npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01702-2

A multivariate model incorporating subharmonic measurements for evaluating vocal roughness

Itsuki Kitayama[1], Kiyohito Hosokawa[1,2] ✉, Shinobu Iwaki[3], Misao Yoshida[4], Akira Miyauchi[5], Kenji Aruga[1], Takanari Kawabe[1], Toshihiro Kishikawa[1], Hidenori Tanaka[1], Takeshi Tsuda[1], Takashi Sato[1], Yukinori Takenaka[1], Makoto Ogawa[1] & Hidenori Inohara[1]

The assessment of voice quality plays a critical role in the clinical evaluation of hoarseness. However, no study has established a highly accurate method for the auditory–perceptual judgment of vocal roughness, which is one of the major components of hoarseness. In this study, we developed a multivariate acoustic model for quantifying vocal roughness using a tailored fundamental frequency (f_o) estimation algorithm. The newly devised parameters enabled the classification and quantification of subharmonics, a key component of rough voice. Furthermore, we introduce the acoustic roughness index (ARI), a predictive acoustic model that integrates these parameters with existing acoustic parameters. The ARI demonstrates high diagnostic accuracy and a strong correlation with auditory–perceptual roughness, establishing it as a robust index for the evaluation of vocal roughness.

Hoarseness is an auditory–perceptual feature of the voice-output spectrum that correlates with the aperiodic vibration of the vocal folds during voice production[1,2]. The irregular vibrations of the vocal folds and supraglottic structures produce nonperiodic low-frequency noise, which is perceived as roughness. In contrast, vocalization with inadequate glottal closure results in harmonic reduction and a mixture of irregular noises, attributed to expiratory leakage, which is perceived as breathiness[1,2]. Therefore, hoarseness is generally described as a voice abnormality related to the auditory–perceptual judgments of roughness and breathiness. These factors exert the greatest impact on the overall assessment of hoarseness[3–5] and are implemented in standardized and quantified perceptual assessment tools, such as the Grade, Rough, Breathy, Asthenic, and Strained (GRBAS) scale[1,6] as well as the Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V)[7]. The GRBAS scale is one of the most widely adopted tools for auditory–perceptual judgment. The CAPE-V protocol standardizes tasks, speech stimuli, and recording methods. It consists of six scales: overall severity, roughness, breathiness, strain, pitch, and loudness. However, the assessment of voice quality using these scales involves subjective aspects, and variability in judgments both within a rater (intra-rater) and between different raters (inter-rater) remains a persistent challenge.

To improve the validity and reliability of voice-quality assessment, objective quantitative assessments using regression models with several acoustic analysis parameters were implemented. One such model is the Acoustic Voice Quality Index (AVQI), which is calculated using the freely available software package Praat (Paul Boersma and David Weenink; Institute of Phonetic Sciences, University of Amsterdam, The Netherlands: http://www.praat.org/)[8]. The AVQI is a multivariate regression model comprising six acoustic measurements that are strongly correlated with the GRBAS G-score[9–11].

The Cepstral Spectral Index of Dysphonia is another indicator that is calculated using the Analysis of Dysphonia in Speech and Voice software (ADSV model 5109; KayPENTAX, Montvale, NJ). Notably, this indicator exhibits a strong correlation with the overall severity of hoarseness on the CAPE-V scale[12,13]. Additionally, Barsties et al. reported on several studies aimed at establishing objective quantitative assessments of breathiness and roughness. For breathiness, the Acoustic Breathiness Index (ABI), a regression model consisting of nine acoustic parameters, was developed as a quantification method that correlated well with the B score[14]. In contrast, several studies have been conducted to establish a prediction equation with roughness as the objective variable. However, even when combining various existing acoustic parameters, it is difficult to quantify roughness. This difficulty may be attributed to the complex acoustic structure and the presence of various subtypes[15,16].

Acoustic voice analysis is an essential aspect of voice research and the clinical assessment of dysphonia; however, its accuracy largely depends on whether or not the voice samples used for analysis exhibit periodic

[1]Department of Otorhinolaryngology and Head & Neck Surgery, The University of Osaka Graduate School of Medicine, Suita-city, Osaka, Japan. [2]Department of Otorhinolaryngology, Osaka Police Hospital (currently, Osaka International Medical and Science Center), Osaka-city, Osaka, Japan. [3]Department of Rehabilitation, Kobe University Hospital, Kobe-city, Hyogo, Japan. [4]Department of Rehabilitation, Sakai Heisei Hospital, Sakai-city, Osaka, Japan. [5]Department of Surgery, Kuma Hospital, Kobe-city, Hyogo, Japan. ✉e-mail: khosokawa@ent.med.osaka-u.ac.jp

waveforms. Regarding the speech periodicity, the summary statement of the National Center for Voice and Speech Workshop on Acoustic Voice Analysis (1995) describes a system for classifying voice samples into three types: Type 1 (almost periodic signals, with no qualitative changes in the analyzed segments), Type 2 (signals with qualitative changes (bifurcations) in the analyzed segments or with subharmonic or modulation frequencies whose energies are close to that of the fundamental frequency (f_o)), and Type 3 signals (signals with no notable periodic structure (chaos))[17]. In particular, subharmonic structures appear in the power spectrum as distinct peaks between two successive harmonic structures, corresponding to f_o. These subharmonic structures generally divide the harmonic interval into several equal intervals (e.g., 1/2, 1/3, and 1/4)[18,19]. The accurate quantification of these various subtypes of subharmonics would contribute to a further understanding of the acoustic diversity of vocal roughness.

Notably, the perception of roughness is influenced by changes in the modulation frequency of approximately 15–300 Hz, which are associated with relatively slow fluctuations in the temporal envelope of the sound waveform[20,21]. Subharmonics form a temporal-envelope change in frequency according to the frequency difference from the f_o, also known as the modulation frequency. In particular, for the frequency-modulated pure tones, the maximum perceived degree of roughness is reported to occur around the 70-Hz modulation frequency. Moreover, the degree of roughness depends on the sound pressure level (SPL)[21].

According to laryngeal aerodynamics, subharmonics can originate from two or more different oscillators (multi-oscillators)[18,22–24], and when two non-linear oscillators combine, they exhibit very complex dynamics, including subharmonics, different types of modulation, and deterministic chaos[25,26]. Additionally, Titze et al. reported the appearance of subharmonics over two to three cycles in asymmetrical vocal-fold oscillations, when two states of period and amplitude alternate[27]. Tokuda et al. conducted a simulation using a two-mass model of the vocal folds and reported that the spectrum for the vocal-fold vibration cycle (denoted as the limit cycle) consists of f_o, expressed as the reciprocal of the vocal-fold vibration period, and higher-order harmonics that are integer multiples of f_o[28]. Conversely, studies have shown that even if the vocal-fold vibration cycle is a limit cycle, subharmonics, which are harmonic components at one-half or one-third multiples of f_o, appear in addition to the f_o of the vocal fold vibration when the vibration patterns of the left and right vocal folds form slightly different patterns.

Several acoustic parameters have been developed to detect subharmonics. The degree of subharmonics measure, implemented in the Multidimensional Speech Program Acoustic Analysis software (Kay-PENTAX, USA), estimates the temporal dominance of subharmonics. However, its accuracy depends on the f_o detection and is likely to be inaccurate for hoarse samples[29]. The Diplophonia Diagram is another measure that assesses the combined quality of single and multiple oscillators; however, it is considered to have limited clinical relevance because of its computational complexity[30]. Furthermore, the method proposed by Awan et al., which performs a two-stage cepstrum analysis by dividing the analysis frequency band[31], exhibits high accuracy in diagnosing roughness. However, the validation experiments revealed limitations in the quantification of subharmonics and emphasized the importance of f_o estimation in hoarse speech[32]. Therefore, Kitayama et al. developed a Spectral-Based f_o Estimator Emphasized by Domination and Sequence (SFEEDS), a robust f_o estimation algorithm independent of the type and degree of hoarseness, as a Praat script (Japanese Patent Application Number 2024-087318)[33]. It effectively reduced f_o estimation errors due to subharmonics in voice samples with roughness, achieving more reliable f_o estimations compared to conventional methods. However, no parameters currently exist that can automatically classify subharmonics by subtype and quantify their spectral intensities. This could explain the difficulty associated with identifying acoustic parameters that are highly relevant to the perception of roughness. Therefore, we developed a script in Praat for acoustic parameters that allows the quantification of spectral intensity according to the subtypes of subharmonics.

The first objective of this study is to develop acoustic parameters for quantifying the frequency components and average spectral intensity of subharmonics using the f_o estimated by SFEEDS. The second objective is to develop an acoustic multivariate model for the automatic quantification of vocal roughness that complements the auditory–psychological judgment of roughness. The third objective is to investigate the validity of the acoustic multivariate model using a concatenated speech sample. The results obtained in this study will contribute to providing a clear definition and foundational technology for the subtypes of roughness and subharmonics, which have been difficult to establish statistically because of acoustic diversity.

Results

Validation of the diagnostic properties of the developed parameters

First, the concurrent validity of the four parameters and SubSUM was evaluated. As presented in Table 1, for Rscore (R_{cs}, R_{sv}, and R_{total}), ChaoN exhibited a strong correlation (0.60–0.79) with R_{score}, and SubSUM exhibited a strong (0.60–0.79) to moderate (0.40–0.59) correlation with Rscore. In contrast, the single parameters Sub2, Sub3, and SubS all exhibited correlations with Rscore ranging from moderate (0.40–0.59) to weak (0.20–0.39) (Table 1). Additionally, the diagnostic accuracy of each parameter was assessed by the area under the curve (AUC) values of the receiver operating characteristic (ROC) curve. In this case, ChaoN and SubSUM could detect Rscore with moderate accuracy (0.70–0.90). Conversely, the detection rate of Rscore was low (0.50–0.69) for the single parameters Sub2, Sub3, and SubS (Table 2). This means that the single parameters Sub2, Sub3, and SubS have low diagnostic accuracy for R_{score} alone; however, they are useful in combination.

Furthermore, the same validation was conducted on voice samples concatenating continuous speech (CS) and sustained vowels (SV), which proved that ChaoN and SubSUM had a strong correlation with R_{score} (0.60–0.79) and a moderate diagnostic accuracy (0.70–0.9).

Acoustic roughness index

Second, a multiple regression analysis was performed to construct a statistical model showing the best combination of acoustic predictors for the

Table 1 | Correlation between the parameters and the type of voice samples

	Sub2	Sub3	SubS	ChaoN	SubSUM
R_{sv}	0.528	0.413	0.320	−0.700	0.616
R_{cs}	0.477	0.413	0.318	−0.613	0.581
R_{total}	0.556	0.473	0.396	−0.732	0.636

Spearman's rank correlation coefficient was used for the correlation between the parameters and the type of voice samples ($n = 454$).

Table 2 | Diagnostic accuracy of roughness ($R_{score} \geq 0.5$)

	Sub2	Sub3	SubS	ChaoN	SubSUM
R_{sv}	0.673 (0.366 / 0.997)	0.589 (0.190 / 0.987)	0.553 (0.107 / 1.000)	0.851 (0.718 / 0.870)	0.708 (0.440 / 0.966)
R_{cs}	0.684 (0.460 / 0.891)	0.604 (0.212 / 0.638)	0.564 (0.153 / 0.974)	0.830 (0.656 / 0.883)	0.743 (0.582 / 0.557)
R_{total}	0.722 (0.271 / 0.860)	0.634 (0.271 / 0.996)	0586 (0.179 / 0.992)	0.866 (0.767 / 0.381)	0.767 (0.633 / 0.848)

Values indicate AUC (sensitivity/specificity).
ROC analysis was implemented to assess the diagnostic accuracy of roughness ($R_{score} \geq 0.5$) ($n = 454$).

Table 3 | Regression equation for acoustic roughness index (ARI)

Parameter	Coefficient
Intercept	4.869405796723514
ChaoN	0.068586
SubS	0.107947
Sub3	0.000516
Sub2	0.017007
finalhfno6000	−0.226746
finalhnrd	−0.023793
cpps	0.035208
shimmerLocaldB	2.019074
psd	545.539678
gneMaximum	−1.054173
slope	0.003609
tilt	0.052504

ARI = 4.869405796723514 + 0.068586 × (ChaoN) + 0.107947 × (SubS) + 0.000516 × (Sub3) + 0.017007 × (Sub2) − 0.226746 × (finalhfno6000) − 0.023793 × (finalhnrd) + 0.035208 × (cpps) + 2.019074 × (shimmerLocaldB) + 545.539678 × (psd) − 1.054173 × (gneMaximum) + 0.003609 × (slope) + 0.052504 × (tilt).

degree of roughness. As a regression equation with R_{total} as the objective variable, a linear equation called the acoustic roughness index (ARI) was calculated as Table 3.

We have provided a GitHub repository of the ARI scripts running on Praat (https://github.com/LarynxOsaka/ARI). The automatically calculated results using Praat are depicted in Fig. 1.

Third, the concurrent validity of the ARI was examined. The correlation between the ARI results and R_{total} was RS = 0.807 ($P < 0.0001$), indicating a high degree of concordance. The proportional relationship between R_{total} and ARI is illustrated in Fig. 2. The R^2 value was 0.652, indicating that 65.2% of the variation in R_{total} was explained by the ARI.

Additionally, the diagnostic utility of the ARI was assessed by the AUC values of the ROC curve (Fig. 3). The AUC was 0.916, indicating a high discriminative power to distinguish smooth voices from rough voices. Further, the ROC curve was employed to achieve the optimal balance between sensitivity and specificity and to identify a cutoff score for the optimal discrimination of rough and smooth voices. The ARI threshold value of 2.09 was chosen as the optimal cutoff score. First, a high specificity of 92.4% was achieved, and many subjects with smooth voices were correctly classified. A high sensitivity of 76.0% was also obtained, and subjects with rough voices were correctly classified. Second, the likelihood analysis for this ARI threshold showed the best balance of discriminative power in the positive and negative likelihood-ratio statistics (LR+ and LR−). The LR+ was 9.97, slightly below the recommended value of LR+ ≥ 10. This indicates that an ARI score above 2.09 is indicative of subjects with a rough voice. The LR− was only 0.26, slightly above the recommended value of LR− ≤ 0.1. Generally, the lower the LR−, the more likely it is that a person with a clinical ARI score below the threshold (i.e., below 2.09) has a smooth voice. Equivalent results were obtained in cross-validation (Supplementary Information).

Discussion

Subharmonics are a known hallmark of pathological rough voice quality[20,21] but have been historically difficult to quantify especially in running speech[34]. In this study, we utilized the robust SFEEDS algorithm for f_o estimation and addressed this challenge by introducing new acoustic parameters that measure the frequency distribution of subharmonic components relative to the fundamental frequency and the average spectral intensity of these subharmonics. In the validity tests of the parameters, ChaoN and SubSUM were highly effective, whereas each type of subharmonic parameter alone was less effective. ChaoN is a parameter that reflects the average intensity of the spectrum between f_o and $2^* f_o$ in frames without subharmonics, after the f_o has been accurately estimated by SFEEDS. The concept is similar to that of the harmonics-to-noise ratio but differs in that the f_o is accurately estimated and the region featuring subharmonics is excluded from the analysis. The type-specific subharmonic parameters Sub2, Sub3, and SubS had a low correlation with R_{total} separately, and the roughness detection accuracy was also low. There are two main reasons for this. First, each parameter is appropriate for high-accuracy roughness perception, but not essential. Each parameter represents the ratio of the presence of each type of subharmonic to the sum of the analyzed frames, implying that when a subharmonic is detected, the roughness quality is accumulated using a point addition method. Second, because subharmonics are classified by type, the contribution to the roughness of each type parameter is low. SubSUM, the simple sum of those single parameters, exhibited a strong (0.6–0.79) correlation with R_{total} and a moderate (0.7–0.9) diagnostic accuracy. In other words, the effectiveness of each type of subharmonic parameter is improved by weighting them individually and subsequently combining them. These measures lay the groundwork for improved roughness detection and can be integrated into the multivariate model.

Furthermore, by combining the above parameters with conventional parameters, an optimal combination of acoustic predictors for the roughness degree, the ARI, was established. As stated previously, no study has established a statistical model that can adequately explain vocal roughness, even with the combination of various conventional acoustic parameters[15,16]. In this study, we established a statistical model that can sufficiently explain the roughness by combining ChaoN and parameters that subdivide subharmonics by type into a regression equation. In other words, it was statistically proven for the first time that the percentage of subharmonics present and the type of subharmonics are significantly related to the auditory–psychological judgment of vocal roughness. The ARI is intended to complement rather than replace human judgment. Auditory roughness ratings by trained listeners are considered a gold standard, yet they are subjective and can vary between raters or sessions. Our model provides an objective quantification that can support clinicians by reducing reliance on purely subjective assessment.

We evaluated the validity of our acoustic model using concatenated speech samples, which combine continuous speech with sustained vowels. The use of such concatenated samples is supported by previous research: combining continuous speech and a sustained vowel has been shown to yield more reliable and ecologically valid voice quality assessments [10]. Our results indicate that ARI maintains its validity on concatenated samples.

In our validation, ARI scores tracked the perceptual roughness grades with high accuracy (RS = 0.807, $P < 0.0001$). This level of correspondence approaches the performance of established multivariate dysphonia measures for overall voice quality[9–11].

As a limitation, in voices with severe breathiness, subharmonic structures may be misidentified, as illustrated in Fig. 4. In many cases, these detections represent biphonation—also referred to as a "torus" in nonlinear dynamics—rather than true subharmonics[28]. Biphonation produces nonperiodic waveforms, and its power spectrum reveals two independent fundamental frequencies and their harmonics. An example occurs when the left and right vocal folds oscillate asynchronously at different frequencies[28]. Furthermore, strong chaotic noise in severely breathy voices increases the likelihood of accidentally meeting subharmonic detection criteria, resulting in false positives (Fig. 4). These errors can be avoided by integrating conventional parameters into the statistical model.

The parameters developed in this study focused on the subharmonic-type classification and do not reflect the specific modulation frequency for f_o as a parameter value. As mentioned above[20,21], the perceived roughness peaks at a modulation frequency of 70 Hz. Therefore, future studies should consider the frequency difference between the f_o and subharmonics peaks in the parameter development. As the ARI was only evaluated on SV samples and text reading at rest, the parameters may need to be adjusted for voice samples with dynamic f_o transitions, such as singing or speech with emotion.

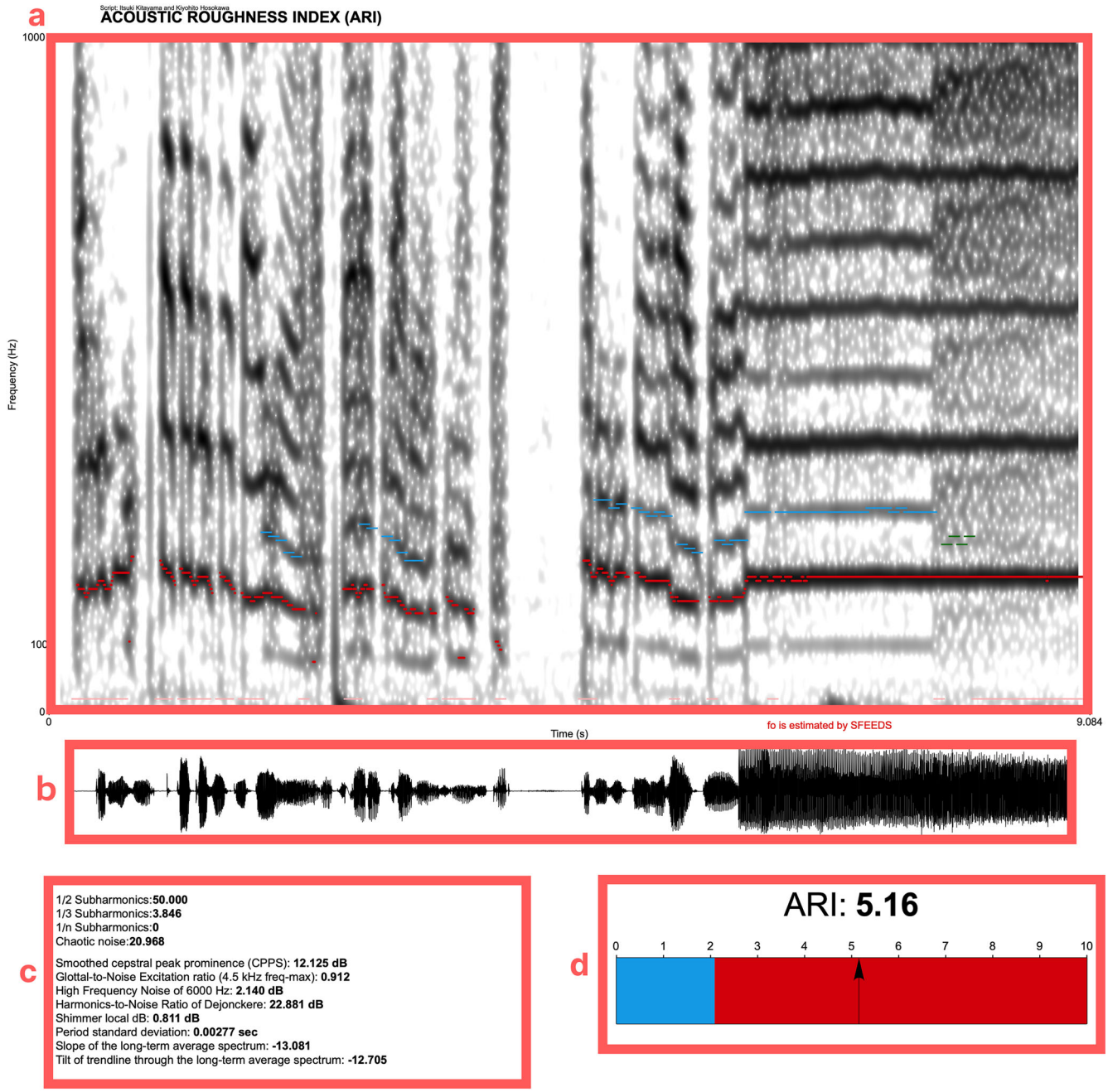

Fig. 1 | Results of ARI calculations using the Praat software. a Narrow-band spectrogram of the analyzed voice samples. The red line shows the f_o transition estimated by SFEEDS, and the other colored lines indicate the detection region of each subharmonic subtype. **b** Oscillogram of the analyzed speech sample. **c** Calculated values of each of the acoustic parameters. **d** Calculated ARI values. In this voice sample, the ARI values are above the green range (below 2.09), indicating a high degree of roughness.

The algorithm is currently limited to validation in Japanese, and future validation in multiple languages is required. For clinical use, this parameter is intended for analysis with audio recorded in a soundproof room and where the signal-to-noise ratio is appropriately maintained. Therefore, there is a physical limitation that prevents accurate measurement in noisy ambulatory environments. The use of aerodynamic testing instead of voice recording, which is less susceptible to external noise, may solve this limitation.

In conclusion, parameters for subharmonics were developed using the SFEEDS algorithm to estimate the f_o of speech samples. Combined with conventional acoustic parameters, a statistical model, ARI, was established to enable the automatic quantification of vocal roughness. Furthermore, the known association between subharmonics and perceptual roughness was corroborated with statistical evidence. Incorporating ARI into conventional acoustic measurement standards will enable more accurate clinical voice assessment and improve the determination of therapeutic efficacy in voice medical treatment.

Methods

Preparation of the voice dataset

A total of 454 voice recordings were collected from a dataset previously used in another study[35]. Of these, 288 recordings were from participants with various organic and non-organic speech disorders showing different degrees of dysphonia, 55 were from participants with no speech complaints, and 111

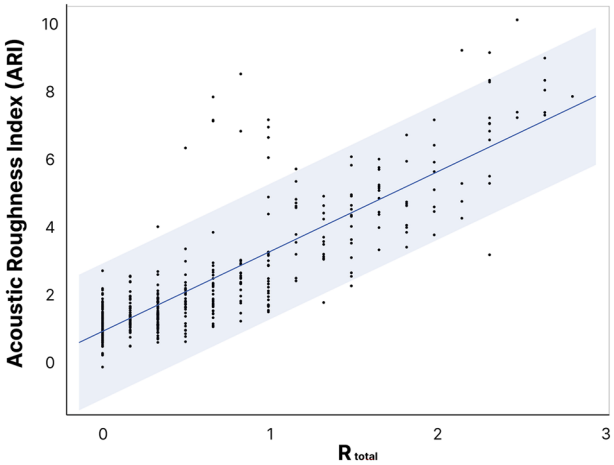

Fig. 2 | Scatterplot showing the concurrent validity of the ARI. The colored ranges above and below the regression line indicate the upper and lower limits of the 95% confidence interval, respectively.

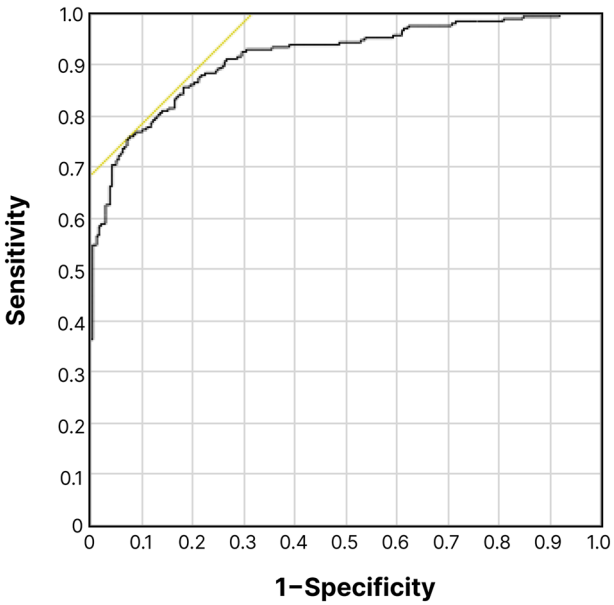

Fig. 3 | The ROC curve illustrates the diagnostic accuracy of the ARI. The intersection with the yellow line shows the coordinates of the sensitivity and (1-sensitivity), corresponding to the threshold that affords the best diagnostic accuracy.

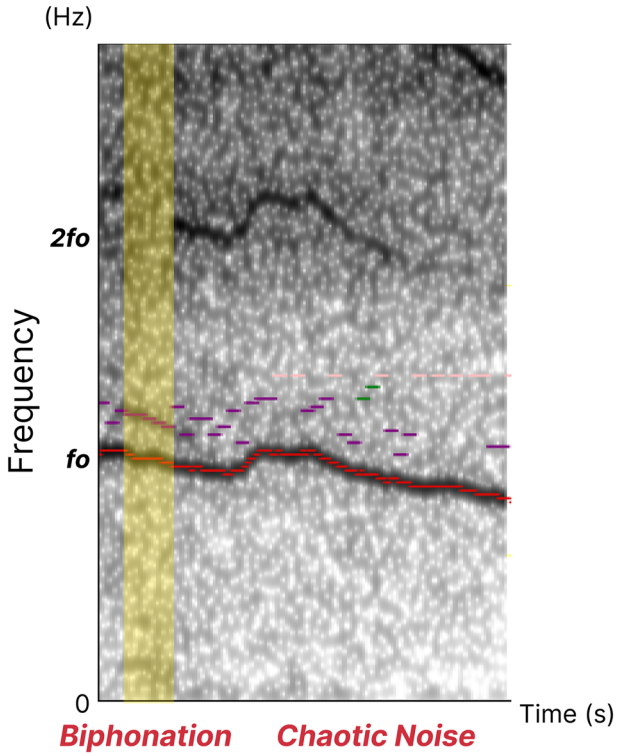

Fig. 4 | Limitations of the analysis in voices with severe breathiness. The auditory–perceptual judgment revealed a breathy voice, with no element of roughness. However, the algorithm misdirects super subharmonics, which are an element of roughness (purple line).

were from participants who were more than 3 months post-treatment. Table 4 provides the diagnostic details for all 343 participants: 181 males and 273 females, mean age 57.5 years, range 9–94. Each participant was asked to sustain the vowel /a:/ for at least 3 s and then read aloud the Japanese translation of "The North Wind and the Sun" at a comfortable pitch, loudness, and tempo.

The procedures for preparing the CS and SV samples were the same as those required to calculate the AVQI in Japanese[36]. Thus, the CS sample consisted of 30 syllables from the beginning of the first sentence to the eighth syllable of the second sentence (/aruhi kitakaze to taiyo: ga chikara kurabe wo shimashita tabibito no gaito: wo/). For the SV samples, the mid vowel was extracted for 3 s, excluding the start and end parts. Subsequently, the CS and SV samples were concatenated for analysis. The data included information on sex, age, and diagnosis, as well as the results of auditory–perceptual judgments of *G* scores, *R* scores, and *B* scores for the SV and CS samples by three raters whose intra- and inter-rater reliabilities were ensured in a previous study[35]. The mean of G_{cs} and G_{sv} furnished by the three raters was defined as G_{total}, and R_{total} and B_{total} were calculated similarly. Hoarseness, roughness, and breathiness were considered present if the G_{total}, R_{total}, and B_{total} values were 0.5 or more, respectively. A summary of the degree of hoarseness for each is given in Fig. 5. The *G*, *R*, B_{score} less than 0.5 is considered normophonic, and all others are considered dysphonic. We confirm that permission to use the dataset originally described by Hosokawa et al.[35] has been granted by the authors.

Recording methods for speech samples

All audio samples were recorded in a sound-treated room using a head-mounted condenser microphone SE50 (Samson Technologies Corp.), positioned 2 cm from the participant's lips. Recordings were digitized at a 44.1 kHz sampling rate with 16-bit resolution using an H4n linear PCM recorder (Zoom Corp.), which incorporated built-in microphone preamps and phantom power (+48 V). All samples fulfilled the generally required level of the signal-to-noise ratio (>30 dB)[37,38].

This study was conducted in accordance with the Helsinki Declaration of 1975 and its amendments, as well as the applicable laws and regulations of Japan. An opt-out approach was employed; therefore, the requirement for obtaining written informed consent from participants was waived. The study protocol, including this waiver, was reviewed and approved by the Institutional Review Board of The University of Osaka (Approval No. 15497), the Institutional Review Board of Osaka Police Hospital (currently Osaka International Medical and Science Center; Approval No. 568), and the Institutional Review Board of Kuma Hospital (Approval No. 20120614-1).

Subharmonics classification

As noted above, the modulation frequency affects the perceived degree of roughness[20,21]. If we let f_o denote the fundamental frequency, a subharmonic frequency f_{sub} was defined by $f_{sub} = f_o/n$, where n is an integer. In this study,

Table 4 | Voice diagnosis and interventions for the eligible normophonic and dysphonic patients

Diagnosis	Number of patients Before	Number of patients After	Interventions
Paresis/paralysis	116	50	Autotherapy (26), Neurorrhaphy (8), AA + MT (7), Collagen injection (3), AA + MT +Neurorrhaphy (2), MT (2), AA (1), Fat injection (1)
Polyp	29	15	Laryngeal microsurgery (14), Autotherapy (1)
Polypoid	28	8	Laryngeal microsurgery (8)
MTD	20	5	Voice therapy (3), PPI medication (2)
Nodules	17	5	Laryngeal microsurgery (5)
Presbylarynx	14	3	Fat injection (3)
Cyst	13	4	Laryngeal microsurgery (4)
Acute laryngitis	13	4	Medication (4)
Glottic tumor/cancer	11	10	Laryngeal microsurgery (6), Concurrent chemoradiotherapy (4), Fat injection (1)
Sulcus vocalis	6	1	Fat injection (1)
Vocal fold scar	4	1	Steroid injection (1)
Ventricular hypertrophy	3	0	—
Laryngeal web	2	2	Endoscopic day surgery (1), Laryngeal microsurgery (1)
Framework trauma	2	1	Open reduction and internal fixation (1)
Phonasthenia	2	0	—
Post radiotherapy	2	0	—
Laryngeal amyloidosis	1	1	Laryngeal microsurgery (1)
Vocal fold granuloma	1	1	PPI medication (1)
ADSD	1	0	—
Androphobia	1	0	—
Hysterical aphonia	1	0	—
Tremor	1	0	—
Normophonic controls	55	0	—
Total	343	111	—

Note: Numbers in parentheses indicate the patient numbers for the respective interventions. Abbreviations: *AA* arytenoid adduction, *ADSD* adductive spasmodic dysphonia, *MT* medialization thyroplasty, *MTD* muscular tension dysphonia, *PPI* proton pump inhibitor.

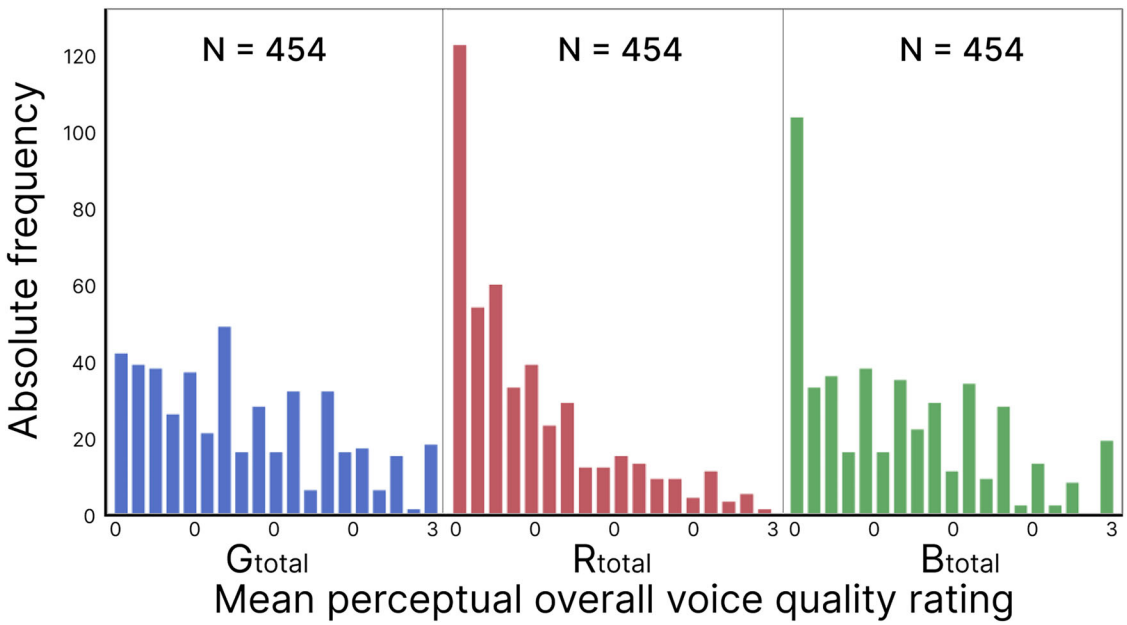

Fig. 5 | Frequency distribution of the mean auditory–perceptual overall voice-quality ratings. The distributions of the auditory perceptual evaluated G_{total}, R_{total}, and B_{total}, respectively, in the voice samples used are shown.

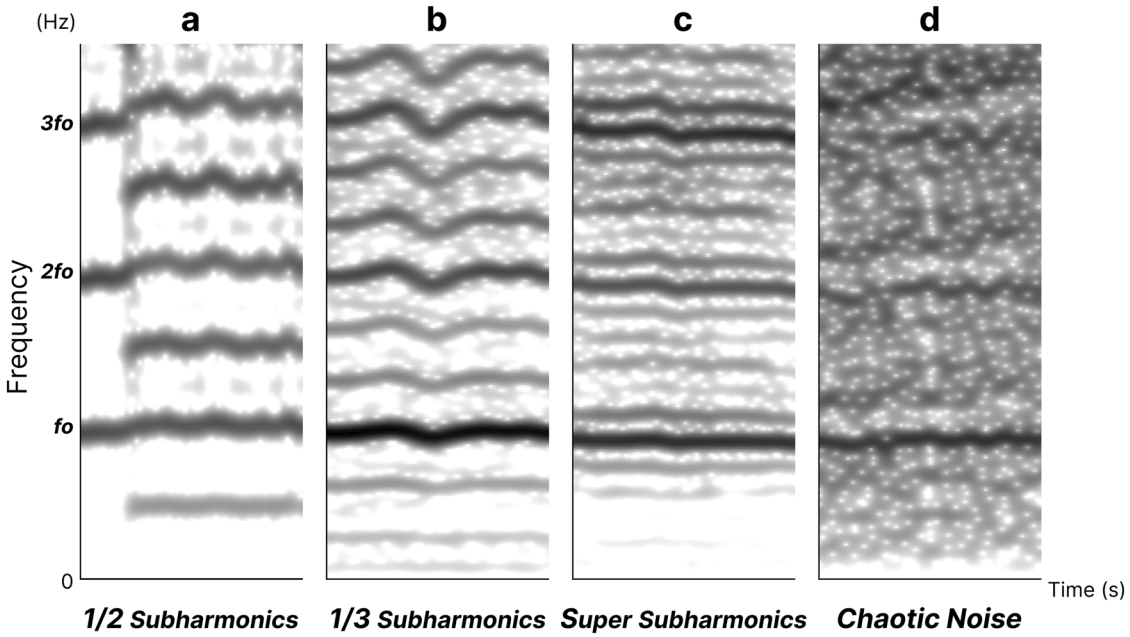

Fig. 6 | Subharmonics were classified into three subgroups for convenience, and nonperiodic noise was classified as chaotic noise. a–d Examples of narrow-band spectrograms for each subtype. **a** Frequency structures of half of the f_o are contained between significant harmonic structures. **b** Frequency structures of a third of f_o are contained between the significant harmonic structures. **c** Frequency structures with a fractional value of 4 or more of f_o are contained between significant harmonic structures. **d** No noticeable frequency peaks are formed between the significant harmonic structures, and only irregular noise is observed.

we classified subharmonics into four distinct groups to reflect their differing characteristics and ensure a more comprehensive analysis: 1/2 subharmonics, 1/3 subharmonics, super subharmonics with more finely divided harmonic intervals, and chaotic noise (no clear subharmonic structure) (Fig. 6).

New parameters for subharmonics

A script was developed in Praat that automatically detects the parts of the speech sample that correspond to the subharmonics classification and calculates the relative spectral intensity to f_o. Details are provided below.

1. Adjusts the average intensity to 70 dB, independent of the sample recording level.
2. Frame segmentation (default: frame length 0.1 s, step 0.0033 s, Gaussian window).
3. Estimates f_o values in all frames using the f_o estimation algorithm SFEEDS.
4. To reduce the computational cost, 20 frame intervals (step 0.066 s) are excerpted. This step width is shorter than the 0.1 s analysis frame length, and the temporal overlap of the analysis window maintains estimation accuracy despite the reduced computational cost. The subharmonic structures in the frequency domain between f_o and 2* f_o are evaluated in the spectral waveform using the estimated f_o values obtained with SFEEDS.
5. To make the spectral waveform easier to analyze, it is converted into a long-term average spectrum (Ltas), a method of analyzing frequencies by dividing them into arbitrary segments (bins) (Fig. 7a).
6. By comparing neighboring bins, the spectral envelope shape for each type of subharmonics was verified (Fig. 7b). Subharmonics types were classified according to the fitting of the shape of the spectral bin and the frequency region where the spectral bin peaks are located. For example, if the top of the spectral envelope is in the frequency range corresponding to 4/3 f_o, it is detected as 1/3 subharmonics.
7. Each frame is provisionally classified by subharmonic type (Fig. 7c).
8. To reduce the risk of misclassifying temporary noise as subharmonics, we implement a test whereby subharmonics are deemed genuine only if the same type of subharmonic, exhibiting a closely matching peak frequency, is detected in two consecutive (temporally adjacent) frames.
9. From the SFEEDS results, first extract the voiced sound frames. Then, calculate the temporal proportion of each of the four subharmonics (Sub2, Sub3, SubS, ChaoN) within these voiced frames, and use these proportions as the corresponding acoustic parameter values (Fig. 7d).

Calculation of other parameter values

Furthermore, each candidate's existing acoustic parameters were calculated to generate regression equations that correlate well with the roughness. The parameters calculated in this study are listed below. These are the parameters that were incorporated as variables when calculating the AVQI and ABI: jitterLocal (percent jitter), shimmerLocal (percent shimmer), shimmerLocaldB (shimmer in dB), CPPS (the cepstral peak prominence), hfno6000 (relative level of high-frequency noise between energy), psd (period standard deviation), gneMaximum (glottal-to-noise excitation ratio), slope (slope of the Ltas), tilt (tilt of the trend line through the Ltas), and hnr (harmonics-to-noise ratio).

The parameters developed in this study focus only on the spectrum between f_o and 2*f_o in all cases. Therefore, estimation errors were expected in voices where the harmonic structure was attenuated and the noise was enhanced in the high-frequency range, such as in voices with severe breathiness. Conventional acoustic parameters already used in AVQI and ABI are integrated into the ARI to compensate for their disadvantages.

Additionally, when calculating conventional parameters, such as CPPS, the diagnostic accuracy is improved by extracting the silent or unvoiced parts using zero-cross rates or other preprocessing methods[39]. However, the fact that the analysis range is largely removed for voices with severe breathiness poses a challenge. Therefore, in a corpus containing a wide variety of hoarseness, trimming samples using a zero-cross rate is considered unsuitable. Further, the analysis in this study was performed with voice samples including silent and unvoiced parts.

Statistical analysis

First, the Shapiro–Wilk normality test showed that the association between the values calculated for the parameters and R_{total} violated the normality assumption ($p < 0.001$), necessitating a non-parametric test. Sub2, Sub3, and SubS represent the temporal proportions of the type-specific subharmonics in the speech sample, respectively. Therefore, their sum, SubSUM, was

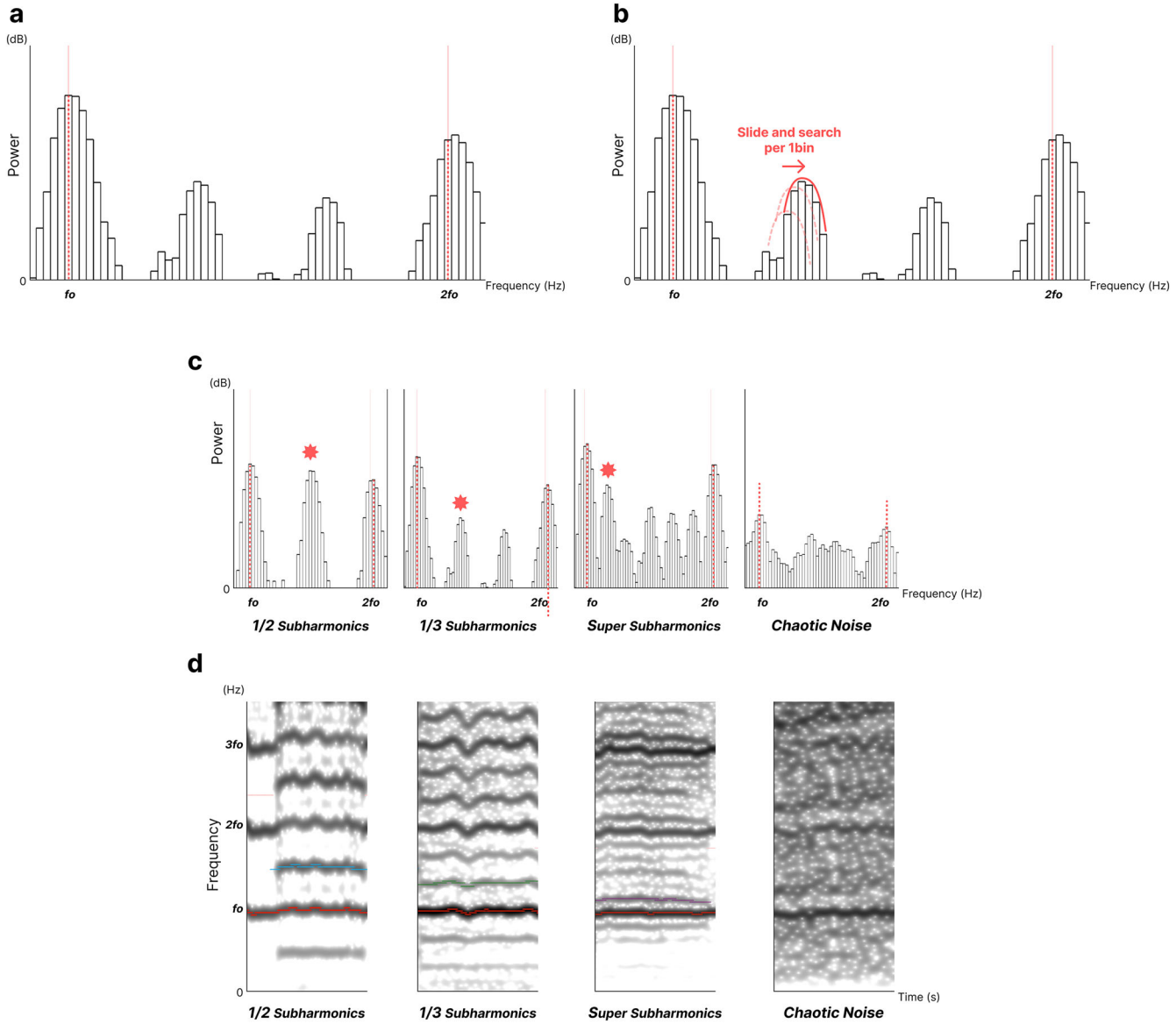

Fig. 7 | Schematic diagram of the method for classifying the subharmonic subtypes and calculating their respective parameters. a Convert into Ltas, which is a method for analyzing frequencies by dividing them into arbitrary segments (bin) (presenting a frame containing 1/4 subharmonics). **b** Compare neighboring bins and estimate the subtype of subharmonics by the shape of the spectral envelope (presenting a frame containing 1/4 subharmonics). **c** Tentatively classify each frame by subharmonic subtype (examples of each subtype are presented). **d** Red line: f_o estimated by SFEEDS. Blue line: frequency region featuring the 1/2 subharmonic. Green line: frequency region featuring the 1/3 subharmonic (depending on the value of f_o, 1/4 subharmonics may be included). Purple line: frequency region featuring more subharmonics.

produced and added to the analysis. To examine the simultaneous validity of the four parameters, the correlation between R_{total} and the four parameters, as well as SubSUM, was established using Spearman's rank correlation coefficient (RS) and the coefficient of determination (R^2). The RS results were interpreted based on Frey et al.'s guidelines[40]. Additionally, the diagnostic usefulness of each parameter was assessed by the AUC values in the ROC curve analysis. The threshold, sensitivity, and specificity for achieving the best accuracy were estimated using the Youden index.

Second, a multiple regression analysis was performed to construct a statistical model showing the best combination of acoustic predictors for the degree of roughness. Considering the complexity of variable selection and the multicollinearity between the parameters, an elastic net was employed to select the explanatory variables. Furthermore, the Cauchy distribution was adapted for the response variable, as the objective variable, R_{total}, is an auditory–psychological judgment that depends on the subject's senses and is prone to outliers. The Akaike information criterion was adopted to balance the goodness of fit and complexity (number of parameters) of the model. To simplify the clinical interpretation, the model was linearly rescaled so that the equation results in a score between 0 and 10. The higher the value, the more advanced the degree of roughness. This final model is referred to as the ARI.

Third, to examine the concurrent validity of the ARI, the correlation between R_{total} and the ARI was calculated using RS and R^2. Furthermore, the diagnostic utility of the ARI was assessed by the AUC values of the ROC curve. All statistical analyses were performed using the JMP version 17.0.0 software package (SAS Institute, Cary, NC). All results were considered statistically significant at $p < 0.05$.

Validation using different data sets

The recording conditions of the voice samples used in the cross-validation were the same, and we extracted up to 100 voice disorder cases recorded from July 2019 to February 2021 at the University of Osaka. In addition,

parameters were calculated for a total of 114 cases, including 14 cases of speech from healthy individuals without voice disorders recorded from September 2021 to February 2022. Excluding 3 cases in which the jitter, shimmer, and psd used in the ARI could not be calculated, 111 voice samples were used for validation. In addition, the correlation between R_{total} and ARI was calculated to examine the concurrent validity of ARI in 111 voice samples for validation. Furthermore, the diagnostic validity of the ARI was evaluated by the AUC value of the ROC curve (Supplementary Information).

Data availability

Data supporting the results of this study can be found in the text and in Supplementary Information. Recorded audio samples are available for research purposes from the corresponding author upon reasonable request. Source data are provided with this paper.

Code availability

ARI scripts running on Praat are available in the following GitHub repository: https://github.com/LarynxOsaka/ARI.

Received: 8 March 2025; Accepted: 3 May 2025;
Published online: 20 May 2025

References

1. Hirano, M. Psyco-acoustic evaluation of voice. *Clinical Examination of Voice* 81–84 (Springer-Verlag, 1981).
2. Eskenazi, L., Childers, D. G. & Hicks, D. M. Acoustic correlates of vocal quality. *J. Speech Hear. Res.* **33**, 298–306 (1990).
3. Dejonckere, P. & Lebacq, J. Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. *Orl.* **58**, 326–332 (1996).
4. DeBodt, M. S., Wuyts, F. L., VandeHeyning, P. H. & Croux, C. Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *J. Voice* **11**, 74–80 (1997).
5. Yamaguchi, H., Shrivastav, R., Andrews, M. L. & Niimi, S. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. *Folia Phoniatr. Logop.* **55**, 147–157 (2003).
6. Isshiki, N., Okamura, H., Tanabe, M. & Morimoto, M. Differential diagnosis of hoarseness. *Folia Phoniatr. Logop.* **21**, 9–19 (1969).
7. Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J. & Hillman, R. E. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am. J. Speech Lang. Pathol.* **18**, 124–132 (2009).
8. Boersma, P. Praat, a system for doing phonetics by computer. *Glot. Int.* **5**, 341–345 (2001).
9. Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N. & De Bodt, M. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *J. Voice* **24**, 540–555 (2010).
10. Barsties, B. & Maryn, Y. The improvement of internal consistency of the Acoustic Voice Quality Index. *Am. J. Otolaryngol.* **36**, 647–656 (2015).
11. Barsties, B. & Maryn, Y. External validation of the acoustic voice quality index version 03.01 with extended representativity. *Ann. Otol. Rhinol. Laryngol.* **125**, 571–583 (2016).
12. Awan, S. N., Roy, N. & Dromey, C. Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model. *Clin. Linguist. Phonetics* **23**, 825–841 (2009).
13. Awan, S. N., Roy, N., Jette, M. E., Meltzner, G. S. & Hillman, R. E. Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V. *Clin. Linguist. Phonetics* **24**, 742–758 (2010).
14. Latoszek, B. B. V., Maryn, Y., Gerrits, T. & De Bodt, M. The acoustic breathiness index (ABI): a multivariate acoustic model for breathiness. *J. Voice* **31**, 511.e11–511.e27 (2017).
15. van Latoszek, B., De Bodt, M., Gerrits, E. & Maryn, Y. The EXploration of an Objective Model for Roughness with Several Acoustic Markers. *J. Voice* **32**, 149–161 (2018).
16. Latoszek, B. V., Maryn, Y., Gerrits, E. & De Bodt, M. A meta-analysis: acoustic measurement of roughness and breathiness. *J. Speech Lang. Hear. Res.* **61**, 298–323 (2018).
17. Titze, I. *Workshop on Acoustic Voice Analysis: Summary Statement*. (National Center for Voice and Speech, Salt Lake City, UT, USA, 1995).
18. Dejonckere, P. H. & Lebacq, J. An analysis of the diplophonia phenomenon. *Speech Commun.* **2**, 47–56 (1983).
19. Baken, R. J. *Clinical Measurement of Speech and Voice* (Taylor & Francis, 1987).
20. Terhardt, E. On the perception of periodic sound fluctuations (roughness). *Acta Acust.* **30**, 201–213 (1974).
21. Fastl, H. & Zwicker, E. *Psychoacoustics: Facts and Models*. 22 (Springer Science & Business Media, 2006).
22. Ward, P. H., Sanders, J. W., Goldman, R. & Moore, G. P. Lxvii diplophonia. *Ann. Otol. Rhinol. Laryngol.* **78**, 771–777 (1969).
23. Neubauer, J., Mergell, P., Eysholdt, U. & Herzel, H. Patio-temporal analysis of irregular vocal fold oscillations: Biphonation due to desynchronization of spatial modes. *J. Acoust. Soc. Am.* **110**, 3179–3192 (2001).
24. Kimura, M. et al. Arytenoid adduction for correcting vocal fold asymmetry: high-speed imaging. *Ann. Otol. Rhinol. Laryngol.* **119**, 439–446 (2010).
25. Berge, P. ', Y. Pomeau and C. Vidal. *Order within Chaos*. (Wiley, New York, 1986).
26. Glass, L. & Mackey, M. C. *From Clocks to Chaos: the Rhythms of Life*. (Princeton University Press, 1988).
27. Titze, I. Fluctuations and perturbations in vocal output. *Principles Voice Production* 209–306 (Prentice Hall, 1994).
28. Tokuda, I. T. Non-linear dynamics in mammalian voice production. *Anthropol. Sci.* **126**, 35–41 (2018).
29. Deliyski, D. D. Acoustic model and evaluation of pathological voice production. In *Third European Conference on Speech Communication and Technology* 1969–1972 (1993)
30. Aichinger, P. et al. Towards objective voice assessment: the diplophonia diagram. *J. Voice* https://doi.org/10.1016/j.jvoice.2016.06.021 (2017).
31. Awan, S. N. & Awan, J. A. A two-stage cepstral analysis procedure for the classification of rough voices. *J. Voice* **34**, 9–19 (2020).
32. Kitayama, I. et al. Validation of subharmonics quantification using two-stage cepstral analysis. *J. Voice,* published online (2023).
33. Kitayama, I. et al. Robust fundamental frequency-detection algorithm unaffected by the presence of hoarseness in the human voice. *J. Acoust. Soc. Am.* **156**, 4217–4228 (2024).
34. Kramer, E., Linder, R. & Schönweiler, R. A study of subharmonics in connected speech material. *J. Voice* **27**, 29–38 (2013).
35. Hosokawa, K. et al. Acoustic breathiness index for the Japanese-speaking population: validation study and exploration of affecting factors. *J. Speech Lang. Hearing Res.* **62**, 2617–2631 (2019).
36. Hosokawa, K. et al. The Acoustic Voice Quality Index Version 03.01 for the Japanese-speaking population. *J. Voice* **33**, 125.e1–125.e12 (2019).
37. Deliyski, D. D., Shaw, H. S. & Evans, M. K. Adverse effects of environmental noise on acoustic voice quality measurements. *J. Voice* **19**, 15–28 (2005).
38. Deliyski, D. D., Shaw, H. S. & Evans, M. K. Regression tree approach to studying factors influencing acoustic voice analysis. *Folia Phoniatr. Logop.* **58**, 274–288 (2006).

39. Kitayama, I. et al. Intertext variability of smoothed cepstral peak prominence, methods to control it, and its diagnostic properties. *J. Voice* **34**, 305–319 (2020).
40. Frey, L. R. et al. *Investigating Communication: An Introduction to Research Methods*. (1991).

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP21K16842 and JP25K12765.

Author contributions

I.K. and K.H. designed the study, including the key conceptual ideas and the proof-of-concept outline. H.T., T. Ki, T.S., T.T., and Y.T. collected the data. M.O., S.I., and Y.M. evaluated the auditory-perceptual judgments. I.K., K.A., and T. Ka analyzed the data and prepared the paper. H.I. and A.M. supervised the project. All authors participated in the discussion of the results and critically reviewed and approved the final paper.

Competing interests

The authors declare no competing interests. I.K., K.H., and H.I. have applied for a patent related to this work (patent applicant, The University of Osaka; names of inventors, Itsuki Kitayama, Kiyohito Hosokawa and Hidenori Inohara; application number, 2024-087318; status of application, substantive examination; specific aspect of paper covered in patent application, fundamental frequency estimation algorithm that is robust and unaffected by voice quality).

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01702-2.

Correspondence and requests for materials should be addressed to Kiyohito Hosokawa.