Tradutor Texto-Voz baseado em autômatos adaptativos

D. P. Shibata, F. R. Koike

***Resumo* — Este trabalho apresenta um tradutor texto-voz desenvolvido para a língua portuguesa baseado na tecnologia de autômatos adaptativos [1]. O tradutor é constituído por dois módulos de software desenvolvidos em paralelo e que integrados realizam o processo de síntese de voz a partir de um texto escrito na língua portuguesa. A primeira parte traduz o texto para uma seqüência fonética utilizando o Alfabeto Fonético Internacional desenvolvido pela Associação Fonética Internacional [2], enquanto a segunda parte recebe essa seqüência fonética e sintetiza a voz a partir de um processo de síntese por concatenação. Devido às diversas variações encontradas para a língua portuguesa, a língua falada na cidade de São Paulo foi escolhida como base para a saída gerada pelo tradutor.**

***Palavras chave* — Alfabeto Fonético Internacional, Autômatos Adaptativos, Tradutor Grafema-Fonema, Tradutor Texto-Voz.**

I. Introdução

ESTE trabalho apresenta um software desenvolvido para realizar a tradução texto-voz de textos escritos na língua portuguesa, baseado em autômatos adaptativos. O software apresentado se encontra dividido em dois módulos desenvolvidos separadamente. O primeiro módulo é responsável pela tradução grafema-fonema dos textos, recebendo como entrada um texto escrito na língua portuguesa e gerando como saída uma seqüência de fonemas que represente o texto de entrada. O segundo módulo sintetiza a voz a partir do método de síntese por concatenação.

Há dois fatores principais que motivam o desenvolvimento deste trabalho, sendo o primeiro o de aplicar os autômatos adaptativos para resolver alguns problemas relacionados à sensibilidade a contexto inerentes às linguagens naturais, utilizando assim todo o seu poder computacional, e o segundo, o caráter social do trabalho que pode ser utilizado para facilitar a comunicação de pessoas com deficiências que afetem o processo de comunicação por meio da fala ou até mesmo pessoas interessadas em utilizar o software para realizar a leitura de arquivos de texto enquanto realizam outras tarefas.

O trabalho inicia-se pela descrição de características de implementação do software, seguida das descrições do tradutor grafema-fonema baseado em Autômatos Adaptativos e do sintetizador de voz, apresentando de forma simplificada as técnicas utilizadas nesses processos. A seguir, encontra-se a discussão sobre os resultados obtidos pelo tradutor e pelo sintetizador e alguns dos problemas encontrados no processo de desenvolvimento dos dois softwares. Por fim, há a conclusão com discussão de trabalhos futuros e solução para possíveis problemas.

II. Características de Implementação

Tanto o tradutor grafema-fonema como o sintetizador de voz foram desenvolvidos utilizando a linguagem Java. Os softwares foram criados como APIs, e até o momento não há interface gráfica associada às bibliotecas.

No entanto, o tradutor pode ser executado por meio de linhas de comando, passando como parâmetro o arquivo com o caminho para o arquivo em que se encontra o texto a ser traduzido, o tipo de saída indicando se deve executar o som ou se deve gerar um arquivo de áudio no padrão WAV e, nos casos em que é gerado um arquivo de áudio, um caminho para esse arquivo.

O software ainda se encontra em versões iniciais de teste sem data definida para início de distribuição. Por essa razão, ainda não foi definida uma licença para distribuição do mesmo.

III. Tradutor Grafema-Fonema

Esta é a parte do software responsável pela recepção dos arquivos de entrada, escritos na língua portuguesa, e que gera como saída seqüências de símbolos escritas no Alfabeto Fonético Internacional, utilizada posteriormente pelo sintetizador de voz.

O software de tradução grafema-fonema foi baseado no modelo de autômato adaptativo apresentado em [3]. Este modelo foi utilizado como base para criar um tradutor grafema-fonema para palavras da língua portuguesa, e esse tradutor de palavras foi utilizado como base para traduzir textos da língua portuguesa.

Esse tradutor de palavras foi associado a um bloco de controle, responsável pela separação do texto em seqüências de palavras (incluindo seqüências de caracteres como números, datas, etc) e que fornece cada uma das palavras ao autômato para que possa ser executada gerando a saída fonética referente à palavra. A Figura 1 ilustra o tradutor texto-voz, enfatizando a separação dos blocos internos ao tradutor grafema-fonema.

D. P. Shibata (danilo.shibata@gmail.com).
F. R. Koike (fabio.koike@gmail.com).

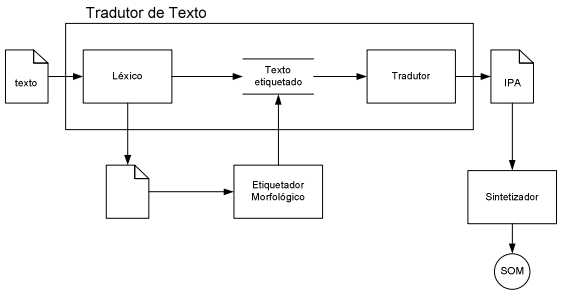

Fig. 1 – Ilustração do Tradutor Texto-Voz focando a tradução grafema-fonema.

Como pode ser visto pela Figura 1, o tradutor grafema-fonema possui um bloco inicial que funciona como analisador léxico e etiquetador, descobrindo as partes mínimas do texto que devem ser analisadas e transformadas em seqüências fonéticas e etiquetando cada uma dessas palavras.

Esse analisador léxico pode utilizar o auxílio de um etiquetador morfológico externo, gerando etiquetas que determinam as características morfológicas das palavras que são posteriormente utilizadas para definir qual o som mais provável para uma palavra considerando essa característica. O uso dessa funcionalidade não é obrigatório e o tradutor grafema-fonema é capaz de gerar uma saída mesmo sem o etiquetador morfológico ou em caso de falhas do mesmo. Nesse trabalho foi utilizado o etiquetador morfológico desenvolvido em [4].

As palavras são então submetidas ao autômato, e se o autômato obtiver mais de uma seqüência fonética para representar aquela palavra, uma delas será escolhida em função da etiqueta morfológica (se houver) ou qualquer uma delas será escolhida se não houver etiquetas.

A. *Tradutor de Palavras*

O tradutor grafema-fonema de palavras utiliza-se do modelo de autômato adaptativo definido em [3]. Esse módulo recebe como entrada uma palavra da língua portuguesa e gera como saída um conjunto de seqüências fonéticas que representa essa palavra.

Inicialmente, a palavra é submetida ao analisador léxico que gera uma seqüência de átomos semelhantes às sílabas definidas pelas regras de separação da língua portuguesa e coloca um símbolo indicando o final da cadeia após a seqüência. Essa seqüência de átomos é utilizada como entrada para o autômato que a traduz para o conjunto adequado de saídas.

O autômato adaptativo é composto por duas sub-máquinas chamadas de *Reconhecedor* e *Tradutor*. O *Reconhecedor* lê os átomos que representam sons e realiza chamadas de função adaptativa que alteram o *Tradutor*, e ao consumir o átomo que indica o final da cadeia executa uma chamada para o *Tradutor*. O *Tradutor* executa com a cadeia vazia e a partir de escrita e leitura de símbolos realiza a transdução da seqüência de átomos para as seqüências fonéticas que a representam.

As seqüências de chamada realizadas pelo *Reconhecedor* criam, no *Tradutor*, blocos de transições com a informação contextual necessária para definir o som adequado para um determinado átomo. Desta forma, toda a informação contextual sobre os átomos são escritas na cadeia e consumidas no devido momento, gerando a saída correta em função do contexto definido.

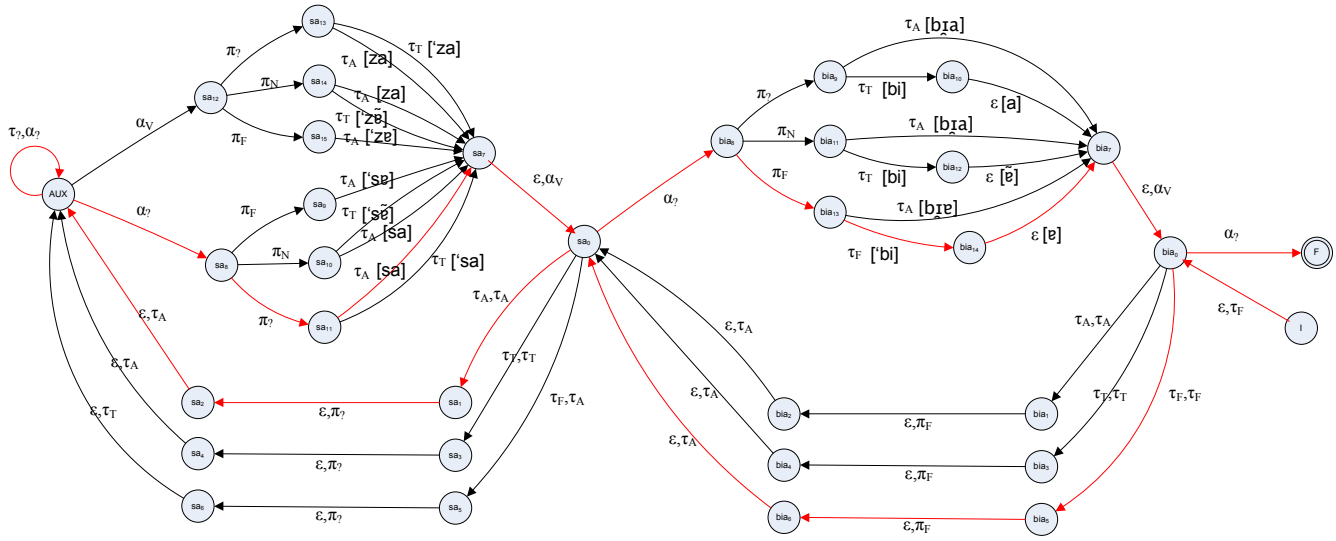

Fig. 2 – Transdutor alterado para representar a palavra "sabia".

A Figura 2 apresenta um exemplo de configuração do *Trasdutor*, quando alterado para representar a palavra "sabia". A palavra é dividida em dois átomos "sa" e "bia". O primeiro átomo é traduzido para o som /sa/, o som é átono pois a tônica se encontra no outro átomo, a letra 's' recebe o som /s/ por iniciar a palavra e o som a não sobre alterações por influências posteriores. O segundo átomo é separado em dois sons /'bi/ e /ɐ/ por ser final de uma palavra não acentuada. A letra 'a' recebe o som desvozeado por ser final, a letra 'b' não é alterada por influência anterior e o som de "bi" é tônico por ser anterior a uma letra 'a' final.

O processo é disparado pela transição que sai do estado "I" escrevendo o símbolo τ_F, que indica que o átomo que o lê é final em uma palavra não acentuada. Daí para frente, segue-se o processo de execução pelas transições marcadas em vermelho na figura. A partir da leitura de um símbolo define-se a ação a ser tomada (escrever novos símbolos ou escrever uma seqüência fonética na saída). As regras contextuais de cada átomo, que definem as ações realizadas a partir da leitura de um símbolo, são definidas pelas chamadas de funções executadas no *Reconhecedor*. O significado dos símbolos α, π e τ são definidos em [3].

IV. Sintetizador de Voz

O sintetizador de voz se encontra dividido em três partes, o Analisador Léxico, o Gerenciador e a Camada de Síntese de Voz. O Analisador Léxico e a Camada de Síntese realizam as tarefas necessárias ao processo de síntese, enquanto o módulo Gerenciador é responsável pela interface entre os dois outros módulos.

No Analisador Léxico identificam-se todos os *tokens* representados pelas sílabas. Adotou-se a sílaba como unidade de concatenação, por proporcionar efeito de coarticulação menores que a concatenação de fonemas [6]. Cada sílaba foi mapeada para sua respectiva pronúncia através da pré-gravação do áudio com a fala natural e cada arquivo de áudio teve que ser normalizado para possuir mesmas características sonoras com freqüência de 44.100 Hz, 16 bits, estéreo e com amplitude máxima de 1 db. Cada Token reconhecido é então armazenado numa estrutura de fila que é retornada para o gerenciador. No Analisador Léxico também são reconhecidas a tonicidade e as pontuações gramaticais, que orientam o processo de definição de entonação da voz sintetizada.

O gerenciador, de posse das unidades fonéticas reconhecidas, repassa-as para a camada de síntese de voz. A camada de síntese fica responsável pela recuperação dos respectivos arquivos de áudio de cada unidade, decodificando os mesmos de MP3 para WAV e extraindo os seus streams de áudio. Caso a unidade fonética seja tônica é aplicado um filtro de amplificação no stream para se conseguir o efeito desejado, esse processo se repete até finalizar a concatenação de todas as unidades fonéticas de uma sentença. Ao final de cada sentença verifica-se se existe alguma pontuação gramatical como interrogação, por exemplo, e aplica-se um novo filtro, agora em todos os streams já concatenados para se obter o efeito de entonação.

Como resultado deste processo, dependendo do comando determinado pelo módulo gerenciador, obtém-se um arquivo de áudio com a fala sintetizada para o texto de entrada ou reproduz-se o áudio. A Fig. 2 ilustra o processo de tradução texto-voz enfocando o processo de síntese a partir da representação fonética.

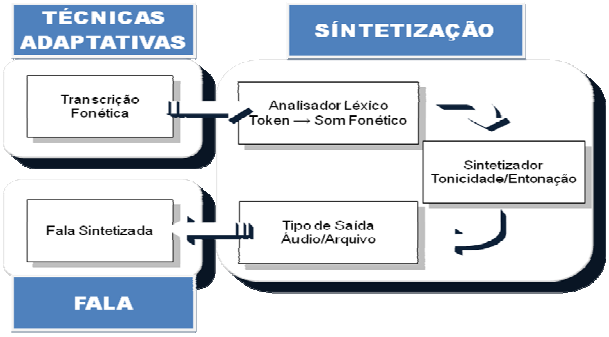

Fig. 3 – Ilustração do Tradutor Texto-Voz focando o processo de síntese.

V. Resultados Obtidos

Apesar de utilizar métodos simples para a tradução grafema-fonema e para o processo de síntese de voz a partir da linguagem fonética, o tradutor texto-voz apresentado obteve resultados satisfatórios e inteligíveis para pessoas com bom conhecimento da língua portuguesa.

Os maiores problemas em relação à tradução de texto para voz se encontram na definição do som da letra x que pode ter quatro sons diferentes (máximo, lixo, exato e fluxo), e das letras 'e' e 'o', quando estas duas últimas se encontram em uma sílaba tônica, sendo que alguns dos problemas com as vogais citadas foram resolvidas a partir do auxílio do etiquetador morfológico.

Eventualmente, a letra 'a' de uma sílaba átona que antecede uma letra nasal (o primeiro 'a' de "caminha" pode ter o som aberto ou fechado) ou uma junção de vogais anterior à sílaba tônica ("ia" forma ditongo em "financiamento", mas não em "adiamento") podem gerar problemas, mas são menos freqüentes e mais difíceis de resolver pois a escolha da forma a ser utilizada é opcional.

Em testes iniciais, foi obtida uma amostra de textos etiquetados com aproximadamente 9000 elementos (formados pelo par palavra e etiqueta). Desses 9000 elementos, aproximadamente 7000 tiveram tradução correta, 1600 geraram dúvidas por existência de 'e', 'o' ou 'x' e 400 falhas relacionadas à existência de 'a' pré nasal ou ditongo pré-tônico.

Para os casos em que houve dúvida, a escolha da saída mais provável gerou uma porcentagem de acerto um pouco superior a 50%, enquanto o uso do etiquetador levou a uma taxa de acertos próxima de 70%. Somando os resultados, o tradutor grafema-fonema obteve aproximadamente 87% de acerto sem o etiquetador e 90% com ajuda desse último.

Das traduções falhas, apenas 8 tiveram uma alteração grave que dificulta significativamente o entendimento (por exemplo, a palavra "porque" que é tônica final ou "anti" que não é, ao contrário do que definem as regras da língua porguesa), sendo

todas as outras falhas que não prejudicam significativamente o entendimento do usuário.

Os resultados obtidos apresentam uma significativa melhora em relação aos que foram obtidos anteriormente no trabalho de [5], que não separa os processos de tradução grafema-fonema, utiliza um autômato mais simples e não realiza processos de tratamento na saída gerada.

Apesar dessas melhoras, acredita-se que seja possível melhorar ainda mais os resultados a partir de um estudo mais aprofundado em outras técnicas de síntese e de melhorias pontuais no autômato existente.

VI. CONCLUSÕES

Este trabalho descreve um software para tradução texto-voz desenvolvido com base no modelo de autômatos apresentado em [2]. O software é dividido em duas partes, um tradutor grafema-fonema baseado no mesmo modelo e um sintetizador de voz baseado no processo de síntese por concatenação.

O software se encontra em estado inicial no processo de desenvolvimento e pode ser melhorado, com maior integração entre os módulos e com melhorias pontuais nos módulos de tradução grafema-fonema e de síntese de voz.

O software comprova a viabilidade do uso de autômatos adaptativos como base para a tradução texto-voz de textos escritos na língua portuguesa, e pode ser um incentivo para a utilização dessa mesma tecnologia em outras aplicações relacionadas ao processamento de linguagens naturais.

REFERÊNCIAS

[1] J. José Neto, “Adaptive Automata for Context-Sensitive Languages”, SIGPLAN NOTICES, Vol. 29, n. 9, pp. 115-124, September, 1994.

[2] International Phonetics Association. http://www.arts.glas.ac.uk/IPA/

[3] D. P. Shibata, R. L. A. Rocha, “An Adaptive Automata based method to improve the output of text-to-speech translators”, The Sixth Congress of Logic Applied to Technology, 2007.

[4] F. N. Kepler, “Um etiquetador morfo-sintático baseado em Cadeias de Markov de tamanho variável”, dissertação de mestrado, orientada por Marcelo Finger, Instituto de Matemática e Estatística, Univesidade de São Paulo,. 2005.

[5] D. A. Alfenas, A. J. B. Castro, D. P. Shibata, H. T. Soejima, “Sintetizador Texto-Voz com Autômatos Adaptativos”, trabalho de conclusão de curso, orientado por R. L. A. Rocha, Escola Politécnica, Universidade de São Paulo, 2004.

[6] F. R. Koike, L. R. E. Oliveira, R. H. S. Santos, T. M. Rizzo, “Síntese de Fala com apoio de Técnicas Adaptativas”, trabalho de conclusão de curso, orientado por R. L. A. Rocha, Faculdade Engenheiro Celso Daniel, Centro Universitário Fundação Santo André. 2007.

Danilo Picagli Shibata nasceu em São Paulo, Brasil em 22 de dezembro de 1981. Formou-se Engenheiro de Computação pela Escola Politécnica da Universidade de São Paulo em 2004 e cursa Mestrado em Engenharia de Computação pela mesma.

Exerceu profissionalmente pela Fundação para o Desenvolvimento Tecnológico da Engenharia (FDTE), na área de análise de confiabilidade e segurança de sistemas metro-ferroviários. Tem como áreas de interesse, Computação Formal e Processamento de Linguagens Naturais.

Fábio Robson Koike nasceu na cidade de Lins interior do estado de São Paulo, Brasil em 24 de novembro de 1972. Formou-se em Técnico em Processamento de dados em 1990 pelo Colégio Objetivo, Tecnólogo em Tecnologia da Computação em 2005 e Engenheiro da Computação em 2007 ambos no Centro Universitário Fundação Santo André - Faculdade Eng° Celso Daniel.

Atua na área de TI desde 1990 ocupando várias responsabilidades desde Estagiário em Processamento de Dados, passando por Programador Jr., Analista Sênior até Coordenador de Projetos e Arquiteto de Soluções em TI na atual ocupação. Tem experiência com projetos de Automação Industrial, Logística, Computação Embarcada, Aplicações Corporativas entre outros. Além de experiência profissional em TI trabalhou no exterior, especificamente no Japão onde domina o idioma, durante muitos anos exercendo a profissão de: Operador/Programador de Robô, Líder de produção em circuitos impressos e Gerente de Recursos Humanos.

Tem como áreas de interesse: Pesquisa e Desenvolvimento de: Processamento de Linguagens Naturais, Metodologia e Processos de Engenharia de Software.