

# Métricas Subjetivas para Avaliação da Usabilidade de Sistemas de Transcrição Automática de Laudos em Radiologia

Valéria Farinazzo Martins<sup>1</sup>, Lincoln de Assis Moura Jr.<sup>2</sup>

<sup>1</sup>Grupo de Processamento Gráfico – Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie – São Paulo, Brasil

<sup>2</sup> Assis Moura eHealth -São Paulo – SP, Brasil  
valeria.farinazzo@mackenzie.br, lincoln.moura@zilics.com.br

***Abstract.** This article describes an usability evaluation for Automatic Transcription System of Radiology Reports through subjective metrics proposed and applied to a case study, using a plan of usability tests in order to validate the process. This review aimed to be considered secondary low cost because they were used, first, volunteers and users only after the metrics were applied to the system worked in the hospital, with real users. This is important because it reduces the time that radiologists have a limitation in their activities.*

***Resumo.** Este artigo descreve uma avaliação de usabilidade de Sistemas de Transcrição Automática de Laudos em Radiologia, através de métricas subjetivas propostas e aplicadas a um estudo de caso, utilizando um plano de testes de usabilidade, a fim de validar o processo. Esta avaliação teve como objetivo secundário ser considerada de baixo custo, pois foram utilizados, primeiramente, usuários voluntários e somente depois as métricas foram aplicadas ao sistema instalado no ambiente hospitalar, com usuários reais. Isso é importante porque diminui o tempo em que os radiologistas têm uma interferência em suas atividades.*

## 1. Introdução

Nos últimos anos, a área de Interface Homem-Computador (IHC) tem se desenvolvido e um grande esforço tem sido empregado em relação a questões de projeto e avaliação de interfaces não convencionais, tais como em Realidade Virtual, Realidade Aumentada, Interfaces por Gestos e Interface do Usuário Baseada em Voz - do inglês Voice User Interface (VUI). Aplicações comerciais de VUI tais como atendimento eletrônico de chamadas telefônicas, agendamento e consulta de horário de vôos e agendamento de hotéis, começam a aparecer no Brasil, mas ainda enfrentam certa barreira do usuário quando ele necessita “conversar” com uma máquina.

Analizando, especialmente, a área da Saúde, a utilização de reconhecimento de voz em sistemas da Saúde de propósito geral, tais como em emergência, não é eficiente devido ao grande vocabulário do domínio - sabe-se que um profissional de saúde utiliza, no seu cotidiano, mais de 100 mil itens no seu vocabulário. Além disto, as possibilidades diagnósticas presentes no SNOMED (Systematized Nomenclature of MEDicine) (2012),

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, tais como em sistemas de ditado como em sistemas de transcrição automática de laudos (STAL); sistemas com unidades de resposta audível baseados em voz (IVR) e sistemas para controlar equipamentos médicos [Durling and Lumsden 2008]. Isto significa que o vocabulário é consideravelmente menor, proporcionando uma precisão mais alta no reconhecimento dos termos específicos.

Porém, muito pouco tem sido explorado em questão da avaliação destes sistemas de voz para este uso na área da Saúde. 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 mais efetivo e eficaz. Tampouco se têm estudado questões como a satisfação do usuário final. Assim, este trabalho apresenta uma avaliação de usabilidade realizada através de entrevista, observação e questionário focada nos usuários do sistema, ou, em alguns casos, a um conjunto de potenciais usuários.

O trabalho está organizado da seguinte maneira. A seção 2 apresenta uma descrição dos Sistemas de Transcrição Automática de Laudos. Na seção 3 são apresentadas as métricas subjetivas propostas para a avaliação destes sistemas. Já a seção 4 apresenta o plano de testes de usabilidade utilizado. A seção 5 traz os resultados dos testes. Finalmente, na seção 6 são descritas as conclusões do trabalho.

## **2. Sistemas de Transcrição Automática de Laudos em Radiologia**

A fim de facilitar o trabalho de digitação dos radiologistas, hospitais e centros médicos tem recorrido a um serviço de transcrição humana destes laudos. O processo consiste em o radiologista interpretar imagens e gravar um laudo em áudio. Este áudio é digitado por uma pessoa, resultando num laudo escrito. Este laudo é considerado preliminar, mas, em algumas práticas, ele já pode ser compartilhado por médicos nesta fase. O radiologista então finaliza o laudo transcrito depois de tê-lo revisado (às vezes sem revisar as imagens) e supõe-se que o texto esteja correto. O laudo final e as imagens são então disponibilizados para os médicos. Este é um processo que pode levar várias horas [Vorbeck et al 2000], [Bhan et al 2008].

Entre as principais desvantagens em relação ao uso de sistemas de transcrição humana, é possível destacar [Hundt et al 1999], [Durling and Lumsden 2008], [White 2005]:

- Demora na devolução dos laudos digitalizados. Este tempo, chamado de *turnaround time* (TAT), é de cerca de 70 horas. Os serviços tradicionais de transcrição são conhecidos por, às vezes, produzirem acúmulos de fitas de ditado esperando tanto para serem transcritas quanto validadas.
- Alguns pacientes necessitam de uma investigação radiológica imediata para fornecer intervenção apropriada. Entretanto, como os laudos transcritos podem consumir bastante tempo, os médicos necessitam esperar um tempo bastante grande antes de aplicar a terapia correta ao paciente.
- As questões ligadas à privacidade das informações dos pacientes podem ser questionadas, pois pessoas não vinculadas à saúde estão tendo acesso a estas informações.

- Adicionalmente, uma consequência não esperada é o fato de que secretárias necessitam assumir mais uma obrigação: de compartilhar o tempo com suas obrigações com o tempo nas transcrições.

A fim de encurtar o tempo necessário para escrever laudos, a tecnologia de reconhecimento de voz aparece como uma alternativa para o serviço de transcrição.

Os STAL são sistemas que visam facilitar a transcrição de linguagem falada em texto para sua inclusão em documentos médicos, utilizando, geralmente, um vocabulário extenso. Um estudo realizado por Rodger e Pendharkar (2007) mostrou que a precisão do sistema utilizado variou de 85% a 99%, dependendo de fatores como idade do usuário, sexo, familiaridade com sistemas de VUI. Já Mohr et al (2003) estudaram a comparação de qualidade e precisão de documentos médicos gerados por transcrição humana e por sistemas baseados em reconhecimento de voz. Mohr et al (2003) concluíram que os níveis de precisão da tecnologia de reconhecimento de voz, naquele momento, forçavam a edição em grande proporção dos documentos transcritos e, portanto, esta tecnologia não deveria ser proferida como redução assegurada nos custos da entrada de dados.

A possibilidade de usar transcrição de interpretações radiológicas controladas por voz foi descrita pela primeira vez em 1981 (Robbins et al 1987), mas a tecnologia disponível naquela época era insuficiente e economicamente inviável para se trabalhar. Naquele momento, era necessário entrar com cada palavra pausadamente. Atualmente, os sistemas de reconhecimento de voz têm trabalhado de forma contínua, permitindo que seus usuários falem de maneira mais natural [Durling and Lumsden 2008].

Alapetite, Boje e Morten (2009) relataram suas expectativas em relação ao aumento da qualidade dos registros do paciente, com o uso de sistemas de reconhecimento de voz, desde que os médicos agora chequem e validem os registros escritos imediatamente após suas conclusões, que ainda estão recentes na memória.

Outro fator importante para o uso de sistemas de reconhecimento de voz é a possibilidade de integrá-los aos sistemas já existentes, incluindo Picture Archiving and Communication System (PACS), Sistemas de Informação Radiológica (RIS) e Sistemas de Informação Hospitalares (HIS).

Portanto, os principais propósitos de se utilizar reconhecimento de voz nos departamentos de radiologia são o de diminuir o TAT e, conseqüentemente, os custos operacionais; conseguir uma qualidade mais alta dos prontuários do paciente e; integrar o sistema de reconhecimento de voz com os sistemas já existentes (PACS, RIS e HIS) [White 2005], [Alapetite, Boje and Morten 2009].

As principais etapas, que compõe um STAL atualmente comercializado, são mostradas na Figura 1 e explanadas a seguir.



**Figura 1: Etapas de um STAL. (Fonte: autor)**

- Entrada de Voz: utilização de um equipamento apropriado para a entrada de voz do radiologista.
- Captura do Laudo: um servidor capaz de armazenar os arquivos de áudio gerados durante o ditado do radiologista.
- Reconhecimento de Voz: um módulo de reconhecimento de voz com o vocabulário e gramática próprias da área de radiologia.
- Correção e Edição de Laudo: verificação de possíveis erros de reconhecimento de voz que possam ter acontecido na fase anterior. Estes erros podem ser simplesmente a falta de pontuação ou erros que comprometam o entendimento do laudo.
- Laudo Disponível: já corrigido, o laudo é disponibilizado para aqueles que possam ter acesso a ele.

### 3. Métricas Subjetivas

Uma avaliação mais completa dos STALs deve levar em consideração não somente a taxa de reconhecimento de voz, mas questões mais amplas que verifiquem, por exemplo, a satisfação do usuário ou a interferência sonora vinda do ambiente.

As métricas subjetivas são aquelas que podem medir a opinião dos usuários a respeito do sistema. Alguns trabalhos sobre avaliação subjetiva de VUI são propostos por Walker, Passonneau e Boland (2001), Möller (2005), Larsen (2003), Dybkjaer e Bernsen (2001), Hartikainen, Salonen e Turunen (2004). Para STALs, foram consideradas as seguintes medidas subjetivas: satisfação do cliente, naturalidade da fala do usuário, diversidade e percepção humana, minimização da sobrecarga de memória, adequação de *feedback*, adequação do sistema de ajuda e visibilidade do sistema. Estas métricas estão ligadas à facilidade do usuário em interagir com o sistema e, portanto, à sua capacidade de ser “amigável”.

Todas as métricas subjetivas tratadas neste trabalho podem ser visualizadas a seguir (Quadro 1 a 4). Um *template* foi criado, pelos autores, a fim de organizar os dados.

Nome da Métrica	SATISFAÇÃO DO CLIENTE	NATURALIDADE DA FALA DO USUÁRIO
Métodos de Avaliação	Questionário	Observação e questionário
Importância	Alta	Alta
Dificuldade de Avaliação	Complexidade Média	Complexidade Média
Sintomas a procurar/ Métricas a utilizar	Facilidade de uso, valor agregado, sucesso da tarefa.	Se o usuário falar com naturalidade, o sistema é capaz de entendê-lo?

**Quadro 1: Avaliação da Satisfação do Cliente e Modalidade Apropriada.**

Nome da Métrica	DIVERSIDADE E PERCEPÇÃO HUMANA	MINIMIZAÇÃO DA SOBRECARGA DE MEMÓRIA
Métodos de Avaliação	Observação e questionário	Observação
Importância	Alta	Alta
Dificuldade de Avaliação	Complexidade Média	1
Sintomas a procurar/ Métricas a utilizar	Verificar se usuários com graus de experiência diferentes conseguem realizar suas tarefas de maneira apropriada. Para isto o sistema deve fornecer formas diferentes de acesso às funcionalidades.	Verificar qual a diferença de desempenho entre o usuário ditar o laudo enquanto olha para a tela e quando não olha.

**Quadro 2: Avaliação da Diversidade e Percepção Humana e Minimização da Sobrecarga de Memória.**

Nome da Métrica	ADEQUAÇÃO DO SISTEMA DE AJUDA	VISIBILIDADE DO SISTEMA
Métodos de Avaliação	Observação e Questionário	Observação e Questionário
Importância	Média	Alta
Dificuldade de Avaliação	Baixa Complexidade	Complexidade Média
Sintomas a procurar/ Métricas a utilizar	Verificar se o sistema de ajuda satisfaz as necessidades do usuário, caso esteja com dúvidas.	Verificar se o sistema é capaz de informar ao usuário o que está acontecendo em determinado momento.

**Quadro 3: Avaliação da Eficiência do Sistema de Ajuda e da Visibilidade do Sistema.**

Nome da Métrica	ADEQUAÇÃO DO FEEDBACK
Tipo de Avaliação	Subjetiva
Métodos de Avaliação	Inspeção e Questionário
Importância	Alta
Dificuldade de Avaliação	Alta Complexidade
Sintomas a procurar/ Métricas a utilizar	a) o sistema não deve fornecer <i>feedbacks</i> que atrapalhem a capacidade de raciocínio do usuário; b) deve estar presente, de maneira que o usuário, em tempo oportuno, saiba de erros que tenham ocorrido no sistema enquanto ele ditava o laudo (tal como uma falha no reconhecimento de alguma palavra)

**Quadro 4: Avaliação do *Feedback* do Sistema.**

#### 4. Plano de Testes

O objetivo desta avaliação foi verificar quão bem funciona o STAL, desejando-se mensurar o quanto o sistema está adaptado às métricas estabelecidas na seção 3 deste trabalho. Para tanto, serão utilizados dois sistemas STAL, que serão descritos a seguir. A fim de facilitar os testes, economizar tempo e custos, além de incomodar, o mínimo possível, os médicos radiologistas, optou-se por se realizar a maioria dos testes com usuários voluntários interessados na área tecnológica. Assim, os testes foram realizados em dois momentos: em um ambiente não hospitalar, com participantes não pertencentes à área médica. Isso é aceitável, visto que vários requisitos aqui abordados se referem a questões de usabilidade de um sistema de ditado, não especificamente de laudos. E, posteriormente, os testes são aplicados ao sistema implantado no hospital, foco do estudo.

O primeiro sistema, o Speech Magic<sup>TM</sup> versão 6.1 (Nuance 2010), foi instalado em uma máquina fora do ambiente hospitalar para fazer os testes com usuários não radiologistas. Embora haja alguns inconvenientes no uso desta abordagem, tais como o sistema não estar instalado como cliente-servidor, não permitir um aprendizado mais eficaz da máquina de reconhecimento de voz e, ainda, não permitir a entrada de novas palavras -, para vários tipos de testes, esta abordagem já é suficiente.

O segundo sistema, o Philips Speech Magic versão 2.1 (Nuance 2010), é o que está instalado no hospital foco do estudo. A Figura 2 mostra um médico radiologista no seu ambiente de trabalho, utilizando o sistema.

Depois de preparado o material, definidas as tarefas, selecionados os participantes, entre as outras tarefas que compõe este tópico, foram realizados dois testes-piloto a fim de verificar possíveis incongruências no teste - um realizado pela autora e outro por um participante voluntário.



**Figura 2: Radiologista transcrevendo o laudo. (Fonte: autor)**

#### 4.1. Preparação dos materiais de testes

Esta subseção detalha a configuração dos equipamentos utilizados durante os testes com o software SpeechMagic (Nuance 2010), assim como os materiais necessários: Laptop Philips: Processador Core 2 Duo, Memória RAM de 4GB, HD de 320 GB; SpeechMike Pro™ Philips; Headset Philips TM SHM 3300; Câmera Fotográfica; Cronômetro; 10 Laudos de Radiologia do Tórax; Questionário de pré-teste; Carta de Apresentação dos Testes de Usabilidade; Questionários de pós-teste, baseado na metodologia Questionnaire User Interface Satisfaction (QUIS) proposta em [Chin, Diehl and Norman 1988]; e Formulários de Observação do Usuário.

#### 4.2. Criação das tarefas

As tarefas preparadas para os participantes executarem fazem parte do contexto de atividades que o radiologista teria que executar no seu dia-a-dia. Sendo assim, foram selecionados laudos escritos, sem identificação do paciente, para os participantes ditarem.

A seguir serão destacados os casos de teste para testar a Naturalidade da Fala do Usuário, Minimização da Sobrecarga de Memória e Adequação do *Feedback* do Sistema, organizados através de um *template* concebido pelos autores.

Nome da Métrica	Naturalidade da Fala do Usuário
Participantes	6 usuários: 3 do sexo masculino e 3 do sexo feminino para o teste "a" e 2 usuários, um do sexo feminino e um do sexo masculino, com sotaque diferente da cidade de São Paulo para o teste "b"
Material Necessário	1 laudo de radiologia, 1 equipamento SpeechMike TM
Roteiro	<p>a) Através da leitura de laudos:</p> <ul style="list-style-type: none"> <li>✓ Verificar a diferença de precisão entre o usuário falar mais pausadamente ou mais rapidamente;</li> <li>✓ Verificar a diferença de precisão entre o usuário falar com tom baixo ou alto;</li> <li>✓ Usuário lê um laudo e usuário fala um laudo sem ler;</li> </ul> <p>b) Através da leitura de laudos:</p> <ul style="list-style-type: none"> <li>✓ Verificar a diferença de precisão entre o usuário falar mais pausadamente ou mais rapidamente;</li> <li>✓ Verificar a diferença de precisão entre o usuário falar com tom baixo ou alto;</li> </ul>

**Quadro 5: Caso de Teste para Avaliar a Naturalidade da Fala do Usuário.**

Nome da Métrica	Minimização da Sobrecarga de Memória
Participantes	6 participantes
Material Necessário	1 laudo de radiologia, 1 equipamento SpeechMike TM

Orientações Participantes	aos	Se você não fez o treinamento inicial da fala, faça-o agora; <input checked="" type="checkbox"/> Leia um laudo sem a pontuação. <input checked="" type="checkbox"/> Leia um laudo e faça a pontuação correta.
Roteiro		Verificar, através de observação, como o usuário age: Há aumento de carga cognitiva significativa?

**Quadro 6: Caso de Teste para Avaliar a Minimização da Sobrecarga de Memória.**

Nome da Métrica	Adequação do Feedback
Participantes	6 participantes
Material Necessário	1 laudo de radiologia, 1 equipamento SpeechMike TM
Roteiro	a) o sistema não deve fornecer feedbacks que atrapalhem a capacidade de raciocínio do usuário: <input checked="" type="checkbox"/> O sistema interrompe o ditado; <input checked="" type="checkbox"/> O sistema inicia processos de alertas; b) deve estar presente de maneira que o usuário, em tempo oportuno, saiba de erros que tenham ocorrido no sistema enquanto ele ditava o laudo: <input checked="" type="checkbox"/> O sistema permite verificar, de alguma forma, que ocorreu uma falha no reconhecimento de alguma palavra;

**Quadro 7: Caso de Teste para Avaliar a Adequação do Feedback do Sistema.**

### 4.3. Seleção dos participantes

Para os testes fora do ambiente hospitalar 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. Já para o ambiente hospitalar, foram observados e entrevistados dois médicos radiologistas.

### 4.4. Condução dos testes

Nos testes de usabilidade, ocorridos fora do ambiente hospitalar, 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. Duração de 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.

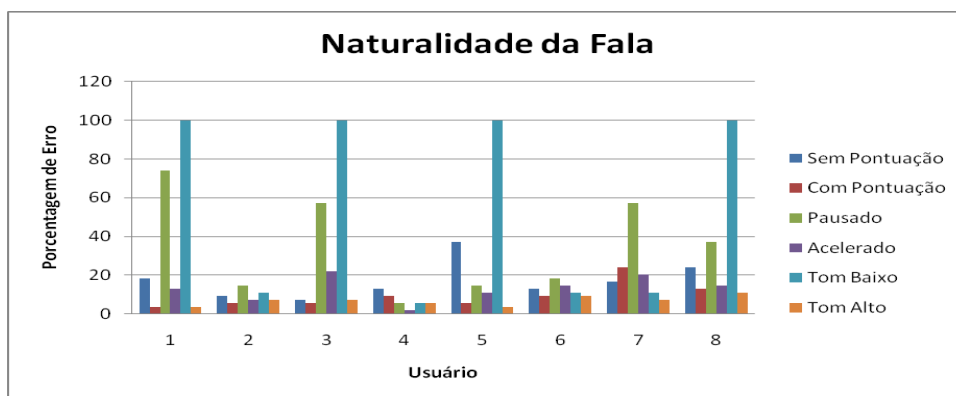
## 5. Resultados

Os dados provenientes de observação foram formatados em gráficos, conforme é exemplificado no Gráfico 1 da seção 5.1. Uma pequena introdução foi apontada para a correta interpretação do gráfico. Posteriormente foram realizadas as análises sobre o gráfico, como mostrado no texto. Já os dados provenientes de questionário, podem ser visualizados no Quadro 8.

### 5.1. Naturalidade da Fala (dados de Observação)

Para estes testes foi utilizado um laudo com 54 palavras. Os resultados podem ser visualizados no Gráfico 1. Os usuários 1, 3, 5 e 8 são mulheres. Os usuários 7 e 8 são, respectivamente, um homem com sotaque e uma mulher com sotaque.

**Gráfico 1: Naturalidade da Fala de Oito Voluntários**



Através deste gráfico é possível concluir que:

- Embora esteja na especificação do sistema, dada pelo fornecedor, que o sistema faz a inserção de pontuação, o sistema não foi capaz de fazê-lo, colocando conjunções tais como “com” e “e” que, em muitas vezes, afetavam o sentido das frases.
- A precisão de reconhecimento da voz, pela leitura do texto com pontuação explícita foi maior entre as mulheres; entre todos os participantes, exceto os que tinham um sotaque bastante pronunciado (usuário 7 e 8), foi superior a 90%.
- Ficou claro que o sotaque influencia enormemente a taxa de reconhecimento, necessitando que estes usuários façam um treinamento mais intensivo e que o sistema aprenda à medida que o usuário utilize o sistema de forma contínua.
- O uso do sistema para uma fala natural do usuário sempre se mostra melhor do que a inserção dos dados através de um ditado em velocidade diferente da habitual.
- O uso do sistema para um tom de voz mais baixo ou mais alto também afeta o reconhecimento, devendo o usuário sempre utilizar seu tom normal de voz. Entre as mulheres, o uso de um tom mais baixo de voz impediu a entrada do laudo, visto que o sistema foi incapaz de captar a entrada de áudio.

Através de questionário (Quadro 8), foi possível investigar a sensação do usuário quanto à naturalidade de sua fala.

Experiência de uso do sistema					
	Concordo Fortemente	Concordo	Indeciso	Discordo	Discordo Fortemente
As formatações, tais como “ponto final”, “vírgula”, “negrito”, são de fácil assimilação, não me fazendo pensar muito para utilizá-las.	3	1			
Quando faço hesitações, tais como "mmh", "Aaah", o sistema ignora, captando somente palavras pertencentes a meu vocabulário para o laudo.	1	2			2
É fácil fazer a formatação do texto enquanto dito um laudo, tal como mudar de linha e pontuar.	3	1			

**Quadro 8: Parte do Questionário da Experiência de Uso do Sistema.**



É possível perceber, através dos dados, que os usuários não se importam em ditar também as formatações do texto. Porém, quando há hesitações na fala, eles percebem que, em muitos casos, o sistema interpreta, de maneira errônea, a fala do usuário.

### 5.3. Considerações sobre as métricas utilizadas

A seguir é possível verificar algumas considerações sobre a análise de cada métrica utilizada para os testes de usabilidade:

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

## 6. Conclusões

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, é necessário verificar o quão estes sistemas são fáceis e prazerosos de serem utilizados.

Para que os hospitais não necessitassem disponibilizar tempo nem incomodar seus radiologistas, optou-se por utilizar testes com voluntários, no que fosse possível analisar sem a presença dos radiologistas.

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.

## Referências

- Alapetite, A., Boje, A.H. and Morten, H. (2009) Acceptance of speech recognition by physicians : a survey of expectations, experiences, and social influence, In: *International journal of human-computer studies*, ISSN 1071-5819, v. 67, n. 1, p. 36-49.
- Bernsen, N.O. and Dybkjaer, L. (2001) Usability evaluation in spoken language dialogue systems and their components. In: *Proceedings of the workshop on Evaluation for Language and Dialogue Systems*, v. 9, Toulouse, France.

- Bhan, S.N., Coblenz, C., Norman, G.R. and Ali, A.H. (2008) Effect of voice recognition on radiologist reporting time. In: *CARJ*, v. 59, n. 4, October.
- Chin, J. P., Diehl, V. A. and Norman, K. L. (1998) Development of an instrument measuring user satisfaction of the human-computer interface. In: *ACM CHI'88 Proceedings*, p. 213-218.
- Durling, S. and Lumsden, J. (2008) Speech Recognition Use in Healthcare Applications. In: *Proceedings of MoMM2008*, MoCoHe, ACM 978-1-60558-269-6/08/0011.
- Dybkjaer, L. and Bernsen, N.O. (2001) Usability evaluation in spoken language dialogue systems. In: *Proceedings of the ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems*.
- Hartikainen, M., Salonen, E. and Turunen, M. (2004) Subjective evaluation of spoken dialogue systems using SER VQUAL method. In: *INTERSPEECH-2004*, p.2273-2276.
- Hundt, W. et al. (1999) Speech processing in radiology. In: *Eur Radiol* n.9, p. 1451-1456.
- Larsen, L.B. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In: *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*, St. Thomas, U.S. Virgin Islands, USA, 2003.
- Mohr, D.N. et al. (2003) Speech recognition as a transcription AID: A randomized comparison with standard transcription. In: *Journal of Medical Informatics*, v. 10, n. 1, p. 85-93.
- Möller, S. (2005) Quality of telephone-based spoken dialogue systems, New York, NY: Springer.
- Nuance, Disponível em: <http://nuance.com/naturallyspeaking/>. Acesso em 20/12/2010.
- Robbins, A.H. et al. (1987) Speech-controlled generation of radiology reports. In: *Radiology* n. 164, p. 569-573.
- Rodger, J. A. and Pendharkar, P. C. (2007) A field study of database communication issues peculiar to users of a voice activated medical tracking application. In: *Decision Support Systems*, v. 43, n. 1, p. 168-180.
- Snomed. Disponível em: <http://www.ihtsdo.org/snomed-ct/>. Acesso em 07/03/2012.
- Walker, M. A., Passneau, R. and Boland, J. E. 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.
- White, K. S. (2005) Speech recognition implementation in radiology, New York: Springer-Verlag, 10.1007/s00247-005-1511-x.