Effect of experience levels on voice quality ratings

Christel de Bruijn, University of Central England
Sandra Whiteside, University of Sheffield

1 Introduction Measuring interrater- and intrarater reliability is an important way of validating perceptual judgments of voice quality. One of the fields of research for which this validation is particularly important, is that which attempts to establish correlations between perceptual measures of voice on the one hand, and physiologic or acoustic parameters on the other. If reliability of a perceptual label or judgement is low, interpretation of a correlation with a perceptual parameter is complicated.

It is often assumed that when listeners who are more experienced, are making judgements about voice quality, they will achieve a higher rating. The literature, however, provides a number of examples where different groups of listeners with different levels of experience in judging voice quality, obtain similar levels of interrater reliability (Rabinov et al., 1995; Yamaguchi et al., 2003; Bassich & Ludlow, 1986). Comparison of reliability across different studies, however, is complicated due to the different methodologies employed, e.g. the use of different rating instruments, parameters, anchoring and the degree of pathology of the voice samples.

The current paper presents a comparison of voice ratings from speech and language therapists (SLTs) specialised in voice, with those of final year speech and language therapy (SLT) students, using the same methodology in each group. The comparison was carried out as part of a wider study on voice fatigue and use of speech recognition software (De Bruijn, 2007).

2 Methodology Conversation fragments were collected from 25 speakers before and after carrying out a 2-hour dictation task using speech recognition software. The speaker group consisted of 14 men and 11 women with no reported voice problems. The age ranged from 19 to 59. Speakers were categorised as having either a high or a low daily vocal load, and using either a discrete or continuous speech recognition system, the cross-over of which resulted in 4 different groups of speakers. For further details about the motivation and methodology for this (wider) study, the reader is referred to De Bruijn (2007).

The voice recordings were evaluated by a panel of 11 listeners, which included 5 SLTs and 6 final year SLT students. All therapists were specialised in voice and had a minimum of 2 years experience in their specialisation. The choice of perceptual parameters (table 1) was based on the symptoms reported in studies on use of speech recognition software and voice quality, and on other studies on vocal fatigue. Additional parameters were chosen from the GRBAS scale. The parameters were evaluated either on a 5-point scale (used for parameters which may or may not be present in a voice sample, such as breathiness) or a 9-point scale (for bipolar parameters, i.e. parameters which are always present in a voice, such as pitch).

Parameter	Scale range	Parameter	Scale range
Breathiness	0-4	Asthenicity/ voice weakness	0-4
Roughness	0-4	Lacking sonority	0-4
Creak/ glottal fry	0-4	Overall instability	0-4
Strain/ vocal effort	0-4	Overall vocal deviation	0-4
Hard glottal attack	0-4	Pitch	0-8
Monotony	0-4	Loudness range	0-8
Audible breath	0-4	Hypo/ hyperfunctionality	0-8

Table 1: Perceptual parameters and rating scales.

All ratings were carried out relative to a selection of voice samples that listeners were trained on, i.e. ratings were normalised to the range of qualities in the present experiment. For example, a rating of 0 on the breathiness scale means “not breathy at all”, where 4 means “most breathy compared to the training set”. Listeners were encouraged to use the full range of the scale. This approach was chosen because changes in voice quality were expected to be fairly small. Intrarater reliability was calculated (Pearson’s r) for each listener by duplicating 30% of the voice samples, for every parameter. If a correlation below .60 was obtained, the ratings for that parameter and listener were excluded from further analysis.

Ratings from the 2 groups were compared for conversation fragments recorded after the dictation task, as follows. First, a repeated-measures ANOVA was carried out with profession (therapist or student) and perceptual parameter as within-subject factors, followed by paired samples t-tests and correlations in order to locate any differences and relationships. Finally, intracorrelations between all perceptual parameters were calculated for students and therapists separately.

3 Results Assumptions of normal distribution and sphericity were checked with the Kolmogorov-Smirnov and Mauchly’s tests. In cases of sphericity violation, the Greenhouse-Geisser correction was applied. The significant main effect for profession indicated that there was a significant difference between the ratings from SLTs and students, but the interaction with the factor parameter indicates that this difference is parameter specific (table 2).

Source	Sphericity correction	Type III Sum of Squares	df	Mean Square	F	Sig.
Profession	Sphericity assumed	3.728	1	3.728	8.045	0.010
parameter	Greenhouse-Geisser	759.919	2.647	248.684	61.007	0.000
profes * parm	Greenhouse-Geisser	12.302	5.182	2.036	14.553	0.000

Table 2: ANOVA results

Paired-samples t-tests (table 3) showed significant differences in ratings between students and therapists for the parameters creak, overall instability, overall vocal deviation and hypo/hyperfunctionality. Therapist ratings were higher for hypo/hyperfunctionality, whereas student ratings were higher for the other parameters. However, this difference is not necessarily an indication that students and therapists rate in different ways. Table 3 shows that for all parameters yielding a significant difference, high correlations were also found between student and SLT ratings. This suggests that students and SLTs are rating in a similar fashion, but have different baselines.

	Mean difference (therapist-student)	t	df	Sig. (2-tailed)	Correlation	Sig.(2-tailed)
creak	-.5174	-4.479	22	.000	.753	.000
overall instability	-.3884	-4.936	22	.000	.752	.000
overall vocal deviation	-.4029	-4.735	22	.000	.666	.001
hypo-hyperfunctionality	.5594	4.126	22	.000	.749	.000

Table 3: significant results of paired-samples t-tests and correlations

In order to get further insight into the rating behaviours of students and therapists, intra-correlations were calculated between all perceptual parameters, e.g. correlations were calculated between breathiness and roughness ratings, breathiness and creak ratings etc. for students and therapists separately. The purpose of this was to find out if students and therapists display similar patterns in what they perceive to be distinct or non-distinct parameters of voice quality. A high correlation between 2 parameters indicates that for that listener (or group of listeners in this case) these 2 parameters basically capture the same information, and are therefore not distinct (De Krom, 1994). Intra-correlation results are shown in table 4. Correlations of .66 and higher are highlighted (in grey).

	breath	rough	creak	Strain	gl. attack	Monotony	Aud breath	asthenicity	Lack sonor	instability	deviation	pitch	Loud range	Hyper /hypo
B		.086	-.167	-.129	-.625	.010	.503	.730	.294	.338	.294	.556	-.220	-.280
R	-.195		.721	.113	.108	.166	-.086	.291	-.128	.593	.503	-.203	-.378	-.477
C	-.249	.669		.282	.329	.231	-.260	.041	-.269	.549	.638	-.354	-.223	-.351
S	-.181	.112	-.036		.481	-.179	-.486	.030	-.181	.292	.364	-.158	.019	.439
GA	-.565	-.119	-.199	.500		-.059	-.210	-.469	-.274	-.026	.066	-.457	.189	.241
M	.268	.381	.271	-.321	-.403		-.005	.326	.632	.501	.085	-.375	-.577	-.382
AB	.732	-.279	-.233	-.038	-.473	.015		.259	.273	.060	-.054	.488	-.010	-.370
A	.804	-.284	-.258	.048	-.374	.190	.520		.579	.696	.563	.322	-.393	-.346
LS	.547	.351	.219	.087	-.393	.600	.313	.676		.269	.030	-.037	-.431	-.194
I	.431	-.049	.106	.191	-.241	-.052	.207	.671	.593		.756	.022	-.496	-.365
D	.408	-.067	.116	.296	-.195	-.274	.277	.575	.383	.816		.034	-.250	-.156
P	.123	-.763	-.539	.365	.172	-.557	.291	.327	-.204	.214	.264		.451	-.253
LR	-.646	-.287	-.125	.295	.577	-.669	-.333	-.619	-.719	-.339	-.217	.368		.041
H	-.589	.199	.010	.627	.553	-.272	-.359	-.499	-.323	-.285	-.306	.116	.589	

Table 4: intracorrelations between parameters for therapists and students.
Values for therapists are listed above the diagonal, and for students below the diagonal.

If we consider correlations between 0.00 and 0.33 to be weak, between 0.33 and 0.66 to be moderate, and 0.66 and 1.00 to be strong, than the therapists displayed 4 strong intracorrelations, namely for the parameter pairs creak-roughness, asthenicity-breathiness, instability-asthenicity and deviation-instability. This same pattern of strong intracorrelations was found for the students. In addition, the students perceived the parameters audible breath and breathiness to be strongly correlated, as well as the parameters asthenicity and lack of sonority thus considerably overlapping. The therapists showed moderate correlations for the latter two pairs.

To summarise, the results show that there were significant differences between student and therapist ratings of voice quality, in particular creak, instability, overall deviation and hypo/hyperfunctionality. Students scored the voice samples higher on creak, instability and overall deviation, but lower on hypo/hyperfunctionality than the therapists. However, the analysis also showed high correlations between the student and therapist ratings, suggesting that students and therapists employ similar rating strategies, but have different baselines. In addition, intracorrelations calculated between the different parameters revealed that students and therapists appear to have similar concepts of the perceptual parameters.

Not many studies have been published about the potential differences in perceptual strategies of listeners with varying amounts of experience in voice evaluation. The findings from the current study correspond to results reported by Murry et al. (1977) who carried out a multidimensional scaling study to determine the perceptual attributes of a group of non-normal voices. They used graduate speech pathology students and experienced clinicians as listeners. When the weightings of the five resulting dimensions were plotted, there were no clear demarcations between the weightings of the students and the clinicians. Therefore, all listeners used the same perceptual categories to similar degrees in their ratings of voice similarity. The findings of the current study, as well as those of Murry et al. (1977), however, appear to be in contrast with those of Kreiman et al. (1990). A counter argument to their hypothesis can be found in De Bruijn (2007).

To conclude then, it is proposed in this study that perceptual strategies between more and less experienced listeners are not different, but rather that these listeners adopt different baselines during perceptual tasks.

4 References

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

Yamaguchi, H., Shrivastav, R., Andrews, M. L. & Niimi, S. (2003) A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale*. Folia Phoniatrica et Logopaedica,* vol. 55, 147-157.

Bassich, C. J. & Ludlow, C. L. (1986) The use of perceptual methods by new clinicians for assessing voice quality. *Journal of Speech and Hearing Disorders,* vol. 51, 125-133.

De Bruijn, C. (2007) *Voice quality after dictation to speech recognition software: a perceptual and acoustic study*. University of Sheffield, UK: Doctoral dissertation.

Verdonck-de Leeuw, I. M. (1998) *Voice characteristics following radiotherapy: the development of a protocol.* University of Amsterdam, The Netherlands: Doctoral dissertation.

De Krom, G. (1994) *Acoustic correlates of breathiness and roughness: experiments on voice quality.* OTS, Research Institute for Language and Speech, ISBN 90-5434-021-5, Utrecht University, The Netherlands: Published doctoral dissertation.

Murry, T., Singh, S. & Sargent, M. (1977) Multidimensional classification of abnormal voice qualities. *Journal of the Acoustical Society of America,* vol. 61 (6), pp. 1630 – 1635.

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990) Listener experience and perception of voice quality. *Journal of Speech and Hearing Research,* vol. 33, pp. 103-115.