An Evaluation Methodology for Automatic Transcription System of Radiology Reports

Valéria Farinazzo Martins[1] and Lincoln de Assis Moura Jr.[2]

[1] Universidade Presbiteriana Mackenzie, Grupo de Processamento Gráfico
Rua da Consolação, 930, São Paulo – SP, Brazil
[2] Escola Politécnica da USP and Zilics eHealth
São Paulo – SP, Brasil
valeria.farinazzo@mackenzie.br, lincoln.a.moura@gmail.com

Abstract. This research combines knowledge from Computer Science and Health Science in order to propose an evaluation methodology for Automatic Transcription System of Radiology Reports. This methodology was designed based on Voice User interface requirements and specific requirements of automatic transcription systems of Radiology report. The same methodology was previously validated through some inspections and usability tests outside the hospital environment and, afterwards, it was used in two hospitals in São Paulo city. This approach aims to reduce costs of testing and available time by radiologists interviewed. Thus, the final product in this work consists of a set of criteria for evaluation of usability, comprising the name of the metric, evaluating method, steps to be followed and material to be used. By the use of this set, the evaluators can process the results of each requirement from the software.

Keywords: Usability Evaluation, Voice User Interface, Automatic Transcription System of Radiology Reports, Dictation System.

1 Introduction

Since the end users of computer systems began to be non professional people in computing, the Human Computer Interface (HCI) field has had a fundamental role in the success of computer products on market. Thus, in addition to meeting the desired features, an application should have intuitive and friendly interfaces to those users.

Emerging technologies have been integrated into interfaces available on the market: touch screen interfaces, three-dimensional navigation environment and voice use as a way to interact with the device are some examples.

Even though dialogue speech systems have appeared in the 1950's, during the onset of Artificial Intelligence research [1-3], a significant growth in the production of systems with users interface based on voice took place in the past decade, especially for commercial use via telephone, such as airplane ticket and hotel reservations, flight schedule queries and accessing bank accounts.

The research for usability evaluation of voice recognition systems is still quite new. The Methodology and suggested methods to evaluate Voice User Interfaces

M.M. Cruz-Cunha et al. (Eds.): CENTERIS 2011, Part III, CCIS 221, pp. 194–203, 2011.

(VUIs) come from the present knowledge of User Interface (UI) evaluation, related to the work of some researchers that developed methods to investigate their specific projects, trying to generalize and to propose reference models for such applications. This is the case of PARADISE [4], EAGLES [5] and DISC [6].

If we analyze the use of voice recognition systems in the Health general purpose, such as in emergency, it is possible to see that they have not been effective due to the domain large vocabulary - it is known that the common vocabulary of the area has more than 100 thousand items. In other words, the information available inside Health field is extremely varied.

Thus, the voice recognition technology has been used for more specific purposes, such as automatic transcription system reports (ATSR) in Radiology field. It means that the vocabulary is considerably smaller, providing a higher accuracy in recognizing specific terms. Although usability is a quality attribute of software that aims to ensure that user requirements are attempted [7, 8], speech recognition systems in healthcare, there are mostly analyzed by methods complete, complex and well-established usability evaluation. The evaluators of such systems are still focused on evaluating only the accuracy or detecting mistakes in these systems [9- 12]. There are too many works in the literature that establishes the specific requirements of this area that must be met in order to make use speech recognition effectively and efficiently.

One of the main problems shown in the literature [9, 10, 13-15] is the delay in radiology reports due to the time spent from the moment of entry of recorded reports to its return in textual form for the radiologist to assess.

The automatic report transcription systems (that use VUI) have been thought of as a solution to decrease this time (Turn Around Time) and also to decrease the running costs of the radiology department. To verify the efficiency of the use of automatic report systems, not only the VUI general requirements must be evaluated but also the specific demands in the area, to see if the available commercial products have been used correctly by the users.

The objective of this article is, therefore, to organize concepts of voice recognition and also voice recognition systems evaluation aiming at proposing a useful set of methods that are feasible, practical and suitable. Thus, a specific methodology to evaluate this category of applications will be suggested.

This article is organized as follows: the second section covers the materials and methods used in research; section 3 presents results and discussions of methodology, and finally, the limitations and advantages of the proposed methodology are discussed and the future work is analyzed.

2 Materials and Methods

In order to make possible the development of a methodology to assess, simply and inexpensively the ATSRs in Radiology, it was necessary to establish three steps, namely: 1) identification of VUI generic requirements and ATSR specific requirements, 2) generation of a methodology for evaluating ATSR systems, and 3) application of the methodology. The following subsections detail these steps.

2.1 Identification of VUI Generic Requirements and ATSR Specific Requirements

Based on Nielsen evaluation [7] and [19], we propose that the following set of requirement should be formally assessed when evaluating VUI-based report transcription systems.

- **Accuracy:** It is one of the most important requirements, because wrong information can compromise report quality, alter a diagnosis and compromise a treatment.
- **Vocabulary Size:** vocabulary can neither be too larger-in order to lower the rate of word recognition nor too small for it does not consider the words in the application's dominion.
- **Specific Dictionary for Radiology Information System (RIS):** the system must consider words used daily in radiology reporting.
- **Noisy Interference:** depending on the area, hospitals can be very noisy, but this should not interfere on the efficiency of recognition.
- **Continuous Recognition:** the user must be able to dictate the report naturally, without having to worry about pauses between words, i.e., user must be able to speak in a natural and continuous way.
- **Integration with Hospital Systems:** Picture Archiving and Communication System (PACS), Hospital Information System (HIS) e RIS.
- **Help:** is linked to the ease with which users, especially beginners, will have to learn and access the help system to get to dictate a report efficiently.
- **Hand and Recover Error:** it is connected, for example, how the system works when it does not recognize a word dictated the user.
- **Adequacy of Feedback:** the system should not provide feedbacks that impair reasoning ability from user, but it is able to generate a report about errors dictated or unrecognized words.
- **Response Time of Feedback:** the transcripts must occur in real time without the delay that could interfere with cognitive load of the end user.
- **Adequacy of TAT:** time must be at least shorter than the human transcription systems.
- **Customer Satisfaction:** This requirement is linked to the pleasure of using a ATSR Radiology, measured by questionnaires.

Although it is known that most of these requirements must be thoroughly tested in order to verify their real value as the system itself - carried out by development enterprise - this paper is related to usability, which can be measured by a moderate amount of users and / or specialists even so as not to raise too much and to negate the cost evaluation.

2.2 Generation of a Methodology for Evaluating ATSR Systems

The proposed evaluation methodology must be able to:

- Use additional usability and inspection tests to provide a lower cost and shorter assessment time.

- Be applied to previously implemented systems.
- Act as a guide to evaluating the usability in this class of systems.
- Investigate the difficulties on evaluating specific requirements.
- Group the proposed requirements according to their characteristics.
- Propose metrics for evaluating each of those requirements.

To assess automatic report systems, the following classes (Table 1) were here defined in a modified way from what Möller [18] proposed in his work about general purpose voice recognition system evaluation.

Table 1. Classes defined in a modified way from what Möller [18]

Class	Requirements
Class 1 Achievement Requirements associated to the correct operation of the application without degrading its achievement	Accuracy, vocabulary size, specific dictionary for RIS, noisy environment, user's naturalness of speech (continuous recognition)
Class 2 Usability Efficiency and efficient requirements, decreasing the user's cognitive load	Minimization of memory overloads, adequate modality, time for the report to be ready
Class 3 Hardware and Integration	Requirements connected to physical achievement: Separateness between keyboard and dictation, use of proper architecture (client-server or browser-server), integration with existing systems, quality of audio system, and quality of database entries
Class 4 Human Factors	Requirements connected to the user's pleasure in using the system and the will to continue to use it
Class 5 Feedback	System's feedback time, system's visibility, feedback's adequacy, message exit quality
Class 6 Handling Error and Help	Requirements that are related to the capacity of the system in correcting not only errors found but also correcting a dictation, may it be in real or posterior time

The requirements were classified according to the level of assessment difficulty (Level 1 – low complexity, Level 2 – medium complexity and Level 3 - high complexity) as an example: accuracy; vocabulary size; noisy environment; continuous recognition fall in complexity Level 1, as in Table 2.

A method for analyzing each requirement was developed. A template was created for each requirement in order to facilitate the assessment, as illustrated in Table 3, for Customer Satisfaction.

Table 2. Complexity of the requirements

Complexity	Requirements
Low	Accuracy; vocabulary size; noisy environment; continuous recognition, time turn aroud.
Medium	Help system, Hand and Recover Error; quality of audio system, time turn aroud
High	System's feedback time, system’s visibility, feedback's adequacy, customer’s satisfaction

Table 3. A template for Customer Satisfaction

Customer Satisfaction	
Kind of Evaluation	Subjective
Evaluation Methods	Questionnaire
Importance	High
Difficulty in Evaluation	Level 3
Evidence to look for / Metrics to use	Ease of use, aggregated value, success of the task

2.3 Application of the Methodology

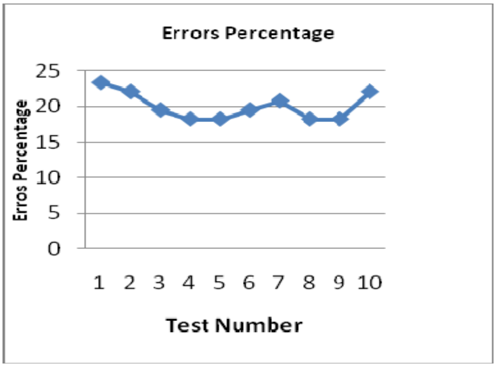
We apply the Usability Evaluation Methodology for the Automatic Transcription Systems Reports in Radiology in two ways:

- First, analyzing all possible requirements, with a stand-alone system (not inside a hospital). These requirements were analyzed using the techniques of satisfaction questionnaires, observation and inspection of usability in order to facilitate testing, saving time and costs, and disturb the least possible the radiologists.
- Second, analyzing the other requirements that could not be analyzed outside the production environment - i.e. inside hospital - with real users - Radiologists - using observation techniques and questionnaires of satisfaction. Then, we selected a user who uses the system more frequently.

Planning the Usability Inspection for Stand-alone System. Inspection of ATSR usability was performed according to the following steps:

- Conception of heuristics: for experts to check compliance with what the system was established.
- Choice of experts.
- Preparation of inspections: these inspections spend about 20 hours, one of the main reasons for not using the second professional for all inspections and also of not using more than two experts.
- Generation of lists: aims to generate the results and analysis of each particular test case for inspection, as shown in Table 4. At this stage, we observed the metrics listed in Section 2.1.

Table 4. Results of inspection

Metric	Hand and Recover Error
Evaluation Method	Inspection
Tester	1 expert
Materials	1 text with 77 words, with many words out of vocabulary specific to the area of radiology, a device SpeechMike.
Steps	Verify, by inspection, the system acts as if the user uses words that are not in the dictionary application.
Results	The text used in this report for the test has 77 words. The test was repeated 10 times (Graphic 1). This test spent about 17 minutes.
	Graphic 1. Error Percentage 
Analysis	In the case, the average error was 4.54%, with average deviation of 1,237%. However, this test was 20% with average deviation of 1.95%. This was expected, since the vocabulary for the system is specific to the area of Radiology.

Planning the Usability Tests for Stand-alone System. We used a methodology for preparation of tests adapted for Diah et al [16], Nielsen [7] and Mitchell [17], this plan consists of the following activities performed consecutively:

- Planning for usability testing: the tests were conducted inside a non-hospital environment, through non medical participants. We conducted two pilot tests to check possible inconsistencies. The main goal was to generate the results and analysis of each particular test case for the tests with end users.
- Preparing test materials: in addition to inspection of materials used, we also need: photography camera, pre-and post-test questionnaire and forms for user observation.
- Tasks establishment: The tasks aimed to validate the metrics (section 2.1).
- Participants' selection: we select six people (three men and three women) and two experts in inspection. Two criteria were determined: different tones of voice, both male and female, and people with different accents of the native people of São Paulo city.

- Conducting usability tests: the sessions were composed of four parts, with an average duration of 45 minutes.
- Analysis of usability problems, as shown in Table 4.

3 Results

We divided the discussions of usability tests results into two parts, one concerning the stand-alone system and another using a system deployed in a hospital in Sao Paulo city.

3.1 Stand-Alone System

Through all testing performed on this system, we can describe the following conclusions.

- The ATSR system was very efficient in relation to the speech recognition accuracy (93% on average), even including in the group people with pronounced accents.
- The voice accuracy reached 95%, even with minimal training of voice.
- The system is sensitive to changes in speed of speech.
- There is no significant difference in the rate of recognition without training and with training. Hence, the system could be used without being carried minimal voice training, with acceptable rates of voice recognition.
- The system delay to display the text on the window causes for people a feeling that the system was not working.
- Interference noise affects mainly the accuracy of speech recognition.
- The two devices used for the entry of the reports - HeadSet Philips SpeechMike and Philips - have proven successful. We had expectations of, according to information from the supplier,to have a big difference in sound quality; however, the recognition accuracy was better with the headset in low-noise interference. Among high noise interference, the device HeadSet was better SpeechMike. The significant cost difference between the two and ergonomic equipment must be taken as important aspect.
- Lack of visibility and adequacy of feedback were pointed by the users as uncomfortable.

3.2 System Used in a Hospital

We observed a end user and can to see:

- The Turn-Around time is about 5 minutes.
- The system is very sensible to environment noise.
- The radiologist needs a a high degree of concentration because it displays three screens simultaneously, as we can see in Figure 1.

Fig. 1. A radiologist using the system

3.3 Comparison

Through observation of the use of this system, we can conclude:

- The voice recognition rate was lower than shown by stand-alone system, maybe because its system version.
- When the typing service is used, the radiologist must review the medical image to confirm the report, already in use ATSR, it becomes unnecessary because the image is real time.
- Emergency Reports can be generated by the system more quickly (20 minutes x 5 minutes in average).

4 Conclusions

This article focuses on the evaluation of automatic transcription system for radiology reports. Various specific requirements in this class of systems that are not taken into consideration, either by the classic evaluation methodologies of usability or by the new VUI evaluation methods were identified. These requirements have been neglected when these applications are evaluated.

The methodology to provide these peculiar requirements based on usability inspection and usability tests was proposed, in order to assure a lower cost and a higher efficiency. It aimed at reducing costs with usability testing, to be known in the literature of HCI, this is a cost that may be impeding the evaluation of many systems.

Since the ATSR have been mooted as a solution to reduce the time for the report is ready, and also as a reduction of overall costs of the Department of Radiology, a methodology for evaluating these products is essential. The proposed methodology contributes to the choice of a system that faces the needs of the market relative to its end user. Thus, this methodology takes into consideration many aspects that go beyond the recognition rate of these systems, addressing issues such as size of vocabulary used, the environment used and user satisfaction.

Hence we tried to use, as much as possible, the inspection techniques that do not use these end users and significantly decrease the time and cost evaluations.

Also, for hospitals that did not require cumbersome and time available for its radiologists, we chose to use the inspection by experts in usability; also, volunteers were selected as users for usability testing, when the inspection was not most appropriate method. Only the observation of end users - radiologists - and filling, for them, a satisfaction questionnaire were the techniques used to analyze the usability when the need to be "in the field." This questionnaire took no more than 3 minutes from the time of the radiologist.

Thus, this work serves as a guide for IT field in hospitals and radiology clinics when evaluating whether to purchase systems for automatic transcription of reports, increasingly common in the domestic market. It can also be used to check when working with customizations such systems, the usability they want to reach and if it is currently in force.

It is desirable that the usability evaluation proposed by this methodology is carried out or led by experts in usability, it is necessary even for an expert, a sizable amount of hours primarily to the evaluation of inspection.

5 Future Work

We suggest as future work having more automation of tests, both with usability experts, and with volunteers and end users, which was not the initial focus of this work. Second, the use of intelligent agents that can capture the reports, change the dictation by synthesizing voice. This would reduce greatly the time of inspection evaluation, considered one of the key drawbacks to this methodology.

Third, we can indicate the use of a more extensive vocabulary of the reports in order to have a more accuracy measure of voice recognition.

References

1. Allen, J., Perrault, C.: Analysing intention in utterances. Artificial Intelligence 15, 143–178 (1980)
2. Kamm, C., Walker, M., Rabiner, L.: The role of speech processing in human computer intelligent communication. Speech Communication 23, 263–278 (1997)
3. Price, P.: Evaluation of spoken language systems: the ATIS domain. In: Proceedings of the Third DARPA Speech and Natural Language Workshop. Morgan Kaufmann, San Francisco (1990)
4. Walker, M.A., Litman, D., Kamm, C., Abella, A.: Evaluating spoken dialogue agents with PARADISE: Two case studies. Computer Speech and Language 12(3), 317–347 (1998)
5. Gibbon, D., Moore, R., Winski, R. (eds.): Hand-book of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, New York (1997)

6. Dybkjaer, L., Bernsen, N.O., Dybkjaer, H.: Generality and objectivity: central issues in putting a dialogue evaluation tool into practical use. In: Hirschberg, J., Kamm, C., Walker, M. (eds.) Interactive Spoken Dialog Systems. Proceedings of a Workshop Sponsored by the Association for Computational Linguistics, Madrid, Spain, pp. 17–24. ACL (1997)
7. Nielsen, J.: Usability Engineering. Academic Press, Cambridge (1993)
8. Sommerville, I.: Software Engineering, 6th edn. Addison Wesley, New York (2001)
9. Kanal, K.M., Hangiandreou, N.J., Sykes, A.M.G., Eklund, H.E., Araoz, P.A., Leon, J.A., Erickson, B.J.: Initial Evaluation of a Continuous Speech Recognition Program for Radiology. Journal of Digital Imaging 14(1), 30–37 (2001)
10. Kimberly, D.V.: A Methodology of Error Detection Improving Speech Recognition. Simon Fraser University Library, Canada (2006)
11. Voll, K., Atkins, S., Forster, B.: Improving the Utility of Speech Recognition Through Error Detection. Journal of Digital Imaging 21(4), 371–377 (2008)
12. Paulett, J.M., Langlotz, C.P.: Improving language models for radiology speech recognition. Journal of Biomedical Informatics 42, 53–58 (2009)
13. White, K.S.: Speech recognition implementation in radiology. Springer, Heidelberg (2005)
14. Bhan, S.N., Coblentz, C., Norman, G.R., Ali, A.H.: Effect of Voice Recognition on Radiologist Reporting Time. CARJ 59(4) (October 2008)
15. Durling, S., Lumsden, J.: Speech Recognition Use in Healthcare Applications. In: Proceedings of MoMM 2008, MoCoHe. ACM, New York (2008) 978-1-60558-269-6/08/0011
16. Diah, N.M., Ismaili, M., Ahmad, S., Dahari, M.K.M.: Usability testing for educational computer game using observation method. In: CAMP 2010, The First International Conference on Information Retrieval and Knowledge Management, Shah Alam, Malaysia (2010)
17. Mitchell, P.P.: A step-by-step guide to usability testing. iUniverse, Lincoln, NE (2007)
18. Möller, S.: A new taxonomy for the quality of telephone services based on spoken dialogue systems. In: Proc. 3rd SIGdial Worksh. on Discourse and Dialogue, US, Philadelphia, pp. 142–153 (2002)
19. Salvador, V.F.M., Oliveira Neto, J.S., Kawamoto A.L.: Requirement Engineering Contributions to Voice User Interface. In: First International Conference on Advances in Computer-Human Interaction, Sainte Luce, pp. 309–314 (2008)