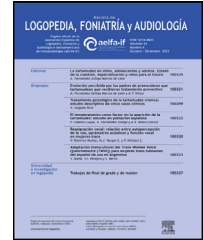




Revista de  
**LOGOPEDIA, FONIATRÍA y AUDIOLOGÍA**

[www.elsevier.es/logopedia](http://www.elsevier.es/logopedia)



ORIGINAL ARTICLE

# User profiles and auditory-perceptual evaluations upon the launch of the All-Voiced app for voice evaluation: Initial insights and training potential



Neus Calaf

Department of Basic, Developmental and Educational Psychology, Autonomous University of Barcelona, Spain

Received 8 October 2024; accepted 10 December 2024

Available online 13 January 2025

## KEYWORDS

Auditory-perceptual evaluation;  
Voice disorders;  
CAPE-V;  
All-Voiced app;  
Voice assessment training

## Abstract

**Introduction:** Auditory-perceptual evaluation is key for diagnosing voice disorders, but variability in judgments underscores the need for improved training materials. The All-Voiced app was developed to enhance consistency in evaluations through real-time feedback and data-driven training.

**Objective:** To present preliminary findings from the first 75 days following the launch of the latest version of the app (September 2024), focusing on user profiles and voice evaluations, exploring the app's potential for deeper insights as more data is gathered.

**Method:** The latest version of the All-Voiced app, launched in September 2024, includes a fully integrated backend for data collection from users who provide consent. The app enables users to practice voice evaluations, receive feedback, and contribute to research. Descriptive statistics were used to analyze user profiles and evaluations from the first 75 days post-launch of this latest version. Box plots and scatter plots were used to compare the evaluations of All-Voiced users with PVQD ratings (Walden, 2022) across different competence levels.

**Results:** A total of 264 participants registered in the app, with daily registration patterns showing consistent activity. Most participants were aged 21–30 (49%), identified with “she/her” pronouns (88%), and were from the United States (51%). The majority were speech-language pathologists (78%), and 40% were beginners in terms of competence level. A total of 557 evaluations were collected across 112 voice samples. Analysis of the three most frequently evaluated voice samples revealed distinct patterns of consensus among evaluators with varying levels of expertise, hinting at trends that could have significant implications as more data is collected.

E-mail address: [ncalaf@gmail.com](mailto:ncalaf@gmail.com)

<https://doi.org/10.1016/j.rlfa.2024.100511>

0214-4603/© 2024 Elsevier España, S.L.U. y Asociación Española de Logopedia, Foniatría y Audiología e Iberoamericana de Fonoaudiología. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

*Conclusion:* The All-Voiced app has been well received, particularly by speech-language pathologists, highlighting its promise as a tool for auditory-perceptual training. By collecting and analyzing large-scale data, the app holds potential to address limitations in the subjective nature of evaluations, enhance reliability, and support evidence-based practices in voice disorders management.

© 2024 Elsevier España, S.L.U. y Asociación Española de Logopedia, Foniatría y Audiología e Iberoamericana de Fonoaudiología. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

## PALABRAS CLAVE

Evaluación  
auditivo-perceptiva;  
Trastornos de la voz;  
CAPE-V;  
Aplicación All-Voiced;  
Formación en  
evaluación de la voz

## Perfiles de usuario y evaluaciones auditivo-perceptivas tras el lanzamiento de la aplicación All-Voiced para la evaluación de la voz: perspectivas iniciales y potencial formativo

### Resumen

*Introducción:* La evaluación auditivo-perceptiva es fundamental para el diagnóstico de los trastornos de la voz, pero la variabilidad en los juicios subraya la necesidad de mejorar los materiales de formación. La aplicación All-Voiced fue desarrollada para potenciar la consistencia en las evaluaciones a través de la retroalimentación en tiempo real y la formación basada en datos.

*Objetivo:* Presentar los hallazgos preliminares de los primeros 75 días tras el lanzamiento de la última versión de la aplicación (septiembre de 2024), enfocándose en los perfiles de los usuarios y las evaluaciones de la voz, explorando el potencial de la aplicación para obtener las perspectivas más profundas a medida que se recopilan más datos.

*Método:* La última versión de la aplicación All-Voiced, lanzada en septiembre de 2024, incluye un *backend* completamente integrado para la recolección de datos de los usuarios que otorgan su consentimiento. La aplicación permite a los usuarios practicar evaluaciones de la voz, recibir la retroalimentación y contribuir a la investigación. Se utilizaron estadísticas descriptivas para analizar los perfiles de los usuarios y las evaluaciones de los primeros 75 días después del lanzamiento de esta última versión. Se emplearon diagramas de cajas y gráficos de dispersión para comparar las evaluaciones de los usuarios de All-Voiced con las calificaciones de PVQD (Walden, 2022) en diferentes niveles de competencia.

*Resultados:* Un total de 264 participantes se registraron en la aplicación, con patrones de registro diarios que mostraban actividad consistente. La mayoría de los participantes tenían entre 21 y 30 años (49%), se identificaban con el pronombre «ella» (88%), y eran de EE. UU. (51%). La mayoría eran logopedas (78%), y el 40% eran principiantes en términos de nivel de competencia. Se recopilaron un total de 557 evaluaciones a través de 112 muestras de voz. El análisis de las tres muestras de voz más evaluadas reveló patrones distintos de consenso entre los evaluadores con diferentes niveles de experiencia, insinuando tendencias que podrían tener implicaciones significativas a medida que se recopilen más datos.

*Conclusión:* La aplicación All-Voiced ha sido bien recibida, especialmente por logopedas, destacando su potencial como herramienta para el entrenamiento auditivo-perceptivo. A través de la recopilación y el análisis de datos a gran escala, la aplicación tiene el potencial de abordar las limitaciones inherentes a la naturaleza subjetiva de las evaluaciones, mejorar la fiabilidad y apoyar prácticas basadas en la evidencia en el manejo de los trastornos de la voz.

© 2024 Elsevier España, S.L.U. y Asociación Española de Logopedia, Foniatría y Audiología e Iberoamericana de Fonoaudiología. Se reservan todos los derechos, incluidos los de minería de texto y datos, entrenamiento de IA y tecnologías similares.

## Introduction

Auditory-perceptual evaluation is a critical diagnostic tool in assessing and documenting voice disorders (Barsties & De Bodt, 2015; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Oates, 2009; Roy et al., 2013). Despite its widespread

use, this method faces significant challenges, particularly due to variability influenced by both the listener's perception and the characteristics of the vocal stimulus (Kreiman, Vanlancker-Sidtis, & Gerratt, 2003). To address these challenges and improve the reliability of judgments, the CAPE-V (Consensus Auditory-Perceptual Evaluation of Voice) was

developed as a standardized tool for consistently evaluating voice quality across clinical settings (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009).

To ensure broader applicability across diverse populations, the CAPE-V has been adapted into multiple languages, including Brazilian Portuguese (Behlau, 2004; Behlau, Rocha, Englert, & Madazio, 2022), European Portuguese (de Almeida, Mendes, & Kempster, 2019; Jesus et al., 2009), Persian (Salary Majd et al., 2014), Italian (Mozzanica, Ginocchio, Borghi, Bachmann, & Schindler, 2014), Spanish (Núñez-Batalla, Morato-Galán, García-López, & Ávila-Menéndez, 2015), Mandarin (Chen et al., 2018), Turkish (Ertan-Schlüter, Demirhan, Ünsal, & Tadihan-Özkan, 2020; Özcebe, Aydinli, Tigrak, İncebay, & Yilmaz, 2019), Kannada (Gunjawate, Ravi, & Bhagavan, 2020), Hindi (Joshi, Baheti, & Angadi, 2020), Japanese (Kondo et al., 2021), Tamil (Venkatraman, Mahalingam, & Boominathan, 2022), Malay (Mohd Mossadeq, Mohd Khairuddin, & Zakaria, 2022), Catalan and Spanish (Calaf & Garcia-Quintana, 2024), and European French (Pommée, Mbagira, & Morsomme, 2024). These adaptations allow for greater cultural and linguistic relevance while maintaining the tool's core objective of delivering reliable perceptual assessments. The language-specific versions have significantly expanded the CAPE-V's usability, making it accessible to clinicians and researchers worldwide.

However, while these adaptations extend the CAPE-V's reach to a wider range of populations, the reliability of auditory-perceptual judgments remains a key concern. Ensuring strong inter- and intra-rater reliability is critical to the tool's effective application in diverse contexts.

### Inter-rater reliability of the auditory-perceptual judgements

Inter-rater reliability—which measures the consistency of evaluations between different raters—has generally been reported as high for the CAPE-V across various languages and vocal attributes. For instance, the original English version demonstrated strong inter-rater reliability for overall severity (Karnell et al., 2007). Adaptations of the CAPE-V have also shown high inter-rater reliability in certain cases. The European Portuguese version (Jesus et al., 2009) exhibited excellent inter-rater reliability for breathiness and loudness, while the Kannada version (Gunjawate et al., 2020) displayed consistently high reliability across all attributes. Similarly, the Italian adaptation (Mozzanica et al., 2014) reported high inter-rater reliability for overall severity and breathiness, and the Turkish version (Özcebe et al., 2019) demonstrated strong reliability for overall severity and loudness. The Japanese adaptation (Kondo et al., 2021) also showed high inter-rater reliability across all attributes. Finally, the European French version (Pommée, Shanks, Morsomme, Michel, & Verduyck, 2024) demonstrated good reliability for overall severity.

However, not all attributes performed consistently across studies. The original English version exhibited low inter-rater reliability for strain, pitch, and loudness (Zraick et al., 2011). Likewise, the Persian (Salary Majd et al., 2014), Brazilian Portuguese (Behlau et al., 2022), and Turkish versions (Ertan-Schlüter et al., 2020) demonstrated low

inter-rater reliability for pitch and loudness. Roughness also showed variability, with lower inter-rater reliability reported in the English (Kelchner et al., 2010; Zraick et al., 2011), Brazilian Portuguese (Behlau et al., 2022), Tamil (Venkatraman et al., 2022), and Hindi (Joshi et al., 2020) versions. The Tamil version further showed lower inter-rater reliability for pitch and strain. Additionally, the bilingual Catalan/Spanish version (Calaf & Garcia-Quintana, 2024), inter-rater reliability for breathiness and loudness was notably low, further highlighting the challenge of achieving consistency in perceptual voice assessments. Similarly, the European French version (Pommée, Shanks, et al., 2024) showed low inter-rater reliability for most parameters except overall severity,

### Intra-rater reliability of the auditory-perceptual judgements

Intra-rater reliability—which measures the consistency of the same evaluator over time—has generally been reported as high for the CAPE-V across different languages and vocal attributes, though some variability persists. For instance, the original English version demonstrated strong intra-rater reliability for overall severity (Karnell et al., 2007). Subsequent adaptations, such as the Italian version (Mozzanica et al., 2014), consistently showed high intra-rater reliability across attributes like overall severity, roughness, and breathiness. Similarly, the Spanish (Núñez-Batalla et al., 2015), Turkish (Özcebe et al., 2019), Japanese (Kondo et al., 2021), and Kannada versions (Gunjawate et al., 2020) displayed strong intra-rater reliability across most attributes. Finally, the European French version (Pommée, Shanks, et al., 2024) also demonstrated good intra-rater reliability for overall severity, strain, and pitch.

However, some studies have reported lower intra-rater reliability in specific dimensions. The English version exhibited lower intra-rater reliability for strain (Zraick et al., 2011). Likewise, the Persian version (Salary Majd et al., 2014) showed reduced intra-rater reliability for pitch and loudness, while the Brazilian Portuguese version (Behlau et al., 2022) also reported lower intra-rater reliability for breathiness and loudness. The Tamil version (Venkatraman et al., 2022) demonstrated moderate intra-rater reliability for pitch and strain, while the bilingual Catalan/Spanish version (Calaf & Garcia-Quintana, 2024) reported particularly low intra-rater reliability for loudness, highlighting the ongoing need for tools and strategies that can help evaluators achieve more consistent auditory-perceptual judgments.

### The need for training materials development

The variability in auditory-perceptual judgments underscores the need for further standardization and the development of reference materials to improve reliability. Efforts such as the simulations created by the University of Wisconsin-Madison (Connor, Bless, Dardis, & Vinney, 2008), the Perceptual Voice Qualities Database (Walden, 2022), and the recent anchor and training sample set for auditory-perceptual voice evaluation (Labaere, De Bodt, & Van Nuffelen, 2023) aim to reduce this variability by provid-

ing structured materials that help standardize perceptual judgments.

The University of Wisconsin-Madison simulations consist of 45 cases, each including CAPE-V sentence samples, conversational speech, and sustained vowels /a/ and /i/. Each case also provides a detailed patient history, which can be used separately for class discussions or reflections on how medical conditions or symptoms might influence a patient's voice and what further assessments are needed for a complete voice evaluation. While the quality of this material is excellent, the application of the CAPE-V remains on paper, is not digitized, does not collect data, and offers limited feedback, as it only allows users to compare their evaluations to a single reference.

The Perceptual Voice Qualities Database (Walden, 2022) offers a greater potential, with each sample being evaluated twice by 3–4 evaluators. However, the raw data from these evaluations require processing to be effectively used in training programs. Unlike the simulations, this resource presents a larger dataset and the possibility of deeper analysis, but its full integration into educational settings remains dependent on further development to transform the raw data into actionable training materials.

The anchor and training sample set developed by Labaere et al. (2023) uses the GRBAS scale (Hirano, 1981), which, while still widely used, is now superseded by the standardized CAPE-V. The CAPE-V is internationally recognized for its standardized protocol for sample collection, refined vocal attributes, and use of a visual analog scale (VAS), allowing parametric statistical tests and improving reliability across clinical settings. A strength of Labaere et al.'s material is the strict reliability criterion for anchor samples, requiring 90% expert agreement on at least one attribute. However, relying on the GRBAS scale limits its alignment with more up-to-date practices like the CAPE-V, which offers better standardization in auditory-perceptual evaluations and allows for a more nuanced assessment than a four-category scale.

In response to these challenges, the All-Voiced app (Calaf, 2024b) was developed to provide a more accessible, scalable, and data-driven platform for both training and research in voice evaluation. The app, which is currently free to use, focuses on improving the consistency and accuracy of auditory-perceptual ratings by allowing users to practice their evaluations, receive real-time feedback, and compare their assessments with those of previous evaluators. This feedback system, combined with the app's interactive features, is designed to enhance reliability in both inter-rater and intra-rater assessments, addressing key gaps in the field.

The All-Voiced app is designed not only as a training tool but also as a research platform, with the goal of gathering large volumes of perceptual data from users with varying levels of expertise. This data collection is intended to support ongoing research aimed at refining auditory-perceptual voice evaluation techniques and identifying patterns that could inform the development of more targeted training materials and clinical tools in the future.

The primary aim of this study is to present preliminary findings on the user profiles and perceptual judgments collected during the first 75 days following the launch of the latest app All-Voiced app version. Specifically, this study

analyzes the user registration patterns and voice evaluations of the most frequently assessed samples, and highlights the potential of the app to provide valuable insights into the perception of vocal qualities as more data is collected over time.

## Method

### App description

#### Development and versions

The All-Voiced app (Calaf, 2024b) has been developed iteratively, with significant milestones marking its progression. The app is built using modern web technologies, with React for the frontend and Node.js with Express for the backend, ensuring a responsive and scalable user experience. The first version of the app was tested from January to April 2024 with smaller groups of users, including students and participants in research projects led by the author.

On April 16th, 2024, coinciding with World Voice Day, the app was publicly launched with limited functionalities. A major update followed on May 31st, 2024, with the release of CAPE-V resources, published after obtaining permission from ASHA, and aligned with the 53rd Annual Symposium of the Voice Foundation. Notably, the All-Voiced app was cited during the special session *Describing Voices*, moderated by Nancy Solomon (Nagle, 2024, May 31).

The most recent version of the app, which is the focus of this study, was launched on September 4th, 2024. This version includes a fully integrated backend system allowing users to create accounts and contribute to science and education by sharing their evaluation data. This version was presented on September 6th at the Voice Lab of the Pan-European Voice Conference (PEVOC) in Santander, Spain (Calaf, 2024a, September 6).

#### User roles and access

The app can be accessed partially without the need for a user account, allowing users to explore some features freely. However, creating an account unlocks additional functionalities. Upon registration, users can choose from four options to declare their voice competence level:

1. **Beginner:** Limited or no experience in the field of voice.
2. **Intermediate:** Some practical experience with a basic understanding of voice concepts.
3. **Advanced:** Good experience in voice, able to handle most cases independently.
4. **Expert:** Extensive experience in voice, recognized as a specialist in the field.

All users, when registering, can choose to agree that their data may be used in research projects, in accordance with the app's Privacy Policy. For users who declare their voice competence level as "Advanced" or "Expert," an additional option is also available: they can choose to share their evaluations anonymously with the community, providing valuable feedback to other users. Before sharing each evaluation, these users are prompted to give explicit consent, ensuring that they consider that specific evaluation valuable as feedback for others.

### Features and functionalities

All registered users have access to the full range of resources available in the app, regardless of their declared competence level. These resources include:

- **Resources for Autonomous Training:** These resources allow users to practice voice evaluations independently. After completing each evaluation, users receive feedback in the form of comparison data between their own evaluations and those provided by other evaluators. The samples in this section are presented randomly from a pre-existing database, allowing users to gain experience with diverse voices.
- **Resources for Teaching and Learning:** Instructors choose specific voice samples for students to evaluate, guiding the training toward particular learning goals. These resources create an interactive and engaging environment, where participants can compare and discuss their evaluations in group sessions, encouraging thoughtful reflection. Students also have access to home training, where they evaluate instructor-selected samples and record their results in the required format, reinforcing the skills learned in class.
- **Resources for Research:** This section is dedicated to gathering essential resources and offering collaborations and personalized services for auditory-perceptual evaluation studies. In contrast to the other two resource types, users in the research section utilize their own voice samples, contributing original data to studies, rather than drawing from the pre-existing database.

The app's comprehensive set of resources ensures that users at all levels—whether beginners or experts—can benefit from structured learning, autonomous training, or active contributions to research. By integrating opportunities for feedback, collaboration, and data sharing, the app serves as a valuable platform for advancing skills in auditory-perceptual evaluation and promoting continuous improvement in the field.

### Vocal samples

In this first phase of the app implementation, 120 vocal samples were selected from the Perceptual Voice Qualities Database (PVQD) (Walden, 2022) for use within the app. This selection was carefully curated to ensure a balanced representation of age groups, gender, diagnoses, and varying severities of dysphonia, providing a comprehensive dataset for auditory-perceptual evaluations. To standardize the listening conditions, all vocal samples were normalized to 70 dB, ensuring a consistent loudness across evaluations. Additionally, the sociodemographic labels were standardized—for example, unifying variations such as "f", "F", "female", and "Female" under a single label for gender, and harmonizing diagnosis terms to improve consistency and comparability across different groups.

### Linguistic adaptations

To ensure accessibility across different regions, the app supports multiple languages. All versions of the CAPE-V

available within the app have been included with permission from the authors and copyright holders. The currently available CAPE-V versions are English (Kempster et al., 2009), Spanish (Calaf & Garcia-Quintana, 2024), Catalan (Calaf & Garcia-Quintana, 2024), French (Pommée, Mbagira, et al., 2024), and Japanese (Kondo et al., 2021). The entire app has been translated into these languages, ensuring comprehensive accessibility. To ensure cultural and linguistic appropriateness, the app includes a feedback form in the footer, inviting users to contribute to improving the language adaptations.

At the time of writing this manuscript, negotiations are ongoing with the authors to include additional language adaptations of the CAPE-V.

### Study design

#### Participants

The study involved 264 registered users of the All-Voiced app who consented to the use of their data for research purposes. Participants come from diverse professional backgrounds, including speech-language pathologists, SLP students, and other voice-related professionals. They also represent a wide range of experience levels in voice evaluation, categorized as beginner, intermediate, advanced, and expert.

Only users who provided explicit consent for research participation are included in this study. All evaluations were analyzed anonymously to ensure privacy and compliance with ethical standards. No identifying information was used in either the analysis or reporting of results.

#### Instruments

Auditory-perceptual voice evaluations were conducted using the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) (Kempster et al., 2009). The CAPE-V was integrated into the app, enabling users to evaluate vocal attributes such as overall severity, roughness, breathiness, strain, pitch, and loudness, along with other specific vocal characteristics like diplophonia, vocal fry, and hypernasality.

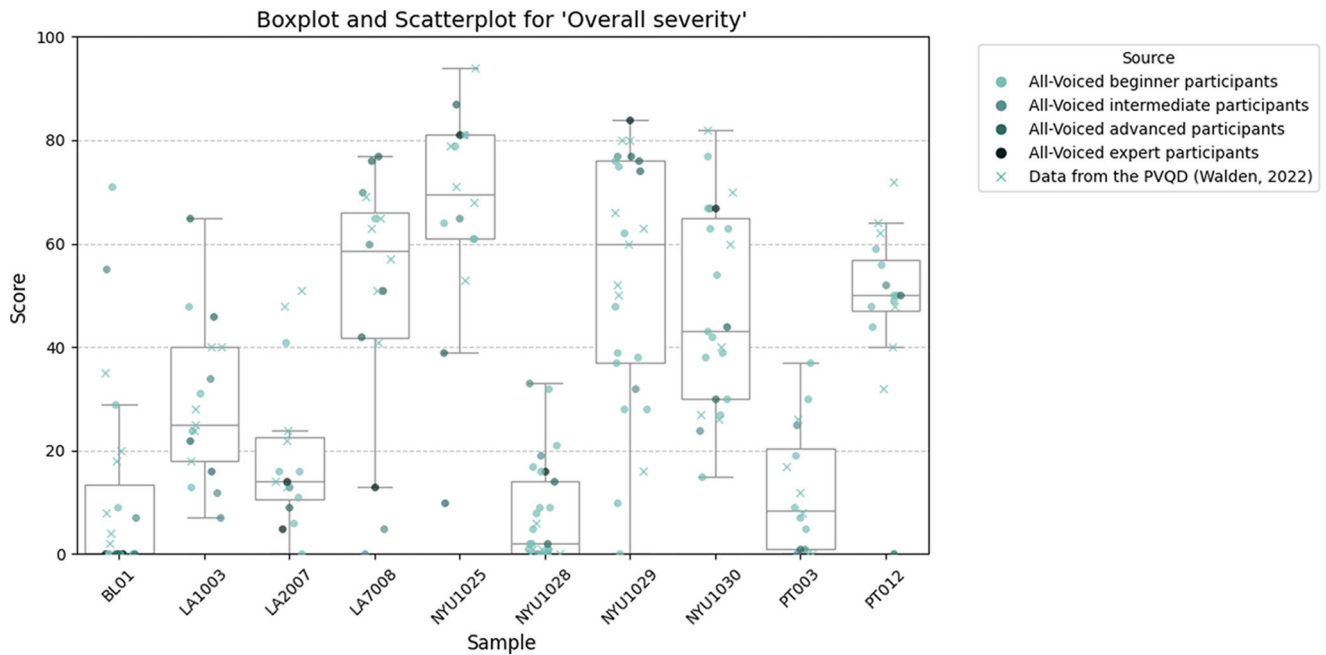
#### Procedure

Upon registration in the All-Voiced app, users are prompted to provide consent for their data to be used for research. Only those who agree to participate are included in the study. Evaluations are submitted during regular app usage, either for personal practice or classroom assignments, and are securely stored for analysis. Data analysis focuses on evaluations submitted during the first 75 days following the launch of the most recent version of the app on September 4th, 2024.

#### Statistical analysis

The statistical analyses and visualizations were conducted using Python (version 3.10.12) in a Google Colab environment. Key libraries included pandas (version 2.2.2) for data manipulation, matplotlib (version 3.8.0) and seaborn (version 0.13.2) for visualizations.

For the participants' sociodemographic data, descriptive statistics were calculated to summarize the distribution



**Figure 1** Distribution of overall severity scores for samples with 10 or more unique all-voiced raters, showing expertise levels of all-voiced participants and PVQD data. *Note.* PVQD = Perceptual Voice Qualities Database (Walden, 2022).

of age, gender pronouns, country of residence, language preference, profession, and voice competence levels. The results are presented in counts and percentages.

To analyze the gathered auditory-perceptual evaluations, scatter plots and box plots (Figs. 1–4) were created to compare the evaluation results from All-Voiced users across different voice competence levels (beginner, intermediate, advanced, and expert) with those of the Perceptual Voice Quality Database (PVQD) by Walden (2022) for standard perceptual attributes such as Overall Severity, Roughness, Breathiness, and Strain.

In this comparison, all available evaluations from the PVQD database were utilized, as each rater provided two independent assessments per sample. This approach ensured the maximum use of the limited data available from this source (3–4 raters per sample). In contrast, for the All-Voiced database, only one evaluation per rater was included to avoid potential biases from raters contributing multiple assessments, given the larger number of raters in this dataset. This methodological choice balanced the differences in dataset structures while preserving the integrity and richness of the PVQD data.

The evaluations from All-Voiced users were visually compared to the PVQD expert ratings, with voice competence levels distinguished by color. For each voice sample and attribute, variability across evaluators was illustrated. Cut-off values for overall severity, as established by Calaf and Garcia-Quintana (2024), were applied to facilitate the interpretation of the data. In addition to the visual analyses, descriptive statistics were calculated to further explore the evaluation data for voice samples, enabling the analysis of both standard perceptual attributes and less commonly evaluated features such as pitch, loudness, hypernasality, and fry.

## Results

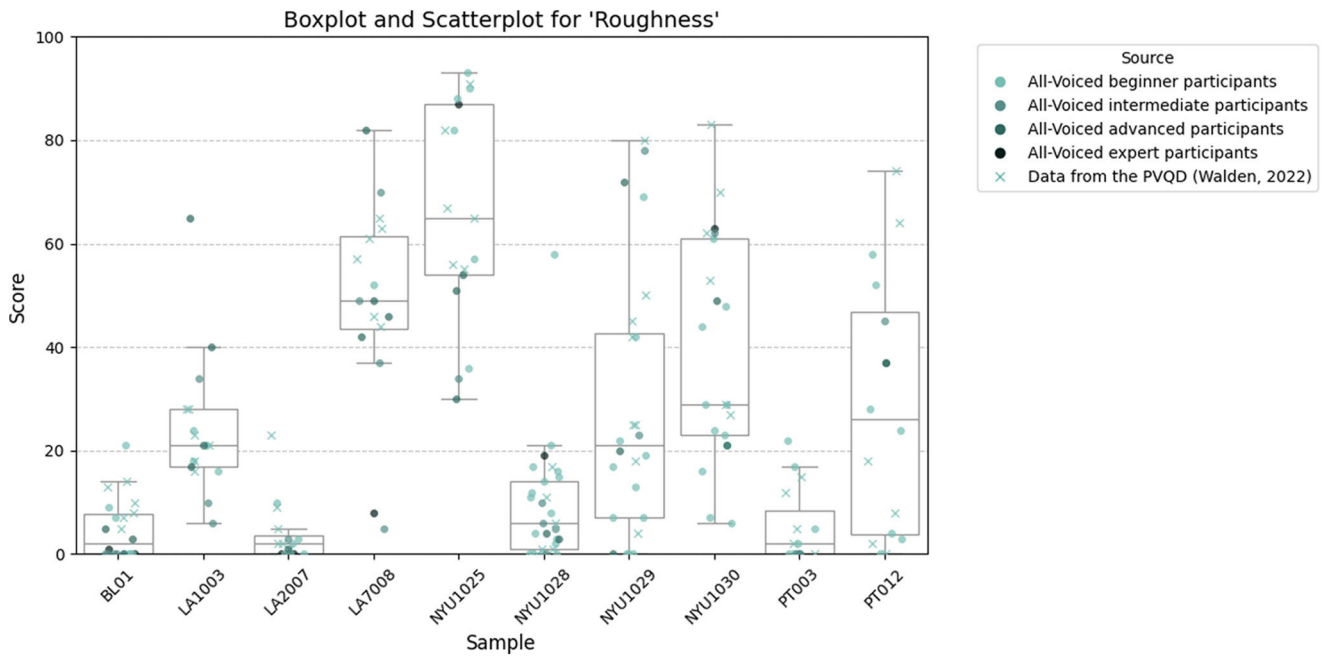
### Participants' sociodemographic characteristics

A total of 264 participants took part in the study, with at least one new user registering each day since the app's launch. In terms of age distribution, the majority of participants were between 21 and 30 years old (49%), followed by 31–40 years (22%), 41–50 years (17%), and 51–60 years (7%). A smaller proportion of participants were under 21 years old (1%), 61–70 years (3%), and only 2 participants (1%) were between 71 and 80 years. There were no participants over 80 years old. With regard to gender pronouns, 88% of participants identified with "she/her," while 8% identified with "he/him." Additionally, 2% preferred "she/they," 1% preferred "he/they", and 1 user preferred other pronouns.

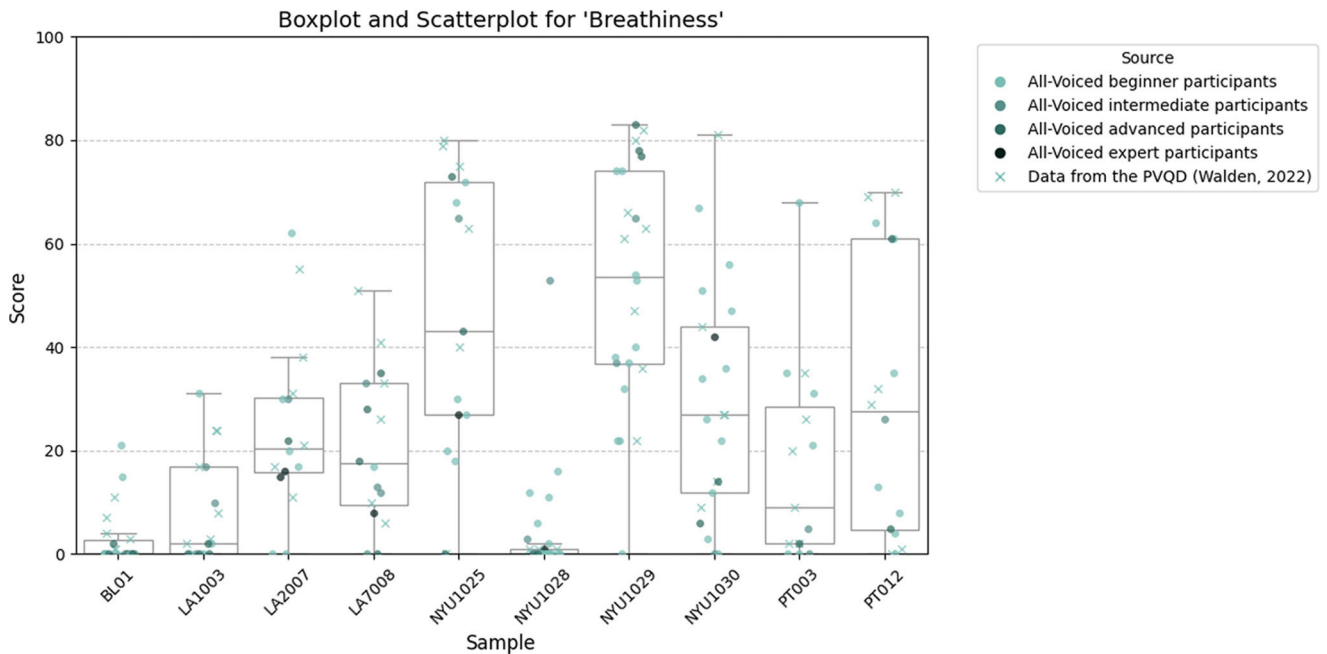
Looking at the country of residence, the majority of participants were from the United States (51%), followed by Spain (18%) and Ecuador (12%). Other countries with notable representation included Canada, Peru and France (7 users each, 3%), as well as the United Kingdom, and Australia (6 users each, 2%). Countries with fewer participants included Italy, Albania, and Belgium (2 users each, 1%) and a diverse range of countries with one participant each, including Estonia, Colombia, Greece, Pakistan, Brazil, Chile, Turkey, Argentina, Bolivia, Costa Rica, and Slovakia.

In terms of language preference, 57% of participants chose English, followed by Spanish (25%) and Catalan (6%). French was also notably represented (4%), while Italian and Portuguese were each selected by 2 participants (1%). Other languages, such as Persian, Turkish, Greek, Pashto, Tamil, and Slovak, were each selected by 1 participant.

Regarding profession, the majority of participants were speech-language pathologists (78%). Other professions



**Figure 2** Distribution of roughness scores for samples with 10 or more unique all-voiced raters, showing expertise levels of all-voiced participants and PVQD data. *Note.* PVQD = Perceptual Voice Qualities Database (Walden, 2022).



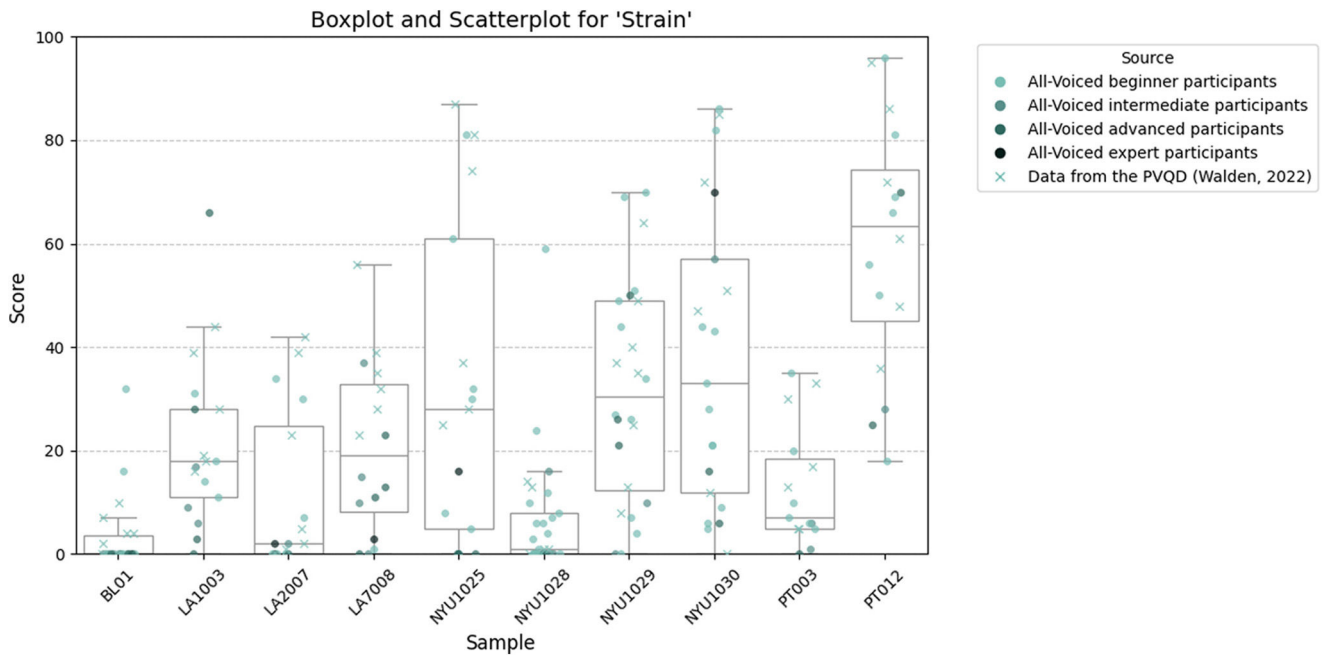
**Figure 3** Distribution of breathiness scores for samples with 10 or more unique all-voiced raters, showing expertise levels of all-voiced participants and PVQD data. *Note.* PVQD = Perceptual Voice Qualities Database (Walden, 2022).

included "other" (17%), singing teachers (1%), one otolaryngologist (ENT specialist), and one vocal coach. Finally, regarding voice competence level, 40% of participants classified themselves as beginners, 25% as intermediate, 20% as advanced, and 14% as experts.

**Gathered evaluations**

A total of 557 evaluations were collected from users, reflecting a strong initial engagement with the platform.

These evaluations encompassed 112 different voice samples, providing a diverse dataset for analysis. 60 evaluations, spanning 42 unique vocal samples, were voluntarily shared with the community, contributing to a collaborative feedback loop that enhances the training process for future participants. While the majority of samples received 1 or 2 evaluations (32 samples had 1 evaluation and 33 samples had 2 evaluations), certain voice samples were evaluated significantly more frequently, with some receiving evaluations from up to 23 unique raters.



**Figure 4** Distribution of strain scores for samples with 10 or more unique all-voiced raters, showing expertise levels of all-voiced participants and PVQD data. Note. PVQD = Perceptual Voice Qualities Database (Walden, 2022).

Figs. 1–4 illustrate the distribution of Overall Severity, Roughness, Breathiness, and Strain scores for samples evaluated by 10 or more unique All-Voiced raters, compared with evaluations from the Perceptual Voice Quality Database (PVQD) by Walden (2022). In these graphs, the voice competence levels of the study participants are distinguished by different colors. This visual representation highlights how evaluators with different levels of voice competence assessed the same voice samples.

Complementing these visualizations, Table 1 presents descriptive statistics for the same attributes across three sources: All-Voiced raters, PVQD raters, and the combined dataset (All-Voiced + PVQD). This table offers detailed insights into central tendencies (mean, median) and variability (standard deviation [Std], interquartile range [IQR]) for each source.

As shown in Fig. 1 and detailed in Table 1, Overall Severity ratings reveal varying levels of agreement among raters across samples. High agreement is observed in samples such as PT012 (IQR = 10), LA2007 (IQR = 12), BL01, and NYU1028 (both IQR = 14). Conversely, substantial dispersion is evident in NYU1029 (IQR = 39) and NYU1030 (IQR = 35). Notably, PT012 and LA2007 present smaller IQRs in the All-Voiced dataset (3 and 9, respectively) compared to the PVQD dataset (22 and 26). In contrast, samples like LA7008 and NYU1029 exhibit greater variability among All-Voiced raters (IQRs of 49 and 44) than PVQD raters (IQRs of 12 and 18).

As depicted in Fig. 2 and detailed in Table 1, samples with lower Roughness levels, such as LA2007, BL01, and PT003 (median = 2 for all), exhibit the least variability, with IQRs of 4, 8, and 9, respectively. Conversely, PT012, which showed the highest agreement for Overall Severity (IQR = 10), exhibits the greatest variability for Roughness (IQR = 43). In this sample, agreement is higher among All-Voiced raters (IQR = 34) compared to PVQD raters (IQR = 49). Conversely,

NYU1025 shows lower overall agreement (IQR = 33), with PVQD raters exhibiting greater consensus (IQR = 20) than All-Voiced raters (IQR = 44).

As illustrated in Fig. 3 and summarized in Table 1, samples with lower levels of Breathiness (median = 0), such as NYU1028 and BL01, exhibit the least variability, with IQRs of 1 and 3, respectively. Interestingly, PT012 displays the highest variability for Breathiness (IQR = 56), consistent with its high variability in Roughness (IQR = 43). However, this same sample displayed the highest agreement in Overall Severity (IQR = 10). Comparisons between datasets reveal differing patterns: for LA1003, All-Voiced raters showed greater consensus, while for NYU1029, NYU1030, NYU1025, and PT003, PVQD raters exhibited higher agreement.

For Strain, samples with lower medians, such as BL01 (median = 0, IQR = 4) and NYU1028 (median = 1, IQR = 8), exhibit the least variability. In contrast, NYU1029, NYU1030, and NYU1025 show the greatest variability, with IQRs of 37, 45, and 56, respectively. Notably, NYU1029 had greater agreement among PVQD raters (IQR = 20) compared to All-Voiced raters (IQR = 40). Conversely, PT003, NYU1025, and LA2007 show higher agreement among All-Voiced raters (Fig. 4, Table 1).

Finally, outliers were observed across samples and attributes, originating from both datasets (Figs. 1–4). Some correspond to All-Voiced raters with varying levels of expertise (e.g., BL01, LA2007, LA7008, NYU1025), while others stem from PVQD raters (e.g., BL01, LA2007, PT012).

As shown in Table 2, twenty voice samples with at least three evaluations were identified for less commonly assessed features, including Pitch (Too High, Too Low), Loudness (Too Loud, Too Soft), resonance-related comments (e.g., Hypernasality), and additional features (e.g., Fry). The number of evaluations per sample ranged from 3 to 9,

**Table 1** Summary statistics for samples evaluated by 10 or more unique all-voiced raters, including comparisons with PVQD and combined (all-voiced + PVQD) datasets.

Sample	Attribute	Source	Mean	Std	Min	Q1	Median	Q3	IQR	Max	Count
BL01	Overall severity	All-voiced	10	21	0	0	0	7	7	71	17
BL01	Overall severity	PVQD	15	12	2	5	13	20	15	35	6
BL01	Overall severity	Combined	11	19	0	0	0	14	14	71	23
BL01	Roughness	All-voiced	3	6	0	0	0	4	4	21	16
BL01	Roughness	PVQD	10	4	5	7	9	12	5	14	6
BL01	Roughness	Combined	5	6	0	0	2	8	8	21	22
BL01	Breathiness	All-voiced	2	6	0	0	0	0	0	21	16
BL01	Breathiness	PVQD	5	4	1	2	4	6	5	11	6
BL01	Breathiness	Combined	3	6	0	0	0	3	3	21	22
BL01	Strain	All-voiced	3	9	0	0	0	0	0	32	16
BL01	Strain	PVQD	5	4	0	3	4	6	4	10	6
BL01	Strain	Combined	3	8	0	0	0	4	4	32	22
LA1003	Overall severity	All-voiced	29	18	7	15	24	40	26	65	11
LA1003	Overall severity	PVQD	29	9	18	24	27	37	13	40	6
LA1003	Overall severity	Combined	29	15	7	18	25	40	22	65	17
LA1003	Roughness	All-voiced	25	16	6	17	21	29	13	65	11
LA1003	Roughness	PVQD	22	5	16	19	22	27	8	28	6
LA1003	Roughness	Combined	24	13	6	17	21	28	11	65	17
LA1003	Breathiness	All-voiced	6	10	0	0	0	6	6	31	11
LA1003	Breathiness	PVQD	13	10	2	4	13	22	18	24	6
LA1003	Breathiness	Combined	8	10	0	0	2	17	17	31	17
LA1003	Strain	All-voiced	18	18	0	8	14	23	16	66	11
LA1003	Strain	PVQD	27	12	16	18	24	36	18	44	6
LA1003	Strain	Combined	22	17	0	11	18	28	17	66	17
LA2007	Overall severity	All-voiced	13	11	0	7	12	16	9	41	10
LA2007	Overall severity	PVQD	29	17	13	16	23	42	26	51	6
LA2007	Overall severity	Combined	19	15	0	11	14	23	12	51	16
LA2007	Roughness	All-voiced	2	3	0	0	1	3	3	10	10
LA2007	Roughness	PVQD	7	8	1	2	4	8	6	23	6
LA2007	Roughness	Combined	4	6	0	0	2	4	4	23	16
LA2007	Breathiness	All-voiced	21	18	0	15	19	28	13	62	10
LA2007	Breathiness	PVQD	29	16	11	18	26	36	18	55	6
LA2007	Breathiness	Combined	24	17	0	16	21	30	15	62	16
LA2007	Strain	All-voiced	8	13	0	0	1	6	6	34	10
LA2007	Strain	PVQD	19	19	1	3	14	35	32	42	6
LA2007	Strain	Combined	12	16	0	0	2	25	25	42	16
LA7008	Overall severity	All-voiced	46	30	0	20	56	69	49	77	10
LA7008	Overall severity	PVQD	58	10	41	53	60	65	12	69	6
LA7008	Overall severity	Combined	50	24	0	42	59	66	24	77	16
LA7008	Roughness	All-voiced	44	24	5	38	48	51	13	82	10
LA7008	Roughness	PVQD	56	9	44	49	59	63	14	65	6
LA7008	Roughness	Combined	49	20	5	44	49	62	18	82	16
LA7008	Breathiness	All-voiced	16	12	0	9	15	26	17	35	10
LA7008	Breathiness	PVQD	28	18	6	14	30	39	25	51	6
LA7008	Breathiness	Combined	21	15	0	10	18	33	24	51	16
LA7008	Strain	All-voiced	11	12	0	2	11	15	13	37	10
LA7008	Strain	PVQD	36	11	23	29	34	38	9	56	6
LA7008	Strain	Combined	20	17	0	8	19	33	25	56	16
NYU1025	Overall severity	All-voiced	63	23	10	61	65	81	20	87	10
NYU1025	Overall severity	PVQD	74	14	53	69	75	81	12	94	6
NYU1025	Overall severity	Combined	67	21	10	61	70	81	20	94	16
NYU1025	Roughness	All-voiced	64	25	30	44	57	88	44	93	11
NYU1025	Roughness	PVQD	69	14	55	58	66	78	20	91	6
NYU1025	Roughness	Combined	66	21	30	54	65	87	33	93	17
NYU1025	Breathiness	All-voiced	40	25	0	24	30	67	43	73	11
NYU1025	Breathiness	PVQD	56	31	0	46	69	78	32	80	6

Table 1 (Continued)

Sample	Attribute	Source	Mean	Std	Min	Q1	Median	Q3	IQR	Max	Count
NYU1025	Breathiness	Combined	46	28	0	27	43	72	45	80	17
NYU1025	Strain	All-voiced	21	28	0	0	8	31	31	81	11
NYU1025	Strain	PVQD	55	28	25	30	56	79	49	87	6
NYU1025	Strain	Combined	33	32	0	5	28	61	56	87	17
NYU1028	Overall severity	All-voiced	9	10	0	1	5	16	15	33	23
NYU1028	Overall severity	PVQD	1	2	0	0	1	1	1	6	6
NYU1028	Overall severity	Combined	7	10	0	0	2	14	14	33	29
NYU1028	Roughness	All-voiced	10	12	0	3	6	15	12	58	23
NYU1028	Roughness	PVQD	6	7	0	1	4	10	9	17	6
NYU1028	Roughness	Combined	9	12	0	1	6	14	13	58	29
NYU1028	Breathiness	All-voiced	5	12	0	0	0	3	3	53	23
NYU1028	Breathiness	PVQD	1	1	0	0	1	1	1	1	6
NYU1028	Breathiness	Combined	4	10	0	0	0	1	1	53	29
NYU1028	Strain	All-voiced	7	13	0	0	1	8	8	59	23
NYU1028	Strain	PVQD	5	7	0	0	1	10	10	14	6
NYU1028	Strain	Combined	6	12	0	0	1	8	8	59	29
NYU1029	Overall severity	All-voiced	51	26	0	32	48	76	44	84	17
NYU1029	Overall severity	PVQD	58	20	16	52	62	70	18	80	8
NYU1029	Overall severity	Combined	53	24	0	37	60	76	39	84	25
NYU1029	Roughness	All-voiced	24	27	0	5	18	28	23	78	16
NYU1029	Roughness	PVQD	36	23	4	23	34	46	23	80	8
NYU1029	Roughness	Combined	28	26	0	7	21	43	36	80	24
NYU1029	Breathiness	All-voiced	49	24	0	36	47	74	38	83	16
NYU1029	Breathiness	PVQD	57	21	22	44	62	70	25	82	8
NYU1029	Breathiness	Combined	52	23	0	37	54	74	37	83	24
NYU1029	Strain	All-voiced	31	23	0	9	27	49	40	70	16
NYU1029	Strain	PVQD	34	18	8	22	36	42	20	64	8
NYU1029	Strain	Combined	32	21	0	12	31	49	37	70	24
NYU1030	Overall severity	All-voiced	46	18	15	30	43	63	33	77	17
NYU1030	Overall severity	PVQD	51	23	26	30	50	68	37	82	6
NYU1030	Overall severity	Combined	48	19	15	30	43	65	35	82	23
NYU1030	Roughness	All-voiced	34	19	6	21	29	49	28	63	15
NYU1030	Roughness	PVQD	54	22	27	35	58	68	33	83	6
NYU1030	Roughness	Combined	39	22	6	23	29	61	38	83	21
NYU1030	Breathiness	All-voiced	28	22	0	9	26	45	36	67	15
NYU1030	Breathiness	PVQD	34	26	9	17	27	40	23	81	6
NYU1030	Breathiness	Combined	29	23	0	12	27	44	32	81	21
NYU1030	Strain	All-voiced	35	28	5	13	28	51	38	86	15
NYU1030	Strain	PVQD	45	33	0	21	49	67	46	85	6
NYU1030	Strain	Combined	38	29	0	12	33	57	45	86	21
PT003	Overall severity	All-voiced	13	13	0	2	8	24	22	37	10
PT003	Overall severity	PVQD	11	10	0	2	10	16	14	26	6
PT003	Overall severity	Combined	12	12	0	1	9	21	20	37	16
PT003	Roughness	All-voiced	5	8	0	0	0	5	5	22	9
PT003	Roughness	PVQD	6	6	0	1	4	10	10	15	6
PT003	Roughness	Combined	5	7	0	0	2	9	9	22	15
PT003	Breathiness	All-voiced	18	23	0	0	5	31	31	68	9
PT003	Breathiness	PVQD	16	14	2	4	15	25	21	35	6
PT003	Breathiness	Combined	17	19	0	2	9	29	27	68	15
PT003	Strain	All-voiced	10	11	0	5	6	10	5	35	9
PT003	Strain	PVQD	17	12	5	7	15	27	20	33	6
PT003	Strain	Combined	13	12	0	5	7	19	14	35	15
PT012	Overall severity	All-voiced	46	17	0	48	50	52	3	59	10
PT012	Overall severity	PVQD	53	15	32	42	55	64	22	72	6
PT012	Overall severity	Combined	49	16	0	47	50	57	10	72	16
PT012	Roughness	All-voiced	29	21	0	9	33	43	34	58	10
PT012	Roughness	PVQD	28	33	0	4	13	53	49	74	6

**Table 1** (Continued)

Sample	Attribute	Source	Mean	Std	Min	Q1	Median	Q3	IQR	Max	Count
PT012	Roughness	Combined	28	25	0	4	26	47	43	74	16
PT012	Breathiness	All-voiced	28	26	0	6	20	55	49	64	10
PT012	Breathiness	PVQD	34	31	0	8	31	60	52	70	6
PT012	Breathiness	Combined	30	27	0	5	28	61	56	70	16
PT012	Strain	All-voiced	56	26	18	34	61	70	36	96	10
PT012	Strain	PVQD	66	22	36	51	67	83	31	95	6
PT012	Strain	Combined	60	24	18	45	64	74	29	96	16

Note. Std = standard deviation; Min = minimum; Q1 = first quartile (25th percentile); Q3 = third quartile (75th percentile); IQR = interquartile range (Q3–Q1); Max = maximum.

**Table 2** Summary statistics for samples with at least 3 evaluations of pitch, loudness, resonance comments, or additional features.

Sample	Attribute	Mean	Std	Min	Q1	Median	Q3	IQR	Max	Count
NYU1028	Hypernasality	45	19	30	30	41	56	26	69	4
PT133	Hypernasality	12	6	7	9	10	13	5	20	4
LA1003	Fry	25	16	7	15	26	35	20	42	4
LA7006	Fry	6	6	1	3	4	8	6	12	3
LA7008	Fry	48	32	8	35	40	65	30	91	5
LA1003	Pitch too low	11	4	8	9	10	13	4	15	3
LA7008	Pitch too low	50	35	16	33	50	68	35	85	3
NYU1016	Pitch too low	21	8	12	18	24	26	8	27	3
NYU1025	Pitch too low	38	29	18	18	20	48	30	94	7
NYU1028	Pitch too low	9	3	5	8	10	11	3	13	9
NYU1029	Pitch too low	45	30	4	26	50	69	43	77	5
PT004	Pitch too low	36	35	13	16	19	48	32	77	3
PT087	Pitch too low	36	31	4	22	40	53	31	65	3
NYU1028	Loudness too loud	14	4	7	12	14	14	2	19	9
LA2007	Loudness too soft	14	0	14	14	14	14	0	14	3
LA5003	Loudness too soft	40	3	37	38	39	41	3	43	3
NYU1025	Loudness too soft	39	28	5	23	32	59	36	76	5
NYU1029	Loudness too soft	27	13	11	21	28	34	13	49	8
PT003	Loudness too soft	8	4	4	7	9	11	4	12	3
PT012	Loudness too soft	30	16	14	16	26	43	27	51	6

Note. Std = standard deviation; Min = minimum; Q1 = first quartile (25th percentile); Q3 = third quartile (75th percentile); IQR = interquartile range (Q3–Q1); Max = maximum.

with most evaluated by 3 to 5 raters and a few by as many as 7 or 9.

Variability in the evaluation of these less commonly assessed features is considerable. For example, for “Pitch too low,” the IQR ranges from 3 to 43, revealing a significant disagreement among raters for some samples, “Loudness too soft” shows IQRs ranging from 0 to 36, while “Fry” exhibits IQRs ranging from 6 to 30, reflecting varying levels of dispersion across samples. With respect to “Hypernasality”, despite significance is limited, as it was evaluated in only 2 samples, its assessment shows a notable variability, with IQRs ranging from 5 to 26.

## Discussion

The registration data highlights a strong interest in the app, with a consistent daily enrollment since the launch of the latest version in September 2024. This successful turnout

suggests that the application meets an existing need within the professional and academic communities—specifically, the need for reliable, high-quality material for training auditory-perceptual skills.

These findings are consistent with prior research that concluded the need for consensus-oriented training, structured forums for aligning internal standards, and the use of external reference points to improve the reliability of auditory-perceptual voice evaluations (Calaf & Garcia-Quintana, 2024; Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Iwarsson & Reinholt Petersen, 2012; Nagle, 2022; Nagle, Kempster, & Solomon, 2024). The app appears to address this identified need, providing valuable resources for both professional development and educational training, thus contributing to the broader goal of standardizing and improving auditory-perceptual evaluation practices.

Regarding profession, 78% of participants identified as speech-language pathologists (SLPs), though this percentage

could be higher, as some participants did not declare their profession. The strong representation of SLPs is likely due to the app's focus on tools like the CAPE-V (Kempster et al., 2009), widely used in speech-language pathology practice, and its promotion through social media networks, where many contacts are SLPs. Additionally, 17% of participants selected 'other' as their profession, which may include both professionals who chose not to specify their field and students. To reduce any ambiguity, a revision of the registration form is planned, encouraging students to select the discipline they are studying and ensuring clarity in professional categorization.

The fact that participants from other professions, including singing teachers (4 participants), one otolaryngologist, and one vocal coach, have also shown interest is a positive sign. This highlights the potential for the app to expand its appeal beyond the strict speech-language pathology practice. Moving forward, it will be worth exploring the possibility of adding resources tailored to the needs of these other professional groups, further broadening the app's utility.

### Educational training

The variation in the number of evaluations per sample suggests that the app is being actively used in academic settings. While the majority of the 120 samples had received only 1 or 2 evaluations 75 days after the launch, certain voice samples were evaluated significantly more frequently, with some reaching up to 23 unique raters. This disparity indicates that instructors may have selected specific samples for use in structured classroom activities, leveraging the app as a teaching tool. This aligns with the app's dual purpose: supporting autonomous training through randomly selected samples and facilitating guided instruction by enabling educators to target specific voice characteristics. These findings underline the app's potential for enhancing both individual learning and collaborative classroom-based training.

The distribution of voice competence levels among participants further supports the idea that the app is being incorporated into educational environments. A notable 40% of participants identified themselves as beginners, followed by 25% as intermediate, 20% as advanced, and 14% as experts. The high proportion of beginner users aligns with the app's potential use in teaching settings, where students learn auditory-perceptual skills under the supervision of instructors.

Importantly, it should be noted that registering for the app is not required for classroom use, yet many beginner participants chose to register and provide consent for their data to be used in research. This suggests that these users found additional value in the app beyond their initial class activities, motivating them to engage more deeply with the platform.

### Cultural diversity

The distribution of participants by country and language can be attributed to several factors. First, the majority of participants in this study were from the United States (51%), which can largely be explained by the fact that the vocal

sample database used in the app (Walden, 2022) is in English, making it more accessible to English-speaking professionals. While the app has been developed to support multiple languages with different CAPE-V adaptations, the sentences in the vocal samples remain limited to English at this stage. Efforts are underway to gather vocal samples in additional languages or to source from global databases that meet ethical requirements, including informed consent. This step-wise approach ensures immediate availability and utility, while laying the groundwork for greater linguistic diversity in future updates.

Additionally, Spain accounted for 18% of participants, with a significant portion selecting Catalan as their preferred language (half of them). This distribution is likely influenced by the professional connections of the app's developer, who has a strong network in Spain, including many Catalan speakers, which likely contributed to early registrations from this region.

Beyond these main regions, the app is also attracting users from a diverse array of countries and languages. This diversity highlights the app's growing international reach and its potential to engage a broad spectrum of users across linguistic and cultural contexts, even as efforts continue to expand the linguistic diversity of the vocal sample database.

### Comprehensive analysis of vocal attributes

An advantage of the data collection process in All-Voiced lies in its ability to digitize the CAPE-V comprehensively, capturing not only the core attributes of overall severity, roughness, breathiness, and strain but also additional features such as distinctions in pitch (e.g., too high or too low) and loudness (e.g., too loud or too soft), resonance-related comments, and additional features. While these attributes are already part of the CAPE-V framework, they are often overlooked in research due to the absence of tools that facilitate their systematic collection. By enabling the efficient capture of all CAPE-V attributes, All-Voiced allows for the identification of voice samples that exemplify these characteristics, which can be utilized in training programs and further studies. This comprehensive approach represents a significant step forward in fully leveraging the CAPE-V's potential for evaluation and research.

To this respect, the results of pitch and loudness evaluations from the PVQD (Walden, 2022) have not been included because it is unclear whether their ratings for inadequate pitch or loudness refer to an alteration that is too high, too low, too loud, or too soft. Such ambiguity complicates the study of reliability in these perceptual judgments. However, by introducing the directional distinctions in the data collection process, the All-Voiced app offers the potential to gather robust data for future studies on the reliability of perceptual judgments also for pitch and loudness alterations, thus addressing a critical gap in the current body of research.

The preliminary results of this study provide further insight into the variability observed in less commonly assessed features, underscoring the complexity of these evaluations. For instance, while some samples showed high consensus among raters, such as for "Pitch too low" with IQRs as low as 3, others exhibited substantial dispersion,

with IQRs up to 43 for the same feature. Similarly, “Loudness too soft” displayed variability ranging from 0 to 36, and “Fry” had IQRs from 6 to 30. This variation suggests that certain attributes are inherently more challenging for raters to assess consistently, emphasizing the need for continued research to better understand the factors contributing to this variability and to refine training approaches for these specific perceptual dimensions.

### Insights from the gathered evaluations

The analysis of Overall Severity ratings reveals patterns of agreement among raters that do not fully align with previous published reports, which suggested greater agreement when evaluating either normal voices or severely dysphonic voices compared to mild and moderate dysphonias (Gerratt et al., 1993; Kreiman & Gerratt, 2000; Kreiman et al., 1993). In this study, the samples BL01 and NYU1028, classified as normal based on the cutoff values proposed by Calaf and Garcia-Quintana (2024), exhibit high agreement (IQR = 14), which is consistent with this tendency. However, LA2007 (classified as mild, IQR = 12) and PT012 (classified as moderate, IQR = 10) show an even greater agreement, presenting a notable incongruence with the literature. This discrepancy suggests that factors beyond overall severity, such as specific perceptual attributes, sample characteristics, or evaluator familiarity with certain vocal patterns, may influence agreement levels, underscoring the need for further investigation.

When analyzing specific perceptual attributes, the tendency for normal or less altered samples to achieve higher agreement among raters becomes more evident. For instance, Roughness evaluation in samples LA2007, BL01, and PT003 show the least variability (median = 2, IQRs = 4, 8, and 9, respectively), consistent with their closer alignment to non-altered voice qualities. Similarly, for Breathiness, NYU1028 and BL01 (median = 0, IQRs = 1 and 3, respectively) exhibit a high level of agreement. For Strain, BL01 (median = 0, IQR = 4) and NYU1028 (median = 1, IQR = 8) show a strong agreement, whereas LA2007 (median = 2, IQR = 35) presents greater variability, which may be attributed to its classification as mild in Overall Severity compared to BL01 and NYU1028, which were classified as normal. Further investigation is needed to fully understand the factors influencing variability, particularly in attributes like Strain, where agreement levels appear to fluctuate depending on the sample’s overall severity classification.

Despite these insights, a key limitation of this study is the limited volume of data at this early stage, which restricts the scope of inter-rater and intra-rater reliability analyses. Many voice samples received only a single evaluation, making it impossible to assess reliability across evaluators for those samples. Moreover, the current dataset size does not yet allow for meaningful exploration of trends based on demographic factors such as age, gender, profession, or language. However, as data collection expands, these analyses will become feasible, offering deeper insights into the patterns of agreement and variability observed in this study.

The observed variability in evaluations across samples and attributes, even among expert evaluators, aligns with findings from prior studies on CAPE-V adaptations across different languages. While high inter-rater reliabilities have

been reported for certain attributes like overall severity (Karnell et al., 2007; Mozzanica et al., 2014; Pommée, Mbagira, et al., 2024), attributes such as strain, pitch, and loudness consistently exhibit lower inter-rater reliability across multiple languages, including English (Zraick et al., 2011), Brazilian Portuguese (Behlau et al., 2022), and Tamil (Venkatraman et al., 2022). Similarly, intra-rater reliability has varied across studies, particularly for attributes like pitch and loudness (Behlau et al., 2022; Salary Majd et al., 2014).

Outliers observed across all levels of expertise, including expert evaluators, further underscore the inherent challenges in achieving consistency. Factors such as differences in internal standards and interpretations of perceptual attributes likely contribute to these discrepancies, reflecting the persistent difficulty of ensuring consensus among evaluators, even within the structured framework of the CAPE-V.

The results for PT012 exemplify how All-Voiced can identify voice samples that pose special challenges for evaluators. Despite displaying the highest agreement in Overall Severity ratings (IQR = 10), this sample exhibited substantial variability in the evaluation of specific perceptual attributes, such as Breathiness (IQR = 56), Roughness (IQR = 43), and Strain (IQR = 29). This discrepancy suggests differences in how evaluators interpret and apply these specific perceptual attributes, reflecting persistent challenges in achieving consensus. According to Nagle, Kempster, et al. (2024), clinicians have reported difficulties with attributes like strain, overall severity, and pitch in the context of the CAPE-V, indicating a need for clearer definitions and guidelines. To address this issue, the same authors have proposed the CAPE-Vr [Pre-print] (Kempster, Nagle, & Solomon, 2024), a revised version of the CAPE-V framework that aims to resolve these conceptual ambiguities and provide more precise guidance for clinical practice.

While standardizing definitions and procedures is essential, the findings presented here emphasize the need to address the inherently subjective nature of auditory-perceptual evaluations. To this respect, All-Voiced emerges as a useful tool to identify voice samples with poor expert consensus, providing a unique opportunity to investigate the root causes of these discrepancies. Such disagreements may stem from disparities in raters’ cognitive representations of voice attributes, or from specific acoustic features in the samples that influence raters’ perceptions differently. Future work will focus on implementing targeted training modules and validated auditory anchors, aiming to create a shared training platform that enhances the reliability of auditory-perceptual evaluations through consensus building and the adoption of universally shared criteria.

### Conclusions

The All-Voiced app has been immediately well received, particularly by speech-language pathologists, highlighting its potential as a valuable resource for auditory-perceptual training. The steady influx of daily registrations underscores the demand for standardized tools to enhance evaluation reliability, a need well-documented in previous research. The high number of beginner-level users, along with fre-

quently evaluated samples, suggest the app is being used in academic settings, reinforcing its role as a promising educational tool. Through large-scale data collection and analysis, the All-Voiced app provides a foundation for addressing broader limitations in the subjective nature of auditory-perceptual evaluations, and bears the potential to enhance consistency and to support evidence-based practices in voice disorder management.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used GPT-4, a generative AI tool developed by OpenAI, to assist in the drafting and refinement of the manuscript. After using this tool, the author reviewed and edited the content as necessary, taking full responsibility for the content of the publication.

## Funding

The author declares no funding.

## Conflict of interest

The author is the developer of the app discussed in this manuscript. This potential conflict of interest has been addressed by ensuring that the data and analysis are presented objectively, with no financial interests involved.

Given his role as Editorial Board Member in the journal, Neus Calaf had no involvement in the peer review of this article and has no access to information regarding its peer review.

## Data availability statement

The datasets generated and analyzed during the current study are available from the corresponding author upon request.

## Acknowledgments

Special thanks to Noel Warren for the introduction to app development, which made this project possible, and to David Garcia-Quintana for his valuable assistance in revising the manuscript.

## References

- Bartsties, B., & De Bodt, M. (2015). Assessment of voice quality: Current state-of-the-art. *Auris Nasus Larynx*, 42(3), 183–188. <https://doi.org/10.1016/j.anl.2014.11.001>
- Behlau, M. (2004). Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), ASHA 2003 [Refletindo sobre o novo]. *Revista Da Sociedade Brasileira de Fonoaudiologia*, 9, 187–189.
- Behlau, M., Rocha, B., Englert, M., & Madazio, G. (2022). Validation of the Brazilian Portuguese CAPE-V Instrument—Br CAPE-V for auditory-perceptual analysis. *Journal of Voice*, 36(4) <https://doi.org/10.1016/j.jvoice.2020.07.007>, 586.e15–e20
- Calaf, N. (2024a). All-voiced [conference presentation]. In *Voice Lab 3, 15th Pan-European Voice Conference (PEVOC)*. September 6.
- Calaf, N. (2024b). All-Voiced [Web application]. Available from: <https://www.all-voiced.com/>
- Calaf, N., & Garcia-Quintana, D. (2024). Development and Validation of the Bilingual Catalan/Spanish cross-cultural adaptation of the consensus auditory-perceptual evaluation of voice. *Journal of Speech, Language, and Hearing Research*, 67(4), 1072–1089. [https://doi.org/10.1044/2024\\_JSLHR-23-00536](https://doi.org/10.1044/2024_JSLHR-23-00536)
- Chen, Z., Fang, R., Zhang, Y., Ge, P., Zhuang, P., Chou, A., & Jiang, J. (2018). The Mandarin version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) and its reliability. *Journal of Speech, Language, and Hearing Research*, 61(10), 2451–2457. [https://doi.org/10.1044/2018\\_JSLHR-5-17-0386](https://doi.org/10.1044/2018_JSLHR-5-17-0386)
- Connor, N., Bless, D., Dardis, C., & Vinney, L. (2008). Voice disorders: Simulations. *Resources for Teaching & Learning*. Available from: <https://slpsims.csd.wisc.edu/> Accessed 07.6.23
- de Almeida, S. C., Mendes, A. P., & Kempster, G. B. (2019). The consensus auditory-perceptual evaluation of voice (CAPE-V) psychometric characteristics: II European Portuguese Version (II EP CAPE-V). *Journal of Voice*, 33(4) <https://doi.org/10.1016/j.jvoice.2018.02.013>, 582.e5–e13
- Ertan-Schlüter, E., Demirhan, E., Ünsal, E. M., & Tadihan-Özkan, E. (2020). The Turkish version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V): A reliability and validity study. *Journal of Voice*, 34(6) <https://doi.org/10.1016/j.jvoice.2019.05.014>, 965.e13–e22
- Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing Internal and External Standards in Voice Quality Judgments. *Journal of Speech, Language, and Hearing Research*, 36(1), 14–20. <https://doi.org/10.1044/jshr.3601.14>
- Gunjawate, D. R., Ravi, R., & Bhagavan, S. (2020). Reliability and validity of the Kannada version of the Consensus Auditory-Perceptual Evaluation of Voice. *Journal of Speech, Language, and Hearing Research*, 63(2), 385–392. [https://doi.org/10.1044/2019\\_JSLHR-19-00020](https://doi.org/10.1044/2019_JSLHR-19-00020)
- Hirano, M. (1981). *Clinical examination of voice*. New York: Springer-Verlag.
- Iwarsson, J., & Reinholt Petersen, N. (2012). Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *Journal of Voice*, 26(3), 304–312. <https://doi.org/10.1016/j.jvoice.2011.06.003>
- Jesus, L. M. T., Barney, A., Santos, R., Caetano, J., Jorge, J., & Couto, P. S. (2009). *Universidade de Aveiro's voice evaluation protocol*. *Interspeech 2009*. <https://doi.org/10.21437/interspeech.2009-289>
- Joshi, A., Baheti, I., & Angadi, V. (2020). Cultural and linguistic adaptation of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) into Hindi. *Journal of Speech, Language, and Hearing Research*, 63(12), 3974–3981. [https://doi.org/10.1044/2020\\_JSLHR-20-00348](https://doi.org/10.1044/2020_JSLHR-20-00348)
- Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailley, S. A., & Hoffman, H. T. (2007). Reliability of Clinician-Based (GRBAS and CAPE-V) and Patient-Based (V-RQOL and IPVI) Documentation of Voice Disorders. *Journal of Voice*, 21(5), 576–590. <https://doi.org/10.1016/j.jvoice.2006.05.001>
- Kelchner, L. N., Brehm, S. B., Weinrich, B., Middendorf, J., deAlarcon, A., Levin, L., & Elluru, R. (2010). Perceptual Evaluation of Severe Pediatric Voice Disorders: Rater Reliability Using the Consensus Auditory Perceptual Evaluation of Voice. *Journal of Voice*, 24(4), 441–449. <https://doi.org/10.1016/j.jvoice.2008.09.004>
- Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18(2), 124–132. [https://doi.org/10.1044/1058-0360\(2008/08-0017\)](https://doi.org/10.1044/1058-0360(2008/08-0017))

- Kempster, G. B., Nagle, K. F., & Solomon, N. P. (2024). Development and rationale for the consensus auditory-perceptual evaluation of voice – Revised (CAPE-Vr). [Pre-print]. *PsyArXiv Preprints*, <https://doi.org/10.31234/osf.io/e84tn>
- Kondo, K., Mizuta, M., Kawai, Y., Sogami, T., Fujimura, S., Kojima, T., ..., & Haji, T. (2021). Development and validation of the Japanese version of the Consensus Auditory-Perceptual Evaluation of Voice. *Journal of Speech, Language, and Hearing Research*, 64(12), 4754–4761. [https://doi.org/10.1044/2021\\_JSLHR-21-00269](https://doi.org/10.1044/2021_JSLHR-21-00269)
- Kreiman, J., & Gerratt, B. (2000). Sources of listener disagreement in voice quality assessment. *Journal of the Acoustical Society of America*, 108(4), 1867–1876. <https://doi.org/10.1121/1.1289362>
- Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research*, 36(1), 21–40. <https://doi.org/10.1044/jshr.3601.21>
- Kreiman, J., Vanlancker-Sidtis, D., & Gerratt, B. R. (2003). *Defining and measuring voice quality*. In *ISCA tutorial and research workshop on voice quality: Functions, analysis and synthesis*.
- Labaere, A., De Bodt, M., & Van Nuffelen, G. (2023). Construction of an anchor and training sample set for auditory-perceptual voice evaluation with the GRBAS-scale. *Journal of Voice*, <https://doi.org/10.1016/j.jvoice.2023.09.033>
- Mohd Mossadeq, N., Mohd Khairuddin, K. A., & Zakaria, M. N. (2022). Cross-cultural adaptation of the Consensus Auditory-perceptual Evaluation of Voice (CAPE-V) into Malay: A validity study. *Journal of Voice*, 38(6), 1527–e27. <https://doi.org/10.1016/j.jvoice.2022.05.018>
- Mozzanica, F., Ginocchio, D., Borghi, E., Bachmann, C., & Schindler, A. (2014). Reliability and validity of the Italian version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Folia Phoniatrica et Logopaedica*, 65(5), 257–265. <https://doi.org/10.1159/000356479>
- Nagle, K. F. (2022). Clinical use of the CAPE-V scales: Agreement, reliability and notes on voice quality. *Journal of Voice*, <https://doi.org/10.1016/j.jvoice.2022.11.014>
- Nagle, K. (2024). Describing voice quality in 5 minutes: Balancing feasibility and psychometric validity [Conference presentation]. In *53rd annual symposium of the voice foundation*. May 31.
- Nagle, K. F., Kempster, G. B., & Solomon, N. P. (2024). Survey of Voice-Focused Speech-Language Pathologists' Usage of the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V). *Journal of Voice*, <https://doi.org/10.1016/j.jvoice.2024.08.032>
- Núñez-Batalla, F., Morato-Galán, M., García-López, I., & Ávila-Menéndez, A. (2015). Adaptación fonética y validación del método de valoración perceptual de la voz CAPE-V al español. *Acta Otorrinolaringológica Española*, 66(5), 249–257. <https://doi.org/10.1016/j.otorri.2014.07.007>
- Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatrica et Logopaedica*, 61(1), 49–56. <https://doi.org/10.1159/000200768>
- Özcebe, E., Aydinli, F. E., Tiğrak, T. K., İncebay, Ö., & Yılmaz, T. (2019). Reliability and validity of the Turkish version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Journal of Voice*, 33(3) <https://doi.org/10.1016/j.jvoice.2017.11.013>, 382.e1–e10
- Pommée, T., Mbagira, D., & Morsomme, D. (2024). French-language adaptation of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Journal of Voice*, <https://doi.org/10.1016/j.jvoice.2024.03.011>
- Pommée, T., Shanks, M., Morsomme, D., Michel, S., & Verduyck, I. (2024). Validation of the European French version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-Vf). *Journal of Voice*, <https://doi.org/10.1016/j.jvoice.2024.10.021>
- Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D., & Hillman, R. (2013). Evidence-based clinical voice assessment: A systematic review. *American Journal of Speech-Language Pathology*, 22(2), 212–226. [https://doi.org/10.1044/1058-0360\(2012/12-0014\)](https://doi.org/10.1044/1058-0360(2012/12-0014))
- Salary Majd, N., Maryam Khoddami, S., Drinnan, M., Kamali, M., Amiri-Shavaki, Y., & Fallahian, N. (2014). *Validity and rater reliability of Persian version of the Consensus Auditory Perceptual Evaluation of Voice*. *Audiology*, 23(3), 65–74.
- Venkatraman, Y., Mahalingam, S., & Boominathan, P. (2022). Development and validation of sentences in Tamil for psychoacoustic evaluation of voice using the consensus auditory-perceptual evaluation of voice. *Journal of Speech, Language, and Hearing Research*, 65(12), 4539–4556. [https://doi.org/10.1044/2022\\_JSLHR-22-00169](https://doi.org/10.1044/2022_JSLHR-22-00169)
- Walden, P. R. (2022). *Perceptual Voice Qualities Database (PVQD)*. *Mendeley Data*, V4. <https://doi.org/10.17632/9dz247gnyb.4>
- Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., Thrush, C. R., & Glaze, L. E. (2011). Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology*, 20(1), 14–22. [https://doi.org/10.1044/1058-0360\(2010/09-0105\)](https://doi.org/10.1044/1058-0360(2010/09-0105))