Tecnologias para o Desenvolvimento de Sistemas de Diálogo Falado em Português Brasileiro

Jefferson Morais, Nelson Neto e Aldebaro Klautau

[1]Laboratório de Processamento de Sinais – LaPS
Universidade Federal do Pará – UFPA
Rua Augusto Correa, 1 – 660750-110 – Belém, PA, Brasil
http://www.laps.ufpa.br

{jmorais,nelsonneto,aldebaro}@ufpa.br

Abstract. *This work discusses the integration of available technologies for developing spoken dialog systems in Brazilian Portuguese. As a proof-of-concept, it describes a system for non-visual and on-line Web search on Windows. The prototype system is based on Microsoft's Speech Application Programming Interface (SAPI), which provides an interface that allows the establishment of a dialog, where the system asks the site and query word. The system then reads aloud the page contents. The system itself coordinates the interaction with the user and is currently limited to query by the name of countries.*

Resumo. *Este trabalho discute a integração das tecnologias disponíveis para o desenvolvimento de sistemas de diálogo falado em Português Brasileiro. Como exemplo, o mesmo apresenta o protótipo de um sistema para busca não-visual e on-line na Web, em ambiente Windows. Com base na interface fornecida pela Microsoft de reconhecimento e síntese de voz denominada Speech Application Programming Interface (SAPI), o sistema estabelece um diálogo falado com o usuário, questionando-o sobre o site e a palavra que deseja consultar via síntese de fala. Como resposta, o conteúdo principal da página é lido. O próprio sistema coordena as interações com o usuário e atualmente é limitado à busca pelo nome de países.*

1. Introdução

Os sistemas de diálogo falado (SDS, de "spoken dialog systems") vêm evoluindo [Borodin et al. 2007, Bohus et al. 2007], contudo a interação homem-máquina ainda é bastante distinta de uma conversa informal entre duas pessoas e os obstáculos encontrados, hoje, pelas tecnologias de fala (ou voz) são diversos. Este trabalho concentra-se em discutir como integrar as tecnologias de reconhecimento automático de voz (ASR) e síntese de voz (TTS) disponíveis para o desenvolvimento de SDS em Português Brasileiro (PB). Como prova do conceito, foi desenvolvida uma aplicação simples, que permite realizar pesquisas de países na Web com base na interface *Speech Application Programming Interface* (SAPI) da Microsoft[1], que recentemente disponibilizou para avaliação o seu reconhecedor de voz para PB em versão beta[2]. O protótipo desenvolvido é um primeiro passo no desenvolvimento de um sistema para navegação Web não-visual.

[1]http://www.microsoft.com/speech/
[2]http://www.microsoft.com/portugal/mldc/default.mspx

2. SAPI da Microsoft

A SAPI é uma interface fornecida pela Microsoft de reconhecimento e síntese de voz para o desenvolvimento de aplicações baseadas nos sistemas operacionais Windows. Essa tecnologia faculta aos programadores acesso ao serviço de voz fornecido por um *engine* de síntese e reconhecimento de voz, conforme ilustra a Figura 1.

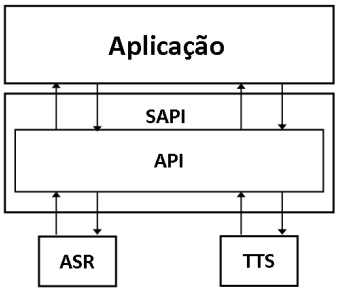

Figura 1. Arquitetura de uma aplicação utilizando a SAPI.

3. O Protótipo Desenvolvido

O protótipo aqui apresentado utiliza a voz como principal modalidade de interação com o usuário, tanto como interface de entrada de dados (ASR), como interface de saída ou *feedback* (TTS). Uma das telas do aplicativo pode ser visualizada na Figura 2.

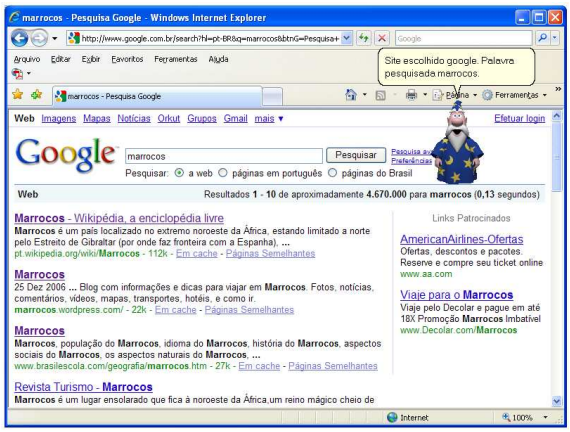

Figura 2. Uma das telas do aplicativo para pesquisa não-visual na web.

A rede com as transições de estado do sistema consiste dos seguintes passos:

- Primeiramente, o usuário deve dizer qual o *site* (*google* ou *wikipedia*) e a palavra (país) que deseja consultar.
- Caso o *site* de busca escolhido seja o *wikipedia*, o conteúdo (texto) principal da página resultante da pesquisa é sintetizado. Por fim, a opção de uma nova busca é disponibilizada.
- Caso o *site* escolhido seja o *google*, uma lista com as âncoras resultantes da pesquisa é sintetizada ao usuário. Em seguida, é solicitado que o usuário escolha um dos *links* disponíveis. Finalmente, o conteúdo principal da página escolhida é sintetizado e o usuário tem a opção de acessar outro *link* ou realizar uma nova pesquisa.

Como já dito, a SAPI 5.1 foi a principal interface de programação utilizada, porém o emprego de outras DLLs[3] se mostrou necessário durante a elaboração de alguns passos do sistema, são elas: Shdocvw e Mshtml. A interface Shdocvw encontra-se diretamente relacionada ao navegador *Windows Internet Explorer*, ou seja, todas as funcionalidades desse aplicativo podem ser controladas pelo componente Shdocvw, incluindo as opções de navegação, gerenciamento de histórico, entre outras. Já a interface Mshtml, está ligada a Shdocvw e é capaz de analisar e renderizar um documento estilo HTML, o que permite ao programador referenciar todos os objetos presentes em um determinado *site*.

Atualmente, a Nuance disponibiliza um sintetizador de voz gratuito na língua portuguesa[4]. Sua arquitetura suporta SAPI 4.0 e é licenciado pela Microsoft especificamente para uso com o *Microsoft Agent*[5] (MSagent). Em 2007, a Microsoft iniciou as gravações para o novo sintetizador de voz natural em Português Europeu. Porém, sua versão beta ainda não encontra-se disponível. Assim, a síntese de voz desta aplicação é realizada através de agentes animados (*Agents*). Apesar de não ser parte da SAPI, o MSagent é uma tecnologia diretamente relacionada, já que permite criar poderosos *Agents* e empregá-los em aplicações para a plataforma Microsoft Windows, além de associá-los aos mecanismos de síntese e reconhecimento de voz (similar à descrita em [Rodrigues et al. 2004]).

Para executar ASR, uma gramática precisa ser definida para que o aplicativo saiba que ação executar quando uma determinada palavra lhe for enviada. Existem dois tipos de gramática: livre de contexto e para ditado. Na gramática livre de contexto, as palavras passíveis de reconhecimento estão limitadas às regras que informam que palavras podem ser ditas, ou seja, possuem um domínio específico e limitado. Já a gramática para ditado, trabalha com a idéia de que todas as palavras selecionadas precisam ser identificadas. As aplicações para ditado continuam distantes do desejável diálogo espontâneo.

A interface SAPI dá suporte a gramática livre de contexto e para ditado. No entanto, o *engine* para reconhecimento em PB da Microsoft em sua versão beta, utilizado nesta aplicação, ainda não dispõe da gramática para ditado. Assim, uma gramática XML seguindo o padrão de texto SAPI[6] foi construída. Uma amostra da gramática elaborada pode ser conferida abaixo. Além da criação de regras gramaticais fixas, regras dinâmicas contendo a lista com as âncoras resultantes de pesquisas realizadas no *site google* também são criadas ao longo da interação aplicação-usuário.

```
<RULE NAME="pesquisa" TOPLEVEL="ACTIVE">
   <L>
       <P>pesquisa</P>
       <P>pesquisar</P>
   </L>
   <O><P>no</P></O>
   <L PROPNAME="SITE" PROPID="SITE">
       <P VALSTR="google">google</P>
       <P VALSTR="wikipedia">wikipedia</P>
   </L>
</RULE>
```

[3] msdn.microsoft.com/en-us/library/aa741313(VS.85).aspx
[4] http://www.nuance.com
[5] http://www.microsoft.com/msagent
[6] http://msdn.microsoft.com/en-us/library/ms723635(VS.85).aspx

É sabido que sistemas de diálogo baseados exclusivamente em processamento de voz têm alguns inconvenientes que podem resultar em interações pouco eficientes [Delgado e Araki 2005]. Além dos possíveis erros de reconhecimento por parte do *engine*, dado as restrições de vocabulário e domínio, os usuários podem "fugir" do contexto da aplicação, fato que também contribui para o diminuição da taxa de acerto de palavras [Williams e Young 2007]. Na tentativa de prevenir esses erros, esta aplicação impõe uma taxa de confiança de reconhecimento de 0,7 ao *engine*, e busca, sempre que possível, confirmar explicitamente os dados solicitados pelo usuário, permitindo que ele volte ao estado anterior e corrija sua solicitação.

O próprio sistema coordena as interações, ou seja, a aplicação guia o usuário ao longo das transições de estado, o que diminui a possibilidade de solicitações fora do contexto, estratégia conhecida na literatura como *system-directed interaction* [Delgado e Araki 2005]. Em uma tentativa de reduzir as limitações em termos de flexibilidade impostas por essa estratégia, criou-se a opção do usuário interromper o TTS do conteúdo da página pesquisada via comando de voz sempre que achar necessário. Finalmente, em virtude da simplicidade dos diálogos presentes nesta aplicação, não foi observado um problema freqüente em sistemas de diálogo: dificuldade dos usuários em entender o fluxo do diálogo, o que causa problemas em saber o que fazer e o que dizer.

4. Conclusões

Construir sistemas de diálogo confiáveis e naturais é um desafio para a engenharia em função das limitações impostas pelo atual estágio do ASR. O sistema desenvolvido exemplifica a construção de aplicações relativamente simples baseadas em voz para ambiente Windows, via interface de programação SAPI e utilizando recursos disponíveis, contornando a atual escassez de recursos para o PB. Uma grande dificuldade encontrada foi a elaboração de estratégias de diálogo mais amplas e amigáveis, muito em função da não utilização aqui de ferramentas específicas para modelagem de diálogos, além do reconhecedor para PB da Microsoft ainda não suportar sistemas de ditado. Como trabalho futuro, pretende-se adaptar este protótipo ao *framework Olympus*[7] e ao padrão VoiceXML[8].

Referências

D. Bohus, S. G. Puerto, V. Keri D. Huggins-Danies, G. Krishma, R. Kumar, A. Raux e S. Tomko (2007). Conquest - an open-source dialog system for conferences. *North American Chapter of the Association for Computational Linguistics*.

Y. Borodin, J. Mahmud, I.V. Ramakrishnan e A. Stent (2007). The hearsay nonvisual web browser. *International World Wide Web Conference*.

R. Delgado e M. Araki (2005). *Spoken, Multilingual and Multimodal Dialogue Systems*. John Wiley & Sons, Ltd.

Paula Lucena Rodrigues, Bruno Feijão e Luiz Velho (2004). Expressive talking heads: uma ferramenta de animação com fala e expressão facial sincronizadas para o desenvolvimento de aplicações interativas. Em *Proceedings of Webmedia. SBC*.

J. D. Williams e S Young (2007). Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language 21 (2007) 393-422*.

[7] www.ravenclaw-olympus.org/

[8] www.w3.org/TR/voicexml20