SUBJECTIVE EVALUATION OF HIGH QUALITY AUDIO SYSTEMS

T. Grusec, Ph.D., Communications Research Centre, 3701 Carling Ave. Ottawa, Ont., K2H 8S2

Radio-frequency spectrum restrictions in radio broadcasting, and space limitations in other applications (such as audio storage on high cost media) demand bit-rate reduction. At the Communications Research Centre, we have been evaluating the subjective quality of audio codecs that reduce digital audio bit-rate by factors of 4 or more, depending on the application. The algorithms used to achieve such reductions are based on psychoacoustic models of hearing so that, theoretically, codecs should be able to operate transparently.

With the best of these codecs, operating at their higher bit-rates, such transparency is likely to be true for average listeners. However, very subtle differences among these codecs can become magnified, for example, by operating several of these codecs in tandem, or by post-processing applied after bit-rate reduction (typical in broadcasting). Differences among the codecs may, thus, become more obvious at the end of complex broadcast chains. Thus, the codecs are not necessarily fully equivalent, and it is essential to make fine-grained comparisons before choosing among them for broadcast, and other critical applications.

For achieving such comparisons, we use special conditions and procedures. We believe these special features are adaptable for the subjective evaluation of any high quality audio devices.

First, our listening room for subjective testing minimizes the effect of room reflections. From 160 Hz upward, our room yields a reverberation time of about 0.20 seconds. This rises slightly at lower frequencies to about 0.35 seconds below 125 Hz. These low values help to ensure that artifacts are not masked while still retaining enough reverberation to make the listening enjoyable. The other important room characteristic is background noise. We have attained an NR rating of 15.

We run listeners one at a time under blind conditions in our evaluation experiments. An innovative disk-based playback system permits seamless switching among three alternative versions of audio materials that each listener compares on each of the trials in a typical experiment. One alternative on each trial ("A") is (usually) an unprocessed reference material. The listener knows that "A" is the standard against which he or she must compare each of the other two versions heard on that trial. One of these other two versions ("B" or "C") is the same audio selection as the reference but processed through one of the codecs under evaluation. The second alternative ("C" or "B") is an "hidden reference," fully identical to the reference "A". The specific assignment of hidden reference and processed versions to "B" or "C" on any trial is not known to the subject. This assignment varies across the different trials in an experiment so that it is unpredictable to the listener.

Each subject must evaluate both "B" and "C" by comparing it to "A" on each trial of the experiment, using a 5-grade rating scale. We instruct subjects to treat this as a continuous scale to single decimal place resolution (a 41-point scale, in effect). Subjects do the switching with a mouse by operating three buttons "A", "B" and "C" seen on a computer screen and corresponding to the audio material version. We design the sessions to take no more than one-half hour for completion by each subject. The actual session length is under the control of the listener. He or she can switch freely among the three versions for as long as needed to decide the scores for each trial. Each 10 to 15 trial session contains all the codecs in the experiment. Each codec is usually presented several times within a session, intermixed unpredictably in the trial to-trial sequence with the other codecs.

Up to three listeners can be run in one afternoon. Two or three sessions of 10 to 15 trials each are usual, with each listener resting while the other two complete a session. Before these individual rating sessions, a group training phase takes place in the morning. In training, all the listeners for a given day can work together, along with a resource person, to become thoroughly familiar with the materials they will be rating in the afternoon. They can interact freely with each other and with the resource person, accessing and discussing all the materials they will be rating later. During training, in contrast to the blind conditions of the afternoon rating sessions, all listeners explicitly know which items are reference and which are processed versions. Thus, they can maximize their sensitivity to the often subtle differences between the versions, learning from each other and from the resource person. Training usually takes up the entire morning.

A crucial process precedes the actual beginning of a subjective evaluation experiment. This is the choosing of "critical" audio materials for use in the experiment. There are no *a priori* methods of making these choices. While work is underway to develop more efficient ways, it is now a tedious, empirical search among standard, commercial CDs and reference or test recordings. The materials must be "fair" ones, so that one should not use artificial materials explicitly designed to "break" a codec. On the other hand, the materials must stress each codec since most "statistically representative" materials will fail to reveal anything due to the high quality of the present generation of codecs.

The method used is simply one of bringing together a number of highly knowledgeable expert listeners and a large library of CD and other materials. Included are versions of these materials processed through the codecs under test. These experts then audition the various materials to find ones that stress each codec to reveal coding artifacts. They try to find a minimum of two stressful materials per codec. Since some materials stress more than one codec, the experts usually find as many materials in total as there are codecs for testing in an experiment rather than twice that number. In the ensuing subjective tests themselves, subjects evaluate each codec against all the materials found for all the codecs. The critical materials search can take up to a month, or more, for 5 or 6 codecs. This search time is usually longer than the time it will take to run the following evaluation experiment.

For the experiments themselves, it is essential that all the subjects are sufficiently sensitive to make the fine discriminations needed to evaluate the codecs reliably. A traditional approach for this purpose is to use pre-screening methods, such as audiometric testing, to choose subjects.

While pre-screening is useful, there are limitations if one uses this approach exclusively. For one thing, one does not usually know whether any given cut-off criterion used for inclusion and exclusion of subjects in conjunction with pre-screening is the most suitable one for a given experiment. If you set the criterion too rigidly, then you may exclude subjects who might have been entirely satisfactory for a given test. On the other hand, a criterion set too loosely may lead to the inclusion of too many deficient listeners. The second of these possibilities is the more serious error for sensitive experiments. From pre-screening alone, there is no way of checking on whether one or the other of these selection errors has occurred, and if so, what its magnitude was.

Also, even though pre-screening suggests that a given subject will do well in an experiment, his or her performance *at the actual time of the experiment* may not be up to that subject's usual capability.

A practical consideration is that formal pre-screening, such as audiometric testing, is costly and time-consuming.

To deal with factors like these, we use a two-fold set of criteria for listener selection. First, we pre-screen in a very loose way by choosing subjects mostly from occupational and interest groups that ought to contain many good listeners. These include audio professionals of various kinds, audiophiles and musicians.

The second step is to measure the actual performance of listeners as shown during an experiment. As mentioned, listeners give ratings to both of the two versions presented on each trial - i.e., to the item they believe to be the hidden reference, and to the one they have concluded is the processed or coded version. Which versions were the true hidden references and which were the coded ones is, of course, known to us as the designers of the experiment. Thus, for all the trials of each subject, we can compare the distribution of scores for the true references with the distribution for the coded items.

We may compare the means or averages of these two distributions with each other statistically by use of a *t*-test that takes into account the correlation between trials, and the variability of the distributions. If the mean for the coded version distribution is significantly different from the one for the reference, then one can infer that the subject was truly discriminating between these two versions. In that case, then, one can conclude that the subject's sensitivity was adequate for the task of the experiment. One can include his or her data along with those of other similarly sensitive subjects in the final analysis of experimental outcomes. On the other hand, if those two means are statistically identical, then one cannot reject the hypothesis that the subject was guessing, overall, rather than properly discriminating between the coded and the hidden reference versions. In this case, then, one can omit that subject's data from the experiment on the grounds of insufficient sensitivity to the experimental task.

Rather than working with the two distributions of scores, there is a fully identical alternative process. One can subtract one of the scores on a trial (reference or coded) from the other one (coded or reference) and work with the resulting single distribution across all trials for each listener. This subtraction procedure automatically takes the correlation between trials into account. A *t*-test would show if the mean of this distribution is statistically zero (indicating guessing) or different from zero (indicating true discrimination).

Over the years, we have built up a pool of listeners and have been able to track the performance over time of those listeners who have been in more than one experiment. Regarding our two stage process of listener selection, we can report that some listeners who were quite adequately sensitive in one experiment are occasionally deficient in another one. This argues that any pre-screening criterion used alone is insufficient to ensure that only good listeners contribute data. A given listener may be good or not, as seen in our *t*-test, depending on factors such as the relative difficulty of detection of the coding artifacts in a specific study.

Although the scores used to measure listener sensitivity are the same ones that we use to draw conclusions about the codecs, we use these scores in independent ways for these two purposes. The listener sensitivity measure (*t*-test) basis is only whether a subject correctly judged items as hidden reference or coded, and not on how he or she evaluated the quality of specific codecs. To illustrate this, we could take any set of data from one of our completed experiments and alter the assignment of rating numbers to specific codecs without disturbing the sensitivity values at all. We usually achieve statistically significant final evaluations of codecs. This means that sensitive listeners tend to be highly consistent with each other in these evaluations, as one would expect.

Our experiments need very few listeners (sometimes as few as half a dozen) to produce conclusive results. Contributing to this reliability is our exclusive use of within-subject (repeated measures) experimental designs. Such designs eliminate individual difference effects.

CONCLUSIONS

The high sensitivity of our experiments is due to many factors. These include: the listening environment and equipment; the use of critical materials; a carefully conducted training phase; rigorous double blind testing conditions; the use of a listener-controlled, seamless switching playback system; performance-based listener selection; and within-subject experimental designs.

We believe that others can use our methods for sensitive evaluations of any high quality audio system or device. More details about these methods are available in a recent report listed below.

REFERENCE

Grusec, T., Thibault, L. & Beaton, R. Sensitive methodologies for the subjective evaluation of high quality audio coding systems. *Proceedings of the AES UK DSP conference*, London, 14-15 September 1992.