The Acoustic Voice Quality Index Version 03.01 for the Japanese-speaking Population

***,†,‡Kiyohito Hosokawa, §,¶Ben Barsties v Latoszek, ‡Toshihiko Iwahashi, #Mio Iwahashi, **Shinobu Iwaki, ‡Chieri Kato, ††Misao Yoshida, ‡,‡‡Hisanori Sasai, ‡‡Akira Miyauchi, *,‡Naoki Matsushiro, ‡Hidenori Inohara, ‡Makoto Ogawa, and §,§§,¶¶Youri Maryn,** **†‡#Osaka and **††‡‡Hyogo, Japan, §§§Antwerp and ¶¶Ghent, Belgium, and ¶Nijmegen, The Netherlands*

Summary: Objectives. We aimed to determine the most appropriate syllable number for analyzing the Acoustic Voice Quality Index for the Japanese-speaking population (AVQIv3-JP) and to validate AVQIv3-JP using the determined syllable number.

Methods. First, we counted how many syllables should be included in each continuous speech (CS) sample to achieve time-balanced analysis between CS and sustained vowel samples using our previous dataset including 336 CS samples with 58 syllables. From the descriptive statistics of the counted syllable numbers, the most appropriate syllable number was identified. Subsequently, we performed validation procedures of AVQIv3-JP using our latest dataset including 455 recordings.

Results. Thirty Japanese syllables were judged to be the most appropriate syllable number. The concurrent validity of the AVQIv3-JP using 30 syllables was confirmed by Spearman's rho of 0.873. Subsequently, the receiver operating characteristic analysis demonstrated the excellent discriminative capability of AVQIv3-JP, showing the area under the curve of 0.915. The AVQIv3's original threshold of 2.43 in the Dutch language corresponded to sensitivity and specificity of 64.6% and 97.3%, respectively. In the present study, a threshold of 1.41 achieved the best accuracy with balanced sensitivity and specificity of 84.4% and 85.6%, respectively. Furthermore, the 95th percentile of the control participants exhibited a threshold of 2.06, showing sensitivity and specificity of 72.1% and 93.8%, respectively, as well as reasonable positive and negative likelihood ratios of 11.7 and 0.298, respectively.

Conclusion. The AVQIv3 using 30 Japanese syllables is a reliable measurement tool for estimating the severity of voice quality and detecting abnormal voices.

Key Words: Acoustic voice quality index–Voice disorders–Auditory-perceptual judgment–The Japanese language–Multivariate acoustic measurement.

INTRODUCTION

In the clinical course for patients with pathological voice, clinicians such as laryngologists and speech-language therapists should assess patient's vocal and laryngeal functions through multidimensional methodologies to comprehend the problems related to voice abnormalities.[1,2] The guideline from the European Laryngological Society[1] advocated a basic set for the assessment of common dysphonia including the auditory-perceptual judgments; videostroboscopic, acoustic, and aerodynamic assessments; and patients' subjective ratings. In particular, the auditory-perceptual judgment and acoustic measurement are directly concerned with the voice itself, which can be a chief complaint and a primary outcome for therapeutic interventions. These assessments provide different points of view in the estimation of the severity degree of voice quality, which is one of the most essential and crucial evaluations when clinicians diagnose and treat patients with voice problems.

In the actual clinical scene, we can use several kinds of assessment tools for the auditory-perceptual judgments, such as the Grade, Rough, Breathy, Asthenic, and Strained (GRBAS) scale[3] and the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V).[4] In particular, the GRBAS scale has often been used worldwide as well as in Japan for 50 years since its establishment, thanks to the efforts of Isshiki et al,[5] Takahashi and Koike,[6] and the Japan Society of Logopedics and Phoniatrics.[3] On the other hand, the CAPE-V has been advocated by the Consensus Conference on Auditory-Perceptual Evaluation of Voice (2002) and sponsored by the American Speech-Language-Hearing Association.[4] Because the CAPE-V was intended to be a standardized protocol in kinds of tasks, speech stimuli, and recording techniques, the reading contexts for continuous speech (CS) tasks are fixed and written in the English language. Therefore, the CAPE-V has never been used in Japan due to its low English-speaking population.

In addition, the auditory-perceptual judgments have been considered to be the gold-standard method of voice quality evaluation because the abnormality in voice quality is perceived based on the recognition process of brain systems.[7,8] However, due to the

Accepted for publication October 2, 2017.

Conflict of interest: The authors report no conflicts of interest.

From the *Department of Otorhinolaryngology, Japan Community Health care Organization (JCHO) Osaka Hospital, Osaka, Japan; †Department of Otorhinolaryngology, Osaka Police Hospital, Osaka, Japan; ‡Department of Otorhinolaryngology and Head & Neck Surgery, Osaka University Graduate School of Medicine, Osaka, Japan; §Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium; ¶Institute of Health Studies, HAN University of Applied Sciences, Nijmegen, The Netherlands; #Nimura ENT Voice Clinic, Osaka, Japan; **Department of Otorhinolaryngology and Head & Neck Surgery, Kobe University Graduate School of Medicine, Hyogo, Japan; ††Department of Rehabilitation, Nishinomiya Kaisei Hospital, Hyogo, Japan; ‡‡Department of Surgery, Kuma Hospital, Hyogo, Japan; §§European Institute for ORL, Sint-Augustinus Hospital, Antwerp, Belgium; and the ¶¶Faculty of Education, Health & Social Work, University College Ghent, Ghent, Belgium.

Address correspondence and reprint requests to Kiyohito Hosokawa, Department of Otorhinolaryngology, Japan Community Health care Organization (JCHO) Osaka Hospital, 4-2-78 Fukushima, Fukushima-ku, Osaka-city, Osaka 553-0003, Japan. E-mail: khosokawa@wg8.so-net.ne.jp

Journal of Voice, Vol. 33, No. 1, pp. 125.e1–125.e12

0892-1997

https://doi.org/10.1016/j.jvoice.2017.10.003

abovementioned psychological nature, the results of the judgments always fluctuate among multiple raters or even among multiple ratings by an identical rater,[9–15] although relatively higher agreements were reported in the GRBAS scale.[16–18] Therefore, a single evaluation by the methods considerably limits the objectivity of the assessment. To obtain objectivity, researchers have made a lot of effort to reduce the affecting factors related to scale, the listener, and stimulus.[12,13,19–24]

Due to these limitations of auditory-perceptual judgments, various methodologies of acoustic measurements have been developed to objectify specific characteristics in the shapes of acoustic waveform.[25–32] For example, jitter and shimmer measures represent cycle-to-cycle variabilities in the periods and amplitudes of acoustic waveforms, respectively,[6,26–28] whereas the harmonics-to-noise ratio (HNR) quantifies the relative amount of glottal noises.[29] Between these acoustic measures and auditory-perceptual judgment, significant correlations were found in a variety of voice research.[28,31–36] However, several problems were reported with regard to the utility of these traditional acoustic measures in the evaluation of voice quality.[37,38] One of the problems is related to the fact that increases in these traditional measures in dysphonic voices are based on the uniformity of successive waveforms in sustained vowels (SVs) in normophonic speakers.[39] For example, stable measurement of perturbation measures such as jitter and shimmer requires at least 100 successive waveforms.[40,41] Therefore, the voice quality of CS samples is difficult to estimate because CS samples have considerable differences in waveform pattern among contained syllables even in normophonic speakers. According to the meta-analysis for acoustic measures of Maryn et al,[42] however, some measures based on cepstral analysis correlate well with auditory-perceptual judgments even in CS samples. In particular, the smoothed cepstral peak prominence (CPPS) developed by Hillenbrand et al[32] was one of the most correlated measures both in CS and SV samples.

For such situations, Maryn et al established the Acoustic Voice Quality Index (AVQI) that enables estimating the deviance of overall voice quality[43,44] (ie, a single grade of hoarseness level for the concatenated samples of CS and SV tasks). First, AVQI version 01 (AVQIv1) was developed by means of a stepwise multiple regression analysis that used 13 kinds of acoustic measures as independent variables and the overall voice quality as a dependent variable.[43] From the statistical procedure, AVQIv1 was defined as a multivariate construct of six measures of CPPS from "*SpeechTool*" (James Hillenbrand; Western Michigan University, Kalamazoo, MI, USA—http://homepages.wmich.edu/~hillenbr/)[32] and HNR, shimmer local (SL), shimmer local dB (SLdB), general slope of the spectrum (Slope), and tilt of the regression line through the spectrum (Tilt) from "*Praat*" (Paul Boersma and David Weenink; Institute of Phonetic Sciences, University of Amsterdam, The Netherlands—http://www.praat.org/).[45] Subsequently, owing to the recent implementation of CPPS in *Praat*, AVQI was modified as AVQIv2 to be calculated using a single software (*Praat*[45]). The criterion-related concurrent validities and diagnostic accuracies of AVQIv1 or AVQIv2 were further examined for other languages even in different language families, namely, some Western-European languages (ie, Dutch,[44,46,47] German,[47,48] English,[47,49] French,[47] and Finnish[50]), one Altaic language (ie, Korean[51]), one Japonic language (ie, Japanese[52]), and one Indo-European language (ie, Lithuanian[53]). In all these studies, AVQIv2 was shown to be relatively insulated from the effect of interlanguage phonetic differences. Furthermore, the third version of AVQI (AVQIv3) was developed to have balanced out time lengths for analysis between CS and SV samples, leading to more ecologically valid measurement.[54] Indeed in AVQIv2, the time lengths of analyzed CS samples tended to be shorter than those of SV samples.[52,54] The previous research showed that mean durations of analyzed CS samples were 1.99 and 1.53 seconds in the Japanese (with 22 syllables)[52] and the Dutch[54] (with 17 syllables) languages, respectively, whereas those of SV samples were always 3 seconds. These results indicated that a greater syllable number should be included in CS samples for AVQIv3. In the definition process of AVQIv3 in the Dutch language,[54] 60 equally distributed voice recordings in each severity degree were prepared to revise the regression coefficients of the AVQI formula. In this process, each CS sample was trimmed by hand to have an individual syllable number which most approximates its analyzed time length to 3 seconds. Subsequently, "the standardized syllable number" was estimated, which was defined as the most robust syllable number which could be used for all the voice samples to approximate the time lengths of analyzed CS and SV samples. For the validation process of AVQIv3,[55] a large number of voice samples, more than a thousand, were assessed, resulting in the marked concurrent validity and high diagnostic accuracy.

However, we cannot use the same value of the standardized syllable number without any definition process because of the differences between the Dutch and Japanese language characteristics. In particular, we speculated that two different characteristics would affect the value of the standardized syllable number: one was that Japanese essentially allowed only "open syllable," which ends with one vowel (or diphthong) without being followed by any consonant,[56,57] and the other was that the Japanese language was categorized as a syllable-timed language, whereas the Dutch language as a stress-timed language.[56,57] Thus, we expected that a large amount of open syllable and syllable-timed pronunciation in the Japanese language would require a lesser syllable number than the Dutch language.

The first aim of the present study was to determine the appropriate syllable number for AVQIv3 in the Japanese language. The second aim was to validate the AVQIv3 using the above-defined syllable number in the Japanese-speaking population. The focus of the second investigation was to assess the feasibility and robustness of AVQIv3, its concurrent validity, responsiveness to change, and diagnostic accuracy in differentiating between normal and hoarse voices.

PARTICIPANTS AND METHODS

This study comprised two phases of the investigation. The first phase was conducted to examine how many syllables should be included in the Japanese CS reading text to achieve equally balanced time lengths of CS and SV analysis. The second phase was performed to estimate the criterion-related concurrent va-

TABLE 1.
Voice Diagnoses and Interventions for the Eligible Normophonic and Dysphonic Patients

Diagnosis	Number of Patients		Interventions
	Before	After	
Paresis/paralysis	117	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)
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	
Androphonia	1	0	
Hysterical aphonia	1	0	
Tremor	1	0	
Normophonic controls	55	0	
Total	344	111	

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

lidity and responsiveness to change as well as the diagnostic accuracy of AVQIv3-JP using the syllable number determined in the first phase.

Participants for the first phase

For the first phase to determine the standardized syllable number for AVQIv3-JP, we prepared 336 recordings of CS samples with a variety of voice etiologies from our previous study on AVQIv2-JP.[52] These samples were derived from the database with 729 voice recordings at the Department of Otorhinolaryngology and Head and Neck Surgery at the Osaka University Hospital (November 2010 to July 2014), the Osaka Police Hospital (April 2014 to September 2015), and the Department of Surgery at Kuma Hospital (June 2012 to August 2014).

Participants for the second phase

For the second phase to confirm the criterion-related concurrent validity and diagnostic accuracy of AVQIv3-JP, we added the latest 207 voice recordings obtained at the Osaka Police Hospital (October 2015 to September 2016) to the former database (a total of 936 recordings). Subsequently, the eligibility of the voice recordings for the validation study was assessed according to the criteria which were similar to those in our previous study.[52] Namely, the inclusion criteria were as follows: (1) voice recordings at initial assessments and corresponding voice recordings after therapeutic interventions or (2) recordings of normophonic voices from otorhinolaryngological patients with neither lesion or voice complaint. The exclusion criteria were the following: (1) misreadings of the CS or (2) high environmental noise level (ie, signal-to-noise ratio of <30 dB).[58,59]

As a result, a cumulative total of 455 voice recordings from 344 participants (male: 133, female: 211, mean age ± standard deviation (SD): 57.5 ± 15.6) were judged to be eligible for the validation study. Of the 455 voice recordings, 55 were from participants without any voice complaints, 289 were from participants with various organic and nonorganic voice disorders with varying degrees of dysphonia, and 111 were the voice recordings at least 3 months after interventions (332 voice recordings were those used in the first phase). Table 1 summarizes the diagnoses and interventions of the 344 participants for the validation procedure. The institutional review board of each hospital approved this retrospective study.

Voice recordings

All recordings were conducted in the same manner as the previous study,[52] that is, in a sound-treated room with a head-worn microphone SE50 (Samson Technologies Corp., Hauppauge, NY) digitized at a sampling rate of 44.1 kHz and resolution of 16 bits using a linear PCM recorder H4n (Zoom Corp., Tokyo, Japan). The participants sustained a vowel /a:/ for more than 3 seconds, and read a text at a comfortable pitch, loudness, and speed. As a reading text for CS samples, the title and first and second sentences of the Japanese translation of "The North Wind and the Sun" with a total of 65 syllables[39] were used.

The first phase: preparation of voice samples

The aim of the first phase was to determine the standardized syllable number suitable for the Japanese language. In this procedure, the first and second sentences with 58 syllables (/aruhi kitakaze to taiyo ga chikara kurabe wo shimashita/ and /tabibito no gaito wo nugasetaho ga kachitoyukoto ni kimete mazu kazekara hajimemashita/ in Japanese phonemes) were adopted for the analysis because the full text of the title (/kitakaze to taiyo/) was included in the first sentence and regarded to be redundant.

The first phase: determination of the standardized syllable number

In the AVQI script,[43,54] CS samples are analyzed in the form of only-voiced CS (vCS) samples without silent or consonant segments, and the durations of SV samples are fixed to 3.0 seconds in all the cases. Therefore, to achieve time-balanced analysis, the time length of vCS samples should be equivalent to 3.0 seconds.

First, we performed sample-by-sample identification of the hand-marked syllable number for each CS sample with 58 syllables (CS_{58} samples), details of which are illustrated in Figure 1. The hand-marked syllable number was defined as an ordinal number of the last syllable which made each vCS sample most approximated to 3.0 seconds. To generate vCS_{58} samples, we treated all the CS_{58} samples with the preanalysis processing (Part 0 and 1 in the AVQI version 03.01 script[54]). Then, the first 3-second vCS samples were prepared. In this process, 13 vCS samples were excluded from this analysis because the durations were too short (less than 2.9 seconds) to estimate the accurate syllable number which made them most approximated to 3.0 seconds. Subsequently, we judged whether the full length of the last syllable should be included or discarded, based on the comparison of the durations between vCS samples with and without

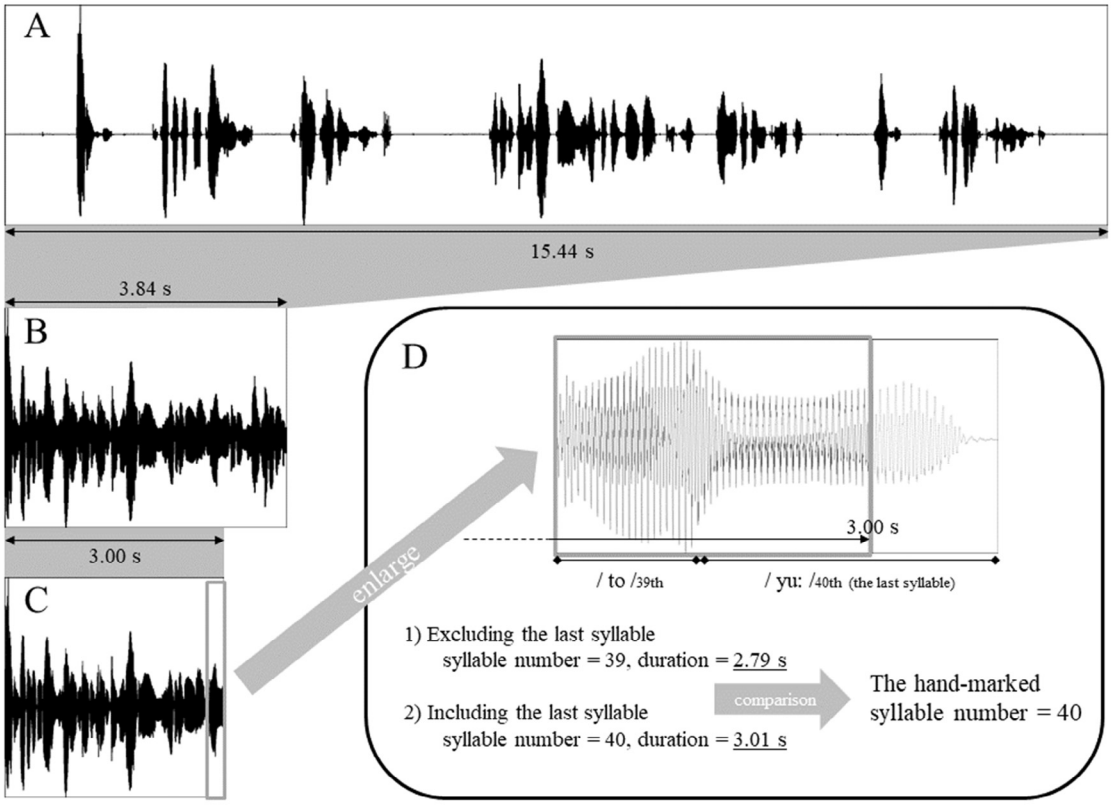

FIGURE 1. Schematic image of the methods identifying the hand-marked syllable number. Waveform A is a CS sample with 58 syllables. Waveform B is the corresponding only-voiced CS sample with 58 syllables (vCS_{58} sample). Waveform C is the first 3 seconds of the vCS_{58} sample. Waveform D is the enlarged waveform of the last part of the first 3 second of the vCS_{58} sample, which reveals that Waveform C terminates at the middle of the 40th syllable. Comparison of the durations between vCS_{39} and vCS_{40} samples revealed that inclusion of the 40th syllable achieves the most approximated duration to 3 seconds, which determines the hand-marked syllable number for the recording.

the full length of the last syllable. Then, the hand-marked syllable numbers for all the 323 CS samples were counted. All the trimming and sound processing described above were performed using the *Praat* software.[45]

Next, in the previous study developing AVQIv3 for the Dutch language,[54] Barsties et al. determined the standardized syllable number, which was a single number applicable for any individuals, to facilitate sample preparation for AVQI analysis. Therefore, we proceeded to define the standardized syllable number for the Japanese language. To determine the standardized syllable number, we calculated the descriptive statistics of the hand-marked syllable numbers. In the original Dutch study,[54] the rounded integer of the lower confidence interval (CI) of the hand-marked syllable number (ie, 34) had matched best to the standardized syllable number. In this study, we regarded the rounded lower CI as well as its adjacent integers as the candidates for the standardized syllable number. Subsequently, we compared the durations of vCS samples and AVQI values with the hand-marked syllable number with those with the three syllable number candidates to define the syllable number with the least difference in durations to be the standardized syllable number. Finally, we confirmed the consistency between the AVQI values using the above-defined standardized syllable number and those using the hand-marked syllable numbers.

The second phase: preparation of voice samples

Using the standardized syllable number determined above, we investigated the validity of AVQIv3-JP. For the validation procedure, all the CS samples for 455 recordings were trimmed to the CS samples with the standardized syllable number (CS_{st} samples). Subsequently, the CS_{st} samples and the SV samples of the 3-second mid-vowel segment proceeded to the calculation of AVQIv3. For the SV samples, an aphonic part, as well as the beginning and end parts of the vowel sample, was allowed to be included only in cases where a patient could sustain for less than 3 seconds.

The second phase: auditory-perceptual judgment

For auditory-perceptual judgments, the CS_{st} and SV samples were concatenated with a 1-second soundless interval in a similar manner to that described by Maryn et al.[43] Two otolaryngologists and three speech therapists who are native Japanese speakers and are experienced in the diagnosis, evaluation, and therapy for voice disorders rated the concatenated samples using the GRBAS scale.[3]

For the judgments, the "GRBAS checker 7.3" (Figure 2), developed by the first author (K.H.) using the Visual Basic for Applications 7.1 integrated in Microsoft Excel 2013 (Microsoft Corp., Redmond, WA), was utilized to facilitate the rating procedure and increase the rater reliability. The GRBAS checker 7.3 enabled the raters to do the following: (1) listen to all the concatenated samples in a random order, (2) make independent judgments for the CS_{st} and SV samples listening to a concatenated sample, (3) be blinded for any information related to the participants, (4) listen to samples anytime they needed to, (5) always revise previous ratings, (6) re-rate 45 duplicated samples (10% of 455 samples in this study) at the end of the procedure to assess the intra-rater reliability, and (7) refer to anchor

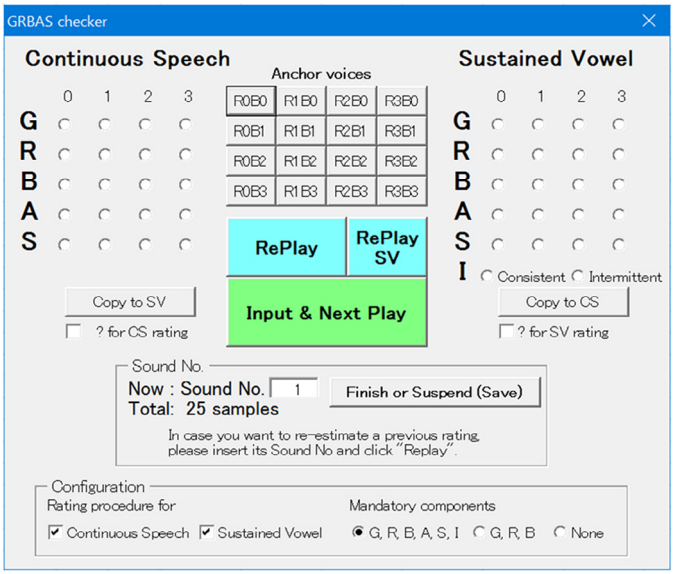

FIGURE 2. Print screen of the Microsoft Excel VBA for the GRBAS checker 7.3.

voices anytime they needed to. The anchor voices represented SV samples corresponding to every severity combination of R and B score (0 to 3 each). As to the anchor voices, two raters with the highest intra- and inter-rater reliabilities in the previous study[52] selected them from our database or an educational software for evaluating the GRBAS scale (*Douga de miru Onseishougai* version 1.0, which means "The voice disorders examined with movies," Interuna Publishers, Inc. Tokyo, Japan). To control the internal affecting factors such as fatigue, attention, and lapse in concentration,[7] the raters were allowed to take a short break after every 25th rating.

Next to the auditory-perceptual judgments, the intra- and inter-rater reliabilities of five raters were assessed to exclude unreliable judgments from further analysis. After the exclusion, the averages of G scores among the raters were calculated separately for the CS_{st} and SV samples (G_{cs} and G_{sv}, respectively). Subsequently, the overall voice quality (G_{total}) for each concatenated sample was defined as the average of the G_{cs} and G_{sv}.[60]

The second phase: calculation of AVQI for the validation procedure

The AVQIv3 also uses the six acoustic parameters (viz., CPPS, SL, SLdB, HNR, Slope, and Tilt) in the regression model with different coefficients from those in the older versions.[54] The AVQIv3 for 455 voice recordings were calculated using a *Praat* script described by Barsties and Maryn.[54] In this script, the regression formula of AVQIv3 was defined as $(4.152 - (0.177 \times \text{CPPS}) - (0.006 \times \text{HNR}) - (0.037 \times \text{SL}) + (0.941 \times \text{SLdB}) + (0.01 \times \text{Slope}) + (0.093 \times \text{Tilt})) \times 2.8902$.

Using these variables for the 455 voice recordings, we estimated the criterion-related concurrent validity and validity related to the responsiveness to change as well as the diagnostic accuracy with the best threshold of AVQIv3-JP. In addition, to generalize the results, the above-defined validity should be confirmed again using a cohort without patients with paresis/

paralysis because too much number of recordings from the patients with paresis/paralysis was included in the cohort of 455 recordings (36.7%: 167 paresis/paralysis in the 455 voice recordings). Therefore, the above-defined validity was also assessed for a cohort of 288 voice recordings without these patients.

Statistical analysis

First, to compare the durations between the vCS_{hm} samples and the vCS samples with the syllable numbers around the lower CI, we performed the Wilcoxon signed-rank test. The best candidate for the standardized syllable number was determined by the least effect size among them. Subsequently, we calculated the intraclass correlation (1, 1) to confirm the consistency between the AVQI values using the standardized and the hand-marked syllable numbers.

Second, to evaluate the intra- and inter-rater reliabilities for the auditory-perceptual judgment in the G_{cs}/G_{sv} scores, we calculated the Cohen kappa (Cκ) and the Fleiss kappa (Fκ) coefficients, respectively. Raters were regarded to be unreliable in the case where their auditory-perceptual judgments in the G_{cs} or G_{sv} were below the guidelines for the interpretation of the κ statistics, which were provided by Landis and Koch.[61]

Third, the criterion-related concurrent validity of the AVQIv3-JP and the validity related to the responsiveness to change were estimated using the Spearman rank-order correlation coefficient (r_s) and the coefficient of determination (r^2). The concurrent validity was confirmed between AVQI and G_{total} on 455 samples, whereas the responsiveness to change was validated between ΔAVQI and ΔG_{total} (the difference between pre- and postintervention) among 222 samples from 111 patients (ie, before and after interventions). Interpretation guidelines for r_s were provided by Frey et al.[62]

Fourth, to evaluate the perceptual diagnostic accuracy of the AVQIv3-JP, the receiver operating characteristic (ROC) curve analysis was performed. In this procedure, either $G_{cs} \geq 0.5$ or $G_{sv} \geq 0.5$ were regarded as an indicator of dysphonic voices because a participant should be considered as having abnormal quality if either of CS or SV samples was judged to be dysphonic by the majority of the raters. In the ROC curve analysis, three definitions of threshold level were examined. An AVQI of 2.43, which had been proposed to be the best threshold level in the Dutch language,[54] was a candidate for suitable threshold level also for the Japanese language. Moreover, the best threshold levels provided by the Youden Index and the upper 95th percentile of AVQI for the normophonic voices were also considered to be possible thresholds. The Youden Index produces the coordinate providing the maximum of (sensitivity + specificity − 1), whereas the 95th percentiles in normal cohorts are generally employed as upper limits of clinical measurements. Subsequently, their applicabilities for clinical decision making were examined by the balance between the "likelihood ratio for a positive result" (LR+) and "likelihood ratio for a negative result" (LR−), which are defined as the sensitivity/(1 − specificity) and (1 − sensitivity)/specificity, respectively. As a general guideline, the diagnostic value of a measure is considered to be high when LR+ is ≥10 and LR− is ≤0.1.[63]

The assessments related to the effect sizes and the kappa coefficients were done using *R version 2.8.1* (R Core Team, Vienna, Austria). The other statistical analyses were performed using the *JMP Pro version 12.2.0* (SAS Institute Inc., Cary, NC). All results were considered statistically significant at $P < 0.05$.

RESULTS

The first phase: determination of the standardized syllable number

Figure 3 indicates the frequency distribution of different hand-marked syllable numbers for 323 CS samples. The 95% CI ranged from 30.1 to 31.3 syllables, whereas the mean and median values

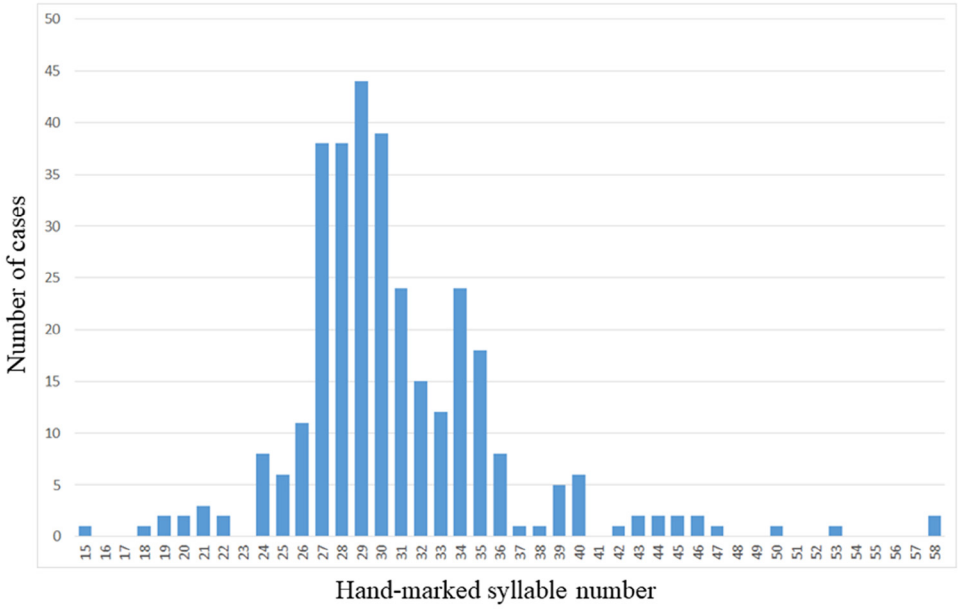

FIGURE 3. Frequency distribution of different hand-marked syllable numbers.

TABLE 2.
Comparison of the Durations and AVQI Values Between vCS_{hm} Samples and vCS Samples With Each Candidate of the Standardized Syllable Number

	Duration				AVQI			
Sample Type	Mean	SD	*P* Value	Effect Size	Mean	SD	*P* Value	Effect Size
vCS_{hm} samples	2.95 s	0.35			2.45	2.53		
vCS_{29} samples	2.88 s	0.71	0.004	0.157	2.43	2.55	0.167	0.076
vCS_{30} samples	3.02 s	0.75	0.104	0.087	2.44	2.56	0.513	0.036
vCS_{31} samples	3.14 s	0.76	<0.001	0.282	2.42	2.54	0.043	0.110

Abbreviations: AVQI, Acoustic Voice Quality Index; SD, standard deviation; vCS_{hm}, only-voiced continuous speech sample with a hand-marked syllable number. *Notes:* vCS_{29}, only-voiced continuous speech sample with 29 syllables; vCS_{30}, only-voiced continuous speech sample with 30 syllables; vCS_{31}, only-voiced continuous speech sample with 31 syllables.

resulted in 30.7 and 30 syllables, respectively. Therefore, we considered that 30 syllables and its adjacents of 29 and 31 syllables were appropriate candidates for the standardized syllable number. To identify the best candidate, we compared the time lengths of the vCS samples with 29, 30, and 31 syllables (vCS_{29}, vCS_{30}, and vCS_{31}, respectively) with those of the vCS_{hm} samples. Table 2 shows that the vCS_{30} samples achieved the least effect size in durations compared with the vCS_{hm} samples, demonstrating that the application of 30 syllables for CS samples was the best substitute for sample-by-sample identification of the hand-marked syllable number, namely, the standardized syllable number. Furthermore, we compared the AVQIv3 values using 30 syllables with those using the hand-marked syllable numbers (Table 2). The AVQI values using the CS_{30} samples were proved to be the nearest to those using the CS_{hm} samples, supported by the least effect size between them. In addition, Figure 4 illustrates the almost perfect consistency with the intraclass correlation (1, 1) of 0.998.

The second phase: reliability of the auditory-perceptual judgments

The intra-rater reliabilities in the five raters for the G_{cs} and G_{sv} were separately assessed (Table 3). The Cκ values in the G_{cs} ranged from 0.578 to 0.756, whereas those in the G_{sv} ranged from 0.481 to 0.805, demonstrating moderate to high reliabilities. Next, the inter-rater reliabilities were evaluated among them. The Fκ values in the G_{cs} and G_{sv} were 0.399 and 0.407, respectively, judged to be of borderline reliabilities (Table 4). Therefore, we decided to discard the ratings that were mostly responsible for the decrease in reliability. The comparison of Fκ values among any combinations of four raters revealed that Rater 5 most affected the inter-rater reliability. Subsequently, Rater 6, who was also a native Japanese-speaking speech therapist, joined the rater panel. The intra-rater reliabilities of Rater 6 showed the Cκ values of 0.479 and 0.573 in the G_{cs} and G_{sv} ratings, respectively, judged to be moderately reliable. Consequently, the Fκ values in the G_{cs} and G_{sv} among Raters 1, 2, 3, 4, and 6 were 0.441 and 0.458, respectively, demonstrating moderate reliabilities. Finally, the G_{total}, as well as the G_{cs} and G_{sv}, was calculated using the averages among Raters 1, 2, 3, 4, and 6. Figure 5 illustrates the frequency distributions of the G_{cs} and G_{sv} as well as the diagnostic status of dysphonia, whereas Figure 6 shows those of the G_{total} ratings.

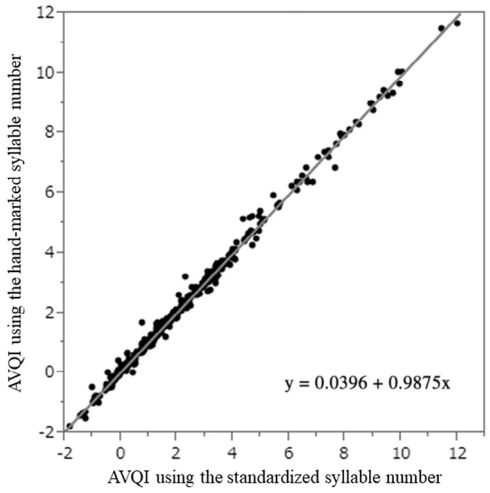

FIGURE 4. The scatter plot illustrates the almost perfect consistency between the AVQI values using the hand-marked syllable number and those using the standardized syllable number.

The second phase: criterion-related concurrent validity

We calculated the AVQIv3-JP values for the 455 voice recordings using the standardized syllable number of 30 syllables. Only one case with the G_{total} of 3.0 could not produce the AVQIv3-JP value due to the failure in calculating the time-domain measures (SL, SLdB, and HNR). Therefore, a total number of 454 voice recordings were used for the validation procedures of AVQIv3-JP.

TABLE 3.
The Intra-rater Reliabilities of Auditory-perceptual Ratings for All the Raters

Cohen's κ	G_{cs}	G_{sv}
Rater 1	0.578	0.757
Rater 2	0.630	0.577
Rater 3	0.756	0.805
Rater 4	0.657	0.655
Rater 5	0.695	0.481
Rater 6	0.479	0.573

Abbreviations: G_{cs}, voice quality rating for continuous speech; G_{sv}, voice quality rating for sustained vowels.

TABLE 4.
Different Inter-rater Reliabilities of Auditory-perceptual Ratings for Each Combination of the Raters

Fleiss κ							
G_{cs}	G_{sv}	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6
0.399	**0.407**	○	○	○	○	○	×
0.366	0.377	×	○	○	○	○	×
0.423	**0.417**	○	×	○	○	○	×
0.395	0.386	○	○	×	○	○	×
0.367	0.375	○	○	○	×	○	×
0.434	**0.472**	○	○	○	○	×	×
0.441	**0.458**	○	○	○	○	×	○

Note: ○ or × indicates that the corresponding raters are included or excluded, respectively. The bolded values indicate acceptable reliability.
Abbreviations: G_{cs}, voice quality rating for continuous speech; G_{sv}, voice quality rating for sustained vowels.

The criterion-related validity of AVQIv3-JP was assessed using a concurrent measure of the G_{total}. Figure 7 shows the proportional relationship between G_{total} and AVQIv3-JP. The linear regression line and r^2 statistic was estimated as $y = -0.145 + 2.827x$ and 0.817, respectively. In addition, the bivariate correlation $r_s = 0.873$ indicated a strong concurrent validity of AVQIv3-JP.

Moreover, we evaluated the validity related to whether AVQIv3-JP linearly responded to the change in overall voice quality after interventions. The proportional relationship between the ΔG_{total} and ΔAVQIv3-JP resulted in the linear regression line and r^2 statistic as $y = -0.188 + 3.040x$ and 0.778, respectively (Figure 8). In addition, the bivariate correlation $r_s = 0.878$ indicated a strong validity of the AVQIv3-JP with respect to responsiveness to change.

The second phase: diagnostic accuracy

We further estimated the diagnostic accuracy of AVQIv3-JP using the standardized syllable number of 30 syllables based on the presence of dysphonic voices. This analysis defined either the $G_{cs} \geq 0.5$ or $G_{sv} \geq 0.5$ as an indicator of abnormal voice quality. The ROC analysis with the area under the curve (A_{ROC}) of 0.915 demonstrated the excellent diagnostic accuracy of AVQIv3-JP (Figure 9). The original AVQIv3 threshold of 2.43[54] exhibited a sensitivity of 64.6% and specificity of 97.3% with a higher LR+ of 23.6 and LR− of 0.364. In contrast, the best diagnostic accuracy was achieved by the Youden Index, giving a decreased cutoff value of 1.41 with a sensitivity of 84.4% and specificity of 85.6% as well as an LR+ of 5.87 and LR− of 0.182. Furthermore, the normative threshold of AVQIv3-JP defined by the 95th percentile of the control participants was calculated to

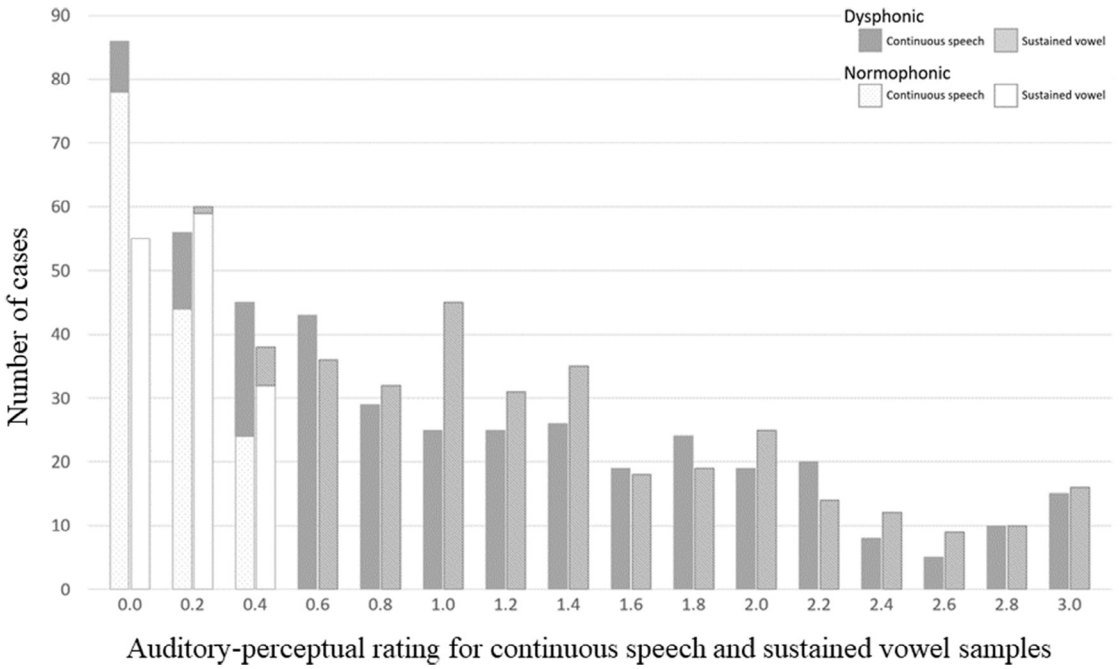

FIGURE 5. Frequency distributions of the G_{cs} and G_{sv} ratings as well as the diagnostic status of dysphonia.

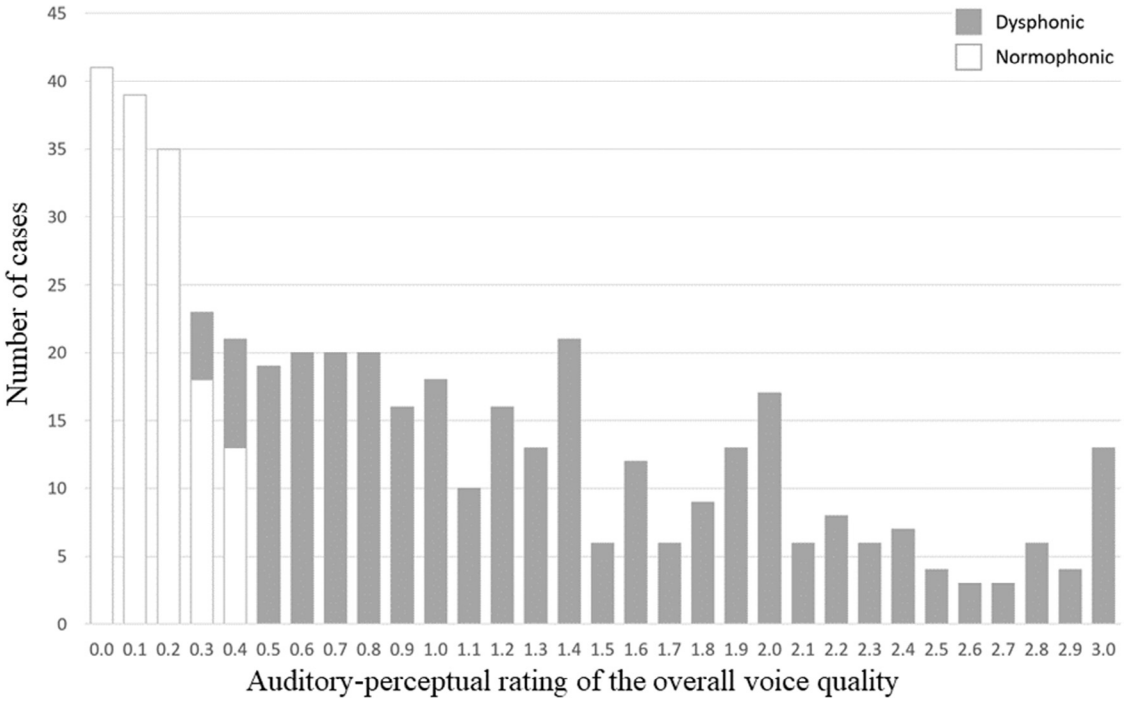

FIGURE 6. Frequency distribution of the G_{total} ratings as well as the diagnostic status of dysphonia.

be 2.06, producing a sensitivity and specificity of 72.1% and 93.8% with balanced LR+ and LR− of 11.7 and 0.298, respectively.

The second phase: confirmation of the results

To generalize the results of the validation procedure, we verified whether three kinds of validity were also relevant for a cohort of 288 voice recordings from the participants without paresis/paralysis. The criterion-related concurrent validity and validity related to responsiveness to change were indicated by the bivariate correlation of $r_s = 0.845$ and 0.812, respectively. Moreover, the excellent diagnostic accuracy was confirmed by A_{ROC} of 0.906. The threshold of AVQIv3-JP of 2.06 showed a sensitivity of 65.6% and specificity of 97.1% with LR+ of 22.6 and LR− of 0.354.

DISCUSSION

To achieve a higher ecological validity, the AVQI was first developed to be a comprehensive voice assessment including both SV and CS materials.[43] However, the previous versions of AVQI tended to include less time length in CS than that of SV samples for the calculation process of the six acoustic measures. In this

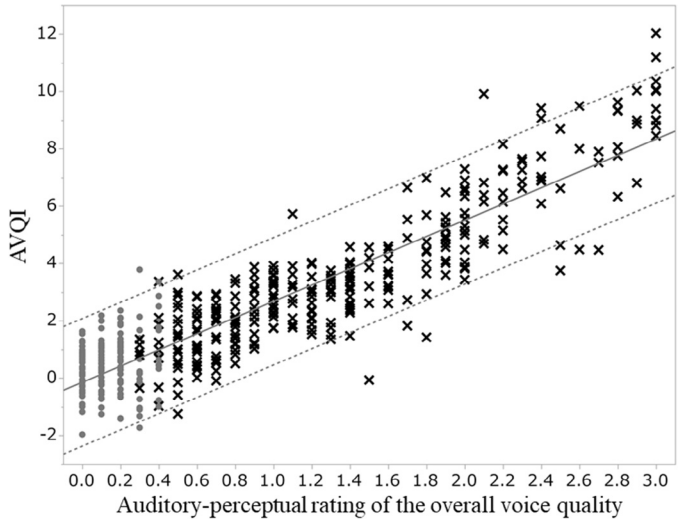

FIGURE 7. The scatter plot illustrates the concurrent validity of AVQI. The two dotted lines above and under the regression line delineate the upper and lower boundaries of 95% confidence interval, respectively.

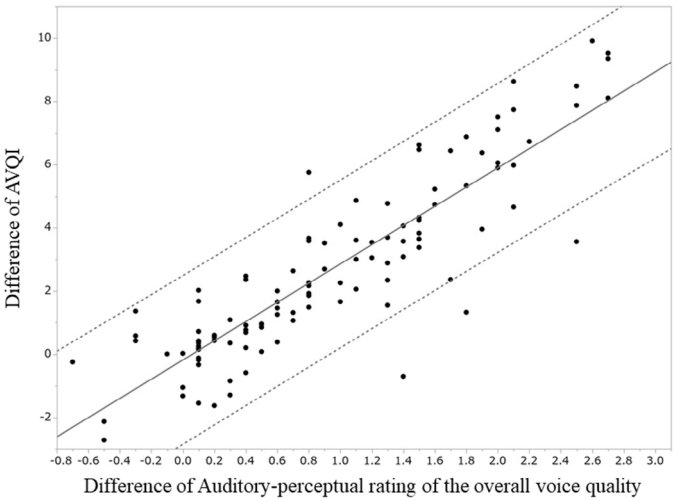

FIGURE 8. The scatter plot illustrates the validity related to the responsiveness to change of AVQI. The two dotted lines above and under the regression line delineate the upper and lower boundaries of 95% confidence interval, respectively.

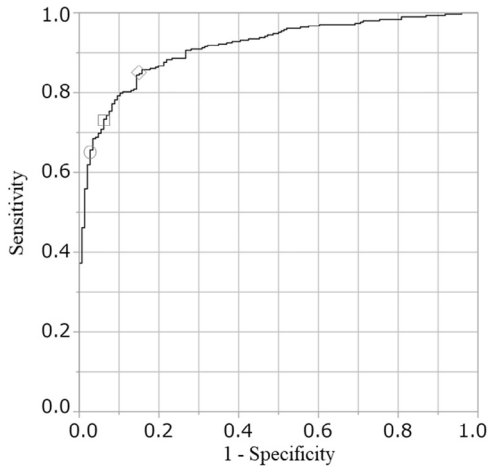

FIGURE 9. The ROC curve illustrates the excellent diagnostic accuracy of AVQI. The ○, ◊, and □ indicate the coordinates of the sensitivity and (1 − sensitivity) corresponding to the threshold derived from the original study in the Dutch language, the threshold giving the best diagnostic accuracy, and the threshold determined by the 95th percentile from the control participants.

situation, the values of AVQI dominantly represented the disability in sustaining voices, leading to underestimation of disturbance in daily conversation. To improve the ecological validity further, AVQIv3 was developed to equally estimate the severity degree of both speech types by taking a balance of time lengths between them.[54] As a result, the original AVQIv3 in the Dutch language determined the most suitable syllable number for the CS reading text as 34 syllables. In contrast to analyses for sustaining vowels, however, analytic research using reading texts could not ignore articulatory and phonatory differences between languages.[47] Thus, to examine the utility of AVQIv3 for the Japanese-speaking population, we should define the standardized syllable number for our reading text.

Standardization of included syllable number in the reading text of AVQI

To equalize analytical time length of CS samples with that of SV samples as possible, we summarized the hand-marked syllable numbers for individual vCS samples which made the time lengths most approximated to 3 seconds. The descriptive statistics of mean, median, and the 95% CI range showed the inclusion of 30 syllables to be the most suitable standardized syllable number, which was shorter than that in the Dutch language as we predicted.[54] Fortunately, we could easily segment a border between 30 and 31 syllables because the 30th syllable was at the end of a clause in our text. Furthermore, the values of AVQIv3 using the 30-syllable segmentation almost perfectly coincided with those using the hand-marked individual syllable numbers. These results demonstrated that utilizing 30 syllables as the standardized syllable number could be a highly reliable and facilitated method to estimate both speech types equally.

Criterion-related concurrent validity and responsiveness to change

The results of the first phase enabled us to examine the criterion-related concurrent validity of the AVQIv3-JP (ie, how well AVQIv3-JP estimates the severity of dysphonic voices). The previous AVQI studies on Japanese[52] and other languages showed high validities represented by r_s ranging from 0.78 to 0.911 against hoarseness levels.[43,44,46–51,53–55] With regard to AVQIv3-JP, linear regression and correlation analysis revealed a strong linear relationship between AVQIv3-JP and G_{total} ($r_s = 0.873$ and 0.845 in the total cohort and the cohort without paresis/paralysis, respectively), showing marked concurrent validity. Moreover, a regression analysis between the pre to post differences of AVQIv3-JP and G_{total} showed an improved validity for responsiveness to change ($r_s = 0.878$ and 0.812 in the total cohort and the cohort without paresis/paralysis, respectively), compared with that in the AVOIv2-JP ($r_s = 0.767$).[52] Thus, the higher levels of the concurrent validity and validity related to responsiveness to change demonstrated that AVQIv3-JP has a remarkable ability to estimate the severity level of an individual patient's voice quality as well as the voice outcome after a specific therapy.

Diagnostic accuracy

With respect to the diagnostic accuracy in detecting dysphonic voices, the original AVQIv3 in the Dutch language showed an excellent result of an A_{ROC} of 0.923.[54] In the present study, an excellent discriminative capability between the normophonic and dysphonic voices was confirmed by the ROC analysis with an A_{ROC} of 0.915. Even in the cohort without paresis/paralysis, an A_{ROC} resulted in 0.906, which also showed a substantial diagnostic accuracy. As regards suitable threshold level, the application of the original AVQIv3 threshold of 2.43[54] to the present study produces an excellent specificity of 97.3% with a lower sensitivity of 64.6%. In contrast, the best diagnostic accuracy was achieved using the decreased threshold level of 1.41 with balanced sensitivity and specificity (84.4% and 85.6%, respectively) as well as a relatively lower LR+ of 5.87. We consider that the AVQIv3-JP threshold of 2.06 defined by the normative 95th percentile is the best-balanced threshold represented by a reasonable sensitivity of 72.1% and excellent specificity of 93.8%, with a high LR+ of 11.7. Moreover, in the cohort without paresis/paralysis, the threshold of 2.06 showed a similar sensitivity and specificity to those by the original Dutch threshold for the present total cohort. Therefore, we consider that the AVQIv3 value of 2.06 was the most reasonable threshold to distinguish between the normophonic and dysphonic voices for the Japanese-speaking population.

Auditory-perceptual judgments

In this study, we renewed the rating system of the auditory-perceptual judgments. All of the studies on AVQI including our previous study[43,44,46–55] have applied a single G score for a concatenated acoustic signal of CS and SV samples for an individual patient. However, in our previous study,[52] the inter-rater reliability of single G scores ended up being judged to be fair but slightly lower ($F\kappa = 0.367$). We considered that one of the reasons for decreased reliability was due to the difficulty in estimating

a single score for a combined CS and SV sample. Therefore, we applied the separated judgments of CS and SV samples while listening to a combined voice sample.

Moreover, our previous study employed the anchor voices that contained both speech types recorded from Dutch speakers,[52] which might cause confusion in the judgments of CS samples using the Japanese text readings. Therefore, we tried to make anchor voices recorded by Japanese speakers to raise the reliability. We completed the table of the anchor voices with any combination of roughness and breathiness levels for SV samples according to the definition of roughness and breathiness recorded in an educational sound source *The Sample Tape of Hoarseness* (Interuna Publishers, Inc., Tokyo, Japan) developed by the Japan Society of Logopedics and Phoniatrics. However, we avoided making CS anchor voices due to a lack of consensus in the determination of severity levels. As a result of the renewed judgment system, we achieved a moderate level in the inter-rater reliability as well as moderate to high levels in the intra-rater reliabilities in both of the SV and CS samples.

In addition, due to the different methods of evaluating auditory-perceptual judgments from those in the past studies,[43,44,46–55] we have one limitation when we compare our results with those of the other AVQI studies. Because we defined the overall voice quality as the average of the G_{cs} and G_{sv}, referred to as the G_{total}, the calculation of G_{total} generated the values from 0.0 to 3.0 with 0.1 intervals, whereas our previous study[52] employed the overall voice quality with 0.2 intervals. It could be considered that the finer unit of intervals we applied in this study possibly contributed to the better correlation results compared with our previous study. However, the large improvement in the validity in relation to the responsiveness to change could not be explained by the limitation.

CONCLUSION

Our results confirm that AVQIv3 is also an ecologically valid measure to judge overall voice quality in spite of the interlanguage difference. However, to achieve a balanced-out dominance of both speech types, the suitable syllable number in a reading text should be standardized in every respective language used in the measure. In conclusion, AVQIv3 using 30 Japanese syllables is a reliable measurement tool for estimating the severity of voice quality and voice outcomes after interventions as well as discriminating abnormal and normal voice quality in the Japanese-speaking population.

REFERENCES

1. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). *Eur Arch Otorhinolaryngol*. 2001;258:77–82.
2. Carding PN, Steen IN, Webb A, et al. The reliability and sensitivity to change of acoustic measures of voice quality. *Clin Otolaryngol Allied Sci*. 2004;29:538–544.
3. Hirano M. Psycho-acoustic evaluation of voice. In: Arnold GE, Winckel F, Wyke BD, eds. *Disorders of Human Communication 5. Clinical Examination of Voice*. Vienna, Austria: Springer-Verlag; 1981:81–84.
4. Kempster GB, Gerratt BR, Verdolini AK, et al. Consensus Auditory-Perceptual Evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18:124–132.
5. Isshiki N, Okamura H, Tanabe M, et al. Differential diagnosis of hoarseness. *Folia Phoniatr (Basel)*. 1969;21:9–19.
6. Takahashi H, Koike Y. Some perceptual dimensions and acoustical correlates of pathologic voices. *Acta Otolaryngol Suppl*. 1976;338:1–24.
7. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Lang Hear Res*. 1993;36:21–40.
8. Oates J. Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatr Logop*. 2009;61:49–56.
9. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res*. 1990;33:103–115.
10. Kreiman J, Gerratt BR, Precoda K, et al. Individual differences in voice quality perception. *J Speech Hear Res*. 1992;35:512–520.
11. Fex S. Perceptual evaluation. *J Voice*. 1992;6:155–158.
12. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *J Acoust Soc Am*. 2007;122:2354–2364.
13. Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res*. 2005;48:323–335.
14. Sofranko JL, Prosek RA. The effect of experience on classification of voice quality. *J Voice*. 2012;26:299–303.
15. Lu F-L, Matteson S. Speech tasks and interrater reliability in perceptual voice evaluation. *J Voice*. 2014;28:725–732.
16. De Bodt MS, Wuyts FL, Van de Heyning PH, et al. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997;11:74–80.
17. Yamaguchi H, Shrivastav R, Andrews ML, et al. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. *Folia Phoniatr Logop*. 2003;55:147–157.
18. Dejonckere PH, Obbens C, de Moor GM, et al. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr (Basel)*. 1993;45:76–83.
19. Dejonckere PH, Remacle M, Fresnel-Elbaz E, et al. Reliability and clinical relevance of perceptual evaluation of pathological voices. *Rev Laryngol Otol Rhinol (Bord)*. 1998;119:247–248.
20. Kreiman J, Gerratt BR. Comparing two methods for reducing variability in voice quality measurements. *J Speech Lang Hear Res*. 2011;54:803–812.
21. Gurlekian JA, Torres HM, Vaccari ME. Comparison of two perceptual methods for the evaluation of vowel perturbation produced by jitter. *J Voice*. 2016;30:506, e1-8.
22. Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res*. 2002;45:111–126.
23. Eadie TL, Kapsner-smith M. The effect of listener experience and anchors on judgments of dysphonia. *J Speech Lang Hear Res*. 2011;54:430–448.
24. Ghio A, Dufour S, Wengler A, et al. Perceptual evaluation of dysphonic voices: can a training protocol lead to the development of perceptual categories? *J Voice*. 2015;29:304–311.
25. Lieberman P. Some acoustic measures of fundamental periodicity of normal and pathologic larynges. *J Acoust Soc Am*. 1963;35:344–353.
26. Koike Y. Vowel amplitude modulations in patients with laryngeal diseases. *J Acoust Soc Am*. 1969;45:839–844.
27. Koike Y. Application of some acoustic measures for the evaluation of laryngeal dysfunction. *Studia Phonologica*. 1973;7:17–23.
28. Kitajima K, Gould WJ. Vocal shimmer in sustained phonation of normal and pathologic voice. *Ann Otol Rhinol Laryngol*. 1976;85:377–381.
29. Yumoto E, Gould WJ, Baer T. Harmonics-to-noise ratio as an index of the degree of hoarseness. *J Acoust Soc Am*. 1982;71:1544–1549.
30. Kasuya H, Ogawa S, Mashima K, et al. Normalized noise energy as an acoustic measure to evaluate pathologic voice. *J Acoust Soc Am*. 1986;80:1329–1334.
31. Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. *J Speech Hear Res*. 1994;37:769–778.
32. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *J Speech Hear Res*. 1996;39:311–321.

33. Wolfe VI, Ratusnik DL. Acoustic and perceptual measurements of roughness influencing judgments of pitch. *J Speech Hear Disord*. 1988;53:15–22.
34. Heman-Ackah YD, Michael DD, Goding GS. The relationship between cepstral peak prominence and selected parameters of dysphonia. *J Voice*. 2002;16:20–27.
35. Halberstam B. Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *ORL J Otorhinolaryngol Relat Spec*. 2004;66:70–73.
36. Jannetts S, Lowit A. Cepstral analysis of hypokinetic and ataxic voices: correlations with perceptual and other acoustic measures. *J Voice*. 2014;28:673–680.
37. Zhang Y, Jiang JJ. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *J Voice*. 2008;22:1–9.
38. Yiu EM-L. Limitations of perturbation measures in clinical acoustic voice analysis. *Asia Pac J Speech Lang Hear*. 1999;4:155–166.
39. Hosokawa K, Iwahashi T, Ogawa M, et al. The principles and practice of acoustic analyses. *Larynx Japan*. 2016;2:78–87.
40. Titze IR. Workshop on Acoustic Voice Analysis: Summary Statement. Iowa City, National Center for Voice and Speech; 1995.
41. Titze IR, Horii Y, Scherer RC. Some technical considerations in voice perturbation measurements. *J Speech Hear Res*. 1987;30:252–260.
42. Maryn Y, Roy N, De Bodt M, et al. Acoustic measurement of overall voice quality: a meta-analysis. *J Acoust Soc Am*. 2009;126:2619–2634.
43. Maryn Y, Corthals P, Van Cauwenberge P, et al. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *J Voice*. 2010;24:540–555.
44. Maryn Y, De Bodt M, Roy N. The Acoustic Voice Quality Index: toward improved treatment outcomes assessment in voice disorders. *J Commun Disord*. 2010;43:161–174.
45. Boersma P. Praat, a system for doing phonetics by computer. *Glot Int*. 2001;5:341–345.
46. Maryn Y, Weenink D. Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and Acoustic Voice Quality Index. *J Voice*. 2015;29:35–43.
47. Maryn Y, De Bodt M, Barsties B, et al. The value of the Acoustic Voice Quality Index as a measure of dysphonia severity in subjects speaking different languages. *Eur Arch Otorhinolaryngol*. 2014;271:1609–1619.
48. Barsties B, Maryn Y. The Acoustic Voice Quality Index. Toward expanded measurement of dysphonia severity in German subjects. *HNO*. 2012;60:715–720.
49. Reynolds V, Buckland A, Bailey J, et al. Objective assessment of pediatric voice disorders with the Acoustic Voice Quality Index. *J Voice*. 2012; 26:672.e1–672.e7.
50. Kankare E, Barsties B, Maryn Y, et al. A preliminary study of the Acoustic Voice Quality Index in Finnish speaking population. 11th Pan European Voice Conference.;2015 August 31-September 4; Florence, Italy; 2015.
51. Maryn Y, Kim HT, Kim J. Auditory-perceptual and acoustic methods in measuring dysphonia severity of Korean speech. *J Voice*. 2016;30:587–594.
52. Hosokawa K, Barsties B, Iwahashi T, et al. Validation of the Acoustic Voice Quality Index in the Japanese language. *J Voice*. 2017;31:260.e1–260.e9.
53. Uloza V, Petrauskas T, Padervinskis E, et al. Validation of the Acoustic Voice Quality Index in the Lithuanian language. *J Voice*. 2017;31:257.e1–257.e11.
54. Barsties B, Maryn Y. The improvement of internal consistency of the Acoustic Voice Quality Index. *Am J Otolaryngol*. 2015;36:647–656.
55. Barsties B, Maryn Y. External validation of the Acoustic Voice Quality Index version 03.01 with extended representativity. *Ann Otol Rhinol Laryngol*. 2015;4–6.
56. Kindaichi H, Hirano U. *The Japanese Language*. 2nd ed. Tokyo, Japan: Tuttle Publishing; 2010.
57. Ohata K. Phonological differences between Japanese and English: several potentially problematic areas of pronunciation for Japanese ESL/EFL learners. *Asian EFL J*. 2004;6:Available at: http://asian-efl-journal.com/.
58. Deliyski DD, Shaw HS, Evans MK. Adverse effects of environmental noise on acoustic voice quality measurements. *J Voice*. 2005;19:15–28.
59. Deliyski DD, Shaw HS, Evans MK, et al. Regression tree approach to studying factors influencing acoustic voice analysis. *Folia Phoniatr Logop*. 2006;58:274–288.
60. Maryn Y, Roy N. Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. *J Soc Bras Fonoaudiol*. 2012;24:107–112.
61. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
62. Frey LR, Botan CH, Friedman PG, et al. *Investigating Communication: An Introduction to Research Methods*. Englewood Cliffs, NJ: Prentice-Hall; 1991.
63. Dollaghan CA. *The Handbook for Evidence-Based Practice in Communication Disorders*. Baltimore, MD: Brookes; 2007.