WEB-based listening test system for speech synthesis and speech conversion evaluation

Laurent Blin, Olivier Boeffard and Vincent Barreaud

IRISA - Institut de Recherche en Informatique et Systèmes Aléatoires
Université de Rennes 1, Enssat, Lannion, France
{Laurent.Blin,Olivier.Boeffard,Vincent.Barreaud}@univ-rennes1.fr

Abstract

In this article, we propose a web based listening test system that can be used with a large range of listeners. Our main goals were to make the configuration of the tests as simple and flexible as possible, to simplify the recruiting of the testees and, of course, to keep track of the results using a relational database. This first version of our system can perform the most widely used listening tests in the speech processing community (AB-BA, ABX and MOS tests). It can also easily evolve and propose other tests implemented by the tester by means of a module interface. This scenario is explored in this article which proposes an implementation of a module for Comparison Mean Opinion Score (CMOS) tests and conduct of such an experiment. This test allowed us to extract from the BREF120 corpus a couple of voices of distinct supra-segmental characteristics. This system is offered to the speech synthesis and speech conversion community under free license.

1. Introduction

The quality of synthesised speech or speech conversion can be rated on various levels: its intelligibility, its naturalness and, in the case of voice conversion, its proximity to a target voice. The Blizzard Challenge (Black and Tokuda, 2005), for instance, compares corpus based speech synthesis systems with a set of tests designed to evaluate the speech quality on these levels. These tests are described as follows. To measure the similarity of two voices, the ABX test is frequently used (Duxans et al., 2004). For this test, the testee listens to three sentences (A, B and X) and decides which one of A or B is closest to X. Classically, X is a converted voice and A (respectively B) is the source (respectively target) voice (Kain, 2001). This test is widely used but has a fault: it cannot state if the converted voice is distinguishable from the target voice (which is the goal of voice conversion). In order to obtain this information, similarity tests (AB-BA, pair comparison) must be performed. For this test, the testee listens to pairs of voices saying different sentences and grades their similarity on a defined scale (Duxans et al., 2004). Finally, to measure the naturalness and the intelligibility of a voice, the Mean Opinion Score (MOS) test is used (Suendermann et al., 2005),(ITU-T, 1996). Other tests can be performed to assess the intelligibility of a voice, such as the *Standard Segmental Test*, or the *Diagnostic Rhyme Tests* (François, 2002),(Jekosch, 1993).

Setting up these tests faces some important logistical constraints: the gathering of listeners, the necessary audio hardware, the amount of time needed to realize the test and to process the results, etc. Therefore, making perceptive tests easier to perform would be help many research works. For instance, the Blizzard Challenge went beyond the geographical restraints using telephone communications for the testee sessions. In the same way, we propose a web based listening test system to ease such a task. The elements of the test (number of testees, audio stimuli to be used, etc.) are described with the help of a simple interface, while the testees can connect to the platform on the web and realize the test from anywhere at any time.

Section 2. of this document introduces the platform. Section 3. presents the software design, while section 4. describes a first experiment and its results.

2. System description

The proposed system is designed to accept a large number of listeners. The test scenarios are deliberately made extensible. Even if we propose the use of AB-BA, ABX and MOS type tests, the system can easily be extended to deal with other tests by adding the corresponding modules, as described in section 4. More classically, the test parameters (such as the required number of testees, the number of speech files to be tested, etc.) are totally configurable by the tester.

2.1. Test design

The system manages testees and tests through *panels*. A panel corresponds to a group of testees to which the system suggests different *sessions*, i.e. different versions of the same test (some different stimulis with the same evaluation goal). To avoid a bias in the results of the test, the size M of the panel is defined in the test configuration by its conceptor. It can be adapted to the number of test sentences available.

A session is considered *invalid* as long as the testee has not finished it. A test is not finished until all testees have completed their tasks, and statistical results can't be obtained until then.

A single test therefore is formulated as M sessions, each session being composed of N *steps*. A step is a question asked of the testee. It is composed of p stimuli (typically sentences) the testee must listen to (for an AB test, 2 stimuli are used; for an ABX test, 3 stimuli; and for a MOS test, 1 stimulus).

The order and nature of the stimuli used in a session can be chosen, randomly generated, or follow a Latin square. In the same way, the number N of test steps played to the testee during a session is configurable.

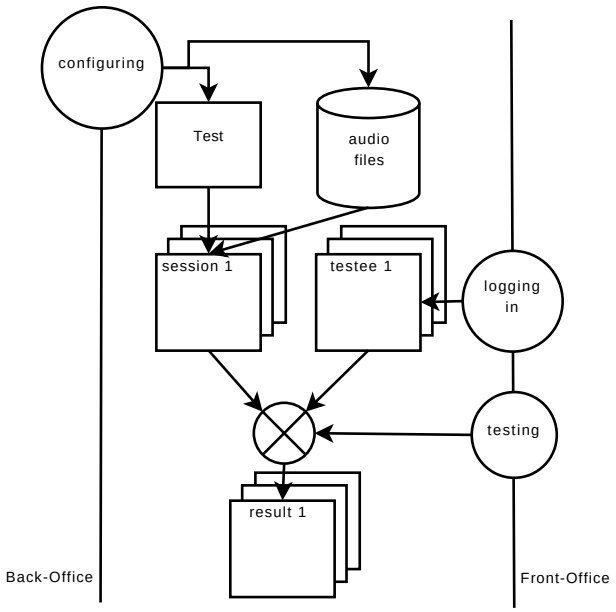

Figure 1: Platform workflow

Moreover, a testee can pause his session at anytime and resume it automatically, and, the system can ensure that two testees do not undertake the same session of a test, if that is required. Finally, the software is able to suggest more than one perceptive test, possibly of different kind, to the same testee, if several unfinished tests are available on the platform.

2.2. Testee access

The interface with the testees is also flexible. In order to reach a large number of testees, the interface of the system is web based, running inside a standard web browser. This allows the testee to work from his own computer when he is free, and it does not imply special software to be installed. Obviously, this kind of interface does not permit the tester to control the listening test conditions. That is why the testee is asked to provide information about his own listening conditions for them to be taken into consideration during the result evaluation if necessary. If accoustic test control is crucial, it is of course possible to gather the testees in a same physical place, providing each one with a computer to record his perceptive choises.
Whatever the listening conditions are, the system keeps track of testees, tests, sessions and results. This is done via a relational database. Flexibility is again built in the database, the number of tables and their structure being hidden behind an interface of simple accessor methods.

3. Software design

Figure 1 represents the posssible interactions of the tester and the testees on the System. This is done through two interfaces: the back and front office. The system software was written in Perl. The portability of this language, its platform independence and its easy integration in an Apache HTTP server were the main factors influencing this choice. The system is based on two widely used Perl modules: the CGI module for the web interface control, and the DBI module for the connection with the relational database management system. The way the software is implemented allows the system administrator to use any type of database.

3.1. Test configuration description

A test configuration file must describe a test dynamically. For instance, its structure must be independent of the (conversion or synthesis) system that generated the speech to be evaluated. Therefore, it must be possible for an operator to create it via a text editor . The grammar of this configuration file is very simple, it relies on the AppConfig Perl lib. This class controls the entry types of the configuration file. The configuration information includes:

- the name of the test author, date of origination, description of the test objectives;
- the test type (AB, ABX, MOS; expandable);
- the number M of requested testees to compose a panel;
- the number N of steps one testee should perform during a test session;
- the ordering of the stimuli (fixed, random or using graeco-latin squares);
- one (MOS tests), two (AB tests) or three (ABX tests) groups of addresses to the audio stimuli composing the test:
  - these groups are introduced by a brief description of the experiment which created this set of stimuli;
  - for AB and ABX tests, the stimuli of rank i in each group are to be submitted together during the test.

The test is then introduced to the platform back-office: the configuration file is analysed to feed and set the database. Figure 2 is a screen shot of the back office. On this screen, the tester can upload a new test and monitor ongoing tests (validity of panels, etc...). M sessions are then generated composed of N steps, each one confronting one or more acoustic stimuli (chosen among the ones provided) according to the test type. The first half of figure 1 illustrates this main back-office operation.

3.2. Testee interface

From the testee viewpoint, the system is accessible using a simple web navigator. The acoustic stimuli are transmitted using basic HTML 4.0 with minimum control options and no specific audio player defined (tested on Firefox 2.0 and Internet Explorer). Therefore the testee's default audio player is automatically selected. If no player is available, the onus is on the testee to install one.
This configuration results in easy installation of the system by the test manager and simple access to the tests for testees in diverse locations.
The second half of figure 1 illustrates the front-office operations, i.e. the testee interface. A testee can record his profile in the system if it's the first time he(she) accesses

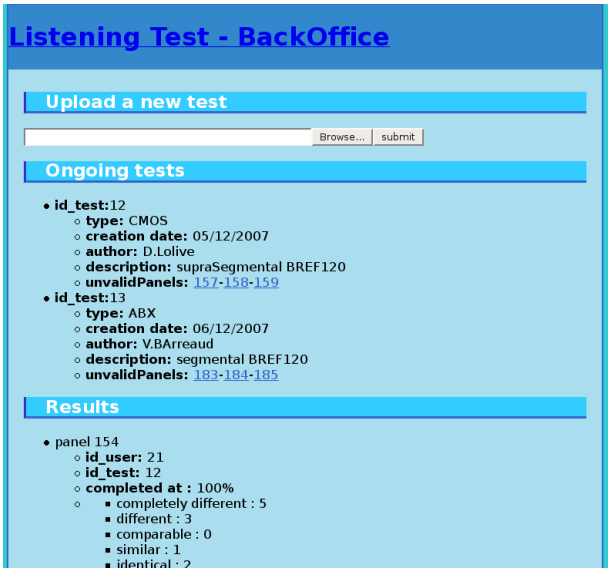

Figure 2: The tester can monitor ongoing tests and create new test instances

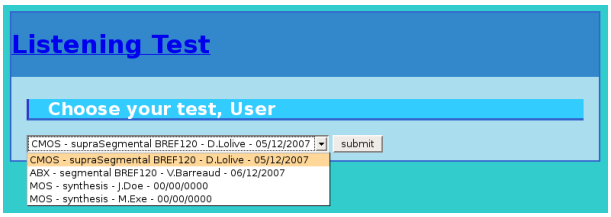

Figure 3: The platform proposes several tests to the listeners

the platform, or log in if he(she) returns. During the first operation, the testee fill in some personal data as his(hers) mother tongue, age, location, etc... Afterwards, the testee can perform one session of the tests selected by the system (see figure 3). Note that a testee cannot perform the same test twice.

The answers to each of the questions are recorded along with a reference to the current session and on the testee's identity. Figure 4 is a screen shot of question ask to the testee during a test.

When the number of valid sessions of a test reaches the

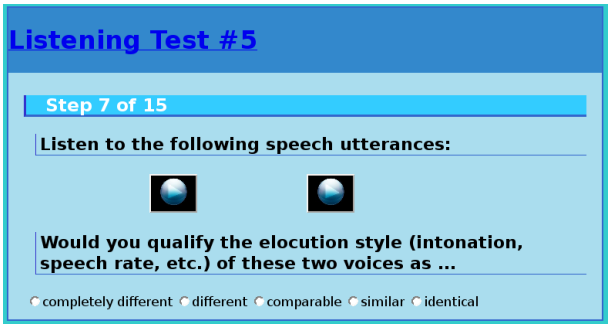

Figure 4: A step in the subjective evaluation

quota fixed by the author of the test, it is no longer accessible by the testees: the panel is completed.

3.3. Database content

The system database consists of six tables:

1. a *user* table for the testee references (identity, age, sex, listening conditions, etc.);
2. a *test* table for the test references (type, creation date, description, etc.);
3. an *audio* table for the acoustic stimuli references used in the tests (file path, experiment belonging, etc.);
4. a *panel* table to tie some user groups to the different available tests, and to record how complete each test is;
5. a *session* table to set the identity and order of the acoustic stimuli to be submitted to a given testee for a given test;
6. a *result* table to record the testee choices at each step.

The *test*, *audio*, *panel* and *session* table data are set during the test configuration file analysis, while the *user* and *result* table data are recorded at registration time and during the tests themselves.

4. Perceptive test experiment

This test platform has been used for a study we are conducting on speech conversion from a supra-segmental viewpoint. The goal of the test was to extract from all the pairs of speakers available in the BREF120 corpus (Lamel et al., 1991) the one showing the greatest prosodical difference among the largest amount of sentences. This pair of speakers should be used later in the study to evaluate supra-segmental conversion methods.

4.1. CMOS Test

The goal of the planned experiment does not exaclty fit the workframes of the common AB-BA, ABX or MOS evaluations. Indeed, our goal here is to assess the supra-segmental difference between two speeches (prosody, elocution rate, etc.). So we added a CMOS type of test (Comparison Mean Opinion Score) (ITU-T, 1996) to the platform with its associated Perl module. Its integration was made easy since this new module only had to implement a standardized interface. This interface gathers three functions. The first function manages the way the audio file paths mentionned in the configuration files are stocked in the database. The second function handles each step (the question to be asked to the testee and the possible choices). The last function exploits the results and presents them.

During a CMOS type test, the testees have to extract a pair of voices from a set of given pairs. The set is limited in size and described in the test configuration file (as illustrated by figure 5). For each session, the testee evaluates the dissemblance degree of the two voices being presented at each step. The question asked is: "*Would you qualify the elocution style (intonation, speech rate, etc.) of these two*

voices as...", and the testee values the voices distance on a 0 to 4 scale (0 for "*completely different*", 1 for "*different*", 2 for "*comparable*", 3 for "*similar*", 4 for "*identical*"). Each step presents two versions of the same sentence, the ones at the same i^{th} index in the two sentence lists of the test configuration file (list $[A]$ and $[B]$ in figure 5). The sentences' presentation order is randomly defined in the CMOS Perl module.

4.2. Test configuration

Figure 5 presents an excerpt of the test configuration file. We used audio files from the BREF120 corpus (Lamel et al., 1991), and the structure of the file directory names follows the original corpus structure. Female speaker IDs end with letter f as in `j8f`, while male speaker IDs end with letter m as in `i6m`. We pre-selected 40 voice pairs, with a distinction between native french speakers and non-native ones. The pair distribution of the selection is:

- 9 native male vs native male pairs,
- 21 native female vs native female pairs,
- 9 native male vs non-native male pairs,
- 1 native female vs non-native female pair.

The speakers implied in these pairs have been chosen among the 120 ones of the corpus, as long as they recorded at least 600 sentences. Male pairs and female pairs were composed with the speakers sharing the most sentences, and ranked following this criterion. Finaly, the 40 first pairs were selected for the experiment. For each voice pair, 5 sentences were randomly selected (resulting in a total amount of 200 different sentences).
As illustrated in the test configuration file excerpt, 9 sessions are requested (*nbrListeners* field), each one composed of 35 steps (*nbrTestSentences* field). Statistically speaking, each testee will then hear 35 different voice pairs amoung the 40 available ones. With this mixing, each voice pair should be equally distributed in the sessions. At the end of the test, each pair is given a score calculated as the mean of the values it was attributed at each step it appeared.

4.3. Results analysis

Results of this comparison test are delivered in tables 1 and 2. The first column introduces the voice pairs using their BREF120 names. Each speaker is presented with its education level into brakets (2: *junior high scholl*, 3: *trade school*, 4: *professional high school*, 5: *high school degree*, 6: *2 year university*, 7: *4 year university or more*) (Lamel et al., 1991). The 5 following columns present the number of votes each pair received for the 5 possible scores (0: *completely different*, 1: *different*, 2: *comparable*, 3: *similar*, 4: *identical*), and the last column gives the weighted mean score.
It can be observed that each pair was presented 7.88 times on average (minumum to 4, maximum to 12). Consequently, some pairs were not evaluated by every testee. The panel size was indeed relatively limited due to the ongoing plateform development at the experiment time. These results also show that the discrimination is more salient among male speaker pairs (mean to 1.29, variance to 0.25) than with female pairs (mean to 1.49, variance to 0.4).
Furthermore, a classic conclusion of subjective tests can be observed, that pairs of speakers of equivalent education level reveived greater score (i.e. presented more similar elocution style). By contrast, it's not possible to conclude about the smoking factor, since pairs composed of a smoking speaker and a non-smoking one are uniformaly distributed in the table (these pairs are shown on grey background in table 1).

```
type = CMOS
description = 'supraSegmental BREF120'
date = 05/12/2007
author = D.Lolive
nbrListeners = 9
nbrTestSentences = 35
[A]
comment = 5 sentences for each speaker
/BREF/j8f/j8fm/j8fm0252.S1.wav
/BREF/j8f/j8fk/j8fk0831.S1.wav
/BREF/j8f/j8fj/j8fj0985.S1.wav
/BREF/j8f/j8fk/j8fk0838.S1.wav
/BREF/j8f/j8fk/j8fk0896.S1.wav
/BREF/j6f/j6fm/j6fm0121.S1.wav
...
[B]
comment = 5 sentences for each speaker
/BREF/k2f/k2fm/k2fm0252.S1.wav
/BREF/k2f/k2fk/k2fk0831.S1.wav
/BREF/k2f/k2fj/k2fj0985.S1.wav
/BREF/k2f/k2fk/k2fk0838.S1.wav
/BREF/k2f/k2fk/k2fk0896.S1.wav
/BREF/k0f/k0fm/k0fm0121.S1.wav
...
```

Figure 5: Excerpt of the "Comparison Mean Opinion Score" (CMOS) test configuration file.

5. Conclusion

This article describes a web-based listening test system for speech synthesis and speech conversion evaluation. It integrates the widely used AB-BA, ABX and MOS tests, but can be easily be extended to deal with other tests. Indeed, we extended this system with a CMOS test as illustrated in the conducted experiment. The platform isolates front-office and back-office operations. It lies upon a MVC (Model-View-Controller) pattern widely used in web application developments. The software is written in Perl ans is very flexible. Tests are realized through standard web navigators, while test descriptions and results are recorded in a relationnal database. Moreover the configuration of a test is simple and flexible, and the web interface simplifies the recruiting of the testees. This system will soon be used to make subjective measures on GMM-based voice transformation systems developed in our laboratory and applied to the ARTIC database. It is offered to the research community under free license.

Speakers		0	1	2	3	4	wtd mean
JNF(5)	JMF(2)	4	3	0	0	0	0.43
I8F(2)	JMF(2)	2	2	1	0	0	0.80
IYF(7)	IZF(5)	3	2	2	0	0	0.86
I9F(7)	JNF(5)	2	0	2	0	0	1.00
IIF(7)	JWF(5)	2	7	0	1	0	1.00
IBF(5)	JOF(7)	1	4	0	1	0	1.17
J9F(7)	K3F(6)	2	4	2	0	1	1.33
IEF(6)	JSF(6)	3	3	3	2	0	1.36
J5F(5)	JZF(5)	1	5	1	0	1	1.38
JLF(6)	JKF(6)	1	2	1	1	0	1.40
I9F(7)	JMF(2)	2	1	1	0	1	1.40
JXF(7)	JWF(5)	0	6	1	2	0	1.56
I8F(2)	I9F(7)	1	3	1	2	0	1.57
IIF(7)	JXF(7)	2	4	1	3	1	1.73
J6F(6)	K0F(6)	1	2	3	2	0	1.75
I3F(3)	I2F(7)	1	5	1	3	1	1.82
JPF(5)	JOF(7)	1	3	1	2	1	1.88
I5F(2)	JHF(2)	1	1	3	1	1	2.00
IBF(5)	JPF(5)	0	4	4	4	0	2.00
I1F(7)	JIF(7)	0	3	3	2	1	2.11
IXF(7)	IWF(5)	0	1	1	2	0	2.25
ICM(7)	IDM(6)	5	3	1	0	0	0.56
I6M(6)	JJM(7)	3	4	1	0	0	0.75
ICM(7)	JQM(5)	1	4	1	0	0	1.00
IHM(7)	IGM(7)	4	5	2	0	1	1.08
K6M(7)	JCM(7)	3	2	2	1	0	1.13
J7M(7)	K1M(7)	3	2	3	0	1	1.33
IDM(6)	JQM(5)	1	3	3	1	0	1.50
I0M(6)	I7M(7)	0	3	0	3	0	2.00
ISM(7)	ITM(7)	0	1	3	3	2	2.67

Table 1: CMOS test results, discrimination of native speaker vs native speaker (top: female speakers; bottom: male speakers)

Speaker		0	1	2	3	4	wtd mean
J8F(6)	K2F(2)	2	3	0	0	0	0.60
IFM(7)	JTM(6)	10	2	0	0	0	0.17
IHM(7)	JUM(7)	8	3	1	0	0	0.42
JFM(7)	JEM(7)	6	2	0	1	0	0.56
IGM(7)	JVM(7)	3	3	2	0	0	0.88
JVM(7)	JUM(7)	2	3	2	2	0	1.44
IHM(7)	JVM(7)	0	5	0	0	0	1.00
K7M(6)	JDM(7)	0	3	2	0	0	1.40
IGM(7)	JUM(7)	0	1	1	1	0	2.00
I4M(7)	JGM(7)	0	2	3	1	3	2.56

Table 2: CMOS test results, discrimination of native speaker vs non-native speaker (top: female speakers; bottom: male speakers)

6. References

A.W. Black and K. Tokuda. 2005. The blizzard challenge-2013 2005: Evaluating corpus-based speech synthesis on common datasets. In *EUROSPEECH*, pages 77 – 80, Lisbon, Portugal, September.

H. Duxans, A. Bonafonte, A. Kain, and J.P.H. Van Santen. 2004. Including dynamic and phonetic information in voice conversion systems. In *International Conference on Spoken Language Processing*.

H. François. 2002. *Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue*. Ph.D. thesis, Université de Rennes 1.

ITU-T. 1996. Itu-t recommendation p.800: Methods for subjective determination of trnasmission quality.

U. Jekosch. 1993. Speech quality assessment and evaluation. In *Proceedings of Eurospeech 93*, pages 1387 – 1394.

A. Kain. 2001. *High Resolution Voice Transformation*. Ph.D. thesis, OGI School of Science and Engineering at Oregon Health and Science University,.

L. Lamel, J.-L. Gauvain, and M. Eskenazi. 1991. Bref, a large vocabulary spoken corpus for french. In *European Conference on Speech Communication and Technology*, pages 505–508, 24-26 September.

D. Suendermann, A. Bonafonte, H. Duxans, and H. Hoege. 2005. Tc-star: Evaluation plan for voice conversion technology. In *DAGA 2005, 31st German Annual Conference on Acoustics*, Munich, Germany, March.