BrowserVox: Uma Extensão De Interface De Voz Para Um Navegador Open-Source

Elizabete Munzlinger [1] elizabete@elizabete.com.br
Fabricio da Silva Soares [1] fabricio@fabricio.net.br
Carlos Henrique. Q. Forster [1] forster@ita.br

[1] Instituto Tecnológico de Aeronáutica
Divisão de Ciência da Computação
Praça Marechal Eduardo Gomes, 50
São José dos Campos, SP Brasil 12.228-900

RESUMO
The use of voice in computational systems interfaces have been utilized for many users who can enjoy the voice to communicate. The performance of current voice technologies and the business demand to improve usability and accessibility in their products and services has been increasing the use of voice interaction in multimodal interface (MMI) applications [11]. This paper presents the previous work improvement [6] through a Voice Interface (VUI) extension for Mozilla Firefox Web Browser. The BrowserVox extension is a plugin with improvements, where some problems of paper [6] was solved. During BrowserVox developement arose project questions about interface and interaction. This questions are described in this paper.

Categories and Subject Descriptors
H.5.2 [**User Interfaces**]: *Evaluation/methodology, Voice I/O*.
H.1.2 [**User/Machine Systems**]: *Human factors*.

General Terms
Design, Human Factors, Languages.

Keywords
Interface multimodal, reconhecimento de fala, síntese de texto, navegador *web*.

1. INTRODUÇÃO

1.1 Interfaces Multimodais
Aplicações da fala em interfaces de sistemas diversos vem sendo bem aceita pela maioria das pessoas que podem usufruir dessa forma de comunicação [8]. O crescente uso das Interfaces de Voz (VUI, Voice User Interface) [2][9] é também impulsionado por fatores como: **1)** Necessidade dos fabricantes em melhorar a usabilidade e acessibilidade de seus produtos e serviços, incorporando a *hardware* e *software* diversas formas de tratamento das necessidades especiais dos usuários, bem como aplicações especiais para alguns tipos de deficiências [11]. **2)** Robustez das atuais tecnologias de fala [2], como o Reconhecimento Automático de Fala (ASR, Automatic Speech Recognition) e a Síntese de Texto em Fala (TTS, Text to Speech) em associação com o uso de Gramáticas Auxiliares [7][8]. O fato de uma VUI ser invisível ao usuário possibilita que a mesma possa ser agregada como extensão aos mais diversos tipos de interfaces de sistemas novos ou pré-existentes, tornando-os sistemas de Interfaces Multimodais (MMI, Multi-Modal Interface), que utilizam múltiplos canais de comunicação para interação entre usuário e sistema [1][8]. Uma corrente ascendente de MMI, se dá pela aplicação de VUI em união com a dominante Interface Gráfica (GUI, Graphical User Interface) [8]. Como a GUI usa estrutura de representações visuais por meio de signos para representação de tarefas do sistema, os mesmos podem auxiliar no controle por voz, pois os processos e opções de comandos podem ser acompanhados visualmente.

1.2 Comparação entre Navegadores MMI
Há ainda hoje, uma carência de navegadores MMI com VUI para o idioma Português-Brasileiro. Dois esforços para suprir esta lacuna são: **1)** o conjunto de aplicativos DosVox, Linx e WebVox [12] que utilizam de ASR e TTS, para o Português-Brasileiro, mas são voltados para deficientes visuais e não possuem a Interface Gráfica; **2)** O navegador *web* multimodal BrowserVox [6], usa ASR e TTS para o Português-Brasileiro e possui Interface Gráfica e Interface de Voz podendo ser utilizado por usuários que não possuem deficiência visual mas que apresentam outras sérias deficiências, como as motoras. Este último navegador, já apresentado e avaliado em trabalho anterior como um aplicativo, é apresentado neste trabalho na forma uma extensão para o navegador Mozilla Firefox, onde são apresentadas alterações e melhorias para alguns problemas apresentados anteriormente [6].

Em geral, os navegadores MMI apresentam limitações que dificultam seu uso pelos usuários sejam eles portadores ou não de deficiências. As extensões FoxVox, FoxyVoice e SpeakIt [4], todas para o navegador Mozilla Firefox possuem somente TTS, sendo que a navegação se dá apenas por mouse. Opera 8.0 [10] e Conversa Web [3] possuem suporte a ASR e TTS mas apenas para o idioma Inglês. BrowserVox possui suporte a ASR e TTS em Português-Brasileiro. Conversa Web [3] e BrowserVox aceitam comandos para navegação por *links* em imagens através de um número apresentado sobre a mesma. BrowserVox aceita ainda a navegação por números nos *links* textuais. Quanto à TTS, o Opera 8.0 pode sintetizar todo o conteúdo textual, podendo ser exaustivo, ou apenas textos selecionados, sendo necessário uso de mouse/teclado, inviável para deficientes. BrowserVox efetua a leitura através de comandos variados como “Leia o texto”, “Quero ouvir o texto”, “Leia a notícia”, etc. Em relação à navegação, o Opera 8.0, funciona pelo comando “Opera next link”. BrowserVox aceita comandos com formações variadas para os links textuais, e ainda navegação numérica. Alguns navegadores MMI podem apresentar restrição de tipos de documentos, como o Opera 8.0, que aceita apenas documentos nos padrões XHTML+Voice [13]. BrowserVox consegue abstrair os *links* de qualquer linguagem de marcação que utilize as marcas de âncora <a>. No quesito plataforma, FoxVox funciona em SO Linux e Windows, e as demais apenas em Windows, pois em geral fazem

uso da TTS Engine do próprio SO. SpeechWeb [5], Opera 8.0 assim como BrowserVox utilizam das bibliotecas IBM Via Voice.

2. EVOLUÇÃO DO BROWSERVOX

O processo de evolução do BrowserVox, neste trabalho, teve como principal objetivo, previamente definido em [6] a migração do mesmo em uma extensão para o navegador Mozilla Firefox. Alguns dos problemas levantados em [6] são resolvidos. Neste procedimento, elegemos e resolvemos algumas novas questões de projeto de interação e de projeto de interface.

2.1 Descrição e Características

BrowserVox permite a navegação através de comandos no idioma Português-Brasileiro para *websites* neste idioma mas também de outros idiomas através de sistema de numeração por etiquetas. Suas características e diferenciais atuais, alcançados a partir deste trabalho e de trabalhos efetuados em [6][7][8] são: **a. Extensão para o navegador Mozilla Firefox:** adicionar uma funcionalidade extra para um aplicativo já conhecido do usuário pode oferecer facilidade de uso e adaptação. **b. Interface multimodal (MMI):** composta por GUI + VUI, oferece mais de um canal de interação, possibilitando a navegação pelos *websites* através da interação por meio do uso de *mouse*/teclado (convencional) e mais interação por comandos de voz. **c. Capacidade multiusuário:** permite que qualquer usuário possa utilizá-lo sem a necessidade de qualquer treinamento prévio para que o sistema possa reconhecer seus comandos de voz, em associação com gramáticas auxiliares. **d. Possibilidade de variação de comandos:** Algumas palavras que não pertencem aos comandos, ou ainda variações de sinônimos ou regionalismo podem ser ditas pelo usuário, proposital ou involuntariamente. Através da adição de palavras complementares nas gramáticas auxiliares os comandos se tornam mais naturais, sem prejudicar seu reconhecimento (habilidade esta desconhecida em outros sistemas com interfaces que usam VUI). **e. Navegação com comandos no idioma Português-Brasileiro:** para sites neste idioma, os comandos podem ser textuais referentes aos textos que representam os *links*. **f. Navegação em sites de outros idiomas:** através do sistema de etiquetas numéricas, é possível continuar a navegação com comando numéricos em Português-Brasileiro. **g. Navegação em links de imagens e links repetidos:** também através do sistema de etiquetas numéricas. **h. Síntese de texto:** textos podem ser sintetizadas em idioma Português-Brasileiro.

2.2 Linguagens e Tecnologias Usadas

O sistema é desenvolvido nas Linguagens de Programação Java e JavaScript, utilizado para manipulação do DOM para a criação de um sistema de etiquetas numéricas para os *links*. Usa como componente ASR e TTS o sistema IBM Via Voice por disponibilizar versão em Português-Brasileiro. Ainda o componente API IBM Java Speech Technology (JSAPI), que provê acesso e tratamento direto ao motor de reconhecimento (*engine*) do sistema IBM Via Voice.

2.3 Modelo de Arquitetura Seguido

A extensão BrowserVox segue o modelo de arquitetura (Figura 1) desenvolvido e utilizado em [8], e apresentado em [6].

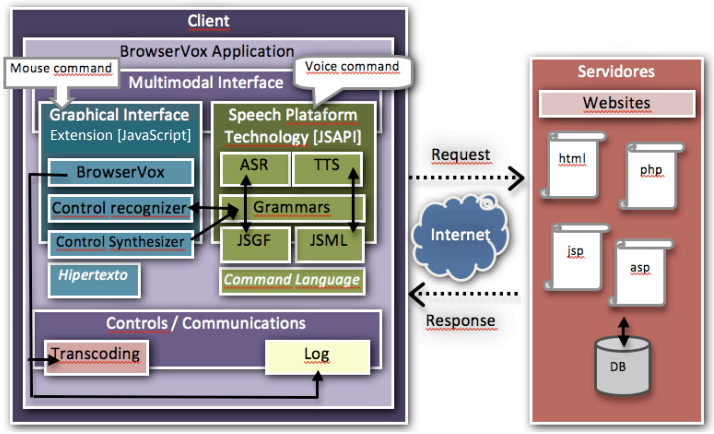

Figura 1. Arquitetura do sistema BrowserVox

No bloco *Graphical Interface* do módulo *Multimodal Interface* o elemento *Swing* que na versão anterior implementava a janela do aplicativo, foi substituído por classes em linguagem JavaScript para a implementação da extensão agregada ao navegador Mozilla Firefox que já possui GUI. As classes fazem a comunicação com bloco *Speech Plataform* da VUI e também a manipulação do documento HTML através do DOM (Document Object Model) [13] para inserir o sistema de etiquetas numéricas nos *links* da página apresentada.

2.4 Projeto da Gramática de Reconhecimento

A gramática mantém a mesma estrutura (Figura 2) criada para a versão anterior do navegador BrowserVox, definida e apresentada em [8]. O modelo de gramática foi desenvolvido a partir dos resultados obtidos no estudo sobre a influência do projeto de gramática em sistemas de voz [7] e mantido por prover a capacidade multiusuário e permitir trabalhar com comandos variados.

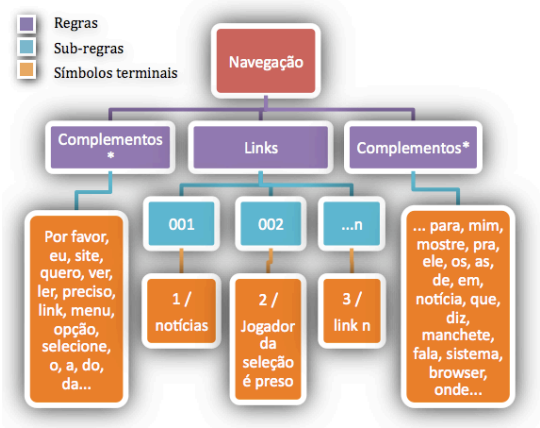

Figura 2. Estrutura de regras da gramática representada através da árvore

As gramáticas geradas alimentam os sistemas ASR e TTS e seguem os padrões JSGF (Java Speech Grammar Format) para a gramática de reconhecimento dos comandos e JSML (Java Speech Markup Language) para as gramáticas de síntese de texto do conteúdo do site.

2.5 Problemas Anteriores Resolvidos

a. Migração para uma extensão de um navegador open-source: Disponibilizar o BrowserVox como uma extensão para um navegador Open-Source popular é uma forma de difundir e disponibilizar o uso da interação por voz. Por vezes, usuários podem apresentam dificuldade ou resistência em instalar e aprender a usar novos aplicativos em sua máquina. Porém, adicionar novas funcionalidades aos aplicativos já instalados e conhecidos, torna-se uma característica agradável. **b. Tratamentos de elementos não textuais**: O texto dos *links* apresentam muitas vezes alguns elementos, aqui chamados de não textuais, por serem caracteres não pronunciáveis durante a leitura dos textos dos *links*. Um conjunto inicial destes símbolos, contendo, dentre outros, itens como hífen (-), ponto (.), vírgula (,), ponto-e-vírgula (;), setas (>,<), exclamação (!), interrogação (?), parênteses (()), colchetes ([]), chaves ({}), aspas (‘, “), foi tratado diretamente em código para serem eliminados das regras de gramática. **c. Tratamento de símbolos pronunciáveis:** Alguns elementos simbólicos pronunciáveis como arroba (@), porcento (%), reais (R$), dentre outros, foram também substituídos em código no momento da geração da gramática pelo seu correspondente nome por extenso. **d. Navegação em links de imagens**: Imagens são amplamente utilizadas para criação de *links* e não existe uma representação textual que componha o comando. Assim, foi criado um sistema que incorpora dinamicamente etiquetas numéricas para cada imagem. Dessa forma, para seguir o *link* correspondente à imagem, basta usar o comando numérico. **e. Navegação em links duplicados**: Muitos *links* em uma mesma página possuem texto repetido, como “clique aqui”, “veja aqui”, “saiba mais”, etc que acabam sendo repetidos nas regras da gramática. Assim, o sistema anterior reconhecia qualquer comando com este texto do *link* como sendo referente à primeira regra da gramática, mesmo que a intenção do usuário fosse acessar o último *link*. Com o sistema de etiquetas numéricas criado, este problema também foi contornado.

2.6 Questões de Projeto de Interação

As seguintes questões-chave foram considerados durante a fase de projeto de interação:

2.6.1 Não interromper a navegação em sites de diferentes idiomas:

Uma das características observadas durante a navegação é que ao seguir *links*, naturalmente, os usuários podem acessar uma página em outro idioma. Este evento não deve interromper o processo de interação entre usuário e sistema. Assim, foi projetado um **sistema de etiquetas numéricas** para os *links*. O usuário que se depara com uma página em outro idioma, consegue manter a navegação através dos comandos numéricos em Português-Brasileiro.

2.6.2 Preservar a privacidade na navegação:

A interação com o sistema deve ser agradável. Apesar de ser natural a navegação através da fala, por vezes esta opção pode gerar constrangimento para o usuário dependendo do conteúdo do *site* visitado. O sistema de etiquetas numéricas auxilia na preservação de sua privacidade, tornando a atividade mais confortável.

2.7 Questões de Projeto de Interface

2.7.1 Apresentação das etiquetas numéricas:

Ensaios foram efetuados buscando a melhor forma de incorporar as etiquetas numéricas ao conteúdo das páginas em associação com os *links*. Duas formas foram testadas:

a. Etiquetas flutuantes: com o objetivo de não alterar a estrutura da página foi testado uma forma de apresentação das etiquetas numéricas flutuantes, posicionadas à esquerda ou direita superior do *link*. A etiqueta é criada usando um elemento de divisão do HTML (tag) que é inserida no documento. A etiqueta é posicionada em virtude de informações da posição do elemento de *link* extraídas através de manipulação do DOM usando a linguagem JavaScript. Com os atuais recursos como CSS (Cascading Style Sheets) [13], que permite que os elementos (links, imagens, textos, etc) sejam posicionados visualmente na página independente de sua ordem no código HTML, o posicionamento das etiquetas flutuantes apresentou problemas de sobreposição e deslocamentos (Figura 3), apresentando-se de forma confusa ao usuário.

Figura 3. Recortes de áreas com links em um portal de notícias usando etiquetas flutuantes

b. Etiquetas inseridas nos links: outra forma de apresentação das etiquetas numéricas foi através de sua inserção dentro do elemento do *link* (tag <a>), também pela manipulação do DOM. Contudo, esta altera de alguma forma a apresentação dos elementos, como alteração no seu tamanho ou posicionamento original, etc. Ainda assim, essa configuração se mostrou mais adequada, pois apresenta as etiquetas relacionadas ao *link* de forma clara ao usuário (Figura 4).

Figura 4. Recortes de áreas com links em um portal de notícias usando etiquetas inseridas nos links

2.7.2 Visibilidade das etiquetas numéricas:

Buscando acessibilidade, diferentes formatações com variações de cor e contraste das mesmas foram testadas (Figura 5) e futuramente poderão ser configuráveis pelo usuário.

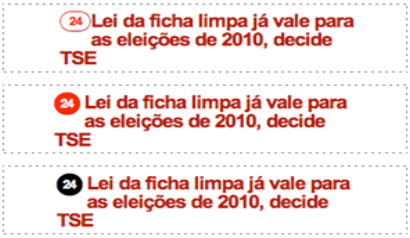

Figura 5. Recortes de áreas com links e etiquetas com diferentes formatações

2.7.3 Comandos de controle de janela

Além de utilizar na gramática comandos compostos pela nomenclatura da janela do *Firefox*, foram adicionados às regras da gramática sinônimos e comandos utilizados para representar as mesas tarefas em outros navegadores com o objetivo de tornar o controle da janela mais familiar ao usuário, principalmente se o mesmo não for usuário deste navegador.

3. CONCLUSÃO E TRABALHOS FUTUROS

Pelo presente trabalho, conclui-se que é viável o uso da extensão VUI para o navegador Mozilla Firefox. O sistema de etiquetas numéricas vem auxiliar na resolução de problemas recorrentes, e permite que o usuário possa navegar em *sites* através de comandos do Português-Brasileiro, mesmo que o *site* esteja em outro idioma. As questões levantadas sobre de Projeto de Interface e de Interação são importantes para auxiliar na usabilidade do sistema. Um cuidado especial foi tomado na forma com a qual as etiquetas numéricas são incorporadas ao documento HTML, de modo que mantenham uma perceptível associação com o *link*. As principais limitações atuais do trabalho se referem aos relacionados à construção das páginas HTML, em especial coma aquelas que não seguem os Padrões Web e as diretivas de acessibilidade projetadas pela W3C. Exemplos são páginas construídas totalmente com animações ou frames (divisões no navegador). Também sites com código mal escrito que gera *links* quebrados ou escondidos. Como o foco da extensão até o momento é a navegação pelos *links* do *site*, controle de comandos para conteúdo multimídia como vídeos, som, etc, e navegação para preenchimento de formulários são atuais desafios. Trabalhos futuros incluem ainda o melhoramento no *feedback* da interface da extensão. Continuação dos tratamentos a símbolos, abreviaturas entre outros também são necessários. Ainda o estudo da migração da extensão para outros navegadores populares.

4. REFERÊNCIAS

[1] BELLIK, Yacine. and TEIL, Daniel. 1993. **A multimodal dialogue controller for multimodal user interface management system application: a multimodal window manager**. In INTERACT '93 and CHI '93. ACM, New York, 93-94. DOI= http://doi.acm.org/10.1145/259964.260124.

[2] COHEN, Michael H.; GIANGOLA, James P.; BALOG, Jennifer. **Voice User Interface Design**. Reading: Addison Wesley, 2004.

[3] CONVERSA WEB. **Conversa Web - Conversational Computing. 1998.** Disponível oline em: [http://www.conversa.com] Acesso em maio de 2010.

[4] FIREFOX. **Add-ons for Firefox.** Disponível online em: [https://addons.mozilla.org/]

[5] FROST, R. A. 2002. **Speech web: a web of natural-language speech applications**. In *Eighteenth National Conference on Artificial intelligence* (Edmonton, Alberta, Canada, July 28 - August 01, 2002). R. Dechter, M. Kearns, and R. Sutton, Eds. American Association for Artificial Intelligence, Menlo Park, CA, 998-999. 2002.

[6] MUNZLINGER, Elizabete; FORSTER, Carlos Henrique Quartucci. **Desenvolvimento e Avaliação de um Sistema Multimodal e Multiusuário de Navegação Web**. In WebMedia '08. ACM, New York, NY, 29-32. DOI= http://doi.acm.org/10.1145/1809980.1809989.

[7] MUNZLINGER, Elizabete; SOARES, Fabricio da Silva; FORSTER, Carlos Henrique Quartucci . **Evaluation of a Multi-user System of Voice Interaction Using Grammars**. In: INTERACT 2007. Springer Berlin: LNCS Lecture Notes in Computer Science, 2007. v. 4663. p. 452-455.

[8] MUNZLINGER, Elizabete. **Extensões Multimodal e Multiusuário de Interface Gráfica e Interface de Voz Baseadas em Tecnologias de Fala e Modelos de Interação**. 2009. 108f. Dissertação (Mestrado em Engenharia Eletrônica e Computação) Divisão de Ciência da Computação, Instituto Tecnológico de Aeronáutica, São José dos Campos: ITA, 2009.

[9] OLIVE, Joseph P. 1999. **The Voice User Interface**. In: Proceedings of the IEEE, 1999. GLOBECOM '99 (Rio de Janeiro, Brazil, pgs: 2051-2055 vol.4) DOI: 10.1109/GLOCOM.1999.827565.

[10] OPERA. **Control Opera Using Your Voice.** 2005. Disponível online em: [http://www.opera.com/browser/tutorials/voice/]. Acesso em junho de 2010.

[11] ROCHA, Heloísa Vieira da; BARANAUSKAS, Maria Cecília Calani. **Design e Avaliação de Interfaces Humano-Computador**. São Paulo, NIED, UNICAMP, 2003.

[12] SACI. **SACI.** Disponível oline em: [http://www.saci.org/]

[13] W3C. **The World Wide Web Consortium**. Disponível online em: [http://www.w3.org].