ALGORITMO PARA LEITURA DE SIGLAS EM UM SINTETIZADOR DE VOZ

Filipe Leandro de F. Barbosa [1], Fernando Gil Resende [1], Carlos Alexandre Gonçalves [2], Maria Carlota Rosa [2]

DEL/EE [1], Faculdade de Letras [2]
Universidade Federal do Rio de Janeiro
Caixa Postal 68504, 21945-970, Rio de Janeiro, RJ
BRASIL
filipe@lps.ufrj.br, gil@lps.ufrj.br, carlexandre@bol.com.br, carlota@centroin.com.br

Abstract: Neste artigo é apresentado um algoritmo que determina a seqüência fonética que deve ser gerada por um sintetizador de voz na leitura de cada sigla. O algoritmo é descrito através de uma árvore de decisão, cuja entrada é um texto escrito e a saída é a correspondente seqüência de fones. O algoritmo proposto foi testado usando um extrato da base textual do CETEN-Folha com 359 siglas. Para esta base, o percentual de acerto na pronúncia foi de 98.05%.

1. Introdução

Sistemas de conversão texto-fala têm atraído grande interesse tanto no mercado comercial como na área científica. Nestes sistemas, a primeira etapa é a transcrição grafema-fone. Parte de pesquisa que visa a desenvolver um sistema de síntese texto-fala para o português do Brasil, este trabalho apresenta um algoritmo para a leitura de siglas. Uma forma simplificada de resolver o problema da leitura das siglas seria, através de um banco de dados, identificar a sigla e a sua respectiva leitura. Um inconveniente de atualização do banco de dados surgiria com o aparecimento de novas siglas. O desenvolvimento de um algoritmo para a leitura de siglas soluciona este problema.

Em [1] criou-se um conjunto de regras para um transcritor grafema-fone, testado com a base de dados CETEN-Folha [2]. Na leitura-teste dos textos compreendidos por esse *corpus*, parte dos erros então revelados derivava da ininteligibilidade produzida quando o transcritor tinha de gerar uma cadeia sonora para as siglas. O algoritmo proposto neste artigo foi testado usando um extrato da base de texto [2] com 359 siglas. Em 98.05% das siglas a transcrição fonética foi realizada corretamente.

A organização deste trabalho é descrita a seguir. Na Seção 2 é discutida a definição que o algoritmo proposto utiliza para detectar siglas. Na Seção 3 a leitura dos diferentes tipos de siglas são apresentadas. Na Seção 4 resultados são comentados e, na Seção 5, estão as conclusões e questões em aberto. Neste artigo usamos o SAMPA [3] para representação fonética.

2. As siglas do português

A noção de *sigla* é definida como seqüência entre espaços e/ou pontuação, constituída apenas de maiúsculas, ou de uma minúscula e, no mínimo, duas maiúsculas, sendo o contexto precedente e o que se segue não constituído apenas por maiúsculas. A definição de sigla focalizada neste trabalho deixa de lado, por conseguinte, formações como *Anatel, Embrapa*, que seguem a leitura padrão, focalizada pelo transcritor mais geral.

Dois são os parâmetros levados em conta na leitura de siglas: (a) o número de letras envolvidas; e (b) as diferentes possibilidades de composição das seqüências de letras, no tocante à combinação de vogais e consoantes. Desta combinação resultará ou uma leitura soletrada ou uma leitura como sílaba e, associados a essas realizações, diferentes padrões de acentuação.

3. O algoritmo proposto

A Figura 1 mostra o algoritmo de leitura das siglas, que está descrito na forma de uma árvore de decisões. Em linhas gerais, tal recurso consiste numa lista hierárquica de perguntas com respostas simples, sim/não, formuladas a partir do estabelecimento de regras categóricas. As regras são propostas de modo a determinar que tipo de seqüência fonética se deve associar às siglas presentes no texto escrito. O nó inicial da árvore verifica se a palavra é uma sigla. A seleção dos nós subseqüentes dependerá da resposta, *sim* ou *não*, sobre o contexto, até que não caibam mais perguntas. Os símbolos 'C' e 'V', no algoritmo de leitura de siglas, significam consoante e vogal, respectivamente. Assim, a sigla FAB (Força Aérea Brasileira), por exemplo, é classificada como sendo do tipo 'CVC'. O símbolo 'Cn' mencionado na Figura 1, onde 'n' é um número natural diferente de zero, se refere à n-ésima consoante da seqüência que caracteriza a sigla, se contada da esquerda para a direita.

As folhas da árvore são constituídas por três diferentes ações: *Soletrar*, *Silabar* e *Exceções*. A função *Soletrar* associa cada grafema da sigla à sua respectiva realização acústica, conforme descrito na Tabela 1. Por exemplo, no caso da sigla BR (Brasil), o mapeamento dos grafemas para os fones levaria a [beERi].

A função *Silabar* realiza o mapeamento dos grafemas para unidades acústicas da mesma forma que uma dada palavra que não seja caracterizada como sigla, ou seja, comporta-se como o transcritor grafema-fone descrito em [1]. Quando há mais de quatro grafemas em uma palavra reconhecida como sigla, a função *Silabar* será utilizada, a menos que quatro ou mais consoantes estejam em seqüência, quando a rotina *Soletrar* será acionada. As siglas de mais de quatro grafemas que são soletradas devem ser tratadas como exceções, o que justifica o nome desse tipo de folha.

Com a finalidade de atender o caso de plurais de siglas, as funções *Soletrar* e *Silabar* fazem a verificação se o último grafema de uma dada sigla é <s> minúsculo. Em caso afirmativo, é adicionado o fone [s] ao final de cada uma das transcrições grafema-fone.

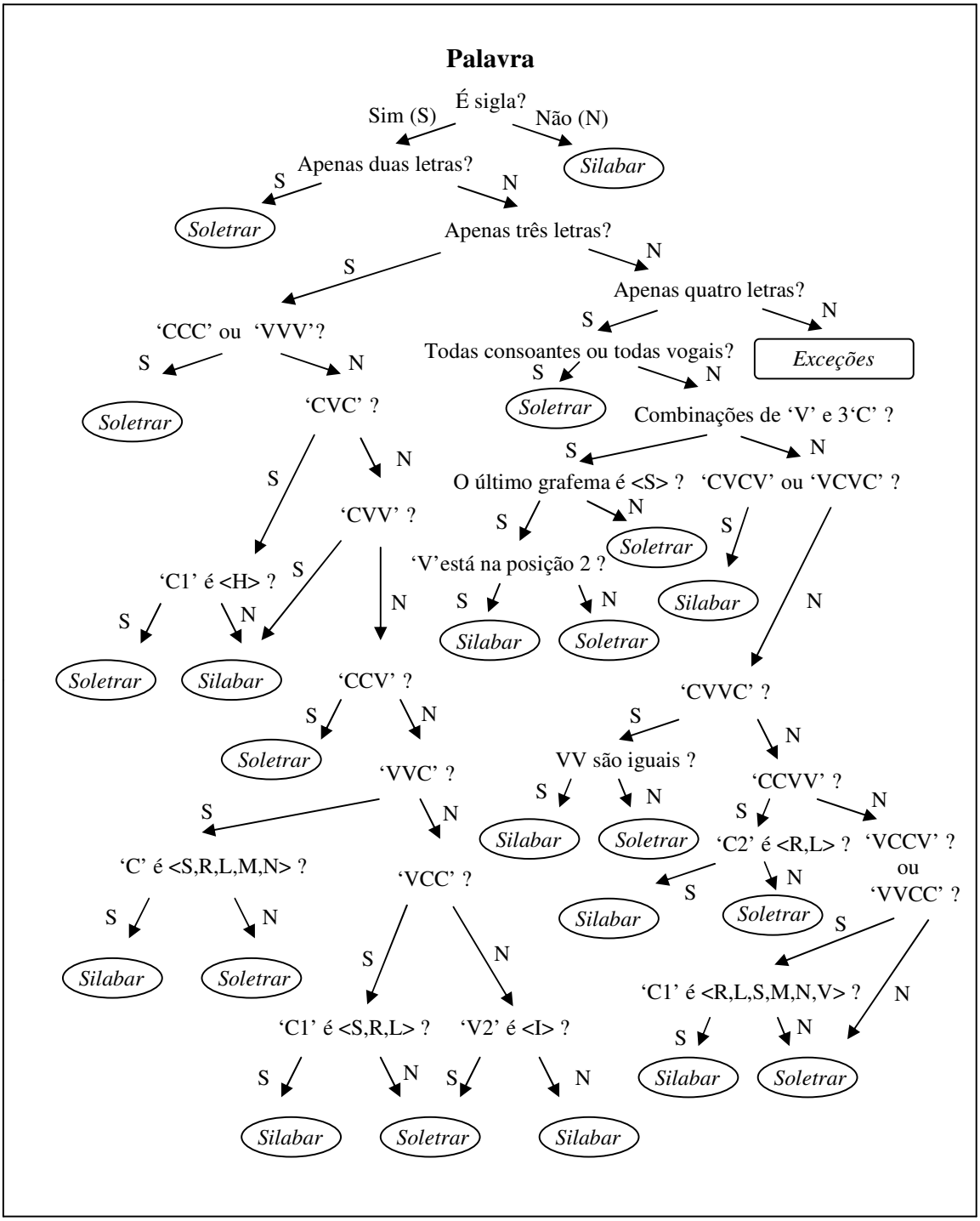

Fig. 1: Descrição da árvore de decisão para leitura de siglas no Português

No caso da sigla possuir hífen, a análise da leitura será feita considerando duas siglas independentes. É importante ressaltar que se a parte da sigla resultante for formada por um único grafema, este será soletrado.

Tabela 1: Descrição da seqüência de fones associada a cada grafema, no comando *Soletrar.*

Grafema	Fone Associado	Grafema	Fone Associado
<A>	[a]	<N>	[eni]
<B>	[be]	<O>	[O]
<C>	[se]	<P>	[pe]
<D>	[de]	<Q>	[ke]
<E>	[E]	<R>	[ERi]
<F>	[Efi]	<S>	[Esi]
<G>	[Ze]	<T>	[te]
<H>	[aga]	<U>	[u]
<I>	[i]	<V>	[ve]
<J>	[ZOta]	<W>	[dabliu]
<K>	[ka]	<X>	[Sis]
<L>	[Eli]	<Y>	[ipsilo~]
<M>	[emi]	<Z>	[ze]

4. Resultados Experimentais

Com a finalidade de testar o algoritmo proposto foram realizados testes utilizando a base de dados de texto CETEN-Folha [2]. Essa base de dados foi construída a partir das 365 edições do jornal brasileiro Folha de São Paulo no ano de 1994 e possui aproximadamente 24 milhões de palavras.

A Tabela 2 mostra os resultados obtidos ao utilizarmos o algoritmo de decisão de leitura aplicado a um extrato selecionado aleatoriamente da base de texto do CETEN-Folha. Neste extrato foram encontradas 359 ocorrências de siglas. As ocorrências de uma mesma sigla não foram descartadas para que haja uma correspondência entre a estatística de acerto e a freqüência relativa de aparecimento das siglas nos textos escritos. A Tabela 2 mostra os resultados incluindo e não incluindo as repetições de siglas.

A maioria dos erros de leitura deveu-se às siglas estrangeiras, tais como: NY (*New York*), DJ (disk jokey) e LA (*Los Angeles*). Tais problemas de mapeamento grafema-fone podem ser resolvidos através de um banco de exceções que daria a forma correta de leitura a essas siglas.

Tabela 2: Resultados experimentais para análise de um extrato do CETEN-Folha.

Total de siglas	Total de erros	Acerto (%)
359	7	98.05
Total de siglas diferentes	Total de erros	Acerto(%)
132	5	96.21

Dentre os problemas não resolvidos pela árvore de decisão podem ser citados os algarismos romanos, que seriam detectados como siglas, já que se encaixam nesse contexto. Para solucionar esse problema, pode ser realizada uma análise de contexto, tal como desenvolvido em [4].

4. Conclusões

Neste artigo é apresentado um algoritmo para leitura de siglas por um sintetizador de voz. As regras são baseadas no número de letras envolvidas e nas suas possíveis combinações, particularmente, em relação a vogais e consoantes. O programa desenvolvido foi testado usando um extrato da base textual CETEN-Folha com 359 siglas, tendo acertado em 98.05% dos casos. O tratamento dos erros, tais como a leitura de números romanos, é assunto de pesquisa em andamento.

Referências

[1] Barbosa, F.L.F.; Pinto, G.O.; Resende Jr., F.G.V.; Gonçalves, C.A.; Monserrat, R.; Rosa, M.C. "Grapheme-phone transcription algorithm for a Brazilian Portuguese TTS", to appear in *Proc. of 6th Workshop on Computational Processing of the Portuguese Language – Written and Spoken*, Faro, Portugal, June 2003.

[2] *Corpus de Extractos de Textos Electrônicos NILC/Folha de São Paulo (CETEN-Folha).* http://acdc.linguateca.pt/cetenfolha/

[3] *Speech Assesment Methods Phonetic Alphabet (SAMPA).* http://www.phon.ucl.ac.uk/home/sampa

[4] Barbosa, F.L.F.; Ferrari, L.; Resende Jr., F.G.V. "A methodology to analyze homographs for a Brazilian Portuguese TTS system", to appear in *Proc. of 6th Workshop on Computational Processing of the Portuguese Language – Written and Spoken*, Faro, Portugal, June 2003.