Comparison and Analysis of Listening Test Methods for Development of Perceptual Speech Quality Assessment

Shi-Han Chen, Shun-Ju Chen, and Chih-Chung Kuo

Advanced Technology Center, ICL, ITRI, Hsinchu, Taiwan
korochen@itri.org.tw

Abstract

Reliability of the listening test design has a great influence on the performance of the quality estimation model. In this paper we compare four different listening test designs by Monte Carlo simulation. Three common problems of interval scale ratings are included in the simulation, and their influences on the performance of estimating the underlying true quality are investigated. It turns out that in these methods, randomly choosing partial trials for Scaled Comparison could be the most reliable way to perform listening test under the influences of interval scale ratings problems.

1. Introduction

Objective evaluations of speech have been developed and used in many different research areas. In the last decade, perceptual-based quality assessments that bridge the gaps between objective measures and human perception of speech quality have drawn much attention. Methods in this area such as PESQ [1] and TOSQA [2] are very well known and have good performances. In these methods, a quality estimation model trained by a large number of auditory test results is often included to predict subjective quality from a vector of objective comparison results [1] [2]. Therefore, reliability of the listening test design has a great influence on the performance of the quality estimation model.

One of the most widely used subjective test paradigm in telecommunications is Absolute Category Rating (ACR) [3]. However, ACR tends to give lower sensitivity in distinguishing stimuli with similar quality [2] [3] [4], and this could bring a less precise quality estimation model. Moreover, since there is no reference stimulus, rather high demands are made on listener's capability of judgment.

On the other hand, relative assessments such as Degradation Category Rating (DCR) [3] and Paired Comparison (PC) [5] [6] offer higher sensitivity than ACR. In addition, PC requires only a binary preference of the pair, and experience showed that it was preferred by non-trained listeners [6]. However, PC is much slower than ACR and DCR, and in [8] the authors used sorting algorithm to reduce the necessary trials. One major drawback of PC is that listeners compare stimuli on an ordinal scale, not interval scale. It means that only the ranking of stimuli is available, distances between pairs are actually unknown and further algorithms and assumptions are needed to estimate the location of each stimulus on the z-scale [5]. This could make combination of scores of different test sessions difficult, and may not be suitable for development of quality estimation models that require a large number of subjective test results.

One solution is to extend the basic PC by using an interval scale such as that in Comparison Category Rating (CCR) [3] to grade not only the binary preferences but also the *dissimilarities* between pairs. This metrology is called Scaled Paired Comparison [9] or simply Scaled Comparison (SC) [10]. In this way the obtained information is on interval scale, and sensitivity is also maintained by the relative judgment. However, a large number of comparisons are still required.

DCR and SC seem to be the better choices for development of a quality estimation model due to their sensitivity and interval scale, and now we want to know which one is more reliable. In this paper, we compare and analysis their reliabilities of estimating underlying true subjective quality by using Monte Carlo Simulation. Knowing that rating with interval scales have some common problems [13], we analysis each of these problems separately to investigate their influences on the different listening tests. Furthermore, we have tried two different approaches to reduce the required number of trials of SC in each session. The first one is selecting trials in each test session by Quicksort algorithm [11], and another one is to randomly choose partial trials in each test session. Unidimensional scaling of the resulting incomplete comparison matrices can then be solved by quadratic programming [12]. Simulation results are presented in Mean Absolute Error*s* between the estimated quality and the simulated underlying true quality versus the total number of listened pairs.

Four different listening test designs, which are DCR, SC, SC + Quicksort (SCQ), and SC + random (SCR), are simulated in this paper. CCR is not compared because there is no order effect in DCR or CCR [3] and the simulation result should be the same as DCR.

The paper is organized as follows. Section 2 first describes the general concept of listening test methods, and in Section 3 we present the common problems of interval scale ratings. In Section 4 we introduce the unidimensional scaling method, and Section 5 describes two methods used to reduce the required trials for SC. Simulation results are given in Section 6, and a discussion of the results concludes the paper.

2. Listening test methods

Listening test methods generally contain two successive procedures, *rating* and *scaling*. In the rating procedure, listeners are presented with different stimuli, and give their ratings (scores) of stimuli on the designated rating scale. It is important to note that scores provided by a listener depend both on the perceived quality of the stimuli and the rating criterion being used by that listener. These two factors may vary between listeners, or even with time. Therefore a good simulation of rating should always simulate these factors as close as possible.

When the rating is completed, in the scaling procedure the resulting scores of different listeners are transformed into values indicating the perceptual quality of the stimuli, as well

as their relative positions on the perceptual dimension. In ACR, for example, the scaling simply means averaging the scores of each stimulus, and the resulting score is represented in the form of Mean Opinion Score (MOS). For comparison tests, the scaling process is more complicated since the scores are all relative. In Section 4 we will describe the unidimensional scaling process for SC.

3. Rating with interval scale

In this paper, we are concerned about the reliability of DCR and SC, which use interval scales in ratings. SC can be treated as a generalized version of PC. Like PC, all possible combinations of pairs of the given stimuli will be judged. However, PC gives only binary preferences, while in SC the binary preferences become the more informative dissimilarities due to the interval scale being used. SC integrates the useful interval scale into the reliable and sensitive PC, and therefore it may be more suitable for the development of a perceptual assessment model. Unfortunately, it also inherits some problems from the use of interval scales. In [13] it is mentioned that there are three common problems in interval scales rating. In the following we will give some explanations on those phenomena as well as how we simulate them in terms of mathematical equations.

3.1. Lack of intra-listener consistency:

A listener could give different ratings to the same stimulus at different times. Generally it is assumed that both the perceived value and the judgment criterion for the same stimulus are changing over time to time. This variation is assumed to occur because of random errors, and therefore it is usually described by adding a random variable of normal distribution to each stimulus's true quality every time it is being listened.

3.2. Unequal-interval judgment scale:

Intervals between rating categories on the underlying psychological dimension will not necessarily be equal, which means the quality difference between ratings 5 and 3 may not be equal to difference between 7 and 5. One obvious reason is called "end-point problem", which is usually caused by someone giving a lowest/highest score to a stimulus that is not the worst/best, and after that she has to rate another worse/better stimulus. This phenomenon results in a sigmoid relation between rated scores and underlying true scores. In comparison ratings, for example, it means stimuli with extreme perceptual differences and with less extreme differences will be both given the same rating. Note that the end-point problem is not the only source of unequal-interval scales, which means interval sizes may completely a random variable for any listener.

3.3. Lack of inter-listener correspondence:

Lack inter-listener of correspondence could result from both the differences in perception and different rating criteria across listeners. Linear differences include different origins and interval sizes of the scoring criteria. For example, in comparison ratings listener A may give two different pairs the scores 2 and -4, while listener B rates those pairs 4 and -2. We say their rating criteria are shifted by 2. On the other hand, interval sizes affect the ranges in which stimuli are given the same ratings. For example, pairs P_1 and P_2 may be rated 3 and 4 by listener C, while listener D thinks they are both 3. In this case, listener D has a larger interval size than listener C in the interval labeled "3".

3.4. Simulation method for comparison ratings

First we assume that each stimulus has an underlying true quality q_j. To simulate the Problem 3.1, we add a normal random variable R_1 to each stimulus' quality every time it is being listened. The *perceived dissimilarity* between stimulus *j* and *k* is defined as

$$\lambda_{jk} = (q_j + R_1) - (q_k + R_1) \tag{1}$$

where λ_{jk} represents the perceived dissimilarity, and $R_1 \sim N(0, \sigma_1)$. $\lambda_{jk} < 0$ means the perceived quality of stimulus *j* is worse than stimulus *k*.

Next, in order to obtain the *rated dissimilarity* δ_{jk}, λ_{jk} has to be mapped onto the criterion scale used by each listener. First, we assume that criterion scale is symmetric for positive and negative λ_{jk}, and we generate the size of each interval on one side of the criterion scale by

$$Isize(i) = \begin{cases} ini\_size\ , & 0 \le i < Csize \\ slope \cdot (i - Csize) + ini\_size\ , & Csize \le i < Inumber \end{cases} \tag{2}$$

where *Isize* and *Inumber* are interval sizes and total number of intervals (number of rating categories), and *ini_size* is the interval size of interval *i* from 0 to *Csize*-1. *slope* controls the increase rate of size of unequal-size intervals between interval *i* from C*size* to *Inumber*, and *Csize* controls the number of equal-size intervals. Increasing *Csize* reduces the end-point problem, while increasing *slope* deteriorates the problem. Note that *ini_size* and *slope* must ensure that sum of all interval sizes equal to the maximum comparison score:

$$\sum_{i=0}^{Inumber-1} Isize(i) = \max(q_j - q_k) \tag{3}$$

After the interval sizes are all generated, the rated dissimilarity δ_{jk} of the perceived dissimilarity λ_{jk} can then be calculated by minimizing distances between λ_{jk} and the category judgment scale o_i

$$\delta_{jk} = o_I \tag{4}$$

where

$$I = \arg\min_i \{abs(\lambda_{jk} - o_i)\}\ ,\ 0 \le i < Inumber \tag{5}$$

and the distance between λ_{jk} and o_I must satisfy

$$abs(\lambda_{jk} - o_I) < Isize(I)/2 \tag{6}$$

There are *Inumber* points on the category judgment scale o_i. In DCR, for example, *Inumber* = 5 and o_i ranges from 5 to 1. To simulate Problem 3.3, we can add a uniform and even random variable to both *Isize* and *I* for each listener:

$$Isize(i) = Isize(i) + R_2 \quad , \quad 0 \le i < Inumber \qquad (7)$$

and Eq. (7) becomes

$$I = \arg\min_i \{abs(\lambda_{jk} - o_i)\} + R_3 \quad , \quad 0 \le i < Inumber \qquad (8)$$

Note that R_2 not only change the interval sizes between listeners, but also makes those unequal within each listener. We summarize the variables controlling the comparison process and the three interval scale rating problems in Table 1. In Figure 1 and 2 we plot the influences of *Csize* and *slope* on end-point problem, and influences of R_2 and R_3 on inter-listener inconsistency and unequal-interval problem, respectively.

Variable	Function	Description
R_1	Intra-listener inconsistency	Normal distribution. Add to each stimulus' quality every time it is being listened.
Csize	End-point problem	Number of equal-size intervals. Increasing *Csize* to reduce the end-point problem.
slope	End-point problem	Increasing rate of size of unequal-size intervals. Increasing *slope* to deteriorate the problem.
R_2	Inter- listener inconsistency	Uniform distribution. Add to *Isize* (interval sizes)
R_3	Inter- listener inconsistency	Uniform distribution Add to *I* (interval centers)

Table 1: Summary of different variables used in the simulation of rating by interval scales

4. Scaling with unidimensional scaling

Unidimensional scaling is the special one-dimensional case of multidimensional scaling. It is assumed that the psychological dimension is one-dimensional, which is generally accepted and used in listening tests. Given a matrix symmetric *D* of rated dissimilarities and another non-negative symmetric matrix *W* of weights. Both *W* and *D* have a zero diagonal. Unidimensional scaling finds the estimated quality x_j for *n* stimuli such that

$$\sigma(x) = \sum_j^n \sum_k^n w_{jk}(\delta_{jk} - (x_j - x_k))^2 \qquad (9)$$

is minimized [12]. δ_{jk} are the rated dissimilarities given in the rating procedure. Eq. (9) can be solved by quadratic programming, which finds x_j by minimizing

$$\varsigma(x) = \frac{1}{2} x^T H x + f^T x \qquad (10)$$

with the constraints

$$\begin{aligned} A \cdot x &\le b \\ Aeq \cdot x &= beq \\ \min(q_j - q_k) &\le x \le \max(q_j - q_k) \end{aligned} \qquad (11)$$

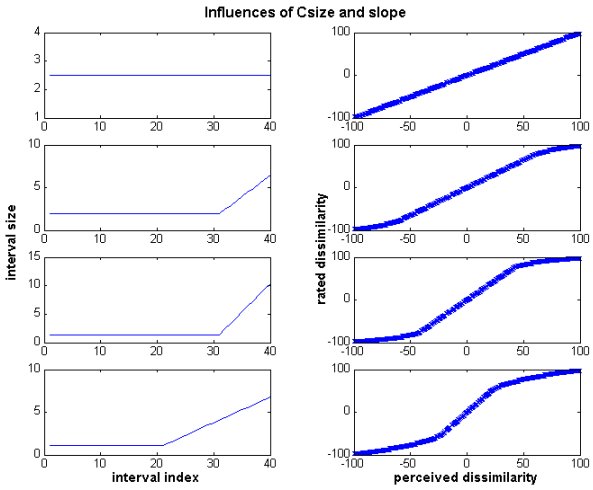

Figure 1: Influences of *Csize* and *slope* on end-point problem. The first column is interval size, and the second column is mapping relationship between *perceived dissimilarities* and *rated dissimilarities*. Variable settings from top to bottom:
Csize = *Inumber*, *slope*= 0;
Csize = 0.75**Inumber*, *slope*= 0.5;
Csize = 0.75**Inumber*, *slope*= 1.0;
Csize = 0.50**Inumber*, *slope*= 0.3

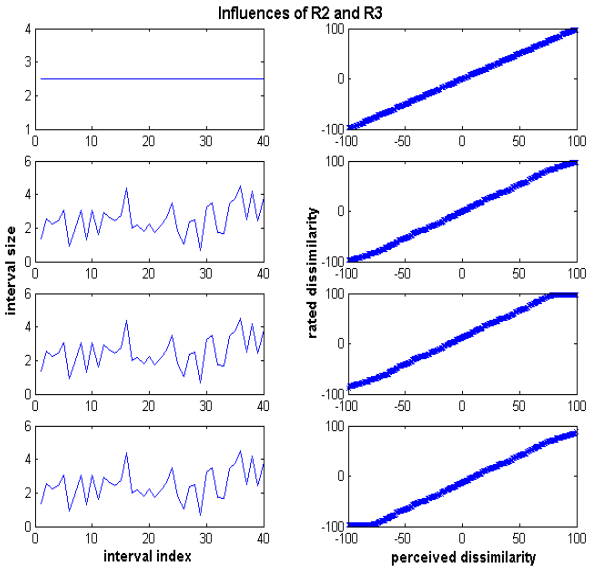

Figure 2: Influences of R_2 and R_3 on inter-listener inconsistency and unequal-interval problem. The first column is interval size, and the second column is mapping relationship between *perceived dissimilarities* and *rated dissimilarities*. R_2 = 0 for the first row, and is a uniform-distributed random variable from -2~2 for the second to fourth row. R_3 = 0, 0, 5, -5 from the first to the last row.

f and H can be derived by

$$H = 2\begin{bmatrix} \sum_k w_{1k} & -w_{12} & \dots & -w_{1n} \\ -w_{21} & \sum_k w_{2k} & \dots & -w_{2n} \\ \dots & \dots & \dots & \dots \\ -w_{n1} & -w_{n2} & \dots & \sum_k w_{nk} \end{bmatrix} \quad (12)$$

$$f = 2\begin{bmatrix} \sum_k g_{1k} \\ \sum_k g_{2k} \\ \dots \\ \sum_k g_{nk} \end{bmatrix} \quad (13)$$

where

$$g_{jk} = \delta_{jk} \cdot w_{jk} \quad (14)$$

A, b can be defined if the order of input stimuli is found by sorting algorithms. Aeq, beq can be defined by the reference stimulus whose quality is always known since the reference stimulus is the unprocessed signal. In this paper, elements of W are all set to one for simplicity.

5. Reducing the number of trials for SC

In general, the required number of trials for SC in each session is larger than DCR. The number of trials required to do each comparison goes up with the number of stimuli:

$$Trials = \frac{n(n-1)}{2} \quad (15)$$

In reality, it makes the required time to finish each session very long. In [3] it is mentioned that in no case should a session exceed 45 minutes. In [6] it says the length of a session should be limited to approximately 30 min. Therefore it is necessary to reduce the number of trials in each session. One solution is to sub-divide one SC session into two or more sessions. In this paper, we have tried two different approaches to reduce the number of trials for SC in each session. The first one is selecting trials in each test session by Quicksort algorithm [11], and another one is to randomly choose partial trials in each test session.

5.1. Using Quicksort to reduce the number of trials

It is well known that it is not necessary to compare every pair to derive the order of a list of numbers, and sorting algorithms can be used to choose the necessary trials from all possible comparisons. In [8] the authors have tried sorting algorithms to reduce the required trials for PC. For SC, each comparison result δ_{jk} includes the order and the rated dissimilarity of that pair. These orders are used to select the required comparisons and sort the stimuli, and the corresponding rated dissimilarity can be recorded to form a partial matrix of dissimilarities. The quality of each stimulus can then be estimated by unidimensional scaling which is described in the previous section.

There are many different kinds of sorting algorithms, and in this paper we decide to use Quicksort. Quicksort is the fastest sorting algorithm in average. The best pivot, which is the median of the list of numbers in each "divide" operation, could be easily chosen since we have category rating for each stimulus. In fact, if no problem described in Section 3 is presented among listeners, which means the underlying true perceptual quality of each stimulus is perfectly recognized without any error, using only those partial trials selected by the sorting algorithms will give exactly the same estimated quality as the underlying true quality. However, since the aim of sorting algorithms is only to find the order, the selected trials are not guaranteed to be the most representative and reliable trials for that session. In the next section simulation results of this method will be given and analyzed.

5.2. Random selection of trials

In order to find out whether it is good or not to select trials by Quicksort, in the following we try to randomly choose partial trials in each test session. In order to make these two methods comparable, the number of selected trials in each session is set to be the same as Quicksort. The quality of each stimulus is also estimated by the same unidimensional scaling method.

6. Monte Carlo Simulation

The four listening test designs, DCR, SC, SC + Quicksort (SCQ), and SC + random (SCR) that we mentioned before are simulated. We simulate an experiment where there are 20 stimuli in each session, and among which there is always one unprocessed signal as the reference stimulus whose quality is set to 0. Each stimulus' true quality value is chosen from a uniform random interval that spanned from 0 to 100. All of the problems described in Section 3 are simulated. Simulation results are presented in Mean Absolute Error*s* of 20 sessions versus the total number of listened pairs.

6.1. The influence of number of rating categories

It's interesting to first find out what is the relationship between the number of rating categories and the performance. In Figure 3 we plot the performances of different *Inumber,* and $R_1 \sim N(0,1)$. From the figure it shows that DCR is more easily influenced by the number of categories, while the comparison ratings are less influenced. It is reasonable since the number of categories represents the rating resolution. Comparison ratings provide more information than DCR about the relative distances of stimuli, and therefore are more likely to restore the original position of each stimulus when resolution is decreased. It has to be mentioned that it is not feasible for listeners to judge on a scale with 50 different categories. Performances in this simulation are all relative to the range of simulated true quality. In this case rating with 50 categories simply means quantizing the original continuous true quality with step size of 2.

In order to investigate each of the interval scale rating problems on the same basis, and since the number of categories is always fixed for every listening test method, in all the following simulations we set *Inumber* to 40, which give about the same performance for all the four designs under consideration in this case.

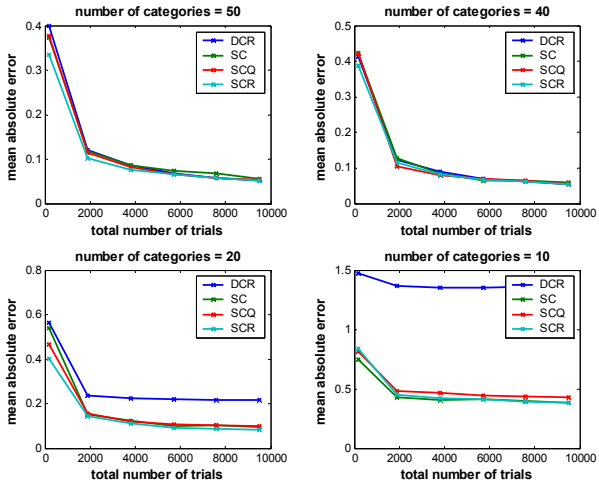

Figure 3: Influences of number of categories

6.2. The influence of intra-listener inconsistency

We plot the performances of different values of R_1 in Figure 4. *Csize* is equal to *Inumber*, and *slope,* R_2, R_3 are all equal to 0. When only the problem of intra-listener inconsistency is simulated, which means the interval of judgment scale for each listener is perfectly equal, and scoring criteria across listeners are exactly the same, the four methods result in very similar performances. Reliability is decreased when the inconsistency becomes worse. Note that DCR, SCQ and SCR require fewer trials in each session than SC while their performances are similar.

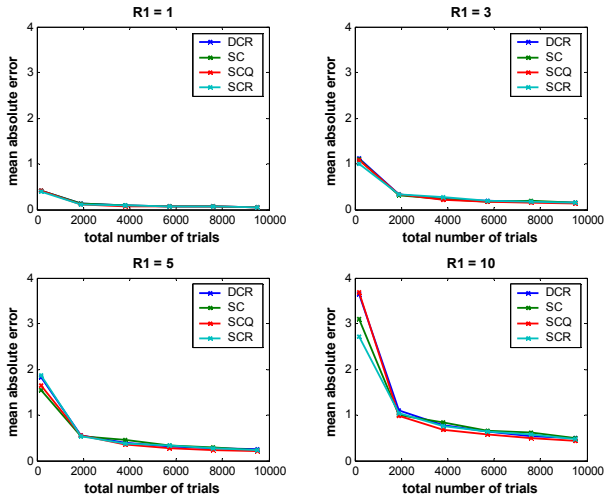

Figure 4: Influences of intra-listener inconsistency

6.3. The influence of end-point problem

Here we simulate the end-point problem by varying *Csize* and *slope* to control the different degree of end-point problem. $R_1 \sim N(0,1)$, and R_2, R_3 are both equal to 0. We plot the results in Figure 5. It shows that three comparison methods are getting more and more reliable over DCR when the endpoint problem is getting worse. This is reasonable, since stimuli located near the two extreme sides of scale will be given the same score in DCR due to the endpoint problem. On the other hand, comparison methods also compare these stimuli, and more information could be used in the final scaling process, which gives a more reliable estimation of scores. SCQ gives a slightly worse performance than the other two comparison methods, which may due to the fact that sorting algorithms always tend to compare closer stimuli. This gives an unbalanced comparison, and it was already discussed in [8]. If we reorder stimuli according to their true quality, sorting algorithms will mostly choose those adjacent pairs that lie near the diagonal region of Figure 6.

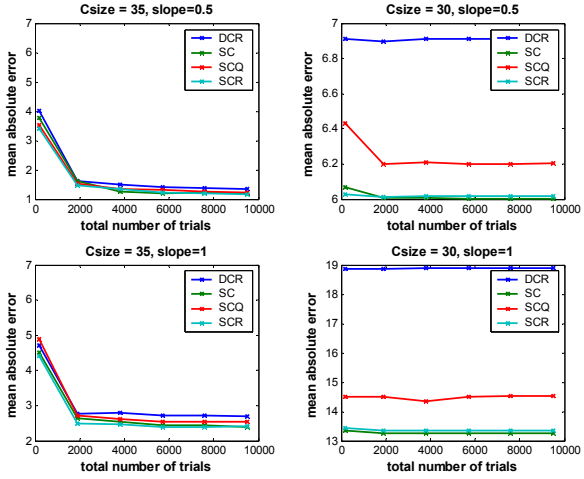

Figure 5: Influences of the end-point problem

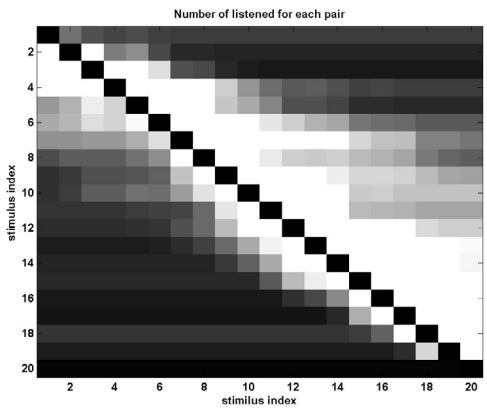

Figure 6: Listened times for each pairs, darker area means lower times of listening.

6.4. The influence of inter-listener inconsistency

The inter-listener inconsistency is simulated by different random amounts of R_2 and R_3 across listeners. $R_1 \sim N(0,1)$, *Csize* is equal to *Inumber*, and *slope* is set to 0. Results are shown in Figure 7. DCR and SCR always give the best performances in this case, while SC and SCQ are worse. The reason is that the required trials in each session are largest for SC, which means it has the fewest listeners than others when the total listened trials are the same. Since R_2 and R_3 are even uniform-distributed random variables, more reliable

estimation could be obtained by using comparison results from more listeners. Therefore DCR and SCR that requires fewer trials than SC in each session give better results. In the figure it also shows that when there are shifts in origins between listeners, SCQ are worse than both DCR and SCR. As mentioned before, errors due to shifts in scorings will be mostly restricted in adjacent pairs when using sorting algorithms. This will prevent quadratic programming from using other shifted distances of non-adjacent pairs, and it makes estimated positions of stimuli shift from the correct positions more seriously.

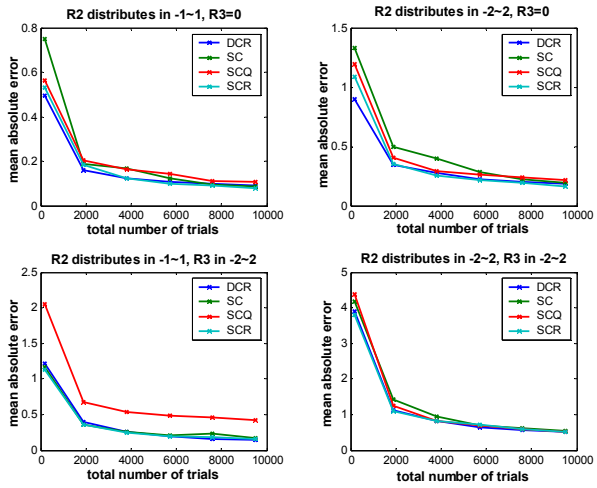

Figure 4: Influences of inter-listener inconsistency

7. Conclusions

From the above observations, we can have several conclusions:

1. DCR may be more easily influenced by the number of categories, while the comparison ratings are less influenced. It means when the resolution of a category rating method is not high enough, more comparisons between stimuli should be judged.
2. Reliability is decreased whenever one of the interval scale rating problems is getting worse.
3. The four methods result in very similar performances when there is intra-listener inconsistency. However, DCR, SCQ and SCR require fewer trials in each session than SC while their performances are similar
4. In the case of end-point problem, it shows that the three comparison methods are getting more and more reliable over DCR when the endpoint problem is getting worse. SCQ gives a slightly worse performance than the other two comparison methods, which may due to the unbalanced comparison.
5. DCR and SCR give the best performances when there is inter-listener inconsistency. SCQ are influenced more than other methods when there are shifts in origins between listeners. SC requires the most trials, and it makes SC the most unreliable method in this case.

To conclude, SC integrates the informative interval scale into the reliable and sensitive PC, and it may be more suitable for the development of a perceptual assessment model. Randomly choosing partial trials for SC could be the most reliable way in these methods to perform listening test under the influences of interval scale ratings problems.

8. References

[1] ITU-T Recommendation P.862, “PESQ: An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs”, ITU, 2001

[2] ETSI EG 201 377-1 V1.2.1, “Speech Processing, Transmission and Quality Aspects (STQ); Specification and measurement of speech transmission quality; Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks”, ETSI, 2002

[3] ITU-T Recommendation P.800, “Methods for subjective determination of transmission quality”, ITU, 1996

[4] S. Dimolitsas. "Subjective Quality Quantification of Digital Voice Communication Systems," *IEE Proceedings, Part-I, Communications, Speech and Vision,* Vol. 138, No. 6, pp. 585-595, Dec. 1991.

[5] ITU-T "*Handbook on Telephonometry*" Geneva, 1993

[6] IEC 60268-13, “Sound system equipment Part 13: Listening tests on loudlisteners”, IEC, 1998

[7] Thurstone, L.L., “The method of paired comparisons for social values”, *Journal of Abnormal and Social Psychology*, 1927, 21, 384-400

[8] D. A. Silverstein and J. E. Farrell, “Efficient method for paired comparison”, *Journal of Electronic Imaging*, 10(2), April 2001, pp. 394-398

[9] Jussi Hynninen, Nick Zacharov, “GuineaPig - A generic subjective test system for multichannel audio”, *AES 106th Convention*, May 8-11 1999, Munich, Germany

[10] R., Quinn, K., O'Sullivan, D., Lewis, D., Wade V.P., “On the Application of Paired Comparison to Trust”, *2nd International Workshop on Managing Ubiquitous Communications and Services (MUCS)*, Dublin, 2004

[11] Hoare, C. A. R. "Partition: Algorithm 63," "Quicksort: Algorithm 64," and "Find: Algorithm 65." *Comm. ACM* 4, 321-322, 1961

[12] Brian Everitt, David Howell, *Encyclopedia of Statistics in Behavioral Science*, 2nd Edition, Wiley, 2005

[13] Brown, Thomas C.; Daniel, Terry C, “Scaling of ratings: concepts and methods”, RM-293. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, 1990