Live Listening Tests @SoundExpert

Okt 2007

Actually, most people don't care much about sound quality they hear from stereos, computers and mp3 players. This is normal. To be honest, Music itself is so great it can impress even through poorly sounding equipment. There is very similar situation with movies. Most home TV sets have equivalent resolution of 1920×1080 pixels at best. Hiding a lot of details and nuances of moving pictures they still afford pleasure to viewers. And now try to remember your last visit to Cinema Theater. How about that feeling of almost reality made by huge, detailed and quality picture? No doubt audio installation of high definition will do the same for your music. Even good headphones and mp3 players make difference.

But how to tell the right audio devices from bad ones. And what those *good* and *bad* mean for each particular listener. Unfortunately there are no simple answers to those questions. May be the easiest solution would be to ask for advice from some experienced listener who knows your musical and listening tastes. If you are not lucky to have such acquaintance the only way out is to listen, listen and listen yourself. And, please, don't pay too much attention to technical parameters such as Total Harmonic Distortion (THD), Signal-to-Noise Ratio (S/N), frequency response and others. Some relationship between those parameters and perceived audio quality can be traced only for analogue audio devices such as amplifiers, speakers (headphones), tape recorders, vinyl players and so on. For CD and mp3 players, sound cards/processors and other digital equipment this relationship is minimal. For all technologies of lossy audio compression such as *mp3*, *aac*, *wma*, *ogg* etc. the relationship between any measurable technical parameters and perceived sound quality is absent as a matter of principle. So once again – you have to listen.

It seems simple – just listen and compare. But in real life you can't compare all contending devices at once. Probably you listen one at friend's home, a few in specialized shops and one or two at exhibition if lucky. Taking into account different listening environments and different music used, real quality comparison is just impossible in the case. But even if we imagine some ideal comparison situation with all the contenders in one place playing the same music, appearance of device, its price and brand name will affect your listening impression greatly weather you want it or not. In order to be meaningful any quality comparison has to be *blind*. You must not know what particular model sounds at the moment.

Now you understand how serious the problem of choice on audio market is for those who want real quality. The core of the problem is lack of knowledge and information. Thousands of professional consultants, experts and advisers in audio salons and magazines earn their money on the problem. Unfortunately for consumers distinction between consulting and advertising in this business is completely eliminated. The most appropriate word is *manipulation*. We intentionally avoid discussion of "High-End" phenomenon here because of its more psychological nature rather than technical or even musical.
But of course, the scope of this extremely overpriced phenomenon would be much smaller if reliable and inexpensive methods of measuring audio quality exist.

The situation began to change since appearance of the internet with its unprecedented possibilities of information interchange. Now if we want we can easily get help from different advisers and compare their opinions. We can directly talk with manufacturers of audio equipment we are interested in and with people who have already purchased it. New on-line audio magazines less depend on manufacturers' advertising budgets and more depend on consumers' attention. Their reviews are more trustable. And this is not the whole story yet.

The internet made it possible to combine efforts of millions of people for solving various extremely complicated problems. Such internet projects are called Distributed Human Projects (DHP). Most known are:

- SETI@home – Search for Extraterrestrial Intelligence by means of internet-connected computers of ordinary users; pioneer of DHPs
- fightAIDS@home – designing new drugs to fight AIDS
- distributed.net – cracking data encryption schemes
- DIMES – studying the structure and topology of the internet
- DMOZ – building the most comprehensive human-edited directory of the Web
- Wikipedia – collaborative project to produce a complete encyclopedia; a good place to start learning about DHPs themselves
- FreeDB - database that stores meta data about music CDs; anyone can contribute information about new CDs
- P2P – peer-to-peer networks can also be considered as DHP projects; they help to establish very efficient distribution of a digital content over the internet by sharing storage capacities and bandwidth of connected computers

SoundExpert is one of the kind. It combines efforts of people who are interested in objective picture of sound quality of various audio devices and technologies on the market. To be more specific it combines peoples' hearing abilities in order to rate perceived audio quality of those devices. SoundExpert (SE) service is completely interactive; its results – ratings – are calculated in real time and immediately available for all interested in. In other words, ratings are created by consumers and for consumers. In order to take part in testing, it suffices to download a short sound file, listen it and feed back a grade. As participants don't know what particular device they test the whole procedure of testing is blind and thus final ratings of

perceived audio quality are objective. Only sound quality matters for SE ratings calculation.

Today SoundExpert maintains ratings of various lossy audio codecs like mp3, aac, wma and others at different bitrates – more than 100 testing items at the moment. First few portable players were added not so far ago but for sure this group of devices is the most interesting for both SE researchers and audio enthusiasts – visitors of the site. And yes, you can propose device for testing because the service is yours and for you.

Inside SoundExpert

As you might have guessed all audio devices and technologies are tested the same way at SoundExpert. Let's call any of that codecs, players etc. by the term commonly used in special literature – *device under test* (DUT). As principles of listening tests are well developed and widely used for measuring audio quality in professional area they formed the basis of SoundExpert testing methodology but were adapted to special conditions of interactive on-line testing with broad participation of non prepared listeners.

Nine short sound samples are used for testing. They represent different types of real-life audio material:

1. J.S.Bach, "Easter-Oratorio" (symph. orchestra) [BAH]
2. Bass (singing voice) [BAS]
3. Castanets [CST]
4. French Male Speech [FMS]
5. Glockenspiel [GLK]
6. Harpsichord [HRP]
7. Postscriptum, "You were here" (lo-fi old tape recording) [LOF]
8. Mike Oldfield, "Music From The Balcony" (electronic music) [MOF]
9. Quartet (singing voices) [QRT]

Most of these reference samples are from SQAM (Sound quality assessment material) disk issued by European Broadcasting Union (EBU) specially for listening tests. Others are taken from high quality CD recordings. One reference sample (7) of low quality represents home made and digitized old tape recordings. More details and the samples themselves can be found on SE Sample page.

In general case the use of natural sounds is preferable in listening tests as human hearing system is more sensitive to natural instruments and human voice sounding. So any drawbacks of a sound reproduction system – device or technology - will be revealed on that sound material more likely.

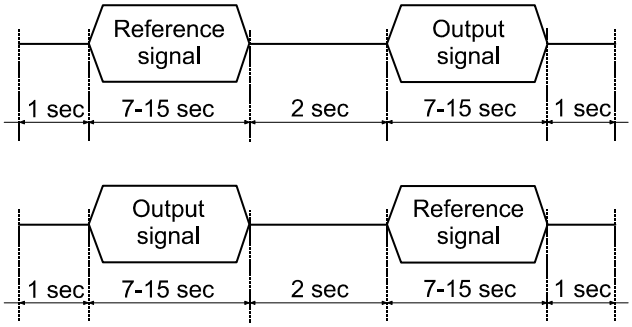
Operation of SE testing engine can be divided into three phases:

1. Preparation of test files
2. Testing by visitors of the site
3. Calculation and visualization of test results

Only the first phase - preparation of test files - is performed by human operator off-line. Other phases are interactive and controlled by the system in real time. Let's have a closer look at those phases.

1. Preparation of test files

At this point all nine test samples are fed into DUT and nine corresponding output signals are recorded digitally. For each of these nine testing items a pair of sound files is created as follows:



Double quantity of test files (18 in total) is necessary to hide the order of samples from listener. Every time downloading a test file it is unknown which sample, reference or output, is the first and which one is the second. This is one of blind test requirements.

Each sound file (*se_test.wav*) is packed into *zip* archive together with *readme.htm* file. Resulting zip archive weights 1.5 – 3.5 Mb depending on sound sample. The whole group of test files corresponding to one DUT (18 files) is added to global rotation of test files in the system. From this moment newly added test files are ready for downloading by visitors of SoundExpert. Usually all DUTs in the rating system are available for testing concurrently. During a dedicated listening test some group of DUTs may be tested exclusively. For example during *128 kbit/s listening test* files of coders@128kbit/s are downloaded and graded only. Such special listening tests are clearly announced on SE main page.

2. Testing by visitors of the site

After clicking on *Download a test file* link the system randomly chooses one of the test files being rotated at the moment, assigns it a random name and gives it out for downloading.

Probability of choosing one file over another depends on number of grades already returned to those files. The more grades – the less probability of the file choosing. In that way the system performs uniform testing of all DUTs and produces ratings of approximately equal *ripeness*.

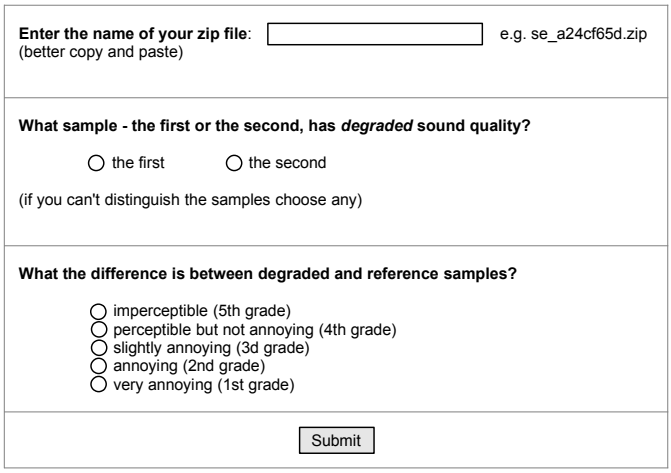
Test file offered by the system may look, for example, like *se_d7e92a58.zip*. It could be downloaded as usually with any browser or download manager. IN NO WAY it can be identified with particular device under test (even time stamp is fictitious). This is the most important condition for blind testing.

Listening test itself is very simple. While playing *se_test.wav* try to determine which one of two sound samples – the first or the second – has degraded sound quality and what is the level of that degradation.

It is recommended:

- to use headphones for listening
- to switch off all sound enhancers like SRS or Dolby
- to set equalizer to *flat* position
- to set the volume you normally listen your music

You might need to listen your test file for several times until you can answer the questions. Also you may fail to distinguish two samples. Don't worry, this result is valid as well. You can feedback your answers filling the form in *readme.htm* file:



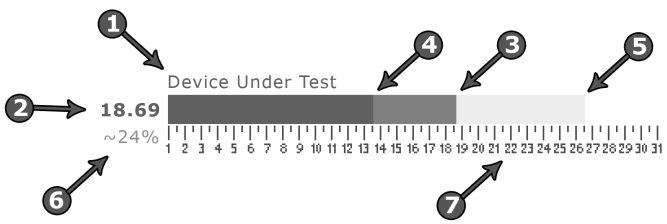
Please note that the system needs exact name of your test file in order to *unblind* the DUT you've tested. After filling all fields of the form and pressing *Submit* button you'll get confirmation page with a short summary of your listening session along with the name of device you've just tested. If SE refuses to accept results you will get the page with some possible reasons of refusal.

And finally, sending back your results to SoundExpert, don't be afraid of mistakes. The system has strong fool-proof and anti-kidding mechanisms. Deliberately false grades won't spoil resulting ratings but can prolong testing period. On the contrary more accurate grades help to achieve reliable ratings faster.

3. Calculation and visualization of test results

For each DUT SE calculates nine different *local* ratings corresponding to the number of test sound samples. It is important to notice that DUTs may produce audible artifacts on some sound samples and be transparent on others. This is most typical for lossy codecs but other DUTs also show similar behavior to less extent though. Final rating is an average of those nine local ratings.

After you have sent back your grade the system determines (*unblinds*) the DUT the test file belongs to, recalculates corresponding local rating, recalculates final rating and displays new rating bar for the DUT:



1 – Device under test (DUT) - device or technology being tested.

2 – Value of actual perceived audio quality rating which is indicated by bar length 3. Anchor points could be interpreted as follows:

In most cases using this device/technology:

1.0 – you will hear heavily distorted sound
2.0 – you will hear unpleasant sound artifacts
3.0 – you will hear distinctly audible but tolerable sound artifacts
4.0 – you will hear faintly discernible sound artifacts
5.0 – you will not hear any sound artifacts
above 5.0 – all sound artifacts will be beyond threshold of human perception with corresponding perception margin; see SARTAMP chapter for details

4 and 5 – The "high" and the "low" of a rating. As each device is tested under nine different sound samples, there are nine different local ratings for a device. In fact, the actual rating 2 is an average of those nine local ratings. The highest and the lowest ones are indicated. Big gap between them means that sound quality of device/technology is not consistent enough. It will vary with type of sound material: music of different genres and complexity, voice with or without music, noisy/clear recordings etc. The lowest local rating is more important in this sense as it indicates worst case behavior of DUT.

6 – Accuracy of rating. It is also indicated by the color of bar - more accurate ratings have less percentage values and darker bars. Accuracy depends on number of grades returned by participants. In most cases 5% or less is OK.

Rating of a DUT is recalculated every time a test file of this DUT is graded by participant. SE system keeps last *n* values of a rating and computes accuracy *Er* [%] as relative width of error tube:

$$Er = \frac{R^n_{\max} - R^n_{\min}}{R} \times 100$$

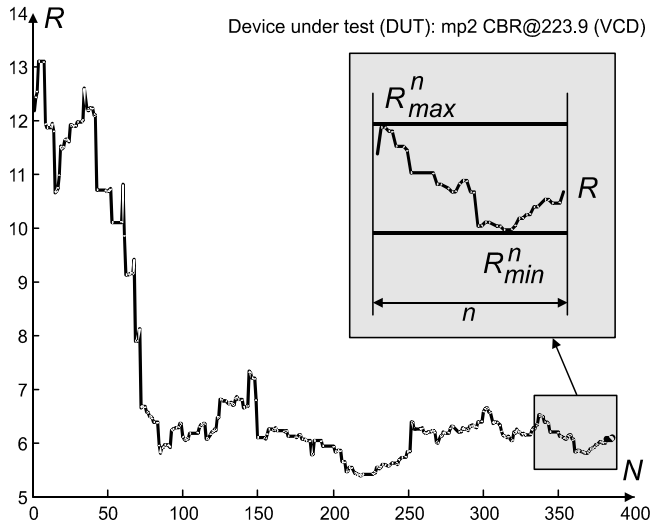
n – number of last rating values (length of error tube); now it equals to number of testing files for DUT: if DUT is tested naturally (without SARTAMP technology) *n = 18*; in case of using SARTAMP number of testing points is higher and it takes more grades to be returned by participants in order to achieve steady-state value of rating; typical *n* for the case is *54* (three testing points for each sound sample).

Rmax, *Rmin* – max and min values among last *n* ratings of DUT.

R – current (the latest) value of DUT rating

In short, accuracy parameter shows fluctuation of rating over last *n* grades returned for its DUT. While number of returned grades increases, fluctuation decreases and rating tends to its final value as it is shown in the Diag.1.

7 – Ruler for convenient measuring of highs and lows.



Diag.1: The rating gradually tends to its final value

Sound artifacts amplification (SARTAMP)

Described above simple method of ratings calculation works fine for all DUTs that produce sound artifacts audible to an average listener – typical participant of SE testing. Good example of such DUT is a lossy codec at bitrate below 100 kbit/s. Audio devices with less audible artifacts such as high bitrate codecs, sound cards and audio tracts of portable players will most likely be graded with *five* by majority of unprepared listeners. Having the same five point ratings all these devices would seem equal to SE system. In common practice in order to rate such devices special listening tests are prepared. Those tests have to meet certain conditions like controlled listening environment, prepared (trained) listeners, high definition audio installation and some others. Obviously there is no possibility to satisfy even some of those conditions in any internet testing. For the purpose standard testing methodology was improved in SoundExpert. This made possible to use ordinary hearing abilities and usual sound equipment of an average internet user for testing and rating of high quality audio devices. The main idea of the improvement is to arm human auditory system with a kind of *sound magnifying glass* which helps to discern even the slightest drawbacks of sound.

Let's resort to sound/image analogy once again. Algorithms of lossy compression for images like JPEG were developed earlier than the ones for sound and widely used nowadays. But even the slightest compression artifacts are easily seen with appropriate zooming of image. In other words there is a simple instrument for estimation and comparison of image artifacts. Similar instrument for sound artifacts is just non existent.

Zooming of waveforms/spectrums/sonograms in audio editors can't serve the purpose because there is no direct relationship between visual representation of a signal in time/frequency domain and auditory image the signal produce in brain. Human hearing is much more complicated than oscilloscope or spectrometer.

This is one of the reasons why devices/technologies for video/images playback are less debatable than the ones for audio. The phenomenon is most noticeable when we compare discussions on video and audio lossy compression. Have you ever seen heated debates on choosing appropriate format and level of compression for storing photos on PC/iPod or about video bitrate for ripping DVDs into DivX? Usually there is little to discuss – nature and size of image artifacts can be clearly seen, measured or compared at any level of compression.

Similar approach is taken in SoundExpert for hearing and measuring of sound artifacts. The technology is called *Sound artifacts amplification* (SARTAMP) and applied to DUTs that introduce low distortion or almost transparent for sound signal. Output signals of those DUTs are specially processed – sound artifacts are amplified to the extent they become audible to an average listener. Corresponding ratings are computed in a way that not only returned grades but also amplification gain (*zooming factor*) is taken into account.

Artifacts amplification can be applied either to all test items or to selected ones depending on audibility of artifacts for each test item. For example artifacts of some DUT could be easily audible on *glockenspiel* and *castanets* samples and not noticeable on others. Then SE test files with glockenspiel and castanets will be produced using natural output signals; others will be produced using output signals processed by SARTAMP. In case of artifacts amplification each natural output signal originates (at least) three processed signals where artifacts are amplified to different extent thus tripling the number of required listening tests (see supplementary diagram for illustration, pdf: 980k). In other words SE testing methodology substitutes one complex task where *golden ears are required* for multiple tasks of lower complexity where *ordinary ears are sufficient*. Decision about using SARTAMP for each particular test item is taken by human operator preparing test files at his solo discretion. Numerous listening tests at SE showed that SARTAMP can be safely used also for test items where artifacts are clearly audible – resulting ratings stay the same but require more listening sessions. So, operator's choice doesn't affect final ratings but helps to lessen number of listening sessions only.

Basically sound artifacts amplification includes three operations: time/phase alignment of reference and output signals, subtraction of some portion of reference signal from output one and post-filtering of resulting difference signal in order to eliminate *parasitic* frequency components. Amplification gain is controlled by objective parameter *Difference level* which could be considered as an extension of THD parameter for non-periodic signals. More strict and detailed description of the technology and the method of ratings calculation can be found in AES paper: Difference Level. An objective audio parameter. (pdf: 694k)

The best and most natural way to go into SARTAMP technology is to hear it by yourself. You can download the package below containing four sound files (44100, 16bit, stereo):

1_ref.wav – reference sound signal (harpsichord excerpt)
2_out.wav – output of mp3 coder (CBR@192kbit/s, decoded)
3_out+6db.wav – the same output with artifacts amplified to 6dB
4_out+12db.wav – the same output with artifacts amplified to 12dB

Download the package (zip: 4.56 Mb)

The samples are self-explanatory; they show what this technology could be helpful for.

In case of artifacts amplification, analytically computed ratings are usually located above 5th point on impairment scale, indicating that the artifacts are beyond the threshold of human audibility. This could be interpreted as quality margin or quality headroom of audio device. You may ask what the purpose of the margin is if sound artifacts are no longer audible. There are four reasons at least why this is important:

- In general case perceived audio quality of a device/technology depends on sound samples used for testing. In ideal case a listening test has to be performed with infinite number of sound samples in order to prove for sure that tested device will not produce unexpected *surprises* on real-world audio material. In practice a limited set of typical or problem (*killer*) sound samples is used. Then testing results are just generalized on all audio. Obviously quality margin makes that generalization more grounded and lessens the probability of getting artifacts on audio material not used during the test.

  Well known technical audio parameter – THD is used quite similarly: measured on pure sine wave and poorly corresponding to perceived audio quality it has to be very low (far beyond human abilities to hear such low distortions of pure waves) in order the equipment to sound acceptable. In other words sound equipment must have substantial quality margin on sinusoidal signals in order to behave smoothly on real world audio material.

- Very often audio devices/technologies are used in chains – connected one after another. In most cases this accumulates sound degradation throughout the chain. Quality margin of each device is highly desirable in order to lessen overall distortion level.

- Such post processors as equalizers, spatializers, SRS and many others usually reveal sound artifacts inaudible without them. Some quality headroom helps to use all those popular sound enhancements safely without danger of discovering drawbacks of other audio components.

- Human hearing abilities differ from person to person. Averaged results of any listening test have to be applied with great caution to someone's particular situation, especially if that someone has *golden ears*. Such person needs audio equipment with greater quality margin in order to be satisfied.

- In some cases hearing abilities of a person are improved with the lapse of time. For example after curing of hearing disabilities or due to beginning of healthier life style. While old audio equipment could be easily replaced in such cases, it's wise to have some audio recordings (especially the rear ones) with substantial quality headroom initially.

Whether we like it or not the quality margin exists objectively, just because any sound equipment/technology can be improved almost infinitely far beyond abilities of human hearing. And some parameter measuring this quality margin is necessary. The only problem is choosing appropriate audio metrics and corresponding measurement technology.

SE testing mechanism in combination with SARTAMP technology makes it possible to test and rate even reference quality audio equipment easily without assistance of listening experts with extraordinary hearing abilities. Sound artifacts in SE test files are more or less audible to non-prepared listeners and the main goal of SE listening tests is to *grade* the annoyance of those artifacts, not to *catch* them. So any internet user with headphones and soundcard can participate. You can make sure of this right now - here is a link to a test file of current SoundExpert listening test. You can download it, grade it and … become an expert for a moment!

Visit SE Testing Room and download a file
(zip: 1.5 – 3.5 Mb)

Keeping audio clear together

As you see, participation in SE listening tests is really easy and entertaining. SounExpert tirelessly collects all listeners' contributions and grade by grade creates objective picture of perceived sound quality of various devices and technologies. Having that picture in mind audio consumers can make deliberate, less emotional purchasing and other decisions related to sound quality. In turn, audio manufacturers and music producers will get stronger feedback about audio quality consumers prefer. Who knows, may be production side will start to pay more attention to Research and Development, where sound quality is actually made, and not to Marketing Departments where that quality is just fabricated. After all, the matter concerns listener's satisfaction and Music as an art.

Links

- Supplementary diagram: SE testing engine (pdf: 980kb)
- More papers about SoundExpert
- SoundExpert web site