	6. How natural does the speech flow sound? (Please consider tempo, pauses, and stress, ignoring accents.) (*Naturalness*)

multiple variations of a speech task. To maximize data utilisation, these files are split into two separate subsets per patient. These subsets contain randomized, balanced stimuli across all domains and tasks, ensuring full coverage of the dataset while avoiding reusing the same voice recordings. To maintain the evaluator's concentration, the full questionnaire is divided into smaller subsurveys inhereting speech files in all domains of four patients. Each subsurvey is lasting about 20 to 25 minutes.

The questionnaire is structured so that, initially, several speech files are presented to allow evaluators to adjust the volume of their headphones to a comfortable but rather loud level. This is depicted as the first section in Figure 3. This is followed by an explanation page, describing the upcoming questions and showing examples of good, medium, and bad speakers (17, 20, and 22) for this domain, along with mean ratings from four experienced evaluators (knowledge in logopedics and neurology) to train inexperienced raters. After this, the corresponding speech samples are presented for evaluation. This process is repeated for each domain, as shown in the three blocks in Figure 3. Finally, demographic data including age, gender, native language, and hearing impairment, as well as feedback of the evaluators are collected. Each evaluator assigns an anonymous code to their responses

to ensure data linkage if they complete multiple subsurveys.

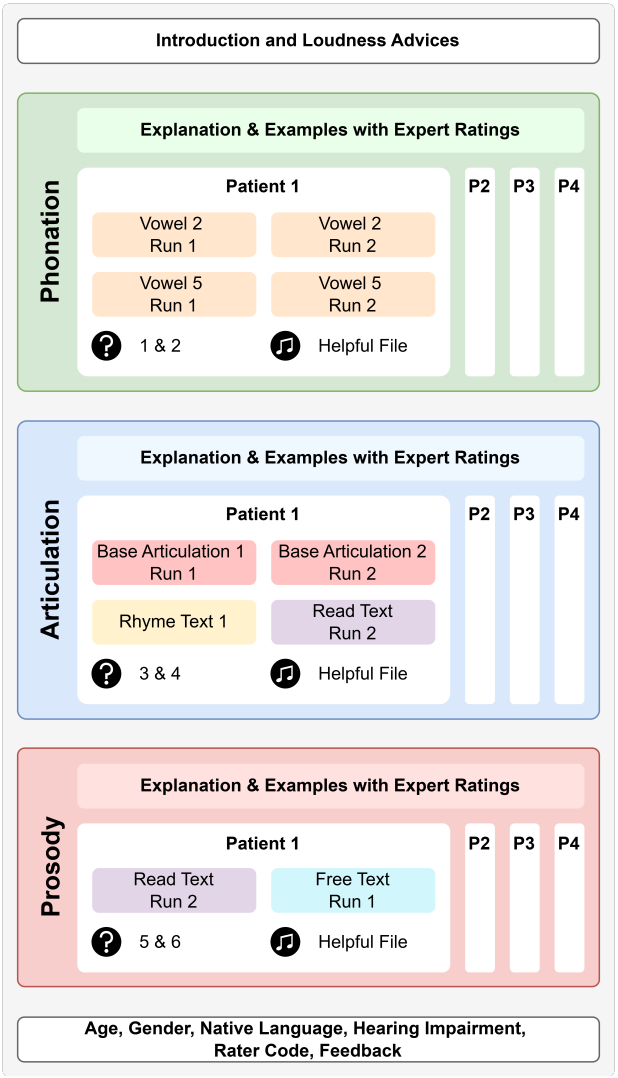

Figure 3: Concept of the perceptual survey design.

The rating scale ranges from 1 to 6, with 6 always representing the best score. For the first phonation question, for example, 1 corresponds to *very shaky* and 6 to *very steady*. The decision to use a six-point scale was made in line with the speech index of the extended NTID scale [14]. This format was chosen to provide sufficient sensitivity to detect clinically relevant differences, while also avoiding a neutral midpoint.

Evaluation and Discussion

Each subsurvey was completed by 14 to 17 naive listeners (23 unique raters in total) without prior training in linguistics or related fields, capturing unbiased perceptual judgments of speech quality. We analyzed the relationship between the two questions across domains and compared ratings with patients' assessment using the VHI. To evaluate reproducibility, a subset of four patients from the first subsurvey was assessed again with different recordings and evaluators (partly). Evaluators were not informed that these speakers had been previously assessed, and there was at least one hour of time between evaluating different subsurveys, minimizing recognition bias.

Outliers were defined as raters whose ratings deviated by more than two standard deviations from the mean in over 15 % of cases. Interrater variability was substantial across domains. In the phonation domain, standard deviations (SD) were 0.6 for *Stability* and 0.8 for *Consistency* (SD Phonation: 0.9), as shown in Figure 4, where scores are shown in ascending order. Large discrepancies were observed for several patients, particularly patients 7 and 2, who exhibited pronounced voice tremor. However, additional medical data captured only general tremor and not voice-specific tremor. In the articulation domain, *Intelligibility* and *Clarity* showed similar variability (SDs 0.9 and 0.8; SD Articulation: 0.9), but ratings of both questions were more consistent (Figure 5), suggesting that a single item may suffice. In contrast, *Dynamics* and *Naturalness* showed comparable variability (SDs 0.8 and 0.8; SD Prosody: 0.9) but greater divergence between questions across patients (Figure 6).

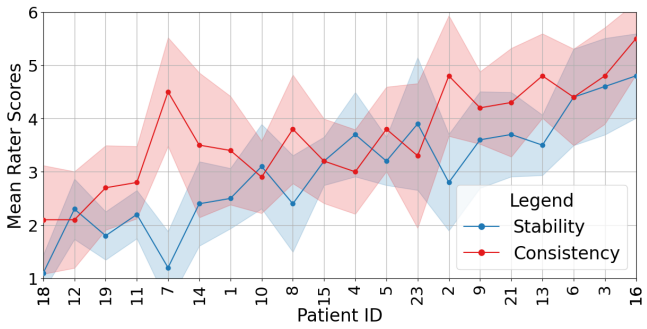

Figure 4: Results of phonation domain rating.

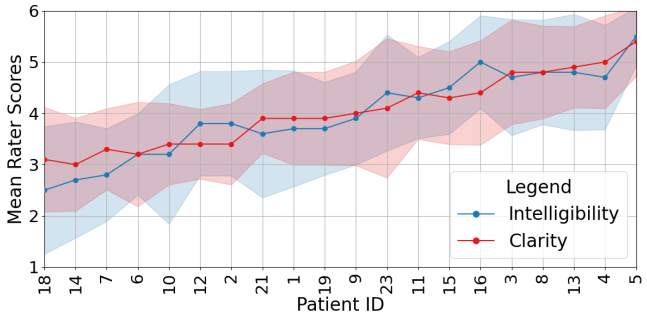

Figure 5: Results of articulation domain rating.

No meaningful correlation was observed between mean perceptual ratings and overall VHI scores: for example speakers with moderate handicap displayed medium to high scores of nearly 3 to 5, exemplarily shown for the articulation domain in Figure 7. Repeated assessments of the same speakers with different spech file sets showed small score variations (0.2 - 0.4), indicating good reproducibility. Overall, perceptual ratings can reliably differentiate between good, moderate, and poor speech quality in pathological speakers, despite high interrater variability reflecting differing rater opinions. Increasing the number of raters and analyzing demographic data of speakers

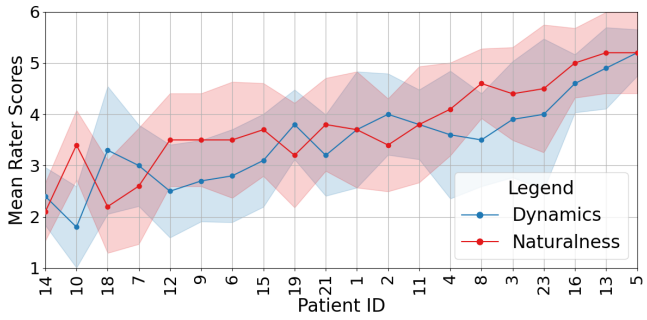

Figure 6: Results of prosody domain rating.

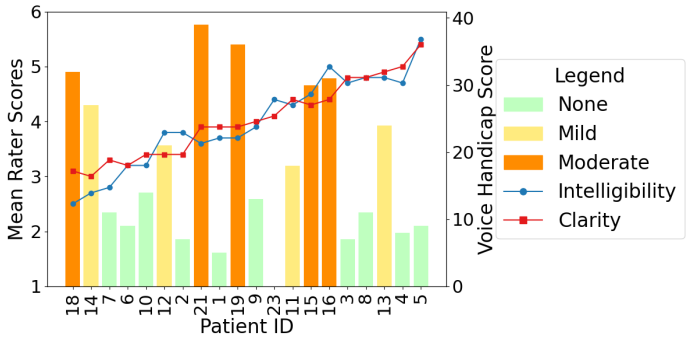

Figure 7: Results of articulation domain rating, including the VHI of all speakers without the example speakers.

and raters, as well as rating behaviors could improve reliability and help identify outlier raters. The rating scale was often not fully utilized, suggesting potential for refinement or benchmarking with healthy speakers.

Summary and Outlook

In this work, a database containing various speech tasks together with additional synchronous movement data was established. Furthermore, a survey was developed to evaluate three voice domains. The results indicate that perceptual ratings can reliably differentiate between good, medium, and poor voice quality. A relationship between speakers self-assessment and the external voice ratings could not be found.

Future work will focus on expanding the survey to include a larger number of speakers and raters, thereby enabling more demographic analyses. In addition, the correlations will be examined in greater detail, and perceptual ratings will be linked to acoustic features using machine learning methods. Feature importance analyses will also be conducted to assess the contribution of objective acoustic parameters to the perception of pathological speech.

Acknowledgments

We sincerely thank our expert raters, PD Dr. Steffen Paschen, Janika Krafthöfer, Katrin Sanne, and Dr. Isabel S. Schiller, for their invaluable evaluations. We also gratefully acknowledge the contributions of all other raters who participated in this study.

References

[1] W. Ariyanti, K.-Y. Chen, S. M. Siniscalchi, H.-M. Wang, and Y. Tsao, Towards Robust Assessment of Pathological Voices via Combined Low-Level Descriptors and Foundation Model Representations, *IEEE Journal of Biomedical and Health Informatics*, 2025. doi: 10.1109/JBHI.2025.3644692.

[2] M. J. Finck, K. Krüger, S. Paschen, A. Helmers, and G. Schmidt, Analyzing Impaired Speech in Context of Magnetic Resonance-guided Focused Ultrasound Using Convolutional Neural Networks," in *Proc. DAGA*, Hamburg, Germany, 2024. urn: urn:nbn:de:gbv:8:3-2024-00700-7.

[3] O. P. Wischhoff, V. Gouraram, T. J. Chumbley, B. Liu, and J. J. Jiang, Auditory-Perceptual Validation of Acoustic Chaos Parameters in Healthy and Dysphonic Voices, *J Speech Lang Hear Res.*, 2025, doi: 10.1044/2025_JSLHR-25-00155.

[4] K. Krüger, P. Piepjohn, and G. Schmidt, A Real-time Objective Speech Analysis Tool for Analysis of Impaired Speech, in *Proc. DAGA*, Germany, 2025.

[5] B. Schneider and W. Biegenzahn, *Stimmdiagnostik: Ein Leitfaden für die Praxis*, Wien: Springer-Verlag, 2007.

[6] S. S. Hammer and A. Teufel-Dietrich, *Stimmtherapie mit Erwachsenen*, Berlin: Springer, 2017. ISBN-10: 3-662-53976-4 / 3662539764.

[7] Universitätsklinikum Schleswig-Holstein. Neurologie Kiel. Available: https://www.uksh.de/neurologie-kiel/

[8] E. v. Wallenberg and B. Kollmeier, Reimtest in deutscher Sprache: Erstellung und Evaluation von Testlisten, *Audiologische Akustik*, 2/1989.

[9] K. Fellbaum, Hörversuche zur Beurteilung der Sprachqualität von Sprachsynthesesystemen für die deutsche Sprache, in *Proc. DAGA*, 1994.

[10] Delsys Incorporated. Trigno Wireless Biofeedback System Users Guide, 2021.

[11] Deutsche Gesellschaft für Phoniatrie und Pädaudiologie (DGPP). Voice Handicap Index (VHI), Deutsche Version, 2006. Available: https://dgpp.de/de/wp-content/files/vhi-dt_2006.pdf

[12] Microsoft Surface Pro 6. Available: https://support.microsoft.com/de-de/surface/surface-pro-6-spezifikationen-und-features-ade5cfc2-e99a-6fd1-abbe-c0e8a8a3942d

[13] Beyerdynamic. FOX Mikrofon. Available: https://www.beyerdynamic.de/p/fox?srsltid=AfmBOopJoGvxcGh-KoEpXo7pfugraC99-YzTkTgJZpIqWyo2XlVFjA7W

[14] D. K. Raymond, *Intelligibility in Speech Disorders - Theory, Measurement and Management*, Amsterdam/Philadelphia: John Benjamins Publishing Company, 1992.