Automating Test Oracles for Systems with Complex Outputs

Rafael A. P. Oliveira
Instituto de Ciências Matemáticas e de Computação (ICMC/USP)
13560-970 São Carlos, SP, Brazil.
rpaes@icmc.usp.br

***Abstract*—In Software Testing, systems with complex outputs (GUIs, web applications, and *Text-to-speech* (TTS) systems) make the test automation a difficult job and may require from the tester a large amount of time to determine whether the current output is correct. Even so, there is still no known general method to define test oracles for such systems. Regarding TTS systems, in which audio files are given as output, the literature describes several techniques useful for measure their quality, but they mostly involve opinion scores, informal interpretations and human intervention. This Ph.D. research aim to use test oracles to deal with this problem, extending the framework O-FIm that uses CBIR concepts to automate test oracles. Using feature extractors implemented specifically for systems with complex outputs (image and audio), this research provides an effective automated test oracle technique, supporting the quality assessment of such systems. We expect this new approach to reduce the human efforts of testing systems with complex outputs.**

***Keywords*-Test Oracle; Text-to-Speech; CBIR; Test Automation**

I. A characterization of the problem

Software Testing activities are supposed to include a mechanism to decide about whether a particular execution is considered as failure or not. This decision about a software execution correctness is provided by the *test oracle*. In practice, some domains make the oracle automation a difficult job and the tester has to expend some extraordinary amount of time to determine whether outputs are correct or not [4]. In extreme complex cases, test oracles do not exist and the software system is considered as non-testable. For example, when the SUT's (*System Under Test*) output is given in a non-trivial format, such as an image, a virtual environment, a sound or a *Graphical User Interface* (GUI), usually, testers are responsible to check whether the current outputs are acceptable (human oracle).

This research aim to develop automated test oracles for systems with complex output formats. The central theme is *Text-to-speech* (TTS) systems, which converts written text into human speech [3], generating an audio file as output. Many embedded systems use TTS applications nowadays to read e-mails or social network updates, to read books or headlines to blind people, to read traveling directions, news, stories, weather forecasts and, in user interactions in dialogue systems. However, the mainstream adoption of TTS is severely limited by its quality. Pronunciation and intonation problems make the speech synthesized highly unnatural. Users expect that the system should sound just like a human, creating a full range of speech. Existent studies on quality aspects in TTS systems use manual and *ad-hoc* techniques [3], fact that enriches this research.

II. Background

In order to get flexible test oracles, this research will extend the open source framework O-FIm (*Oracle for Images*)[1], which was developed and evaluated in previous works [2]. O-FIm uses CBIR (*Content-Based Image Retrieval*) concepts to support the automation of oracles for programs with graphical outputs. CBIR is any technology which helps organize digital images using their visual content [1]. In an image database, according to one or more provided criteria, CBIR systems locate images that are similar to a query image. Similarity criteria are obtained by extracting features from the image such as color, texture and shape. A set of extracted characteristics creates a feature vector that is considered in its retrieval. Then, the system measures how similar are two feature vectors, using some similarity function.

O-FIm is a software framework that adapts CBIR concepts to provide a dynamic way of creating test oracles, using feature extraction, similarity functions and object comparison. Given a model image, O-FIm uses feature extraction to obtain information or data from an output image. Afterwards, O-FIm compares the model feature vector with the output feature vector. The tester chooses which features and how they should be extracted using an oracle description.

O-FIm allows the installation of plug-ins: feature extractors and similarity functions. Testers may implement their plug-ins and design flexible testing oracles using the framework's environment. The flexibility provided by O-FIm allows testers to obtain a Java program that compares two objects, responding if they are similar or not, according to a threshold. Then, the plug-ins are the main contribution from the testers to build a new oracle. Similarly to the extraction of image characteristics, speech signals can also be processed and have their features analyzed. In this research, regarding TTS systems, a catalog of feature extractors to speech signal (output of TTS systems) will be created and used in automated test oracles.

III. Material and Methods

Images and audio files are different in their essence, however both domains allow comparisons among extracted features. In this sense, using O-FIm resources, one can create oracle applications for systems with audio as output. In

[1] see: http://ccsl.icmc.usp.br/pt-br/projects/o-fim-oracle-images

978-0-7695-5194-4/14 $31.00 © 2014 IEEE
DOI 10.1109/ICSTW.2014.48

order to do that, it is necessary to identify and implement the necessary audio feature extractors. In this research, we have identified the following audio features as useful plug-ins in test oracles: location of vowels, phonemes and fundamental frequencies and measures of energy, pause, durations, formant frequencies, and reader characteristics. The similarity functions may not differ from those used in CBIR systems.

However, one of the counterparts in using O-FIm is the need for a model from which it is possible to establish a baseline for comparing objects that are under test. For example, when it is desired to evaluate the appearance of a GUI in a particular L&F (*Look and Feel*), it requires an environment from which it is possible to establish a standard image for comparison during the test. Regarding TTS systems two empirical strategies are possible: *(1)* use two different systems and compare their outputs; and *(2)* compare the output produced by a TTS system with the audio of the voice of a person who reproduces the text input. So far, we have implemented two audio feature extractors and some empirical studies have been conducted using strategy number *1*.

IV. Current Status

Technical and theoretical activities of this research were divided in two major activities: *(1)* studying TTS applications and speech synthesis to determine the useful features to be extracted from audio signals; and *(2)* extending the O-FIm to extract, implementing O-FIm plug-ins and test oracles. The first major activity was concluded and its main result is a *Systematic Mapping* (SM) with 115 papers on quality of TTS systems. The studies included in this SM shows that informal human interpretations and different manual approaches are the most common ways to evaluate the outputs of TTS systems. The MOS (*Mean Opinion Score*) strategy is the most common technique identified.

Regarding the second activity, O-FIm components were completely reimplemented to accept audio plug-ins. In addition, two audio feature extractors were implemented and empirically evaluated. Using a TTS systems written in Brazilian Portuguese, this research have implemented a *vowel* feature extractor and a *phoneme* extractor. In both cases, using parameters, the tester can set a test oracle using numbers and positions of specific vowels or phonemes.

V. Early Results

Our preliminary results were obtained using a test scenario in which one of the systems represents the oracle and the other represents the SUT. The two systems used in this study were *CPqD Text-To-Speech (version 3.3)* regarded as the oracle and *Google Translate Text-to-Speech* regarded as the SUT[2]. In order to generate a data set, 100 Portuguese words were selected. These words were generated by CPqD system loaded with three different news from an aleatory popular news website. The choice of words did not follow any restriction, and it was obtained following the order of occurrence. At the end of the process, 100 WAVE files were generated. After that, the selected words were generated using "Google Text-to-Speech" and the same extraction process was carried out on the selected words. The two extractors developed were applied in this words and the set of words was analyzed manually to check the real occurrences of vowel and phonemes of interest. The results were considered satisfactory in both cases. Table I presents the results obtained using the vowel extractor. An empirical study is being conducted and an automated oracle is being evaluated through the association of the two extractors and the Euclidean similarity function.

Table I
Summary for the vowels feature extractor

Vowel	Number of occurrences					
	CPqD system			Google TTS		
	0	**1**	**> 1**	**0**	**1**	**>1**
A	71%	85%	57%	60%	68%	64%
E	65%	76%	60%	73%	61%	50%
I	89%	72%	50%	74%	59%	25%
O	75%	69%	40%	61%	72%	40%
U	68%	68%	100%	93%	40%	100%
hits	**74%**	**74%**	**61%**	**72%**	**60%**	**56%**

VI. Conclusions and Expected Contributions

This research aim to use CBIR concepts, by means of the O-FIm framework, to automate test oracles to systems with complex outputs. TTS systems, despite their importance, have had their quality assessed by manual and unproductive processes. This research has implemented plug-ins and oracles useful as components for an automated test of TTS systems. At the end of this study, this strategy will be evaluate by means of empirical studies, highlighting its strengths and identifying limitations. It is expected that the proposed method works as complementary to traditional testing techniques, mitigating human efforts and adding general knowledge to this field.

Acknowledgments

Currently Rafael is a student visitor at University of Maryland and Prof. Atif Memon is his supervisor. Rafael is supported by FAPESP (Grant Numbers 2012/06474-1 and 2013/01775-6). He also thanks Prof. Marcio E. Delamaro and Prof. Fátima L.S. Nunes, the advisors of this research.

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[2] M.E. Delamaro, F.L.S. Nunes, and R.A.P. Oliveira, "Using concepts of content-based image retrieval to implement graphical testing oracles", *STVR*, vol. 23, no. 3, pp. 171–198, 2013.

[3] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, Cambridge, U.K., pp. 603, 2009.

[4] T. Yu, A. Sung, W. Srisa-an, and G. Rothermel, "Using property-based oracles when testing embedded system applications," in *ICST 2011*. Berlin, Germany, 2011, pp. 100–109.

[2] CPqD is the major Brazilian provider of telecommunications and IT solutions. *CPqD Text-To-Speech* is a TTS system able to synthesize texts written in Portuguese in speech signals near human speech see: http://www.cpqd.com.br/