A Machine-Learning Algorithm for the Automated Perceptual Evaluation of Dysphonia Severity

[#,*]Benjamin van der Woerd, [†]Zhuohao Chen, [†,1]Nikolaos Flemotomos, [‡]Maria Oljaca, [§]Lauren Timmons Sund, [†,§]Shrikanth Narayanan, and [§]Michael M. Johns, *Hamilton, Canada, and †‡§Los Angeles, California*

Summary: Objectives. Auditory-perceptual assessments are the gold standard for assessing voice quality. This project aims to develop a machine-learning model for measuring perceptual dysphonia severity of audio samples consistent with assessments by expert raters.

Methods. The Perceptual Voice Qualities Database samples were used, including sustained vowel and Consensus Auditory-Perceptual Evaluation of Voice sentences, which were previously expertly rated on a 0–100 scale. The OpenSMILE (audEERING GmbH, Gilching, Germany) toolkit was used to extract acoustic (Mel-Frequency Cepstral Coefficient-based, n = 1428) and prosodic (n = 152) features, pitch onsets, and recording duration. We utilized a support vector machine and these features (n = 1582) for automated assessment of dysphonia severity. Recordings were separated into vowels (V) and sentences (S) and features were extracted separately from each. Final voice quality predictions were made by combining the features extracted from the individual components with the whole audio (WA) sample (three file sets: S, V, WA).

Results. This algorithm has a high correlation (r = 0.847) with estimates of expert raters. The root mean square error was 13.36. Increasing signal complexity resulted in better estimation of dysphonia, whereby combining the features outperformed WA, S, and V sets individually.

Conclusion. A novel machine-learning algorithm was able to perform perceptual estimates of dysphonia severity using standardized audio samples on a 100-point scale. This was highly correlated to expert raters. This suggests that ML algorithms could offer an objective method for evaluating voice samples for dysphonia severity.

Level of Evidence. 4

Key Words: Machine learning–Voice evaluation–Perceptual voice evaluation–Automation–Artificial intelligence.

BACKGROUND

Structured voice evaluation is a critical component of assessing patients with dysphonia. Comprehensive assessment typically includes both perceptual and instrumental assessments. Auditory-perceptual analysis represents the gold standard for the assessment of dysphonia severity. It is inexpensive and robust.[1,2] This method of voice assessment is widely accepted in clinical applications as well as research purposes.[2-4]

Despite the widespread use of perceptual evaluations, it remains a subjective assessment and raters will develop their own internal reference standards with inherent biases, which impact the judgment of future voice samples.[5] These internal standards can vary across time and between different raters, highlighting one critique of this form of voice assessment, namely reliability. Through standardized scales, such as the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) tool, high levels of consistency within and across raters can be achieved.[6,7] With tools such as this, small-scale changes from sample to sample can be reliably detected.[4,6,8] Reliability of auditory-perceptual evaluations has been extensively researched and, when confounding variables are controlled, they have been proven a robust form of voice assessment.[2,4,6,8,9]

Expert raters are important in the reliability of these assessments.[2,10-12] Speech pathology assessment is a time-limited resource and voice evaluations are limited to the times patients can provide voice samples. Furthermore, these assessments typically rely on in-person voice samples, though some research suggests that remote sample collection from non-optimized settings may be adequate for clinical assessment.[13,14] These restrictions indicate a resource bottleneck in these evaluations. A computer-automated perceptual evaluation tool may provide an opportunity to relieve the resource limitations and objectively measure voice samples. This might allow for interval evaluations between in-person visits, which could increase the total number of assessments, and ultimately could allow for within-person normative values as targets for tracking therapeutic improvement or decline.

Recent advancements in machine learning methods have led to many medical applications, including applications

Accepted for publication June 7, 2023.

Presented as podium presentations at the Canadian Society of Otolaryngology Annual General Meeting 2022 in Vancouver, British Columbia, Canada and Fall Voice Conference 2022 in San Francisco, California, USA.

From the #Department of Surgery, Division of Otolaryngology—Head & Neck Surgery, McMaster University, Hamilton, Ontario, Canada; †Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, California; ‡Keck School of Medicine, University of Southern California, Los Angeles, California; and the §Department of Otolaryngology—Head & Neck Surgery, University of Southern California, Los Angeles, California.

Address correspondence and reprint requests to Benjamin van der Woerd, Department of Surgery, Division of Otolaryngology—Head and Neck Surgery, McMaster University, 50 Charlton Avenue East, Office G839, Hamilton, Ontario L8N 1Y3, Canada.

1Nikolaos Flemotomos is now working at Apple Inc.

Journal of Voice, Vol xx, No xx, pp. xxx–xxx

0892-1997

https://doi.org/10.1016/j.jvoice.2023.06.006

within otolaryngology—head and neck surgery.[15-17] With respect to voice assessments, researchers have developed tools to categorize samples according to gender and evaluate dysphonia using sustained vowel samples.[15] However, the traditional auditory-perceptual assessment includes both sustained vowels and connected speech to evaluate voice parameters across a variety of laryngeal behaviors.[18,19] To date, there are no machine-learning global assessments of dysphonia severity using both sustained vowels and sentence samples calibrated to known expert ratings. In this study, the authors seek to develop a machine-learning algorithm for evaluating dysphonia severity on a 100-point scale using previously collected and expertly rated voice samples of sustained vowels and connected speech.

METHODS

This study was designed using a previously labeled data set to train a machine-learning model on. The Perceptual Voice Qualities Database (PVQD) includes audio samples (n = 295) which were professionally captured at participating voice centers. These samples include sustained vowels and connected speech (CAPE-V sentences) segments. Furthermore, each sample was previously rated by three experts on a 0–100 scale according to the standards of the CAPE-V. The labeled data set allows the computer to know how the samples are supposed to be rated and to fit different criteria for the prediction model result in similar estimates. The primary goal was to teach the model to categorize voice samples according to the same scale.

To do this, we used the OpenSMILE open-source toolkit with the emobase2010 configuration.[20] This was used to extract acoustic and prosodic features, as well as pitch onsets and recording duration. More explicitly, we extracted a base of 34 low-level descriptor (LLD) features (including Mel-Frequency Cepstral Coefficient features, logarithmic power of Mel-frequency bands, normalized intensity, etc) with 34 corresponding delta coefficients appended and applied 21 different statistical functions (such as standard deviation, arithmetic mean, skewness, kurtosis, etc) to these, which resulted in 1428 features. Next, 19 functionals were applied to four LLD features based on the pitch (F0final, jitterLocal, jitterDDP, and shimmerLocal) as well as their corresponding four delta coefficients, resulting in 152 features. Finally, we appended pitch onsets and recording duration features. All LLD features are extracted based on a frame-by-frame analysis, using windows of 25 ms with 10 ms frameshifts.

In consideration of the nature and size of our dataset, we adopted a support vector machine (SVM) instead of a recursive neural network. An SVM is a method often adopted for categorization problems, looking to draw lines between categories that maximize the margin between the line and the closest data points in each category. Within the context of the SVM, different features were tried to achieve the highest correlation and lowest root mean square error (RMSE) to the expert raters' data. The features were analyzed separately for different parts of the voice recordings, and then combined to make the final predictions.

Our feature set, extracted through OpenSmile, consists of descriptive statistics related to low-level characteristics. These are widely utilized in various speech-related tasks such as emotion recognition. The SVM model was implemented with Gaussian kernel, and we tried different configurations of the penalty parameter $C \epsilon [10^{-1}, 1, 10^{1}]$ and kernel coefficient $\gamma \epsilon [10^{-4}, 10^{-3}, 10^{-2}]$. The optimal parameters were selected based on the five-fold cross-validation. This entails splitting the data set into five parts. The model is then trained and tested five times, each using a different part as the test data and the remainder as the training data. Since the data is limited, we selected the K best features using a univariate F test by the scikit-learn Python module,[21] where K $\in$ }50, 100, 150, 200, …, 950, 1000}. The F test measures the impact of each feature on the output. The features with the highest scores were selected for the model and the number of features selected was chosen from the set of values ranging from 50 to 1000.

We separated the recordings into two segments: sustained vowels (V) and connected sentences (S). For each segment, we separately extracted features and also generated a third dataset by extracting features from the whole, unsegmented audio file (WA). To improve the predictive quality of the model, we concatenated the features from these three datasets before selecting the K best features. We selected the combined audio set as it had the best correlation and lowest RMSE. After feature selection, we found jitter-related features of the entire audio file were consistently selected. Jitter, influenced by both sustained vowels and connected sentences, is closely linked to overall voice quality. Extracting jitter features separately from these segments before combining them resulted in slightly lower accuracy, likely due to the short duration of some individual audio files.

RESULTS

The algorithm developed for this study was found to have a high correlation (r = 0.847) with estimates of expert raters when combining features of the three audio sets as shown in Table 1. The model showed enhanced performance when

TABLE 1.
Mean Error, RMSE, and Pearson's Correlation Coefficient

	Mean Error	RMSE	Pearson's
Whole audio	10.484	13.423	0.767
Vowels only	12.337	16.098	0.822
Sentences	10.996	14.233	0.846
Combined (WA, VO, and SO)	10.227	13.362	0.847

Abbreviations: SO, sentences only; VO, vowels only; WA, whole audio.

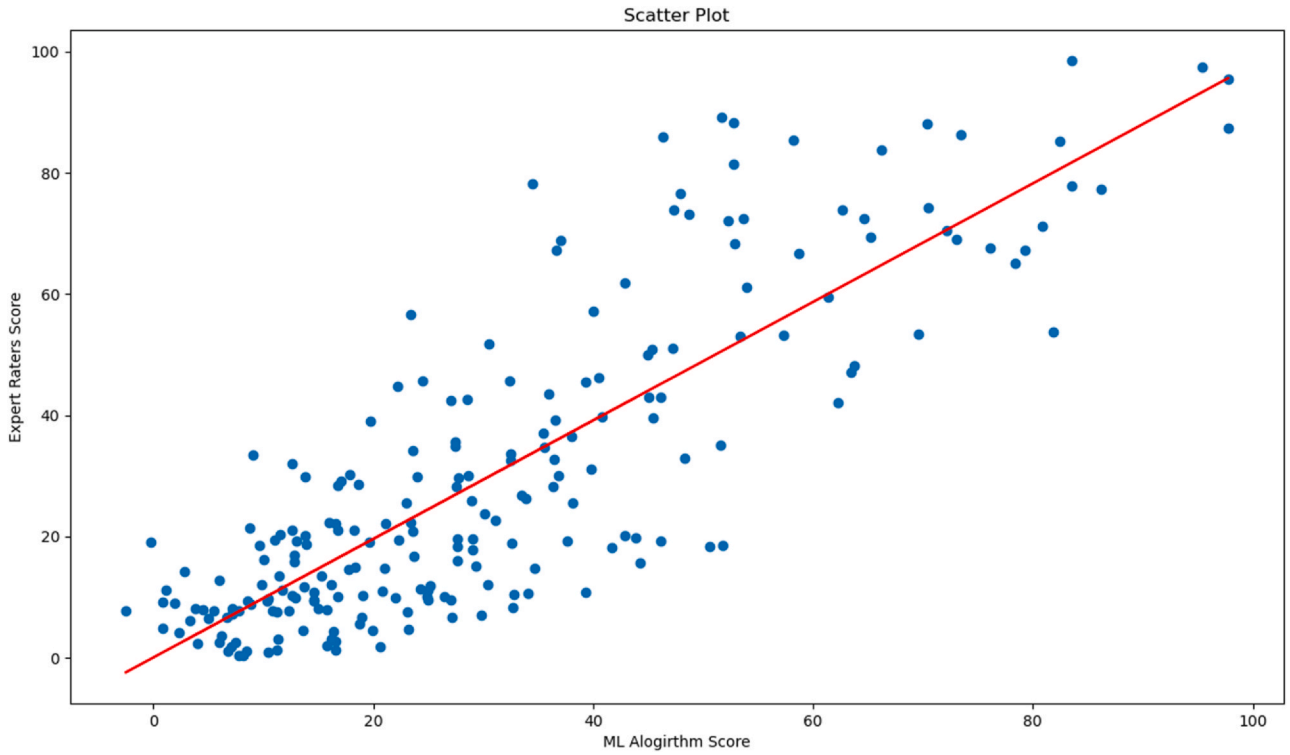

FIGURE 1. Scatter plot of the combined audio set data showing the expert raters' score on the *y*-axis and the ML algorithm score on the *x*-axis. The red line represents the line of best fit. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

exposed to the augmented dataset that is comprised of combined features from (S, V, and WA). This was indicated by higher correlation coefficients and lower RMSE values. The sustained vowel segments alone showed Pearson's correlation coefficient of 0.767. The sentences segment alone had a correlation of 0.822, and unsegmented audio (WA) had a correlation of 0.846.

Mean error when combining three sets of features (V, S, WA) was 10.22, with RMSE of 13.36 (Table 1). Increasing signal complexity resulted in better estimation of dysphonia, whereby combining the features outperformed WA, S, and V sets individually.

The mean of the two combined data sets was 29.88 (standard deviation (SD) 21.24) for the algorithm and 29.32 (SD 25.17) for the expert raters. The intra-class correlation coefficient was 0.83 (95% CI (confidence interval) 0.79, 0.97, $P < 0.001$) between the two data sets, suggesting they are strongly aligned (Figure 1).

DISCUSSION

Voice evaluations are a crucial part of evaluating laryngology patients. The current gold standard for voice evaluation is through auditory-perceptual assessments performed on recorded audio samples collected during the process of in-person office visits. Furthermore, expert voice raters are required for perceptual evaluations, representing a bottleneck in the scalability of ratings. These limitations expose a resource-intensive portion of the laryngology patient examinations.

Automation using computers represents a potential pathway toward reducing the cost of voice evaluation and increasing potential measurement opportunities. Furthermore, machine learning techniques have been applied to tackle a wide variety of predictive tasks within medicine, including the identification of skin malignancies and disordered voices, as well as predicting binary gender from voice samples in gender-affirming voice care.[15-17] Current-day computing power has led to innovation in voice assessments, particularly in the field of modern acoustic voice evaluations including spectral and cepstral analyses.

In this study, the authors have sought to develop an algorithm for automating portions of the auditory-perceptual evaluation of voice, namely Overall Severity. The high correlation coefficient and intra-class correlation with expert raters in the voice samples of the PVQD database indicate this algorithm is able to accurately estimate dysphonia severity. Moreover, the RMSE of 13.36 further attests to the accuracy of the algorithm. A low RMSE, a measure indicating the precision of the predicted versus observed values of a model, indicates a more accurate model. In this case, an RMSE of 13.36 further supports the utility of this algorithm for estimating dysphonia severity in clinical use.

As far as the selection of the features and data sets, this was determined during the training of the model. The 5-fold cross-validation process revealed that the combined audio set was the most highly correlated and had the lowest RMSE of the group (WA, V, and S). In more detail, after

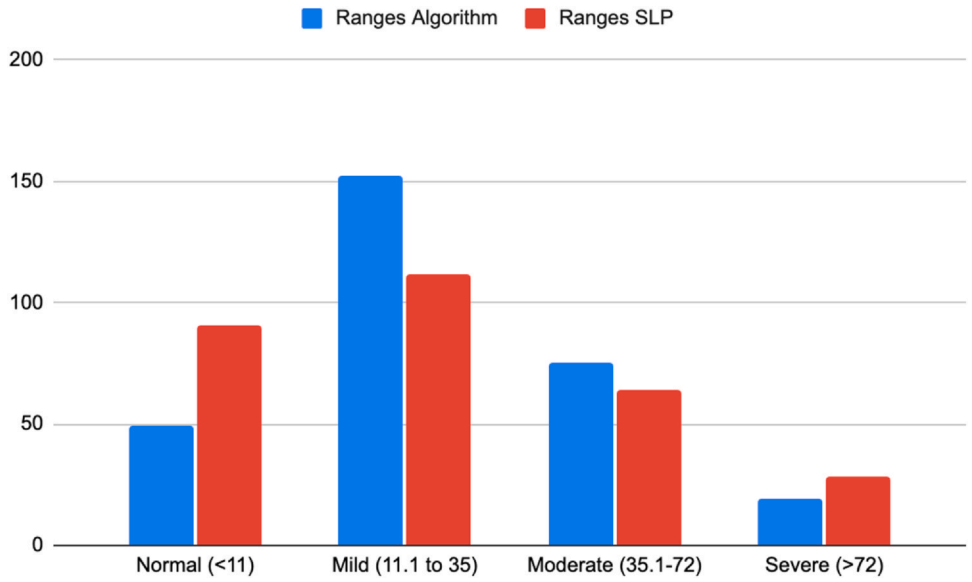

FIGURE 2. Categorical ranges algorithm ratings versus speech-language pathologist (SLP) ratings. Here are the ratings divided into categories of normal (< 11), mild (11.1–35.0), moderate (35.1–72), and severe (> 72). The *y*-axis represents the number or samples found within each category.

performing feature selection on the combined version, the authors observed that the jitter-related features of WA are consistently chosen. This is a result of the fact that jitter is a characteristic influenced by both V and S, and results in independently different measures. The measurement of jitter as part of the whole audio file was more closely linked to overall vocal quality. Should we have used the jitter features extracted from S and V individually before combining them, the measurement performed slightly less accurately, presumably because of the short duration of the individual audio files in some samples.

An area in that we saw the model stray from the experts was with the normal and severe ranges of the CAPE-V system. These edge cases were more likely to be classified as mildly or moderately dysphonic instead of normal and severe respectively (Figure 2). Further study of normal and severely dysphonic samples would be important to improve the validity of this tool. Until then, these limitations need to be recognized when evaluating the quality of this algorithm.

For a tool like this to achieve maximal utility to clinicians, it would be applied to mobile and remotely collected voice samples. This study was performed on high-quality audio samples, but future studies should include non-optimized audio files for analysis. This algorithm performed very well with this data set and opens the doors for exciting opportunities to measure voice samples in an automated fashion, thereby reducing the resource requirements of perceptual voice. This could also lead to the opportunity of collecting measurements outside of traditional evaluation settings, such as clinical visits. If these possibilities can be achieved, within subject normative values and measurement trending could be used to objectively track voice quality throughout treatment courses and time. These types of remote assessments could provide a method for biofeedback-driven therapeutic approaches. Lastly, the predictive capabilities of an algorithm like this could eventually recommend in-person visits for patients with acute voice changes when compared to their own normal samples.

This study is limited by the development of this algorithm using the high-quality, professionally captured audio samples of the PVQD. This data set is small (n = 295), which represents one limitation of the study's generalizability and future studies should include a higher number of samples for assessment of the model. Secondly, the samples were vetted to be of high recording quality, which would not be representative of voice samples collected remotely or in non-optimized settings, and a secondary study should be completed before this can be applied to these settings. Finally, the data used to train and test the model were not prospectively captured as part of a clinical assessment or within subjects' measurements. The repeatability across time within individuals should be further studied.

CONCLUSIONS

This study represents a significant step toward the automation of dysphonia severity evaluations. The machine learning algorithm developed in this study has shown remarkable consistency with expert raters in evaluating the overall severity of dysphonia. The high correlation coefficient (r = 0.847), intra-class correlation (0.83), and small RMSE (13.36 points) between the algorithm's ratings and the expert raters' ratings suggest that this algorithm has the potential to become a reliable and objective tool for dysphonia evaluation. With respect to comprehensive voice evaluations, dysphonia severity represents one small part. Future studies should evaluate the same algorithm as it applies to other parts

of the CAPE-V to fully automate a perceptual evaluation. This study represents a significant achievement toward building confidence in automated voice evaluations. Importantly, this evaluates voice samples similar to the way clinicians do in many prominent voice centers (100-point CAPE-V scale). Categorization in this manner is far more complicated than would be found in previous studies that performed a similar task utilizing the GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) scale.

This algorithm was trained on a dataset consisting of previously rated voice samples of the CAPE-V sentences and sustained vowels from the PVQD. Although the results are encouraging, more work is necessary to confirm these findings in a larger clinical population. In particular, future studies should assess the repeatability of the measures and compare them with expert raters' ratings. Moreover, the algorithm's performance may vary depending on the voice samples' quality, as this study only used optimized voice samples. Therefore, future studies should evaluate the algorithm's performance using non-optimized voice samples.

The developed algorithm has the potential to automate remote evaluations of dysphonia, which can be particularly useful in areas where access to clinical experts is limited. Furthermore, the algorithm provides an objective estimation of dysphonia severity, which can aid in developing treatment plans and evaluating the effectiveness of interventions. By training a model on expected values from expert raters who utilized the CAPE-V scales, while it does not truly represent the full spectrum of the perceptual evaluation provided by a trained listener, it provides an objective estimation with contextual information of the CAPE-V scales. This number alone does not have meaning but provides a common nomenclature to communicate dysphonia severity by recognizing how the algorithm was trained to perform the estimation.

In conclusion, this study has demonstrated that automated perceptual evaluations of dysphonia severity are possible using current computing power and machine learning techniques. With further development and validation, this algorithm has the potential to become an important tool for clinical evaluations of dysphonia.

LAY SUMMARY

This project develops and evaluates a machine learning model for the perceptual assessment of dysphonia severity of audio samples consistent with assessments by expert raters. The development of this model is a proof-of-concept for the automation of perceptual voice analyses using machine-learning approaches.

Declaration of Competing Interest

The authors affirm that they have no conflicts of interest, financial or otherwise, that could be perceived as potentially influencing the objectivity or integrity of the research presented in this publication. We hereby declare that no competing interests exist, ensuring that this work has been conducted with complete transparency and in accordance with ethical guidelines.

References

1. Kent RD. Hearing and believing. *Am J Speech-Lang Pathol.* 1996;5:7–23. https://doi.org/10.1044/1058-0360.0503.07.
2. Oates J. Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatr Logop.* 2009;61:49–56. https://doi.org/10.1159/000200768.
3. Eadie TL, Doyle PC. Classification of dysphonic voice: acoustic and auditory-perceptual measures. *J Voice.* 2005;19:1–14. https://doi.org/10.1016/j.jvoice.2004.02.002.
4. Karnell MP, Melton SD, Childes JM, et al. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice.* 2007;21:576–590. https://doi.org/10.1016/j.jvoice.2006.05.001.
5. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res.* 1993;36:21–40. http://www.ncbi.nlm.nih.gov/pubmed/8450660.
6. Nemr K, Simões-Zenari M, Cordeiro GF, et al. GRBAS and Cape-V Scales: high reliability and consensus when applied at different times. *J Voice.* 2012;26:812.e17–812.e22. https://doi.org/10.1016/J.JVOICE.2012.03.005.
7. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Am J Speech-Lang Pathol.* 2011;20:14–22. https://doi.org/10.1044/1058-0360(2010/09-0105).
8. Webb AL, Carding PN, Deary IJ, et al. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Oto-Rhino-Laryngol.* 2004;261:429–434. https://doi.org/10.1007/s00405-003-0707-7.
9. Dejonckere PH, Obbens C, de Moor GM, et al. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr.* 1993;45:76–83. http://www.ncbi.nlm.nih.gov/pubmed/8325573.
10. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res.* 1990;33:103–115. https://doi.org/10.1044/JSHR.3301.103.
11. Kreiman Jody, Gerratt BR, Precoda K, et al. Individual differences in voice quality perception. *J Speech Lang Hear Res.* 1992;35:512–520. https://doi.org/10.1044/jshr.3503.512.
12. Wuyts FL, De Bodt MS, Van de Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice.* 1999;13:508–517. https://doi.org/10.1016/S0892-1997(99)80006-X.
13. Uloza V, Ulozaitė-Stanienė N, Petrauskas T, et al. Accuracy of acoustic voice quality index captured with a smartphone – measurements with added ambient noise. *J Voice.* 2021;37:465.e19–465.e26. https://doi.org/10.1016/J.JVOICE.2021.01.025.
14. van der Woerd B, Wu M, Parsa V, et al. Evaluation of acoustic analyses of voice in nonoptimized conditions. *J Speech Lang Hear Res.* 2020;63:3991–3999. https://doi.org/10.1044/2020_JSLHR-20-00212.
15. Bensoussan Y, Pinto J, Crowson M, et al. Deep learning for voice gender identification: proof-of-concept for gender-affirming voice care. *Laryngoscope.* 2021;131:E1611–E1615. https://doi.org/10.1002/LARY.29281.
16. Dai X, Spasic I, Meyer B, et al. Machine learning on mobile: An on-device inference app for skin cancer detection. Presented at: 2019 4th International Conference on Fog and Mobile Edge Computing, FMEC 2019; 2019:301–305. https://doi.org/10.1109/FMEC.2019.8795362.
17. Kojima T, Fujimura S, Hasebe K, et al. Objective assessment of pathological voice using artificial intelligence based on the GRBAS scale. *J Voice.* 2021. https://doi.org/10.1016/J.JVOICE.2021.11.021.
18. ASHA. Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ASHA Special Interest Group 3, Voice and Voice Disorders; 2009.

19. Kempster GB, Gerratt BR, Verdolini Abbott K, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech-Lang Pathol.* 2009;18:124–132. https://doi.org/10.1044/1058-0360(2008/08-0017).
20. Eyben F, Wöllmer M, & Schuller B. OpenSMILE - The Munich versatile and fast open-sourceaudio feature extractor. Presented at: MM'10 - Proceedings of the ACM Multimedia 2010 International Conference; 2010:1459–1462. https://doi.org/10.1145/1873951.1874246.
21. Pedregosa F, Weiss R, Brucher M, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.