Desenvolvimento de uma ferramenta para a produção de mídias utilizando personagem animado com síntese de voz

Rodrigo Lins Rodrigues[1], Alexandre Magno Andrade Maciel[1], Edson Costa de Barros Carvalho Filho[1]

[1]Vocal Lab Sistemas de Informação – www.vocallab.com.br

Caixa Postal 50.730-260 – Recife – PE – Brasil

{rodrigo.lins,alexandre,edson}@vocallab.com.br

Abstract. *This research shows the conception and design of a tool for automated production of video-classes, using the generation of animated characters and synthetic voice. For this, is demonstrated a method of conception used as a prototyping process and the solution development. At least, is showed the results of the user tests.*

Resumo. *Este trabalho apresenta a concepção e design de uma ferramenta de autoria para a produção automatizada de vídeo aulas, através da geração de personagens animados e voz sintética. Para tal o trabalho demonstra o método de concepção utilizado bem como o processo de prototipagem e o desenvolvimento da solução, por último é mostrado os resultados do teste com usuário.*

1. Introdução

A partir dos anos 90, é notável o crescente interesse pelo estabelecimento de relações sociais entre seres humanos e interfaces virtuais ou robóticas, e isto têm despertado o interesse da comunidade científica, que cada vez mais propõe novos sistemas especializados neste tipo de interação humano-computador (DUARTE & Costa, 2005).

Diversos grupos de pesquisa têm desenvolvido interfaces para personagens virtuais interativos que apresentem características sociais, de tal forma a permitir o estabelecimento de relacionamentos humano-personagem ou humano-máquina, destinados a atividades específicas, em diversas áreas do conhecimento humano.

No âmbito educacional, essas interfaces através de personagens animados interagem com o estudante em ambientes de aprendizagem baseada em computador, estimulando e encorajando o aprendizado. Esses personagens aumentam a efetividade das aplicações de educação e treinamento, e podem também ser empregados em muitas outras aplicações interativas e que auxiliem usuários (PROLA & VICCARI, 2003). Utilizam técnicas de entretenimento interativo, que tornam os estudantes mais interessados em aplicações com propósitos educacionais e em algumas aplicações que possam gerar apresentações mais interessantes e prazerosas.

Entre essas caraterísticas, pode-se mencionar o estabelecimento de expressões faciais em personagens virtuais utilizados para demonstrar determinados estados emocionais, reações de afeto e amizade, inclusão de uma personalidade virtual que se assemelha a personalidades humanas, assimilação de convenções sociais, interação com crianças e pessoas idosas, e até mesmo, demonstrações de carinho. Quando estes personagens são inseridos em um ambiente virtual de simulação, o aluno pode aprender e praticar habilidades no mundo virtual. Com estes personagens, o computador pode interagir com os alunos através de iniciativa mista, diálogo tutorial no papel de professor ou companheiro aprendiz. Ele pode se comunicar de forma verbal e não verbal.

Normalmente a construção destes personagens em softwares educacionais é feita com ferramentas que exigem um conhecimento específico de artes, modelagem e animação. O processo de inserção da voz geralmente é feito de forma artesanal, ou seja, são criadas animações em diferentes contextos de interação, e posteriormente é gravado a voz de um locutor que ao final é inserida na animação desenvolvida. Esse processo, na maioria das vezes, demanda de um alto investimento e alta carga de trabalho, pois necessita de locutores, estúdios de gravação de voz, artistas, animadores e roteiristas.

Tendo em vista essa problemática, o objetivo deste trabalho é conceber uma ferramenta de fácil utilização para a produção de vídeo aulas utilizando personagens animados e voz sintética.

2. Revisão bibliográfica

Nesta seção é apresentada a revisão bibliográfica necessária para o desenvolvimento desta pesquisa, bem como, os conceitos fundamentais para se entender os processos e tecnologias envolvidos na concepção do produto proposto, dentre esses conceitos temos o entendimento do processo de construção de materiais para EAD, utilização da tecnologia de síntese de voz na produção de material educacional e por fim a apresentação de técnicas de animação em personagens virtuais.

2.1. Produção de material didático em Educação à Distância

O material didático assume na Educação à Distância o papel de instrumento para o diálogo permanente entre alunos, professores e o conhecimento. É evidente que o material didático precisa estar bem alinhado à proposta pedagógica do curso.

No momento do planejamento é importante levar em conta o tipo de material didático que se deseja produzir. Essa diferenciação leva em conta o meio de veiculação e consequentemente a forma de interação dos materiais com os alunos. Ruiz e Cordero (1998) referem-se a preocupações que devem existir na elaboração de material didático para a EAD, considerando principalmente as definições das formas de comunicação e estratégias da narrativa a serem aplicadas aos diálogos, assim como a linguagem audiovisual e as ferramentas auxiliares utilizadas para o processo de ensino-aprendizagem.

De acordo com (ABREU, 2010) O processo de produção clássico para a produção desses materiais, utilizado normalmente por instituições acadêmicas e

corporativas se constitui em quatro fases: planejamento, produção, Implementação e Avaliação figura 1.

Figura 1. Processo macro de construção de material em EAD

Este processo por sua vez é subdividido e gerado subprocessos que compõem etapas menores, como mostra a Figura 2.

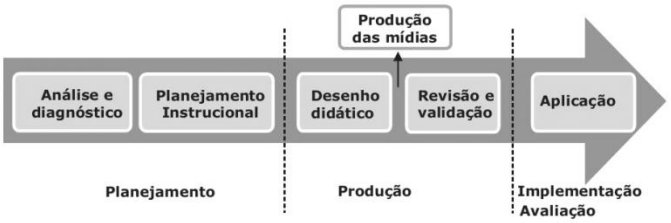

Figura 2. Subprocessos de construção de material em EAD

O caminho percorrido para a realização do processo, em sua amplitude, inclui seis etapas: análise e diagnóstico, planejamento instrucional, desenho didático, produção das mídias (esta etapa é de responsabilidade da equipe de design, TI, vídeo e áudio), validação e revisão e aplicação.

A **produção das mídias** é de responsabilidade da equipe de design, tecnologia da informação, programação, vídeo e áudio. Para cada mídia é necessário que a equipe seja composta por um grupo de profissionais especialistas na produção da mesma. Essa etapa é uma das mais trabalhosas no processo, pois é necessário ter profissionais de diversas áreas do conhecimento interagindo para a confecção de um produto. Normalmente essa fase exige uma demanda de tempo para a integração do grupo, alto custo financeiro e elevado tempo de produção das mídias.

Basicamente os artefatos dessa etapa mais críticos são a modelagem e animação do personagem e a construção do diálogo através da voz natural. A construção destes dois artefatos pode ser automatizada utilizando técnicas pré-moldadas e a geração de voz sintética em tempo real, desta forma é possível minimizar tempo e custo no processo de produção de materiais pra EAD especificamente na etapa de produção de mídias.

2.2. Utilização de síntese de voz em artefatos educacionais

Síntese de Fala pode ser definida como a utilização de mecanismos artificiais para a produção de um sinal de fala. Pode também ser definida como um processo de geração automática de formas de onda de voz que é projetado para responder a um pedido de informação utilizando mensagens faladas (SIMÕES, 1999).

Os sistemas de conversão texto-fala (no inglês: *Text-To-Speech* - TTS) são capazes de gerar fala sintetizada a partir de uma mensagem escrita. A utilização desse tipo de sistema é extremamente abrangente, pois, em princípio, qualquer tipo de

mensagem pode ser representada na firma textual e, portanto, qualquer tipo de mensagem pode ser sintetizada. A qualidade do sinal de voz sintetizado por um sistema de conversão texto-fala geralmente é inferior àquela gerada por meio de amostras pré-gravadas. Uma das razões disso é o fato de que nem sempre o módulo de processamento linguístico é capaz de fornecer transcrição fonética correta de todas as palavras do texto.

O processo de síntese de um sistema TTS é composto de duas fases principais. A primeira consiste no Processamento de Linguagem Natural (PLN), onde a mensagem de entrada é transcrita em uma representação de nível fonético e a segunda consiste na Geração da Fala em que as formas de onda de voz são geradas e a saída acústica é produzida. Essas duas fases são respectivamente chamadas de síntese de alto-nível e síntese de baixo-nível (MAIA, 2006). Uma versão simplificada do processo é apresentada na Figura 3.

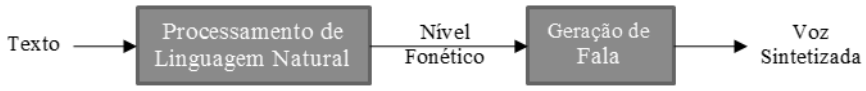

Figura 3. Descrição do processo de síntese de fala

O Processamento de Linguagem Natural é responsável por traduzir o texto de entrada de uma dada linguagem em uma representação fonética que inclui informações a respeito das unidades acústicas a serem produzidas (por exemplo: fonemas, sílabas) juntamente com características textuais (por exemplo: tonicidade, indicadores de fronteiras das frases/palavras).

A geração de fala é responsável por receber as informações da locução e gerar voz. Considerando o caso geral, o processo de geração de fala pode ser dividido em duas subfases. De acordo com essa divisão, a primeira subfase é responsável por processar as informações da locução e gerar um conjunto especifico de parâmetros necessários para a geração do sinal de fala - a segunda subfase (MACIELl, 2012).

Algumas iniciativas de desenvolvimento de software educacionais utilizando esta tecnologia foram desenvolvidas, dentre elas as de mais destaques são as ferramentas de leitura de tela, como por exemplo a ferramenta Jaws[1]. Um leitor de tela para deficientes visuais. Grande parte das ferramentas educacionais que utilizam este tipo tecnologia de síntese e reconhecimento de voz são focadas na construção de artefatos para tecnologias assistivas, ou seja, tecnologias que contribuem para proporcionar ou ampliar habilidades funcionais de pessoas com deficiência e consequentemente promover Vida Independente **e** Inclusão.

2.3. Animação facial através da construção de visemas

A partir de esforços multidisciplinares de pesquisa e desenvolvimento nas áreas de reconhecimento de voz, processamento da linguagem natural, inteligência artificial, síntese da fala, computação gráfica e animação, é possível implementar personagens

[1] http://www.freedomscientific.com/products/fs/jaws-product-page.asp

virtuais capazes de capturar mais facilmente a atenção do usuário e tornar a atividade de interação mais atrativa e envolvente (GRATCH, 2009).

O desenvolvimento *detalking heads* leva em consideração o papel de destaque que a face ocupa na comunicação humana. Desde o nascimento somos treinados nos mecanismos de comunicação face a face e, estimulados por experiências sociais, nos tornamos capazes de interpretar e identificar estados emocionais transmitidos pela face, utilizando sua informação visual para complementar a compreensão da mensagem transportada pelo sinal acústico da fala.

A animação facial por computador sincronizada com a fala permite a implementação de cabeças virtuais que podem contribuir para tornar interfaces humano-computador mais eficientes e atraentes.

Neste sentido, um dos objetivos da animação facial gerada por computador é conferir a uma face virtual a aparência, a movimentação e o comportamento de uma face real. Esta capacidade pode ser qualitativamente expressa em termos do grau de vídeo-realismo alcançado pela animação, ou seja, sua capacidade de ser confundida com o vídeo de uma face real. Assim, uma animação facial vídeo-realista, além da reprodução fotográfica das características estáticas da face (como rugas e textura da pele), é também capaz de reproduzir os movimentos articulatórios da fala em sincronia e harmonia com a locução.

A reprodução realista dos movimentos articulatórios da fala é obtida levando-se em consideração os mecanismos de produção da mesma. A realização acústica dos diversos fonemas de uma língua se dá através de configurações típicas do trato vocal que, entre outros elementos articuladores, inclui as cordas vocais, o palato, a cavidade nasal, a língua e os lábios. No entanto, apenas uma parcela dos movimentos realizados pelos órgãos articuladores é visualizada na face através, principalmente, da movimentação dos lábios e da região em torno deles. Assim, a modelagem dos movimentos articulatórios faciais visíveis pode ser realizada através de visemas.

No desenvolvimento deste trabalho, utilizamos a definição de (COSTA, 2009), que define um visema como sendo uma postura labial estática que é visualmente contrastiva a outra e que pode ser associada à realização acústica de um fonema.

3. Processo de concepção e desenvolvimento da ferramenta proposta

O método de design utilizado neste trabalho se baseou na literatura de Interação Humano-computador, especificamente no método de design da interação. Essencialmente, o processo foi composto por quatro atividades básicas: (1) *identificação das necessidades do usuário*; (2) *desenvolvimento de alternativas de design*; (3) *construção de versões iterativas* e (4) *avaliação do design*.

3.1. Identificação das necessidades e concepção da ferramenta

A fase de identificação das necessidades do usuário envolveu uma pesquisa da situação atual para identificar necessidades e oportunidades de concepção, a fim de determinar as características do produto de design. O principal objetivo desta fase foi à

identificação das necessidades do usuário e o levantamento de requisitos. Para tal utilizamos da técnica de construção de cenários caricaturados que serviram para criar uma situação de uso da ferramenta a ser concebida.

- **Prototipagem de baixa fidelidade**

O processo de prototipagem foi utilizado para antecipar ao usuário final características da interface que puderam ser testadas, validadas e modificadas pelos *stakeholders*. Neste processo, foi criada uma interface semelhante à interface final, partindo dos requisitos iniciais figura 4.

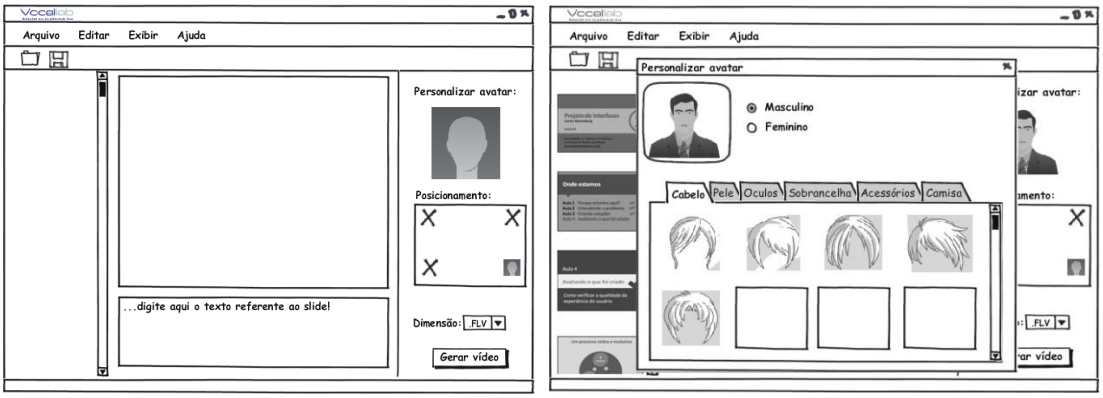

Figura 4. Protótipo de baixa fidelidade

Ao finalizar a prototipagem de baixa fidelidade, foi feito os primeiros testes com usuários afim de identificar possíveis problemas na interface, assim como possíveis problemas técnicos em relação a implementação. Nesta etapa foi gerado algumas modificações que foram implementadas na versão interativa.

3.2. Desenvolvimento da versão interativa

A partir da especificação feita e construção de protótipos que demonstraram as alternativas de design, a terceira fase do processo de desenvolvimento foi à construção e implementação do protótipo em uma versão interativa, que pôde ser testada pelos usuários, onde os mesmos tiveram a possibilidade de reagirem ao design e sugerirem mudanças. O software foi desenvolvido utilizando a linguagem java através da plataforma J2SE Figura 5.

Figura 5. Versão interativa da ferramenta

Foi identificado, através de entrevistas com o especialista em interfaces que a escolha entre apenas dois personagens deixaria o software limitado em termos de afetividade, neste caso na primeira tela tivemos a modificação do item de tela *RadioButton* para um botão que levaria a ativação de uma nova tela de escolha de personagens.

O protótipo foi modificado e disponibilizado personagens com caraterísticas físicas diferenciadas, permitindo ao usuário escolher entre os diversos tipos de personagens com aparências pré-definidas figura 6.

Figura 6. Localização do personagem

A tela de localização do personagem permaneceu inalterada em relação a prototipagem inicial, nesta tela é possível o usuário escolher entre: superior esquerdo, superior direito, inferior esquerdo e inferior direito.

- **Bibliotecas de integração utilizadas no desenvolvimento**

A arquitetura do ambiente foi construída utilizando três API para o desenvolvimento, a primeira utilizada foi a *Apache POI*[2], essa API teve a função de implementar a extração de caraterísticas do slide, tais como as notas inseridas pelo usuário e a transformação de slides em imagens. A segunda foi o VLSynthesizer[3], uma biblioteca responsável pela transformação das notas extraídas nos slides em voz sintética. A terceira biblioteca utilizada foi a FFMPEG[4] responsável por transformar as imagens extraídas e a voz sintetizada em vídeo, a biblioteca possibilitou a exportação em diversos formatos de vídeo, tais como avi, mpeg, flv, swf.

- **Técnica de animação por visemas**

A técnica de animação que serviu de base para este trabalho foi desenvolvida por Costa (2009), onde apresentou um método de síntese de animação facial 2D baseado em imagens cujo desenvolvimento foi guiado por dois objetivos principais: a reprodução realista da movimentação articulatória visível da fala, incluindo os efeitos da

[2] http://poi.apache.org/

[3] Motor responsável pela síntese de voz http://www.vocallab.com.br

[4] FFmpeg is a complete, cross-platform solution to record, convert and stream audio and video http:// http://ffmpeg.org/

coarticulação. A técnica desenvolvida baseia-se em uma base de imagens de visemas dependentes de contexto para o Português do Brasil e adota a técnica de metamorfose entre visemas para a síntese da animação facial. A abordagem representa uma estratégia de síntese capaz de reproduzir a movimentação articulatória visível da fala, incluindo os efeitos da coarticulação, a partir de uma base de 34 imagens intituladas como visemas como mostra a figura 7 (COSTA, 2009).

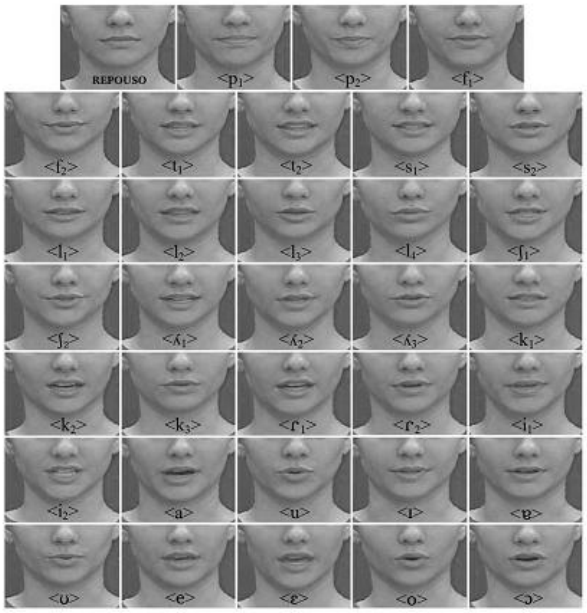

Figura 7. Representação dos visemas

A síntese da animação foi implementada tendo-se como parâmetro de entrada a transcrição fonética temporizada da fala a ser visualmente animada. A partir das informações fornecidas pela transcrição fonética. A animação foi sintetizada através do apropriado sequenciamento, concatenação e apresentação de quadros resultantes do processamento de imagens da base.

Essa sequencia de visemas foi estabelecida de acordo com a *API VLSynthesizer* que transforma texto em voz e nos fornece uma *string* de fonemas de acordo com o áudio gerado, a partir daí é possível obter a imagem referente ao *visema* e fazer o sequenciamento permitindo a animação e a gesticulação labial do personagem.

Foi necessário a criação de uma tabela de conversão entre a nomenclatura proposta por Costa (2009) e a nomenclatura utilizada para a implementação neste trabalho, tendo em vista que existia a presença algarismos arábicos que tiveram de ser convertidos em nomenclaturas entendidas pela linguagem de programação utilizada.

3.3. Processo de avaliação

O processo de avaliação foi dividido em duas etapas. A primeira etapa foi avaliação de usabilidade da ferramenta desenvolvida, nesta etapa foi estabelecida duas tarefas para o usuário a fim de identificar possíveis problemas no fluxo de interação da atividade. A

segunda avaliação foi relacionada ao nível de satisfação do usuário em relação ao produto final gerado pela ferramenta.

- **Avaliação de usabilidade do design**

Um pequeno estudo de usabilidade realizado com quatro professores de cursos de graduação a distância veio confirmar a eficácia das interfaces desenvolvidas neste trabalho. O teste teve como objetivo encontrar erros em duas tarefas específicas: (1) Inserir um slide na ferramenta e (2) Exportar uma nova vídeo aula.

	Solicitou ajuda	**Tempo da tarefa**	**Erros**	**Cliques**
Usu.1	3	8,4 min	1	8
Usu.2	3	5 min	2	13
Usu.3	2	7,3 min	1	11
Usu.4	2	4,8 min	0	10

Tabela 1: Teste de usabilidade

A tabela 1 mostra os resultados das duas tarefas em sequencia, podemos observar que em média o entrevistador foi solicitado 2,5 vezes, dentre essas solicitações todas foram relativas a dúvidas do tipo de arquivo que a ferramenta suportaria. Em relação ao tempo de conclusão ta tarefa pudemos identificar que o tempo médio para a construção de uma vídeo aula, utilizando a ferramenta desenvolvida, é em média de 6,4 min. Se comparado ao processo tradicional de construção de vídeo aulas, podemos considerar que esse tempo relativamente baixo, pois os processo de construção da voz e animação do personagem é automatizado pela ferramenta. A quantidade de erros cometidas durante o teste foi em média 1 erro por tarefa executada e de 10,5 cliques para a conclusão da mesma.

- **Avaliação de satisfação do usuário**

Nesta segunda etapa do teste, o objetivo foi avaliar a satisfação do usuário. Foi feito uma entrevista estruturada onde os usuários responderam perguntas referentes ao potencial da ferramenta e a satisfação em relação ao vídeo gerado.

	Personagem	**Qualidade da voz**	**Animação**	**Sincronia**
Usu1	Ótima	Aceitável	Boa	Boa
Usu2	Boa	Aceitável	Boa	Boa
Usu3	Ótima	Insatisfeito	Aceitável	Boa
Usu4	Ótima	Aceitável	Ótima	Ótima

Tabela 2: Teste de satisfação

De acordo com o teste de satisfação, um dos principais problemas encontrados foi a qualidade da voz, os quatro usuários entrevistados, no primeiro momento, tiveram uma expectativa abaixo da esperada. Em aos personagens, todos os entrevistados se disseram satisfeitos, pois consideraram a diversidade de estilos físicos um fator preponderante na personalização dos vídeos gerados. Todos consideraram boa a sincronia e a animação facial dos personagens no momento da fala.

4. Considerações finais

Através dos resultados desta pesquisa inicial, pudemos identificar que a solução concebida tem potencial para ser desenvolvida em maior escala, tendo em vista que os testes com os usuários foram satisfatórios do ponto de vista da interface e interação, e dentro da perspectiva de satisfação do usuário. Como trabalhos futuros pretende-se abordar as melhorias nos problemas identificados, tais como o problema da qualidade da voz sintética, tendo em vista que grande parte dos usuários relataram um estranhamento da voz sintética se comparada a voz natural humana, no entanto todos afirmaram entender perfeitamente as palavras e frases ouvidas no decorrer das vídeo aulas.

Referências

ABREU, D. (2010). *Produção de Material Didático para EaD.* Paraná: Editora UFPR.

COSTA, P. D. (2009). *Animação facial 2D sincronizada com a fala baseada em imagens de visemas dependentes do contexto fonético.* Tese de mestrado - UNICAMP.

DUARTE, G. D., & Costa, A. C. (2005). Uma Proposta para Modelagem de Personagens Virtuais Emotivos Utilizáveis em Ambientes de Educação a Distância. *XXV CSBC*.

GRATCH, J. (2009). Creatinginteractivevirtualhumans:some assemblyrequired. *IEEE*.

MACIELl, A. M. (2012). *Investigação de um ambiente para o desenvolvimento integrado de interface de voz.* Recife: Tese de doutorado.

MAIA, R. S. (2006). Speech Synthesis and Phonetic Vocoding for Brazilian Portuguese based on Parameter Generation from Hidden Markov Models. Tese (Doutorado em Engenharia). Nagoya Institute of Technology.

PROLA, M. T., & VICCARI, R. M. (2003). *Modelagem de um Agente Pedagógico Animado para um Ambente Colaborativo: considerando fatores sociais relevantes.* Porto Alegre.

RUIZ, T. B., & CORDERO, J. M. (1998). *Guia para el diseño, elaboración y evaluación de material escrito.* Brasília/ Madrid: UnB/ Uned: Apostila para o Curso de Especialização em Educação Continuada e à Distância da Faculdade de Educação da Universidade de Brasília.

SIMÕES, F. O. (1999). *Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil.* Campinas: Dissertação (Mestrado em Engenharia Elétrica). Universidade Estadual de Campinas.