# A COMPARISON OF ANCHORED AND UNANCHORED RATING PROTOCOLS

*Christine Steutel and *Guus de Krom*
** Research Institute for Language and Speech, University of Utrecht, the Netherlands*

## ABSTRACT

It has been shown that the consistency and reliability of roughness ratings on synthetic vowels may be higher in an anchored rating protocol than in an unanchored one [1]. In the present study, synthetic and natural vowels were rated on roughness, using the two protocols. For both types of stimuli, mid-scale ratings were more consistent in the anchored than in the unanchored protocol. Rating reliability was also slightly higher in the anchored protocol.

## INTRODUCTION

In the perceptual evaluation of pathological voice quality, listeners usually rate voices on a number of aspects (e.g. roughness, breathiness), using Equal-Appearing Interval scales (EAI). However, the consistency and reliability of the obtained ratings may be rather low [2], which suggests that a listener's internal criteria are unstable, and that dissimilar criteria may be applied by different listeners.

As an alternative, listeners may be asked to match stimuli against a set of selected reference (anchor) signals, which constitute a continuum on the scale to be evaluated. Results obtained in a study with synthetic vowels indicated that roughness ratings were more consistent and reliable when obtained in such an anchored protocol [1].

The question remains whether a perceptual evaluation of natural voice samples would also benefit from an anchored rating protocol, as natural voices typically differ on more than one (i.e. the anchor stimulus) dimension. In this study, the consistency and reliability of roughness ratings was compared for anchored and unanchored protocols, using both synthetic and natural vowels.

## METHODS

### Material

The material that was used consisted of vowels /a/ with a duration of one second (78 natural and 25 synthetic vowels).

The *natural* vowels were produced by 78 speakers (males and females), including 57 voice patients suffering from different types and degrees of dysphonia. All recordings were made at comfortable pitch and loudness. The vowels were segmented to a duration of one second, beginning at the onset, and were given 12.5 ms linear ramped offsets to prevent audible clicks (see [2], for details).

The 25 *synthetic* vowels were produced with a voice synthesiser that allowed manipulation of several source and filter parameters [3]. For all vowels, F0 was set to 155 Hz (a value in between typical F0 values for males and females). The F0 vibrato frequency was 5 Hz (frequency modulation depth 3%, amplitude modulation 5%). The Open Quotient (open time / closed time) and Speed Quotient (opening time / closing time) of the glottal pulse were .5 and 1.16, respectively. The center frequencies of the formants F1 to F5 were .79, 1.35, 2.6, 4.0, and 5.0 kHz, with a bandwidth of 10% of the center frequency.

To create a series of vowels differing in roughness, the source parameters jitter and shimmer were varied, with all other parameters constant. Jitter was defined as a random fluctuation in the duration of source signal periods, and was expressed as a percentage of the mean period duration (%jitter). Shimmer was defined similarly, referring to fluctuations in the peak amplitudes of signal periods (%shimmer). The synthetic vowels were given 12.5 ms ramped onsets and offsets.

Ten of the 25 synthetic vowels were selected by the authors to serve as reference stimuli in the anchored protocols. We tried to create an equal perceptual distance between two successive reference signals. Figure 1 gives the %shimmer and %jitter values for the synthetic signals.

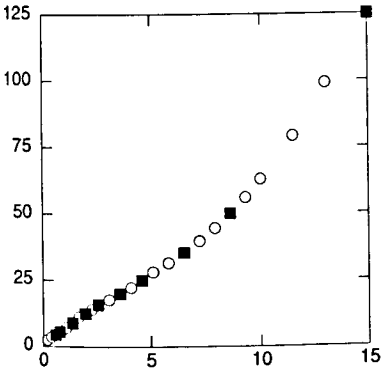


*Figure 1. %jitter (x-axis) against %shimmer (y-axis) for the synthetic vowels. Reference stimuli are marked with squares.*

### Listeners

The listeners were 24 females (12 students of Speech Pathology in their final grade and 12 speech pathologists). The listeners were paid for their co-operation.

### Perceptual evaluation

The perception experiments were performed with the help of a personal computer. In both types of protocols, a given stimulus was repeated until it had been assigned a rating.

In the *unanchored* protocol, the listeners were asked to press a key (1 to 10) corresponding to the degree of roughness of the presented stimulus: 1 was defined as "no roughness at all", and 10 as "the maximum degree of roughness conceivable".

In the *anchored* protocol, the listeners were asked to select the reference stimulus that sounded most like the current test stimulus. The reference stimuli could be made audible by pressing a key (key 1 produced the reference stimulus with the lowest shimmer/jitter values, and so forth).

Each listener participated in all four experiments (synthetic and natural stimuli, anchored and unanchored protocols). All stimuli were rated twice by each listener. The stimuli were randomised in each block. The four blocks were presented in two larger sessions A and B, separated by at least a week. In both sessions, an unanchored rating task preceded an anchored one. Session A began with the synthetic stimuli, session B with the natural stimuli. Half of the listeners began with session A, the other half with B.

### Statistical analyses

The statistical analyses were performed with a multilevel analysis program [4], which allowed the simultaneous estimation of variance components at three data levels (i.e. the listeners, the stimuli, and the replicas [first and second ratings of a given stimulus]).

*Confidence intervals* reflect the effects of rating inconsistencies. If confidence intervals around any two scale values overlap, these values cannot be considered distinct. Hence, confidence intervals provide information on the true resolution of the roughness scales. The determination of the confidence intervals

is described in detail in a previous study [2].

*Rating reliability coefficients* relate the magnitude of the variance component of interest (i.e. the stimulus variance [sv]) to the magnitudes of other variance components (i.e. the listener variance [lv], and the replica variance [rv]) [5]. Reliability coefficients ρ were defined as given in Equation 1:

$$\rho \equiv \frac{sv}{sv + (lv/24) + (rv/2)} \qquad [1]$$

The division of the lv and rv estimates by 24 and 2 is a compensation for the number of independent measurements at the listener and replica levels [2].

## RESULTS

### Correlations

Pearson correlation coefficients between ratings in the anchored and unanchored protocols were .96 for natural stimuli and .99 for synthetic stimuli. These high values suggest that the listeners had judged similar aspects in both protocols, and that the data obtained in these two protocols were therefore comparable.

### Confidence intervals

Between- and within-listener confidence intervals (95%, two-tailed) for **synthetic** stimuli are given in Figures 2 and 3.

As can be observed, the bandwidths of the confidence intervals were narrower in the anchored than in the unanchored protocol for the larger part of the scales. The effect was stronger for between-listener data than for within-listener data. This is not surprising; one may expect that internal criteria applied by different listeners are more dissimilar than internal criteria applied by a given listener on different listening trials. Providing different listeners with anchor stimuli elimates a part of these dissimilarities.

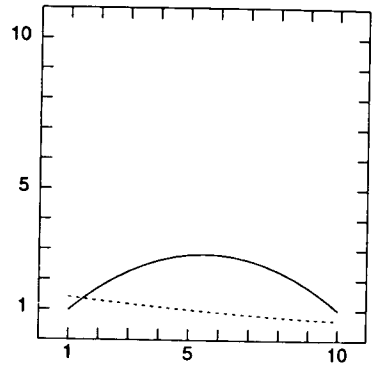


*Figure 2. Between-listener confidence intervals (vertical) for ratings on **synthetic** stimuli. Solid: unanchored protocol; Dotted: anchored protocol.*

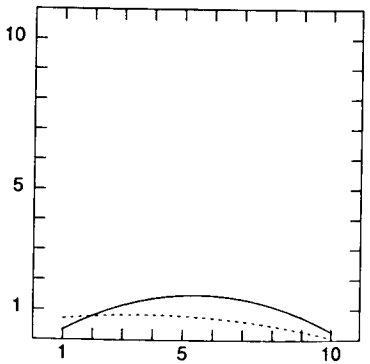


*Figure 3. Within-listener confidence intervals (vertical) for ratings on **synthetic** stimuli Solid: unanchored protocol; Dotted: anchored protocol.*

Confidence intervals for **natural** stimuli are given in Figures 4 and 5.

As for the synthetic stimuli, confidence intervals were narrower for anchored ratings than for unanchored ratings along the larger part of the scales. The effect was again most marked for between-listener data. Overall, the effect of anchoring was more pronounced for the synthetic vowels, which may relate to



the fact that the reference stimuli were drawn from the same continuum.

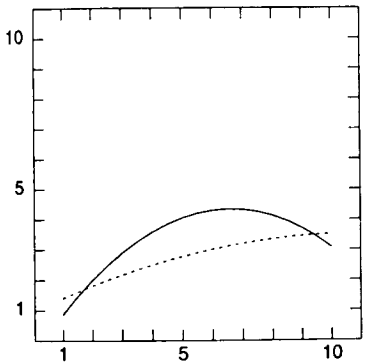


*Figure 4 Between-listener confidence intervals (vertical) for ratings on **natural** stimuli. Solid: unanchored protocol; Dotted: anchored protocol.*

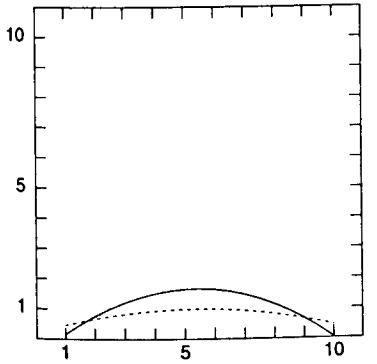


*Figure 5. Within-listener confidence intervals (vertical) for ratings on **natural** stimuli. Solid: unanchored protocol; Dotted: anchored protocol.*

### Reliability coefficients

Reliability coefficients are given in Table 1. These data indicate that anchored ratings were slightly more reliable than unanchored ratings. The magnitude of the effect of anchoring on rating reliability is closely comparable for synthetic and natural vowels. Overall, however, synthetic stimuli were rated more reliably than natural stimuli.

*Table 1. Reliability coefficients ρ for roughness ratings in anchored and unanchored protocols, using natural and synthetic stimuli.*

| vowel type | protocol | ρ |
|---|---|---|
| natural | unanchored | .84 |
| natural | anchored | .87 |
| synthetic | unanchored | .93 |
| synthetic | anchored | .97 |

## CONCLUSIONS

The data indicate that the presentation of (synthetic) anchor stimuli may improve the consistency and reliability of roughness ratings on synthetic and natural vowels. Anchoring seems to have a stronger positive effect on between-listener rating consistency than on rating consistency within a listener.

## REFERENCES

[1] Gerratt, B.R., Kreiman, J., Antoñanzas-Barraso, N., & Berke, G.S. (1993). Comparing internal and external standards in voice quality judgements. *Journal of Speech and Hearing Research, 36,* 14-20.

[2] de Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of vowel fragments. *Journal of Speech and Hearing Research, 37,* 985-1000.

[3] Pabon, J.P.H. (1994). A real-time singing voice synthesizer (Alto). *Proc. SMAC 1993,* Stockholm, pp. 288-293.

[4] Prosser, R., Rasbash, J., & Goldstein, H. (1991). ML3-software for three-level analysis. *Users' Guide for V.2.* London: University of London, Institute of Education.

[5] Asendorpf, J., & Wallbott, H.G. (1979). Maße der Beobachter-übereinstimmung: ein systematischer Vergleich. *Zeitschrift für Sozialpsychologie, 10,* 243-252.