MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010
© 2010 IMIA and SAHIA. All rights reserved.
doi:10.3233/978-1-60750-588-4-1246

Evaluation Methodology for Automatic Radiology Reporting Transcription Systems

Valéria Farinazzo Martins Salvador[a], Lincoln de Assis Moura Jr.[b]

[a] *Escola Politécnica da USP and Universidade Presbiteriana Mackenzie*
[b] *Escola Politécnica da USP and Zilics eHealth*

Abstract

This article describes a usability evaluation methodology for automatic transcription system used for radiology reporting. In order to assess this class of system's limitations and strengths, a review of the concepts involved in this kind of system is done in a critical way. Specific requirements, for this category of application, that are forgotten when a product is launched in the market, are listed and a methodology for their evaluation is presented.

Keywords:

Voice user interface, Usability evaluation, Automatic transcription system for reports

Introduction

Voice User Interface (VUI) uses technology of recognition and voice synthesis to provide access to information to its users, allowing them to perform transactions and offering communication support. Even though dialogue speech systems have appeared in the 1950's, during the onset of Artificial Intelligence research [1, 2, 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 reservations, hotel reservations, flight schedule queries, accessing bank accounts and others. The state of the art in voice technology (recognition and synthesis) allows for the development of automatic systems that work in real conditions, even though the systems are quite simple, meaning they have a limited vocabulary and therefore a high rate of recognition – close to 95%. Companies such as Philips [4], IBM [5] and Nuance [6] have invested in the development of voice recognition technologies for restricted domain [7-10].

According to [8], this uncontrolled growth is due to a series of factors: clients dissatisfaction when using touch tones systems, information access through mobile devices growth, the company's need to provide clients with a more efficient way and with lower costs for their clients demands. Although the authors state that this growth is due to the development of voice recognition and synthesis technology that have become stronger and more capable of sustaining effective interactions, one must realize that the present systems have a restricted capacity for voice recognition, a limited vocabulary and grammar coverage, and a limited ability of tolerating and correcting mistakes. Voice recognition systems, based on tasks that have a high level of variability, with a large vocabulary and that have to work with an open grammar, similar to man-to-man communication, make real voice recognition more difficult.

In Health Care, the use of voice recognition in general-purpose systems, in emergencies for instance, does not work efficiently because of the large domain's vocabulary (it is known that a Health Worker uses more than 100 thousand vocabulary words in their daily routine) In addition, there are over 60 million diagnostic possibilities in SNOMED [11]. The information available in Health Care is quite diverse.

Nevertheless, VUI has been used in voice recognition systems for more specific purposes, such as the Automatic Transcription of Reports in Radiology. That means the vocabulary is considerably smaller and gives a higher precision of recognition of specific denominations.

Regarding the usability of Voice Recognition Systems as a whole, there is yet much to determine as an evaluation methodology. In his research, Nielsen [12] proposed rules and determined heuristics to allow the interfaces to be analyzed in regard of usability. The authors got his inspiration from the user's graphic interfaces that were (and still are) in widespread use. But VUIs system evaluation is different from GUIs, mainly when it comes to voice transience that affects major usability factors such as transparency, learning, cognitive overload, error handling and user's control.

The research for usability evaluation of voice recognition systems is still quite new. The Methodology and suggested methods to evaluate VUIs come from the present knowledge of UI evaluation, related to the work of some researchers that developed methods to investigate their specific projects, and that tried to generalize and proposed reference models for such applications. That is the case of PARADISE [13], EAGLES [14] and DISC [15].

An even more incipient case is the usability evaluation of automatic transcription system. There are many specific usability issues that have not being analyzed as part of more concrete evaluation methodologies. There are two main reasons for that. First of all, owing to the fact that the classic usability evaluation methodologies cannot cope with voice systems, significant changes must be done for these methodologies to become efficient for that purpose. Secondly, the evaluators of such systems are still focused on evaluating only the accuracy or detecting mistakes in these systems [16-19]. In other words, there are many other usability issues that are being neglected. The evaluation of automatic report transcription systems is a relatively different task because:

- Health Care vocabularies are more extensive than commercial vocabularies, and also contain specific terms, increasing the likelihood of a lower rate of recognition.
- The dialogue between the user and the system is much simpler, since the system must only generate a text from the user's speech, with no questions to be made (through speech) to the user.
- The handling and prevention of mistakes are quite different. The system must be able to show, somehow, the words in the text that are misunderstood, but must not interrupt the user or ask for confirmation. The radiologist does not wish to be interrupted by an error message while they are dictating a report.
- The quality of the messages and the adequacy of the outcoming speech should be replaced by the text's accuracy.

There are some other important considerations applicable to the automatic report transcription system.

- Flaws in the report that were caused by a wrong recognition of words can be catastrophic to the patient, for it can lead to an inaccurate diagnosis.
- It is necessary to think of how to create a methodology that takes into consideration the main requirements for this kind of system.

The challenges found when assessing automatic report transcription systems are:

- Define the VUI requirements that have to be considered to evaluate this system;
- Determine which, among the many requirements presented, are said to be essential and viable of evaluation;
- Determine how to measure each requirement pointed out as essential for these systems;
- Define how to evaluate these systems in a viable way, with acceptable costs and time for the Health Care organizations and/or systems supplier;

Thus, any suitable evaluation methodology for these systems must take into account the issues mentioned above in order to decide if a product is appropriate in matters of use efficiency, user's satisfaction and functionality, or if the product has only a good rate of voice recognition.

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

This article is organized as follows: The following part points out the material and methods used in the research. After that, a critical review of the concepts involved in this work will be made: Voice recognition and synthesis, technology evaluation, automatic transcription system for report and its specific demands. Then, a methodology for the evaluation of these systems is described. In the end, the limitations and advantages of the proposed methodology are discussed.

Materials and Methods

This article is enclosed in a wider context of a doctorate's research for the evaluation of user interface based on voice with the purpose of accomplishment of the following activities:

- Bibliographic review of the themes in use in the Project, including: VUI, radiology information systems with VUI and traditional methods of usability evaluation;
- Identification of generic demands for the users interfaces based on voice and the users interface requirements for voice-based system in Health Care, especially for radiological information systems.
- Proposition of a methodology that is able to provide usability evaluation focused on automatic transcription system for radiology reports.

Conceptual Foundations

Voice Recognition

The main characteristic of the VUI applications is the interaction, through voice, of a user with a system. This kind of interface includes elements such as: *prompts* or system messages, grammar and logical dialogue or call flow. Prompts are all the pre-recorded or synthesized voice messages that must be executed during the dialogue with the user. Grammar defines all the words, sentences or phrases that can be said by the user in answer to a prompt. The logical dialogue defines all the actions that should be taken by the sys-tem in a specific moment of the interaction, such as the access to the database [8, 20, 21].

According to [22], when an application with a users interface based on voice is developed, there are some issues that shouldn't be neglected for the application to be successful.

- The vocabulary affects the voice recognition through its size and subject field coverage. So, an extensive vocabulary with good subject field coverage is appealing, for it is capable of recognizing more words. Nevertheless, smaller vocabularies provide an enlargement in recognition accuracy.
- The users influence the voice recognition through clarity and consistency in the pronunciation of words. User- dependent systems have a higher rate of voice recognition than the systems that are user independent, but the formers need training sessions and can be more sensitive to noise, microphone and voice variations.
- A noisy environment affects the voice recognition in two ways: a) voice signal distortions cause difficulty to distinguish the pronounced words; b) when in a noisy environment, users tend to alter their voices and doing so, cause distortion in or alter the voice signal.
- All voice recognition systems are based on the statistical standards principles. However, in spite of its similarities, systems differ in their voice signal parameterization, the acoustic model of each phoneme and the language pat-

tern used in the choice of words accord with the words spoken and stored previously. Thus, many systems bring about differences in relation to the recognition errors, even when they have similar rates of recognition.

Voice Synthesis

Voice synthesis is the process that converts text into voice. The synthesizer receives a piece of text in digital form and vocalizes it. A Voice Synthesis program is useful to vocalize information that comes from data base queries and when the user cannot divert his attention to read something or does not have access to the written text; a system with the users interface based on voice can use a module for voice synthesis or use pre-recorded messages when there is no variation in the information given to the user [8].

It is worth noting that until now, the voice synthesizers cannot represent intonation and are still quite poor when compared to voice dialogue among humans.

Usability in Voice Recognition Systems

Usability is a system quality requirement that contains aspects related to the efficiency when using the system, ease of learning, subjective satisfaction from the user and adequacy to specific patterns; it is the process of assuring interface usability and guarantee that the user's demands be met [12, 23- 25]. Although the aspects for usability mentioned above are conceptually clear, it is difficult to use these definitions in practice. When the evaluation is made through empirical studies, the researchers need to decide about metrics for each factor [26].

If the companies in general have not been preoccupied about following usability patterns in its websites, established many years ago, these issues are even more serious in VUIs, for it is an even newer and less settled form of interaction with the user.

General Usability Requirements for Voice Recognition Systems

One way of evaluating voice recognition systems usability is through general heuristics proposed by Nielsen [12] such as: simple and natural dialogue, use of feedbacks and handling and preventing errors. Nevertheless, more specific criteria are necessary to evaluate VUI specific issues, such as the ones proposed by [27- 31]. Those criteria include:

- Output phrasing adequacy: The output content of the system must be correct, relevant and informative enough, without providing information overload to the user. The system's way of communicating with the users must be clear and unambiguous and the language must provide an appropriate and familiar terminology to the user.
- Output voice quality: this quality is connected to issues of clearness and intelligibility (right intonation, emotion, appropriate speech pace and pleasure when heard).
- Input recognition adequacy: appropriate voice recognition means that the system rarely misunderstands the user's entry. But this is associated to many factors in the environment (as level of noise) and also to user's factors: gender, age, accent, depth or shrillness of voice and the voice quality as received by the system.
- Adequacy of dialogue initiative: it is necessary for the system to choose, in a reasonable way, the dialogue initiative established between it and the user. This is related to the user's level of knowledge of the system.

Voice Recognition Systems in Radiology

One of the main problems shown in the literature [16, 17, 32-35, 37] is the delay in radiology reports due to the time spent from the moment of entry of a 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.

Specific Usability Requirements for the Automatic Transcription System for Radiology Reports

In addition to the general requirements for interfaces and to the general requirements found in voice recognition systems, there are specific requirements for automatic transcription system for radiology reports. We propose that the following set of requirement should be formally assessed when evaluating VUI-based report transcription systems

1. Accuracy: It is one of the most important requirements, because the wrong information can compromise report quality, alter a diagnosis and compromise a treatment.

2. Vocabulary Extent: It is a very important requirement, as the vocabulary can neither be too big in order to lower the rate of word recognition nor too small for it not to consider the words in the application's dominion.

3. Specific Dictionary for RIS: The system must consider the words used daily in radiology reporting

4. Hospital environment: depending on the area, hospitals can be very noisy, but that should not interfere with the efficiency of recognition.

5. Continuous Recognition: the user must be able to dictate the report naturally, without having to worry about pauses between words, i.e., the user must be able to speak in a natural and continuous way.

6. Desirable separation between the keyboard and the dictation system: The user should be able to dictate through a specific voice capture device, through a cellular or regular phone, allowing the application to be ubiquitous;

7. Use of client-server or browser-server architectures: For the radiologists to be free to move from one station to another in a hospital or clinic, or even across hospital units in a network of health care providers:

8. Integration with exiting systems: PACS, HIS e RIS;

9. Time for the report to be ready: must be at least shorter than the human transcription systems;

10. User's naturalness of speech: The user must be able to speak in a natural and continuous way, as if the user were recording an audio tape;

11. Resolution of ambiguity for homonyms: Words that have the same pronunciation but different spelling should not affect the application.

Proposed Methodology

The proposed evaluation methodology should be able to:

- Use additional usability and inspection tests to provide a lower cost and shorter assessment time.
- Be applied to previously implemented systems;
- Function as a guide to evaluating the usability in this class of systems;
- Investigate the difficulties in evaluating specific requirements;
- Group the proposed requirements according to their characteristics;
- Propose metrics for evaluating each of those requirements.

This methodology proposes that the interface evaluation should be done by inspection[1] whenever possible, without involving the user in order to decrease the usability tests session prices. Use the usability tests when it is verified that the inspection is not enough.

To assess automatic report systems, the following classes were here defined in a modified way from what Möller [31] proposed in his work about general purpose voice recognition system evaluation:

- 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.

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 1, for Client's Satisfaction.

Table 1 – Template of Client's Satisfaction

CLIENT'S 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

The corresponding questionnaire was based on SUMI (Software Usability Measurement) of University College Cork.

Conclusion

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.

A 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.

As future work, this methodology should be applied to several case studies in order to be perfected and validated for real cases.

References

[1] Allen J, Perrault C. Analysing intention in utterances. Artificial Intelligence 15, 143– 178, 1980.

[2] Kamm C, Walker M, and Rabiner L. The role of speech processing in human computer intelligent communication. Speech Communication 23 (1997), pp. 263–278.

[3] Price, P. Evaluation of spoken language systems: the ATIS domain. Proceedings of the Third DARPA Speech and Natural Language Workshop, Morgan Kaufmann, 1990.

[4] Philips, http://www.dictation.philips.com/

[5] IBM, http://www.ibm.com/us/en/

[6] Nuance Dragon Naturally Speaking Solutions, http://www.nuance.com/naturallyspeaking/.

[7] Mctear MF. Spoken Dialogue Technology: Enabling the Conversational User Interface, ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 90–169.

[8] Cohen MH., Giangola JP, Balogh J. Voice User Interface Design, Addison Wesley, 2004.

[1] Usability inspection is a set of methods where an evaluator inspects a user interface. It can generally be used early in the development process by evaluating prototypes or specifications for the system that can't be tested on users. It generally considered to be cheaper to implement than testing on users [35].

[9] San-Segundo R, Montero JM, Macías-Guarasa J, Ferreiros J, Pardo JM. Knowledge-Combining Methodology for Dialogue Design in Spoken Language Systems, Int J of Speech Tech 8, 45-66, Springer Science + Business Media, 2005.

[10] Shneiderman B. The Limits of Speech Recognition, Communications of the ACM, Vol. 43, No 9, pp 24 – 27, 2000.

[11] Snomed. http://www.ihtsdo.org/snomed-ct/.

[12] Nielsen J. Usability Enginnering. Academic Press, Cambridge, MA, 1993.

[13] Walker MA, Litman D, Kamm C, and Abella A. Evaluating spoken dialogue agents with PARADISE: Two case studies. Comput Speech and Language 12,3, 317–347, 1998.

[14] Gibbon D, Moore R., and Winski R., EDS. 1997. Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, New York, NY.

[15] Dybkjaer L., Bernsen NO, and Dybkjaer H. 1997. Generality and objectivity: central issues in putting a dialogue evaluation tool into practical use. In Interactive Spoken Dialog Systems. Proceedings of a Workshop Sponsored by the Assoc for Computational Linguistics (Madrid, Spain), J Hirschberg, C Kamm and M Walker, Eds. ACL, 17–24.

[16] Kanal KM., Hangiandreou NJ, Sykes AMG, Eklund HE, Araoz PA, Leon JA., Erickson BJ. Initial Evaluation of a Continuous Speech Recognition Program for Radiology, J of Digital Imaging, Vol. 14, no 1 (March), 2001: pp 30-37.

[17] Kimberly DV. A Methodology of Error Detection Improving Speech Recognition In Radiology, Springer, 2001.

[18] Voll K, Atkins S, Forster B. Improving the Utility of Speech Recognition Through Error Detection, Journal of Digital Imaging, Vol. 21, No 4, 2008, pp 371-377.

[19] Paulett JM, Langlotz CP. Improving language models for radiology speech recognition, Journal of Biomedical Informatics, 42 (2009), 53-58.

[20] Hunt A, Walker WA. fine Grained Component Architecture for Speech Application Development. 2000.

[21] Lauesen S. User Interface Design – A Software Engineering Perspective, Pearson Education Limited, Great Britain, 2005, ISBN 0 321 18143 3

[22] Alapetite A., Boje AH, Morten H. Acceptance of speech recognition by physicians : A survey of expectations, experiences, and social influence, International journal of human-computer studies, ISSN 1071-5819, vol. 67, no1, pp. 36-49, 2009.

[23] Avouris NM., An introduction to software usability. In Workshop on Software Usability – Proceedings of the 8 th Panhellenic Conference on Informatics. v. 2, p. 514-522, 2001.

[24] Sommerville I. Software Engineering – 6th Edition, Addison Wesley, ISBN 0-201-39815-X, 2001.

[25] ISO 9241-11. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) Part 11: Guidance on usability. ISO 1997

[26] Skov MB, Stage J, Supporting problem identification in usability evaluations, Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future, November 21-25, 2005, Canberra, Australia.

[27] Dybkjaer L, Bernsen NO. Usability Evaluation in Spoken Language Dialogue Systems, Proceedings of the ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems, 2001.

[28] Salvador VFM, Oliveira Neto JS, Kawamoto AL. Requirement Engineering Contributions to Voice User Interface. In: First International Conference on Advances in Computer-Human Interaction, 2008, Sainte Luce. First International Conference on Advances in Computer-Human Interaction, 2008. p. 309-314.

[29] Komatani K, Ueno S, Kawahara T, Okuno HG. Flexible Guidance Generation using User Model in Spoken Dialogue Systems, Proceedings of the 4lst Annual Meeting of the Association for Computational Linguistics, 2003.

[30] Walker MA., Passnneau R., Boland JE. Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems, In: Proceedings of the 39rd Annual Meeting on Association for Computational Linguistics, Toulouse, France, 2001.

[31] 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, 142-153, 2002.

[32] White KS., Speech recognition implementation in radiology, Springer-Verlag, 2005.

[33] Bhan SN, Coblentz C, Norman GR, Ali AH. Effect of Voice Recognition on Radiologist Reporting Time, CARJ Vol 59, No 4, October 2008.

[34] Durling S, Lumsden J. Speech Recognition Use in Healthcade Applications. Proceedings of MoMM2008, MoCoHe, 2008 ACM 978-1-60558-269-6/08/0011

[35] Nielsen J. Usability Inspection Methods. New York, NY: John Wiley and Sons, 1994

Address for correspondence

Valéria Farinazzo Martins SalvadorRua Maranhão, 43, apto 114, São Paulo – SP, CEP 01240-001 Brazil
E-mail: valeria.farinazzo@mackenzie.br