Conversão Fala-texto para o Português com Segmentação Sub-silábica e Vocabulário Ilimitado

Francisco J. Fraga

Instituto Nacional de Telecomunicações - Santa Rita do Sapucaí - MG - Brasil

***Resumo* — A tarefa de implementação de um sistema de conversão fala-texto com vocabulário ilimitado, para o português falado no Brasil, pode ser realizada mediante duas etapas consecutivas: O reconhecimento de fonemas por meio de uma segmentação sub-silábica e a conversão da seqüência de fonemas em texto. Este artigo apresenta um sistema de reconhecimento automático de fala com estas características, descrevendo brevemente o método de segmentação juntamente com o reconhecedor de fonemas e mais detalhadamente o algoritmo de conversão fonológico-grafêmica. Este último 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 com relação ao tamanho do vocabulário abrangido.**

***Palavras–Chave* — Reconhecimento de fonemas, modelos ocultos de Markov, sistemas fala–texto.**

***Abstract* — It is possible to implement a speech–to–text system with unlimited vocabulary by connecting two subsystems: A phoneme recognizer, which is performed by sub–syllabic segmenting the incoming speech, and a phonologic–graphemic converter. This paper presents an automatic speech recognition system with these features. The segmentation method and the phoneme recognizer are briefly described while the phonologic–graphemic converter is detailed. The algorithm that allows the transition from the phoneme level to the word level is based on rules obtained from the structure of the portuguese language. This task is achieved without any kind of pronouncing tables, which allows the system to recognize any word that belongs to the portuguese lexicon, without limitation on the size of the vocabulary.**

***Index Terms* — Phoneme recognition, hidden Markov models, speech–to–text systems.**

I. INTRODUÇÃO

Quando se pensa em construir um sistema de reconhecimento de fala visando a geração automática de texto, logo se coloca a questão das unidades fonéticas a serem utilizadas. A técnica dos modelos ocultos de Markov *("Hidden Markov Models - HMM")* produz resultados excelentes quando são utilizados modelos de palavras inteiras [1], porém a necessidade de treinar o sistema pronunciando-se diversas vezes cada palavra, necessariamente limita o tamanho do vocabulário a poucas centenas de palavras. A solução se encontra no uso de modelos (unidades fonéticas) menores, tais como sílabas, semi-sílabas, fonemas ou alofones.

No entanto, na medida em que é reduzido o número de unidades a serem treinadas, quase sempre se aumenta a taxa de erros, pois fenômenos tais como as nasalizações e a coarticulação entre os diversos fonemas são difíceis de modelar, devido ao elevado número de diferentes contextos nos quais ocorrem [2].

Observando-se a estrutura fonética do português falado no Brasil, percebe-se que o fato de as vogais ocuparem um papel de destaque (a proporção entre o número de vogais e o de consoantes é maior que a de outros idiomas) sugere o uso de unidades de reconhecimento baseadas nas vogais, como é o caso das sílabas, pois a vogal é sempre o centro da sílaba. Sendo assim, pode-se dividir cada sílaba em uma parte ascendente, do início até o centro da vogal, e outra descendente, do centro até o final. A estas partes da sílaba assim definidas dá-se o nome de semi-sílabas. O número de semi-sílabas do português é inferior a 700; dividindo-se este conjunto em grupos correspondentes a cada vogal, obtém-se sub-vocabulários da ordem de algumas dezenas, que é o tamanho ideal para o emprego bem sucedido da técnica HMM.

Outra questão relacionada com o tamanho do vocabulário é a necessidade do uso de modelos de pronúncia das palavras. 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 a cada unidade fonética [3]. Em geral, as unidades fonéticas correspondem aos fonemas da língua na qual pretende-se realizar o reconhecimento de fala. Para os sistemas de vocabulário gran-

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

de, 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 [4]. 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.

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) [5]. 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 [6]. 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, através de uma segmentação do sinal de fala em semi-sílabas. 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.

II. RECONHECIMENTO DE FONEMAS

II.1 FONEMAS E SEMI-SÍLABAS

A abordagem empregada para realizar o reconhecimento de fonemas foi a de segmentar cada palavra a ser reconhecida em unidades sub-silábicas, que posteriormente são convertidas em uma seqüência fonêmica. Mediante esta segmentação prévia, o reconhecimento de fala contínua pode ser feito com o mesmo método usado para palavras isoladas, sem a necessidade do uso de algoritmos de busca, tais como o *Level Building* [7].

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	**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 1: Fonemas da Língua Portuguesa

Não faz parte do escopo deste artigo a descrição detalhada do método utilizado na segmentação do sinal de fala. A descrição e o detalhamento da implementação do sistema completo foram apresentados como tese de doutorado [8]. A abordagem empregada no referido sistema foi a de segmentar cada palavra a ser reconhecida em semi-sílabas, que posteriormente são convertidas em uma seqüência fonêmica, contendo também a indicação da posição do acento tônico dentro da seqüência.

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 1. Esta notação difere daquela padronizada pelo Alfabeto Fonético Internacional apenas por razões de ordem prática, visando facilitar a implementação computacional do algoritmo de conversão fonológico-grafêmica.

A fim de conseguir a realização de todas as etapas da conversão fala-texto, visando um vocabulário limitado apenas pela estrutura da língua portuguesa, tivemos de fazer algumas restrições. São elas a exigência de pronúncia pausada e clara, principalmente

[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.

das vogais, e a eliminação dos encontros vocálicos (ditongos e tritongos). Estas restrições foram feitas a fim de facilitar a tarefa de segmentação, que exige a identificação da vogal central de cada sílaba, e diminuir o número de possíveis semi-sílabas.

A seguir apresentamos (Tabela 2) a relação das sílabas simples (consoante - vogal - consoante) que foram as escolhidas para esta fase de desenvolvimento do sistema. Nesta relação elas já figuram divididas em semi-sílabas ascendente e descendente, pois assim serão modeladas e treinadas. As semi–sílabas ascendentes são formadas pela concatenação da consoante inicial com a vogal central (indicadas com *VOG* na tabela. Analogamente, as semi–sílabas descendentes são formadas pela seqüência de vogal central e consoante final (indicadas com *CF* na Tabela 2).

CONSOANTES INICIAIS																		*VOG*	*CF*	
-	**b**	**d**	**f**	**G**	**j**	**k**	**l**	**L**	**m**	**n**	**N**	**p**	**r**	**R**	**s**	**t**	**v**	**x**	**z**		**-**	**r**	**s**
																				a			
																				E			
																				e			
																				i			
																				O			
																				o			
																				u			
																				~ a		**X**	
																				~ e		**X**	
											X									**~ i**		**X**	
								X												**~ o**		**X**	
								X												**~ u**		**X**	

Tabela 2 : Semi-sílabas simples

Pela Tabela 2 pode-se notar claramente a vantagem do uso da semi-sílaba, pois na língua portuguesa qualquer sílaba termina em vogal (seguida ou não de r ou s) ou em ditongo decrescente (seguido ou não de s). Os ditongos decrescentes não aparecem na tabela por causa das restrições adotadas nesta fase de implementação do sistema. Sendo assim, existem poucas combinações possíveis para as semi-sílabas descendentes, mesmo porque várias combinações vogal-consoante não são permitidas no português, o mesmo acontecendo com algumas combinações consoante-vogal, ambas assinaladas com um **X** na Tabela 2. Verificamos também que, quando futuramente forem considerados os encontros consonantais e os ditongos, o número destas combinações proibidas cresce consideravelmente, aumentando ainda mais a viabilidade do presente enfoque.

II.2. BASE DE DADOS DE FALA

A base de dados de treinamento constitui um ponto chave para o bom desempenho de um reconhecedor baseado em HMM, principalmente se forem empregadas funções densidades de probabilidade contínuas, como é o caso deste projeto. Para treinar satisfatoriamente os modelos, foram necessárias dezenas de gravações de cada uma das sílabas [2], todas elas pronunciadas por um único locutor do sexo masculino. Para captação do sinal de voz foi utilizado um microfone profissional unidirecional, em ambiente sem isolação acústica (diante de um microcomputador). A amostragem foi efetuada por uma placa de som Sound Blaster 16-ASP da Creative Labs à taxa de 22 kHz, com 16 bits por amostra.

II.3. PARÂMETROS DO PROCESSAMENTO DIGITAL DA VOZ.

As elocuções foram analisadas por meio de uma janela Hammig de 20 ms, aplicada a cada 10 ms em um sinal pré-enfatizado pela função de transferência $1 - 0.95z^{-1}$. A seguir foram extraídos, para cada janela, 12 coeficientes mel–cepstrais [9], aos quais foi acrescentado o valor do logaritmo da energia total de cada segmento, normalizada pela energia máxima de cada elocução. Também foram calculadas os parâmetros delta–mel e delta–energia (derivadas de primeira ordem), de acordo com a expressão

$$\Delta m_t = \frac{\sum_{\tau=1}^{Q} \tau (m_{t+\tau} - m_{t-\tau})}{2\sum_{\tau=1}^{Q} \tau^2}$$

onde m_t é o coeficiente mel–cepstral extraído na janela de tempo com índice t. Nesta implementação usamos $Q = 2$, ou seja, consideramos as diferenças entre os coeficientes de janelas adiantadas e atrasadas de 1 e de 2 intervalos de 10 ms. Portanto, o vetor de observação é composto por 12 coeficientes mel–cepstrais, 12 coeficientes delta–mel, log–energia e delta–log–energia, formando um vetor de 26 componentes.

II.4. TOPOLOGIA DOS MODELOS OCULTOS DE MARKOV

Foram considerados três modelos para cada sílaba: Um modelo de vogal, destinado a reconhecer a região central da sílaba, um modelo para a semi-sílaba ascendente e outro para a descendente. Para todos eles foi utilizada a topologia *left-right* conforme ilustra a figura 1, sendo considerados 4 estados para as regiões

ascendente e descendente e 3 estados para a região central.

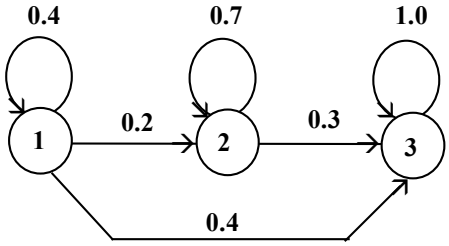

Figura 1 : HMM usado para as vogais

Os valores decimais que aparecem junto a cada uma das transições indicam a probabilidade de mudança ou de permanência em cada estado. Observe-se que a somatória das probabilidades das transições de cada estado é sempre igual a 1. A cada estado está associada uma função densidade de probabilidade composta por uma mistura de 4 distribuições gaussianas, com matriz de covariância diagonal. O uso preferencial da matriz de covariância diagonal obedece a razões de ordem prática, pois o tempo de reconhecimento aumentaria bastante caso fossem utilizadas matrizes de covariância completas, o mesmo ocorrendo em escala muito maior com o tempo de treinamento.

III. ALGORITMO DE CONVERSÃO FONOLÓGICO-GRAFÊMICA

A entrada deste algoritmo será uma seqüência fonológica (fonêmica) de acordo com a notação apresentada na Tabela 1 e contendo a indicação da posição do acento tônico por meio de um apóstrofo. 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 [10], 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 / [11].

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 4.

Para se ter uma idéia da complexidade deste algoritmo de conversão fonológico-grafêmica, apresentamos na Tabela 3 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 3 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 3, nas regras de decisão onde são averiguados os fonemas anteriores ou posteriores ao fonema / *s* /.

A Tabela 4 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 4, 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 [8].
As possibilidades grafemáticas de saída mostradas na Tabela 4 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 [12].

Como pode-se observar, no caso da seqüência fonológica **/ses`~aw/,** apresentada na tabela, mais de uma palavra ortograficamente correta é assinalada.

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

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 3: Regras para o fonema medial / *s* /

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.

IV. 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.

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 4: Exemplos de conversão fonológico-grafêmica realizadas pelo algoritmo

Primeiramente mostraremos (Tabela 5) 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.

A seguir apresentamos (Tabela 6) os resultados de reconhecimento sem levar em conta a correção ortográfica, tomando as ***n*** primeiras possibilidades

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

Tabela 5: Resultados de reconhecimento para as primeiras grafias corretas

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 [13].

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

Tabela 6: 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 5: 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.

V. CONCLUSÃO

Descrevemos neste artigo a implementação de um sistema de conversão fala–texto, com segmentação prévia da fala em semi–sílabas antes de proceder ao reconhecimento dos fonemas. Com mais detalhe, descrevemos o algoritmo que converte a seqüência de fonemas reconhecidos 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 como etapa posterior 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 [14], porém adaptadas para esta finalidade específica, como já se tentou fazer para o caso da língua alemã [15]. Essa sugestão, integrando ao atual sistema algumas ferramentas de inteligência artificial, seria a melhor continuação que poderia fazer-se do presente trabalho.

REFERÊNCIAS

[1] L. R. Rabiner, B. H. Huang; An Introduction to Hidden Markov Models, IEEE ASSP Magazine, january 1986.

[2] Denis Jouvet; Modèles de Markov pour la reconnaissance de la parole, France, 1996. X. D. Huang, Y. Ariki, M. A. Jack; Hidden Markov Models for Speech Recognition, Edinburgh University Press, 1990.

[3] 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.

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

[5] 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.

[6] RABINER, L. Fundamentals of Speech Recognition, Prentice Hall Press, 1993.

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

[8] DAVI, S. ; MERMELSTEIN, P. Comparision of Parametric Representations for Monosyllabic Word Recognition. IEEE Trans. ASSP, vol. 28, pp 357-366, 1980

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

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

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

[12] 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.

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

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

SOBRE O AUTOR

Francisco José Fraga da Silva nasceu em São Paulo, em 25 de setembro de 1965. Formou–se em Engenharia Eletrônica com ênfase em Telecomunicações pela Escola Politécnica da Universidade de São Paulo (POLI-USP), em 1987. Foi pesquisador do Laboratório de Sistemas Integrados (LSI) da Escola Politécnica da Universidade de São Paulo em 1988. Concluiu o Mestrado em Engenharia Eletrônica pela mesma instituição, no ano de 1991. Durante o ano de 1992 foi professor colaborador do Instituto Tecnológico de Aeronáutica (ITA) de São José dos Campos – SP. Em 1998 recebeu o título de Doutor em Ciência pelo Instituto Tecnológico de Aeronáutica (ITA) de São José dos Campos – SP. Em 1999 desenvolveu *softwares* de Reconhecimento Automático de Fala para as empresas LG e Compo do Brasil., com resursos da lei de informática. Atualmente é Professor Adjunto em tempo integral do INATEL, onde desenvolve e ministra cursos nas áreas de Comunicação Digital e Processamento Digital de Sinais, nos níveis de Graduação, Pós-graduação *lato sensu* e Mestrado em Telecomunicações.

E-mail: fraga@inatel.br