Original paper

The Acoustic Voice Quality Index Version 02.02 in the Finnish-Speaking Population

Elina Kankare[1,2], Ben Barsties v. Latoszek[3], Youri Maryn[3,4,5,6], Marja Asikainen[1], Eija Rorarius[1], Sarkku Vilpas[1], Irma Ilomäki[2], Jaana Tyrmi[2], Leena Rantala[7], Anne-Maria Laukkanen[2]

[1]Department of Phoniatrics, Tampere University Hospital, Tampere, Finland

[2] Speech and Voice Research Laboratory, University of Tampere, Tampere, Finland

[3]Faculty of Medicine and Health Sciences, University of Antwerp, Belgium

[4]Sint-Augustinus Hospital, European Institute for ORL-HNS, Antwerp, Antwerp, Belgium

[5]Faculty of Education, Health & Social Work, University College Ghent, Ghent, Belgium

[6]Department of Speech, Language and Hearing Sciences, University of Ghent, Ghent, Belgium

[7]Logopedics, School of Social Sciences and Humanities, University of Tampere, Tampere, Finland

Elina Kankare, Biokatu 14, Box 2000, FI.33521 Tampere, Finland Tel. +358 3 31166834, eliina.kankare@pshp.fi

The AVQI 02.02 in the Finnish-Speaking Population

Abstract

Background: The Acoustic Voice Quality Index (AVQI) is a multiparametric tool for objectively measuring the general acoustic characteristics of voice. The AVQI uses both sustained vowel and continuous speech in its analysis, and therefore, validation is required for different languages. In the present study, validation was performed in the Finnish-speaking population. Methods: The study included 200 native Finnish-speaking participants of whom 115 were voice patients attending a phoniatric clinic, and the remaining 85 subjects participated in the study as healthy controls. Voice samples were recorded, and the auditory evaluation was performed by five speech therapists. An ordinal four point interval scale was used to evaluate the degree of voice abnormality (Grade, G). Several statistical analyses were performed to test the validity and the diagnostic accuracy of the AVQI in the Finnish-speaking population. Results: The inter-rater reliability of four of the five raters was high enough to allow the use of G_{mean} in the validation. There was a statistically significant correlation between the AVQI scores and the evaluation of overall perceptual voice quality (r=0.74). Conclusions: The results confirmed the good discriminatory power of the AVQI in differentiating between normal and abnormal voice qualities. The AVQI 02.02 threshold value for dysphonia was 2.87 in the Finnish-speaking population.

Keywords: AVQI, acoustic analysis, perceptual evaluation, acoustic clinical tool, dysphonia, voice disorder, Finnish language

Introduction

The minimal standards of voice diagnostics in clinical practice are laryngoscopic examination, perceptual evaluation, aerodynamics, acoustic analyses and the patient's subjective voice ratings (1-3). In the evaluation of voice quality, several subjective judgment scales have been generally adopted, such as the Grade, Roughness, Breathiness, Asthenia, and Strain i.e. the GRBAS scale (4), the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) (5), the Australian Perceptual Voice Profile (6), the Swedish Stockholm Voice Evaluation (7) or Danish Dysphonia Assessment (8,9). All of these judgment scales are subjective and the agreement within and between listeners/raters varies from low to high with respect to judgements of voice quality (10). Acoustic analysis is a non-invasive method which objectively measures voice quality (11-13). Furthermore, a sensitive and time-saving acoustic tool would be beneficial in daily clinical practice. Therefore, Maryn et al. (14) developed an acoustic model called the Acoustic Voice Quality Index (AVQI) to measure the overall voice quality or hoarseness. Hoarseness is a voice symptom; it represents a common abnormality in the sound of the voice but it does not specify whether the abnormality is in breathiness, roughness, strain or asthenia (4). AVQI is a correlate of the perceived degree of hoarseness which allows an objective measurement of the overall voice quality. The index comprises six voice acoustic parameters based on a statistical linear regression analysis. The AVQI equation includes the smoothed cepstral peak prominence (CPPS), harmonics-to-noise ratio (HNR), shimmer local (SL), shimmer local dB (SLdB), general slope of the spectrum (Slope), and tilt of the regression line through the spectrum (Tilt). Over time further refinements of the AVQI have been made, and the AVQI version 02.02 is executed in Praat. (Paul Boersma and David Weenink, University of Amsterdam, The Netherlands) (15). The AVQI has been shown to be a valid and useful acoustic tool. Several investigations have revealed that the AVQI is highly sensitive to the voice changes through voice therapy (13,16). The graphical presentation of results of the AVQI is also easy for patients to understand [see figure 1].

The AVQI uses a concatenation of continuous speech and a sustained vowel [a:] in the analyses, thus quantifying the overall acoustic voice quality into one score for the entire vocal sample (14). These two tasks have been found to be indispensable in voice quality assessment because they reflect vocal function both during rapid glottal and supraglottal changes (speech) and in a more steady state (sustained vowel) (12,15,17).

Since continuous speech is used in the AVQI analyses, the phonetic characteristics of languages may affect the acoustic measures of AVQI. Therefore it is necessary to validate the method and to determine a threshold value between normophonic and dysphonic voices individually for different languages. Studies have shown the AVQI to be cross-linguistically robust in the Indo-European language family: in Germanic languages (Dutch, German, English) (13,18-21), in a Romance language (French) (18) and in a Baltic language (Lithuanian) (22). The AVQI has also been validated in Japanese (16) and for the Altaic language family (Korean) (23).

The present study investigated the validity of the AVQI in the Finnish-speaking population. The Finnish language represents the Finno-Ugric language family (24), which differs from all of the other language families in which the AVQI has been validated. The Finnish language has eight vowels: a, e, i, o, u, y, ä, ö, and seventeen consonants: *p*, *t*, *k*, (*b*), *d*, (*g*), *m*, *n*, *ŋ*, (*f*), *s*, (*š* i.e. *ʃ*), *h*, *l*, *r*, *v*, and *j*. The consonants in brackets are used mainly in some foreign loan words. None of the consonants are produced as aspirated. A tremulant /r/ is produced with the tip of the tongue. The consonants /v/ and /j/ are semivowels in Finnish language. Both vowels and consonants appear in short and long variants. The appearance of vowels and consonants in the Finnish standard language is 47.9 % and 52.1 % (25). In a cross-linguistic comparison, the ratio of vowels to consonants can be rated as moderately low (26), and the mean duration ratio of single over double vowels (V/VV) is 1 : 2.3 in Finnish, whereas for example in English it is smaller (1:1.8) (27). The Finnish language utilizes vowel harmony, and it is an agglutinative language (using many suffixes instead of prepositions), and therefore it is characterized by long word constructions.

The present study aimed to investigate whether AVQI would be suited for use as a clinical tool in a Finnish speaking population. The results from validation of AVQI in different languages serve a larger purpose of development of cross-linguistically robust and globally applicable clinical tools for voice evaluation and therapy.

Methods

Participants

The Ethics Committee of Tampere University Hospital provided an ethical approval for the present study (R15014). The study applied the AVQI to 200 native Finnish participants. Forty-eight of the participants were male and 152 were females. Although there was an uneven gender distribution, the male-female ratio in this study was acceptable because it corresponds to the gender ratio of voice patients (28). The mean age of all participants was 47 years (SD 15.1 years, range 19−84). One hundred and fifteen participants were voice patients attending the Department of Phoniatrics in the Tampere University Hospital and they all had a diagnosed voice disorder (86 females, 29 males, mean age 51, SD 15, range 19−84). The main laryngeal diagnoses of the patients are listed in table I. Eighty-five participants were volunteers with normal voices (66 females, 19 males, mean age 42, SD 14, range 19−67). These participants were volunteers who were enlisted from students, teachers and other staff from the university or from visitors to the Voice Research Laboratory in the university who had no diagnosed voice disorders, although nine of them scored more than 38 points in the Voice Activity and Participation Profile (VAPP) self-evaluation questionnaire which is regarded as the maximum score for healthy voice users (29).

[Please insert the table I. near here]

Voice recordings

The recordings were made in Tampere University Hospital, in the University of Tampere, and in the Tampere University of Technology. Thirty-eight of the recordings were conducted under studio conditions in the University of Tampere, while the other recordings were made in field conditions (i.e. a quiet office or a surgery room). The mean signal-to-noise ratio of the recordings was 39.8 dB with SD of 5.6 dB. Thus, all the recordings were consistent with the recommended norm of SNR, > 30 dB, for acceptable conditions for acoustic recordings and analysis (30).

All voice samples were recorded with an AKG C544L head-mounted condenser microphone and digitised at 44,100 samples per second using the Focusrite iTrack Solo audio interface. The 4 cm mouth-to-microphone distance and the 45° azimuthal angle were controlled by using a ruler from the corner of the participants' mouths. The participants were standing during the recording, and the text for reading was placed on a music stand, comfortably adjusted to the height of each participant. The recordings in the hospital were made by one speech therapist and those in the university were recorded by an experienced technician provided with detailed instructions for the present study. The participants undertook two tasks in the recordings. First they read aloud a short text ("The north wind and the sun", in Finnish "Pohjatuuli ja aurinko"), and secondly, they phonated the sustained vowel [a:] three times for at least 5 seconds per phonation. In both recordings, the participants used a comfortable pitch and loudness, and these samples were saved in a WAV format. In cases where the participants tended to raise the pitch towards a singing-like phonation during the vowel task, they were requested to revert to their habitual speaking pitch by producing short phrases or uttering Uh-Huh with a descending pitch.

Acoustic analyses

All voice samples were analysed using the AVQI-script version 02.02 with Praat (5.3.55) software (15). Three medial seconds of the middle [a:] vowel and the first 23 syllables 'Poh-jan-tuu-li ja au-rin-ko väit-te-li-vät kum-mal-la o-li-si e-nem-män voi-maa' of the text were used in the analyses. The regression formula for the analyses of the AVQI version 02.02 was $9.072 - 0.245 \times \text{CPPS} - 0.161 \times \text{HNR} - 0.470 \times \text{SL} + 6.158 \times \text{SLdB} - 0.071 \times \text{Slope} - 0.170 \times \text{Tilt}$. The analysis automatically merged the voiced segments of the text with the sustained vowel and it was possible to obtain one AVQI-score index value per participant on a scale from 0 to 10 (figure 1).

[Please insert figure 1 near here]

Perceptual evaluation

In the present study, the concatenated samples of continuous speech and sustained vowels **from the original recordings** were used for perceptual evaluation. The six-second long listening samples included 23 syllables of continuous speech from the reading task and immediately thereafter a 3-second long sustained vowel. Five Finnish speech therapists with considerable experience of working with voice patients evaluated the voice quality of the samples using the ordinal four-point equivalent to the interval GRBAS scale (4). In this study, only the G-rating from the GRBAS-scale (representing the overall degree of voice abnormality) (4) was utilized in the perceptual evaluation (31). The complete listening task included 225 voice samples and the listening order of the voice samples was randomised. Twenty-five voice samples (i.e. 12.5% of the total number) were presented twice to check the listeners' intra-rater reliability.

In the listening task, the speech therapists were given a set of anchor voice samples representing different degrees of voice dysphonia (32,33). The principal author listened and pre-evaluated the samples, and selected the specimens to be used as anchor samples. Thereafter the suitability of the anchor samples was evaluated in the phoniatric clinic by three speech therapist and two phoniatricians, who were specialised in voice disorders. In the final anchor sample set, there were eight voice samples, two for each degree of the G on the scale from 0 to 3. The speech therapists received the voice samples and listening instructions on a memory stick, and they listened to the voice samples using on their own computer with around-ear headphones. The listeners were asked to make a combined judgement of both sample parts in each case and they listened to the anchor voice samples prior to the listening task and after every twenty-fifth voice sample. In the subsequent analyses, only the mean ratings of the G-scores were used (G_{mean}).

Statistical analyses

All statistical analyses were conducted using SPSS for Windows version 22.0 (IBM Corp., Armonk, NY, USA), and all the results were considered statistically significant at $p \leq 0.05$, except when otherwise stated. Firstly, to evaluate the rater reliability of the perceptual judgments, the Cohen kappa (Cκ) for intra-rater reliability and the Fleiss kappa coefficient (Fκ) for inter-rater reliability were calculated. Guidelines for the interpretation of the κ statistics were provided by Landis and Koch (34). The analyses were calculated with the software package of r-Studio v. 3.0.1 software package (R

Core Team, Vienna, Austria). Furthermore, significant changes (i.e. considered statistically significant at $p \leq 0.01$) in all kappa values were tested using bootstrapping with 1,000 replications (i.e., this mothod consists in q samples of size n with replacement [35]) based on a script devised by van Belle (36). In order to establish a group of raters with a homogeneous and a high level of reliability, the following criteria were applied: (1) no significant differences should be found in the intra-rater Cκ results between all pairs of raters; (2) each rater should reached an intra-rater reliability of a level of Cκ≥0.41 (34) and (3) all remaining raters with representative and comparably high intra-rater reliability were analysed to find a homogenous rater group with an inter-rater reliability of Fκ≥0.41 (34). If the Fκ result was significantly better by excluding one rater, the rater with the highest significant value was to be excluded for the next round. Thus, in each round we used a backward stepwise method to exclude the rater whose kappa value differed most significantly from the Fκ for all tested raters. This procedure was repeated until a minimum kappa value of ≥0.41 was achieved without any significant improvement in the Fκ by excluding one of the raters. After this procedure four raters from five remained in the study.

Secondly, the concurrent validity of the AVQI in the Finnish language was investigated using the Spearman's rank order correlation coefficient (r_s) and the coefficient of determination (r^2) between the mean ratings of the G-scores (G_{mean}) and the AVQI. Interpretation guidelines for r_s were provided by Frey et al. (37).

Finally, the diagnostic accuracy of the AVQI was evaluated using the receiver operating characteristic (ROC) statistic with several estimates. The diagnostic precision of the AVQI was evaluated by its sensitivity (i.e. correctly identified hoarseness when tested positive on the AVQI) and specificity (i.e. correctly identified hoarseness when testing negative on the AVQI). The absence of hoarseness was defined at G_{mean} (0.0–

0.49). The best threshold level for the AVQI in the Finnish language was determined using the Youden Index, with this based on the results of the ROC statistics. The Youden Index was calculated by the maximum of sensitivity + specificity − 1. Furthermore, the applicability of the best threshold for a clinical decision was assessed by the balance between the “likelihood ratio for a positive result” (LR+) and the “likelihood ratio for a negative result” (LR−), which were 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 $LR+\geq 10$ and $LR-\leq 0.1$ (38). Additionally, the ability of the AVQI to discriminate between normal and hoarse voices was assessed by the “area under ROC-curve” (A_{ROC}). An $A_{ROC} = 1.0$ will be obtained for measures that perfectly distinguishes between normal and hoarse voices. An $A_{ROC} = 0.5$ corresponds to a chance-level diagnostic accuracy (39).

Results

The intra-rater reliability based on 25 voices samples out of 200 samples revealed no significant difference between the perceptual raters in Cκ values (Cκ = 0.63 to 0.87, t = 4.35, p = 0.363). An inter-rater reliability was determined for all five raters. The Fκ results revealed a sufficient kappa value but significant differences were detected between the five raters (Fκ = 0.51, t = 17.8, p = 0.004).The inter-rater reliability when there were only four judges showed significantly better Fκ results, i.e. one rater out of the group of five was excluded (see Methods, Statistic analyses) while still achieving a sufficient level of the kappa value (Fκ = 0.55, t=6.624, p=0.159). Therefore, all analyses were conducted with the perceptual G_{mean} ratings of the four raters. The descriptive statistics for the AVQI score are presented in table II. The frequency distribution of the mean auditory-perceptual overall voice quality ratings is presented in figure 2. A substantial correlation was found between the AVQI scores and the overall rating of perceptual voice quality (Spearman's rho = 0.74, *p* = 0.01) (figure 3) (37). In order to evaluate the AVQI's potential to distinguish Finnish-speaking subjects with normal voice qualities from those with an abnormal voice quality, an ROC curve was constructed (figure 4). The A_{ROC} was 0.862 (i.e. 86.2%) thus confirming the high discriminatory power of the AVQI in differentiating between normophonic and hoarse voices in the Finnish population. The AVQI 02.02 threshold level in the Finnish-speaking population was 2.87, which is related to the highest Youden's Index of 0.606. At this threshold, a sensitivity of 0.796 (79.6%) and a specificity of 0.862 (81%) were achieved. The general guideline of the likelihood ratio statistics was not reached (i.e. LR+=4.08 and LR-=0.25).

[Please insert table II and figures 2, 3, 4 near here]

Discussion

This study tested the validity and diagnostic accuracy of the AVQI for the Finnish language. These results confirmed the good discriminatory power of the AVQI in differentiating between normal and abnormal voice quality (i.e., sensitivity of 79.6% and specificity of 81%). The threshold for dysphonia was determined to be 2.87. Thus, the threshold value for Finnish occupies a low position in comparison with the other languages already tested. The values have been found to range for 2.97 in Lithuanian to 3.66 for Dutch with the AVQI version 02.02 and when perceptual ratings have been based on the G_{mean} from GRBAS. The highest ROC statistics in previous AVQI 02.02 validation studies have been presented in the German language (A_{ROC} 0.958) (18). In addition, investigations in the Dutch (A_{ROC} 0.893), French (A_{ROC} 0.869) (18), Lithuanian (A_{ROC} 0.940) (22) and Japanese (A_{ROC} 0.905) (16) languages have higher A_{ROC} values than those found in the present study (AROC 0.862). In this Finnish validation of the AVQI 02.02, the values for sensitivity and specificity were also among the lowest of those languages tested previously. The result of concurrent validity was in line with the investigation on the Lithuanian language (22), but lower than that obtained with the other tested languages (16,18,23).

The lower diagnostic validity may be attributable to the quality and size of the dysphonic participant groups. A large number (n = 25) of voice patients with a diagnosis of adductor spasmodic dysphonia (ADSD) participated in the present study. ADSD is mainly characterized by intermittent voice breaks, and blockages in phonation resulting due instability in the adductory movement of the vocal folds (40). The AVQI may not always detect the poor voice quality in a patient with ADSD, because his/her voice quality is not necessarily rough or breathy but unstable, and only occasionally strained. Figure 3 shows variations in the AVQI data per mean G score. The variation is most prominent at mean G=3 (i.e., a level of G for which variability between raters is expected to be at its lowest). A further analysis of the data showed that this large variability in the AVQI values among samples that were rated as representing G=3 was mainly due to samples from patients with ADSD. The exclusion of these samples slightly improved the correlation between AVQI and perceptual rating (the Spearmans's rank order correlation coefficient increased from r=0.74 to r=0.75). In general, one cannot expect a perfect correlation between the Grade in GRBAS and AVQI. In fact, Grade has been defined as 'the degree of hoarseness or voice abnormality' (4). There are different kinds of vocal abnormalities (such as trembling or breaking quality, abnormally high or low pitch, register shifts) that may not be reflected in the parameters of AVQI as well as various noise characteristics in the voice. A further source of discrepancy between acoustic and perceptual results may stem from variations of voice quality with time (e.g. 41). Listeners may react more to intermittent signs of hoarseness than is apparent in the acoustic analysis. Additionally, some listeners may pay more attention to voice quality either in the continuous speech part or the vowel part of the sample, while the acoustic analysis examines each sample as a whole. In the present study, a relatively high variation in AVQI was also found for those voice samples which were rated as normal (G=0) (see Figure 3). A particularly striking discrepancy is seen in the voice sample whose AVQI was 3.96. This means that all four raters judged this voice to be normal, whereas AVQI suggested that there was hoarseness (i.e., a value higher than the threshold value 2.87). The reason for this discrepancy may stem from the fact that the sample was spoken relatively softly. Earlier results have shown that acoustic measures of hoarseness such as jitter and shimmer values, SN ratio and cepstral measures, react to voice loudness: better values are obtained when speaking more loudly and *vice versa* (e.g. 42-44). To summarize, the results revealed that there is an overlap in the AVQI scores between different levels of perceived abnormality of voice. This has also been found in previous studies (e.g. 13,16,22).

Differences in the AVQI threshold between languages are most likely attributable to differences in their phonetic structure as well as due to cultural aspects in the perceptual evaluation. A high rate of fricatives and strongly aspirated consonants may increase the threshold value. At present the AVQI has been shown to be adaptable with highly comparable validity across studies using very different languages, regardless of the phonetic content of the continuous speech segments, and the 3-second sustained [a:] segments are a relatively constant factor across the different languages. Furthermore, voiceless fragments, including voiceless fricatives, are automatically removed from the continuous speech recordings before the acoustic analysis and the determination of the AVQI. Therefore, the differences in the diagnostic AVQI thresholds seem to lie in perceptual evaluation, i.e. the listeners' language and culture related tolerance for certain linguistic traits e.g. nonmodal breathiness (as in aspirated fricatives). The perceptual ratings in the present study correlated strongly with the AVQI, as has also been found in earlier studies with other languages. However, the inter-rater reliability was moderate, but higher than in the previous AVQI validation studies in which the Fleiss kappa coefficient has been used (16,22). The perceptual evaluation was made on the basis of both a sustained vowel and a text extract. It is possible that the listeners concentrated their attention on different parts of the material, which would explain the inter-rater differences. Instead the AVQI sums up various acoustic characteristics into single value. Therefore, it has the potential to be a more accurate classifier than a rarely calibrated auditory-perceptual rating. In the script of the AVQI 02.02, the vowel dominates in terms of length over the connected speech. In the future, the more balanced AVQI version 03.01 (21) should also be validated with the Finnish language. Future studies should focus on the performance of the AVQI in different types of dysphonia due to various causes. Furthermore comparative studies will be required, for example, comparing the AVQI results and the patients' subjective evaluation of their voices.

Conclusions

There was a significant correlation between the perceptual rating of the degree of voice abnormality (i.e. G_{mean} of the GRBAS scale) and the AVQI. The results suggest that the AVQI 02.02 is valid and specific in distinguishing between normal and dysphonic voices in Finnish speakers. The AVQI version 02.02 threshold in the Finnish language was 2.87.

Acknowledgements

This study was supported by the Finnish Work Environment Fund (Stipend, number 115498).

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

1.Dejonckere PH, Bradley P, Clemente P, Cornut G, Crevier-Buchman L, Friedrich G, 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.Mehta D HR. Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. Curr. Opin Otolaryngol Head Neck Surg 2008; 16: 211-215.

3.Nelson R, Barkmeier-Kraemer J, Eadie T, Sivasankar MP, Mehta D, Paul D, et al. Evidence-based clinical voice assessment: A systematic review. Am J Speech Lang Pathol 2013; 22: 212-226.

4.Hirano M. Psycho-acoustic evaluation of voice. In: Arnold GE, Winckel F, Wyke BD, ed. Disorders of human communication 5. Clinical examination of voice Vienna, Austria: Springer-Verlag; 1981. p. 81-84.

5.Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. Am J Speech-Lang Pathol 2009; 18: 124-132.

6.Oates J, Russell A. Learning voice analysis using an interactive multi-media package: Development and preliminary evaluation. **J Voice** 1998; 12: 500-512.

7.Hammarberg B. Voice research and clinical needs. Folia Phoniatr Logop 2000; 52: 93-102.

8.Iwarsson J, Bingen-Jakobsen A, Johansen DS, Kølle IE, Pedersen SG, Thorsen SL, Petersen NR. Auditory-perceptual evaluation of dysphonia: A comparison between narrow and broad terminology systems. J Voice 2018; 32: 428-436.

9.Barsties B, De Bodt M. Assessment of voice quality: Current state-of-the-art. Auris Nasus Larynx 2015; 42: 183-188.

10.Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. J Acoust Soc Am 1998; 104: 1598-1608.

11.Awan SN, Roy N. Outcomes measurement in voice disorders: Application of an acoustic index of dysphonia severity. J Speech Lang Hear Res 2009; 52: 482-499.

12.Parsa V, Jamieson DG. Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. J Speech Lang Hear Res 2001; 44: 327-339.

13.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.

14.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 2010; 24: 540-555.

15.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.

16.Hosokawa K, Barsties B, Iwahashi T, Iwahashi M, Kato C, Iwaki S, et al. Validation of the Acoustic Voice Quality Index in the Japanese language. J Voice 2017; 31: 260.e9.

17.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.

18.Maryn Y, De Bodt M, Barsties B, Roy N. 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.

19.Barsties B, Maryn Y. The Acoustic Voice Quality Index. Toward expanded measurement of dysphonia severity in German subjects. HNO 2012; 60: 715-720.

20.Barsties B, Maryn Y. The improvement of internal consistency of the Acoustic Voice Quality Index. Am J Otolaryngol 2015; 36: 647-656.

21.Barsties B, Maryn Y. External validation of the Acoustic Voice Quality Index version 03.01 with extended representativity. Ann Otol Rhinol Laryngol 2016; 125: 571-583.

22.Uloza V, Petrauskas T, Padervinskis E, Ulozaitė N, Barsties B, Maryn Y. Validation of the Acoustic Voice Quality Index in the Lithuanian language. J Voice 2017; 31: 257.e11.

23.Maryn Y, Kim H, Kim J: Auditory-perceptual and acoustic methods in measuring dysphonia severity of Korean speech. J Voice 2016; 30: 587-594.

24.Crystal D. The Cambridge Encyclopedia of Language. Cambridge University Press, Cambridge ed.; 1997.

25. Iso suomen kieliopin verkkoversio. Kotimaisten kielten tutkimuskeskuksen verkkojulkaisuja 5, principal editor Maria Vilkuna. Available from: http://scripta.kotus.fi/visk/ URN:ISBN: 978-952-5446-35-7. [Last accessed 8.8.2018]

26.Lehtonen J. Aspects of quantity in standard Finnish. Jyväskylä, Finland: Studia philologica Jyväskyläensia; 1970.

27.Iivonen A. Intonation in Finnish. In: Intonation Systems. A Survey of Twenty Languages, Hirst D & Di Cristo A, ed. Cambridge: Cambridge University Press; 1998. p. 311-327.

28.De Bodt M, Van den Steen L, Mertens F, Raes J, Van Bel L, Heylen L, et al. Characteristics of a dysphonic population referred for voice assessment and/or voice therapy. Folia phoniat logop 2015; 67: 178-186.

29.Kleemola L, Helminen M, Rorarius E, Isotalo E, Sihvo M. Voice Activity and Participation Profile in assessing the effects of voice disorders on quality of life: Estimation of the validity, reliability and responsiveness of the finnish version. Folia Phoniatr Logop 2011; 63: 113-121.

30.Deliyski DD, Shaw HS, Evans MK. Adverse effects of environmental noise on acoustic voice quality measurements. J Voice 2005; 19: 15-28.

31.Nemr K, Simões-Zenari M, Cordeiro GF, Tsuji D, Ogawa AI, Ubrig MT, Menezes MHM. GRBAS and Cape-V Scales: High reliability and consensus when applied at different times. J Voice 2012; 26: 812.e22.

32.Brinca L, Batista AP, Tavares AI, Pinto PN, Araújo L. The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. J Voice 2015; 29: 776.e14.

33.Dos Santos PCM, Vieira MN, Sansão JPH, Gama ACC. Effect of auditory-perceptual training with natural voice anchors on vocal quality evaluation. J Voice. In Press, corrected proof, Available online 10 January 2018.

34.Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174.

35.Vanbelle, S. and A. Albert (2008). A boostrap method for comparing correlated kappa coefficients. J Statist Comput Simul 78, 1009-1015.

36.Van Belle S. Agreement between raters and groups of raters. Unpublished doctoral dissertation, Univeristy of Liège, Department of Mathematics, Liège, Belgium; 2009.

37.Frey LR, Botan, CH, Friedman PG, Kreps, GL. Investigating communication: An Introduction to Research Methods. Englewood Cliffs, NJ, USA: Prentice Hall; 1991.

38.Dollaghan CA. The handbook for evidence-based practice in communication disorders. MD Bookes, Baltimore, USA; 2007.

39.Portney LG, Watkins MP. Foundations of clinical research, applications to practice. 2.th ed. Prentice Hall Health, Upper Saddle River, NJ, USA; 2000.

40.Chen Z, Li J, Ren Q, Ge P. Acoustic and perceptual analyses of adductor spasmodic dysphonia in mandarin-speaking chinese. J Voice. In Press, corrected proof, Available online 12 February 2018.

41.Gaskill CS, Awan JA, Watts CR, Awan SN. Acoustic and perceptual classification of within-sample normal, intermittently dysphonic, and consistently dysphonic voice types. J Voice 2017; 31: 218-228.

42.Orlikoff RF, Kahane JC. Influence of mean sound pressure level on jitter and shimmer measures. J Voice 1991; 5: 113-119.

43.Brockmann-Bauser M, Bohlender JE, Mehta DD. Acoustic perturbation measures improve with increasing vocal intensity in individuals with and without voice disorders. J Voice 2018; 32: 162-168.

44.Awan SN, Roy N, Jiang JJ. Nonlinear dynamic analysis of disordered voice: The relationship between the correlation dimension (D2) and pre-/post-treatment change in perceived dysphonia severity. J Voice 2010; 24: 285-293.

Figure 1. Example of the graphical output of the Praat script for the AVQI 02.02. The graphic presents the AVQI 02.02 result of subject number 115 who was a 42-year old female with idiopathic vocal cord paresis. The AVQI score is 5.03 where as the G_{mean}, the mean of voice abnormality ratings by five speech therapists, was 1.80. The table on the left side illustrates the outcomes of the six separate acoustic measures in the AVQI model. The severity line from 0 to 10 demonstrates the AVQI values next to the table. The higher the AVQI score, the more abnormal is the voice and vice versa.

Figure 2. Frequency distribution of the mean auditory-perceptual overall voice quality ratings (average of the G-scores of the four judges) for the 200 concatenated voice samples.

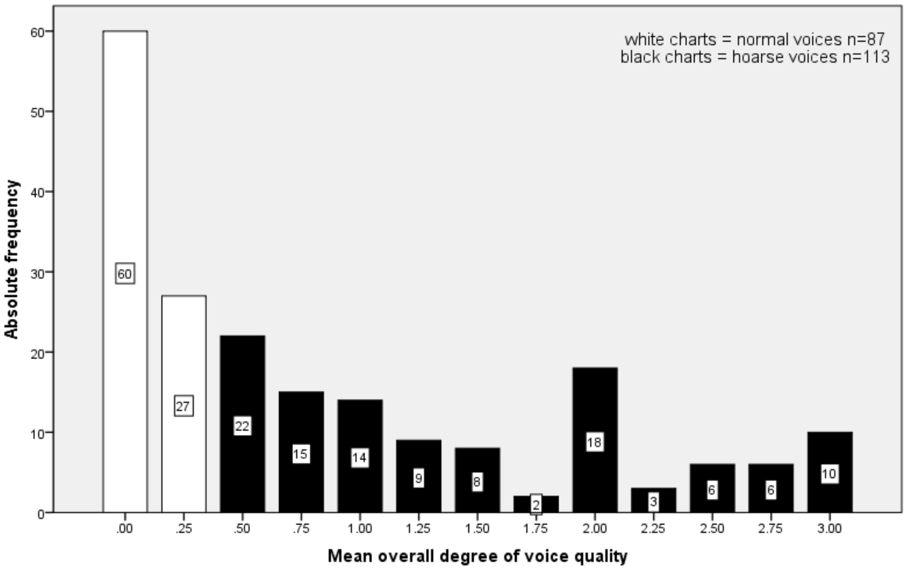

Figure 3. Scatterplot and linear regression line illustrating the proportional relationship between the AVQI version 02.02 and G_{mean} (the two lines above and below the regression fit line delineate the upper and lower boundaries of the 95% prediction interval, respectively).

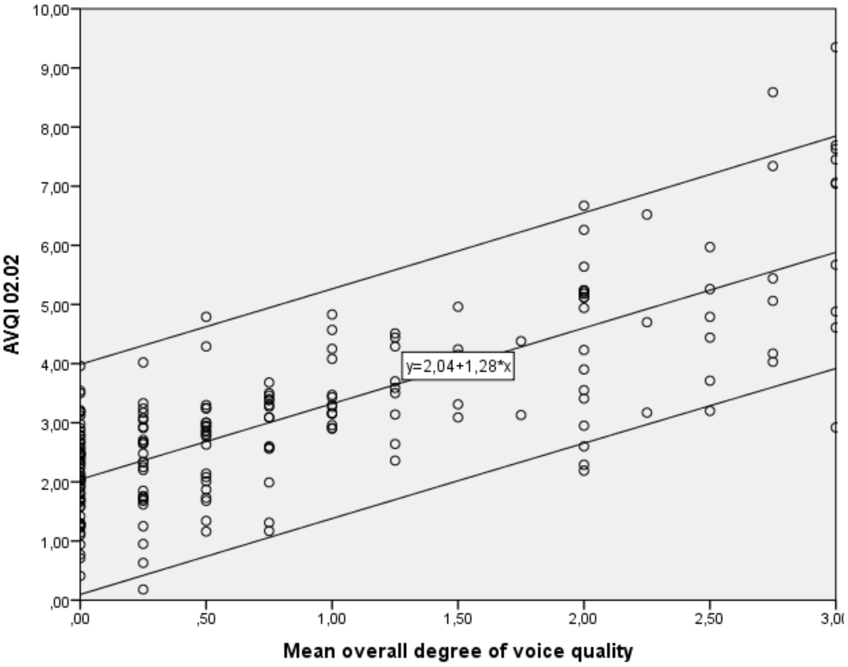

Figure 4. ROC-curve illustrating the diagnostic accuracy of the AVQI.

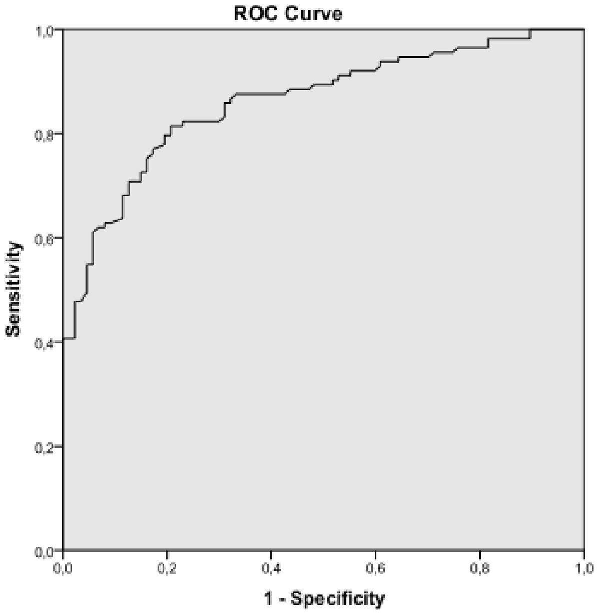

Table I. List of the participants with and without laryngeal diagnoses, ICD-10 codes and the number of participants in each group in the study.

Laryngeal diagnosis	**ICD10 code**	**Detailed definition**	**Number of participants**

Participants with no diagnosed voice disorder			85
Functional dysphonia	R49.01		31
Paralysis/paresis of vocal the cords	J38.0	Different degree of vocal cord paralysis	25
Spasmodic dysphonia	R49.02		25
Chronic laryngitis	J37.0		9
Nodules	J38.2		5
Other diseases of the vocal cords	J38.3	E.g. cyst, bulge in a blood vessel, swelling of the vocal cords	5
Other undefined dysphonia	R49.08	Granuloma, multifactorial voice disorder	4
Larynx irritable	J39.3		2
Transsexualism	F64.0	Male-to-female	2
Polyp of the vocal cord	J38.1		1
Other diseases of the larynx	J38.7	Myoclonus	1
Cough	R05		1
Laryngeal spasm	J38.5		1
Dysphagia	R13	With voice symptoms	1
Ehlers-Danlos syndrome	Q79.6		1
Larynx trauma	Y96.0, W17		1
		Total of voice patients	115
		Total of participants	200

Table II. The mean, standard deviation, and minimum and maximum results of the AVQI 02.02 index values.

	N	AVQI index mean	SD	Min	Max
All participants	200	3.16	1.6	0.18	9.35
Volunteers from the university	85	2.25	0.9	0.18	4.79
Voice patients	115	3.84	1.6	0.94	9.35