JHI JOURNAL OF HEALTH INFORMATICS

Uma metodologia para avaliação da usabilidade para sistema de transcrição automática de laudo em radiologia

An evaluation methodology for automatic transcription system of radiology reports

Una metodología para el sistema de evaluación de usabilidad de la transcripción automática de informes de radiología

Valéria Farinazzo Martins[1], Lincoln de Assis Moura Junior[2]

RESUMO

Este trabalho relaciona elementos das áreas de Computação e Saúde para comporem a elaboração de uma metodologia para avaliação de usabilidade de sistemas de transcrição automática de laudos na área de Radiologia. Esta metodologia é elaborada a partir de requisitos gerais de Voice User Interface e nas peculiaridades dos sistemas de transcrição automática de laudos, sendo, posteriormente, validada através de inspeções e testes de usabilidade, realizados fora do ambiente hospitalar, para, então, ser também aplicada a um hospital da cidade de São Paulo.

Descritores: Informática Médica; Registros Médicos; Sistemas de Informação em Radiologia

ABSTRACT

This work combines knowledge from Computer Science and Health Science in order to propose an evaluation methodology for Automatic Transcription Systems of Radiology Reports. This methodology was designed based on Voice User Interface requirements and specific requirements of automatic transcription systems of Radiology report. This methodology was previously validated through some inspections and usability tests outside the hospital environment and afterward the methodology was used in a hospital in São Paulo city.

Keywords: Medical Informatics; Medical Records; Radiology Information Systems

RESUMEN

Este trabajo enumera los elementos de las áreas de Informática y Salud, que conforman el desarrollo de una metodología para la evaluación de la usabilidad de los sistemas de transcripción automática de informes en el ámbito de la Radiología. Esta metodología se extrae de los requisitos generales de la interfaz de usuario de voz y las peculiaridades de los sistemas de transcripción automática de informes, y posteriormente validados por las inspecciones y las pruebas de usabilidad a cabo fuera del hospital, a continuación, también se aplicará a un hospital de São Paulo.

Descriptores: Informática Médica; Registros médicos; Sistemas de Información Radiológica

[1] Mestre em Ciência da Computaçao. Universidade Presbiteriana Mackenzie e Escola Politécnica da Universidade de São Paulo - USP, São Paulo (SP), Brasil.
[2] Doutor em Engenharia Elétrica. Zilics e Health e Escola Politécnica da da Universidade de São Paulo - USP, São Paulo (SP), Brasil.

Autor Correspondente: **Valéria Farinazzo Martins**
e-mail: **valeria.farinazzo@mackenzie.br**

Artigo recebido: **23/02/2011**
Aprovado: **18/05/2011**

INTRODUÇÃO

Desde que os usuários finais de sistemas computacionais passaram a ser pessoas leigas em Computação, a área de Interface Homem-Computador (IHC) passou a ter um papel fundamental no sucesso de produtos computacionais no mercado. Assim, além de atender às funcionalidades desejadas, uma aplicação deve ter interfaces intuitivas e "amigáveis" para estes usuários.

O uso de tecnologias emergentes nas interfaces é algo que o mercado tem vivenciado continuamente: interfaces *touchscreen*, navegação em ambiente tridimensional e uso de voz como forma de interação com os equipamentos são apenas alguns destes exemplos.

Embora a área de reconhecimento de voz em interfaces mais intuitivas tenha surgido na década de 50[1-2], somente na década passada é que foi possível desenvolver, comercialmente, sistemas que realmente pudessem ser utilizados na prática. Há, ainda, várias lacunas em como se criar e se avaliar sistemas com reconhecimento de voz[1]* de maneira mais eficaz e sistemática.

O estado da arte em avaliação de usabilidade para estes sistemas advém de pequenas contribuições de avaliações que foram desenvolvidas para projetos específicos, e que tentam generalizar e propor recomendações para tais classes de aplicações, como PARADISE[3], EAGLES[4] e DISC[5].

Analisando, especialmente, a área da Saúde, a utilização de reconhecimento de voz em sistemas da Saúde de propósito geral, tais como em emergência, não tem sido eficiente devido ao grande vocabulário do domínio - sabe-se que o vocabulário comum da área é de mais de 100 mil itens. Além disto, as possibilidades diagnósticas presentes no SNOMED[6], por exemplo, são mais de 60 milhões. Ou seja, a informação disponível na área da Saúde é extremamente variada.

Assim, a tecnologia de reconhecimento de voz tem sido usada para propósitos mais específicos, como em sistemas de transcrição automática de laudos (STAL) na área de Radiologia. Isto significa que o vocabulário é consideravelmente menor, proporcionando uma precisão mais alta no reconhecimento dos termos específicos. Embora a usabilidade seja um atributo de qualidade de software que visa garantir que os requisitos dos usuários sejam atendidos[7-8], os sistemas que utilizam reconhecimento de voz na área da Saúde, não são, na sua grande maioria, analisados seguindo métodos completos, complexos e bem-estabelecidos de avaliação de usabilidade. Não há trabalhos suficientes na literatura que estabeleçam os requisitos específicos da área que devem ser atendidos de forma a tornar o uso de reconhecimento de voz efetivo e eficaz.

Estudando as metodologias existentes para a avaliação de usabilidade de sistemas de reconhecimento de voz, num âmbito geral, percebe-se que muitos dos requisitos não podem ser utilizados para se avaliar os STALs, tai como: o diálogo existente entre a máquina e o usuário, por voz; o direcionamento do diálogo a fim de que o usuário diga o que o sistema espera receber como entrada; a amabilidade do pedido da máquina – por voz; esclarecimentos sobre a tarefa que a máquina possa fornecer, por voz. Além disso, a maioria das metodologias existentes utiliza métodos automáticos de verificação de *logs* e gravações que tentam responder às métricas estabelecidas, relacionadas à eficiência do diálogo, qualidade do diálogo, sucesso da tarefa e satisfação do usuário. Ou seja, não englobam, diretamente, a participação do usuário final na sua avaliação, além de estarem preocupadas com questões mais ligadas à meta-comunicação, que foge ao contexto dos sistemas aqui avaliados[9].

Frente ao exposto, torna-se necessário o desenvolvimento de uma metodologia para analisar a usabilidade de STAL que possa abordar não somente falhas no reconhecimento da voz, mas que possa tratar questões mais amplas, como a satisfação do usuário, o fácil aprendizado do sistema e a adaptação do sistema ao nível de conhecimento do usuário.

Este artigo apresenta uma metodologia para a avaliação de usabilidade de STAL e tem como objetivos principais: a utilização de testes de usabilidade e inspeção de usabilidade de maneira complementar a fim de propiciar um menor custo e um menor tempo de avaliação; possibilidade de ser aplicada a sistemas já implementados; a investigação da viabilidade de se avaliar determinados requisitos; o agrupamento dos requisitos propostos segundo classes de características; e a proposta de métricas para avaliar estes requisitos.

Esta metodologia tem o intuito de ser utilizada como um guia para hospitais e clínicas médicas que desejem adquirir sistemas de transcrição automática de laudos em Radiologia, ou visem ao próprio desenvolvimento deste sistema e suas customizações e, que estejam preocupados tanto com a usabilidade quanto com a aceitação destes sistemas por seu corpo de médicos radiologistas.

Este artigo está organizado como se segue. A seção 2 aborda os materiais e métodos utilizados na pesquisa; a seção 3 traz os resultados e discussões acerca da metodologia empregada; finalmente, a seção 4 traz as conclusões do trabalho.

MATERIAIS E MÉTODOS

A fim de tornar possível a criação de uma metodologia que possa avaliar, de maneira simples e com baixo custo, os STALs em Radiologia, foi necessário estabelecer os três passos descritos a seguir:

Identificação dos requisitos genéricos de VUI e requisitos específicos para STAL em Radiologia

Através da revisão bibliográfica sobre IHC, VUI, Sistemas de Informação Radiológica (RIS), foi possível listar os seguintes requisitos no uso de VUI em STAL:

Precisão (Acurácia): razão de acerto no reconhecimento das palavras ditadas pelo usuário e

* São sistemas em que as entradas do usuário são capturadas por um reconhecedor automático de voz. Assim, para atender as necessidades do usuário, a máquina necessita "entender" o que o usuário diz, desempenhar um processo de computação/transação, e responder ao usuário de tal forma que dê prosseguimento à conversação e cumprimento dos objetivos do usuário[9].

transcritas para o laudo, expressa como percentagem.

Tamanho do vocabulário: não pode ser muito pequeno de maneira que não cubra o domínio da aplicação, nem tão amplo que diminua a taxa de reconhecimento.

Dicionário específico para RIS: o vocabulário deve cobrir o domínio da aplicação, no caso, os termos utilizados na área de Radiologia.

Interferência sonora: num ambiente hospitalar, é necessário investigar qual a contribuição do ruído do ambiente tanto na redução da taxa de reconhecimento quanto no aumento da carga cognitiva do usuário final.

Naturalidade da fala do usuário: capacidade do sistema conseguir reconhecer o ditado feito pelo usuário final, referente a um laudo. Como o usuário está construindo as frases em tempo-real, pode haver vários problemas da fala natural, tais como concordância verbal e nominal, inserção de palavras irrelevantes ao texto que terão que ser tratadas pelo sistema.

Ajuda: está ligada à facilidade com que usuários, principalmente iniciantes, terão para aprenderem e acessarem o sistema de ajuda para conseguirem ditar um laudo de maneira eficiente.

Tratamento e recuperação de erros: está ligado, por exemplo, a como o sistema age quando não reconhece uma palavra que o usuário ditou.

Adequação do *feedback*: o sistema não deve fornecer *feedbacks* que atrapalhem a capacidade de raciocínio do usuário; deve estar presente, de maneira que o usuário saiba de erros que tenham ocorrido no sistema enquanto ele ditava o laudo, tal como uma palavra não reconhecida.

Tempo de resposta do *feedback*: as transcrições devem ocorrer em tempo-real, sem que o *delay* possa interferir na carga cognitiva do usuário final.

Visibilidade do sistema: este requisito está relacionado à maneira como o sistema terá para mostrar ao usuário seu *status* atual. Por exemplo, se o sistema encontrar algum erro de reconhecimento ou tiver algum problema no acesso ao dicionário, como será realizada a visibilidade do sistema.

Integração com os sistemas já existentes: importante requisito que evita retrabalho dentro das unidades que compõe o RIS e o PACS.

Qualidade das entradas de dados: este requisito está relacionado a bons dispositivos de captação de voz (microfones) que não causem problemas de interferência sonora ou elétrica que promovam distorções na fala.

Tempo para o laudo ficar pronto: o tempo para o laudo estar disponível para consulta através do uso de um STAL deve ser menor do que através de outros métodos: escrita à mão, digitação ou serviço de transcrição humana.

Satisfação do Cliente: este requisito está ligado ao prazer de se utilizar um STAL em Radiologia, medido através da aplicação de questionários, tais como o apresentado na seção de Resultados e Discussões.

Embora seja sabido que a maioria destes requisitos deva ser testada exaustivamente para se verificar seu valor real na qualidade do próprio sistema – realizado pelo fabricante - este trabalho está relacionado à usabilidade, podendo ser medida através de uma quantidade moderada de usuários e/ou especialistas, mesmo a fim de não elevar demasiadamente os custos e inviabilizar a avaliação.

Geração de uma metodologia de avaliação de interfaces do usuário baseadas em voz

Esta metodologia utiliza técnicas de inspeção e testes de usabilidade. A inspeção é utilizada a fim de baratear os custos com testes que envolvam o usuário final. Quando a inspeção for insuficiente ou não aplicável, então é lançado mão do teste de usabilidade.

Os requisitos de VUI e dos requisitos específicos para STAL em Radiologia foram agrupados em classes de

Tabela 1 - Agrupamento dos Requisitos em Classes.

Classe	Requisito
Classe 1: Desempenho Requisitos ligados ao correto funcionamento da aplicação, em termos de reconhecimento da voz e degradação do seu desempenho.	Precisão, Tamanho do vocabulário, Dicionário específico para RIS, Interferências sonoras, Naturalidade da fala do usuário, Resolução de ambiguidade para homônimos.
Classe 2: Facilidade de Uso Facilidades que o sistema pode fornecer ao usuário, a fim de que sua tarefa possa ser executada de maneira eficaz e eficiente, diminuindo a carga cognitiva do usuário.	Minimização da sobrecarga de memória, Modalidade apropriada, Tempo para o laudo ficar pronto.
Classe 3: Hardware e Integração Requisitos ligados ao desempenho físico, com dispositivos que permitam uma boa entrada de áudio e facilidade de integração com o PACS e HIS.	Integração com sistemas já existentes, Qualidade do sistema de áudio, Qualidade das entradas de dados.
Classe 4: Fatores Humanos Requisitos que estejam ligados ao prazer do usuário em usar o sistema e sua vontade de continuar a utilizar o sistema.	Diversidade e Percepção humana, Satisfação do usuário.
Classe 5: *Feedback* Requisitos ligados às mensagens apresentadas ao usuário sobre o *status* do sistema, assim como o tempo de resposta entre o ditado do usuário e sua transcrição na tela.	Tempo de *feedback* do sistema, Visibilidade do sistema, Adequação do *feedback*.
Classe 6: Tratamento de Erros e Sistema de Ajuda Requisitos relacionados à capacidade do sistema corrigir erros encontrados e do sistema corrigir um ditado, em tempo real ou tempo posterior.	Tratamento e Recuperação de Erros, Sistema de Ajuda.

Tabela 2 - Classificação da complexidade de se avaliar cada requisito.

Complexidade	Requisitos
Grau 1	Precisão; Tamanho do vocabulário; Interferências sonoras; Reconhecimento contínuo; Tempo para o laudo ficar pronto.
Grau 2	Adequação do Sistema de Ajuda, Tratamento e recuperação de erros; Qualidade das entradas de dados; Modalidade apropriada; Naturalidade da fala do usuário.
Grau 3	Minimização da sobrecarga de memória; Tempo de resposta do *feedback*; Adequação do *feedback*; Satisfação do cliente; Resolução de ambiguidade para homônimos.

Tabela 3 - Avaliação da Minimização da Sobrecarga de Memória[5].

Minimização da sobrecarga de memória	
Tipo de Avaliação	Subjetiva
Métodos de Avaliação	Observação
Importância	Alta
Dificuldade de Avaliação	1
Sintomas a procurar/ Métricas a utilizar	Verificar qual a diferença de desempenho entre o usuário ditar o laudo enquanto olha para a tela e quando não olha.

aderência (conforme Tabela 1) e também agrupados segundo seu grau de dificuldade de ser utilizado, sendo que: Grau 1 – Baixa Complexidade; Grau 2: Complexidade Média; e Grau 3: Alta Complexidade. (Tabela 2).

Método de Avaliação de Cada Requisito

Será utilizada a avaliação por inspeção de usabilidade para as medidas objetivas baseadas em heurísticas[7], no estudo de requisitos não funcionais para VUI e nas boas práticas de desenvolvimento de VUI[5,10-11]. Já para as medidas subjetivas, serão utilizados testes e questionários, estes foram elaborados seguindo a metodologia Questionnaire User Interface Satisfaction (QUIS)[12], mas adaptados para o uso em VUIs, de acordo com as peculiaridades dos STAL em Radiologia. Um *template* baseado em Dybkjaer e Bernsen[5] é proposto para melhor organizar as avaliações, tal como exemplificado na Tabela 3.

Aplicação da Metodologia

A aplicação da Metodologia de Avaliação de Usabilidade dos Sistemas de Transcrição Automática de Laudos em Radiologia ocorreu de duas maneiras:

- Analisando todos os requisitos possíveis com um sistema *stand-alone* e não está implantado num hospital. Estes requisitos foram analisados utilizando as técnicas de questionários de satisfação, observação e inspeção de usabilidade, a fim de facilitar os testes, economizar tempo e custos, além de incomodar, o mínimo possível, os médicos radiologistas.

- Analisando os demais requisitos que não poderiam ser analisados fora do ambiente de produção – ou seja, no hospital – com usuários reais – radiologistas – utilizando técnicas de observação e questionários de satisfação. Já para esta fase, foi selecionado um usuário que utiliza o sistema com mais frequência.

Planejamento da Inspeção de Usabilidade para o Sistema *Stand-alone*

A inspeção da usabilidade do STAL foi realizada de acordo com as seguintes etapas:

Criação das heurísticas: para que os especialistas verifiquem a conformidade do sistema com o que foi estabelecido. Para rodos os requisitos listados no item 2.1 deste artigo foram realizados estudos e especificações, conforme é visto nos cinco primeiros itens da Tabela 1.

Escolha dos especialistas: foram convidadas duas especialistas em usabilidade para esta inspeção que se encontram em fase de doutoramento na área de Interfaces Homem-Computador.

Preparação das inspeções: estas inspeções demoraram cerca de 20 horas, sendo uma das principais razões do não uso da segunda profissional em todas as inspeções e também da não utilização de mais de dois especialistas.

Os seguintes materiais foram necessários: *laptop* Philips com Processador Core 2 Duo, Memória RAM de 4GB, HD de 320 GB; SpeechMike Pro TM Philips para as entradas de áudio; *headset* Philips TM SHM 3300; decibelímetro; cronômetro; 10 Laudos de Radiologia do Tórax; softwares SpeechMagic™, Neoc Interactive 6.1 e SmInitialTraining da empresa Nuance[13].

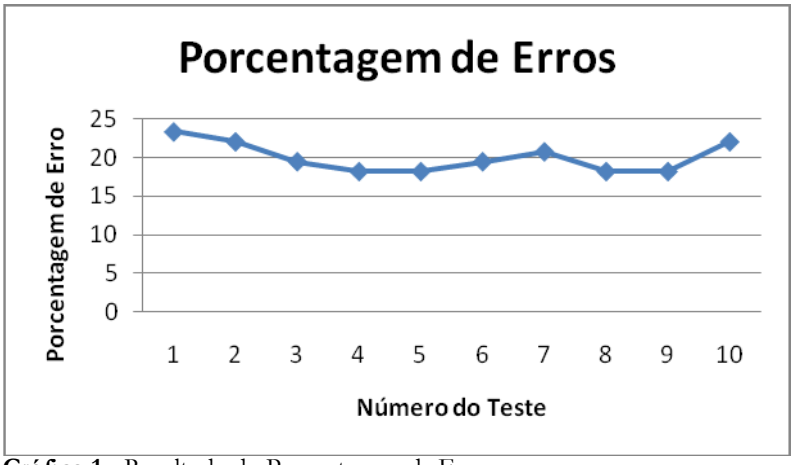
Geração das listas: objetiva gerar os resultados e análise de cada caso de teste determinado para inspeção, tal como mostrado na Quadro 1. Nesta fase, foram observadas as seguintes métricas: avaliação da precisão, recuperação de erros, interferência sonora, resolução de ambiguidade para homônimos, tempo de *feedback* do sistema, qualidade das entradas de áudio, visibilidade do sistema e adequação do *feedback*, adequação do sistema de ajuda, modalidade apropriada, naturalidade da fala do usuário e tamanho do vocabulário.

Planejamento dos Testes de Usabilidade para o Sistema *Stand-Alone*

Utilizando uma metodologia para a elaboração dos testes por observação adaptada de Diah et al[14], Nielsen[7] e de Mitchell[15], este planejamento é composto pelas seguintes atividades executadas de maneira consecutiva:

Planejar o teste de usabilidade: os testes foram realizados em um ambiente não hospitalar, através de participantes não pertencentes à área médica. Foram realizados dois testes-piloto a fim de verificar possíveis incongruências. Objetivou gerar os resultados e a análise

Quadro 1 - Avaliação da Reparação de Erros por Inspeção

Nome da Métrica	Recuperação de Erros
Método de Avaliação	Inspeção
Participantes	1 especialista
Material Necessário	1 texto com 77 palavras, com várias palavras fora do vocabulário específico da área de Radiologia, 1 equipamento SpeechMike ™
Roteiro	Verificar, através de inspeção, como o sistema age se o usuário usa palavras que não estão no dicionário da aplicação.
Resultados	O texto referente ao laudo utilizado neste teste tem 77 palavras. O teste foi repetido 10 vezes. Este teste demorou cerca de 17 minutos.  **Gráfico 1 -** Resultado da Porcentagem de Erro
Análise	No caso da inspeção utilizando palavras do vocabulário, a média de erros foi 4.54%, com desvio médio de 1.237%. Porém, neste teste foi de 20%, com desvio médio de 1.95%. Isso era esperado, visto que o vocabulário para o sistema é específico para a área de Radiologia.

de cada caso de teste determinado para os testes com usuários finais.

Preparar os materiais de teste: além dos materiais utilizados para inspeção, listados na Seção 3.4.1, também foram necessários: câmera fotográfica, questionários pré e pós-teste, carta de apresentação e formulários de observação do usuário.

Criar as tarefas: as tarefas objetivaram validar as seguintes métricas: recuperação de erros, naturalidade da fala do usuário, carga cognitiva, visibilidade do sistema e adequação do *feedback*, tempo de *feedback* do sistema e qualidade do sistema de áudio

Selecionar os participantes: foram selecionados seis voluntários (três homens e três mulheres) e dois especialistas em inspeção. Dois critérios foram determinados: diferentes timbres de voz, tanto masculino quanto feminino e pessoas com sotaque diferente das pessoas da cidade de São Paulo.

Conduzir os testes de usabilidade: as sessões foram compostas por quatro partes, com duração média de 45 minutos: Introdução: explica os motivos para a realização dos testes e como o usuário deve agir. Deve durar cerca de 5 minutos; Questionário de pré-teste: coleta informações relevantes sobre o perfil do usuário, com duração média de 2 minutos; Tarefas: quais as tarefas que o usuário deve executar, com duração média de 30 minutos. Os voluntários fizeram um treinamento mínimo prévio de sua voz, de cerca de 2 minutos, em média; Questionário de pós-teste: coleta as informações referentes à impressão do usuário sobre as tarefas que ele executou, com duração média de 3 minutos.

Analisar os dados do teste de usabilidade: corresponde à análise dos problemas encontrados, tal como mostrado na Quadro 2.

RESULTADOS E DISCUSSÕES

As discussões sobre os resultados dos testes de usabilidade aplicados neste trabalho foram divididas em duas partes, uma referente ao sistema *stand-alone* e outra ao sistema implantado num hospital de São Paulo.

Sistema *Stand-Alone*

Através de todos os testes realizados com este sistema, foi possível concluir várias questões referentes à inspeção e também aos testes de usabilidade fora do ambiente hospitalar, que estão descritas a seguir.

- O STAL em Radiologia se mostrou bastante eficiente em relação à precisão do reconhecimento de voz, ficando acima de 93%, em média, mesmo incluindo, no grupo, pessoas com sotaque pronunciado.

- A precisão da voz chegou a 95% de acerto, mesmo com um treinamento mínimo da voz. Se forem incluídas palavras novas ao vocabulário inicial do sistema e também um treinamento mais aprofundado para os usuários esta precisão pode melhorar. Para exemplificar o ganho que se teria, se a palavra "quilovoltagem" fosse incluída, haveria uma diminuição do erro, em média, de 25.65%.

- O sistema é sensível à mudança da velocidade da fala.

- Não houve diferença significativa na taxa de reconhecimento sem o treinamento e com o treinamento,

Quadro 2 - Caso de Teste para Avaliação da Recuperação de Erros do Sistema por Observação.

Nome da Métrica	Recuperação de Erros
Método de Avaliação	Observação e Questionário
Participantes	6 participantes
Material Necessário	1 laudo de radiologia, 1 equipamento SpeechMike ™
Roteiro	Verificar, através de observação, como o sistema age perante os seguintes erros: a) O usuário usa palavras que não estão no dicionário da aplicação – um laudo será modificado para atingir este objetivo. b) O sistema não consegue reconhecer o que o usuário dita. c) O usuário erra o laudo e deseja refazê-lo.
Resultados	O texto continha 77 palavras, com várias palavras fora do vocabulário específico da área de Radiologia. Os usuários 1, 3, 5 são mulheres. Os resultados podem ser visualizados no Gráfico 2.
Análise	a) Quando o usuário errou o laudo, ele se sentiu bastante desconfortável em relação a como promover a correção. Interjeições foram freqüentes quando o usuário se conscientizou do erro. Também perguntas ao observador de como promover a correção foram comuns. b) Quando o usuário olhava a tela enquanto ditava e o sistema demorava em apresentar o texto – que acontecia com atraso de até 5 segundos – o usuário gaguejava ou tentava recomeçar a fala. c) É possível verificar, através deste gráfico, que houve uma considerável diminuição de precisão com o uso de um texto com palavras fora do vocabulário específico de radiologia. O que era esperado.

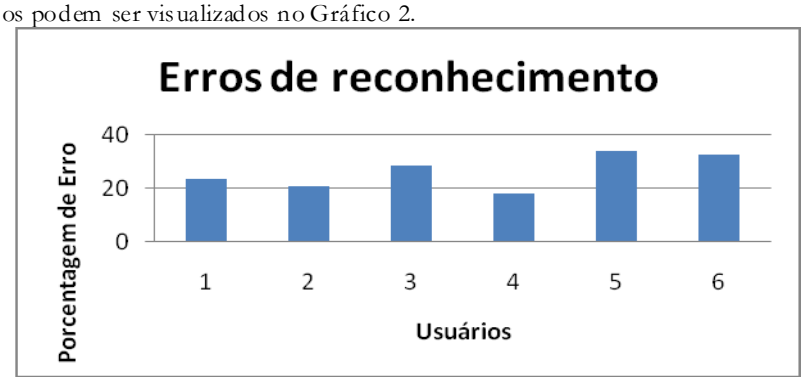


Gráfico 2 - Porcentagem de Erro de Teste com Seis Usuários

Os usuários também responderam a perguntas do questionário a respeito da recuperação de erros e esta questão apareceu até nas questões abertas como críticas ou melhorias.

Tabela 1 - Parte do Questionário sobre Experiência de Uso do Sistema

Experiência de uso do sistema					
	Concordo Fortemente	Concordo	Indeciso	Discordo	Discordo Fortemente
Quando erro no ditado, é fácil fazer o sistema reconhecer o que falo e se adaptar.		2		4	4

Pontos Negativos da Aplicação: O sistema não possui mecanismos de reparo de erros durante o ditado.
Sugestões ou reclamações sobre a aplicação: Tratamento de erro – algum mecanismo para reparar algum erro, solicitado pelo usuário.

como era pensado, ficando a taxa de reconhecimento, sem treino, em quase 93%, em média. O sistema poderia, então, ser utilizado mesmo sem ser realizado um treinamento mínimo da voz, com taxas aceitáveis de reconhecimento da voz.

- O *delay* do sistema em apresentar o texto referente ao que estava sendo ditado foi bastante alto, chegando a quase 5 segundos, causando um sentimento de que o sistema não estava funcionando.

- A interferência sonora influencia fundamentalmente a precisão do reconhecimento de voz, devendo, o ambiente, se necessário, ser modificado para estar de acordo com as normas de ruído estabelecidas pela ABN[T16].

- A massa de dados para testar o requisito "Dicionário específico para RIS" é insuficiente para se afirmar que o dicionário inicialmente inserido no sistema de reconhecimento seja considerado aprovado ou reprovado.

- O requisito "Resolução de Homônimos" foi testado de maneira superficial, pois não foram levadas em consideração todas as palavras homônimas que poderiam aparecer no dicionário do sistema.

- Os dois equipamentos utilizados para a entrada dos laudos – HeadSet Philips e SpeechMike Philips – se mostraram eficientes. Era esperada, segundo informação

do fornecedor, uma grande diferença de qualidade sonora, porém, a precisão no reconhecimento se mostrou melhor com o *headset* em condições de pouca interferência sonora. Já com alta interferência sonora, o equipamento SpeechMike foi melhor. A diferença expressiva de custo e a ergonomia entre os dois equipamentos devem ser levadas em consideração no processo de aquisição.

- O sistema de ajuda foi considerado bastante superficial: o *help* online não funcionou e não havia um sistema de buscas de palavras nem de contexto.

- Em relação à falta de visibilidade e adequação do *feedback*, a criticidade apontada pelos usuários se deu pelo sistema nunca informar ao usuário palavras que, possivelmente, tenham sido reconhecidas com baixa taxa. Uma sugestão proposta foi que estas palavras fossem marcadas – tal como o uso de negrito ou sublinhado – para tornar a visualização mais fácil.

- Em relação à recuperação de erros, os usuários se sentiram perdidos porque não conseguiram consertar erros de ditado utilizando comandos de voz, ou seja, os usuários não conseguiam retroagir no ditado.

Informações do hospital pesquisado

Uma das razões apresentadas por um dos médicos da equipe de Radiologistas do hospital para o não uso da ferramenta refere-se ao fato dele digitar mais rapidamente do que o sistema consegue transcrever. Este médico também fala muito rapidamente, o que ocasionaria uma queda de *performance* no sistema.

Atualmente este hospital gera, no seu departamento de Ressonância Magnética, uma média de 120 laudos por dia, através de 8 médicos radiologistas. Existem 2 digitadoras no departamento, que, segundo o médico entrevistado, possuem gargalo de atendimento, devido à escassez de pessoas.

Quando o laudo necessita de urgência, é possível fazê-lo em 20 minutos utilizando o serviço de digitação, contra 5 minutos utilizando o STAL em Radiologia. O tempo médio para que um laudo seja liberado no hospital é de 3 dias, sendo comum que esta tarefa seja executada em até 5 dias.

Outro ponto importante é que, quando o serviço de digitação é utilizado, o médico precisa rever as imagens médicas para confirmar o laudo; já no uso do STAL, isso se torna desnecessário devido ao imediatismo com que o laudo fica pronto para uso.

Entre os pontos fracos na observação do uso deste sistema, percebeu-se que a taxa de reconhecimento de voz foi bastante inferior ao apresentado pelo sistema *stand-alone*.

Também foi percebido e confirmado pelo médico que, nesta versão do sistema há grande dificuldade em reconhecer epônimos, tal como "T2" que é reconhecido como "tendões".

Segundo o médico entrevistado e observado, que trabalha em outro hospital com o mesmo sistema, em versão diferente, as principais melhorias no uso do sistema, em versão superior são: maior taxa de reconhecimento da voz, menor interferência sonora do ambiente, eliminação do problema da versão 2.1 de espaçamento entre palavras, ser integrado ao sistema PACS e RIS e o sistema aprende mais rápido os epônimos.

CONCLUSÃO

O trabalho visou a redução de custos com testes de usabilidade, por ser conhecido na literatura de IHC, que este é um custo que pode ser impeditivo da avaliação de muitos sistemas.

Visto que os STAL em Radiologia têm sido cogitados como solução para diminuir o tempo para o laudo ficar pronto e também como redutor dos custos globais do departamento de Radiologia, uma metodologia de avaliação destes produtos é essencial. A metodologia proposta contribui para a escolha de um sistema que atenda às necessidades de mercado em relação ao seu usuário final. Para tanto, esta metodologia leva em consideração muitos aspectos que vão além da taxa de reconhecimento destes sistemas, tratando questões como tamanho do vocabulário utilizado, ambiente utilizado e satisfação do usuário.

Assim, procurou-se utilizar, ao máximo possível, as técnicas de inspeção que não utilizam usuários finais e que diminuem significativamente o tempo e os custos das avaliações.

Também, para que os hospitais não necessitassem disponibilizar tempo e incômodos de seus radiologistas, optou-se por utilizar a inspeção realizada por especialistas em usabilidade e, por outro lado, usuários voluntários foram selecionados para os testes de usabilidade, quando a inspeção não fosse o método mais apropriado. Somente a observação dos usuários finais – radiologistas – e o preenchimento, por eles, de um questionário de satisfação foram as técnicas utilizadas para se analisar a usabilidade, quando da necessidade de estar "em campo". Este questionário não tomou mais que 3 minutos do tempo do radiologista.

Sendo assim, este trabalho serve de guia para a área de TI de hospitais e clínicas radiológicas quando se deseja avaliar a compra de sistemas de transcrição automática de laudos, cada vez mais comuns no mercado interno. Também pode ser utilizado para se verificar, quando se trabalhe com customizações de tais sistemas, a usabilidade que se deseja atingir e a que está atualmente vigente.

É desejável que a avaliação de usabilidade proposta através desta metodologia seja realizada, ou comandada, por especialistas em usabilidade, pois é necessária, mesmo para um especialista, uma quantidade bastante grande de horas para, principalmente, a avaliação de inspeção.

REFERÊNCIAS

1. Allen J, Perrault C. Analysing intention in utterances. Artificial Intelligence. 1980; 15(3):143-78.
2. Kamm C, Walker M, Rabiner L. The role of speech processing in human computer intelligent communication. Speech Communication. 1987;(23): 263-78.
3. 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; 2001 jul. 9-11; Toulouse, France; 2001.
4. Gibbon D, Moore R, Winski R (eds). Handbook of standards and resources for spoken language systems. Berlin, New York: Mouton de Gruyter; 1997.
5. Dybkjaer L, Bernsen NO. Usability evaluation in spoken language dialogue systems. In: Proceedings of the ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems; 2001 jul. 6-7;Toulouse, France; 2001.
6. SNOMED. [Citado 2009 jul 03]. Disponível em: <http://www.ihtsdo.org/snomed-ct/>.
7. Nielsen J. Usability engineering. Cambridge: Academic Press; 1993.
8. Sommerville I. Software engineering. 6th. ed. New York: Addison Wesley; 2001.
9. Zukerman I, Litman D. Natural language processing and user modeling: synergies and limitations. User Modeling and User-Adapted Interaction. 2001;11(1-2):129-58.
10. Salvador VFM, Oliveira Neto JS, Kawamoto AL. Requirement engineering contributions to voice user interface. In: Proceeding of the First International Conference on Advances in Computer-Human Interaction; 2008 feb. 10-15; Sainte Luce, Martinique; 2008.
11. Komatani K, Ueno S, Kawahara T, Okuno HG. Flexible guidance generation using user model in spoken dialogue systems. In: Proceedings of the 4lst Annual Meeting of the Association for Computational Linguistics; 2003 jul. 07-13; Sapporo. Japan; 2003. p.256-63.
12. Chin JP, Diehl VA, Norman KL. Development of an instrument measuring user satisfaction of the human-computer interface. In: Proceedings of the ACM CHI'88; 1988 jun.15-19; Washington; 1988. p. 213-8.
13. NUANCE. [Citado 2010 dez 20]. Disponível em: <http://nuance.com/naturallyspeaking/>.
14. Diah NM, Ismaili M, Ahmad S, Dahari MKM. Usability testing for educational computer game using observation method. In: CAMP'10 is the first international conference on Information Retrieval and Knowledge Management. Shah Alam, Malaysia, 2010.
15. Mitchell PP. A step-by-step guide to usability testing. Lincoln: NE: iUniverse; 2007.
16. ABNT – NBR 10152/1987. [Citado 2010 dez 26]. Disponível em: <http://www.filecrop.com/NBR-10152.html>.