CONVERSÃO FALA-TEXTO PARA O PORTUGUÊS DO BRASIL COM VOCABULÁRIO ILIMITADO

Francisco J. Fraga
E-mail: fraga@inatel.br

Grupo de Pesquisa em Processamento Digital de Sinais
Instituto Nacional de Telecomunicações
Santa Rita do Sapucaí, MG, Brasil

RESUMO

A implementação de um reconhecedor automático de fala com vocabulário ilimitado, para o português falado no Brasil, pode ser realizada mediante duas etapas consecutivas: O reconhecimento de fonemas e a conversão da seqüência de fonemas em texto. Este artigo apresenta um sistema de conversão fala-texto com estas características, descrevendo brevemente o reconhecedor de fonemas e mais detalhadamente o algoritmo de conversão fonológico-grafêmica. O algoritmo é inteiramente baseado em regras extraídas da própria estrutura da língua portuguesa, permitindo a passagem do nível dos fonemas para o das palavras sem recorrer a nenhum tipo de tabelas de pronúncia. Desta forma o sistema é capaz de reconhecer qualquer palavra pertencente ao léxico da língua portuguesa, sem limitação alguma com relação ao vocabulário abrangido.

1. INTRODUÇÃO

Alguns sistemas de reconhecimento automático de fala, especialmente desenvolvidos para a língua portuguesa e atualmente implementados com sucesso[1], podem chegar a reconhecer um vocabulário de cerca de 60.000 palavras. Em tais sistemas, o vocabulário pode ser expandido pelo próprio usuário, incluindo cada nova palavra que deseja que o sistema seja capaz de reconhecer. No entanto, por mais extenso que seja o vocabulário assim formado, ele nunca poderá ser classificado como ilimitado, mesmo sendo flexível, pois cada nova palavra deverá ser individualmente acrescentada ao vocabulário.

Esse procedimento está relacionado com o método de reconhecimento de fala utilizado por esses sistemas. O método mais usado, por ser o que apresenta melhor desempenho, é o de associar um modelo oculto de Markov (HMM) a cada unidade fonética [1]. Em geral, as unidades fonéticas correspondem aos fonemas da língua na qual pretende-se realizar o reconhecimento de fala. Para os sistemas de vocabulário grande, além dos modelos de palavras formados pela concatenação das unidades fonéticas, costuma-se usar também um modelo estatístico da língua em questão. O modelo de língua atribui probabilidades aos eventos de sucessão de palavras em frases que façam sentido naquela língua [2]. Estes modelos reduzem a *perplexidade*[2] a algumas dezenas, aumentando consideravelmente a taxa de acertos no reconhecimento e diminuindo o tempo gasto pelo algoritmo de busca.

Para saber quais unidades devem ser concatenadas para formar uma determinada palavra, torna-se necessário gerar modelos de pronúncia para cada palavra do vocabulário. Os modelos de pronúncia são seqüências de fones ou fonemas, que indicam ao mecanismo de reconhecimento quais unidades fonéticas devem ser concatenadas para formar cada palavra do vocabulário.

[1] Por exemplo, o "ViaVoice", da empresa IBM.

[2] O conceito de perplexidade (PP) deriva do conceito de entropia (H), sendo que $PP = 2^H$. A perplexidade, quando aplicada a um modelo de língua, indica o número médio de palavras que podem seguir-se a uma palavra previamente determinada.

Nos sistemas mais modernos, que permitem ao usuário adicionar novas palavras ao vocabulário, a geração dos modelos de pronúncia é feita de forma automática a partir da grafia das palavras. Para tanto, fazem uso do mesmo algoritmo empregado pelos *softwares* de síntese de fala (*text-to-speech*), que a partir da grafia de uma palavra geram a sua pronúncia (seqüência de fonemas) [3]. Uma maneira de fazer com que estes sistemas de vocabulário grande tornem-se sistemas de vocabulário ilimitado, seria elaborar um algoritmo que fizesse o caminho contrário: A partir de uma determinada pronúncia, isto é, de uma seqüência de fonemas, gerar a correspondente grafia da palavra.

O objetivo deste trabalho foi o de averiguar a possibilidade de converter uma seqüência de fonemas em uma ou várias seqüências de grafemas (letras formando palavras), sem usar nenhum tipo de tabelas de pronúncia como se faz habitualmente [4]. Para tanto, o que se fez foi descobrir regras específicas, aplicáveis ao português do Brasil, de transformação de seqüências de fonemas em grafemas, possibilitando o uso de um vocabulário ilimitado. A seção 2 traz uma breve explicação da etapa de reconhecimento dos fonemas do português brasileiro e uma relação completa dos mesmos. Na seção 3 é apresentado o algoritmo de conversão de seqüências de fonemas em possíveis grafemas na língua portuguesa. A seção 4 trata dos resultados obtidos na conversão fala-texto e finalmente a seção 5 faz a conclusão, indicando as vantagens do uso deste algoritmo de conversão fonológico-grafêmica em sistemas de reconhecimento automático de fala.

2. RECONHECIMENTO DE FONEMAS

Não faz parte do escopo deste artigo a descrição detalhada do método utilizado no reconhecimento de fonemas. No entanto, o algoritmo de conversão fonológico-grafêmica, que será explicado na seção seguinte, foi de fato utilizado como etapa final de um sistema de reconhecimento de fala completo. A construção da etapa inicial deste sistema foi apresentada em um artigo no XV Simpósio Brasileiro de Telecomunicações (1997) [5]. A descrição e o detalhamento da implementação do sistema completo foram apresentados como tese de doutorado [6]. A abordagem empregada no referido sistema foi a de segmentar cada palavra a ser reconhecida em unidades sub-silábicas, que posteriormente são convertidas em uma seqüência fonêmica, contendo também a indicação da posição do acento tônico dentro da seqüência.

Fonema	**Exemplo**	**Fonema**	**Exemplo**
a	b**a**sta	**k**	care**c**a
E	d**e**la	**l**	gaze**l**a
e	m**e**smo	**L**	mu**lh**eres
i	d**i**zem	**m**	nenhu**m**a
O	f**o**rte	**n**	**n**a
o	b**o**nito	**N**	so**nh**e
u	g**u**ri	**p**	**p**otente
~a	m**an**da	**r**	se**r**
~e	hom**en**s	**R**	ca**rr**o
~i	**im**porta	**s**	bon**s**
~o	c**on**forto	**t**	fon**t**e
~u	**un**s	**v**	jo**v**em
b	**b**ula	**x**	**ch**efe
d	juran**d**o	**z**	pe**s**a
f	**f**enda	**y**	mã**e**
g	**g**onzos	**w**	pã**o**
j	**g**erentes		

Tabela I: Fonemas da Língua Portuguesa

No entanto, qualquer outro sistema que realize o reconhecimento de fonemas do português brasileiro, pode ser usado como entrada para o algoritmo de conversão fonológico-grafêmica, desde que obedeça a notação apresentada na Tabela I. Esta notação difere daquela padronizada pelo Alfabeto Fonético Internacio-

nal apenas por razões de ordem prática, visando facilitar a implementação computacional do algoritmo de conversão fonológico-grafêmica.

A entrada deste algoritmo será uma seqüência fonológica (fonêmica) de acordo com a notação apresentada na tabela e contendo a indicação da posição do acento tônico por meio de um apóstrofo (caracter `).

3. ALGORITMO DE CONVERSÃO FONOLÓGICO-GRAFÊMICA

Antes de iniciar a transformação fonológico-grafêmica, os fonemas são previamente analisados e através de uma série de restrições impostas à seqüência fonológica obtida, consegue-se uma depuração inicial, que visa eliminar as seqüências de fonemas não permitidas pela língua portuguesa. A seguir os fonemas são classificados, de acordo com a sua posição na seqüência de entrada, em fonemas iniciais, mediais ou finais.

Entram então em jogo as regras específicas para cada fonema e seu contexto, isto é, seus antecedentes e subseqüentes dentro da palavra. Tendo a Lingüística e a Filologia como fontes de conhecimento específicos do português brasileiro [7], produz-se como saída uma série de palavras, ordenadas por critérios probabilísticos extraídos do léxico, que formam o assim chamado conjunto de possíveis grafemas para uma dada seqüência fonológica da língua portuguesa.

Procura-se considerar como "lícitas" diversas variações de pronúncia de algumas regiões brasileiras, mas somente aquelas que acarretam diferenças na transcrição fonológica. As de nível fonético são absorvidas pelo reconhecedor de fonemas na passagem para o nível fonológico. Como exemplo de variação de pronúncia em nível fonético temos o caso da palavra *tia*, na qual o fonema inicial / t / pode ser pronunciado usando-se o som plosivo [*t*] ou sua forma africada [*tch*]. A troca de [*t*] por [*tch*] não altera o significado do signo *tia*; dizemos então que [*t*] e [*tch*] são *alofones* do fonema / t / [8].

Em nível fonológico, por exemplo, temos as variações de pronúncia das palavras *mentira* (pronunciada / *m~it`ira* / ou / *m~et`ira* /) e *homem* (pronunciada como / *`Om~ey* / ou como / *`~om~e* /). O algoritmo foi desenvolvido tendo em conta estas variações de pronúncia em nível fonológico, de forma que elas não impedem a obtenção da grafia correta, como é possível observar pelos exemplos mostrados na Tabela III.

R1	Fonema posterior é vogal ou semivogal
R2	Fonema anterior é vogal oral
R3	Palavra começa com / *e* / ou / *ine* /
R4	Fonema posterior é / *E* /
R5	Fonema post. é / *e* /, / *i* /, / *~e* /, / *~i* / ou / *y* /
R6	Fonema posterior é / *~o* / ou / *~u* /
R7	Fonema posterior é / *E* /, / *e* / ou / *~e* /
R8	Fonema posterior é / *~i* / ou / *~i* /
R9	Fon. anter. é vogal nasal, semivogal ou / *R* /
R10	Fonema anterior é / *b* /
R11	Fonema anterior é vogal nasal
R12	Fon. post. é / *E* /, / *e* /, / *i* /, / *~e* /, / *~i* /, / *y* /
LER FONEMA MEDIAL / *s* /	
Se	~R1 → 2I (**x , s**)
Se	R1R2~R3R4 → 2I (**c , ss**)
Se	R1R2~R3~R4R5 → 3I (**c , sc , ss**)
Se	R1R2~R3~R4~R5~R6 → 3I (**ç , sç , ss**)
Se	R1R2~R3~R4~R5R6 → I (**ss**)
Se	R1R2R3R7 → 2I (**xc , ss**)
Se	R1R2R3~R7R8 → 3I (**c , sc , ss**)
Se	R1R2R3~R7~R8~R6 → 3I (**ç , sç , ss**)
Se	R1R2R3~R7~R8R6 → I (**ss**)
Se	R1~R2R9~R12 → 2I (**ç , s**)
Se	R1~R2~R9R10 → 3I (**c , sc , s**)
Se	R1~R2R9R12~R11 → 2I (**c , s**)
Se	R1~R2R9R12R11 → 3I (**c , sc , s**)
Se	R1~R2~R9~R10~R12 → I (**s**)
Se	R1~R2~R9~R10R12 → I (**c**)

Tabela II: Regras para o fonema medial / *s* /

Para se ter uma idéia da complexidade deste algoritmo de conversão fonológico-grafêmica, apresentamos na Tabela II as regras relativas unicamente ao fonema / *s* / quando presente no meio da palavra (o sinal ~ indica negação lógica). Os comandos 2I e 3I na Tabela II são os responsáveis pela multiplicação de possibilidades grafemáticas, oriundas de uma única seqüência fonêmica de entrada. De fato, o número de possibilidades geradas pelo algoritmo varia muito de acordo com o fonema visado e seu respectivo contexto. Este fato pode ser verificado por meio da Tabela II, nas regras de decisão onde são averiguados os fonemas anteriores ou posteriores ao fonema / *s* /.

Entrada : **/`Om~ey/** Saída : omem **homem** ómen hómen	Entrada : **/m~it`ira/** Saída : mintira **mentira** mintera mentera	Entrada : **/as`Esu/** Saída : **acesso** hacesso assesso hassesso aceço haceço asseço hasseço acesço hacesço assesço hassesço acessu hacessu assessu hassessu aceçu haceçu asseçu hasseçu acesçu hacesçu assesçu hassesçu
Entrada : **/`~om~e/** Saída : omem **homem** ômen hômen	Entrada : **/awz`~eti/** Saída : ausênti hausênti alsênti halsênti auzênti hauzênti alzênti halzênti **ausente** hausente alsente halsente auzente hauzente alzente halzente	
Entrada : **/ses`~aw/** Saída : **sessão** **cessão** **seção** ceção sesção cesção		

Tabela III: Exemplos de conversão fonológico-grafêmica realizadas pelo algoritmo

A Tabela III mostra diversas seqüências de fonemas submetidas ao conversor fonológico-grafêmico e suas respectivas saídas em forma de possibilidades grafemáticas. É interessante observar, em todos os casos mostrados na Tabela III, que todas as possibilidades grafemáticas, quando pronunciadas, convergem para a mesma seqüência fonêmica de entrada.

Isso explica em parte porque é tão fácil errar na escrita das palavras, pois elas de fato podem ser grafadas de muitas maneiras diferentes. Ou seja, a grafia correta é uma convenção, determinada por razões de ordem histórica e filológica. O contrário também pode ser observado em alguns casos, isto é, diferentes seqüências de entrada geram a mesma saída; como no caso da palavra **homem,** que é apresentada na tabela, sendo obtida por meio de dois diferentes modos de pronúncia.

Na mesma tabela, a palavra de saída ortograficamente correta foi assinalada em negrito por um corretor ortográfico de uso comercial[3]. Esta correção ortográfica constitui a última fase do sistema completo de conversão fala-texto, com vocabulário ilimitado, a partir do português falado no Brasil [6].

As possibilidades grafemáticas de saída mostradas na Tabela III são apresentadas pelo algoritmo por ordem de probabilidade, sendo as primeiras aquelas grafias mais freqüentemente relacionadas com a seqüência fonêmica de entrada. Para obter esta hierarquia de possibilidades e as regras de conversão fonema-letra, foi necessário um amplo trabalho de pesquisa e classificação das grafias mais freqüentes, tomando-se por base todo o léxico da língua portuguesa [9].

Como pode-se observar, no caso da seqüência fonológica **/ses`~aw/,** apresentada na tabela, mais de uma palavra ortograficamente correta é assinalada. Nesse e em outros casos semelhantes, que são muito poucos na língua portuguesa, a decisão final entre as palavras só poderia ser feita por meio de uma análise semântica da frase completa.

[3] "ProVerb" versão 2.0 da empresa "PC software".

4. RESULTADOS

A base de dados de teste usada para obter as taxas de acerto finais do sistema completo de conversão fala-texto é composta de 200 frases, pausadamente pronunciadas por um locutor do sexo masculino. As frases continham 1729 palavras e 6988 fonemas, sendo que destes 3496 eram vogais e 3492 eram consoantes. A etapa de reconhecimento de fonemas fornece ao conversor fonológico-grafêmico não apenas uma, mas várias possíveis seqüências fonêmicas para cada palavra falada.

Primeiramente mostraremos (Tabela IV) os resultados para as ***n*** primeiras possibilidades ortograficamente corretas geradas pelo sistema, para cada palavra de cada uma das 200 frases pronunciadas. As distribuições percentuais referem-se ao melhor candidato encontrado desde o primeiro até o ***n***-ésimo.

n	***Fonemas Corretos***	***Palavras Corretas***	***Inserção Fonemas***	***Exclusão Fonemas***
1	**95,9 %**	**87,0 %**	**0,72 %**	**1,07 %**
2	98,0 %	92,8 %	0,66 %	0,72 %
3	98,6 %	94,4 %	0,63 %	0,69 %
6	99,0 %	96,5 %	0,55 %	0,61 %

Tabela IV: Resultados de reconhecimento para as primeiras grafias corretas

A seguir apresentamos (Tabela V) os resultados de reconhecimento sem levar em conta a correção ortográfica, tomando as ***n*** primeiras possibilidades grafemáticas geradas e comparando-as com a palavra realmente falada.

Podemos notar na primeira linha de cada tabela, que os resultados de reconhecimento de palavras são pobres, se comparados aos resultados de reconhecimento de fonemas. Porém, eles melhoram significativamente quando tomamos um maior número de possibilidades, mormente quando é usada a correção ortográfica. Nesse sentido, é interessante notar uma diferença marcante entre o nosso sistema e aqueles outros que utilizam-se de tabelas de pronúncia e modelos de língua, onde a taxa de acerto para as palavras ou frases é sempre *maior* que a taxa de reconhecimento de fonemas [10].

n	***Fonemas Corretos***	***Palavras Corretas***	***Inserção Fonemas***	***Exclusão Fonemas***
1	**87,9 %**	**60,6 %**	**1,33 %**	**1,08 %**
2	90,7 %	67,8 %	1,33 %	1,05 %
3	91,8 %	71,0 %	1,33 %	1,02 %
9	95,2 %	81,4 %	1,30 %	0,93 %

Tabela IV: Resultados de reconhecimento para as primeiras possibilidades grafemáticas

Esta é uma característica própria dos sistemas que visam um vocabulário grande porém *limitado* ; pois esta restrição permite, por eliminação, escolher a palavra ou frase que melhor corresponde à seqüência fonética reconhecida. Pelo contrário, no nosso sistema, que almeja o reconhecimento de palavras pertencentes a um vocabulário *ilimitado*, as taxas de reconhecimento de palavras são e serão sempre *inferiores* às taxas de acerto dos fonemas. Isto porque não existe nenhum modo de restringir a busca pelas palavras corretas a não ser por meio da correção ortográfica, já implementada com sucesso, e da análise sintática e semântica das frases, sugeridas como trabalho futuro.

Por isso queremos chamar a atenção para a última linha da Tabela IV: A significativa melhora dos resultados com o aumento do número de candidatos considerados mostra que, se pudéssemos escolher sempre o melhor dentre eles, alcançaríamos elevadas taxas de reconhecimento de palavras. Acreditamos portanto que as altas taxas de reconhecimento de fonemas (**99%**) e de palavras (**96,5%**), obtidas quando se toma o melhor dentre os primeiros 6 candidatos de palavras, não são uma meta inatingível. Estimamos que estas taxas poderão ser alcançadas por este mesmo sistema, se o texto final gerado for submetido a uma análise sintática e semântica, desenvolvendo-se para tanto as ferramentas de inteligência artificial correspondentes.

5. CONCLUSÃO

Descrevemos neste artigo a implementação de um algoritmo que converte uma seqüência de fonemas em grafemas, inteiramente baseado em regras extraídas da própria estrutura da língua portuguesa, que permite passar do nível dos fonemas para o das palavras sem recorrer a nenhum tipo de tabelas de pronúncia. Utilizando-se como etapa prévia de um determinado reconhecedor de fonemas [6] e como etapa posterior de um *software* de correção ortográfica, atingiu-se uma taxa de acerto razoável no reconhecimento de palavras, considerando que trata-se de um vocabulário *ilimitado*.

Esta taxa aumentaria consideravelmente caso fosse implementada uma etapa posterior de pós-processamento de texto, realizado por um analisador sintático e semântico. Esse pós-processamento seria feito com base nas técnicas de processamento de língua natural [10], porém adaptadas para esta finalidade específica, como já se tentou fazer para o caso da língua alemã [11]. Essa sugestão, integrando ao atual sistema algumas ferramentas de inteligência artificial, seria a melhor continuação que poderia fazer-se do presente trabalho.

6. REFERÊNCIAS

[1] X. D. Huang, Y. Ariki, M. A. Jack; *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[2] CHIEN, L. F. ; CHEN, K-J. ; LEE, L-S. "A Best-First Language Processing Model Integrating the Unification Grammar and Markov Language Model for Speech Recognition Applications". *IEEE Transactions on Speech and Audio Processing*, vol. 1, nº 2, pp 221-239, April 1993.

[3] VAN COILE, B. "On the Development of Pronunciation Rules for Text-to-Speech Synthesis". *Proceedings of Eurospeech Conference*, Berlin, Sep 1993, pages 1455-1458

[4] ZHAO, Y. "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 1, nº 3, pp 345-361, July 1993.

[5] FRAGA, F. J.; SAOTOME, O. "Reconhecimento de Fala com Vocabulário Ilimitado para o Português do Brasil". *Anais do XV Simpósio Brasileiro de Telecomunicações*, pp 10-13, Setembro de 1997.

[6] FRAGA, F. J.; SAOTOME, O. *Conversão Fala-Texto em Português do Brasil Integrando Segmentação Sub-Silábica e Vocabulário Ilimitado.* Tese de Doutorado, ITA, 1998.

[7] SILVA,M.C.; KOCH, I.G. *Lingüística Aplicada ao Português: Morfologia*, Cortez, 1983.

[8] LYONS, J. *Introduction to Theoretical Linguistics.* Cambridge University Press, Cambridge, 1968.

[9] FERREIRA, AURÉLIO B. H. *Novo Dicionário da Língua Portuguesa*, Nova Fronteira, 1975.

[10] MORAIS, E. S.; VIOLARO, F. "Sistema Híbrido ANN-HMM para Reconhecimento de Fala Contínua". Anais do XV Simpósio Brasileiro de Telecomunicações, pp 117-120, Setembro de 1997.

[11] PEREIRA, F.C.N.; GROSZ, B. J. *Natural Language Processing***,** Elsevier, 1993.

[12] MUDLER, J. "A System for Improving the Recognition of Fluently Spoken German Speech". *Proceedings of IJCAI*, pp 633-635, 1983.