

TRUE: an Online Testing Platform for Multimedia Evaluation

Santiago Planet, Ignasi Iriondo, Elisa Martínez, José A. Montero

GPM – Grup de Recerca en Processament Multimodal
Enginyeria i Arquitectura La Salle, Ramon Llull University
Pg. Bonanova 8, 08022 Barcelona, Spain
E-mail: {splanet, iriondo, elisa, montero}@salle.url.edu

Abstract

TRUE (*Testing platfoRm for mUltimedia Evaluation*) is an online platform developed to create and perform subjective tests oriented to the evaluation of stimuli of different nature such as audio, video, graphics and text. Due to the high flexibility that the platform offers to researchers different kinds of tests can be carried out, such as emotion identification or quality assessment of synthesis systems, among others. The results can be used for different purposes depending on the research goals, e.g. to validate the emotional content of multimedia data of a corpus or to measure the expressivity of synthesized elements. TRUE involves all the stages related to the tests lifecycle, from their creation to the results retrieval, and allows the evaluators to answer the tests using any computer with an Internet connection. Making things easy for evaluators helps to minimize negative effects of fatigue, but also allows researchers to focus their efforts on the analysis of the tests results rather than on the supervision.

1. Introduction

Development of audiovisual corpora with authentic emotional content is one of the most challenging issues in the research on emotion and affect. Sources of emotional content can range from natural occurrences to acted performances (Campbell, 2000; Schröder, 2004), and a compromise between authenticity and recording quality should be considered. For this reason, tools to label and validate corpora content are required in order to guarantee a right performance in posterior use.

In general, two kinds of tests can be considered to face this goal: objective and subjective. The former does not require any kind of judgment from evaluators while the latter always involves the action of human raters. Subjective tests can be applied to: i) expressiveness validation of emotional audiovisual corpora, ii) labeling of corpora elements, including audiovisual, text and graphics resources, and iii) evaluation of synthesis systems, by rating individual stimuli or by comparing those synthesized by different techniques. Nevertheless, subjective tests can be useful in many other studies where a human criterion applied to the evaluation of data is required.

However, these subjective tests are usually designed to fit the particular features of the specific research and, in most cases, their designs are not reusable. Furthermore, evaluations tend to be time-consuming and tedious for users, whose fatigue could influence the results. In addition, achieving a high number of evaluators of heterogeneous profile tends to be difficult.

This paper describes TRUE, an online platform for designing and carrying out subjective tests that tries to solve the mentioned drawbacks. Section 2 describes TRUE features. Section 3 details the evolution of the platform from its creation to the current implementation.

Section 4 explains the technology used for its development. Section 5 is devoted to published works that have used TRUE in order to gather subjective information. Finally, the conclusions and future work are presented in Section 6.

2. Description

TRUE copes with the requirements and the drawbacks mentioned in Section 1 by offering a tool that provides a single platform to design customized and reusable tests. Researchers can design the tests according to their needs and retrieve the results from the same tool, while evaluators can take the tests from any computer with an Internet connection. All the creation process of the tests and other management issues are carried out by means of web forms, as it is illustrated in Figure 1, where a form for the creation of a new test is shown.

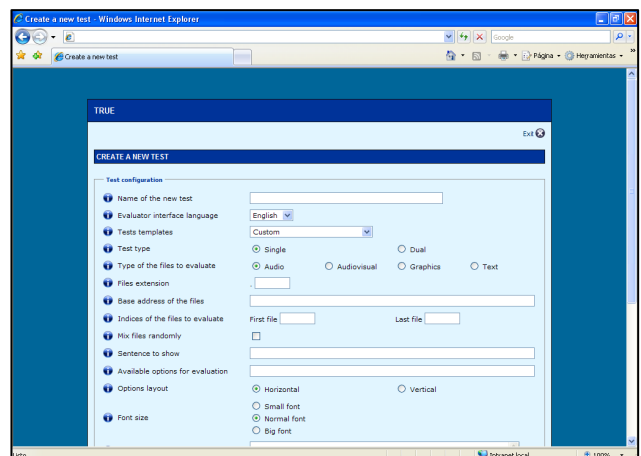
The image shows a screenshot of a web browser window titled "Create a new test - Windows Internet Explorer". The browser address bar shows "http://localhost:10000/". The main content area displays a form titled "CREATE A NEW TEST" with a blue header and a light blue background. The form is organized into sections with expandable/collapsible icons. The "Text configuration" section is expanded, showing the following fields and options: "Name of the new test" (text input), "Evaluator interface language" (dropdown menu set to "English"), "Tests templates" (dropdown menu set to "Custom"), "Test type" (radio buttons for "Single", "Dual", "Audio", "Audiovisual", "Graphics", "Text", with "Single" selected), "Type of the files to evaluate" (radio buttons for "Audio", "Audiovisual", "Graphics", "Text", with "Audio" selected), "Files extension" (text input), "Base address of the files" (text input), "Indices of the files to evaluate" (text input with "First file" and "Last file" labels), "Mix files randomly" (checkbox), "Sentence to show" (text input), "Available options for evaluation" (text input), and "Options layout" (radio buttons for "Horizontal" and "Vertical", with "Horizontal" selected). The "Font size" section is partially visible at the bottom, showing radio buttons for "Small font", "Normal font", and "Big font". The browser's status bar at the bottom shows "Lito" and "100%".

Figure 1: Form for the creation of a new test

The main goal of TRUE is to offer an interface for carrying out online tests. In this sense, TRUE gives a tool

for researchers to set up the tests allowing remote evaluators to rate the stimuli. Their answers are automatically stored in a database and can be recovered whenever the administrator requests them. Files to be tested can be stored in the same server than TRUE or in an external one because TRUE links these files from the test definition and shows the correct stimuli each time. This operation is shown in Figure 2.

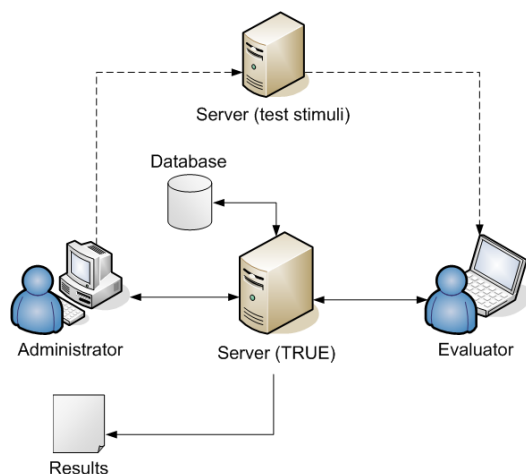


Figure 2: TRUE operation schema

It is important to highlight that different modalities (audio, video, graphics and text) can be used as stimuli and be suitably displayed in the user's web browser. Moreover, in order to avoid the negative effects of users' fatigue on the results, tests designed with the TRUE platform can be postponed and resumed. Another interesting feature is the inclusion of demonstrations at the beginning of a test which guides the answers from evaluators. A survey can be shown at the end of the test asking for user's profile and/or comments about the concluded test. Time spent on taking the test is also recorded.

With the aim of supplying standard tools to test designers, TRUE includes predefined templates for tests. These templates are related to MOS (Mean Opinion Score), CMOS (Comparison Mean Opinion Score) and DMOS (Degradation Mean Opinion Score) tests as defined by the International Telecommunication Union (1996). MOS tests are related to the assessment of perceived quality of various stimuli by means of a numerical indicator; CMOS tests perform a comparison between two stimuli; and DMOS tests are similar to CMOS but they measure the degradation in the stimulus quality when compared to another.

TRUE also allows the inclusion of plug-ins as templates. These plug-ins, such as Flash objects, can define specific tests – e.g. SAM (Self-Assessment Manikins) interface (Bradley & Lang, 1994), specifically oriented to emotion perception. The inclusion of plug-ins opens new ways of evaluating the stimuli far away from a forced answer test. In this sense, theories about representing emotions like

points in small dimensional spaces instead of determining only a set of them, as Schröder (2004) stands, can be applied to online subjective tests. Figure 3 shows two different approaches for evaluating a stimulus: by means of a forced answer test by selecting an option from the radio buttons, or through the SAM based interface plug-in, where three dimensions of emotion are evaluated: activation, valence and dominance.

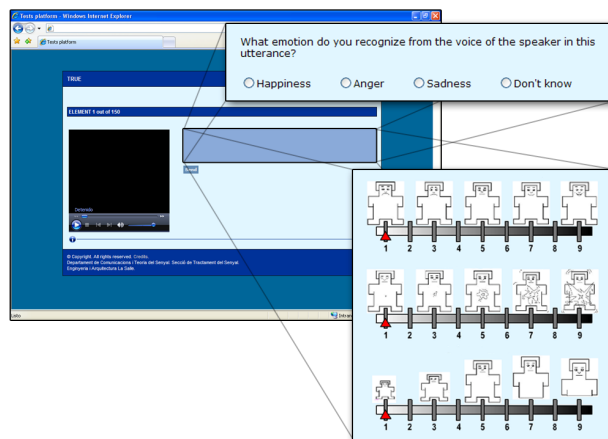


Figure 3: Evaluation by radio buttons and by means of a plug-in (SAM based interface, in this case)

TRUE differs from other test systems because it is not designed for a specific type of stimuli and it is very flexible. Unlike Irtel (n.d.) and NBS (n.d.), TRUE is not focused on a very wide range of areas of psychological research; it focuses on evaluation of corpora elements instead. Because of being focused on a specific area, and only oriented to online tests, its operation is easier than other systems like Empirisoft (2007), and permits the development of new evaluation tools for this research area. In addition to that, TRUE allows a broader audience than other tests that are usually conducted by means of paper forms or in a specific computer, like Schröder (2005). Possible inconsistencies caused by the use of an online system can be avoided by introducing some control elements. These elements can help to measure the rater's coherence.

3. Software evolution

The initial version of TRUE was designed to evaluate an emotional speech corpus in order to validate a study related to automatic emotion recognition (Planet, Morán & Formiga, 2006). The test consisted on a web with an embedded multimedia player where the user had to answer a question regarding the perceived emotion in the audio files. A set of radio buttons was used to show the studied emotional states (happiness, sadness, anger and neutral state). There were some sample audio files in the welcome page and a short survey at the end asking for age, sex and occupation of the evaluator along with his/her criteria when choosing the different options. The results were then compared with the ones from an automatic

classification following the approach of previous similar works as the described in Oudeyer (2003). A test alike to the one described above is shown in Figure 4.

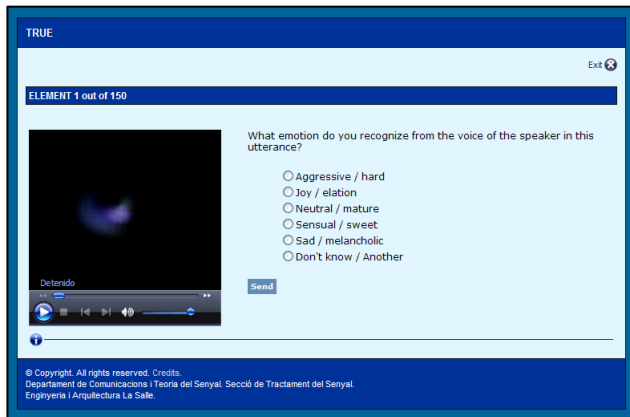


Figure 4: A test for the evaluation of audio stimuli by means of radio buttons

According to the requirements of other research areas, and by tuning the configuration of the embedded player, the test was adapted to allow the evaluation of video files keeping the rest of the features unchanged. Another interesting feature for subjective tests is the comparison between different elements, for example in studies concerned to the evaluation of stimuli synthesized by different algorithms. This was put into service by permitting the inclusion of two embedded players and one or two questions related to the played files. The layout of the items could also be configured by the designer. Posterior versions included graphics and text as other possible evaluable elements. Figure 5 shows a dual test of graphical elements with horizontal layout.

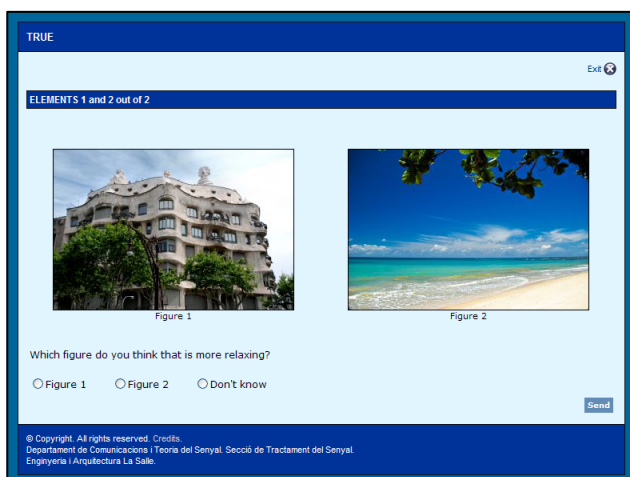


Figure 5: Test of two graphical elements with horizontal layout

Moreover, the welcome page was modified to offer different options: i) simple welcome without initial demonstration, ii) blind demonstration showing certain

elements of the test with no further indications, and iii) full demonstration including some sample files with related comments. Each option can be fully customized by means of an embedded HTML editor in the creation web form or predefined templates can be chosen instead. The goal of these welcome pages is to give guidelines to the evaluators about the test, but also to familiarize them with the stimuli that they are going to rate in order to minimize errors during the evaluation. Figure 6 shows a welcome page with a full demonstration set up, which includes five audiovisual stimuli.

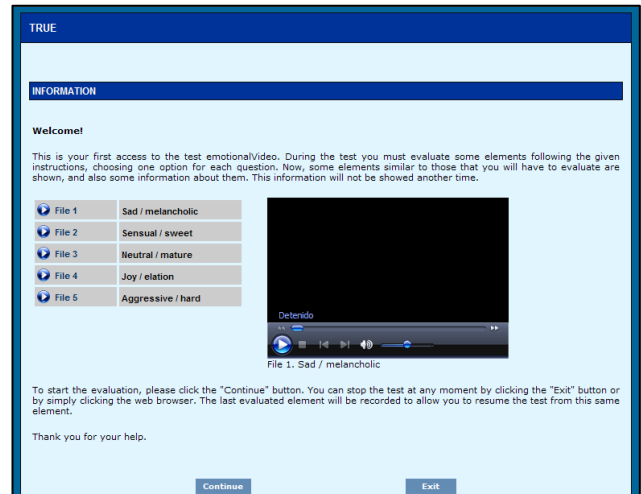


Figure 6: Full demonstration in the welcome page corresponding to an audiovisual test

The final survey can also be customized allowing the selection of the number and type of questions to be asked to the users. These questions can be text fields, text areas, radio buttons and lists. Figure 7 shows a final survey sample.

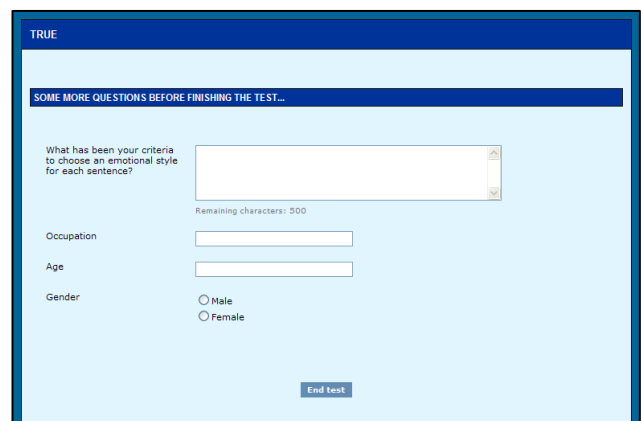


Figure 7: Example of survey at the end of a test

In all the cases, TRUE can be customized for different languages by installing the appropriate configuration files. These files are text documents with translations for the web elements and information related to the set up of the installed plug-ins.

4. Technology

TRUE is a web service implemented in Java©. It provides a tool for building and performing subjective tests, storing the evaluations of the different elements and offers management tools. Stimuli files can be placed in the same server where TRUE is installed or in another external one; this feature entails the storage capacity of the server does not need to be large. Flexibility of the platform is not only shown in the creation process but also in the reusability of previous tests and in the retrieval of results in different formats, like Microsoft Excel© or CSV (Comma Separated Values). These results, as other data related to the tests, are stored in a MySQL© database in the server and downloaded when the user requests this information; nevertheless, internal processes are transparent to end users.

5. Related studies

Several studies carried out within the authors' research group (GPMM) have used the TRUE platform. Concerning to emotion identification from speech, TRUE has been used to validate an emotional speech corpus for its use in an automatic emotional recognition experiment (Planet, Morán & Formiga, 2006). The goal of the study was to evaluate different data mining techniques comparing the performance of the algorithms and the recognition rate achieved by human evaluators.

In a broader sense, in three recent studies (Iriondo et al., 2007c; 2007b; 2007a) the objective was the automatic validation of a whole emotional speech corpus by mapping subjective criteria to automatic classification algorithms. The corpus consisted of 4638 utterances corresponding to five expressive styles: neutral-mature, joy-elation, sensual-sweet, aggressive-hard and sad-melancholic. Only 480 utterances were randomly chosen (96 per style) to be subjected to a forced answer test with the question: *What emotion do you recognize from the voice of the speaker in this utterance?* The possible answers were the five styles plus an extra option *Don't know / Another*, to avoid biasing the results because of confusing cases. The first of the studies revealed differences between classification errors made by the automatic algorithms and human evaluators although in both cases the percentages of identification were very high. The subsequent studies were focused on emulating these subjective results by mapping them to the automatic classification.

In Iriondo, Socoró & Aliás (2007), tests are related to expressive speech synthesis, with the aim of measuring the quality of the synthesized utterances. In Gonzalvo et al. (2007a; 2007b), TRUE is used with the same goal. In the three cases, subjective viewpoint is measured by MOS tests.

Audiovisual analysis and synthesis has been covered in Sevillano, Melenchón & Socoró (2006) and Melenchón (2007).

TRUE can be helpful in a wide range of purposes. As an example, it has been successfully used in research about learning methodologies (Montero et al. 2007). In this work, the subjective tests made by TRUE collected expert knowledge from teachers, which was later modeled by a fuzzy logic system able to rate the teamwork performance of engineering students.

6. Conclusion and future work

This paper has presented TRUE, an online platform designed to develop subjective tests. Although TRUE's main goal was to help in studies related to validation of emotional speech corpora, its flexibility makes it useful for a wide range of stimuli sources and purposes, e.g. evaluation of audiovisual synthesis algorithms. Because of being an online platform, tests are very accessible for users. This is a great advantage for tests designers as this makes it possible to reach a broader audience keeping an easy way of developing their tests.

TRUE has proved to be a very helpful tool in different research fields. Authors wish it can ease other researchers work and invite them to download TRUE from the TRUE website¹. Feedback on its use and suggestions for improvement will be appreciated. TRUE's design allows the inclusion of new features related to the kind of tests to be created. In these sense, different plug-ins are being developed to fit the requirements of different research areas. Moreover, any researcher can design specific plug-ins according to the guidelines provided in the development documentation and include them in their TRUE installation; it does not require any additional change in the platform.

As detailed in Section 5, TRUE has been used in many studies to gather a big amount of subjective data. In all these studies, a later stage consisted on a statistical analysis to obtain relevant conclusions using hypothesis tests such as ANOVA, t-student, Kolmogorov-Smirnov, etc. Current work is focused on embedding these tools in order to allow the users to ask for these statistical analyses during the result retrieval process.

7. Acknowledgements

This work has been partially supported by the Spanish Science and Education Ministry (CICYT TEC2006-08043/TCM).

8. References

- Bradley, M. M. & Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, 25(1), 49--59.
- Campbell, N. (2000). Databases of emotional speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 34--38.
- Empirisoft (2007). Medialab and DirectRT Software for

¹ <http://www.salle.url.edu/tsenyal/true>

- Psychology Experiments. Retrieved April 5, 2008, from <http://www.empirisoft.com/medialab.aspx>.
- Gonzalvo, X., Iriondo, I., Socoró, J. C. & Monzo, C. (2007a). Mixing HMM-based spanish speech synthesis with a CBR for prosody estimation. In *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007, Paris, France, May 22-25, 2007. Lecture Notes in Computer Science*, 4885, pp. 78--75. Springer, Heidelberg.
- Gonzalvo, X., Socoró, J. C., Iriondo, I. & Monzo, C. (2007b). Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castillian Spanish. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6)*, pp. 362--367, Bonn, Germany.
- International Telecommunication Union. (1996). *Methods for subjective determination of transmission quality. ITU-T Recommendation P.800*.
- Iriondo, I., Planet, S., Alias, F., Socoró, J. C. & Martínez, E. (2007a). Validation of an expressive speech corpus by mapping automatic classification to subjective evaluation. In *Computational and Ambient Intelligence. 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings. Lecture Notes in Computer Science*, 4507, pp. 646--653. Springer, Heidelberg.
- Iriondo, I., Planet, S., Alias, F., Socoró, J. C., Monzo, C. & Martínez, E. (2007b). Expressive speech corpus validation by mapping subjective perception to automatic classification based on prosody and voice quality. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*. Saarbrücken, Germany.
- Iriondo, I., Planet, S., Socoró, J. C. & Alias, F. (2007c). Objective and subjective evaluation of an expressive speech corpus. In *Advances in Nonlinear Speech Processing. International Conference on Non-Linear Speech Processing, NOLISP 2007, Paris, France, May 22-25, 2007. Lecture Notes in Computer Science*, 4885, pp. 86--94. Springer, Heidelberg.
- Iriondo, I., Socoró, J. C. & Alias, F. (2007). Prosody modelling of Spanish for expressive speech synthesis. In *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4, pp. 821--824. Honolulu, HI, USA.
- Irtel, H. (n.d.). PXLab: The Psychological Experiments Laboratory. Retrieved April 5, 2008, from <http://www.pxlab.de>.
- Melenchón, J. (2007). Síntesis Audiovisual Realista Personalizable. PhD Thesis. Ingeniería i Arquitectura La Salle, Universitat Ramon Llull.
- Montero, J. A., Alias, F., Garriga, C., Vicent, L. & Iriondo, I. (2007). Assessing students' teamwork performance by means of fuzzy logic. In *Computational and Ambient Intelligence. 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings. Lecture Notes in Computer Science*, 4507, pp. 383--390. Springer, Heidelberg.
- NBS (n.d.). Auditory, Visual and Multi-modal Stimulus Delivery for Neuroscience. Retrieved April 5, 2008, from <http://www.neurobs.com/presentation>.
- Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2), 157--183, special issue on Affective Computing.
- Planet, S., Morán, J. A. & Formiga, L. (2006). Reconocimiento de emociones basado en el análisis de la señal de voz parametrizada. In *Actas da 1a Conferência Ibérica de Sistemas e Tecnologias de Informação, Ofir, Portugal, 21 a 23 de Junho de 2006*, 2, pp. 837--854.
- Schröder, M. (2004). Speech and Emotion Research: An Overview of Emotion Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis. PhD Thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University.
- Schröder, M. (2005). RatingTest - Java software for designing and carrying out listening tests. Retrieved April 5, 2008, from <http://ratingtest.sourceforge.net>.
- Sevillano, X., Melenchón, J. & Socoró, J. C. (2006). Análisis y síntesis audiovisual para interfaces multimodales ordenador-persona. In *Proceedings of the 7th Congreso Internacional de Interacción Persona-Ordenador (Interacción 2006)*. Puertollano, Ciudad Real.