Vowel- and Text-based Cepstral Analysis of Chronic Hoarseness

Cornelia Moers[1], Bernd Möbius[2], Frank Rosanowski[3], Elmar Nöth[4], Ulrich Eysholdt[3], Tino Haderlein[3/4]

[1]Department of Speech and Communication, University of Bonn, Poppelsdorfer Allee 47, 53115 Bonn, Germany

[2]Computational Linguistics and Phonetics, Saarland University, Building C7.2, 66123 Saarbrücken, Germany

[3]Department of Phoniatrics and Pediatric Audiology, University Clinics Erlangen, Bohlenplatz 21, 91054 Erlangen, Germany

[4]Pattern Recognition Lab, University of Erlangen-Nuremberg, Martensstraße 3, 91058 Erlangen, Germany

Corresponding author:

Dr.-Ing. Tino Haderlein

Department of Phoniatrics and Pediatric Audiology, University Clinics Erlangen

Bohlenplatz 21

91054 Erlangen

Germany

E-Mail: Tino.Haderlein@informatik.uni-erlangen.de

Phone: +49 9131 852-7872

Fax: +49 9131 303811

Running title: Text-based Cepstral Analysis of Hoarseness

Keywords: hoarseness, automatic voice evaluation, cepstrum, cepstral peak prominence

SUMMARY

Objectives/Hypothesis: Automatic voice evaluation is usually performed on stable sections of sustained vowels which often cannot capture hoarseness properly. The measures Cepstral Peak Prominence (CPP) and Smoothed Cepstral Peak Prominence (CPPS) do not require exact determination of the cycles of fundamental frequency like established perturbation-based measures. They can also be applied to text recordings. In this study, they were compared to perceptual evaluation of voice quality and the German Roughness-Breathiness-Hoarseness (RBH) scheme.

Study Design: Retrospective data analysis.

Methods: 73 hoarse patients (48.3±16.8 years) uttered the vowel /e/ and read the German version of the text "The North Wind and the Sun". The text recordings were evaluated perceptually by 5 speech therapists and physicians according to the RBH scale. The criterion "overall quality" was measured on a 4-point scale and a visual analog scale. For the human-machine correlation, the automatic measures of the Praat program (vowels only) and the "cpps" software were compared to the experts' ratings. The experiments were repeated for speakers with jitter≤5% or shimmer≤5% (n=47).

Results: For the entire group (n=73), the best human-machine results for most of the rating criteria were obtained for text-based CPP and CPPS (up to $|\rho|=0.73$). For the 47 selected speakers, the correlation was remarkably worse for all measures but still best for text-based CPP and CPPS ($|\rho|\leq 0.50$).

Conclusions: Cepstrum analysis should be performed on a text recording. Then it outperforms all perturbation-based measures, and it can be a meaningful objective support for perceptual analysis.

INTRODUCTION

The lifetime prevalence of a voice disorder is almost 50%[1]. When the disorder becomes chronic, it may have severe psycho-social consequences for the affected person and cause enormous costs for modern communication society[2]. A standardized, efficient method for voice assessment is therefore needed. The validated, multi-dimensional voice protocol of the European Laryngological Society (ELS)[3] includes the demand for the application of objective, automatic methods of voice evaluation. However, these methods are still controversially discussed[4]. For instance, there is still no consensus about a standard set of valid scales for measuring voice quality automatically; actually, there is still a discussion on which kind of data should actually be processed. The topic of this article is automatic evaluation of hoarseness from vowel and speech recordings and its comparison to perceptual evaluation by a set of speech experts.

Despite many attempts to automatize voice assessment, perception-based methods are still the basis for the evaluation of voice pathologies by patients and physicians, and they serve as the reference for objective methods. Perception of voice qualities, however, is too inconsistent among single raters to establish a standardized and unified classification[5]. In this way, it cannot be used for clinical and scientific purposes. For this reason, the average opinion of a group of raters is often chosen as a reference for automatic assessment. This is again not suitable for clinical application. Maryn et al. reported in a review numbers of raters between 1 and 22 with 8 being the average[4]. With this background of methodological shortcomings, simple rating criteria for perceptual evaluation have been established. Among them are usually “hoarseness”, “grade” (e.g. ‘G’ from the GRBAS scale[6]) or “overall severity”, and “overall voice quality”.

Hoarseness is a psycho-acoustically defined measure which was originally believed to be distinct of the other two categories roughness (or harshness) and breathiness[7]. Nowadays, hoarseness is often seen as the superclass of these categories[8]. The Roughness-Breathiness-Hoarseness (RBH) evaluation scheme[9] takes this into account. It is an established means for perceptual voice assessment in German-speaking countries and served as the reference for the automatic assessment presented in this study.

Perception experiments are usually applied to spontaneous speech, standard sentences, or standard texts. Automatic analysis relies mostly on sustained vowels. Maryn et al.[4] reported that 18 out of 25 reviewed studies examined sustained vowels exclusively, four only speech, and three both vowels and speech. For the analysis of speech, mostly one sentence of the English "rainbow passage" was used. Speech recordings have the advantage that they contain onsets, variation of F_0 and pauses[10]. The impression of roughness, for instance, is influenced by the vowel onset fragments[11]. In general, hoarseness is more present and perceptible in long vowels, especially in open vowels, vowels in voiced context, vowels after glottal closure or in strained vowels[12]. In automatic evaluation, however, usually only the stable part of an isolated vowel is examined. This is even recommended by some researchers[13]. Following these recommendations means that a substantial portion of patients whose phonation is highly irregular cannot be evaluated at all. Hence, these methods cannot fill the "diagnostic gap".

Most studies on automatic voice evaluation use perturbation parameters, like jitter and shimmer, and measures like the noise-to-harmonicity ratio (NHR)[4]. However, perturbation parameters have a substantial disadvantage. They require exact determination of the cycles of the fundamental frequency F_0. In severe dysphonia it is hardly possible to find an F_0 due to the irregularity of phonation. Carding et al. reported that about 20% of their patients could not be processed by the

software[14]. This drawback can be eliminated by using the Cepstral Peak Prominence (CPP) and the Smoothed Cepstral Peak Prominence (CPPS) which represent spectral noise[15]. They do not require F_0 detection and are therefore applicable also in the case of strongly dysphonic voices[16]. In this way, the diagnostic gap can be reduced.

CPP and CPPS showed high human-machine correlations in previous studies[15,16,17,18]. On both sustained vowels and read-out text, they correlated better with the GRBAS scale than perturbation-based measures[19]. In this study, text-based automatic evaluation of these measures is compared to the German RBH evaluation scheme for the first time. Additionally, the impact of "unreliable" measures on the evaluation result is examined. This concern was neglected in most previous studies and may have lead to erroneous results. Reliability is of high clinical relevance since automatic evaluation measures should be suitable for every patient. The questions addressed in this article are the following:

How does cepstral-based analysis correspond with the perception-based RBH evaluation?

How do the cepstral-based measures perform in comparison to other introduced measures?

Are there significant differences between the results of automatic vowel and text evaluation?

How do the results change when only voices are evaluated which fulfill certain stability criteria for automatic analysis?

MATERIALS AND METHODS

Patient Group

73 German persons with chronic hoarseness (24 men and 49 women) between 19 and 85 years of age participated in this study. The average age was 48.3±16.8 years. The age distribution is shown

in Table 1. Patients suffering from cancer were excluded. The most frequent pathology was functional dysphonia (Table 2). The subjects were examined by an experienced laryngologist and phoniatrician following the standard protocol of the European Laryngological Society.

The vowel and speech recordings were obtained in the Department of Phoniatrics and Pediatric Audiology at the University Clinics in Erlangen. Each person uttered the vowel /e/ and read the text "Der Nordwind und die Sonne" ("The North Wind and the Sun"[20]), a standard text with 108 words (71 distinct) and 172 syllables. It is frequently used in medical speech evaluation in German-speaking countries. For the automatic text evaluations, the first sentence only (approx. 8-12 seconds, 27 words, 44 syllables) was used. The data were recorded with a sampling frequency of 16 kHz and 16 bit amplitude resolution by a microphone AKG C 420 (AKG Acoustics, Vienna, Austria).

The study respected the principles of the World Medical Association (WMA) Declaration of Helsinki on ethical principles for medical research involving human subjects and has been approved by the ethics committee of the University of Erlangen-Nuremberg.

Perceptual Evaluation

Perceptual evaluation on the text recordings following clinical standards was performed by 5 speech therapists and physicians according to the German Roughness-Breathiness-Hoarseness (RBH) scale[9]. Each of the three criteria can be evaluated on a 4-point scale where '0' means "absent" and '3' means "high degree". In order to capture the fact that hoarseness is the superclass, the H rating must have either the same or a higher rating than R or B. RBH represents a short version of the GRBAS scale[6] where the categories "asthenia" and "strain" are omitted.

Additionally, the criterion “overall quality” was measured both on a 4-point scale (1=“very good” to 4=“very bad”) and a 10 cm visual analog scale (VAS, 0.0=“very good” to 10.0=“very bad”). The 4-point scale rating for quality was the first item on the evaluation sheet for each patient; the VAS was the last item.

Vowel Analysis with Praat

For vowel analysis, sections of at least 0.5 seconds duration of stable phonation excluding onset and offset were evaluated. From 17 speakers, a section of 0.7 seconds could be extracted; from 36 speakers, a full second was available in the sustained vowel recording. The automatic analysis was performed using the software Praat 5.1[21]. The target of the analysis was to find measures which can be used independently of the speaker’s gender, just like the fact that the human raters do not need different evaluation methods for men and women. For this reason, the results of Praat and the computation of the human-machine correlation will be presented for the entire speaker group in this article.

The following measures were computed on the vowel recordings:

a) Jitt loc (= Jitter local): relative period-to-period variability in percent
b) Jitt loc ab (= Jitter local absolute): absolute period-to-period variability in µs
c) RAP (= Jitter rap): Relative Average Perturbation Quotient with a smoothing factor of 3, i.e. average jitter among 3 periods in percent
d) PPQ5 (= Jitter ppq5): Pitch Perturbation Quotient with a smoothing factor of 5, i.e. average jitter among 5 periods in percent

e) Shim loc (= Shimmer local): period-to-period variability of the amplitude in percent

f) Shim db (= Shimmer local, dB): amplitude variability in dB

g) APQ11 (= Shimmer apq11): Amplitude Perturbation Quotient with a smoothing factor of 11, i.e. average shimmer among 11 periods in percent

h) NHR (= Mean noise-to-harmonics ratio): ratio of noisy portions between 1500 and 4500 Hz and harmonic portions between 70 and 4500 Hz

i) HNR (= Mean harmonics-to-noise ratio): ratio of harmonic and noisy portions in the spectrum

Praat does not give any error message when extreme or unreasonable values occur. According to Titze, perturbation values above 5% are not reliable[13]. For this reason, the experiments were repeated after removing all speakers with Jitt loc > 5% or Shim loc > 5% from the original test set. The reduced group of 47 speakers comprised 14 men and 33 women between 19 and 79 years of age. Their average age was 45.6±16.7 years.

Cepstral Analysis

For the cepstrum-based parameters, this kind of speaker selection is not necessary. It is assumed that they are able to handle severely distorted voices since they are not based on period detection. Nevertheless, they were also evaluated on both the entire (n=73) and the selected patient set (n=47).

For the computation of the cepstrum, each short section (usually about 10 ms) of the acoustic signal is converted to the spectrum by a Fourier transform. The spectrum shows the intensity of each frequency in the signal. By applying a logarithmic function and another Fourier transform, the spectrum is converted into the cepstrum. The frequency axis of the spectrum is converted into the

quefrency axis which is in a time domain again. The cepstrum reveals the harmonic structure of the spectrum since it describes the intensity of periodic patterns in the amplitude spectrum. The Cepstral Peak Prominence (CPP) is the difference between the cepstral peak and the regression line over the entire cepstrum at this quefrency. A strongly distorted voice has a flat cepstrum and a low CPP due to its unharmonic structure. The correlation with the human evaluation, especially with the RBH scores, is therefore expected to be negative. The Smoothed Cepstral Peak Prominence (CPPS) is the average of cepstra across a certain number of time or quefrency frames.

The computation of CPP and CPPS for the vowels and text recordings was performed by the free software "cpps"[22] which implements the algorithm introduced by Hillenbrand and Houde[15]. The vowel-based results will be denoted by "CPP-v" and "CPPS-v", the results on the first sentence of "The North Wind and the Sun" by "CPP-NW" and "CPPS-NW". For the analysis of these speech data, sections where patients laughed or cleared their throat, were removed from the recording.

Human-Machine Correlation

Statistical analysis was performed using PASW Statistics 18. The inter-rater reliability for the entire rater group was measured using Cronbach's α. In order to examine human-machine correlation, for each rating criterion the respective automatic measure of each recording was compared to both the average value and the median of the five experts' ratings. The correlations between different groups or measures were computed using Spearman's rank-order correlation coefficient ρ.

RESULTS

Perceptual Data

The average values for the perceptual rating criteria are given in Table 3. The respective inter-rater values were α=0.89 for roughness, α=0.91 for breathiness, α=0.93 for hoarseness, α=0.93 for voice quality (4-point scale), and α=0.93 for voice quality (visual analog scale). Correlations between the rating criteria are given in Table 4 and 5. The criteria roughness and breathiness are only moderately correlated with each other. The quality criterion correlates very well with the hoarseness scoring (ρ>0.9), regardless of the evaluation scale. These correlations are as high as those between the two different types of voice quality rating. On the selected patient group with "reliable" perturbation measures (n=47), the correlation between the rating for breathiness and the other criteria dropped while the correlations between roughness and the remaining criteria increased remarkably.

Automatic Acoustic Analysis

The values computed by Praat and cpps are summarized in Table 6 for the whole set of 73 speakers, and in Table 7 for the 47 selected speakers. The F_0 of men and women was significantly different (Mann-Whitney-U test, $p<0.001$). However, the jitter and shimmer values of men and women did not show significant differences. CPPS-v, CPP-NW, and "Jitter loc ab", however, are significantly different. APQ11 could not be computed for one speaker because of too many irregularities in the voice. The distributions of the values among the speakers appeared to be unimodal, bimodal, or asymmetric; the standard deviation of many measures is larger than the mean value. A Kolmogorov-Smirnov test on normal distribution failed for several variables. Only CPP-v, CPPS-v,

CPP-NW, and mean F_0 showed a Gaussian distribution. The average values and standard deviations of all perturbation measures show large differences between the two subject groups (n=73 vs. n=47). For CPP and CPPS, this effect is smaller.

Human-Machine Correlation

The correlations between the perceptual evaluation and the automatic measures are given in Table 8 and 9 for the entire group of patients, and in Table 10 and 11 for the selected 47 speakers.

In the entire group (n=73), for almost all evaluated criteria the best results are obtained for CPP-NW and CPPS-NW. For the roughness (R) evaluation of the 73 speakers, only Jitt loc ab, NHR, and HNR show a slightly better correlation than CPP-NW. In general, the agreement between the raters and the acoustic measures are better when the average of the raters is used instead of the median. There are just a few exceptions where the median performs slightly better (breathiness: Jitt loc, Jitt loc ab, RAP, PPQ5, APQ11; quality on the 4-point scale: Jitt loc, Jitt loc ab, RAP, PPQ5, Shim loc db, NHR, HNR).

After selection of the speakers according to Titze's 5% recommendation (n=47), the human-machine correlation shows remarkably worse results. For several combinations of human ratings and acoustic measures, virtually no correlation could be measured any more. No other measure reaches the results obtained by CPP-NW and CPPS-NW any more, except for NHR on roughness ($|\rho|=0.37$). The differences between average and median evaluation are larger than for the entire patient group: For breathiness, hoarseness, and quality (VAS), there are some better results with the median, but all these values are $|\rho|\leq 0.18$. For quality on the 4-point scale, however, the median rating was superior for all criteria except CPP-v and CPPS-v.

In general, breathiness, hoarseness, and voice quality are better mapped by the measures than roughness. Furthermore, the text-based CPP-NW and CPPS-NW perform better than the vowel-based CPP-v and CPPS-v.

DISCUSSION

In all kinds of dysphonia, irregularity of vibration is accompanied by higher forces at the voice organs. This may cause secondary pathologies, like nodules or edema. For this reason, it is generally not possible to diagnose a specific type of dysphonia just by the perceived degree of hoarseness. Hence, no special selection of anatomic pathology types was made during acquisition of the patient group. Instead, a representative set of different types of hoarseness was evaluated together (Table 2). It comprises more women than men since women suffer from this type of voice pathology more often[23].

The main purpose of this study was to determine the correlation between the standardized German RBH evaluation scheme and cepstral-based measures. The criterion "overall voice quality" was added because it is used in most of the studies in the literature. The VAS for this criterion was added because it allows a more differentiated evaluation than the 4-point scale. Better human-machine correlations have been reported for the VAS[24]. From our results, however, no recommendations for any particular one of them can be drawn. The average and the median value of all raters for one particular speaker were computed for the same reason; both values are used in the literature. The median has the advantage that it stays in the same range as the original data (e.g. for the 4-point quality scale). The average value may provide a more differentiated view of the

ratings since it is not restricted to the original domain. The human-machine results of this study support the application of the average value.

The results on human-machine correlation with cepstral parameters confirmed some findings of other studies. Hillenbrand and Houde[15] found a significant correlation between these parameters and the perceived degree of breathiness for sustained vowels and speech recordings. This was confirmed in our study, but only for speech recordings. Heman-Ackah et al.[19] reported a correlation of the total degree of dysphonia and CPPS of r=-0.80 on stable vowel sections and r=-0.86 on sentence recordings. Their correlation between vowel- and sentence-based CPPS and the breathiness rating was r=-0.70 and r=-0.71, respectively. Our best sentence-based results for voice quality and breathiness were ρ=-0.73 and ρ=-0.64, respectively. The vowel-based measures, however, reached just around ρ=-0.45. Nevertheless, the capability of cepstral-based measures for voice evaluation from running speech was thus confirmed. It was shown that it outperforms both cepstral- and especially perturbation-based vowel analysis. Vowel analysis requires stable phonation, and the lack of stability in phonation may be the most probable reason for pertinent differences across studies. Often a frame of one second of the vowels /a/ (predominantly), /e/, or /i/ is chosen. Other vowel segment durations from 0.1 seconds up to 3 seconds have been reported[4]. For our study, the minimum duration of stable phonation was set to 0.5 seconds, because some patients were not able to phonate longer without too much irregularity. Our subjects uttered /e/, sometimes shifted towards /ɛ/ which is the adjacent phoneme in the German vowel space. Therefore, the results may not be completely comparable to other studies. On the other hand, these variations in duration and vowel quality show that there will always be inconsistencies in the data obtained from a representative group of patients. If their influence deteriorates the evaluation results to such an extent, then the method cannot be used for clinical purposes. This is another important argument against vowel-based perturbation analysis for voice evaluation.

Even larger differences to results in the literature appeared when Titze's 5% guideline for perturbation measures was applied and the experiments were repeated with the smaller patient group (n=47). Hence, most of the human-machine correlation in the larger group (n=73) was based upon extremal values and outliers in the data. It may be that it was those "unreliable" perturbation values that led to the conclusion of earlier studies that certain measures are suitable for automatic voice evaluation. The 5% rule excluded over 40% of all subjects older than 40 years (Table 1). Automatic methods that cannot be applied for such a high percentage of subjects are not suitable for clinical use. CPP and CPPS are actually excluded from Titze's recommendation since they are no perturbation measures. CPP-NW and CPPS-NW also show the same effect on the correlations when the highly irregular voices are excluded, but to a much smaller degree. The high correlation on "reliable" speakers implies that the high correlation on all patients is influenced much less by outliers. Hence, these measures can be used for the entire spectrum of patients. Additionally, cepstral-based text analysis can still moderately indicate the degree of pathology when only a small interval of the range of possible input values is accepted as reliable. Here, all vowel-based approaches failed.

One aspect that contributes to the problems with the unreliable human-machine agreement is the use of correlation coefficients as agreement measures. They are sensitive to the distribution of the values. When a distribution forms two clusters in which the human and machine values are not correlated with each other, the distance between both clusters can enlarge the overall correlation. Outliers and extreme values have the same effect. This has not been considered sufficiently in most previous publications. Nevertheless, correlation coefficients are still the most frequently used statistical measure in automatic voice evaluation[4]. If the use of other measures, like Cohen's κ or one of its extensions, is intended, the different domains of human and machine evaluation have to

be unified. This means, for instance, that continuous intervals of CPP must be mapped to the discrete values {0,1,2,3} of the RBH components. The definition of the interval boundaries for this mapping is an optimization problem. The measure to optimize is the human-machine correlation. Some studies on automatic voice evaluation present very high correlations as a proof of the reliability of their approach. However, this good agreement is valid only for the particular mapping where the "training set" for the interval search was equal to the "test set" for the human-machine agreement. The method can only be regarded as effective if the mapping also shows good results for another test set that was not involved in the interval search. Hence, agreement measures requiring such a mapping comprise one more potential source of error and misinterpretation and should better be avoided, if the amount of available data is too small to form two distinct sets of reasonable size. When correlation coefficients are used instead, the distribution of the input data should be known.

The problem of interval search occurs also when an automatic method is supposed to perform a binary classification in the two classes "normal speech" and "pathologic speech". This was not the goal of this study. Instead, the continuum of degrees of pathology was supposed to map the continuum of human ratings. Our data covered the full continuum of hoarseness (Table 3). One problem with the classification method is that for each measure a threshold value must be defined above or below which a voice should be regarded as being pathological. However, the average values and high standard deviations for the hoarse speakers in Table 6 and 7 indicate that finding such a value may be actually impossible. Furthermore, the results of this kind of analysis strongly depend on the software[25]. For Praat and for the cepstrum analysis tool, not even normative values were provided. Additionally, it would be questionable whether these normative values would hold for vowel and text samples[26] or speech data across languages. When speech recordings are analyzed, languages with a higher percentage of consonants, like e.g. Slavic languages, might produce higher average perturbation results. Hence, methods addressing the two-class problem

should not be used in clinical practice. They are not reliable and do not provide as much information as continuum-based analysis.

Some aspects for enhancing the human-machine correlations have not been tested in this study. Human perception is often non-linear, like e.g. the bark scale for pitch. Physical scales are often linear, like the frequency measured in Hertz. Better human-machine correlations may be found with non-linear mappings between the two modalities. CPP and CPPS have also not been combined with other measures for our data so far. As single measures, they cannot differentiate between different voice qualities[17], which is necessary for the creation of a voice pathology index[27]. On the other hand, this is consistent with the interaction between different dimensions of human perception: the presence of roughness in a voice does not influence the perception of breathiness. However, the perceived degree of roughness is strongly influenced by the presence of breathiness[28]. Additionally, dysphonic voices with lower fundamental frequencies are perceived as more rough than those with higher F_0[27]. Our results, however, confirm the assumption that roughness and breathiness are perceived as separate dimensions and that hoarseness is the superclass of both[8]. Roughness and breathiness correlate with hoarseness with $\rho>0.7$ while they correlate with each other only moderately. It will be one of the most important aspects in future work to teach the automatic analysis to distinguish roughness, breathiness, and hoarseness as good as human listeners can.

CONCLUSION

Cepstral-based analysis corresponds well with the German perception-based RBH evaluation on a representative group of chronically hoarse patients. However, the correlation is only moderate when speakers, for who the perturbation measures are regarded as unreliable, are excluded. The cepstral-based CPP and CPPS still outperform all introduced perturbation measures in this case and show

more stable correlations in different degrees of pathology. This speaks in favor of their application since the exclusion of certain patients is against clinical practicability[14]. Furthermore, it proves that results obtained with several widely used measures should be handled very carefully. The results are best when cepstrum analysis is performed on a text recording. CPPS alone, however, is still not suitable to provide a full hoarseness index. But in combination with other methods, it may be a meaningful objective support and addition to perceptual analysis.

ACKNOWLEDGMENT

This study was partially funded by the German Cancer Aid (Deutsche Krebshilfe) under grant 107873. We would like to thank Dr. Hikmet Toy for acquiring and documenting the speech data.

REFERENCES

[1] Roy N, Stemple J, Merrill RM, Thomas L. Epidemiology of Voice Disorders in the Elderly: Preliminary Findings. Laryngoscope 2007;117:628-633.

[2] Ruben RJ. Redefining the survival of the fittest: communication disorders in the 21st century. Laryngoscope 2000;110:241-245.

[3] Dejonckere PH, Bradley P, Clemente P, Cornut G, Crevier-Buchman L, Friedrich G, Van De Heyning P, Remacle M, Woisard V; Committee on Phoniatrics of the European Laryngological Society (ELS). 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.

[4] Maryn Y, Roy N, De Bodt M, Van Cauwenberge P, Corthals P. Acoustic measurement of overall voice quality: A meta-analysis. J Acoust Soc Am 2009;126:2619-2634.

[5] Kreiman J, Gerratt BR. The perceptual structure of pathologic voice quality. J Acoust Soc Am 1996;100:1787-1795.

[6] Hirano M. Clinical Examination of Voice. New York: Springer; 1981.

[7] Fairbanks G. Voice and articulation drillbook. New York: Harper; 1960. 2nd ed.

[8] Aronson AE, Bless DM. Clinical Voice Disorders. New York: Thieme; 4th ed., 2009.

[9] Nawka T, Anders L-C, Wendler J. Die auditive Beurteilung heiserer Stimmen nach dem RBH-System. Sprache - Stimme - Gehör 1994;18:130-133.

[10] Parsa V, Jamieson DG. Acoustic Discrimination of Pathological Voice: Sustained Vowels Versus Continuous Speech. J Speech Lang Hear Res 2001;44:327-339.

[11] De Krom G. Some Spectral Correlates of Pathological Breathy and Rough Voice Quality for Different Types of Vowel Fragments. J Speech Hear Res 1995;38:794-811.

[12] Laver J. The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press; 1980.

[13] Titze IR. Workshop on Acoustic Voice Analysis: Summary Statement. Denver: National Center for Voice and Speech; 1995.

[14] Carding PN, Steen IN, Webb A, Mackenzie K, Deary IJ, Wilson JA. The reliability and sensitivity to change of acoustic measures of voice quality. Clin Otolaryngol 2004;29:538-544.

[15] Hillenbrand J, Houde RA. Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. J Speech Hear Res 1996;39:311-321.

[16] 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.

[17] 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.

[18] Awan SN, Roy N, Dromey C. Estimating dysphonia severity in continuous speech: Application of a multi-parameter spectral/cepstral model. Clin Linguist Phon 2009;23:825-841.

[19] Heman-Ackah Y, Michael DD, Goding Jr. GS. The Relationship Between Cepstral Peak Prominence and Selected Parameters of Dysphonia. J Voice 2002;16:20-27.

[20] International Phonetic Association. Handbook of the International Phonetic Association. Cambridge: Cambridge University Press, 1999.

[21] Boersma P, Weenik D. Praat: Doing phonetics by Computer, Version 5.1.33. Available at: http://www.fon.hum.uva.nl/praat. Accessed March 24, 2011.

[22] Hillenbrand J. cpps.exe [computer program]. Available at: http://homepages.wmich.edu/~hillenbr. Accessed March 24, 2011.

[23] Coyle SM, Weinrich BD, Stemple JC. Shifts in relative prevalence of laryngeal pathology in a treatment-seeking population. J Voice 2001;15:424-440.

[24] Yu P, Revis J, Wuyts FL, Zanaret M, Giovanni A. Correlation of Instrumental Voice Evaluation with Perceptual Voice Analysis Using a Modified Visual Analog Scale. Folia Phoniatr Logop 2002;54:271-281.

[25] Maryn Y, Corthals P, De Bodt M, Van Cauwenberge P, Deliyski D. Perturbation Measures of Voice: A Comparative Study between Multi-Dimensional Voice Program and Praat. Folia Phoniatr Logop 2009;61:217-226.

[26] Zhang Y, Jiang JJ. Acoustic Analyses of Sustained and Running Voices From Patients With Laryngeal Pathologies. J Voice 2008;22:1-9.

[27] Wolfe V, Martin DP. Acoustic Correlates of Dysphonia: Type and Severity. J Commun Disord 1997;30:403-416.

[28] Kreiman J, Gerratt BR, Berke GS. 1994 The multidimensional nature of pathologic vocal quality. J Acoust Soc Am 1994;96:1291-1302.

TABLES

TABLE 1. *Distribution of speaker's age in the speaker groups (all 73 speakers and selected 47 speakers with Jitt loc ≤ 5% and Shim loc ≤ 5%)*

age	≤20	21-30	31-40	41-50	51-60	61-70	>70
no. of speakers (n=73)	4	9	13	15	10	15	7
no. of speakers (n=47)	4	6	11	8	7	7	4

TABLE 2. *Diagnoses within the speaker groups (all 73 speakers and the selected 47 speakers with Jitt loc ≤ 5% and Shim loc ≤ 5%)*

diagnosis	no. of speakers (n=73)	no. of speakers (n=47)
functional dysphonia	45	29
organic dysphonia	9	4
organic dysphonia + paresis	1	0
spasmodic dysphonia	1	1
laryngitis	2	2
laryngitis + functional dysphonia	1	1
laryngitis + organic dysphonia	1	1
vocal fold polyp	6	5
paresis	4	1
paresis + Reinke's edema	1	1
Reinke's edema	2	2

TABLE 3. *Perceptual evaluation results for the 73 and the 47 (in parentheses) patients*

	possible range	average	standard dev.	min	max	range
R	0-3	1.56 (1.33)	0.83 (0.77)	0.00 (0.00)	3.00 (3.00)	3.00 (3.00)
B	0-3	1.19 (0.89)	0.81 (0.56)	0.00 (0.00)	3.00 (2.20)	3.00 (2.20)
H	0-3	1.84 (1.51)	0.84 (0.74)	0.00 (0.00)	3.00 (3.00)	3.00 (3.00)
quality (4-point)	1-4	2.54 (2.20)	0.87 (0.71)	1.00 (1.00)	4.00 (4.00)	3.00 (3.00)
quality (VAS)	0.0-10.0	4.74 (3.77)	2.51 (2.07)	0.32 (0.32)	9.50 (8.86)	9.18 (8.54)

TABLE 4. *Correlation ρ between the perceptual ratings (upper triangle: average, lower triangle: median of 5 raters) for all 73 speakers.*

	R	B	H	quality (4-point)	quality (VAS)
R		0.46**	0.79**	0.74**	0.69**
B	0.31*		0.78**	0.77**	0.81**
H	0.73**	0.64**		0.95**	0.92**
quality (4-point)	0.69**	0.62**	0.88**		0.95**
quality (VAS)	0.66**	0.68**	0.89**	0.89**	

* = correlation is significant on the 0.05 level, ** = correlation is significant on the 0.01 level.

TABLE 5. *Correlation ρ between the perceptual ratings (upper triangle: average, lower triangle: median of 5 raters) for the 47 selected speakers.*

	R	B	H	quality (4-point)	quality (VAS)
R		0.51**	0.95**	0.87**	0.84**
B	0.36*		0.69**	0.65**	0.69**
H	0.86**	0.48**		0.93**	0.90**
quality (4-point)	0.85**	0.43**	0.84**		0.92**
quality (VAS)	0.81**	0.51**	0.83**	0.84**	

* = correlation is significant on the 0.05 level, ** = correlation is significant on the 0.01 level.

TABLE 6. *Automatic measures obtained by the Praat software on vowels and the cepstrum analysis software cpps on vowel and text for all speakers (n=73; for APQ11: n=72)*

	average	standard dev.	min	max	range
CPP-v	17.15	4.36	8.78	25.30	16.52
CPPS-v	6.09	2.24	0.91	11.08	10.17
CPP-NW	12.11	1.55	9.05	16.27	7.22
CPPS-NW	4.12	0.95	1.89	6.33	4.44
Jitt loc	1.05	1.08	0.12	5.64	5.52
Jitt loc ab	85.75	140.51	4.81	922.79	917.98
RAP	0.64	0.83	0.05	5.20	5.15
PPQ5	0.75	1.15	0.07	7.94	7.87
Shim loc	6.07	4.75	1.37	20.83	19.47
Shim loc dB	0.54	0.43	0.12	1.83	1.7
APQ11	4.72	4.37	1.11	31.09	29.98
NHR	0.14	0.19	<0.01	0.83	0.82
HNR	12.79	5.39	1.22	26.09	24.88

TABLE 7. *Automatic measures obtained by the Praat software on vowels and the cepstrum analysis software cpps on vowel and text for the selected speakers (n=47)*

	average	standard dev.	min	max	range
CPP-v	18.87	3.42	11.35	25.30	13.95
CPPS-v	6.86	1.78	3.61	11.08	7.47
CPP-NW	12.85	1.21	10.29	16.27	5.98
CPPS-NW	4.60	0.65	3.22	6.33	3.11
Jitt loc	0.58	0.33	0.12	1.48	1.36
Jitt loc ab	39.14	35.14	4.81	166.69	161.87
RAP	0.32	0.19	0.05	0.84	0.79
PPQ5	0.35	0.21	0.07	0.92	0.85
Shim loc	3.15	0.93	1.37	4.67	3.3
Shim loc dB	0.28	0.08	0.12	0.44	0.32
APQ11	2.62	1.02	1.11	5.08	3.98
NHR	0.05	0.03	<0.01	0.15	0.15
HNR	15.72	3.37	9.95	26.09	16.14

TABLE 8. *Spearman's rank-order correlation ρ between perceptual and automatic evaluation for the entire speaker group (n=73); the perceptual result was the mean value of all raters.*

	mean R	mean B	mean H	mean quality (4-point)	mean quality (VAS)
CPP-v	-0.24*	-0.54**	-0.52**	-0.54**	-0.50**
CPPS-v	-0.18	-0.46**	-0.46**	-0.53**	-0.44**
CPP-NW	-0.46**	-0.66**	-0.68**	-0.67**	-0.67**
CPPS-NW	-0.52**	-0.64**	-0.73**	-0.73**	-0.72**
Jitt loc	0.42**	0.58**	0.60**	0.59**	0.56**
Jitt loc ab	0.49**	0.53**	0.60**	0.57**	0.53**
RAP	0.36**	0.54**	0.55**	0.52**	0.48**
PPQ5	0.41**	0.55**	0.58**	0.56**	0.52**
Shim loc	0.39**	0.48**	0.56**	0.56**	0.55**
Shim loc db	0.40**	0.49**	0.57**	0.57**	0.56**
APQ11	0.36**	0.46**	0.53**	0.52**	0.54**
NHR	0.50**	0.52**	0.63**	0.58**	0.54**
HNR	-0.47**	-0.51**	-0.61**	-0.56**	-0.51**

**: correlation is significant on the 0.01 level; *: correlation is significant on the 0.05 level

TABLE 9. *Spearman's rank-order correlation ρ between perceptual and automatic evaluation for the entire speaker group (n=73); the perceptual result was the median of all raters.*

	median R	median B	median H	median quality (4-point)	median quality (VAS)
CPP-v	-0.20*	-0.52**	-0.47**	-0.51**	-0.48**
CPPS-v	-0.16	-0.43**	-0.41**	-0.48**	-0.42**
CPP-NW	-0.39**	-0.65**	-0.62**	-0.65**	-0.66**
CPPS-NW	-0.45**	-0.62**	-0.67**	-0.70**	-0.71**
Jitt loc	0.38**	0.60**	0.53**	0.61**	0.53**
Jitt loc ab	0.43**	0.56**	0.54**	0.61**	0.51**
RAP	0.32**	0.56**	0.47**	0.54**	0.45**
PPQ5	0.36**	0.56**	0.50**	0.58**	0.48**
Shim loc	0.33**	0.47**	0.52**	0.56**	0.52**
Shim loc db	0.35**	0.48**	0.53**	0.57**	0.53**
APQ11	0.30**	0.47**	0.49**	0.50**	0.52**
NHR	0.44**	0.48**	0.58**	0.63**	0.52**
HNR	-0.43**	-0.46**	-0.57**	-0.61**	-0.50**

**: correlation is significant on the 0.01 level; *: correlation is significant on the 0.05 level

TABLE 10. *Spearman's rank-order correlation ρ between perceptual and automatic evaluation for the selected speaker group (n=47); the perceptual result was the mean value of all raters.*

	mean R	mean B	mean H	mean quality (4-point)	mean quality (VAS)
CPP-v	-0.01	-0.18	-0.09	-0.12	-0.08
CPPS-v	0.19	-0.04	0.05	-0.07	0.03
CPP-NW	-0.37**	-0.48**	-0.41**	-0.40**	-0.42**
CPPS-NW	-0.44**	-0.48**	-0.48**	-0.49**	-0.48**
Jitt loc	0.23	0.32*	0.24	0.20	0.20
Jitt loc ab	0.32*	0.32*	0.31*	0.24	0.23
RAP	0.17	0.29*	0.20	0.14	0.13
PPQ5	0.24	0.29*	0.24	0.19	0.18
Shim loc	0.23	0.16	0.18	0.18	0.23
Shim loc db	0.24	0.18	0.20	0.21	0.26*
APQ11	0.21	0.11	0.13	0.10	0.17
NHR	0.37**	0.21	0.27*	0.17	0.14
HNR	-0.36**	-0.19	-0.25*	-0.16	-0.09

**: correlation is significant on the 0.01 level; *: correlation is significant on the 0.05 level

TABLE 11. *Spearman's rank-order correlation ρ between perceptual and automatic evaluation for the selected speaker group (n=47); the perceptual result was the median of all raters.*

	median R	median B	median H	median quality (4-point)	median quality (VAS)
CPP-v	-0.01	-0.24	-0.02	-0.09	-0.04
CPPS-v	0.17	-0.04	0.09	-0.05	0.08
CPP-NW	-0.29*	-0.48**	-0.30*	-0.45**	-0.41**
CPPS-NW	-0.37**	-0.43**	-0.38**	-0.50**	-0.48**
Jitt loc	0.22	0.37**	0.12	0.25*	0.17
Jitt loc ab	0.28*	0.37**	0.18	0.34*	0.22
RAP	0.16	0.34*	0.08	0.18	0.10
PPQ5	0.22	0.34*	0.11	0.23	0.14
Shim loc	0.16	0.19	0.09	0.24	0.23
Shim loc db	0.19	0.20	0.11	0.26*	0.25*
APQ11	0.12	0.18	0.03	0.17	0.15
NHR	0.35**	0.16	0.20	0.33*	0.15
HNR	-0.36**	-0.11	-0.20	-0.32*	-0.11

**: correlation is significant on the 0.01 level; *: correlation is significant on the 0.05 level