The Visual Sort and Rate method for perceptual evaluation in listening tests

Svante Granqvist, Lic Eng
Department of Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden

Abstract

Three methods for perceptual rating of audio stimuli were compared. The Visual Analogue Scale, VAS, was implemented as two computer programs and compared to a VAS where the responses were given on paper. The first program is a straightforward implementation allowing multiple playbacks and re-play of previously heard stimuli. The second introduces the Visual Sort and Rate method, VSR. This method facilitates comparing similar stimuli to each other, thus making the rank ordering of the stimuli easier. The three methods were compared by using two sets of stimuli. The first set was a synthetically generated series of stimuli mimicking the vowel /a/ with different spectral tilts. In this test, a single parameter was rated. The second set of stimuli was a naturally spoken voice. For this set of stimuli three parameters were rated. Results show that the VSR method gave better reliability of the subjects' ratings in the single-parameter tests; Pearson and Spearman correlation coefficients were significantly higher for the VSR method than for the other methods. For the multi-parameter, intra-subject test, significantly higher Pearson correlation coefficients were found for the VSR method than for the VAS on paper.

Introduction

Listening tests are needed for relating acoustical parameters to human auditory perception. They can also be used to quantify perceptual parameters in cases where the acoustical characteristics underlying the sound quality are unknown. For example, in the field of human voice quality evaluation, well-working sets of voice quality terms have been developed (Isshiki et al., 1969, Hammarberg, 1986; Titze, 1994).

In the literature, the reliability of perceptual evaluation has been compared with acoustic measures of voice quality (Sederholm et al., 1993; Gauffin et al., 1995; Rabinov et al., 1995; McAllister et al., 1995, 1996). Some authors are in favour of the acoustic measures, others favour perceptual evaluation. Such comparisons are relevant, but it is not to be expected that either of the methods will eliminate the need for the other. Voice quality is primarily a perceived phenomenon. On the other hand, acoustic measures derived by computer programs are more closely related to physiology and should ideally be more objective. Such programs are easily copied between computers and ideally yield identical results, regardless of who operates them. Thus, both methods are necessary for future development of voice quality judgement and both methods can be developed for better accuracy.

The variance in data obtained from listening tests is sometimes large, and methods of increasing inter- and intrajudge reliability are desirable (Kreiman et al., 1993). This variance has often served as an argument against listening tests for measuring voice quality, thus favouring objective measures. As a rule, many sessions are needed in listening tests to reach statistical significance. However, playing the stimuli many times is tiring for the subjects and there is a limit beyond which fatigue starts to degrade the accuracy of the results. One alternative is to split the test in several listening sessions, which however entails practical problems with scheduling subjects etc.

If it could be shown that part of the variance in the data is due to the listening test situation, rather than to the listeners' abilities, and if the listening test situation can be improved, the value of the perceptual rating method would increase accordingly. Any method that leads to increased accuracy in subjects' responses would be welcome.

Using the Visual Analogue Scale, VAS, where the subject puts a mark on a 10 cm long scale, has become a standard for rating

perceptual parameters, although a considerable variance in the data still remains (for a résumé, see Wewers et al., 1990). In particular, it is generally difficult for listeners to rank order a set of stimuli according to a specific perceptual quality, when no opportunity is offered for direct comparison of stimuli adjacent in the rank order. This difficulty can be overcome by creating all possible pairs of stimuli and asking the subject to rank the order within pairs, only, rather than rating each stimulus separately. The problem with this approach is that the number of pairs increases with N^2 (N is the number of stimuli), which easily leads to an overwhelming number of pairs.

To improve rating consistency, Gerratt et. al. (1993) and Berliner et al. (1978) used anchors as fixed references for the listeners during the test. Both these studies used synthetic stimuli with systematically varied, well-defined acoustic differences and investigated the effect of introducing these anchors on the reliability of the listeners' ratings. In the evaluation of voice qualities of naturally spoken voices though, the anchoring procedure presents some problems. First, obtaining naturally spoken anchors might be difficult, since there mostly are no objective measures of where the anchor *should* to be positioned along a rating scale. Second, the perceptual differences between anchors selected from naturally spoken voices are most probably multi-dimensional. This might cause the subjects to disagree even on the ordering of the anchor stimuli, which of course would lead to confusion. Third, there is a possibility that some stimuli will fall outside the range of the anchors, which would cause ceiling or floor effects.

It is evident that both comparison within pairs and introducing anchor stimuli are associated with certain problems. The present article compares two experimental designs of listening tests with the VAS on paper method. In one design, subjects were free to listen to the stimulus any number of times, and to quickly proceed to adjacent stimuli. In the other design, the Visual Sort and Rate method, the listener's task was first to sort and then to rate the stimuli. It will be shown that this latter method improves listeners' performance.

Stimuli

Two sets of stimuli were used. The first set consisted of synthetic stimuli with a single known parameter changing (spectral tilt), thus making examination of rank ordering possible. The second set was naturally spoken vowels from the same person recorded at different stages of voice therapy.

The first set of stimuli was created with the Addsynt program (Granqvist, 1996). In this program, a voice source was created by adding sinusoids, and thus the spectral tilt of the source could be modified in a controlled manner. The source signal was then fed through a standard set of formant filters in order to mimic a spoken vowel /a/. Spectral tilts for the stimuli were chosen randomly, since an equidistant distribution of tilt values could unfairly improve the results from the VSR method; if the tilt values had been equidistant, it would be sufficient for the subject merely to sort the stimuli correctly and rate them as equally spaced and a perfect match would emerge. A random distribution of tilt values forces the subjects to distinguish between large and small stimuli differences.

The spectral tilt parameter gives the stimuli different degrees of high-frequency content. Generally, variation of overall spectral tilt is induced by variation of subglottal pressure and hence related to vocal loudness as well as pressed or breathy phonation.

Synthetic stimuli for intra-subject reliability test

For this test, 24 random numbers, ranging from -3.1 to -12.2, were generated and those numbers were used for the voice source spectral tilt expressed in dB per octave. From these 24 stimuli, six series, each consisting of 13 stimuli, were selected (Table 1). The selection was made in a semi-random way, so that each stimulus appeared in three of the six series, except for the stimuli with the steepest and the flattest tilts that appeared in all series. The aim of this seemingly complicated procedure was to expose the subjects to the same stimuli several times in different contexts. If the six series would have been identical, there would be a risk that they reproduced the same "picture" of stimuli position with the VSR method, rather than really rating each stimulus.

Synthetic stimuli for inter-subject reliability test

For this test, one of the series of 13 stimuli included in the test just described was used. All subjects rated the same stimuli.

Table 1. The stimuli used in the listening sessions. The stimuli used in the different sessions are listed in columns 1 to 6. The intra-subject tests were performed in 6 sessions for each of 2 subjects. The inter-subject test was performed in 6 sessions with 6 different listeners and the stimuli in column 1 was used for all sessions.

Synthetic stimuli						
Session Stimulus [dB/oct]	1	2	3	4	5	6
3.1	x	x	x	x	x	x
3.5		x	x	x		
3.6				x	x	x
3.9	x		x			x
4.0			x		x	x
4.7	x	x			x	
4.8	x			x		x
5.7		x	x			x
6.6	x	x		x		
7.0		x			x	x
7.3		x			x	x
7.6	x		x		x	
8.0		x	x		x	
8.1				x	x	x
9.1			x	x		x
9.3	x	x	x			
9.6	x	x		x		
9.7		x	x	x		
9.8	x	x	x			
10.0	x			x	x	
10.5	x				x	x
10.8	x			x		x
12.0			x	x	x	
12.2	x	x	x	x	x	x

Human stimuli						
Session Stimulus number	1	2	3	4	5	6
0				x	x	x
1	x	x				x
2				x	x	x
3		x		x	x	
4	x	x				x
5	x			x	x	
6			x		x	x
7	x		x	x		
8		x			x	x
9		x	x			x
10	x		x		x	
11			x	x	x	
12	x	x	x			
13	x			x		x
14	x	x	x			
15		x	x	x		

Human stimuli for intra-subject reliability test

The stimuli in this test emanated from a material consisting of recordings of patients with vocal nodules repeatedly made during the voice therapy period (Holmberg et al., 2000). From this material, 16 recordings of one patient were selected from which a short sentence ("The rainbow is a division of white light into many beautiful colours"), was used as stimuli. From the set of 16 stimuli, six sub-sets were selected in a semi-random manner, such that each stimulus appeared in three of the series. Thus, each subject had six listening sessions and rated each stimulus three times, but in slightly different contexts.

Human stimuli for inter-subject reliability test

One of the last mentioned series, consisting of 8 stimuli, was randomly selected. All subjects rated the same stimuli.

Rating methods

Three methods of perceptual evaluation were examined, a 10 cm VAS on paper (Paper VAS), VAS presented on the computer screen (Computer VAS), and a visual sort and rate method (VSR).

Paper VAS

The stimuli were arranged in a random order and played from the computer at the subject's request. The order was randomised and different

for each listening session. The subject was required to listen to each stimulus at least three times. Ratings were collected as markings along a 10 cm VAS on an answering sheet (Figure 1). This type of playback is similar to listening to a pre-recorded tape and having the opportunity to replay the latest stimulus, a procedure commonly used in VAS tests. In this study, the computer replaced a tape recorder, so that the same hardware conditions were applied for all three methods.

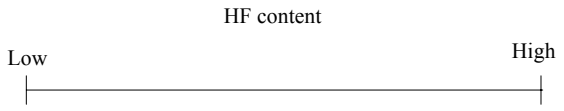

Figure 1. The Paper VAS. The subject rates the stimuli by putting a mark on a 10 cm long horizontal line. Each stimulus was presented at least three times.

Computer VAS

The computer VAS was implemented by standard scrollbars (Figure 2). The implementation on computer offers some advantages. First, the subjects may jump back and forth in the series of stimuli. Second, the computer program can automatically arrange the stimuli in a separate random order for each test session, thus minimising the risk of errors due to inter-stimuli effects. In order to achieve this with tapes, it would be necessary to prepare a separate tape for each listening session. Third, data processing becomes easier since the subjects' responses can be automatically stored in the computer, thus eliminating the risk of errors associated with manual entering of responses into the computer.

VSR

The visual sort and rate method can be regarded as an enhancement of the Computer VAS. Here, the stimuli are represented by icons in the right part of the screen and the subject can listen to the stimuli by clicking on these icons (Figure 3). The subject's first task is to sort the stimuli by vertically moving the icons on the screen so that icons of similar-sounding stimuli lie close to each other on the screen. This facilitates comparisons between stimuli, particularly for stimuli that sound similar. The second task of the subject is to rate the stimuli along the vertical VAS on the left part of the screen.

This procedure will make each stimulus an external reference for the remaining stimuli. The VSR method implies that subjects shift the reference from their internal representation to an external representation constituted by the other stimuli. In some sense this method has the same benefits as the anchoring procedure (Gerratt et al., 1993).

Listening test

Listening tests were carried out to compare the three methods described above with regard to inter- and intra-subject reliability. Six subjects, all speech pathologists with long experience of voice quality rating participated. All six subjects

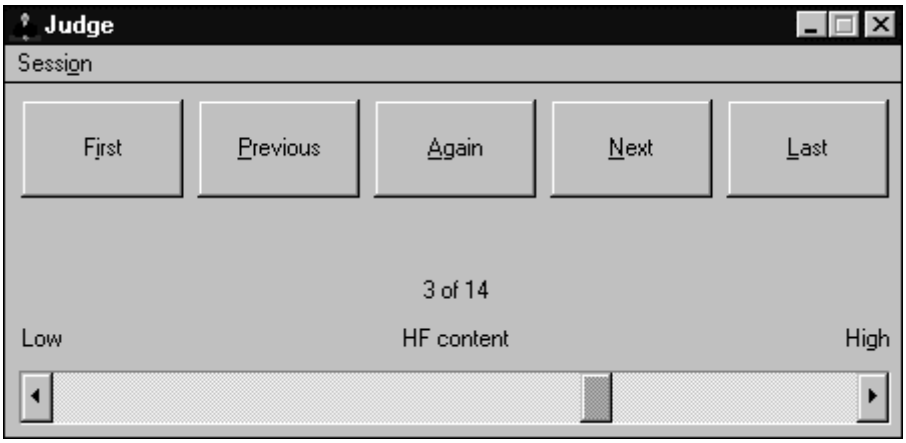

Figure 2. The computer VAS was implemented in a straightforward manner on a computer. The subject rates the stimuli by adjusting a scrollbar instead of putting a mark on a line on a paper. Subjects are free to step back and forth in the list of stimuli and adjust their responses so as to match their auditory perception optimally.

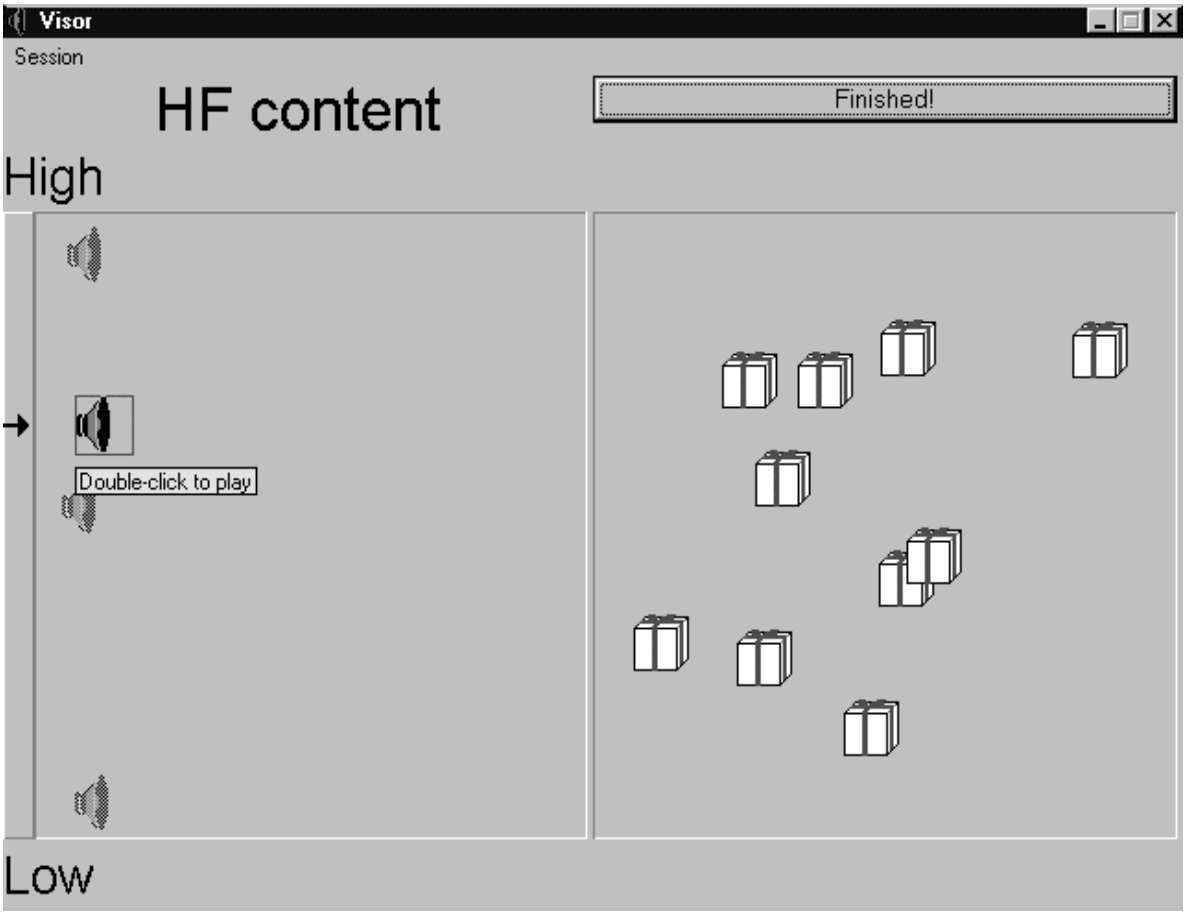

Figure 3. In the Visor program, which implements the VSR method, the subject is supposed to "open" the packages to the right and listen to the stimuli within them. Then, the task is to move them to the VAS to the left according to perceived quality. As a result, similar sounding stimuli will be located close to each other, which facilitates comparisons.

participated in the inter-subject listening tests and four of them in the intra-subject listening tests, two rating the synthetic stimuli and two rating the human stimuli.

In the tests with synthetic stimuli, subjects were asked to rate the high-frequency (HF) content of the stimuli. Stimuli that had steeper spectral tilt were expected to receive lower ratings of HF content. In the test with natural voices, the subjects were asked to rate three parameters; breathiness, hyperfunctionality, and a third, compound parameter consisting of roughness, vocal fry and gratings. These parameters were found perceptually relevant in an informal pre-test listening session.

Before each of the three tests the subjects were asked to play all stimuli in a sequence. The subjects tested the three methods in different orders. In this way, no method should suffer more than the others from effects of subject's fatigue or unfamiliarity with the stimuli.

Evaluation methods

The performance of the three different methods was evaluated in two ways. One measure concerned the relation between ratings from different sessions by means of Pearson's correlation coefficients. The second measure concerned the subjects' ability to rank order the stimuli in the expected order.

Pearson correlation coefficient

Results from different listening sessions with the same method were correlated in all possible pairs. At least three data points are required to calculate a relevant correlation, and since some combinations of sessions have only two points

in common (e.g sessions 1 and 5, human stimuli), these combinations were excluded. The correlation coefficients from these comparisons were analysed by means of a repeated measures analysis of variance (ANOVA).

Spearman correlation coefficient

The synthetically generated stimuli were expected to be rated in a sequence with monotonically increasing HF content. Therefore, the rank order of the responses was correlated against the actual rank order of the stimuli and analysed by means of the Spearman's correlation coefficient. The resulting coefficients were analysed by means of a repeated measures ANOVA. This measure was not derived from the test with the human voice, since the perceptually relevant rank order for these voices was unknown.

Results

Results are summarised in Table 2. All data analysed with ANOVA was tested with Mauchly's test of sphericity. In cases where this test resulted in significance, the degrees of freedom were adjusted with the Greenhouse-Geisser method. An alpha level of 0.05 was used for all statistical tests.

For the human stimuli, the ANOVA showed significant effects of rating method in the inter-subject test. Post-hoc test (Scheffe) revealed a

Table 2. Statistical analyses of the results. Significant results are shown in boldface. In some cases, Mauchly's test of sphericity was significant. In those cases, the degrees of freedom were adjusted with the Greenhouse-Geisser method. Such cases are marked with an asterisk.

Human stimuli, intra-subject				
		Paper VAS	Computer VAS	VSR
Avg Pearson's corr.		0.767	0.796	0.772
Effect of method		F(2,118)=0.142, p=0.867		

Human stimuli, inter-subject				
		Paper VAS	Computer VAS	VSR
Avg Pearson's corr.		**0.685**	0.758	**0.778**
Effect of method		F(2,84)=6.747, **p=0.002**		
Post Hoc	P-VAS	-		
(Scheffe)	C-VAS	P=0.056	-	
	VSR	**P=0.010**	p=0.7885	-

Synthetic stimuli, intra-subject				
		Paper VAS	Computer VAS	VSR
Avg Pearson's corr.		**0.956**	**0.939**	**0.983**
Effect of method		F(1.373,38.432)=11,984, **p<0.001***		
Post Hoc	P-VAS	-		
(Scheffe)	C-VAS	p=0.1833	-	
	VSR	**p=0.0155**	**P<0.0001**	-

Synthetic stimuli, inter-subject				
		Paper VAS	Computer VAS	VSR
Avg Pearson's corr.		**0.911**	**0.930**	**0.978**
Effect of method		F(1.418,19.849)=11.286, **p=0.001***		
Post Hoc	P-VAS	-		
(Scheffe)	C-VAS	p=0.436	-	
	VSR	**P<0.001**	**p=0.010**	-

Synthetic stimuli				
		Paper VAS	Computer VAS	VSR
Avg Spearman's corr.		**-0.914**	**-0.930**	**-0.980**
Effect of method		F(2,30)=13.611, p<0.001		
Post Hoc	P-VAS	-		
(Scheffe)	C-VAS	p=0.496	-	
	VSR	**p<0.001**	**p=0.003**	-

significant improvement of correlation coefficient for the VSR method, as compared to paper VAS.

For the synthetic stimuli, significant effects were found in both inter- and intra-subject correlation test, as well as in the rank ordering test. Post hoc tests (Scheffe) showed significantly improved correlation coefficients for the VSR as compared to the computer VAS in all cases. Post hoc tests also showed significantly improved correlation coefficients for the VSR as compared to the paper VAS in all cases.

Discussion

Implementing a listening test as a computer program entails several advantages. It saves the work of manually entering data from the forms filled in by the subjects, which is also a potential source of error. Furthermore, creating listening tapes is not needed. In order to minimise the risk of inter-stimuli influence, the order of presentation should be different for each listening session. This normally requires the assembling of many different tapes, but if implemented in a computer, randomisation of the order of presentation can be realised automatically.

The greatest advantage, though, is the increased freedom offered to the subjects to play the stimuli as many times as they wish, and in any order. This can increase the subjects' motivation, which in turn can lead to more consistent results.

The most interesting result of this investigation is the performance of the VSR method in the single parameter test. The absolute correlation coefficients for the VSR method was in the range 0.978 - 0.983, for the computer VAS within 0.930 - 0.939, and for the paper VAS within 0.911 - 0.956.

One reason for the improved reliability of the VSR method could be that it allows the subject to compare stimuli of similar ratings to each other, rather than comparing them with some internal standard. In a sense, each stimulus can be said to serve as an external reference in the VSR method. A somewhat similar technique is to include anchor stimuli, i.e. a set of reference stimuli with explicitly specified positions along the scale (Gerratt et al., 1993). One problem with anchor stimuli is that it can be hard to find appropriate stimuli for this purpose, e.g., because the perceptual value of the examined parameter is typically probably unknown. In the case of voice disorders, it is mostly hard to find voices that differ along one dimension only. With the VSR method, no special anchor stimuli are required, unless absolute references are needed. Of course, if valid anchor stimuli *are* available, they can be mixed with the test stimuli without specified positions, thus providing a standard-relative measure of the test material. Anchor stimuli could also be assigned fixed positions along the VAS in the VSR method, which may further improve the reproducibility. This seems an interesting possibility to examine in a future investigation.

The improved Spearman's correlation coefficient when using the VSR method was expected, since the subject had the opportunity to directly listen to the difference between two closely rated stimuli and adjust the responses accordingly. This could be expected to improve the rank ordering ability. The VSR method has similarities with a method where stimuli are arranged in all possible pairs and the subject rates the order within the pairs. The problem with that method is the large number of pairs required ($\sim N^2$, where N is the number of stimuli). With the VSR method the critical pairs are automatically generated, as the subject moves the icons along the VAS.

Informal post-test interviews with the subjects revealed an agreement on the benefits of the VSR method. However, some of the subjects were uncomfortable with the instruction to use the entire scale. For example, in cases where none of the stimuli had a large amount of breathiness, there was a reluctance to put the most breathy stimulus at the high extreme of the VAS. This was probably due to the fact that all subjects had long experience with the paper VAS method with the instruction to rate absolute values of the perceived entities. Thus, they can be assumed to have developed reasonably well-established internal references. This suggests that greater relative performance improvement with the VSR method could be expected for untrained listeners, or with listeners that are trained with the VSR method.

The method implemented by the standard VAS could also be called "Visual Rate" since the subject is asked to visually rate each individual stimulus along the VAS. The VSR method complements this method with a sorting task. After sorting the stimuli, rating should become easier.

A risk with asking subjects to sort *and* subsequently rate the stimuli in the VSR method may occur in cases when the stimuli are not

equidistantly distributed along the perceptual scale. For example, if many stimuli should be rated low perceptually, and only few should be rated high, the subject should ideally place many stimuli at the low end of the VAS. If the instructions are not clear enough, the subject might tend to position the stimuli equally spaced along the VAS, which would lead to a distortion of the response curve. Thus, appropriate instruction of the subjects is important. For instance, it might be worthwhile to remind the subjects, just before finishing the test, that the stimuli should be not only sorted but also rated. Also, introducing a few stimuli duplicates should reduce this risk.

Ratings collected from listening tests can be considered as being either relative or absolute or a combination of both. Absolute rating requires skilled listeners, trained to rate stimuli consistently, regardless of the test situation and other stimuli. Providing anchor stimuli might help the listeners in this task, but as soon as more than one stimulus is included in the test, there is a risk of influence from the other stimuli. The VSR method, on the other hand, can largely be seen as relative. In the typical test setting, the subjects are instructed to utilize the entire VAS, such that at least one stimulus is given the maximum rating and at least one is given the minimum rating. However, if the listeners have well-established internal references, these references may affect the results in spite of the instructions. For example, in the present test there was no extremely breathy stimulus, which may have caused some listeners to choose rather low ratings of this parameter for all stimuli. In this sense, the VSR method is not entirely relative.

The design of the present study deliberately did not favour the VSR method; a number of factors probably were disadvantageous. First, in the intra-subject tests, different sets of stimuli were used in the six different sessions for each subject. This should be a disadvantage for relative methods such as the VSR, since the context in which each stimulus was played differed between sessions. Second, all the listeners had a long experience of rating voice quality and can be assumed to have established reasonably stable internal references. Hence they could be expected to perform better when using the absolute methods. Third, the listeners had a long experience of rating voices by means of visual analogue scales, but only limited experience from VSR tests. In spite of these design disadvantages, the VSR method performed significantly better than the other two methods in the single-parameter test and never performed significantly poorer than the other methods in the multi-parameter test. In fact, in one case it performed significantly better than the P-VAS.

Presently, the VSR method has been used at KTH in Stockholm, Sweden, (House, 2000) as well as at the Huddinge hospital for evaluation of treatment effects during voice therapy (Holmberg et al., 2000). However, the method is currently being used in other investigations. Experiences from these investigations will shed more lights on the potentials of the method.

Conclusions

The Visual Sort and Rate method, VSR, can be used to improve the performance of subjects in listening tests. Experiments show that both inter- and intra-subject correlation was improved in single-parameter tests when the VSR method was used instead of the more commonly used Visual Analogue Scale, VAS. Likewise, in a multi-parameter test significantly better results were observed for the VSR method than for VAS on paper. The benefits of the VSR method applied in single-parameter tests may be due to the fact that it supplies the subjects with external references, even though no anchor stimuli are used.

Acknowledgement

This work was supported by research grants from the Bank of Sweden Tercentenary Foundation. I would also like to thank Johan Sundberg for editorial assistance, Joakim Westerlund for the statistical analyses, and Jan Gauffin, Britta Hammarberg and Stellan Hertegård for their valuable input.

References

Berliner JE, Durlach NI, Braida LD (1978). Intensity perception. IX. Effect of a fixed standard on resolution in identification. *Journal of the Acoustic Society of America* 64/2: 687-689.

Gauffin J, Granqvist S, Hammarberg B, Hertegård S, Håkansson A (1995). Irregularities in the voice, some perceptual experiments using synthetic voices. *Proc of XIII Intl Congress of Phonetic Sciences (ICPhS95),* Stockholm, 2: 242-245.

Gerratt B, Kreimann J, Antonanzas-Barroso N, Berke G (1993). Comparing internal and external standards in voice quality judgements. *Journal of Speech and Hearing Research* 36: 14-20.

Granqvist S (1996). Addsynt, an additive voice synthesiser for PC. *TMH-QPSR, KTH,* 4: 57-60.

Hammarberg B (1986). Perceptual and acoustic analysis of dysphonia. Stockholm: *Dissertation*, Dept of Logopedics and phoniatrics, Karolinska institute, Stockholm.

Holmberg EB, Hillman RE, Hammarberg B, Södersten M, Doyle P (2000). Efficacy of a behaviorally-based voice therapy protocol for vocal nodules *(Accepted 2000 for Journal of Voice).*

House D (2000). Rise alignment in the perception of focal accent and pitch in Swedish. *Proc. Fonetik 2000, Skövde, Sweden*, 73-76.

Isshiki N, Okamura H, Tanabe M, Morimoto M (1969). Differential diagnosis of hoarseness. *Folia Phoniatrica* 21: 9-19.

Kreimann J, Gerrat B (1993). Perceptual evaluation of voice quality: Review, Tutorial, and a Framework for Future Research. *Journal of Speech and Hearing Research* 36: 21-40.

McAllister A, Sederholm E, Ternström S, Sundberg J (1995). Perturbation and hoarseness: A pilot study of six children's voices. *Journal of Voice* 10/3: 252-261.

McAllister A, Sundberg J, Hibi S (1996). Acoustic measurements and perceptual evaluation of hoarseness in children's voices. *TMH-QPSR, KTH*, 4: 15-26.

Rabinov R, Kreiman J, Gerratt B, Bielamowicz S (1995). Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter. *Journal of Speech and Hearing Research* 38: 26-32.

Sederholm E, McAllister A, Sundberg J, Dalkvist J (1993). Perceptual analysis of child hoarseness using continuos scales. *Scandinavian Journal of Logopedics and Phoniatrics* 18: 73-82.

Wewers ME, Lowe NK (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health* 13: 227-236.

Titze I (1994). *Definitions and Nomenclature Related to Voice Quality. Vocal Fold Physiology, Voice Quality Control* San Diego: Singular Publishing Group; 335-342.