Agreement and reliability of voice quality ratings in anchored and unanchored protocols

Guus de Krom[†], **Rianneke Crielaard**[‡]
† University College Utrecht, University of Utrecht, the Netherlands
‡ Utrecht Institute of Linguistics, University of Utrecht, the Netherlands
E-mail: gkrom@ucu.uu.nl, Rianneke.Crielaard@let.uu.nl

ABSTRACT

Ratings of roughness, breathiness, strain, and asthenicity were obtained under two protocols: an unanchored one, in which listeners used conventional 7-point semantic interval scales, and an anchored one, in which listeners rated the same aspects, given an example human voice for each aspect and gradation combination. In order to assess intrarater (test-retest) consistency, listeners rated each voice twice in both protocols. Measures of rating agreement within and between listeners were computed, as well as variance estimates at different levels in the data hierarchy (i.e. listeners, speakers, and replicated ratings of speakers by listeners). Rating reliability coefficients were computed on the basis of these variance estimates for both protocols. For all four aspects, interlistener agreement and rating reliability was higher in the anchored protocol than in the unanchored protocol. The higher reliability could generally be attributed to a reduction of interlistener variance, though anchoring helped to magnify the speaker variance for ratings on the asthenic scale as well. These results suggest that anchoring using natural human voice fragments may help to improve the reliability of voice quality ratings.

1. INTRODUCTION

Previous studies have revealed that individual listeners may be inconsistent in their repeated ratings of a given voice, that different listeners may disagree in their evaluation of a given voice, and that individual listeners have different opinions regarding the number and nature of voice quality dimensions that can be distinguished. Many of these shortcomings relate to the fact that voice quality impressions are subjective by nature. Objective measures have been proposed as an alternative, though these have disadvantages as well. First, it has been argued that objective measures should always be related to subjective impressions, in order to be valid descriptors of voice quality [1]. Second, objectively measured data are not necessarily more reliable than subjective perceptual judgements [2].

For these reasons, we conclude that perceptual ratings will continue to serve as a standard against which objective measures are validated, and that there remains a need for methods and protocols yielding more valid and reliable perceptual data. The present study is aimed at the development and testing of a voice quality rating protocol with enhanced reliability.

A conventional design in voice quality perception experiments is one employing a group of listeners who repeatedly rate a set of voice fragments on a specific voice quality aspect. With this type of design, the rating variance has three origins: actual voice quality differences between the speakers who produced the voice fragments (interspeaker variance = true item variance), differences in the ratings of a given speaker by different listeners (interlistener variance), and inconsistencies in the ratings of a given speaker by a given listener (intralistener variance). Because reliability can be defined as the ratio of a true item variance component and other variance components, high rating reliability can be achieved by (1) increasing the interspeaker variance, (2) reducing the intralistener and / or interlistener rating variance, or (3) a combination of (1) and (2).

Listeners base their voice quality impressions on internal, subjective criteria. Factors like language background, type of education, type and degree of professional experience may differ across listeners, and may cause listeners to weigh similar acoustic information differently, or to develop idiosyncratic perceptual criteria, which contributes to the interlistener variance component.

Intralistener rating variance emerges because voice quality impressions are inherently unstable to a certain degree, giving rise to inconsistent input (stimulus)-output (rating) relations.

The use of anchored rating protocols has been suggested as a possible means to reduce intra- and interlistener variance, and hence to boost rating reliability. The idea is to provide a subject with one or a few explicit reference (anchor) stimuli that can be compared to the actual test stimuli. The anchor stimuli are thus intended to serve as a specimen with given (fixed) properties, for instance having a certain degree of roughness. In speech research, anchored protocols were first used in psychophysical experiments [3, 4]. Gerratt et al. were the first to apply anchoring to voice quality evaluation [1]. They created a series of synthetic /a:/ vowels which represented a roughness continuum. Listeners were asked to rate the degree of roughness, using a five-point interval scale. Their results showed that rating reliability was higher in the anchored than in the unanchored condition. The higher

 ISBN 1-876346-48-5

rating reliability was attributed to an increased rating consistency both within and between listeners, in other words a decrease in both intra- and interlistener variance. Steutel and de Krom [5] extended this experiment. Synthetic /a:/ vowels were varied with regard to jitter and shimmer in order to simulate different degrees of roughness. Rating reliability was compared for anchored and unanchored conditions. The listeners were also asked to rate recordings of naturally produced vowels, using the same synthetic anchors. Again, anchoring had a slight positive effect on rating reliability, for both synthetic and naturally produced test vowels.

The aim of the current study was to investigate whether the use of naturally produced connected speech fragments for anchors would also yield increased rating reliability. We opted for connected speech, because it obviously reflects conversational speech much better than synthetic vowels or stable segments of naturally produced sustained vowels.

2. METHODS

2.1 Recordings

Recordings were made of 32 male and 41 female speakers, aged between 18 and 76 years. All 73 speakers suffered from some type of voice disorder. Recordings were made in quiet rooms in local speech therapy clinics. The speakers were asked to read aloud the Dutch version of "the North wind and the sun" at comfortable pitch and loudness. For this study, we only used the second phrase of the story. Recordings were made using a condenser microphone placed off-axis at a distance of 30 cm from a speaker's mouth. The voice fragments were stored on a digital audio tape, later downsampled to 16 kHz, segmented by hand, and converted to AIFF file format.

2.2 Selection of anchor stimuli

Anchors were selected in a separate session involving three experienced listeners, who were asked to rate all 73 voice fragments on the degree of roughness, breathiness, asthenicity and strain. Each stimulus (voice fragment) was repeated once in each session. Stimuli were presented over headphones in a double-walled, sound-treated booth, under software control. Ratings were given on scales ranging between 0 and 6. Listeners could rate a stimulus by pressing a button in a 4 (quality aspects) × 7 (scale values) matrix presented on a computer screen. Listeners could replay each stimulus as often as they wished. Six ratings were obtained for each stimulus on each aspect (3 listeners × 2 replicated ratings). Mean ratings and 95% confidence intervals were calculated for each stimulus. The widths of the confidence intervals ranged between 0 (at the high end points of the breathiness, strain, and asthenic scales and the low end point of the breathiness scale) and 2.47 (at the center of the roughness scale). Across each scale, the average width of the confidence interval was .99 for breathiness, 1.08 for roughness, .98 for asthenic, and 1.13 for strained. Twenty-eight stimuli had to be selected for use as anchor stimuli in the actual listening test. Anchor stimuli were selected on the basis of (1) the mean rating, which should be as close to an integer scale point as possible, (2) the width of the confidence interval, which should be as narrow as possible, and (3) the mean rating on the other aspects, which should be as low as possible, in order not to contaminate judgements on one aspect with a possibly deviant impression on other aspects. A set of 24 unique fragments was selected to serve as anchors: four fragments were used as anchors for two different aspects. All but three of the 24 fragments were obtained from female speakers; inclusion of three males speakers was needed to cover the higher end of the roughness scale. None of the selected anchor stimuli was included in the set of test stimuli. The resulting set of 49 test stimuli consisted of recordings of 29 males and 20 females.

2.3 Listening experiment

Ten female speech therapists participated in the experiment. Each listener participated in both an anchored and an unanchored session, separated by a time span of at least one week. The order of anchored and unanchored sessions was balanced across listeners.

In the unanchored session, listeners were asked to rate the test stimuli using the same experimental setup as the one used in the pre-test anchor selection listening session. For each test stimulus, listeners were free to determine the order in which the four aspects were to be rated.

In the anchored session, the listeners were asked to compare a test stimulus against the set of anchor stimuli. The experimental setup was similar to the one used in the unanchored session, except that mouseclicking a response button in the matrix made the corresponding anchor stimulus audible. As in the unanchored protocol, the listeners could mouseclick a "repeat stimulus" button as often as they wanted, making the experiment self paced.

3. RESULTS

3.1 Mean ratings and confidence intervals

Calculated across speakers, listeners and replicated ratings, the (overall) mean rating lay between 2.09 and 2.56 for all aspect / protocol combinations, except for asthenic in the unanchored protocol, which had an overall mean rating of only 1.54. On a scale from one to six, these mean values lie less than one full scale point below the center of the scales, indicating that the speakers represented a mixture of severities for most aspects. Few speakers received mean ratings higher than 4.0 on any of the scales: for asthenic in the unanchored protocol, the highest mean rating assigned to a speaker was 3.8. In the anchored protocol, the mean asthenic ratings were more evenly distributed across the scale (i.e higher interspeaker variance), which suggests that listeners were better able to discriminate between speakers.

ISBN 1-876346-48-5

95% Confidence intervals of mean ratings were calculated for each voice fragment, in both anchored and unanchored protocols. The average widths of the confidence intervals, calculated across the entire range of the scales, ranged between .48 (asthenic, anchored), and .64 (strain, unanchored). For all four aspects, the confidence intervals were widest at the centers of the scales., narrowing at the lower, and to a lesser extent also the higher ends of the scales, indicating floor and ceiling effects: These findings agree with results obtained in previous studies [].

3.2 Intralistener agreement: consistency within listeners

For voice quality ratings to be reliable, a sufficiently high intralistener (test-retest) agreement is one of the prerequisites. We calculated the mean difference between a listener's first and second ratings of a given speaker. Half of the listeners first participated in an anchored test, followed by an unanchored one, while the others took the tests in the reverse order. Results are given in Table I.

	UA		AU	
	U	A	U	A
R	.67 (.06)	.65 (.05)	.67 (.05)	.80 (.07)
B	.81 (.06)	.52 (.04)	.60 (.05)	.72 (.06)
A	.68 (.06)	.67 (.04)	.62 (.05)	.74 (.06)
S	.67 (.05)	.74 (.05)	.69 (.05)	.82 (.06)

Table I. Intralistener agreement: mean difference (standard error) between a listener's first and second ratings of a speaker for unanchored (U) and anchored (A) protocols.
R = rough, B = breathy, A = asthenic, S = strained. Protocol orders: UA: unanchored-anchored; AU: anchored-unanchored. N = 245 for each aspect.

Within one protocol order, means for anchored and unanchored data were compared using a paired data t-test. The only significant difference (5% level) was for breathiness in the unanchored-anchored (UA) protocol order, with a marginally lower mean value for anchored ratings (.52 versus .81). T-tests for independent data were used to compare means across protocol orders, because these involved different listeners. Two comparisons yielded a significant difference (at the 1% level). Anchored breathiness ratings in the unanchored-anchored protocol order had a lower mean than anchored breathiness ratings in the anchored-unanchored order. The reverse was found for unanchored breathiness ratings: a lower mean value in the anchored-unanchored protocol order.

3.3 Interlistener agreement: consistency between listeners

Measures of interlistener agreement were calculated, similar to those for intralistener agreement. Again, separate analyses were performed for the two protocol orders. Values are given in Table II.

	UA		AU	
	U	A	U	A
R	1.73 (.05)	1.24 (.04)	1.43 (.04)	1.43 (.04)
B	1.60 (.04)	.93 (.03)	1.34 (.04)	1.38 (.04)
A	1.24 (.04)	1.10 (.03)	1.32 (.04)	1.19 (.04)
S	2.02 (.05)	1.28 (.03)	1.30 (.03)	1.43 (.04)

Table II. Interlistener agreement: mean difference (standard error) between a listener's rating of a speaker and another listener's rating of that speaker for unanchored (U) and anchored (A) protocols. R = rough, B = breathy, A = asthenic, S = strained. Protocol orders: UA: unanchored-anchored; AU: anchored-unanchored. N = 980 for each aspect.

The values in Table II pertain to the mean difference between any two listeners' ratings of a speaker. As could be expected, these values are higher than the values given in Table I, which pertain to rating inconsistencies within listeners. Anchored and unanchored ratings within the anchored-unanchored protocol order did not differ, but in all other cases, the difference was significant at 1%. In the AU protocol order, anchored strain ratings had in fact a significantly higher mean value than unanchored ratings. In all other cases, however, the mean difference between any two listeners's ratings was smaller for anchored ratings than for unanchored ones, indicating higher rating agreement between listeners in the anchored protocol. Table II shows that protocol order clearly matters: the difference between the means for anchored and unanchored ratings is larger if the order is unanchored-anchored than if it is anchored-unanchored. One possible explanation is that having been exposed to anchor stimuli in the first listening session (the anchored-unanchored order) has a long lasting effect on listeners. The following unanchored protocol

3.4 Rating reliability: variance components analysis

Though high levels of rating consistency are a prerequisite for high rating reliability, they do not guarantee this. Even perfectly consistent ratings may be unreliable if there is no true item variance to begin with, i.e. if listeners are not sensitive to actual voice quality differences between speakers. To determine the overall reliability of the ratings, the total rating variance was therefore decomposed to interspeaker (item) variance, intralistener variance, and interlistener variance, using a multilevel analysis technique [6]. The analyses yielded a three-level variance components model (Equation 1)

$$y_{ijk} = \gamma_{000} + (v_k + u_{jk} + e_{ijk}) \quad (1)$$

In equation (1), the fixed part γ_{000} models the effect of the grand mean rating; the random part (in parentheses) describes the variance estimates. The total variance is

ISBN 1-876346-48-5

decomposed to three levels: intralistener variance $s^2(e_{ijk})$, interspeaker variance $s^2(u_{jk})$, and interlistener variance $s^2(v_k)$. By expressing reliability as a ratio of estimated variance components, measures of intra- and interlistener agreement (measures of listener consistency) are now simultaneously related to the estimated magnitude of the inter-speaker (item) variance (a measure of listener sensitivity). This has an additional value over the analyses of intralistener and interlistener agreement reported in Tables I and II, which do not take the interspeaker variance into account. A rating reliability coefficient r_{xx} (in %) was defined on the basis of the ratio of the interspeaker (item) variance estimate and the total variance estimate, which was in turn defined as the sum of the three independent variance estimates (Equation 2).

$$r_{xx} = \frac{s^2(u_{jk})}{s^2(v_k) + s^2(u_{jk}) + s^2(e_{ijk})} \times 100\% \quad (2)$$

Rating reliability coefficients for the different aspects and task orders are given in Figure 1.

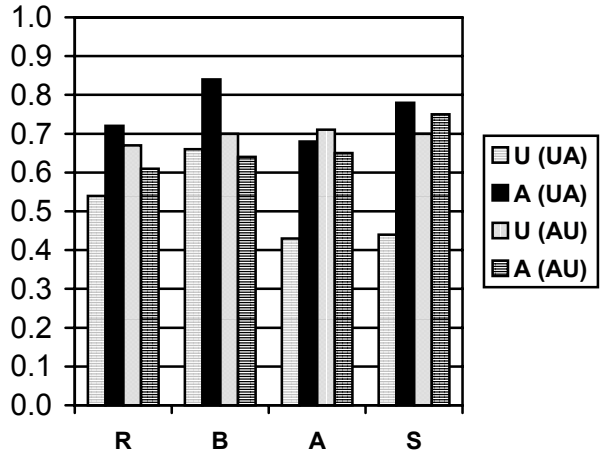

Figure 1. Reliability coefficients for unanchored (U) and anchored (A) ratings in two protocol orders: UA: unanchored-anchored; AU: anchored-unanchored. R = rough, B = breathy, A = asthenic, S = strained.

Figure 1 illustrates that the effect of anchoring on rating reliability is positive if the first listening session is the unanchored one. Rating reliability increases most for the aspects asthenic and strained, with gains of 25 and .34%, and less for rough and breathy (.both gains 18%). In the anchored-unanchored protocol order, the reliability for anchored is actually slightly lower for rough, breathy, and strain (losses of .6%, .7% for strain), whereas the reliability of asthenic ratings increases marginally with 5%.

4. CONCLUSIONS

Anchoring with natural connected speech fragments may have a positive effect on measures of rating agreement and reliability. The main advantage of an anchored over an unanchored protocol is that interlistener agreement is increased, presumably because the anchors serve to create a common frame of reference for the listeners. Intralistener (test-retest) agreement is not affected much by anchoring: apparently, the frame of reference that an individual listener applies when rating the quality of a voice seems to be relatively stable. Indeed, if the main function of an anchor stimulus is to create a common and stable frame of reference, anchoring may be expected to have a larger impact on interlistener agreement than on interlistener agreement, because the ratings of individual listeners tend to differ more from each other than repeated ratings of a given listener.. The increased interlistener agreement has a direct bearing on the reliability of the ratings. All things being equal, a reduction of interlistenr variance alone will already have a positive effect on rating reliability. Obviously, this positive effect is magnified if intralistener variance also decreases, and / or if if speaker variance increases as well. In this study, a clear increase of interspeaker variance was found for asthenic ratings in the anchored protocol. It seems that anchoring helped the listeners to define the perceptual space for this aspect. Asthenic ratings are often reported to have relatively low reliability [7], which may be related to the ill-defined nature of the aspect

REFERENCES

[1] B.R. Gerratt, J. Kreiman, N, Antoñanzas-Barroso and G.S. Berke, "Comparing internal and external standards in voice quality judgements", *J. Speech Hear. Res*, vol. 36, pp. 14-20, 1993.

[2] C.R. Rabinov, J. Kreiman, B.R. Gerratt and S. Bielamowicz, "Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter," *J Speech Hear. Res*., vol. pp. 38, 26-32, 1995.

[3] I. Pollack, "The information of elementary auditory displays. II." *J. Acoust. Soc. Am*., vol. 100, pp. 1787-1795, 1953.

[4] J.E. Berliner, N.I. Durlach and L.D. Braida, "Intensity perception. IX. Effect of a fixed standard on resolution in identification," *J. Acoust. Soc. Am*., vol. 64, pp. 687-689, 1978.

[5] C. Steutel and G. de Krom "A comparison of anchored and unanchored rating protocols," *Proceedings ICPhS95*, pp. 622-625, 1995.

[6] H. Goldstein, *Multilevel statistical models* (2nd ed.), London: Edward Arnold, 1995.

[7] M.S. de Bodt, F.L. Wuyts, P.H. van de Heyning and C. Croux "Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality," *J. Voice*, vol. 11, pp. 320-326, 1997.