A LISTENING TEST ENVIRONMENT FOR SUBJECTIVE ASSESSMENT OF SPEECH AND AUDIO SIGNAL PROCESSING ALGORITHMS

Magnus Schäfer, Christopher Schnelling, Bernd Geiser, and Peter Vary
Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University
{schaefer|schnelling|geiser|vary}@ind.rwth-aachen.de

Abstract:

Despite the availability of objective measures of audio quality, listening tests are still indispensible for a thorough evaluation of new speech and audio signal processing algorithms. Compared to objective measures, they offer great flexibility and can be applied for a multitude of questions at the price of an increased demand in time and effort. Many of the commercially available systems rely on specialized hardware and are far from being platform independent.

The LIStening Test ENvironment (LisTEn) for the subjective assessment of speech and audio signal processing algorithms is presented. This system allows to conveniently set up different types of listening tests (e.g., absolute category rating (ACR), comparison category rating (CCR), and degradation category rating (DCR)) in accordance with the procedures that are, e.g., standardized by the ITU in [2, 3]. Other types of listening tests that are readily available in LisTEn include ABX, rhyme and multiple stimuli with hidden reference and anchor (MUSHRA) [4] tests.

The setup of an example listening test is presented to illustrate the ease of use of the presented LIStening Test ENvironment. This presentation is accompanied by general remarks on setting up a listening test and the motivation behind the choice of various parameters.

A test version of LisTEn is available on www.ind.rwth-aachen.de/listen.

1 Introduction - Listening Tests in General

A necessary task when developing speech and audio processing algorithms (e.g., coding, noise reduction, dereverberation, echo control or bandwidth extension) is the evaluation of the performance of the new algorithm. For this, two basic paradigms are usually applied: listening tests or instrumental measures.

Listening tests are a well-known tool for the subjective evaluation of signal processing algorithms. In order to conduct a listening test, four elements are necessary:

- Test items

These test items are usually short pre-processed audio files that are played back in a defined environment. The exact choice of test items depends on the type of listening test and obviously on the conditions that shall be tested.

- A set of questions that have to be answered or tasks that have to be fulfilled

Again, the exact questions and tasks strongly depend on the system under test and could ask for very general things as well as for very special aspects of the algorithm that is tested.

- A number of participants that answer the questions or fulfill the tasks

The results of listening tests are always of a statistical nature which makes it preferable to have many participants. On the other hand, increasing the number of participants increases the time and effort necessary to conduct the listening test. An example for a study on the numbers of participants necessary to reach reasonable statistical significance can be found in [8].

- Clear and unbiased instructions for the participants

Listening tests should be carried out as unbiased as possible. This can be ensured (partly) by a good technical setup (i.e., low background noise, comfortable loudness for all test items, etc.) and a careful design of the test procedure (i.e., no suggestive questions, randomization to avoid sequence effects, etc.). Finally, the instructions for the participants play an important role in this regard, the conductor of the listening test has to ensure that the instructions are identical for all participants (achievable fairly easily by written instructions) and that the instructions are unbiased.

Even though several approaches have been made in the past to replace time-consuming and expensive listening tests by instrumental measures, no generic measure could be developed so far that can replicate the flexibility of a listening test or evaluate all the different aspects that can be addressed by a listening test. These instrumental measures range from simple quantities like the signal-to-noise ratio (SNR) to complex systems that try to model human perception based on a multitude of features (e.g., Perceptual Evaluation of Speech Quality (PESQ) [5] or speech transmission index (STI) [6]).

The development of these instrumental measures is also based on extensive listening tests. The results of the listening tests are then correlated with features of the test signals and a (possibly multidimensional) regression is applied to obtain an overall objective quality score.

2 User Interface of LisTEn

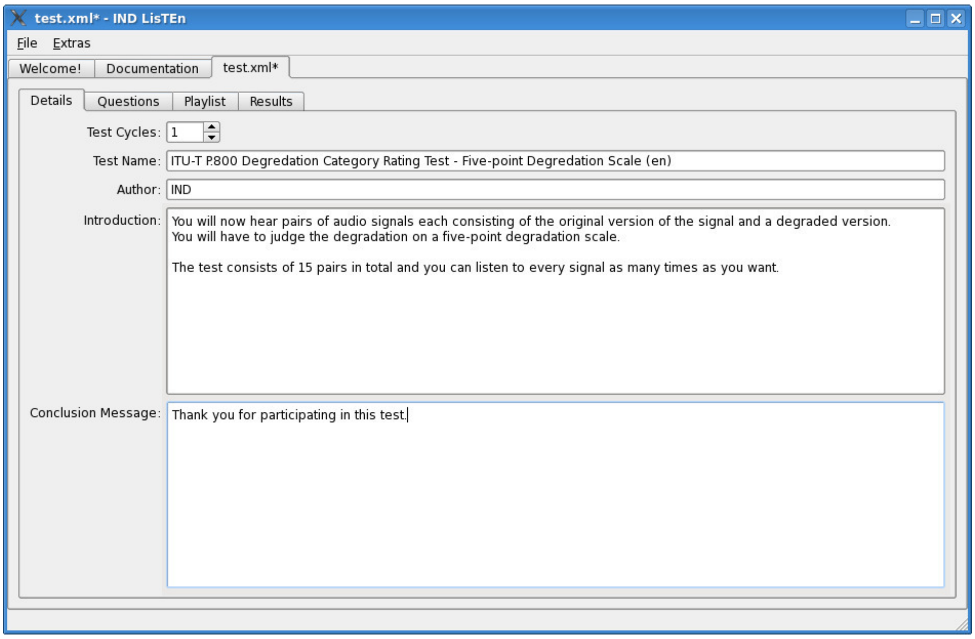
The underlying concept of the presented LIStening Test ENvironment (LisTEn) is a clear separation of the interfaces for the person conducting the listening test and the person participating in the listening test.

The user interface for the conductor is fairly static, the only variable elements are the number of questions and the number of playlists. The number of questions depends on the design of the listening test while the number of playlists depends on the type of listening test.

The basic setting screen for the conductor of the listening test is shown in Fig. 1 for a DCR test according to ITU-T P.800.

Since the DCR test is a comparison test, two playlists are necessary, the playlists and the corresponding options can be seen in Fig. 2.

The layout of the graphical user interface (GUI) for both the conductor and the participant is generated dynamically from the chosen test and the chosen number of questions. There is no end user action necessary in order to get a usable layout for the listening test.



Test Cycles: 1

Test Name: ITU-T P.800 Degredation Category Rating Test - Five-point Degredation Scale (en)

Author: IND

Introduction: You will now hear pairs of audio signals each consisting of the original version of the signal and a degraded version.
You will have to judge the degradation on a five-point degradation scale.

The test consists of 15 pairs in total and you can listen to every signal as many times as you want.

Conclusion Message: Thank you for participating in this test.

Figure 1 - Primary setting screen of LisTEn

3 Implementation and Templates

The software toolbox is implemented in C++ [7]. The Qt framework [1] is used for the GUI. Hence, it can be used on practically any computer platform that may be of interest for the realization of listening tests. In Fig. 4, an overview of the basic structure of LisTEn can be seen. The Implementation is separated into two parts:

- The core program IND LisTEn takes care of the settings, the playlist, the results and also manages the interface to the hardware.
- The choice and configuration of the listening test is conveniently done by XML files that can be generated by a wizard. Various XML templates provide ready-to-use skeletons for a number of standardized tests:

3.1 Tests according to ITU-T P.800

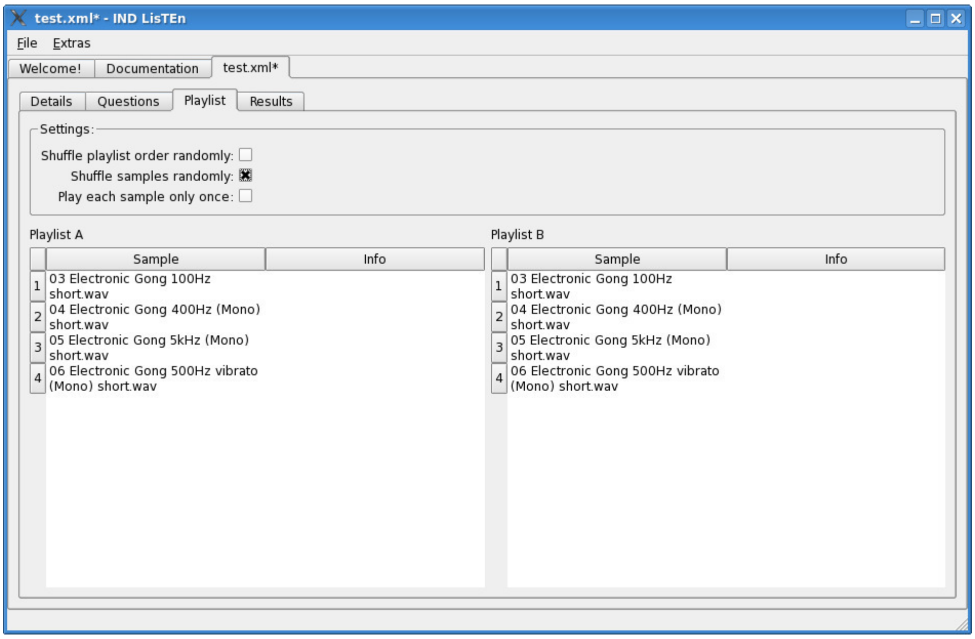
3.1.1 Absolute Category Rating (ACR) Tests

Within this test scenario the subject is asked to rate several test items individually against a fixed rating scale. This test method is used to obtain absolute, subjective opinions on the test items and is a rather general approach. The recommendation defines three different scales:

- Quality of the speech: e.g., Excellent - Good - Fair - Poor - Bad
- Effort to understand the meanings of sentences
- Loudness preference

3.1.2 Degredation Category Rating (DCR) Tests

To provide more detailed results, a comparative method is suggested. This test presents stimuli in pairs to the listeners, the first being a quality reference test item, the second being the same



Settings:

Shuffle playlist order randomly: □
Shuffle samples randomly: ☒
Play each sample only once: □

Playlist A

	Sample	Info
1	03 Electronic Gong 100Hz short.wav	
2	04 Electronic Gong 400Hz (Mono) short.wav	
3	05 Electronic Gong 5kHz (Mono) short.wav	
4	06 Electronic Gong 500Hz vibrato (Mono) short.wav	

Playlist B

	Sample	Info
1	03 Electronic Gong 100Hz short.wav	
2	04 Electronic Gong 400Hz (Mono) short.wav	
3	05 Electronic Gong 5kHz (Mono) short.wav	
4	06 Electronic Gong 500Hz vibrato (Mono) short.wav	

Figure 2 - Playlist and playlist settings

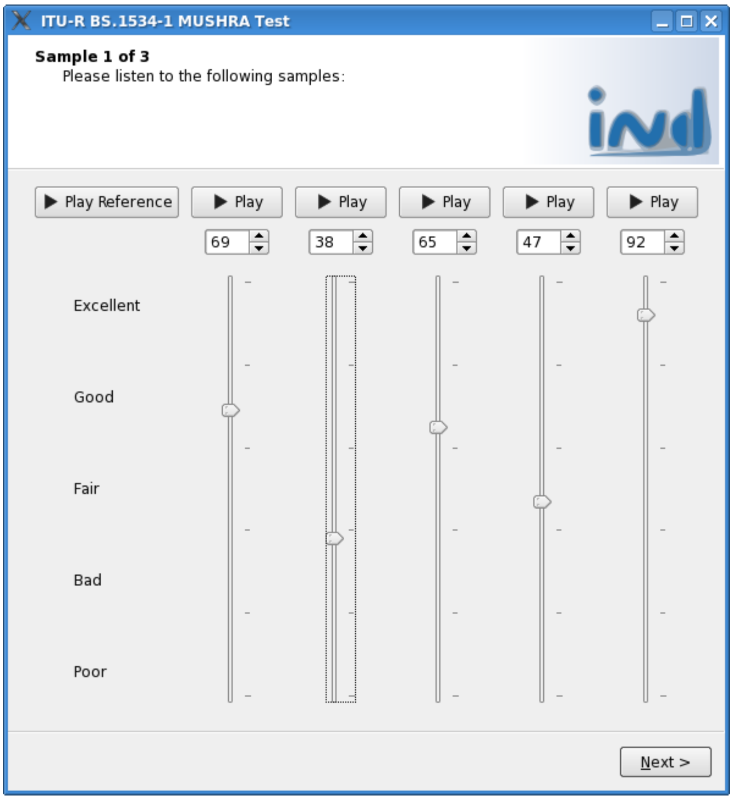
test item processed by the system under test. The subjects are asked to rate the pairs in a comparative manner on a five-point degredation scale:

- Degredation is inaudible
- Degredation is audible but not annoying
- Degredation is slightly annoying
- Degredation is annoying
- Degredation is very annoying

3.1.3 Comparison Category Rating (CCR) Tests

Being similar to the DCR method mentioned above, the CCR method asks for a comparative rating for a pair of stimuli as well. In contrast to a DCR test, in which the second test item is always degraded relatively to the first one, there is no quality ranking that is known in advance here. Hence, the order of the test items is chosen randomly in each trial and the rating scale must provide a neutral option here.

- Much Better
- Better
- Slightly Better
- About the Same (Neutral Option)
- Slightly Worse
- Worse
- Much Worse



ITU-R BS.1534-1 MUSHRA Test

Sample 1 of 3

Please listen to the following samples:

69 38 65 47 92

Figure 3 - User interface of a MUSHRA test

3.2 Multi Stimulus Test with Hidden Reference and Anchor (MUSHRA) according to ITU-R BS.1534-1

To obtain fine-grained results, a more detailed rating scale is necessary, providing a possibility to identify small impairments of the presented stimuli. In each trial, an unprocessed reference signal is presented as such to the subject along with a list of processed versions of this signal. Within this list the unprocessed reference is presented as a hidden reference once again and a hidden anchor is presented as well, being, e.g., a low-pass filtered version of the reference signal. The subject is asked to rate the test items on a numerical scale within the range of 0 to 100, as shown in Fig. 3.

4 Example Listening Test

In order to get an idea of how to conduct a listening test with the presented system, a short overview on the preparation and realization of an example listening test for the evaluation of a new general purpose speech and audio codec is given. In our scenario, there are two candidate proposals and the target of the listening test is to choose the better one.

By using LisTEn, the normally cumbersome preparations of a listening test reduce to a few simple steps:

4.1 Designing the Listening Test

The basic decision to make is the type of listening test that will suit our needs best. Since we have to compare two proposals directly, the obvious choice is the CCR test as defined in [2].

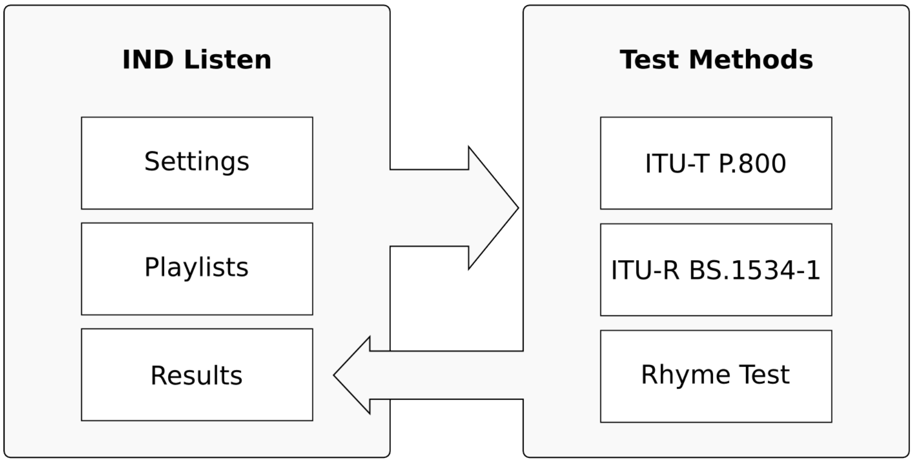


Figure 4 - Overview of the LisTEn system

This test has to be parameterized depending on the specific needs of the test. These parameters have to be set:

- Questions that are asked for each file (pair) with their answers

This is the core of every listening test because the questions are the means that are available to the conductor of the listening test in order to get exactly the information from the participants that are needed for a meaningful evaluation of the algorithm that is tested.

For the standardized tests like the ITU-T P.800 CCR, the questions and answers are already set within the XML templates that are provided by LisTEn. This test gives the participant the possibility to judge the performance of the two proposals on a fine seven-step scale ("The quality of the second audio file compared to the quality of the first is") ranging from "Much better" over "About the same" to "Much worse". This suits our needs with this test so we keep the standard version.

- Number of repetitions

A higher number of repetitions can be advantageous since it allows to reach higher statistical significance with a lower number of participants and also to get reliability information on the participant. A reliable participant will give identical answers on all repetitions while the answers of an unreliable participant will differ from repetition to repetition.
However, a larger number of repetitions will also increase the time that is necessary for every participant and usually the listening test is designed not to exceed a certain time per participant in order to avoid fatigue. Hence the only choice is between having a number of audio files tested twice or having twice the number of audio files tested once.

In our example listening test, we are looking at a system that is not designed for one specific use case but at a system that should work under all circumstances. Hence we choose to play every pair of audio files only once to maximize the number of different signal types for our test.

- Some text items (title and author of the test, introduction, and afterword)

Especially the introduction plays an important role to ensure identical conditions for all participants by having identical introductory explanations.

For our setup, one might give a short explanation on the background of the test and then give the instructions to the participant not to hear for some specific feature of the two signals that are to be compared but to judge the overall quality.

- Additional questions can be defined

For some applications, it is convenient to ask multiple questions for every audio file, e.g., one question about one specific feature and one question about the overall quality.

In our comparison, we are only interested in the final and overall verdict between the two candidate proposals, so we do not need additional questions.

- Fixed or random order of files or file pairs, respectively

Usually, a randomization is advantageous for most cases since it is the easiest way to avoid sequence effects. However, having a predefined order of the audio files can be necessary for certain tests where these sequence effects are explicitly part of the test.

We do not want sequence effects for our test, thus we randomize both the order of the file pairs and the order within the file pairs (i.e., sometimes the first audio file is candidate proposal A, sometimes, it is proposal B).

- Listen once or listen ad libitum

The choice for this parameter depends mostly on the clarity of the differences between the audio files. Allowing only listening once will significantly speed up the listening test and lead to a larger sample size for the statistical analysis. On the other, repeated listening will allow the participant to find even small differences between the audio files and hence lead to more precise and more reliable results.

There are good arguments for both choices here but since we are mostly interested in the real-world applicability of the proposed codecs, we opt for only allowing listening once to get a large sample size of first impressions of possible users of the codec.

4.2 Preparing the Audio Files

The audio files for the listening test have to be processed by the system that shall be evaluated and made available to LisTEn. Depending on the type of test, more than one version of every file is necessary, e.g., in a DCR test, a reference signal (e.g., the clean original) and a degraded version (e.g., the output of a low bit rate speech codec) have to be provided.

The audio files for LisTEn can be conveniently provided by just putting .wav-files for each playlist into one folder. The audio player is very flexible with respect to sampling rate, number of bits per sample, etc. since it gets all the necessary information from the file header

In our example, we need the processed files from both candidate codecs so we encode and decode the same set of test files with both proposals and put them into separate folders for the listening test.

4.3 Setting up the Testing Environment

After the test type with the corresponding parameters is set and the audio files are prepared, the only remaining task is to prepare the testing environment. For our comparison and for most other tests, a place with low background noise, little or no visual distractions and a good audio reproduction system is the best choice.

4.4 Evaluation of the Results

After enough people have participated and the listening test is finished, the results can be exported in a simple structured manner (one row per participant, one column per file (pair)) as comma separated values (.csv files) or MATLAB matrices (.mat files). Thereby, the evaluation and visualization of the results can be done by one of the many available statistical tools (e.g., MATLAB, Gnu R, etc.).

5 Summary

An overview on the newly developed LIStening Test ENvironment (LisTEn) was presented. It is a simple and flexible system for conducting listening tests according to various standards and also easily configurable to allow for specific setups and needs. It is implemented in C++ and Qt and hence platform independent.

The setup of LisTEn was exemplarily presented for an example listening test that could take place in the standardization process of a new general purpose speech and audio codec. The preparation and configuration was shown to be a simple process of setting appropriate parameters and letting LisTEn take care of the tedious details.

A test version of LisTEn is available on `www.ind.rwth-aachen.de/listen`.

References

[1] BLANCHETTE, J. ; SUMMERFIELD, M. : *C++ GUI programming with Qt 4.* Upper Saddle River, NJ [u.a.] : Prentice Hall [u.a.], 2006. – ISBN 0–13–187249–4

[2] ITU: *Methods for subjective determination of transmission quality (ITU-T Recommendation P.800).* International Telecommunications Union, Aug. 1996

[3] ITU: *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems (ITU-R Recommendation BS.1116-1).* International Telecommunications Union, 1997

[4] ITU: *Method for the subjective assessment of intermediate quality level of coding systems (ITU-R Recommendation BS.1534-1).* International Telecommunications Union, 2003

[5] ITU: *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs (ITU-T Recommendation P.862.2).* 2005

[6] STEENEKEN, H. J. M. ; HOUTGAST, T. : A physical method for measuring speech-transmission quality. In: *The Journal of the Acoustical Society of America* 67 (1980), January, No. 1, pp. 318–326

[7] STROUSTRUP, B. : *The C++ programming language - special edition (3. ed.).* Addison-Wesley, 2007. – I–X, 1–1020 S. – ISBN 978–0–201–70073–2

[8] WILLIAMS, W. ; COLIN-THOME, G. : Test Subject Numbers and the Performance of Hearing Protectors. In: *Annual Conference of the Australian Acoustical Society.* Gold Coast, Australia, 2004