Interface do usuário baseada em voz como ferramenta para promover o ensino/aprendizagem de língua estrangeira

André Brasileiro da Silva, Luiz Fernando da Silva Fernandes, Valéria Farinazzo Martins

Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie
Rua da Consolação, 930 Cep 01302-907 - Consolação - São Paulo - SP - Brasil

{andrebrasiliano, lfernandesfotos}@hotmail.com,
valeria.farinazzo@mackenzie.br

Abstract. *This paper studies the use of Voice User Interface (VUI) to aid on foreign language learning – to improve the listening, understanding and pronunciation of English language. Using recognition and synthesis voice technology, an application called VAL is built; this application follows a specific methodology for developing VUI applications.*

Resumo. *Este artigo estuda o uso de interfaces do usuário baseada em voz no auxílio ao aprendizado de língua estrangeira, inserida no contexto de melhorar atividades práticas de escuta e compreensão de textos, bem como a pronúncia correta de palavras. Através de técnicas de reconhecimento e síntese de voz, foi desenvolvida a aplicação VAL, que segue uma metodologia específica para o desenvolvimento de aplicações VUI.*

1. Introdução

A partir do momento em que mais pessoas começam a ter computadores e dispositivos eletrônicos em geral, a usabilidade das interfaces homem-máquina torna-se requisito fundamental em todas as classes de aplicações que sejam executadas nestes equipamentos. O uso das interfaces não convencionais – tais como Realidade Virtual, Realidade Aumentada, interfaces gestuais e interfaces baseadas em voz – surge como uma forma de tentar tornar estas aplicações mais naturais e mais fáceis de serem usadas, gerando maior satisfação aos usuários não especialistas [Oliveira Neto, Salvador e Kawamoto 2010].

A voz pode ser apontada como uma das formas mais naturais de interação entre pessoas. Desde a década de 50, estudos em Inteligência Artificial vislumbram o uso da voz como ferramenta para a interação entre máquinas e pessoas, mas as limitações de hardware e software foram impactantes. Somente depois dos anos 90 é que a tecnologia de reconhecimento de voz contou com uma significativa melhora e sistemas puderam ser efetivamente criados e usados [Mctear 2002], [Cohen, Giangola e Balogh 2004].

As interfaces do usuário baseadas em voz – dos termos em inglês *Voice User Interface* (VUI) e *Speech User Interface* (SUI) – são interfaces em que o sistema capta as entradas *por voz* do usuário, realiza determinada ação correspondente ao *entendimento* da requisição do usuário, e gera uma saída, geralmente, por voz – pré-gravada ou sintetizada [Damasceno, Pereira e Brega 2005].

Tradicionalmente, o desenvolvimento de VUI tem focado em aplicações comerciais, tais como em sistemas para busca e recuperação de informações em mercados de ações, horário e reserva de vôos e reserva de hotéis, auxílio à lista telefônica, guia de restaurantes, bares e filmes. Porém, já existem algumas iniciativas voltadas para o acesso à informação para a Educação [Estabel, Moro e Santarosa 2006]. O objetivo principal deste artigo é apresentar as possibilidades de uso de interfaces baseadas em voz nas práticas de ensino/aprendizagem de línguas estrangeiras, bem como apresentar os benefícios e impactos do uso desta tecnologia como ferramenta pedagógica.

Após o estudo do referencial teórico e o levantamento de requisitos, foi desenvolvida a aplicação VAL – ***V****oice* ***A****pplication* ***L****earning* - um ambiente para exercitar a escuta do idioma estrangeiro, compreensão de textos e prática de pronúncia de palavras.

Este artigo está organizado como descrito a seguir. A seção 2 apresenta os fundamentos conceituais necessários para o entendimento da aplicação criada, a saber: reconhecimento de voz, síntese de voz e VUI; a seção 3 descreve a aplicação *Voice Application Learning*, abordando as principais fases de seu desenvolvimento. Finalmente, a seção 4 encerra o trabalho apontando as conclusões do trabalho e oportunidades de pesquisa futuras.

2. Fundamentos Conceituais

A fim de entender a aplicação desenvolvida *Voice Application Learning*, alguns conceitos fundamentais devem ser aqui explicados: (i) Reconhecimento de Voz, (ii) Sintetização de Voz – do termo em inglês *Text-to-Speech* (TTS) – e, (iii) VUI.

2.1. Reconhecimento de Voz

O reconhecimento de voz, conforme apresentado na Figura 1, consiste de uma série de módulos projetados para capturar uma entrada de voz (emitida pelo usuário), entender o que foi capturado, executar as transações ou tarefas computacionais, e responder de maneira apropriada [Deng e Huang 2004], [Cohen, Giangola e Balogh 2004].

Figura 1. Módulos de Reconhecimento de Voz

O ponto de finalização – *endpointing* - detecta o início e o final da fala – através da captura do silêncio - e determina a forma de onda. A onda é empacotada e enviada para o módulo de extração das características, que transforma a demarcação do que foi

ditado em fonemas e a cada um deles é atribuído um número – chamado de vetores de características. Em seguida, o módulo reconhecedor usa a sequência de vetores de características para determinar as palavras que foram ditas pelo usuário. O módulo de entendimento de linguagem natural atribui significado às palavras que foram ditas, através de um conjunto de blocos de valores. Um bloco é definido para cada item de informação que é relevante para a aplicação, ou seja, palavras-chaves. Assim, o módulo de gerenciamento do diálogo é iniciado. É o gerenciador de diálogo que determina as ações que o sistema deve fazer dentre as várias possibilidades, tais como acesso ao banco de dados ou executar uma transação.

2.2. Sintese de Voz

A tecnologia de sintetização de voz é o processo que converte texto em voz. O sintetizador recebe um texto na forma digital e faz sua vocalização. Um programa de síntese de voz é útil para vocalizar informações resultantes de consultas à base de dados e em situações em que o usuário não pode desviar a atenção para ler algo ou não tem acesso ao texto escrito; um sistema com interface do usuário baseada em voz pode usar um módulo para sintetização de voz ou utilizar mensagens pré-gravadas quando não houver variação da informação a ser prestada ao usuário.

Embora a tecnologia TTS ainda não reproduza com fidelidade a qualidade da voz humana gravada - vale à pena destacar que, até o momento, os sintetizadores de voz não conseguem representar entonação - ela tem melhorado muito nos últimos anos. Tipicamente, o uso da voz humana pré-gravada está atrelado aos *prompts* e ao envio de mensagens para os usuários. No entanto, determinadas aplicações, como leitores de e-mail e notícias, tem dados muito dinâmicos. Nesses casos, uma vez que os textos das mensagens não podem ser previstos, pode-se usar a tecnologia TTS para criar os discursos de saída [Cohen, Giangola e Balogh 2004].

2.3. Voice User Interface

Interface do usuário baseada em voz consiste na interação de uma pessoa com um sistema através de voz, utilizando uma aplicação de linguagem falada. São capazes não somente de reconhecer a voz do usuário, mas compreender o que ele diz e responder a estas entradas, geralmente, em tempo real [Lai 2000], [Shneiderman 2000].

Este tipo de interface inclui elementos tais como: *prompts* ou mensagens do sistema, gramáticas e lógica de diálogo ou fluxo de chamada (*call flow*). Os *prompts* são todas as mensagens de voz pré-gravadas ou sintetizadas que devem ser executadas durante o diálogo com o usuário. Gramáticas definem todas as palavras, sentenças ou frases que podem ser ditas pelo usuário em resposta a um *prompt*. A lógica de diálogo define todas as ações a serem tomadas pelo sistema em determinado ponto da interação, tais como um acesso à base de dados [Cohen, Giangola e Balogh 2004], [Chamberlain et al 2004], [Bosch et al 2004].

Existem diferenças substanciais no desenvolvimento de Interfaces Gráficas do Usuário (GUI) e VUI, desde que a voz não está visível ao usuário como no caso das GUIs e também pelo seu caráter transiente, o que pode aumentar consideravelmente a carga cognitiva exigindo uma memória de curta duração do usuário mais ativa; outro ponto a considerar é que entradas por voz são muito mais rápidas que as entradas via teclado, ao passo que as saídas por voz são mais lentas do que leituras textuais feitas

superficialmente. Além disto, na VUI, a comunicação tende a ser serial e com canal único – ou seja, o usuário não é capaz de escutar e falar ao mesmo tempo, nem é capaz de ouvir mais de uma fala ao mesmo tempo. Todos estes argumentos supracitados levam a crer que projetar VUI é razoavelmente diferente de se projetar GUI e, por isto, necessitam de uma metodologia diferenciada.

3. Voice Application Learning

A *Voice User Application* (VAL) consiste em uma aplicação para o auxílio no aprendizado de língua estrangeira para pessoas não-nativas. Esta aplicação permite que qualquer pessoa interessada em treinar a escuta, a compreensão de textos e a pronúncia correta de palavras em outros idiomas possa o realizar através de seu uso.

O processo utilizado para o desenvolvimento da aplicação, baseado em Cohen, Giangola e Balogh (2004) e em Lamel, Minker e Paroubek (2000), é composto pelas seguintes fases: Definição dos requisitos, Projeto de Alto Nível, Projeto Detalhado, Desenvolvimento, Teste e avaliação e *Tuning*.

3.1. Definição dos Requisitos

Nesta etapa, o principal objetivo é adquirir um conhecimento detalhado da aplicação em questão, suas metas, características e funcionalidades desejadas, usuários finais, as motivações dos usuários para utilizarem a aplicação e o cenário de uso. Deve-se entender, também, o contexto de negócio. Sendo assim, especifica-se:

- Definição dos usuários finais: a) Aluno: qualquer pessoa interessada em utilizar a aplicação, sem que seja necessário haver uma fase de treinamento prévio do sistema; b) Professor: responsável por inserir todos os arquivos de texto e áudio de língua inglesa em cada nível de conhecimento do aluno. Sobre estes usuários, foi possível definir:
    - Frequência de uso: a) aluno: esporádico a diário; b) professor: esporádico;
    - Nível de conhecimento da aplicação e de conhecimento da tecnologia: variável para ambos;
    - Sexo e faixa etária: feminino e masculino, com faixa etária variável, para ambos.
- Definição do Ambiente: o ambiente físico em que se dará o uso desta aplicação poderá ser uma escola de línguas ou em qualquer outro ambiente, preferencialmente em ambiente silencioso;
- Definição das tarefas: definem as regras do negócio. A aplicação permite que sejam introduzidos (pelo professor) áudios sobre determinados assuntos, textos referentes ao áudio e também um questionário (texto que será transformado pelo sistema em um arquivo de áudio) que o aluno deverá responder (por voz) a fim de que se possa verificar se a compreensão sobre a historia foi suficiente. Assim, são tarefas atribuídas aos dois tipos de usuários:
    - Aluno: Conforme mostrado no diagrama de Casos de Uso descrito na Figura 2, para o aluno, são suas opções:

- Só ouvir o arquivo de áudio, correspondente a uma historia;
- Ouvir o arquivo de áudio e acompanhar o texto;
- Ouvir cada pergunta do questionário disponibilizado depois do áudio e responder por voz;
- Ouvir cada pergunta e acompanhar o texto simultaneamente e depois responder por voz.

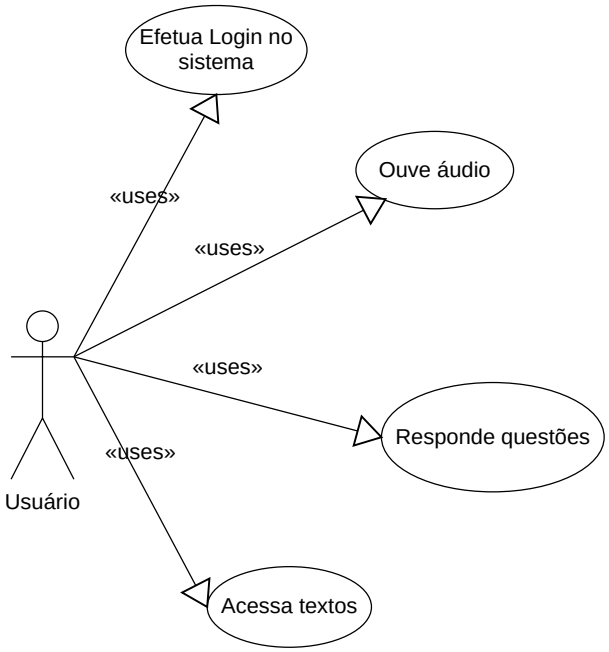

Figura 2. *Use Case* para o ator Aluno

o Professor: As mesmas funcionalidades permitidas para alunos, além de (Figura 3):

- Cadastra Usuários: pode cadastrar os usuários que são alunos;
- Cadastra textos, de áudio e de questionários: cadastro do texto, do áudio, do grau de dificuldade deste texto e também as perguntas que estão relacionadas ao texto;

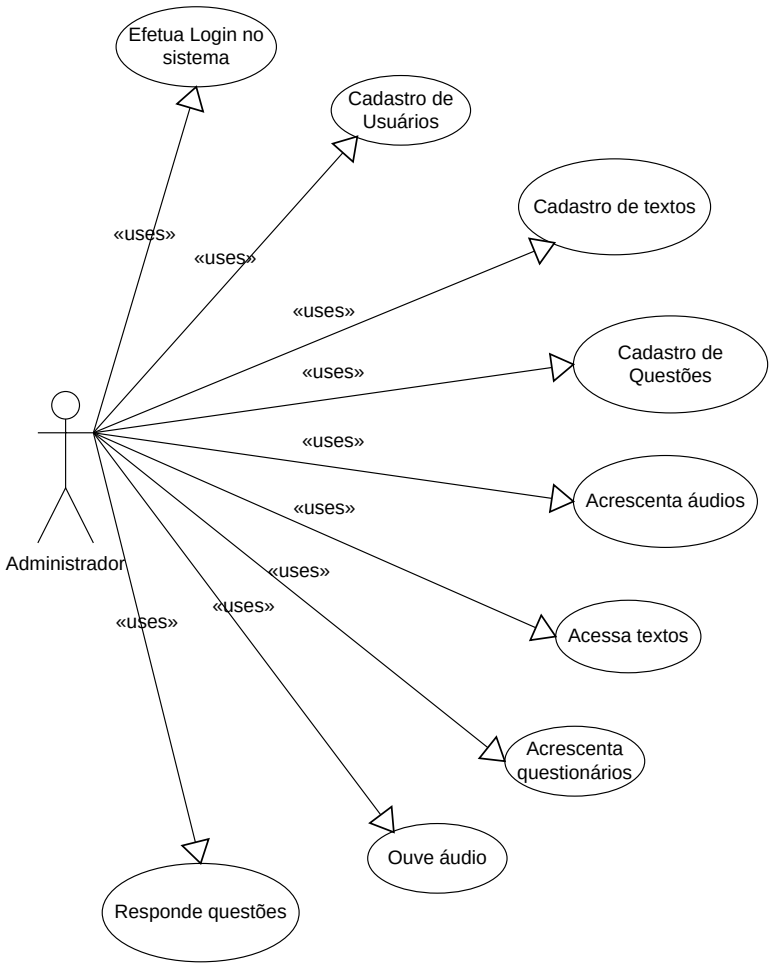

Figura 3. *Use Case* para o ator Professor

3.2. Projeto de Alto Nível

Esta fase é geralmente rápida, mas, tem um papel crucial, visto que cria uma experiência - do usuário - consistente, efetiva e única. O objetivo é encapsular os requisitos de forma mais concreta para guiar o projeto e tomar decisões sobre estratégias e elementos de diálogo que permeiam o projeto, alcançando consistência. Os passos desta fase para a aplicação em questão são:

- Definição dos critérios-chave: a) uma interface bastante intuitiva, sem necessidade de treinamento para seu uso; b) *feedback* adequado; c) alta taxa de reconhecimento de voz.
- Estratégias do diálogo: o diálogo da aplicação consistirá em apresentar, ao usuário, perguntas sintetizadas por *text-to-speech*; a resposta a esta pergunta dada pelo usuário por voz; o *feedback* da aplicação, a respeito da correta ou não resposta dada pelo usuário a cerca da compreensão do texto anteriormente apresentado ao usuário.

- Elementos pervasivos do diálogo: se a aplicação tiver uma taxa de reconhecimento baixa sobre a entrada do usuário, deve haver um mecanismo para solicitar a re-entrada do usuário.

A fim de uma melhor visualização da arquitetura do sistema, é apresentada a Figura 4, a seguir.

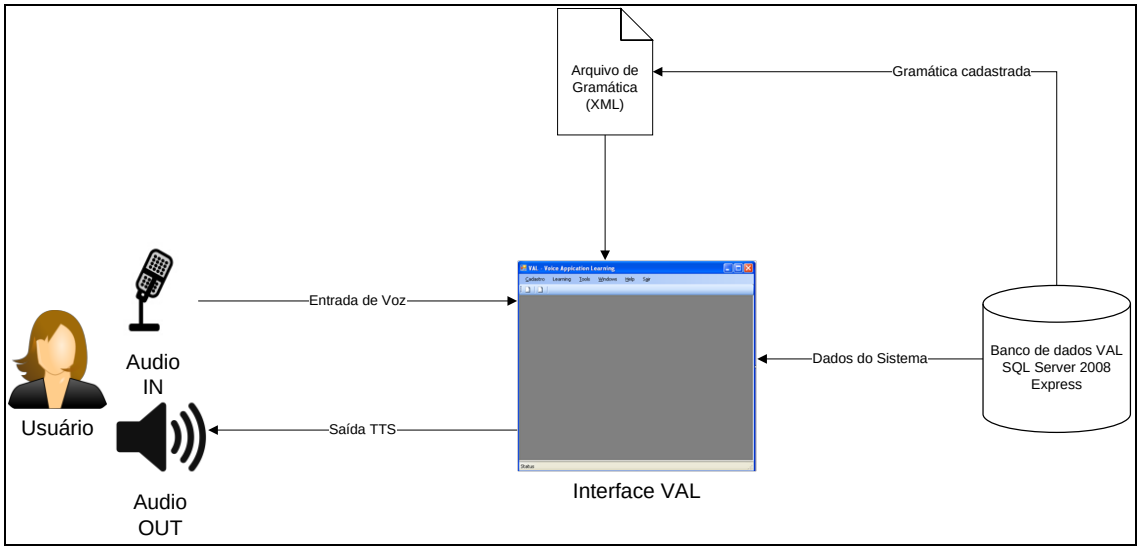

Figura 4. Arquitetura da aplicação

A aplicação carrega o texto referente a uma pergunta e o sintetiza em voz e a dispara para o usuário. O usuário fornece a resposta correspondente a esta pergunta através de um microfone (áudio - IN). A aplicação, que já fez a carga da gramática, realiza a consulta à base de dados e devolve a resposta – confirmando se o usuário acertou ou não a pergunta – por voz.

3.3. Projeto Detalhado

Nesta etapa, para cada diálogo devem ser descritos alguns componentes, entre eles: os *prompts* iniciais, a gramática de reconhecimento, o gerenciamento de erros e a especificação de ação. Um exemplo deste projeto detalhado em relação à aplicação – uso de um questionário - é apresentado a seguir na Tabela 1:

Tabela 1: Exemplo de diálogo entre o sistema e o usuário

SISTEMA:	Where is living John?
USUÁRIO:	He is living in Washington.
SISTEMA:	You are correct. Congratulations.

Se o sistema não entender o que o usuário diz, ou seja, se a confiabilidade do que o usuário está dizendo não alcançar um nível satisfatório – por exemplo, 70% - então o sistema deve requisitar a reentrada da resposta do usuário (Tabela 2):

Tabela 2: Exemplo de erro no reconhecimento da fala do usuário

SISTEMA:	Where is living John?
USUÁRIO:	He is living in Washington.
SISTEMA:	Sorry. Could you repeat this answer, please?

3.4. Desenvolvimento

O arquivo de gramática, que faz uso da tecnologia XML, é preenchido dinamicamente pelo sistema e descreve todas as palavras que estão passíveis de reconhecimento. Quanto maior o número de palavras que a gramática contiver, maior a precisão de reconhecimento de palavras, uma vez que o sistema passa a apresentar um maior número de hipóteses para as palavras ditas pelo usuário. Por outro lado, se a gramática for menor, maior é a chance de reconhecer o que o usuário fala.

O papel do banco de dados no sistema é o de armazenar a gramática entendida pelo sistema e também armazenar os dados cadastrados pelos professores – novos usuários, textos, áudios e questões. Para esta aplicação, o Sistema Gerenciador de Banco de Dados utilizado foi a Microsoft SQL Server 2008 Express.

O cadastro dos textos realizado pelo professor, Figura 5, faz a inserção do texto, arquivo .rtf, referente ao que o aluno deve ler, seu respectivo áudio (arquivo .wav), além do grau de dificuldade (iniciante, intermediário ou avançado). Após o cadastro do texto, o professor deve fornecer as questões que comporão a parte de *listening* e interpretação do texto, conforme mostrado na Figura 6.

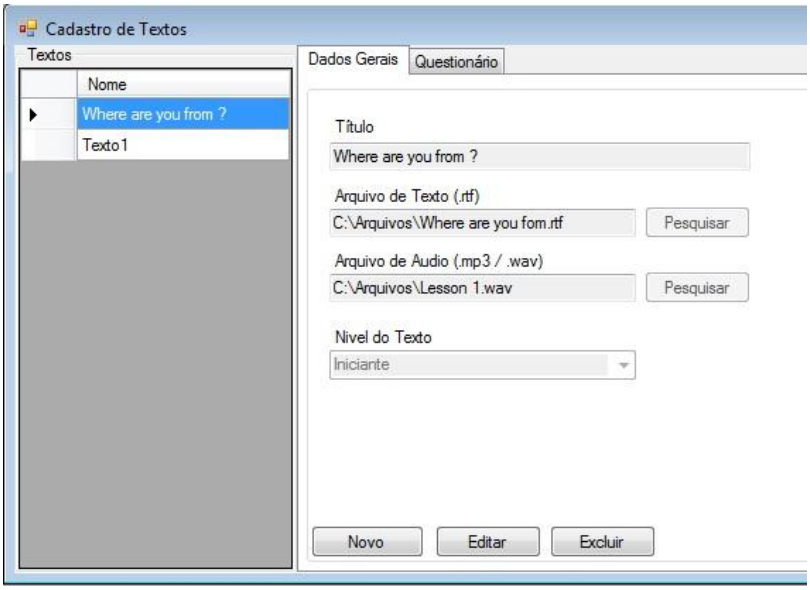

Figura 5. Cadastro de Textos, Áudios e Grau de dificuldade do Texto

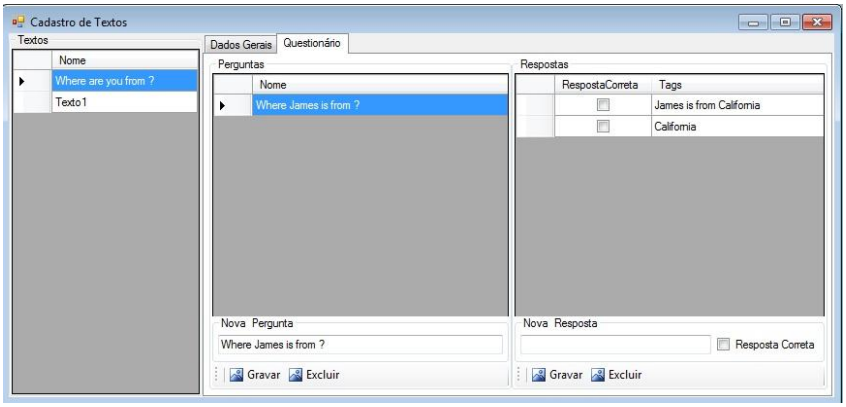

Figura 6. Cadastro do Questionário – Perguntas e Respostas

3.5. Testes e *Tuning*

Depois de desenvolvidos, cada um dos componentes da aplicação (isto é: código de diálogo e código de interação) são testados, isoladamente. Porém, nesta fase, o teste se refere à verificação de todo o sistema. Há uma grande quantidade de testes a serem realizados antes do lançamento do sistema, incluindo teste de aplicação, teste de reconhecimento e teste de avaliação de usabilidade. Todos esses testes são executados com todo o sistema funcionando.

Depois de completar a fase de testes, já se pode instalar o sistema no ambiente do usuário. Esse procedimento geralmente é feito em duas etapas: na primeira fase é instalado o piloto, em que o sistema é lançado para um número limitado de usuários (geralmente centenas). Coletam-se os dados e o sistema é melhorado a partir desses dados. Na maioria das vezes, são realizadas três iterações de coleta de dados. Após a fase de *tuning*, o sistema é finalmente lançado para todos os usuários finais.

Esta aplicação contou com os testes de aplicação, de reconhecimento e de avaliação da usabilidade, porém, o teste com o programa-piloto foi realizado somente entre os integrantes, de maneira informal.

4. Conclusões

O contexto atual de desenvolvimento e pesquisa de interfaces do usuário destaca o uso da voz como uma modalidade de muita relevância na interação entre o ser humano e os dispositivos digitais que cada vez mais fazem parte do nosso cotidiano. Este trabalho reafirma a importância deste tipo de interfaces ao propor uma aplicação baseada em voz para auxiliar o ensino/aprendizagem de língua estrangeira. Ao longo deste trabalho foram discutidas e apresentadas as etapas do desenvolvimento e as tecnologias utilizadas na criação da aplicação VAL.

Podem ser identificados como trabalhos futuros: (i) definição de critérios de usabilidade em aplicações de VUI para o ensino/aprendizagem de língua estrangeira, (ii) inclusão de recursos de colaboração e compartilhamento de informação, e (iii) combinação da modalidade voz com outras modalidades de interação, como gestos.

Referências

BOSCH, L., OOSTDJIK, N., RUITER, J. P. (2004), Turn-taking in social talk dialogues: temporal, formal, and functional aspects, *SPECOM-2004*, pp 454-461.

CHAMBERLAIN, J. ELLIOTT, G., KLEHR, M., BAUDE, J. (2006) *Speech User Interface Guide, RedPaper IBM*, http://www.redbooks.ibm.com/redpapers/pdfs/redp4106.pdf

COHEN, M. H., GIANGOLA, J. P., BALOGH, J. (2004) Voice User Interface Design, Addison Wesley, ISBN 0-321-18576-5, 368 pp.

DAMASCENO, Eduardo F.; PEREIRA, Tatiane V.; BREGA, José R.F. (2005) Implementação de Serviços de Voz em Ambientes Virtuais. In *INFOCOMP Journal of Computer Science*, v.4, n3, p.67-73.

DENG, L., HUANG, X. (2004). Challenges in adopting speech recognition, *Commun. ACM* 47, 1, pp. 69-75.

ESTABEL, Lizandra B.; MORO, Eliane L.; SANTAROSA, Lucila C. (2006) A inclusão social e digital de pessoas com limitação visual e o uso das tecnologias de informação e de comunicação na produção de páginas para a Internet. Ci. Inf., vol.35, n.1.

LAI, J. (2000) Conversational Interfaces, *Communications of the ACM*, Vol. 43, No 9, pp 24 – 27.

LAMEL, L.; MINKER, W.; PAROUBEK, P. (2000) "Towards Best Practice in the Development and Evaluation of Speech Recognition Components of a Spoken Language Dialog System", *Natural Language Engineering*, vol 6 (3-4), United Kingdom Cambridge University Press, pp. 305 - 322.

MCTEAR, M. F. (2002) Spoken Dialogue Technology: Enabling the Conversational User Interface, *ACM Computing Surveys*, Vol. 34, No. 1, pp. 90–169.

OLIVEIRA NETO, J. S. de; SALVADOR, V. F. M.; KAWAMOTO, A. L. (2010) "Aplicações interativas baseadas em voz na Educação: oportunidades e estudo de caso". In: Anita Maria da Rocha Fernandes; Michelle Silva Wangham. (Org.). Livro de Minicursos. Florianópolis , 2010, v. , p. 1-26.

SHNEIDERMAN, B. (2000) The Limits of Speech Recognition, *Communications of the ACM*, Vol. 43, No 9, pp 24 – 27.