Audio Engineering Society

Convention Paper 7176

Presented at the 123rd Convention
2007 October 5–8 New York, NY, USA

On the Use of Graphic Scales in Modern Listening Tests

Sławomir Zieliński, Peter Brooks*, and Francis Rumsey

Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, UK
s.zielinski@surrey.ac.uk, mu42pb@surrey.ac.uk, f.rumsey@surrey.ac.uk

ABSTRACT

This paper provides a basis for discussion of the perception and use of graphic scales in modern listening tests. According to the literature, the distances between the adjacent verbal descriptors used in typical graphic scales are often perceptually unequal. This implies that the scales are perceptually non-linear and the ITU-R Quality Scale is shown to be particularly non-linear in this respect. In order to quantify the degree of violation of linearity in listening tests, the evaluative use of graphic scales was studied in three listening tests. Contrary to expectation, the results showed that the listeners use the scales almost linearly. This may indicate that the listeners ignore the meaning of the descriptors and use the scales without reference to the labels.

1. INTRODUCTION

There are two types of graphic scales that are commonly used in modern audio quality listening tests. The first one, known as the continuous impairment scale, is presented in Figure 1 and is typically used for evaluation of impairments exhibited by a processed sound compared with an unimpaired reference. As it can be seen, it incorporates five adjectives (labels) describing the following levels of impairments: "Imperceptible", "Perceptible, but not annoying", "Slightly annoying", "Annoying," and "Very annoying". It is important to notice that the labels are distributed along the scale at equal geometrical distances. This type of a graphic scale is recommended by the ITU-R BS.1116 standard [1].

Another example of a graphic scale that is in common use is the scale presented in Figure 2. This scale is recommended by the ITU-R BS.1534 (MUSHRA) standard [2]. It is often referred to as a continuous quality scale. It differs from the previously discussed impairment scale in two ways. Firstly, instead of the impairment labels it contains the labels describing five different levels of quality: "Excellent", "Good", "Fair", "Poor" and "Bad". Secondly, instead of defining discrete points, the labels are used to define five intervals on the scale. For example, the term "Excellent" is used to define the top 20% range of the scale. It is important to notice that, similarly to the impairment scale, the labels are spread uniformly along the scale.

* Currently employed by Macquarie Bank Limited, Level 3 Moor House, 120 London Wall, EC2Y 5ET, London, UK

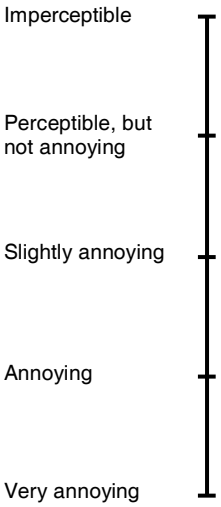

Figure 1 Example of a graphic impairment scale.

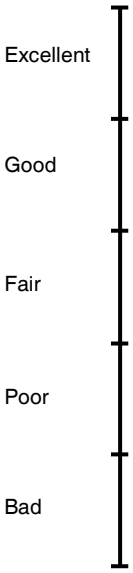

Figure 2 Example of a graphic quality scale [2].

The fact that in the case of both discussed scales the labels are distributed uniformly with equal geometrical distances between the adjacent labels is probably the reason that some researchers implicitly assume that the scales are perceptually linear. However, this may not be the case. In fact there is some evidence, which will be presented below, indicating that the scales may be non-linear in a perceptual sense. A potential departure from linearity may lead to problems with the data analysis and interpretation of the results from listening tests. For example, as a result of the use of the non-linear scale the scores may be biased and hence their absolute values may be distorted. Consequently, it might be impossible to assess the magnitude of the differences between the stimuli and the scales should only be used to rank order the stimuli. Moreover, without knowing the exact nature and the degree of the non-linearity of the scale, the researchers may not be able to make any perceptual inferences from the results. Finally, the researchers may not be able to use parametric methods of statistical analysis as the distorted data may not meet the assumptions required by these techniques. Therefore, in order to avoid the above problems, it is important to check whether the two most commonly used graphic scales are perceptually linear.

The paper is organized as follows. The next section will provide a summary of the research in this area. It will be shown the two scales in question potentially exhibit severe departure from linearity. However, there is also some evidence contradicting the above finding, which will be discussed in Section 3. The experiment designed to investigate this issue in more detail will be described in Section 4 and the results presented in Section 5. The last section includes summary and conclusions.

2. POTENTIAL DEPARTURE FROM LINEARITY

The issue of a potential departure from linearity of the impairment and quality scales was of concern to many researchers, predominantly those working in the area of picture and multimedia quality assessment. Although in those times they were primarily concerned with the potential non-linearity of the (discrete) categorical scales, the results of their studies also apply to the (continuous) graphic scales discussed in this paper, since the ITU graphic scales involve the same labels as the ITU categorical scales.

One of the first reports indicating that the ITU scales may exhibit non-linearity was published in 1986 by Jones and McManus [3]. In their experiment they asked the subjects to graphically scale a group of adjectives, including those used in the quality scale. The task given to the participants was to indicate the meaning of every adjective by putting a mark on a vertical 18 cm scale. The participants were instructed to do it with respect to the two ends of the scale defined as the “best

imaginable" and "worst imaginable" respectively. According to the results obtained by Jones and McManus, the labels used in the quality scale are not equidistantly spaced along the scale. Consequently, it could be concluded that the scale with the equidistantly distributed quality labels does not have equal-interval properties, and hence it is rendered non-linear.

The observed effect of the uneven spacing between the adjectives was caused by the non-uniform semantic differences between the adjacent labels. For example, according to the results obtained by Jones and McManus in American English the terms "Poor" and "Bad" are semantically similar. By contrast, there is a large semantic difference between the terms "Fair" and "Poor". Therefore, the semantic distance between the adjectives "Fair" and "Poor" is much bigger than the difference between the terms "Poor" and "Bad", which may give rise to a non-linear effect in the use of the scale if the terms are spaced equidistantly. In one area of the USA participants scaled the term "Poor" slightly lower than the term "Bad" which implies that, due to the linguistic variations in different regions of the country, the quality scale may not only be non-linear but may not even be monotonic.

The quality and impairment scales discussed above are used internationally with translated versions of the labels. This gives rise to a question about whether a similar non-linear effect could be observed in other languages. Jones and McManus repeated their experiment in Italy and discovered that the effect of irregular spacing of the Italian equivalents of the studied adjectives was even more pronounced than it was the case in the USA [3]. The semantic difference between the terms "Discreto" and "Mediocre" was approximately six times bigger than the difference between the adjectives "Mediocre" and "Cattivo" (see Figure 3).

The above results prompted other researchers to undertake similar studies in their countries. In 1990 the International Telecommunication Union (formerly CCIR) published a report with the results of the experiments conducted in Germany and France [4]. It was shown that the semantic differences between the French equivalents of the studied terms were also non-uniform. However, in contrast to the previous results, it was found that the German equivalents of the quality or of the impairment adjectives were scaled in an almost uniform way (see Figure 3). A similar result of a uniform distribution was also found by Narita in 1993 in the study undertaken in Japan [5]. However, the results of the similar studies undertaken in the Swedish and the Dutch languages revealed a non-uniform distribution of the adjectives [6], [7]. More recently, a similar experiment was undertaken by Watson [8] in England using a group of British English speakers revealing even more uneven distribution distortions compared to that observed by Jones and McManus in the USA. The above results showed that the magnitude of the semantic differences between the adjacent labels used in the quality and the impairment scales is language specific.

Figure 3 illustrates the combined results of a semantic scaling of the terms that are used in the ITU quality scale. As can be seen, the amount of the non-uniformity in the distribution of the adjectives varies across different languages, being the most severe for the British English language (labeled as "UK English") and the least severe for the German language. It is interesting to see that in most of the languages represented in the figure, the semantic difference between the terms "Poor" and "Bad", or their translated equivalents, is much smaller than the differences between the terms "Fair" and "Poor".

Figure 3 does not show any results for the ITU impairment labels, however research has shown that the adjectives used in the ITU impairment scale are also distributed in a non-uniform way and that this effect is language specific [4].

Considering the results of the studies undertaken so far, some researchers reached the conclusion that the standard quality and impairment scales are not linear, perhaps with the exception of the German and Japanese equivalents of these scales. For example, Virtanen et al. [6] questioned the equal-interval property of the ITU quality scale and concluded that it shows "profound non-linearity". Watson also criticized the ITU (5-category) quality scale and stated that it is invalid and should not be used as an equal-interval scale. She also expressed her concern about the fact that this scale is so popular: "That it continues to be used all over the world by telecommunications companies, in face of evidence that it is not a reliable method, is alarming at best" [8]. She also argued that in order to circumvent the problem of the non-linearity, the labels on the scale should be removed. For example, an unlabelled scale consisting of a vertical line 20 cm long with no labels other than a "+" sign at the top and "-" at the bottom could be used as an alternative to the ITU continuous quality scale. In

fact, her proposal is very similar to the "Note 1" in the ITU-R BS.1116 Recommendation. According to this note, the use of predefined anchor points (labels) may introduce bias. As an alternative, it is recommended to use "the number scales without description of anchor points" but only with the indication of the intended orientation of the scale [1]. A similar solution is proposed in the ITU-R 1082 Report [4] and the ITU-T P.910 Recommendation [9]. These documents recommend using a graphic scale with only two labels at its extremes. Moreover, a scale with the labels at the end points and no labels in between was also recommended by Guski [10]. Despite of these recommendations, many researchers still continue to employ scales with labels. A possible reason for this conservative attitude is that label-free scales had not been studied in depth yet and, although potentially promising, their suitability for the evaluation of audio quality still needs to be validated. Consequently, some researchers may have justified concerns about the validity and reliability of this new approach.

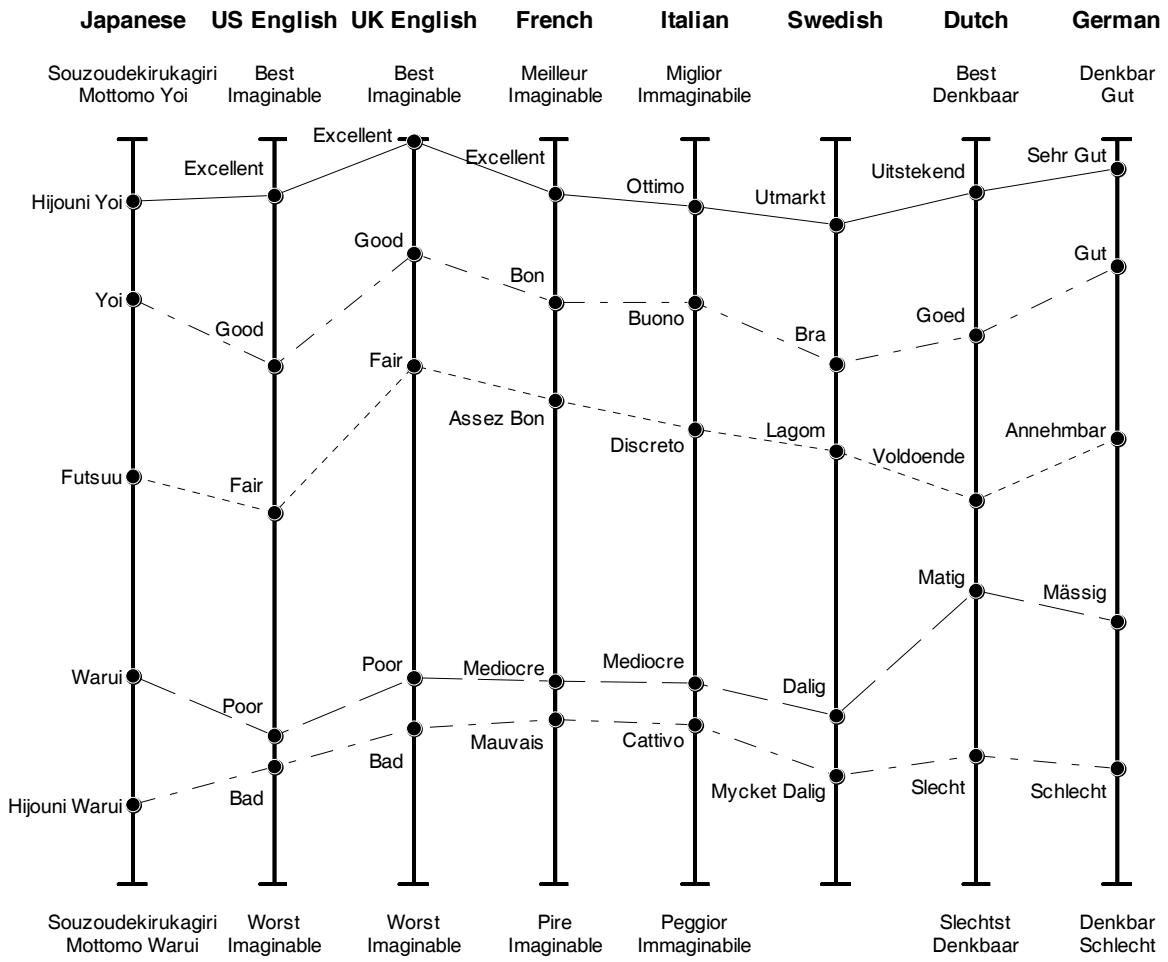

Figure 3 Combined results of scaling of labels used in quality evaluation. Data taken from [3]-[8] (see text for details).

3. CONTRADICTORY FINDING

As concluded above, scales without labels could be used instead of labeled scales if one wanted to avoid a problem with non-linearity of a scale. In order to validate this approach, Watson conducted an experiment where a group of listeners were asked to assess quality of impaired speech recordings using either a 5-category labeled quality scale or a label-free graphic scale ranging from 0 to 100. She found that the results obtained using the two methods followed the same trend and also that the listeners used the label-free scale in a consistent manner. These outcomes confirmed that the label-free scale can be used as a replacement for the labeled scale.

Although Watson stated that the results obtained using both scales "followed the same trend", she did not present any detailed information about the differences in the results obtained using the labeled and the unlabeled scales. This can be considered as a significant omission in the data analysis since the magnitude of the differences between these results could provide information about the degree of non-linearity exhibited by the 5-category quality scale. Therefore, in order to investigate this issue further, the data obtained by Watson was reanalyzed by the authors of this paper using the following procedure. The raw data elicited using the label-free scale, originally presented in Appendix G in [8], was averaged across all the listeners and trials. Since the authors of this paper did not have direct access to the raw data obtained using the labeled scale, it was decided to graphically extract the results shown at the two bottom graphs of Fig. 22 in [8]. The error of the graphical extraction of the data was equal to approximately 3%. Then, the extracted data was averaged across the trials. Similarly to the previous case, the extracted scores represented the values averaged across trials and listeners. Finally, the averaged scores obtained for both scales were plotted against each other using the scatter plot presented in Figure 4. The dashed line shows the result of the linear regression fitted to the experimental data.

As can be seen, most of the scores are scattered along the regression line, indicating a strong linear relationship between the scores obtained using the labeled and the unlabeled scales. In view of the fact that the labeled scale could exhibit a "profound non-linearity" [6], the above result is intriguing as it contradicts the conclusions reached by the researchers investigating this issue so far. This result prompted the authors to undertake further experimental work in this area, which will be described in the next section.

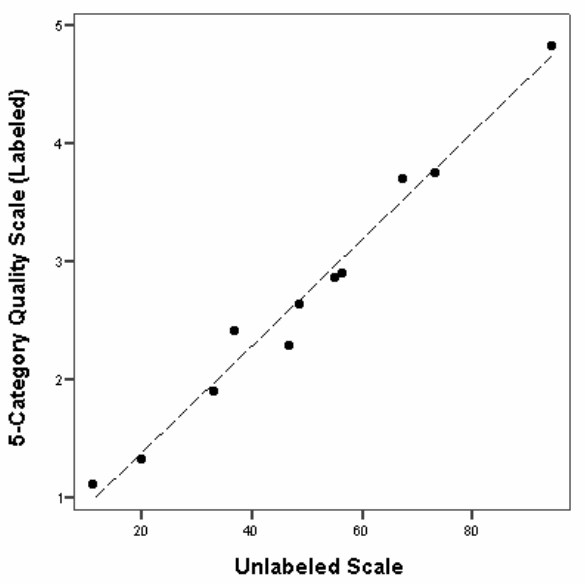

Figure 4 Data extracted from [8]. The dashed line represents the linear regression model fitted to the data.

4. EXPERIMENT

The research questions were as follows:

- What is the degree of the non-linear effect exhibited by the labeled ITU quality and impairment scales in the context of audio quality assessment?
- Are there any differences in the way the listeners use labeled and label-free scales for audio quality evaluation (e.g. in terms of the span of the scores)?

In order to quantify the degree of the perceptual non-linearity exhibited by the ITU quality and impairment scales, three separate listening tests were undertaken. All listening tests were designed according to the MUSHRA recommendation [2]. The same seven audio stimuli were used in all listening tests. All the stimuli were presented to the listeners ten times in order to increase the statistical sensitivity of the test and in order to be able to monitor listeners' consistency. The only difference between the listening tests was the type of the graphic scale used. In the first listening test, the ITU

quality scale was used as recommended in the MUSHRA standard, in the second listening test the ITU impairment scale was used, whereas in the third listening test a label-free scale was implemented. The last case of the label-free scale can be considered as the control condition since it was assumed that the scale is perceptually linear due to the lack of any labels.

The listeners were asked to evaluate "basic audio quality" defined as a single, global attribute describing any and all differences between the reference and the evaluated recordings. In all three listening tests the participants were instructed to assess the hidden reference using the top value of the scale.

In the case of the listening tests involving the label-free scale, the listeners were free to use the scale in their own way. However, as mentioned above, they were instructed to judge the hidden reference using the top end of the scale. In this way the polarization of the scale was defined. No instruction was given with regard to the bottom end of the scale.

The listening tests were undertaken in the control room of Studio 3 at the Institute of Sound Recording, University of Surrey. The acoustical properties of this room were similar to those recommended in [1].

In order to avoid a transfer bias, different groups of listeners were recruited for each listening test. Forty-five listeners were initially invited to participate in the three listening tests to give a total of fifteen listeners for each test. They were all undergraduate students of the Tonmeister Course (Music and Sound Recording) at the University of Surrey. Care was taken in order to have a balanced proportion of first, second and final year students in all the three listening tests. Unfortunately, it was only possible for thirteen subjects to participate in the label-free test, due to the unavailability of two listeners.

4.1. Stimuli

The same seven stimuli were used in all three listening tests. They consisted of the original 2-channel stereo recording and six low-pass filtered versions. The original excerpt chosen was a looped riff section from a popular song which was consistent and homogenous in nature with a broad spectrum.

A low-pass filter was required to produce the MUSHRA 3.5 kHz anchor and was also chosen to provide test stimuli with varying degrees of basic audio quality degradation. It was decided to use an FIR filter with a slope of approximately 390 dB per octave. The selected low-pass filter complied with the MUSHRA standard [2].

Seven target quality scores, evenly spread across the quality scale were selected for the purpose of this study (see Table 1). The Multichannel Quality Advisor [11] was used to predict the cut-off frequency required to represent a certain quality score. Although the Quality Advisor was originally developed to predict the quality of the low-pass filtered multichannel audio signals, it was assumed that it would provide sufficiently accurate results for the 2-channel stereo recording used in this study. This assumption was verified by means of a pilot listening test involving six listeners. Moreover, this assumption was also confirmed in a separate test after the experiment whose results are presented in Appendix. The second column of the table shows the cut-off frequencies predicted by the Multichannel Quality Advisor. These cut-off frequencies were used to filter the original recording. The presentation of the stimuli was randomized for every listener in order to counter any learning effects.

Table 1 Chosen quality scores and their predicted low-pass cut-off frequencies

Target Quality Score	Predicted Cut-Off Frequency [kHz]
100	20 (Reference)
86	14.8
72	13.4
58	11.6
44	8.8
30	5.9
18	3.5 (Anchor)

4.2. User Interfaces

As mentioned above, all three listening tests were designed according to the MUSHRA recommendation and the only difference between them was the type of the scale used in the tests. The user interface used in the first test is presented in Figure 5. As it can be seen, it employed the original ITU quality scale recommended in the MUSHRA standard. The user interfaces used in

the second and the third tests are presented in Figures 6 and 7 respectively. The interface employed in the second listening test incorporated the ITU impairment scale. As illustrated in Figure 7, in the third listening test it was decided to use an interface with a label-free scale in order to remove any non-linear effects caused by the labels.

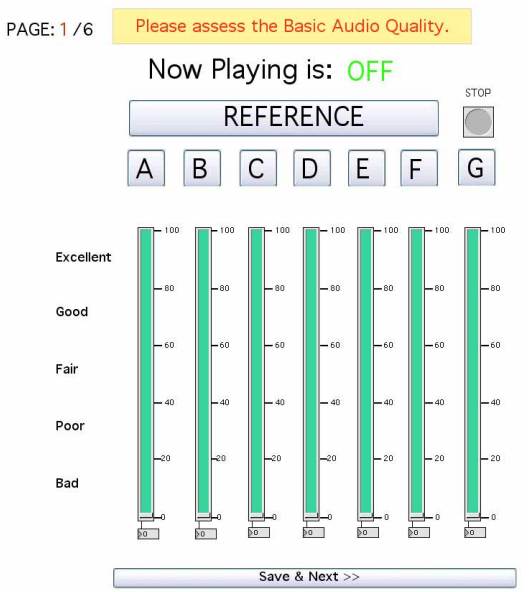

Figure 5 User interface employing the quality scale.

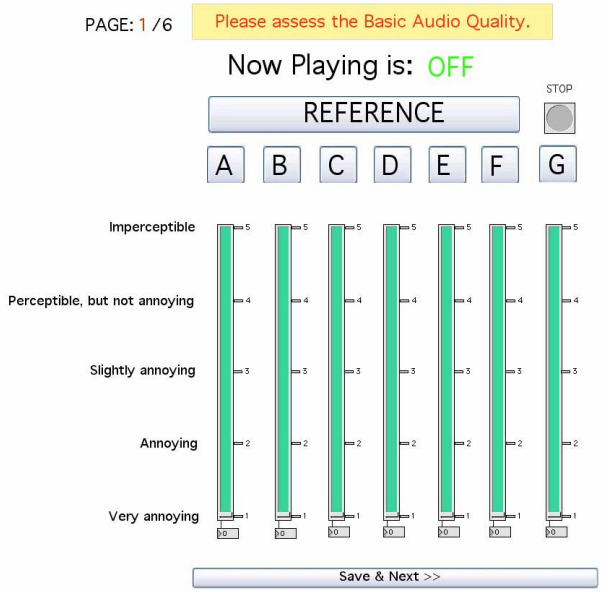

Figure 6 User interface employing the impairment scale.

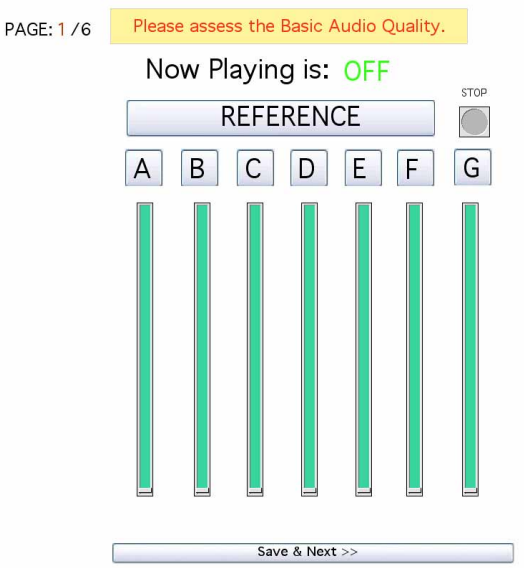

Figure 7 User interface with the label-free scale.

More details about the experimental design can be found in [12].

5. RESULTS

The obtained raw data was examined in order to check whether the listeners could reliability identify the hidden reference and whether they gave consistent grades across the stimuli. It was found that the inconsistency rms error ranged between 2% for the most consistent listener and 12% for the least consistent listener. This degree of the listeners' inconsistency was considered acceptable for the purpose of this study. The majority of the listeners managed to reliably identify the hidden reference. Only one listener made a single mistake. Since this listener made this error only once out of seven trials it was decided not to remove the data from this listener. Consequently, the data was accepted as sufficiently consistent and reliable and there was no need for any post-screening of data.

5.1. Correlation Analysis

In order to check the similarity of the results obtained using different scales it was decided to calculate the correlation coefficients between the scores. Prior to this analysis, the scores were averaged across the listeners and trials. It was found that the scores obtained from the all three listening tests were highly correlated with respect to each other. The correlation coefficient was in

all cases equal to 0.999 and was statistically significant at $p < 0.001$ level. The details of the correlation analysis are presented in Table 2.

Table 2 Results of the correlation analysis.

		Label-Free Scale	Quality Scale	Impairment Scale
Label-Free Scale	Pearson Correlation	1	.999	.999
	Sig. (2-tailed)		.000	.000
	N	7	7	7
Quality Scale	Pearson Correlation	.999	1	.999
	Sig. (2-tailed)	.000		.000
	N	7	7	7
Impairment Scale	Pearson Correlation	.999	.999	1
	Sig. (2-tailed)	.000	.000	
	N	7	7	7

Since the correlation coefficients provide only a global indication of the similarity between the scores without any information about its nature, it was decided to undertake a more detailed examination of the data using scatter plots. Figure 8 shows the scores obtained using the quality scale plotted against the scores obtained using the label-free scale. In addition to the mean values the graph also presents the associated 95% confidence intervals. The dashed diagonal line shows the $y = x$ axis. It can be seen on the figure that the results are scattered almost along the dashed line, which indicated a high similarity of scores obtained in both tests.

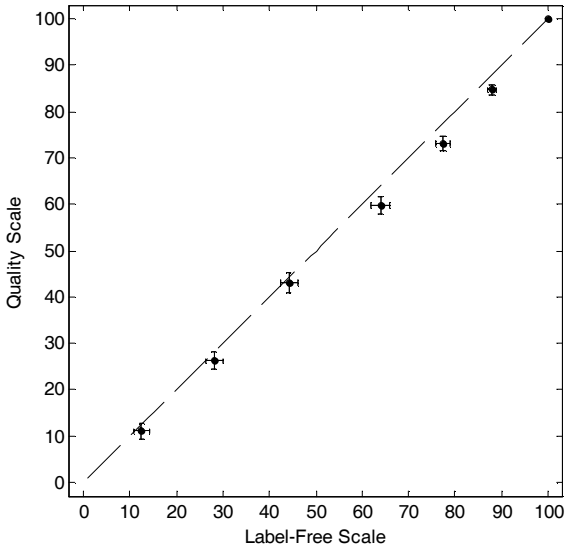

Figure 8 Comparison of the scores obtained using the quality scale and the label-free scale (mean values and associated 95% confidence intervals).

As mentioned above, the listening test employing the label-free scale could be considered to produce bias-free reference data. This assertion is based on the assumption that a scale without any labels is perceptually linear and, in the absence of any other biases, will yield bias-free results. On the other hand, in the case of the labelled scales it is expected that the scores will be biased due to the non-linear effect discussed earlier in Section 2. Consequently, one could expect to see a non-linear relationship between the scores presented in the scatter plot. However, as it can be seen in Figure 8, the relationship between the scores is almost perfectly linear. This result was also confirmed by a regression analysis (not presented here) showing that a linear model is capable of predicting more than 99% of variance in the data.

A similar linear relationship was also observed between the scores obtained using the impairment scale and the label-free scale, which is demonstrated in Figure 9. However, in this case it is possible to see a consistent vertical offset of data, indicating that the scores obtained using the impairment scale were slightly underestimated compared to the data obtained using the label-free scale. In order to quantify this offset a linear regression analysis was performed. According to the obtained results (not presented here) the shift of the data was equal to -6 points. This result was significant at $p < 0.01$ level. Since the proportion of the explained variance in the linear regression model was higher than 99%, the observed relationship between the data can be considered as linear too.

Contrary to the conclusions reached by other researchers summarized in Section 2, the results presented above indicate that the relationship between the investigated labeled scale and the label-free scale is linear. Although one may still argue that it is possible to see some small non-linear effects in our results, especially in Figure 8, the degree of this effect is much smaller than it could be expected on the basis of the literature review. It was checked that the application of the 3rd order non-linear model to the data presented in Figure 8 gives improvement of only 0.3% over the linear model in terms of the percentage of the explained variance. When a similar analysis was applied to the data presented in Figure 9, the percentage of the explained variance in the non-linear model was only 0.2% higher that that explained by the linear model. Consequently, the non-linear effects in our data can be regarded as negligibly small.

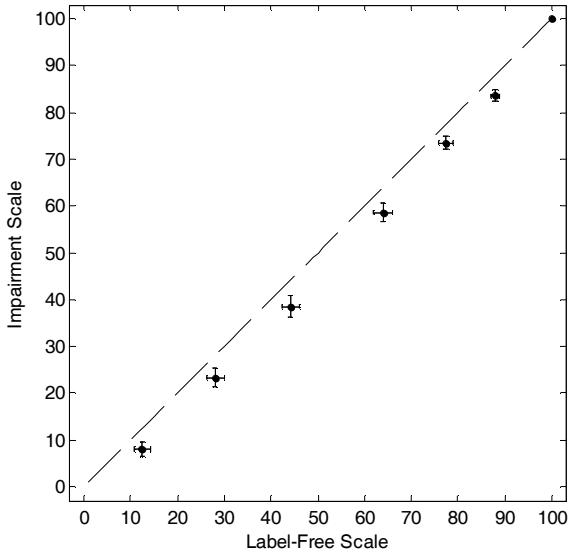

Figure 9 Comparison of the scores obtained using the impairment scale and the label-free scale (mean values and associated 95% confidence intervals).

As it was mentioned above, the results obtained in the listening test employing the label-free scale were regarded as reference data and therefore the scatter plots presented so far always included the data from that test. However, for the sake of completeness it was also decided to present in this paper the results obtained using the quality scale plotted against the data from the impairment scale (see Figure 10).

It is interesting to note that for the high quality stimuli (scored above 50) the results obtained both in the case of the quality scale and in the case of the impairment scale are almost identical. However, for the low quality recordings the scores obtained for the quality scale are slightly greater than the scores obtained using the impairment scale.

The results presented in this section almost completely answer the two research questions posed at the outset of this investigation (see the beginning of Section 4). They show that the degree of the non-linearity of the labeled scales is negligibly small compared to the data obtained using the label-free scale (answer to the first question). In addition, the results indicate that the listeners use all three scales in the similar way in terms of their span. It can also be seen in Figures 8-10 that the confidence intervals have a similar size in all scales, which indicates that the all the scales yield results exhibiting a similar magnitude of the experimental error (answer to the second question). However, it was also found that there are some important differences in the way the listeners used the scales, which will be the discussed in the next section.

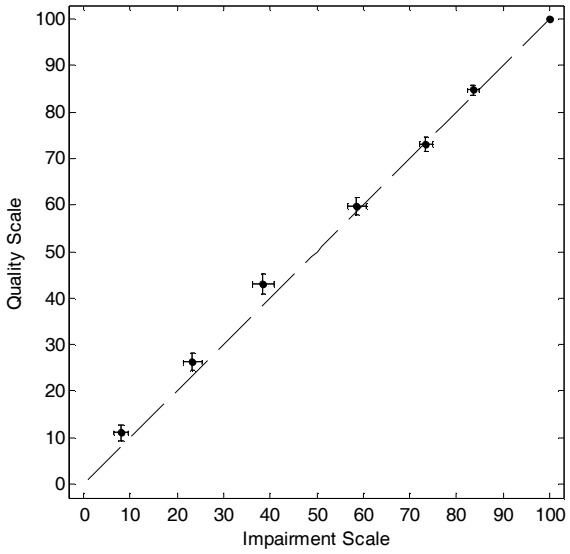

Figure 10 Comparison of the scores obtained using the quality scale and the impairment scale (mean values and associated 95% confidence intervals).

5.2. Distribution Analysis

The histograms presented in Figures 11-13 show the distributions of the raw data obtained using the quality, the impairment and the label-free scales respectively. As it can be seen, there are significant differences between them. For example, in the case of the impairment scale (Figure 12) it is possible to see a strong quantization effect. It manifests itself by the distinct peaks in the histogram near the labels and also near the points located half way through in-between the labels. This effect was caused by the fact the listeners used the points marked by the labels or numbers more often than the other points on the scale. They also frequently used the points located half-way through between the labels. A similar effect, although less pronounced, can be observed in the case of the quality scale (Figure 11). The distribution of the data obtained using the label-free scale also exhibits some peaks in the histogram (Figure 13). However, their number is less than that observed in the two previous cases. Consequently, one may conclude that the label-free scale has a potential

advantage of yielding the data with the less pronounced quantization effect. This conclusion requires further experimental verification.

There are some researchers who assert that different subjects have their own individual ways of using the scale. For example, subjects appear to have preferred ranges or numbers on the response scale that they use most frequently. This phenomenon is referred to as an idiosyncratic scale usage bias [13]. As a result of this bias, the data histograms vary between the listeners. The observed differences in the distribution of the data are so unique that the histograms can be considered to be "finger-prints" of the listeners. For example, the histogram presented on the left-hand side of Figure 14 comes from the listener who only used six distinct points on the impairment scale, leading to a severe quantization effect. The histogram presented on the right-hand side of this figure demonstrates the histogram of the listener who used the same scale in a more continuous way, although he or she used some points of the scale more frequently than others.

Figure 15 shows two examples of the histograms obtained in the case of the label-free scale. On the left-hand side of this figure it can be seen that the selected listener used the scale in the continuous manner. In addition it can be seen that he or she used some intervals of the scale more frequently than the others. In contrast, the listener whose histogram is presented on the right-hand side of the figure used only three intervals along the label-free scale that can be described as high, medium and low. These examples support the assertion that the listeners use the scales in their own individual way. They also show that the differences in the distribution histograms can be affected not only by the differences in the used scale but also by the inter-listener differences in the way they use of the scales. This emphasizes the need for using a large number of participants in the listening test in order to "average out" any adverse effects caused by differences between the listeners. The results showed above also demonstrate that analysis of the raw data can provide useful information about the use of the scale and about the differences between the listeners. However, most of the reports from the listening tests that the authors are aware of, present the results only in terms of the statistics such as the mean values or confidence intervals. Therefore, in order to gain more information from the listening tests, it is recommended that the reports should also contain the description of the distribution of the raw data.

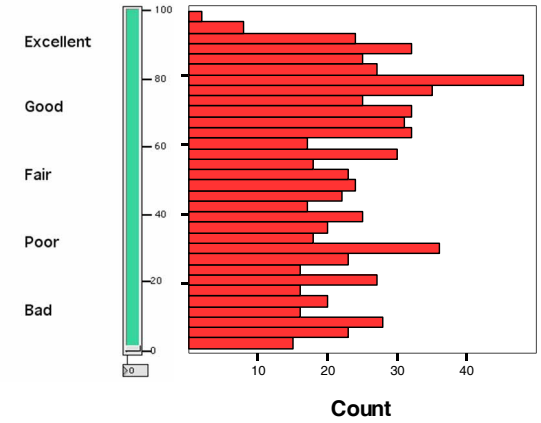

Figure 11 Distribution of scores obtained using the quality scale.

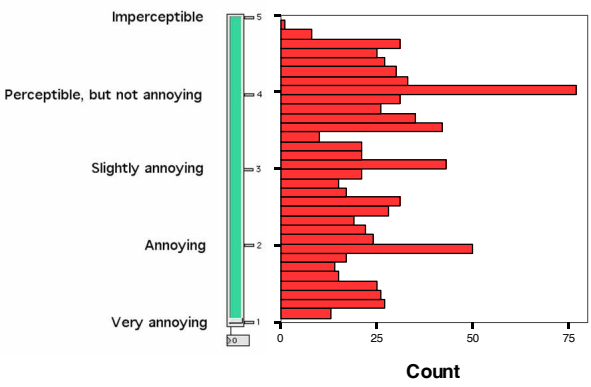

Figure 12 Distribution of scores obtained using the impairment scale.

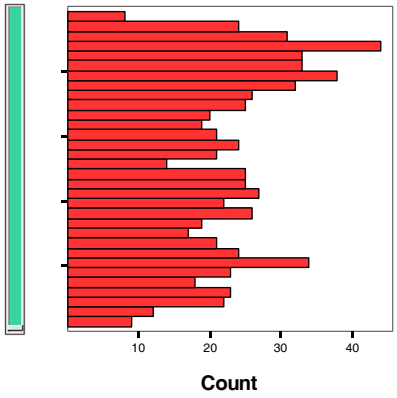

Figure 13 Distribution of scores obtained using the label-free scale.

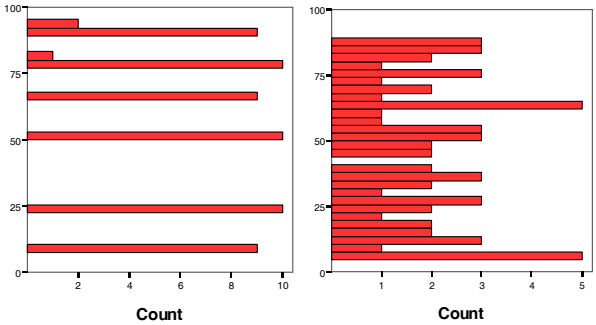

Figure 14 Examples of data from two different listeners obtained using the impairment scale.

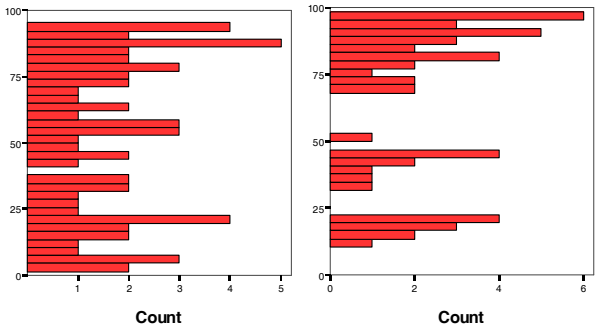

Figure 15 Examples of data from two different listeners obtained using the label-free scale.

6. SUMMARY AND CONCLUSIONS

The research has shown that the semantic differences between the adjacent labels used both in the ITU quality and impairment scales are not equal. The degree of this non-uniformity varies across languages. The above observation undermines the equal-interval properties of the labeled scales. Consequently, some researchers concluded that the ITU quality and impairment scales are non-linear in a perceptual sense.

In order to check the degree of non-linearity exhibited by the ITU quality and impairment scales, three separate listening tests involving three independent groups of listeners were performed. They all shared the same experimental protocol and involved evaluation of the same stimuli. The main differences between the tests were the scales used. In the first test, the ITU continuous quality scale was employed. In the second test, we used the ITU continuous impairment scale, whereas the last test employed the label-free sale. The last case of the label-free scale can be considered as the control condition since it is possible to assume that the scale is perceptually linear due to the lack of any labels.

All three listening tests led to almost identical results in terms of the scores averaged across the listeners. The data revealed that there was an almost linear relationship between the results obtained using the ITU quality scale and the label-free scale. A similar relationship was found between the scores obtained using the ITU impairment scale and the label-free scale. However, when the distribution of the raw data obtained using the three scales was examined, it was found that the label-free scale yielded less quantized data than the labeled scales. The quantization effect was the most pronounced in the case of the ITU impairment scale. It was also found that useful information about the usage of the scale can be acquired by examining the individual listener's histograms.

It is difficult at this stage to provide a reliable explanation for why the listening tests yielded almost the same results in all three cases. However, one possible explanation is that the listeners ignored the meaning of any labels and used the graphic scales without reference to the labels, or perhaps only taking the end point labels into account. If this supposition is correct and if it is confirmed by future experiments, the use of labels may be rendered obsolete and consequently it might be advisable to undertake listening tests using label-free graphic scales.

7. ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Mr. Yu Jiao (Joey) for providing the MUSHRA listening test interface and low-pass filter Matlab code.

8. REFERENCES

[1] ITU-R Rec. BS.1116-1, "Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunications Union, Geneva, Switzerland (1994).

[2] ITU-R Rec. BS.1534-1, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," International Telecommunications Union, Geneva, Switzerland (2003).

[3] B.L. Jones and P.R. McManus, "Graphic Scaling of Qualitative Terms," SMPTE Journal, pp. 1166-1171 (1986).

[4] ITU-R Report BT.1082-1, "Studies toward the Unification of Picture Assessment Methodology," International Telecommunications Union, Geneva, Switzerland (1990).

[5] N. Narita, "Graphic Scaling and Validity of Japanese Descriptive Terms Used in Subjective-Evaluation Tests," SMPTE Journal, pp. 616-622, (1993 July).

[6] M.T. Virtanen, N. Gleiss and M. Goldstein, "On the Use of Evaluative Category Scales in Telecommunications," in Proc. of Human Factors in Telecommunications, Melbourne, Australia (1995).

[7] K. Teunissen, "The Validity of CCIR Quality Indicators along a Graphical Scale," SMPTE Journal, pp. 144-149 (1996 March).

[8] A. Watson, "Assessing the Quality of Audio and Video Components in Desktop Multimedia Conferencing," Ph. D. theses, Department of Computer Science, University College London (1999).

[9] ITU-T Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," International Telecommunications Union, Geneva, Switzerland (1999).

[10] R. Guski, "Psychological Methods for Evaluating Sound Quality and Assessing Acoustic Information," Acta Acustica united with Acustica, vol. 83, pp. 765-774 (1997).

[11] S. Zieliński, F. Rumsey, R. Kassier and S. Bech, "Development and Initial Validation of a Multichannel Audio Quality Expert System," J. Audio Eng. Soc., vol. 53, pp. 4-21 (2005).

[12] P.E.A. Brooks, "On the Scales Used in Modern Listening Tests," Final Year Tonmeister Technical Project, Institute of Sound Recording, University of Surrey (2007).

[13] H.T. Lawless and H. Heymann, *Sensory Evaluation of Food.* Principles and Practices (Kluwer-Plenum, London, 1998).

9. APPENDIX

As mentioned in Section 4, the Multichannel Audio Quality Advisor was used in order to determine the cut-off frequencies used for the filtering of the original, 2-channel stereo recording. Since the prediction tool mentioned above was originally developed for multichannel audio signals [11], there was some risk of an error. Therefore, in order to avoid any experimental error, the low-pass filtered stimuli were assessed in a pilot test by six trained listeners. The obtained results matched closely the predicted results and therefore it was decided to use the above stimuli in the proper listening tests.

After completing the proper listening test, the predicted scores were compared again with the scores of the listening test. For this purpose, the results obtained using the MUSHRA test with the ITU quality scale were used since a similar method was used in the development of the Quality Advisor. It was found again that the actual scores matched the predicted scores well, which is illustrated in Figure 16. As can be seen, for the stimuli scored above 40, the results obtained in the listening test were almost identical to the results obtained from the Quality Advisor. It was only for the two stimuli exhibiting the lowest quality that some discrepancy between the data was observed. However,

considering the small magnitude of the differences (less than 7 points), it can be concluded that the Quality Advisor performed well even when it was applied to the prediction of the quality of filtered 2-channel stereo recordings. This result confirms the validity of the applied prediction tool.

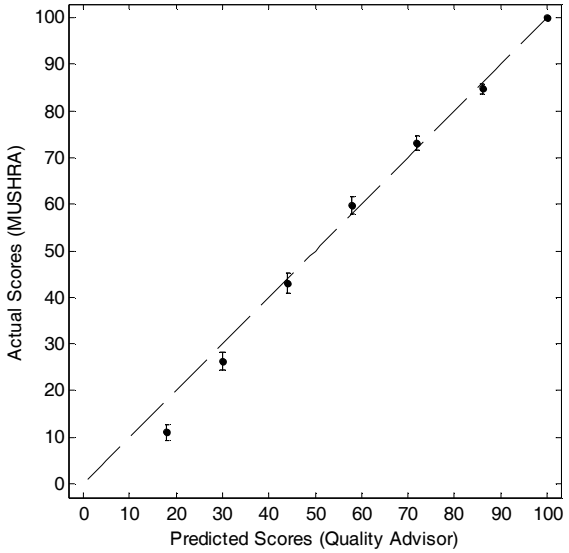

Figure 16 Comparisons of the actual and predicted scores using Multichannel Audio Quality Advisor.