An ISO-Based Quality Model for Evaluating Mobile Medical Speech Translators

Nikos Tsourakis[1] and Paula Estrella[2]

[1] ISSCO/TIM/ETI, University of Geneva, Switzerland
[2] FaMAF, Universidad Nacional de Córdoba, Argentina
Nikolaos.Tsourakis@unige.ch, pestrella@famaf.unc.edu.ar

Abstract. Medical translation systems present an intriguing research area as language barriers can become life-threatening when health issues come into place. There is however a lack of common evaluation techniques, making the fair comparison of such systems a difficult task. In this work we try to remedy this deficiency by proposing a quality model based on the ISO/IEC 9126 standard that could serve as a comparison basis among homologous systems. We focus on the mobile world believing that it suits patients' needs better, as they experience diverse scenarios along the pathway to healthcare. Our work involves the definition of the quality characteristics of the model along with the quantification of their importance based on two target groups of users (12 doctors and 12 potential patients) that demonstrate different needs and goals towards the system.

Keywords: Medical Translation, Quality Model, ISO 9126, Mobile Translators.

1 Introduction

Language barriers often cause inconvenience but when medical issues are involved can become life-threatening. Quantitative studies, e.g. [1], have shown that lack of a common doctor-patient language correlates with an increased probability of negative outcomes. Unfortunately, trained medical translators are both scarce and expensive. Even if a universal speech-to-speech translator still seems an insurmountable problem, the substantial gap between the need for and availability of language services in health care could be bridged through effective medical speech translation systems, such as [2]; a system like this would be far more useful to users if it was available on a hand-held device. Indeed, different systems already are efforts towards the deployment of mobile speech-to-speech translation applications [3], [4], [5].

During the lifecycle of these systems authors provided evaluation results leveraging various computer and human centered metrics. Despite some early efforts towards a common evaluation framework [6] we argue that there is a lack of such methodology that would provide a fair comparison framework for different mobile medical translation systems. Additionally, the lack of appropriate quality assessment techniques can deteriorate user satisfaction. As quality is hard to assess and assure,

K.S. Nikita et al. (Eds.): MobiHealth 2011, LNICST 83, pp. 275–283, 2012.

several models try to address software quality issues by employing a set of quality attributes, characteristics and metrics [7], [8], [9], [10]. In this work we discuss how to evaluate mobile medical translators with a quality model based on ISO/IEC 9126 [11]. Unlike other models, it enjoys the benefits from being an international standard and as it is generic, it can be applied to any kind of software product.

Our work had two stages. Initially we had to create the quality model per se, defining the quality characteristics that constitute the model, either by selecting them among those proposed in ISO/IEC-9126 or by introducing new ones. In the second phase we asked two target groups of users that demonstrate different needs and goals towards the system (12 doctors and 12 potential patients), to quantify their preferences concerning which attributes (i.e. quality characteristics of the model representing desired features of a system) are more important.

The paper is organized as follows: in Section 2 we discuss how a system like this should be realized in a hospital environment. In Section 3 we decompose the model according to our case study. Section 4 presents our methodology for ranking the quality characteristics and Section 5 summarizes the results along with a short discussion. The final section concludes.

2 The Pathway to Healthcare

The path to healthcare as described in [12] may involve different stages besides the typical diagnosis scenario between the doctor and the patient. As illustrated in Fig. 1, we can imagine a patient interacting with other staff in the hospital, for example with a secretary at the welcome reception desk, with a nurse during an examination procedure or hospitalization, etc. All these diverse scenarios indicate just the gamut of possible situations.

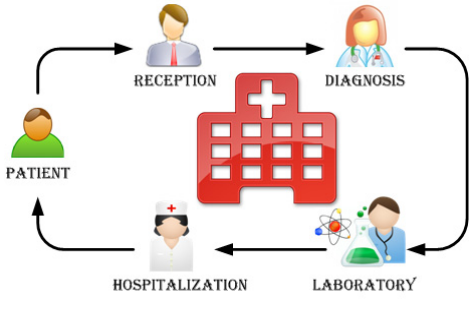

Fig. 1. A typical pathway to healthcare

A fundamental question introduced by Somers [12] is: “Who is the primary user of such a system, the physician or the patient?” On the one hand, there is the doctor who usually has high level education and interacts with the system on a daily basis and conversely, the patient who may use the system solely once in his life. There is no single answer to this question: while many efforts have put the doctor in charge of the dialogue e.g. [2], [13], others have followed a parity oriented approach, where two

separate graphical user interfaces are offered for each one of the two parties [5]. Thus, we should address the following:

- Quality of translation. The genre of the task requires safe critical high quality translation.
- Heterogeneous interaction. Users may interact with different personnel or in diverse environments.
- Mobility. Interaction happens with a mobile device, which per se involves special consideration.
- User physical constrains. Patient's physical disabilities can pose hurdles to the efficient usage of the system.
- Wireless interconnection. The network can cause delays or even connection failures.
- Application availability. It could be preinstalled on a hospital's device or users could install it on their own device.

3 Decomposition of Our ISO Model

A generic quality model is proposed by ISO/IEC 9126 [11] for the evaluation of any software product, thus it can also be used in the evaluation of mobile medical speech translators. We described each external attribute with a friendly, application-specific statement to help doctors or potential users first understand attributes and then weight them; this is based on the hypothesis that these users are not necessarily familiar with the ISO terminology. In the second phase the quality characteristics were compared in pairs and we extracted the corresponding weights by adopting a methodology similar to [14], which uses a mutual comparison method [15] and multi-criteria Analytical Hierarchy Process (AHP) [16]. The customized definitions are summarized below; unless otherwise stated all the quality characteristics were adapted form the ones proposed in ISO/IEC 9126.

1. Functionality

Suitability. The system can be seen either as a replacement when no interpreters are available or as a palliative before resorting to an interpreter. It should therefore support as many languages, domains and diverse usage environments as possible.

Accuracy. In this specific application domain the translation between languages needs to be produced in the most reliable and robust way, achieving high quality.

Interoperability. The system should facilitate interoperability of the different nodes in the pathway, e.g. information about prescribed medication or treatment must be available to the corresponding personnel or system.

Security. The system should guarantee that the information gathered during the interaction is stored and accessed in a restricted manner. Treatment of sensitive medical data should be carefully considered.

Traceability (added attribute). User's activity along the pathway should be traced by the system and may be used to identify possible problems (e.g. delays), perform correct pricing of praxes, etc.

Exploitability (added attribute). Users may be forced to wait their turn for an examination or wait between examinations. This idle time can be used for familiarizing with the application, and thus fostering user's trust.

Controllability (added attribute). If the doctor gives instructions he should also be control of the dialog flow but during diagnosis the weight of control should be equilibrated between the two parties. This also conforms to current clinical theory of patient-centered medicine [17].

2. Reliability

Maturity**.** The special nature of the application demands zero faults therefore the system should aim to minimize the frequency of failures.

Fault tolerance. In case of any faults the system should resort to a backup plan, e.g. trained personnel could take over control and interact with the patient.

Recoverability. The software should be able to recover after a failure either by incorporating a logging mechanism or by storing user's data locally or remotely.

1. Usability

Understandability. Users should understand what the system is supposed to do. Short introductions should be provided along with context dependent prompts. Cultural limitation or physical disabilities should also be addressed.

Learnability. As in any spoken dialogue application, it should give users immediate feedback on the system's intended coverage, particularly when recognition fails.

Operability. As the end user may use a system like this only once in his life the interaction should be based on simplicity.

Attractiveness. Due to the limited lifecycle of the application (used only in a hospital environment), the issue of attractiveness becomes of lesser importance. However, the success of the system may depend on relevant factors.

Uniformability (added attribute). Following the path of healthcare each user should experience a uniform interaction. This will minimize the effort of learning how the system works in different situations and will cause less confusion.

Trustability (added attribute). The system should offer results that are predicable and don't engender any surprise to end-users, so that patients establish trustful relations towards the system.

Customizability (added attribute). As users may vary in a range of literate to complete illiterate, the system should take this into account as well as other special needs (weak sight, hearing problems, etc).

Privacy (added attribute). The system should contemplate issues of privacy, e.g. patients can be reluctant to talk in front of other, even in front of relatives, be embarrassed when using the system unsuccessfully, etc.

2. Efficiency

Time behavior. Time management is very important as the diagnosis should be made as quick as possible. It should also be dependant on the usage scenario, as it may be more urgent to complete a task at the reception than at the laboratory.

Resource utilization. The system should target to efficient utilization of resources (e.g. battery life, wireless connectivity, data access) and also fair sharing among users.

3. Maintainability

As our work is pertinent to external evaluation we won't delve into these attributes, which reflect mainly a technical (i.e. internal) viewpoint, such that of developers.

4. Portability

Adaptability. It should be adaptable to a number of platforms. Proprietary solutions may narrow the possible options, so the design should take into account forthcoming technologies and open source alternatives.

Installability. If end users decide to install the system on their own device this should be as transparent as possible considering that the application's life time could be limited to just the time the patient stays in the hospital.

Co-existence. The system should successfully co-exist with other independent systems working in a common environment and sharing common resources. Issues of conflicts may include the bandwidth usage, interference problems, other wearable medical devices, etc.

Replaceability. As most of the times upgrading the existing software should not be performed by end-users, issues of replaceability are not a subject of their concern.

5. Compliance (for all characteristics)

In our scenario it could happen that the interaction in environments that impose zero noise level may be prohibited, thus being subject to specific hospital regulations. Furthermore, issues related to interference in specific areas should be considered.

4 Relative Importance of Attributes

The relative importance of each quality characteristic in the model is dependent on the user and, as expected, user's perception about product quality varies across user types. For example, end users typically value usability more than developers do. In order to quantify users' preferences we polled two groups that have different needs and goals towards the system. The first group included 12 professional doctors having a different specialization background (excluding specializations that don't involve direct contact with patients) and a second group of 12 non-doctors with different higher academic background. All participants were between 20-40 years old and gender was approximately balanced across conditions.

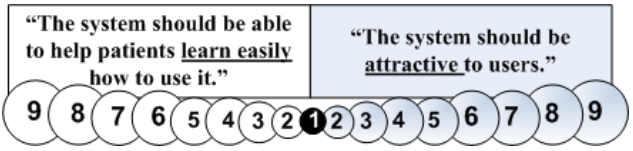

Fig. 2. Learnability vs. attractiveness

For the construction of the survey we considered the fact that participants have limited or no experience with speech-to-speech translation systems, they have no familiarity with the ISO hierarchy and terminology and that they have a busy schedule, so we limited the time devoted to the survey to around 15-20 minutes. Participants were asked to express their opinion by choosing a number in a scale of 1-9 favoring the feature they liked most. An example is shown in Fig 2.

The mutual comparisons were limited within attributes in the same category. For *n* characteristics at a given category *n(n-1)/2* mutual comparisons are needed (e.g. 21 comparisons for *functionality*). Our analysis was based on the Analytical Hierarchy Process [16], which shapes a problem in a hierarchical structure. Accordingly, we shape our ISO model in three levels, where the goal (first level) is to evaluate a mobile speech translator; the quality characteristics of the model constitute the second level (e.g. *functionality*) and the sub-characteristics the third one (e.g. *accuracy*).

5 Results and Discussion

The results of applying the AHP are presented in Table I. For the first group (doctors) the weights are depicted in the left side and for the second group (patients) in the right one with the grey background. From the high level attributes of the model *functionality* seems to be the most important for physicians (38.53%) followed by r*eliability* (18.04%) and *usability* (17.47%). For patients *efficiency* shows the highest weight (30.32%) and surprisingly *usability* the lowest (9.53%). One explanation could be that patients value features related to performance (e.g. response time) more than the ones related to ease of use. Paradoxically, *compliance* (with hospital regulations) is considered more important by patients than by doctors.

At the second level, results corroborated our intuition that *accuracy* is of utmost importance for both target groups (26.29% and 39.17%); it is followed by *security* (22.27% and 17.01%) and last by *exploitability* (5.40% and 4.26% respectively). Also, physicians seem to care about privacy issues more than patients do (22.63% vs. 14.9%), whereas the latter prioritize *customizability* (related with users' special needs). *Attractiveness*, receives the lowest rank among all evaluators.

Another interesting finding is that physicians consider *co-existence* very important, as a newly introduced system shouldn't affect the systems already deployed. Patients

on the other hand prioritize the *replaceability* of the system, despite the fact that they are normally not involved in this process. Finally, both groups agree on the sub-attributes of *efficiency* (clearly favoring *time behavior,* the ability of the system to respond quickly), and they also adopt the same stance for *reliability* prioritizing the elimination of failures (*maturity*). Moreover patients seem to consider the *recoverability* more important than the *fault tolerance*.

Table 1. Weighted Quality Model

Quality factor	Weight %		Quality sub-factor	Weight %	
	Doctors	Patients		Doctors	Patients
Functionality	38.53	10.88	Suitability	18.18	12.28
			Accuracy	26.29	39.17
			Interoperability	10.38	13.37
			Security	22.27	17.01
			Traceability	9.26	8.61
			Exploitability	5.40	4.26
			Controllability	8.22	5.30
Reliability	18.04	13.47	Maturity	55.80	48.70
			Fault tolerance	38.50	6.20
			Recoverability	5.70	45.10
Usability	17.47	9.53	Understandability	4.31	7.63
			Learnability	10.24	9.61
			Operability	15.83	17.42
			Attractiveness	2.37	3.30
			Uniformability	9.43	3.97
			Trustability	11.87	19.91
			Customizability	23.32	23.26
			Privacy	22.63	14.90
Efficiency	5.64	30.32	Time behavior	78.38	69.70
			Resource utilizat.	21.62	30.30
Portability	6.40	15.74	Adaptability	22.14	18.95
			Installability	14.31	15.54
			Co-existence	40.08	26.65
			Replaceability	23.47	38.86
Compliance	13.92	20.06			

As in every human assessment, coherence and consistency are an important matter. This issue is taken into care of calculating a *Consistency Ratio* (CR) provided by AHP (the lower the better), which quantifies how much the evaluators' judgments fulfill the *transitive* property (i.e. if a>b and b>c then a>c). Initially, we started with more than 12 subjects in each of the two target groups; the survey was sent to 16 doctors and 20 potential patients. However, when calculating the CR on the entire dataset for each group we found its CR to be too high, risking useless results. There is also a correlation of individual consistency with the overall consistency of the model. As we eliminated participants with the highest CR, the overall CR (the one obtained after averaging their answers) dropped to an acceptable 10% (approximately). Hence, the study was limited to the 12 most consistent physicians and to the 12 most consistent patients.

It is worth noting that not every subject chosen for the study had CR < 10%. We believe that this is strongly dependent on the number of items under comparison. *Usability* for example, demands 28 pair-wise comparisons, hindering consistent subjective judgments. Another potential source of inconsistency is the formulation of each statement: for high level characteristics the statements embedded multiple concepts and it was there where we encountered most of the inconsistencies. We also observed that physicians exposed lower levels of inconsistency.

Finally, upon completion of the survey each participant was asked to express his/her opinion on different topics, to propose enhancements or to point out deficiencies. In Fig. 3 we present the answers to four of the questions, being *intention of usage*, *intention of buying*, *preference over human interpreters*, *efficient use of system by patients*. Even if both groups seem eager to use a system like this; patients seem reluctant to buy it (only 31% are positives). Less than 31% in both target groups express a clear preference (by answering "Yes") for the system over a human interpreter and lastly, less than half of the participants believe that the system can be used efficiently by the patients.

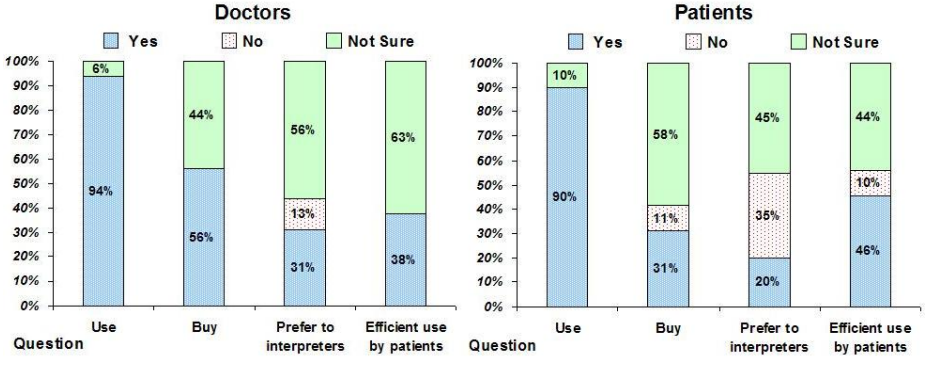

Fig. 3. Subjective opinions of both target groups

6 Conclusions

We defined an ISO quality model suited for mobile medical speech translation systems, which can help evaluators, compare similar systems on a common evaluation ground and could also help developers focus on those aspects of quality that users deem important.

We tried to address some issues related to the design and the implementation of surveys for acquiring the relative importance of attributes and sub-attributes (i.e. weights). We also provided some guidelines that might be useful to others intending to use a similar protocol: formulation of statements, number of comparisons, understanding the tenor of the problem, number of participants and scale for the comparisons, are some of the factors that should be carefully considered.

Finally, the next step of this work involves further decomposing each sub-characteristic if necessary and the accumulation of relevant metrics. Additionally, including weights for the internal attributes is important for the completeness of the model.

References

1. Flores, G.: The impact of medical interpreter services on the quality of health care: A systematic review. Medical Care Research and Review (2005)
2. Bouillon, P., Flores, G., Starlander, M., et al.: A Bidirectional Grammar-Based Medical Speech Translator. In: SPEECHGRAM Workshop, Prague Czech Republic (2007)
3. Zhang, Y., Vogel, S.: PanDoRA: a large-scale two-way statistical machine translation system for hand-held devices. MT Summit, Copenhagen, Denmark (2007)
4. Gao, Y., Zhou, B., Zhu, W., Zhang, W.: Handheld Speech to Speech Translation System. Automatic Speech Recognition on Mobile Devices and over Commun. Networks (2008)
5. Tsourakis, N., Bouillon, P., Rayner, M.: Design Issues for a Bidirectional Mobile Medical Speech Translator. In: SiMPE Workshop, Bonn, Germany (2009)
6. Rayner, M., et al.: A Small-Vocabulary Shared Task for Medical Speech Translation. In: Workshop on Speech Processing for Safety Critical Translation, Manchester, UK (2008)
7. Chidamber, S.R., Kemerer, C.F.: A Metrics Suite for Object Oriented Design. IEEE Transactions on Software Engineering (1994)
8. Wakil, M.E., Bastawissi, A.E., Boshra, M., Fahmy, A.: Object Oriented Design Quality Models – A Survey and Comparison. In: Int. Conference on Informatics and Systems (2004)
9. Hyatt, L.E., Rosenberg, L.H.: Software Metrics Program for Risk Assessment. Elsevier Acta Astronautica 40, 223–233 (1997)
10. Bansiya, J., Davis, C.G.: A Hierarchical Model for Object-Oriented Design Quality Assessment. IEEE Transactions on Software Engineering 28, 4–19 (2002)
11. ISO/IEC. 2001. ISO/IEC 9126-1:2001 (E) - Software Engineering - Product Quality - Part1: Quality Model (2001)
12. Somers, H.: Theoretical and methodological issues regarding the use of Language Technologies for patients with limited English proficiency. TMI (2007)
13. Narayanan, S., et al.: The Transonics spoken dialogue translator: An aid for English-Persian doctor-patient interviews. Dialogue Systems for Health Communication (2004)
14. Behkamal, B., Kahani, M., Akbari, M.: Customizing ISO 9126 quality model for evaluation of B2B applications. Journal of Information and Software Technology (2008)
15. Firyaki, F., Ahlatcioglu, M.: Fuzzy stock selection using a new fuzzy ranking and weighting algorithm. Applied Mathematics and Computation Journal (2005)
16. Saaty, T.L.: The Analytical Hierarchy Process, 2nd edn. McGraw-Hill, New-York (1994)
17. Stewart, M., Brown, J.B., Weston, W.W., McWhinney, I.R., McWilliam, C.L., Freeman, T.R.: Patient-centered medicine: Transforming the clinical method (2003)