

The Effect of Visible Speech in the Perceptual Rating of Pathological Voices

Jan W. M. A. F. Martens, MSc; Huib Versnel, PhD; Philippe H. Dejonckere, MD, PhD

Objective: To test a simple method for improving consistency among raters for the perceptual evaluation of pathological voice quality by providing visible speech (spectrogram) as additional information because, to date, the interrater variability still limits the widespread clinical use of the best available rating system.

Design: Experimental comparison between 2 different ways (with and without the addition of visible speech) of perceptual rating by trained professionals of recorded pathological voices. Furthermore, the correlation between acoustical (jitter, shimmer, and noise-harmonic ratio) and perceptual parameters was investigated in both rating conditions.

Subjects: Six experts evaluated 70 recorded pathological voices using the GIBAS (grade, instability, roughness, breathiness, asthenicity, and strain) scale in 2 separate sessions: first, conventionally, without visible speech as additional information, and several months later, with visible speech as additional information.

Main Outcome Measures: The κ interrater agreement and the correlation coefficient between GIBAS scores and acoustic measures.

Results: We found a significant effect of visible speech on the agreement between the raters. The interrater agreement according to κ statistics was significantly stronger with the addition of visible speech than without for rating grade, roughness, and breathiness. The correlation between acoustical and perceptual parameters showed no significant effect of visible speech.

Conclusions: The addition of visible speech to the perceptual evaluation of pathological voices is an interesting clinical asset to enhance its reliability. The addition of visible speech to the clinical setting is feasible, since affordable computer programs are currently available that can provide the spectrogram in quasi-real time while conversing with the patient. The acoustical analysis might be applied in addition to perceptual rating in a multi-dimensional approach to assess voice quality.

Arch Otolaryngol Head Neck Surg. 2007;133:178-185

Author Affiliations: The Institute of Phoniatics, Department of Otorhinolaryngology, University Medical Centre Utrecht, Utrecht, the Netherlands.

DIAGNOSTIC ASSESSMENT OF voice disorders requires accurate and reliable voice quality measurements. In this scope, perceptual evaluation of the voice by experts such as laryngologists or speech therapists is a primary tool. The GRBAS scale introduced by Hirano¹ has become a commonly used scale for rating severity of deviance, where G stands for grade (overall impression), R for roughness, B for breathiness, A for asthenicity, and S for strain. The parameter I for instability was added by Dejonckere et al² because the quality of a voice can fluctuate over time, thus forming the acronym GIBAS. The GRBAS scale was found to be a reliable perceptual scale^{2,3}; however, it has consider-

able disadvantages. Raters have to be experienced, and judgments of different raters (even experienced ones) might differ considerably.⁴⁻⁶ Furthermore, training is important to reach a satisfactory interrater agreement for grade, roughness, and breathiness,² and agreement among raters is less for pathological voices than for normal voices.^{7,8}

Acoustical analysis of pathological voice quality has several advantages as being quantitative and noninvasive and cost and time efficient. As a disadvantage, most acoustical analyses rely on quasiperiodic waveforms and thus cannot be used on noisy and irregular voices. Historically, a direct relation between perceptual and acoustical entities has been sought. Well-known examples are the close relations be-

tween loudness and amplitude⁹ and between pitch and frequency.¹⁰ No such relation has yet been found for voice qualities that are relevant for pathological conditions. This is best illustrated by inconsistencies between various reports on the correlations between perceptual and acoustical voice parameters.^{2,11,12} For instance, Dejonckere et al² reported a strong correlation between breathiness and shimmer, whereas Morsomme et al¹¹ reported a weak correlation between those parameters. Because of the lack of a one-to-one relationship between acoustical and perceptual voice parameters, perceptual assessment cannot be replaced by acoustical analysis.

In 1946, Koenig et al¹³ described an electronic sound spectrograph, which they had developed at Bell Telephone Laboratories. Among other results, the spectrogram enabled the visualization of speech, thus sometimes referred to as “visible speech,” and it was widely used in voice and speech research.¹⁴ It was also clinically applied to evaluate voice.^{15,16} Yanagihara,¹⁵ for instance, described the relation between hoarseness and noise seen in the spectrogram. The present study examined whether the addition of visible speech would enhance the interrater agreement of perceptual ratings of pathological voices. Since spectrograms reveal acoustical properties, which are related to parameters such as jitter and noise-harmonic ratio, it is conceivable that visible speech increases the correlations between acoustical and perceptual parameters. Therefore, the effect of visible speech on these correlations was also examined in this study.

METHODS

PATHOLOGICAL VOICES

Seventy pathological voices of various causes were digitally recorded. They were collected from the archives of the voice clinic in a university hospital. The recorded voice tracks consisted of a prolonged /a/ with a duration of several seconds and a spoken sentence in Dutch. The sample frequency was 44.1 kHz.

VISIBLE SPEECH

The visible speech consisted of 2 spectrograms (0-4000 Hz) of the sustained /a/, as shown in **Figure 1**. One spectrogram was produced with a fine-frequency resolution (bandwidth, 59 Hz) showing harmonics, and the other was produced with a fine-time resolution (bandwidth, 300 Hz) showing glottal pulses.

To some extent, the GIRBAS parameters can be deduced from the visible speech. The worse the grade of a voice, the more distortion is seen in the spectrogram. Instability might be observed in variations in the spectrogram, although instability is usually perceived in running speech and not in sustained vowels. Roughness is reflected by irregularities in the glottal pulses. These events are more clearly seen in the spectrogram with a fine-time resolution. A secondary acoustic feature of roughness is the presence of subharmonics. Breathiness relates to the presence of noise, which can appear in various spectral regions, depending on the extent of breathiness (in higher frequencies and between or instead of harmonics in upper and lower formants). This can be perceived in the spectrogram with fine-frequency resolution. Furthermore, a voice with asthenicity (and a weak glottal closure) is expected to show less high harmonics, whereas a voice with strain (and a strong glottal closure) should demonstrate more high harmonics.

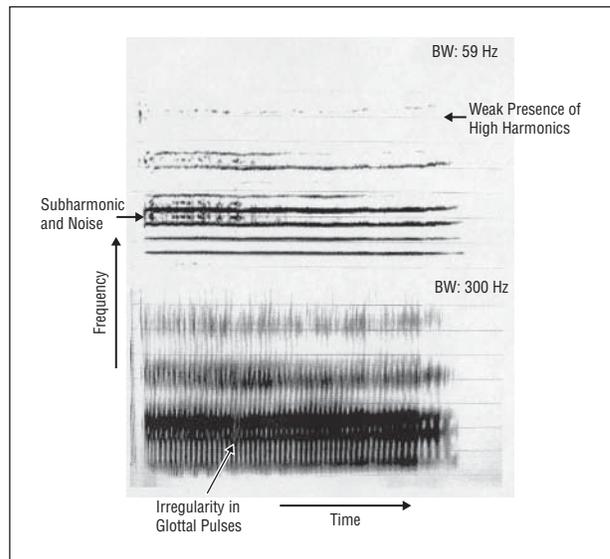


Figure 1. An example of spectrograms (“visible speech”) shown to the experts. The top shows a spectrogram with fine-frequency resolution (filter bandwidth [BW], 59 Hz). The bottom graph shows the spectrogram with a fine-time resolution (filter BW, 300 Hz). The time duration is 1 second, and the frequency range is 4000 Hz. The horizontal lines indicate 500-Hz segments. Arrows point to particular features that are important for the raters’ judgment.

The spectrograms shown to the expert were created with the Digital Sona-graph model 7800 (Kay Elemetrics Corp, Lincoln Park, NJ), and afterwards, were digitally scanned.

PERCEPTUAL EVALUATION

Six raters independently evaluated the voice samples (prolonged /a/ and sentence, all on a compact disc) in 2 sessions, with an interval of 4 to 10 months between sessions. The raters, 4 laryngologists and 2 speech therapists working in different voice clinics, had ample experience in the diagnosis and treatment of voice disorders as well as in GIRBAS parameters (9-20 years of experience). None of them had ever examined or treated any of the patients whose voices they had to evaluate for this study. The experts rated the GIRBAS parameters on a 0 to 3 discrete scale, where 0 indicates normality; 1, a slight deviance; 2, a moderate deviance; and 3, a severe deviance from normal.¹ The ratings were based on the total impression of both sustained /a/ and sentence. The sentence was included because GIRBAS ratings are commonly made on the basis of running speech.¹⁷ A brief explanation on GIRBAS parameters and the rating system was provided because some raters were accustomed to using visual analog scales.

During the second evaluation session, the accessory visible speech of the sustained /a/ was presented to the experts simultaneously with the acoustic presentation of the voice samples (prolonged /a/ and sentence). Their judgment was made on the basis of both the auditory perception of the voice and the visual inspection of the spectrogram. The sequence of voice samples during the second evaluation randomly differed from the sequence during the first evaluation. The experts received a written explanation about the spectrograms, including information on the spectral and temporal resolution and range. There was no instruction on the interpretation of the spectrograms with respect to perceptual parameters.

ACOUSTIC EVALUATION

A variety of acoustic parameters such as jitter, shimmer, and noise-harmonic ratio was calculated using the multidimen-

Table. The Influence of Visible Speech on Mean GIRBAS Scale Ratings for One Specific Pathological Voice*

Parameter	Without Visible Speech	With Visible Speech
G	2.17 (0.41)	3.00 (0.00)
I	0.33 (0.52)	0.67 (0.82)
R	2.00 (0.63)	0.50 (0.84)
B	1.83 (0.75)	2.50 (0.55)
A	0.17 (0.41)	1.33 (0.82)
S	0.67 (0.82)	0.33 (0.52)

Abbreviation: GIRBAS, grade, instability roughness, breathiness, asthenicity, and strain.

*Data are given as mean (SD) across raters.

sional voice program (MDVP) (Kay Elemetrics Corp). A complete list of the acoustic parameters analyzed is available from the authors.

A relatively stable portion of a signal is recommended for acoustic evaluation,¹⁸ so only a relatively stationary part of the prolonged /a/ was used and not the spoken sentence. To obtain a relatively stable portion, the first and last 250 milliseconds (ms) of the signal, which include the onset and offset, were removed.

AGREEMENT

Agreement between perceptual evaluations of 2 experts can be estimated using the parameter κ introduced by Cohen.¹⁹ The Cohen κ statistic corrects for agreement by chance. If raters perfectly agree, $\kappa = 1$; if they totally disagree, $\kappa = -1$; and if their ratings behave independently, $\kappa = 0$. To assess the agreement among the 6 raters, we computed κ according to Fleiss,²⁰ who extended the Cohen κ for more than 2 raters. One should note that κ determines the exact agreement between experts (or between sessions of 1 expert).

To determine whether the agreements found in the conventional and visible speech conditions significantly differ, the 2 κ values were statistically tested using the following equation:

$$(1) \quad z = \frac{k_1 - k_2}{\sqrt{\sigma_{k_1}^2 - \sigma_{k_2}^2}}$$

where k_1 and k_2 denote the κ values and σ_{k_1} and σ_{k_2} denote the corresponding standard deviations. The P value was calculated on the assumption of z being a standardized normal distribution.

ACOUSTICAL VS PERCEPTUAL PARAMETERS

Because the perceptual GIRBAS parameters are ordinal, the correlation between the acoustic and perceptual evaluations was calculated using the Spearman rank correlation coefficient. To determine whether correlation coefficients significantly differ, we applied equation 1 on Fisher-transformed values of ρ , which is expressed as follows:

$$(2) \quad z = \frac{Z_{\rho_1} - Z_{\rho_2}}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

where Z is the Fisher-transformed correlation coefficient and N is the number of evaluated points (in this case $N_1 = N_2 = 70$). The difference z is approximately normally distributed. The corresponding P value is calculated accordingly.

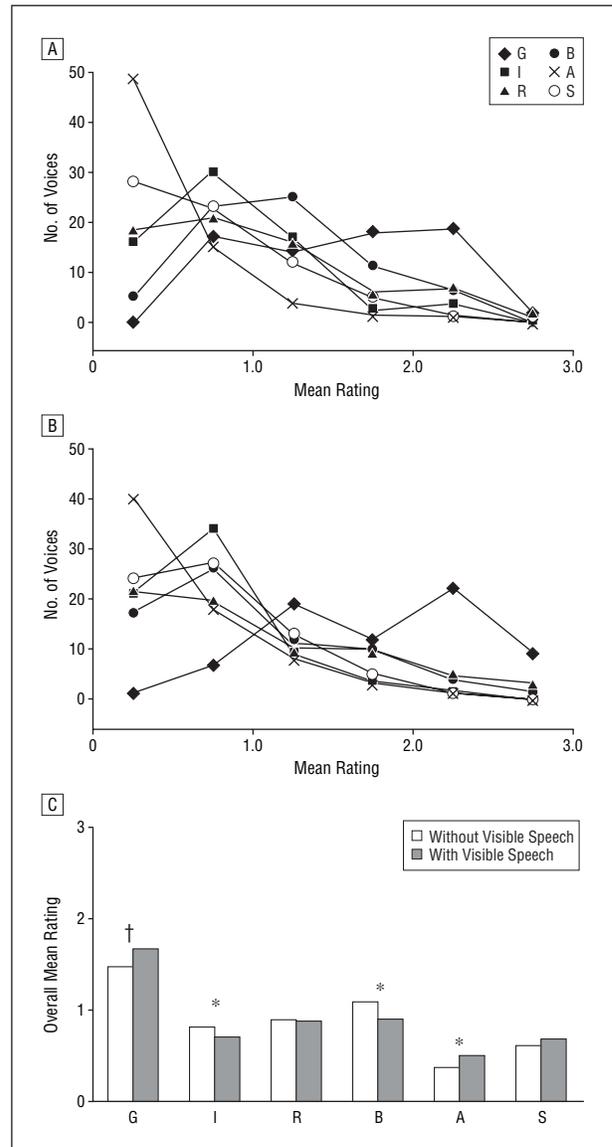


Figure 2. Distributions of GIRBAS (grade, instability, roughness, breathiness, asthenicity, and strain) scale ratings without (A) and with (B) visible speech averaged across 6 raters for 70 voices (the bin width of the distribution is 0.5) and GIRBAS ratings averaged across raters and voices (C). Significant differences between mean ratings with and without visible speech are indicated with an asterisk or dagger (paired t test, * $P < .05$, † $P < .001$).

RESULTS

INTERRATER AGREEMENT

The **Table** gives the ratings, averaged across the 6 raters, without and with the addition of visible speech, for one example of a pathological voice. The ratings with visible speech differed from the ratings without visible speech, with a notable change of roughness from 2.0 to 0.5. **Figure 2** shows the distributions (A and B) and their means (C) of the average GIRBAS ratings of the 70 voices. The severity of voice disorders, as expressed by the mean grade rated conventionally (Figure 2A), is evenly distributed from 0.5 to 2.5. The distributions obtained with visible speech (Figure 2B)

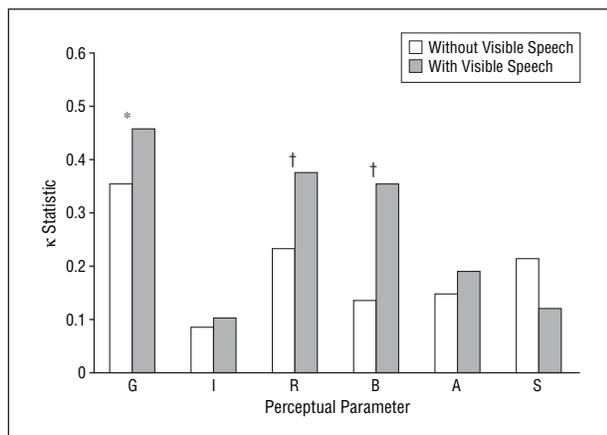


Figure 3. Statistics for 6 raters for GIBBAS scale (grade, instability, roughness, breathiness, asthenicity, and strain) parameters, without and with visible speech. Significant differences between κ with and without visible speech are indicated with an asterisk or dagger (* $P < .05$, † $P < .001$).

differed from the ones obtained conventionally (Figure 2A). In particular, the grade distribution shifted upward, and the breathiness distribution shifted downward. This is also expressed by shifts of the means (Figure 2C), which were statistically significant (paired t test, $P < .01$).

The ratings of the 70 voices were used to calculate κ for 6 raters. The κ values for each perceptual parameter are shown in **Figure 3**. The agreement between ratings was significantly higher with than without visible speech for the perceptual parameters of grade, roughness, and breathiness. In particular, the increase of agreement in rating breathiness was large, from $\kappa = 0.13$ to $\kappa = 0.35$. To illustrate the implication of these κ statistics, the number of voices for which all or all but 1 rater agreed on breathiness increased from 13 to 35. For grade and roughness, the changes in these high-agreement numbers were from 31 to 46 and from 22 to 30, respectively.

THE EFFECT OF VISIBLE SPEECH ON THE RATINGS OF AN INDIVIDUAL RATER

For each rater, the agreement between the 2 rating sessions was calculated. The intrarater agreement reflects the rater's consistency as well as the effect of visible speech on the ratings of the individual rater. A low intrarater agreement might indicate a large effect of visible speech and vice versa. **Figure 4A** shows the intrarater agreement expressed by κ , and **Figure 4B** shows the corresponding intrarater Spearman rank correlation coefficients for the GIBBAS parameters for each rater. It appears that the effect of visible speech on ratings varied considerably between raters. For instance, experts 2 and 5 had different breathiness ratings when they used visible speech (κ about 0), while expert 1 had similar ratings ($\kappa = 0.37$). Averaged across raters, the intrarater agreements on breathiness were relatively low, which reflects the large overall effect of visible speech on the rating of breathiness, as shown in **Figure 3**. The effect of visible speech on rating grade was relatively small for most experts, but it was markedly large for expert 4 ($\kappa = -0.01$). For this expert, the grade ratings with visible speech were shifted

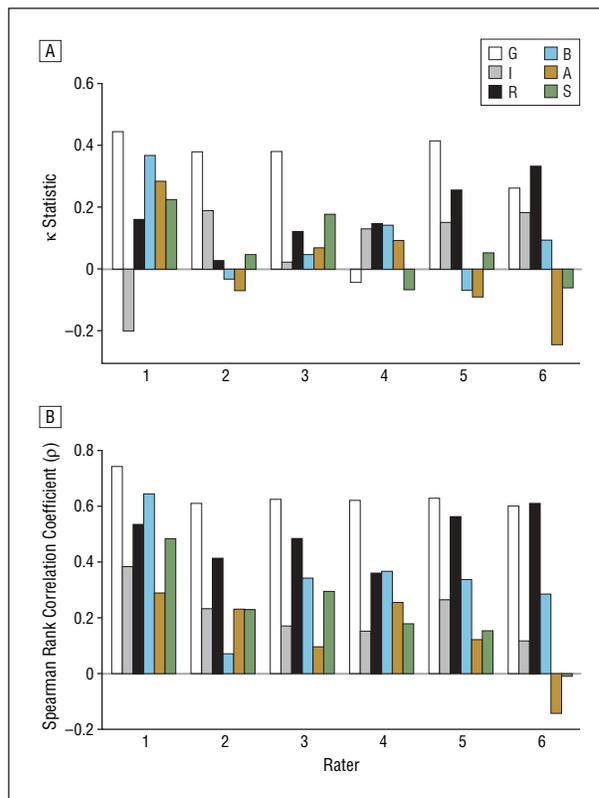


Figure 4. Comparisons between the ratings without and with visible speech of each rater separately. A, Cohen κ statistic; B, Spearman rank correlation coefficient. GIBBAS indicates grade, instability, roughness, breathiness, asthenicity, and strain.

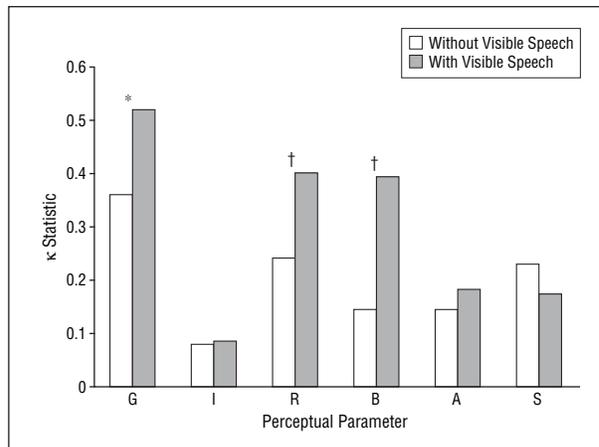


Figure 5. Statistics for 5 raters for GIBBAS scale (grade, instability roughness, breathiness, asthenicity, and strain) parameters, without and with visible speech. Compared with **Figure 3**, rater 4 has been left out. Significant differences between κ with and without visible speech are indicated with an asterisk or dagger (* $P < .05$, † $P < .001$).

systematically up compared with the ratings without visible speech, which is reflected by a relatively high correlation between the ratings ($\rho = 0.62$).

Because the markedly large effect of visible speech on rating grade for expert 4 might solely account for the increase in interrater agreement (**Figure 3**), an analysis of κ without expert 4 was performed. The results of this analysis, shown in **Figure 5**, demonstrate that removing the ratings of expert 4 does not change the outcome.

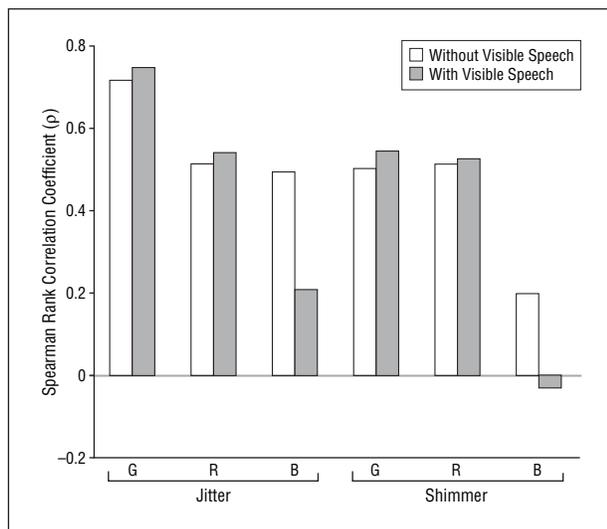


Figure 6. Correlations between perceptual parameters grade (G), roughness (R), and breathiness (B) and the acoustic parameters jitter and shimmer. Differences were not significant ($P > .05$).

If anything, the effect of visible speech on the rating agreement for grade, roughness, and breathiness among the 5 remaining raters (experts 1, 2, 3, 5, and 6) is even larger. The analysis was repeated for each combination of 5 raters, and it appeared that none of the experts had a large influence on the results.

ACOUSTICAL AND PERCEPTUAL PARAMETERS

We correlated the perceptual GIBAS parameters with various acoustical parameters. **Figure 6** compares correlations between subsets of perceptual (grade, roughness, and breathiness) and acoustical (jitter and shimmer) parameters for ratings with and without visible speech. No significant changes in correlation were found. Also, other acoustical parameters were correlated with all 6 GIBAS parameters. These correlations were in the range of those shown in Figure 6 (ρ between -0.4 and 0.7), and in no case was a significant effect of visible speech found.

A relatively stable portion, 250 ms after onset to 250 ms before offset, was used as a standard window to calculate the acoustical parameters. However, the acoustical parameters, and their correlations to perceptual parameters, might depend on the analysis window of the sustained vowel. Therefore, we investigated the effect of different selection windows on the correlations between acoustical and perceptual parameters. We compared the entire vowel including onset ramp and offset damp, the standard window, and a fixed-duration (1 second) window of 250 ms to 1250 ms after onset. These different selection windows did not produce different results on the correlations between acoustical and perceptual parameters.

COMMENT

Our study produced 2 pronounced results. First, the interrater agreement was clearly larger with the addition

of visible speech than without (information as provided in spectrograms of the voice track) for rating grade, breathiness, and roughness. Second, visible speech had no effect on the correlations of the GIBAS ratings with acoustical parameters.

Perceptual rating (particularly grade, breathiness, and roughness) is commonly used^{2,3,5,8,12} and recommended as a basic dimension for assessment of pathological voices, especially for otolaryngologists performing phonosurgery.¹⁷ The addition of visible speech to the clinical setting is feasible, since affordable computer programs can provide quasi-real-time visible speech. Hence, the enhancement of the interrater agreement by the addition of visible speech is an important finding.

FACTORS IN INTERRATER AGREEMENT

In attempting to increase reproducibility and optimize rating settings, effects of various factors have been widely tested. Factors that play a role in the interrater agreement are training and experience of the raters,^{4,5} rating scheme and scale,^{3,21} type of voice material,²² and type and range of voice disorders.²³ Training and experience are factors that have 2 effects on rating: on the one hand they will lead to more rating consistency, but on the other, they might cause individually based rating strategies. Experts have their own internal representation of voice qualities as roughness and breathiness.⁷ The former effect will enhance interrater agreement, and the latter effect will reduce it. De Bodt et al⁵ found a greater intrarater agreement in experienced (>3 years as a professional in voice pathology) than in inexperienced judges, which reflects a larger consistency. They hardly found an effect of experience on the interrater agreement, which might imply that differences between strategies cancel the increase in consistency. Use of an external anchor to overcome the effect of an internal representation has been tested by Chan and Yiu.²⁴ They found that in naive listeners rating roughness and breathiness, external anchors and a training program can enhance the interrater agreement.

A couple of methodological factors are the rating scheme and, within a certain scheme, the rating scale. Webb et al³ compared 3 perceptual rating schemes: the Buffalo voice profile, the vocal profile analysis scheme, and the GRBAS scheme. The agreement among 7 experienced raters, determined by κ and the intraclass correlation, was found to be highest with the GRBAS scheme. Apart from strain, all GRBAS parameters showed relatively good interrater agreements (κ around 0.4), whereas the Buffalo voice profile and vocal profile analysis had only 1 or 2 parameters with such agreements. An interesting method of phonetic labeling has recently been introduced by Revis.²⁵ In this method, each isolated phoneme of a standardized sentence is rated for the presence or absence of deviant characteristics such as breathy, unvoiced, and rough. This method produced a relatively large interrater and intrarater reproducibility. However, it requires a time-consuming preliminary manipulation of the signal.

In our study, we used a conventional ordinal 4-point scale. Alternatively, to enable a finer judgment, one could use more points on the scale or use an analog scale. Wuyts

et al²¹ compared a visual analog scale to an ordinal 4-point scale in an experiment with 29 judges rating the GRBAS parameters in 14 pathological voices. They found a greater interrater agreement with the 4-point scale. Indeed, considering the variability in rating scores in various studies, including ours (17% of the grade-roughness-breathiness rating with the addition of visible speech deviated >1 point), more precision than a 4-point scale might not be expected by using an analog scale, in which case the simpler 4-point scale is preferable.

Another factor is the type of speech segment. The results of a study by de Krom²² indicated that the type of speech (in his experiment, connected speech, complete sustained vowels, or sustained vowels without onset) had virtually no effect on either within- or between-listener consistency of the grade, breathiness, or roughness ratings. The listeners in his experiment had received the same training for voice evaluation.

Finally, rating agreement might depend on the type and severity range of voice disorders. Agreement is generally higher with extremes (normal or severe disorder) than with intermediate disorders.^{8,23} This is also apparent in our data of the roughness, asthenicity, and strain ratings, for which high agreements (5 or 6 out of 6 raters in agreement) were mostly found for the normal category.

The voice disorders in the present study were mostly intermediate (Figure 2) and therefore relatively hard to rate. On the other hand, the settings, the 4-point GIBAS scale, the steady-state vowels, and the experienced raters were generally advantageous for voice rating.

EFFECT OF VISIBLE SPEECH ON INTERRATER AGREEMENT

The enhancement of the interrater agreement found for grade, roughness, and breathiness can be ascribed to specific features in the spectrograms related to grade, roughness, or breathiness. Such features are the presence of noise (indicative of breathiness) and the presence of irregularities in the pulse pattern (indicative of roughness). One cannot exclude that the spectrograms also caused a bias toward a "false" judgment because the spectrograms cannot capture all aspects, which contributes to a perceptual judgment of grade, roughness, and breathiness. The implication would be a larger interrater agreement at the cost of accuracy. However, this is not likely because no systematic shifts have been found with the addition of visible speech: the average rating of grade increased, whereas it decreased for breathiness and did not shift for roughness (Figure 2). More essentially, ratings should distinguish between various degrees of voice disorders (for instance, before and after voice therapy), and therefore consistency (and thus, a large interrater agreement) is important. Considering the wide distribution of ratings (Figure 2B), the ratings with visible speech seem to distinguish well between various voices.

The improvements in reproducibility could have been due to changed rating behavior of one individual expert. However, removing a single rater from the κ statistics, and repeating this analysis for each rater, dem-

onstrated that the enhancement of agreement was not due to an individual rater (Figure 3 and Figure 5).

Compared with findings in the literature on interrater agreement,^{2,3,8,21} the κ values we found without the addition of visible speech were relatively low, in particular for breathiness (Figure 3). Over the 5 GRBAS parameters, our values are comparable with those of Wuyts et al,²¹ but lower than those of Dejonckere et al² and Webb et al.³ A factor that might at least partly explain this difference is the type of training. The raters in the study by Webb et al³ received the same training, and the raters in the study by Dejonckere et al,² although rating independently, were always from the same department and had been working as a team for years. In our study, the experts had different training backgrounds and were from different departments. Furthermore, most voices in our data had an intermediate degree of deviance (Figure 2), which is more difficult to rate than severe or normal voices.^{8,23}

The interrater agreement was significantly enhanced for grade, roughness, and breathiness but not for instability, asthenicity, or strain. If anything, the agreement for strain was even less. The lack of change for instability was expected because instability, if rated, would have occurred during the spoken sentence rather than during the sustained vowel, while visible speech gave only information about the sustained vowel. The presence of strain or asthenicity would be noticeable in the spectrograms by less higher formants for asthenicity and more higher formants for strain; thus, some effect is to be expected. However, we observed no effect, which raises the question of why the reproducibility for the asthenicity-strain rating remained low.

Considering various G(I)RBAS studies, one might say that among the GIBAS parameters, asthenicity, strain, and instability show the lowest interrater agreement values.^{2,3,5,21} A plausible hypothesis for the low values of asthenicity and strain is that behavioral aspects are difficult to evaluate on the sole voice sound. Furthermore, for some raters confusion can exist between asthenicity and breathiness.²¹ Because of the poor reproducibility, asthenicity, strain, and instability were omitted from the basic protocol for functional assessment of pathological voices of the European Laryngological Society.¹⁷ This omission of instability, asthenicity, and strain from the basic protocol might have led to less familiarity with instability, asthenicity, and strain for some of the raters, which could explain the low agreement values and the lack of effect of visible speech.

CORRELATION BETWEEN ACOUSTICAL AND PERCEPTUAL PARAMETERS

The ratings of grade, roughness, and breathiness changed considerably by the addition of visible speech, thus it was possible that correlations of grade-roughness-breathiness ratings with acoustical parameters had changed as well. However, these changes were not found. The correlations we found were generally comparable with those found by others.^{2,11,26} There are several arguments why correlations between

the GIRBAS parameters and acoustical parameters are not high and why visible speech has no effect on these correlations.

First, the GIRBAS parameters are not perceptual correlates of any known acoustical parameter, such as pitch for frequency and loudness for sound level. For instance, roughness relates to both fundamental period perturbations (eg, jitter) and the presence of subharmonics, and breathiness relates to amplitude perturbations (eg, shimmer) and the presence of noise. Moreover, small but abnormal perturbations cannot be perceived, which weakens the correlations. Second, voices were not only rated on the basis of the same sustained vowel, which is used for acoustical analysis, but also on the basis of a spoken sentence. Moreover, the onset and offset of the vowel are left out from the acoustical analysis, and onset and offset might well contribute to the perceptual judgment. De Krom²² found a higher rating reliability for vowel onset and whole-vowel stimuli than for postonset vowels, indicating a role of onset in the vowel quality judgment. The interpretation of the MDVP parameters relies on analysis of the relatively stable voice segment,¹⁸ excluding onset and offset; thus, acoustical analysis of onset based on the MDVP parameters will not be a proper approach to address the correlation between acoustical and perceptual parameters.

Features in the spectrograms that relate to GIRBAS are not captured by the acoustical MDVP parameters. For instance, visible irregularities in the spectrogram with a fine-time resolution corresponding to creaks and indicating roughness do not necessarily correlate to jitter because jitter reflects period-to-period details (aperiodicity in the range of 1% to 5%), which for the most part will not be observed in the spectrograms. The noise-harmonic ratio parameter, which reflects the presence of noise between 1500 and 4500 Hz, might be expected to correlate with breathiness. However, in our sample of voices, this correlation was not present in either rating condition. Visible speech had a large impact on rating breathiness, which suggests that for the most part, breathiness features as visually noted by the experts were related to noise in formant regions.

Our results confirm the notion that perceptual rating cannot be replaced by acoustical parameters, at least as produced by MDVP paradigms. The use of auditory models based on psychoacoustical and neurophysiological phenomena might provide objective measures, which can successfully replace subjective measures. An example of such an approach has been presented by Shrivastav and Sapienza,²⁷ who used a loudness model incorporating cochlear mechanisms to assess breathiness. Their model was partly successful in that the new measure correlated well with breathiness ratings but not better than some existing acoustical parameters. Another useful approach could be the application of artificial neural networks.²⁸

Perceptual and acoustic measures can be considered complementary. Hence, an optimal evaluation of voice quality is achieved according to a multidimensional protocol, including acoustic and perceptual measures.^{12,17,26} Laryngostroboscopy, aerodynamic measures, and self-assessment might be included as well.¹⁷

CONCLUSIONS

This study, based on data from 70 pathological voices perceptually rated by 6 experts on the GIRBAS scale, shows that the addition of visible speech clearly increases the interrater reproducibility for the 3 main parameters grade, roughness, and breathiness. Consequently, it enhances the reliability and relevance of perceptual evaluation, justifying widespread use. Implementation in the clinical setting is feasible, since inexpensive and easy-to-use computer programs can be used to obtain quasi-real-time visible speech. Also, this study shows that visible speech does not influence the correlation between acoustical and perceptual parameters. For clinical purpose, both measurements are useful because both carry complementary information about possible voice abnormalities. Therefore, we propose a multidimensional approach for assessing voice function.

Accepted for Publication: November 2, 2006.

Correspondence: Philippe H. Dejonckere, MD, PhD, Phoniatrics AZU F.02.504, Department of Otorhinolaryngology, University Medical Centre Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands (ph.dejonckere@umcutrecht.nl).

Author Contributions: Drs Martens, Versnel, and Dejonckere had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. *Study concept and design:* Dejonckere. *Analysis and interpretation of data:* Martens, Versnel, and Dejonckere. *Drafting of the manuscript:* Martens and Versnel. *Critical revision of the manuscript for important intellectual content:* Versnel and Dejonckere. *Statistical analysis:* Martens and Versnel. *Study supervision:* Versnel and Dejonckere.

Financial Disclosure: None reported.

Acknowledgment: We thank Michel Sardeman, MSc, for voice data acquisition and Lian Nijland, PhD, for her useful comments on an earlier version of the manuscript.

REFERENCES

1. Hirano M. *Clinical Examination of Voice*. New York, NY: Springer Verlag; 1981.
2. Dejonckere PH, Remacle M, Fresnel-Elbaz E, Woisard V, Crevier-Buchman L, Millet B. Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Rev Laryngol Otol Rhinol (Bord)*. 1996; 117:219-224.
3. Webb AL, Garding PN, Deary IJ, MacKenzie K, Steen N, Wilson JA. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Otorhinolaryngol*. 2004;261:429-434.
4. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res*. 1990;33:103-115.
5. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. 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.
6. Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. *J Acoust Soc Am*. 2000;108:1867-1876.
7. Kreiman J, Gerratt BR, Precoda K, Berke GS. Individual differences in voice quality perception. *J Speech Hear Res*. 1992;35:512-520.
8. Dedititis RA, Barros APB, Queija DS, et al. Interobserver perceptual analysis of smokers voice. *Clin Otolaryngol*. 2004;29:124-127.
9. Green DM. Auditory intensity discrimination. In: Yost WA, Popper AN, Fay RR, eds. *Human Psychophysics*. New York, NY: Springer; 1993:13-55.
10. Moore BCJ. Frequency analysis and pitch perception. In: Yost WA, Popper AN,

- Fay RR, eds. *Human Psychophysics*. New York, NY: Springer; 1993: 56-115.
11. Morsomme D, Jamart J, Wery C, Giovanni A, Remacle M. Comparison between the GIRBAS scale and the acoustic and aerodynamic measures provided by EVA for the assessment of dysphonia following unilateral vocal fold paralysis. *Folia Phoniatr Logop*. 2001;53:317-325.
 12. Bhuta T, Patrick L, Garnett JD. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *J Voice*. 2004;18:299-304.
 13. Koenig W, Dunn HK, Lacy LY. The sound spectrograph. *J Acoust Soc Am*. 1946; 18:19-49.
 14. Truax B. *Handbook for Acoustic Ecology*. 2nd ed. Burnaby, British Columbia: Cambridge Street Publishing; 1999.
 15. Yanagihara N. Significance of harmonic changes and noise components in hoarseness. *J Speech Hear Res*. 1967;10:531-541.
 16. Rontal E, Rontal M, Rolnick MI. Objective evaluation of vocal pathology using voice spectrography. *Ann Otol Rhinol Laryngol*. 1975;84:662-671.
 17. 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.
 18. Titze IR. Workshop on acoustic voice analysis; summary statement. Iowa City: National Center for Voice and Speech, University of Iowa; 1995.
 19. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960; 20:37-46.
 20. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378-382.
 21. Wuyts FL, De Bodt MS, Van de Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? an experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice*. 1999;13:508-517.
 22. de Krom G. Consistency and reliability of voice quality ratings for different types of speech fragments. *J Speech Hear Res*. 1994;37:985-1000.
 23. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am*. 1998;104:1598-1608.
 24. 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.
 25. Revis J. *Approche Phonétique de l'analyse Perceptive des Dysphonies*. Marseille, France: Université de la Méditerranée; 2004.
 26. Speyer R, Wieneke G, Dejonckere PH. Documentation of progress in voice therapy: perceptual, acoustic, and laryngostroboscopic findings pretherapy and posttherapy. *J Voice*. 2004;18:325-340.
 27. Shrivastav R, Sapienza CM. Objective measures of breathy voice quality obtained using an auditory model. *J Acoust Soc Am*. 2003;114:2217-2224.
 28. Schönweiler R, Hess M, Wubbelt P, Ptak M. Novel approach to acoustical voice analysis using artificial neural networks. *J Assoc Res Otolaryngol*. 2000;1: 270-282.