UMA NOVA ABORDAGEM FONÉTICO-FONOLÓGICA EM SISTEMAS DE RECONHECIMENTO DE FALA ESPONTÂNEA

Rubem Dutra Ribeiro Fagundes e Ivandro Sanches

Resumo - O presente trabalho descreve um sistema de reconhecimento de fala de linguagem contínua, empregando uma abordagem fonético-fonológica, para extenso vocabulário. Os fundamentos teóricos e as técnicas de implementação apresentadas para o reconhecimento de padrões são as empregadas em estruturas *Hidden Markov Models* (HMM). Procura-se acentuar o enfoque lingüístico, destacando-se os conceitos, definições e correntes de pesquisa da Lingüística úteis e aplicáveis na construção de um sistema de reconhecimento de fala. O sistema proposto atua no nível fonético-fonológico, melhorando o desempenho final de reconhecimento de fala, através de estruturas fonéticas, ao elevar os escores de similaridades das combinações fonéticas mais freqüentes do idioma utilizado. Esta nova abordagem fonético-fonológica é altamente recomendada para sistemas de reconhecimento de fala espontânea.

Palavras-chave: Reconhecimento de Voz, Reconhecimento de Fala, Modelagem Fonética, Modelos de Markov HMM, Vocabulário Extenso, Linguagem Contínua, Fala Espontânea.

Abstract - This work describes a continuous speech recognition system, using a phonetic-phonological approach, covering a large vocabulary. The basic theory and implementation technique for the acoustic pattern recognition is the well-known Hidden Markov Models (HMM). Furthermore, the linguistic approach is enhanced, spotting the concepts, definitions and research lines in Linguistics, providing useful and applicable elements to build a large vocabulary speech recognition system. This system improves the speech recognition performance with phonetic-phonological modeling. The general likelihood scores are increased, getting a better recognition performance, due to the statistical phonetic structure enhancement of some frequent phonetic combinations usually found in the idiom. The utilization of this new phonetic-phonological approach is strongly recommended in spontaneous speech recognition systems.

Keywords: Speech Recognition, Phonetic Modeling, HMM, Large Vocabulary, Continuous Speech, Spontaneous Speech.

Rubem D. R. Fagundes atua no SISC – Laboratório de Sistemas e Computação, Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brasil. Ivandro Sanches atua no Genius Instituto de Tecnologia, Manaus, AM, Brasil.
E-mails: rubemdrf@attglobal.net, ivandro@genius.org.
Editor de Área responsável: Rui Seara. Artigo submetido em 06/Jul/1999, revisado em 05/Out/1999, 20/Ago/2002, 23/Out/2002, 13/Fev/2003, aceito em 14/Mar/2003.

1. INTRODUÇÃO

O texto a seguir será dividido em seções, nas quais serão feitas as descrições das etapas que compõem o processo de reconhecimento de fala empregado neste trabalho. Assim, o texto terá a seguinte estrutura:

O item 2 é dedicada à descrição do processo de comunicação pela fala, e sua forma de implementação de forma artificial em um sistema de comunicação Homem-Máquina. Neste item também se introduz as definições e conceitos dos Modelos de Markov do tipo Oculto (*Hidden Markov Models ou HMM*). Os itens 3, 4 e 5 são dedicados à Lingüística. Além de conceitos e definições, procura-se dar ênfase à descrição dos modos de aplicação e de representação do conhecimento fonético-fonológico na máquina. Os itens 6, 7 e 8 procuram combinar os conceitos e definições da Lingüística apresentados nas seções anteriores, introduzindo também os tipos de algoritmos de busca empregados nestes sistemas, descrevendo as estruturas dos sistemas de reconhecimento de fala de vocabulário grande. Neste sentido, estes itens são importantes por introduzir novos conceitos de base lingüística, bem como apresentar novas considerações quanto aos sistemas de reconhecimento de fala de vocabulário grande. Além disso, a proposta de um modelo fonético estatístico é também aqui apresentada.

A metodologia adotada, bem como os resultados obtidos são apresentados nos itens 9 e 10, promovendo um estudo comparativo entre os desempenhos obtidos por diferentes tipos de implementações realizadas neste trabalho.

2. PROCESSO DE COMUNICAÇÃO PELA FALA

O processo de comunicação vocal está longe de ser considerado como uma tarefa fácil. A capacidade desenvolvida pelo homem de estabelecer este modo de comunicação é única, no que diz respeito ao grau de complexidade das informações trocadas neste ato, e inerente ao *homo sapiens*, ao caracterizar sua diferença evolutiva enquanto espécie, pela sua forma de representação e veiculação de informação [1]. Assim, visto que o objetivo é o de efetuar uma interação vocal Homem-Máquina, é importante analisar as etapas desta comunicação, de forma a dominá-la e reproduzí-la, na medida do possível, de forma natural.

Uma vez que a fala é um processo de comunicação entre seres humanos, sua compreensão é importante para posteriormente transportá-lo, de forma artificial, para a interação Homem-Máquina. Na Figura 1 procura-se descrever o processo de comunicação vocal do Homem, tanto na etapa de transmissão, quanto na etapa de recepção da mensagem.

- No processo de recepção, os órgãos auriculares no **Nível Fisiológico** captam o sinal acústico. Nesta etapa,

deve-se considerar que o sinal acústico é composto, não só pelo sinal de fala do locutor, como também de todos os sinais acústicos existentes no ambiente. A fim de executar uma correta interpretação, exige-se uma etapa adicional de reconhecimento dos elementos acústicos pertinentes à mensagem. Tal etapa é realizada ao nível acústico.

- No **Nível Acústico**, todo o sinal acústico captado na etapa anterior é avaliado e todos os componentes acústicos da mensagem são discriminados. Assim, os padrões acústicos utilizados pela língua (usualmente chamados de fonemas da língua) são identificados na locução que está sendo ouvida.
- Ao **Nível Lingüístico**, os fonemas relacionados na fase anterior são agrupados e reconhecidos como as palavras que compõem a língua utilizada na comunicação pelo locutor e de domínio do ouvinte. Nesta fase a construção da sentença em palavras exigirá regras gramaticais e de sintaxe, bem como informações contextuais, para uma correta composição da sentença.
- Por fim, ao **Nível Conceitual**, a frase composta será interpretada, em conceitos que expressam o conteúdo informacional transmitido [2].

Evidentemente, vários são os fatores que influenciam o processo de transmissão e de recepção do sinal de fala, tanto fatores externos, como ruído ambiente, quanto fatores contextuais, como nacionalidade dos interlocutores, meio sócio-cultural, condições emocionais, entre outros [3].

O processo de transmissão envolverá as mesmas etapas apresentadas anteriormente, porém com uma seqüência de execução reversa, fazendo uso dos órgãos articuladores da fala.

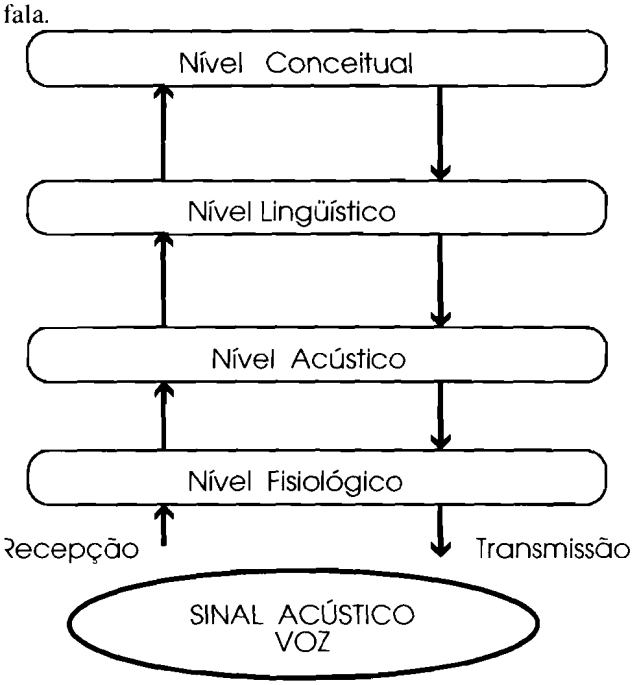


Figura 1. Processo de Comunicação Vocal.

2.1. SISTEMAS DE COMUNICAÇÃO HOMEM-MÁQUINA

Um sistema interativo de fala é um sistema automático de síntese/reconhecimento de fala em linguagem contínua que deverá desempenhar o processo de comunicação descrito no item anterior, executando etapas que, de forma artificial, estejam próximas das etapas naturais.

Assim, a estrutura geral de um sistema de síntese/reconhecimento de fala em linguagem contínua é apresentada no diagrama de blocos a seguir.

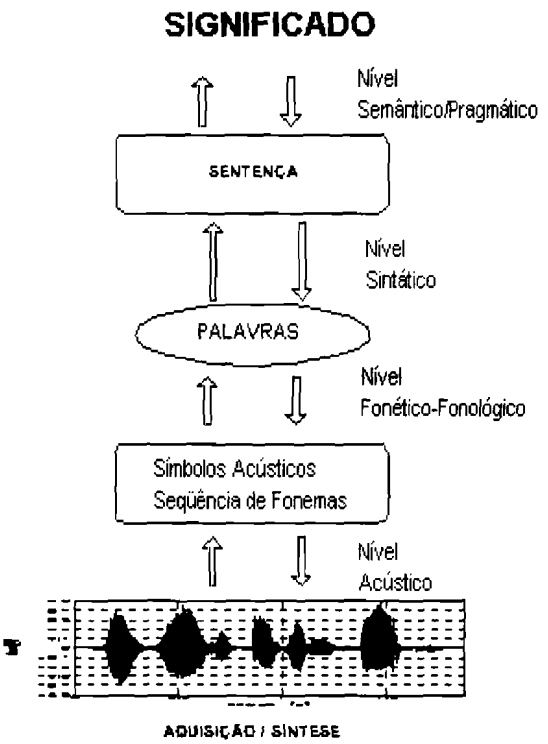


Figura 2. Estrutura de um sistema de síntese/reconhecimento de fala.

A Figura 2 apresenta os níveis do sistema interativo de fala, onde o processamento de cada etapa apresenta diferentes símbolos de representação da informação, dependendo da etapa de realização do processo global. Ainda que o objetivo deste trabalho seja o de estudar apenas um sistema de reconhecimento de fala, a Figura 2 representa também o processo de síntese de fala, de forma a descrever um sistema interativo de fala completo. Para uma descrição dos sistemas de síntese de fala, sugere-se [4-6].

2.1.1. NÍVEL ACÚSTICO

O processador acústico é responsável pela identificação dos padrões acústicos existentes no sinal de fala adquirido na entrada, convertendo-o em uma seqüência de símbolos acústicos que podem representar fonemas ou ainda subunidades fonéticas, isto é, padrões acústicos que compõem o fonema. Em sistemas de reconhecimento de fala de vocabulário pequeno, onde a mínima unidade de informa-ção a ser identificada é uma palavra, o processador acústico funciona como um bloco único com o processador sintático, não existindo o processamento lingüístico [7-8]. O processamento acústico realiza uma comparação de padrões, entre

o sinal de entrada e padrões de referência previamente armazenados, empregando a identificação de padrões por Modelagem Probabilística.

Modelagem probabilística é a técnica de uso de estruturas HMM em identificação de padrões de fala. Deste modo, no comparador de padrões usando estruturas HMM, o padrão de fala de teste é avaliado por modelos que procuram descrever a própria geração do padrão. Para determinar os modelos que fazem a descrição, ou ainda identificação do padrão, torna-se necessária uma abordagem por modelagem estocástica, estabelecendo um paralelo entre o comportamento de geração de locuções de fala, e modelos de processos estocásticos de parâmetros teoricamente definidos.

A identificação, deste modo, é um procedimento de determinação de verossimilhança, entre a seqüência de amostras do padrão de teste, e os modelos de Markov, de parâmetros definidos por padrões de referência em uma etapa de treinamento.

Não será meta deste trabalho apresentar uma extensa descrição matemática dos modelos de Markov HMM, sendo que para uma abordagem mais didática sugere-se [7] [9-11].

2.1.2. PARÂMETROS DE UM MODELO HMM

A Figura 3 apresenta um exemplo de estrutura HMM usualmente empregada em sistemas de reconhecimento de fala.

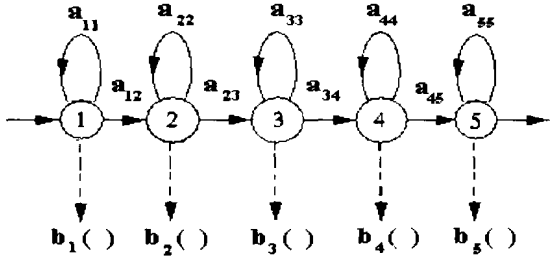


Figura 3. Exemplo de uma estrutura HMM.

Pode-se observar na Figura 3 que um HMM é definido por um conjunto de parâmetros, a saber:

Uma matriz de transição de estados A:

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \\ a_{N1} & \cdots & a_{NN} \end{bmatrix}, \text{ sendo } a_{ij} \geq 0,\ a_{ij} = P(q_j \mid q_i), \quad (1)$$

em que q_i, q_j são estados do modelo e a_{ij} é a probabilidade de transição para o estado j partindo do estado i. A matriz A será uma matriz quadrada de dimensão N^2, sendo que N é o número de estados.

Lembrando ainda que $\sum_{j=1}^{N} a_{ij} = 1$, para quaisquer $i = 1,2,\ldots,N$.

Um vetor de probabilidade inicial π:

$$\pi = [\pi_1 \quad \cdots \quad \pi_N], \quad (2)$$

sendo $\pi_i = P(q_i)$ a probabilidade de iniciar seqüência de estados percorridos na estrutura, pelo estado q_i. Os elementos π_i, são tais que

$$\sum_{i=1}^{N} \pi_i = 1.$$

Para um HMM *left-right*, como o da Figura 3, $\pi_1 = 1, \pi_{2,\ldots,N} = 0$, ou ainda $\pi = [1 \quad 0 \quad \cdots \quad 0]$.

Funções de probabilidade $b_j(y_t)$, de expressão:

$$b_j(y_t) = \sum_{m=1}^{M} c_{jm} \mathrm{N}(y_t, \mu_{jm}, \Sigma_{jm}), \quad 1 \leq j \leq N \quad (3)$$

sendo que:

y_t é um vetor de parâmetros acústicos resultante da etapa anterior de análise de sinal,

$\mathrm{N}(y_t, \mu_{jm}, \Sigma_{jm})$ é uma função densidade de probabilidade Gaussiana com

μ_{jm}, vetor média e Σ_{jm}, matriz de covariância,

c_{jm} são chamados *mixture coefficients*, isto é coeficientes de "mistura", ou ainda de combinação linear das funções gaussianas $\mathrm{N}(y_t, \mu_{jm}, \Sigma_{jm})$. Assim, cada função de probabilidade $b_j(y_t)$, é composta por um número M previamente escolhido de funções gaussianas, combinadas segundo c_{jm}, que deve satisfazer a restrição

$$\sum_{m=1}^{M} c_{jm} = 1, \text{ para } 1 \leq j \leq N,\ 1 \leq m \leq M \text{ e } c_{jm} \geq 0.$$

Com respeito ainda aos elementos da função densidade de probabilidade gaussiana, é usual assumir que não existe uma correlação entre os coeficientes do vetor acústico y_t, então a matriz de covariância Σ_{jm} será uma matriz diagonal[1].

3. NÍVEL FONÉTICO - FONOLÓGICO

O processamento fonético-fonológico fará a montagem das palavras encontradas na locução, a partir da seqüência de fonemas identificados pelo processador acústico. Para tanto, cada palavra que compõe o vocabulário da linguagem deverá ser decodificada na seqüência de fonemas constituintes de sua locução. As formas de representação destas "transcrições fonéticas[2]" comumente empregadas são as bases de dados e os grafos fonéticos.

As bases de dados, também chamadas dicionários fonéticos, são como bases de conhecimento que relacionam o vocabulário representado como grafema (isto é, uma palavra escrita, como "casa") com a transcrição fonética correspon-

[1] De fato, isto não é verdade. No entanto, verifica-se experimentalmente que é melhor utilizar uma matriz de covariância diagonal, e várias funções gaussianas, do que poucas funções gaussianas e uma matriz de covariância plena.

[2] Termo empregado pela Lingüística para definir a representação de uma palavra por sua seqüência fonética.

dente (como k/a/z/a)[2][12]. O exemplo a seguir descreve este tipo de representação.

Palavra	*Transcrição I*	*Transcrição II*
CASA	[kaza]	
CARNE	[kaɾni]	[kaxni]
CANETA	[kaneta]	[kanyeta]
COURO	[kowɾu]	[koɾu]
CORDA	[kɔɾda]	[kɔɽda]
CORPO	[koɾpu]	[koɦpu]
LATIR	[latiɾ]	[latix]
LARGO	[laɾgu]	[laxgu]

Tabela 1: Exemplo de Dicionário Fonético.

No exemplo, existem até duas transcrições possíveis para cada palavra [13][3]. Na prática podem ser usadas várias transcrições, dispostas por colunas segundo critérios referenciais que incluem variantes regionais (sotaque regional), variantes sócio-culturais, variantes específicas do falante (nos casos de sistemas dependentes do locutor ou ainda que empregam alguma descrição auxiliar para classificação de tipos de locutor), entre outros.

Um problema freqüente com este tipo de representação é o de que as transcrições fonéticas são repetidamente apresentadas, mesmo quando existem palavras no vocabulário que possuem grupos de fonemas similares. Tal tipo de representação é usualmente chamada de representação linear, visto que para cada palavra do vocabulário é apresentada uma seqüência linear de fonemas. Além do problema de espaço de armazenamento, esta representação consome muito tempo de processamento, dado que, para cada palavra, o algoritmo de busca fará uma busca seqüencial exaustiva de cada grupo de fonemas repetidos. Além disso, recentes pesquisas têm reportado que um percentual elevado de palavras compartilham da mesma seqüência inicial de fonemas [14]. A forma de representação por grafos fonéticos procura solucionar este problema.

Os grafos fonéticos apresentam a transcrição fonética das palavras como uma estrutura em árvore, onde um fonema é representado por (***ni,f,nf***), sendo ***ni*** o nó inicial, ***f*** o fonema e ***nf*** o nó final [15]. Tal estrutura é apresentada na Figura 4.

A estrutura da Figura 4 apresenta as descrições fonéticas usadas no exemplo anterior. Os nós terminais da árvore (no exemplo os nós 5, 8, 14, 16, 17, 19, 22 e 26) representam o final da busca e a obtenção da palavra do vocabulário. Para tanto, existirá uma tabela relacionando os nós terminais às palavras do vocabulário, que para este está mostrada na Tabela 2.

Várias abordagens diferentes usando grafos fonéticos são hoje pesquisadas para o processamento lingüístico, onde cada grafo sugerido utilizará um algoritmo de busca especificamente adaptado para o grafo proposto [14][16].

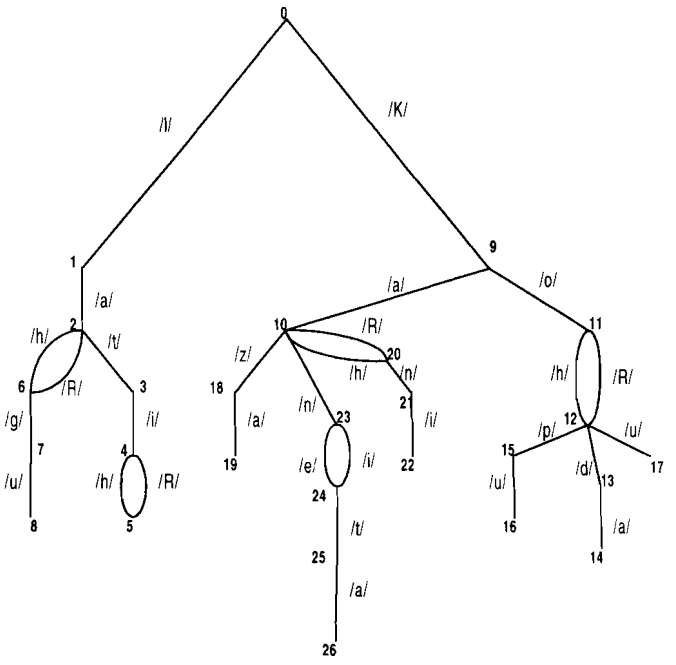


Figura 4. Exemplo de grafo fonético.

Estado Terminal	*Palavra*
5	LATIR
8	LARGO
14	CORDA
16	CORPO
17	COURO
19	CASA
22	CARNE
26	CANETA

Tabela 2: Exemplo de relação nó terminal-vocabulário.

4. NÍVEL SINTÁTICO

De forma similar ao processamento lingüístico, o processador sintático possui regras gramaticais para construção de sentenças válidas[4], que são representadas freqüen-temente por uma estrutura sintática que define, *a priori*, a estrutura geral da sentença. A Figura 5 exibe um exemplo [7].

No exemplo da Figura 5 a estrutura sintática impõe a seqüência de palavras válidas, orientando a busca executada pelo processador lingüístico. Este tipo de sistema é denominado "orientado a comandos", e funciona bem em sistemas de vocabulário e sintaxe definidos em um ambiente específico para aplicação [7] [17-18] onde, neste exemplo, o ambiente proposto é o de automatização de tarefas bancárias.

[3] Exemplo elaborado com a contribuição de [13].

[4] Tais regras são chamadas de modelo de linguagem.

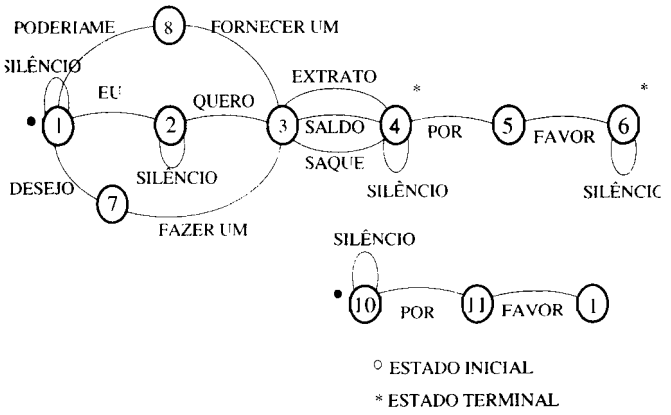


Figura 5. Exemplo de estrutura sintática.

4.1. MODELAMENTO ESTOCÁSTICO: "BIGRAMAS E TRIGRAMAS"

Este modelo de linguagem é usualmente empregado em sistemas de reconhecimento de fala de vocabulário grande. Proposto inicialmente por Jelinek, Bahl e Mercer [19], esta abordagem utilizará medidas de probabilidades de utilização das palavras do vocabulário, obtidas a partir de levantamento estatístico do *corpus* escolhido *a priori* para teste e avaliação do reconhecedor.

Assim, o modelo de linguagem não é uma topologia ou rede conectando as palavras do vocabulário, mas sim uma tabela relacionando as palavras do vocabulário e as suas probabilidades de ocorrência em seqüência numa sentença falada.

Tomando a seqüência W de palavras, tal seqüência é da forma [20]

$$W = w_1, w_2, w_3, ..., w_n . \tag{4}$$

A expressão da probabilidade de ocorrência da seqüência W, $P(W)$ é dada por

$$P(W) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2) \cdot P(w_4 \mid w_1 w_2 w_3) \cdots P(w_n \mid w_1 w_2 w_3 \cdots w_{n-1}) \tag{5}$$

ou ainda

$$P(W) = \prod_{i=1}^{n} P(w_i \mid w_1 ... w_{i-1}) . \tag{6}$$

Percebe-se da expressão (6) que o cálculo de $P(W)$ é intratável, mesmo para um vocabulário pequeno, quanto mais para aplicações de vocabulário grande.

Assim, uma abordagem para contornar o problema é usar o cálculo das probabilidades de ocorrências das seqüências mais recentes, isto é, apenas as últimas palavras da história da seqüência serão consideradas no cálculo das probabilidades do modelo de linguagem. Definindo-se o número de palavras recentes, pode-se calcular as probabilidades condicionais para duas palavras $(n = 2)$, chamados de ***bigramas*** e para três palavras $(n = 3)$, chamados de ***trigramas***. Deste modo, um modelo de linguagem com bigramas é tal que possui calculadas, para cada palavra do vocabulário, a probabilidade de ocorrência desta palavra e as probabilidades condicionais desta palavra, dado que já ocorreram as outras palavras do vocabulário. As expressões (7) e (8), a seguir, ilustram estes modelos de linguagem [20-21]:

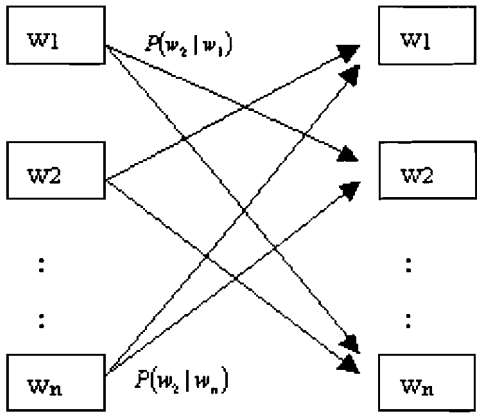
bigramas:

$$P(w_n \mid w_1 w_2 w_3 ... w_{n-1}) \cong P(w_n \mid w_{n-1}) \tag{7}$$

trigramas:

$$P(w_n \mid w_1 w_2 w_3 ... w_{n-1}) \cong P(w_n \mid w_{n-2} w_{n-1}) . \tag{8}$$

A Figura 6 ilustra um grafo de modelo de linguagem tipo bigrama:



Fgura 6. Grafo de modelo de linguagem bigrama.

4.2. ALGORITMO DE PESQUISA

O processo de pesquisa, também chamado de processo de busca, combina as estruturas fonético-fonológicas empregadas para a representação das palavras do vocabulário com os modelos de linguagem, quer seja a sintaxe fixa, quer sejam modelos estatísticos gerando ao final a seqüência de palavras que apresentaram um escore de máxima similaridade. Vários algoritmos podem ser propostos para a execução desta tarefa. Os mais simples deles são o de busca irrestrita, que se subdividem em busca em profundidade *deep search* e busca em largura *breadth search*.

Infelizmente, o emprego de um algoritmo de busca irrestrita não apresenta bom resultado em tempo hábil, uma vez que o número de seqüências de palavras possível cresce exponencialmente com o número de palavras existentes. Assim, algoritmos de programação dinâmica[5] (DP) são extensivamente aplicados em sistemas de reconhecimento de fala, onde variações destes algoritmos são freqüentemente utilizadas, tanto em sistemas de palavras isoladas quanto em linguagem contínua. O motivo mais freqüente para o emprego destes algoritmos reside na velocidade de processamento, uma vez que podem ser executados na medida em que as amostras do sinal estão sendo coletadas, bem como na pequena quantidade de memória exigida para seu funcionamento [22].

[5] *Dynamic Programming.*

Os algoritmo mais empregado é o algoritmo de Viterbi, considerado o algoritmo clássico para sistemas de reconhecimento de fala.

4.3. ALGORITMO DE HERMAN NEY

O algoritmo de Herman Ney (HN), é basicamente o algoritmo um-estágio (*One-step*[6]) que tratará o reconhecimento através do uso de uma representação fonética do vocabulário, e executará a identificação e posterior montagem das palavras e da sentença.

Deste modo, o algoritmo possui características distintas, a saber:

- Uma árvore de representação fonética do vocabulário.
- O algoritmo HN executa a busca numa estrutura em árvore, chamada de "árvore léxica"[7] na qual todo o vocabulário está representado com suas transcrições fonéticas, e os nós finais detêm os índices referentes a cada palavra do vocabulário. Esta árvore deve ser especialmente construída, mapeando com apenas um caminho seqüências comuns de fonemas existentes em mais de uma palavra. Tal técnica visa reduzir o espaço de busca, como visto no início deste capítulo. A Figura 7 exemplifica a árvore léxica. Cabe notar que este grafo é essencialmente similar ao da Figura 4.

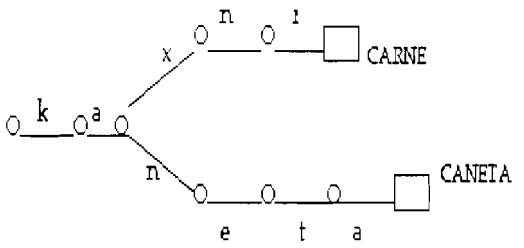


Figura 7. Exemplo de árvore léxica.

Deve-se observar que, com o uso de uma árvore léxica, o algoritmo HN somente determina *a posteriori* qual a palavra identificada. Em outras palavras, a árvore léxica possui um mesmo caminho para as transcrições fonéticas iniciais comuns, pertencentes a mais de uma palavra do vocabulário. Assim, somente quando a busca alcança um nó final é que se verifica qual palavra do vocabulário foi identificada.

5. NÍVEL SEMÂNTICO/PRAGMÁTICO

De posse das palavras encontradas na sentença, bem como de quaisquer informações adicionais fornecidas pelos processos anteriores, o processador semântico/pragmático poderá usar de várias formas de abordagem diferentes, a fim de extrair o significado ou ainda informação contida na locução analisada. As abordagens mais comuns incluem bases de conhecimento, redes semânticas e linguagens descritoras, dentre outros [2] [26-27].

Esta tarefa exige uma extensa base de dados, bem como algoritmos de busca que efetuem a correlação das informações disponíveis na análise, e também conhecimento de senso comum inerente ao falante. Um tratamento detalhado deste problema pode ser encontrado em [28].

Nos tópicos anteriores foram apresentados os aspectos referentes às definições da Lingüística utilizadas em sistemas de reconhecimento de fala, bem como os algoritmos de busca freqüentemente empregados. Aqui, procura-se apresentar outros aspectos importantes e não ainda abordados dos sistemas de reconhecimento de fala que empregam representação fonética.

6. FILOSOFIAS LINGÜÍSTICAS GERATIVISTAS/ EMPIRISTAS

Até o presente momento, a Lingüística tem oferecido os elementos e definições necessários para a manipulação da língua, visto ser essencialmente a sua área de estudo. Deste modo os aspectos lingüísticos apresentados neste trabalho procuraram mostrar, tanto no plano da estrutura gramatical/sintática, quanto no plano fonético/fonológico, a língua como descrição de uma estrutura de regras e relacionamentos pré-definidos. No entanto, mesmo na Lingüística como ciência, o pensamento de que a língua possui uma estrutura inata de regras não é unanimemente aceito pelos pesquisadores. Neste sentido, duas são as correntes de pensamento, ou ainda filosofias, aceitas pelos lingüistas: Gerativista e Empirista.

Por um lado, segundo o lingüista Chomsky e sua "hipótese de competência" (Apud Bresnan, T. e Kaplan,R. M.) [29]:

"(...) a reasonable model of language use will incorporate, as a basis component, the generative grammar that expresses the speaker-hearer's knowledge of the language (...)"

Assim, esta seria uma corrente de pensamento mais racionalista da Lingüística e que declara que a língua possui uma estrutura gramatical gerativa-transformacional, isto é, uma estrutura de regras que descreve o comportamento da língua bem como a sua evolução e transformação desde as línguas mais antigas praticadas pelo homem (começando pela língua raiz de todas - o indo-europeu, e passando pelas línguas como o grego, o latim e o eslavo, tidas como dialetos distintos do indo-europeu) até a língua praticada pelos falantes no momento atual. Além disso, tal modelo de linguagem representaria a forma como é representado o conhecimento da língua no plano do falante-ouvinte.

Seguindo ainda esta filosofia, existiriam níveis em que estas regras se aplicariam: um nível genérico, onde as regras propostas regeriam a estrutura das sentenças da língua, e um nível mais específico, descrevendo as alterações ao nível fonológico das palavras segundo a evolução, ou ainda transformação, subseqüente no processo de evolução da língua.

No entanto, e infelizmente, além desta pesquisa não ser uma tarefa fácil, a busca de tais regras tem sido por muitos considerada por vezes não possível, visto que a observação

[6] O algoritmo *One-step* é uma forma de implementação do Algoritmo de Viterbi e não será descrito neste texto. Para uma apresentação completa sugere-se [23-25].

[7] *Lexical Tree.*

prática, bem como os modelos lingüísticos existentes, têm falhado na descrição de muitos aspectos experimentalmente encontrados nas línguas. Em outras palavras, ainda que se aceite a existência de um certo conjunto de regras, por serem estas verificáveis experimentalmente, parece ser altamente improvável que elas realmente representem e descrevam com fidelidade a evolução da língua, desde o passado até os nossos tempos.

Além disso, os pesquisadores da psicolingüística têm com freqüência observado que os modelos utilizados pelos falantes para a representação da língua não correspondem aos modelos empregados pela Lingüística para descrever a língua. Assim, existe uma outra corrente de pensamento que procura mostrar que a língua não evoluiu e nem é regida por um conjunto inato de regras, mas sim se desenvolve a partir das experiências empíricas e da descrição da realidade adotada pelos falantes ao longo do tempo [29]. Em outras palavras, de fato não existiriam regras ditando a estrutura bem como a transformação da língua, mas sim que a língua seria uma conseqüência da representação da realidade mental dos falantes e então conseqüência direta de suas experiências empíricas. Compreensivelmente, esta é uma posição conhecida como empirista. De modo geral, os empiristas declaram que o estudo da língua exigiria modelos estatísticos bem como estudos mais profundos de psicolingüística, a fim de se determinar as formas de representação existentes na mente dos falantes [29-30].

Do ponto de vista dos pesquisadores que trabalham com reconhecimento de fala, tanto a abordagem gerativista quanto a abordagem empirista têm apresentado bons frutos, ao menos no plano gramático/sintático. Assim, os sistemas que impõem estruturas sintáticas, bem como regras gramaticais para a construção e análise das palavras de uma sentença, atendem a uma abordagem lingüística gerativista, enquanto os sistemas que empregam modelos de linguagem estatísticos com bigramas e trigramas atendem a uma abordagem lingüística empirista.

Entretanto, nas formas de representação fonético-fonológicas os sistemas de reconhecimento de fala ainda assumem apenas a abordagem gerativista, dado que na grande maioria destes sistemas procurar-se-á representar a estrutura fonética das palavras de acordo com regras acústico-articulatórias e estabelecendo relações entre os padrões acústicos com base nestas estruturas de regras. Assim, dadas as duas abordagens concorrentes na Lingüística, é compreensível presumir-se que uma abordagem empirista aplicada na representação fonético-fonológica poderia também render um bom desempenho. Em outras palavras, a proposta de modelos fonéticos estatísticos, isto é empregando bigramas e trigramas, mas agora ao nível fonético, pode conduzir a uma melhoria nos sistemas de reconhecimento de fala que empregam representação fonética.

Por último, o ponto principal que se procura seguir neste trabalho é o de assumir que ambas as correntes de pensamento praticadas pela Lingüística possuem elementos importantes a ser aplicados nos sistemas de reconhecimento de fala, pois este é o tópico principal deste trabalho e, sendo assim, não cabe e não será feito qualquer julgamento sobre a questão "empirismo versus gerativismo", visto que a aplicação destas filosofias no plano da máquina (que não possui psiquê e muito menos pratica a língua desde os tempos mais remotos de nossa história) não tem sentido. Assim, na máquina, deve-se aplicar toda a teoria, quer seja lingüística, matemática, computacional, ou ainda qualquer outra teoria cabível, a fim de resolver o problema real: o do reconhecimento da fala humana.

7. ABORDAGEM *TOP-DOWN* E *BOTTOM-UP*

O processamento computacional de um sistema de reconhecimento de fala é de fato o processamento de um algoritmo de busca (como visto anteriormente), sobre uma base de dados que relaciona fonemas e palavras, de acordo com os tipos de filosofias lingüísticas existentes, bem como as formas de representação fonético-fonológicas apresentadas no item 3. No entanto, para tal processamento, duas abordagens podem ser definidas: *Top- Down* e *Bottom-Up*. A Figura 8 exemplifica as duas abordagens.

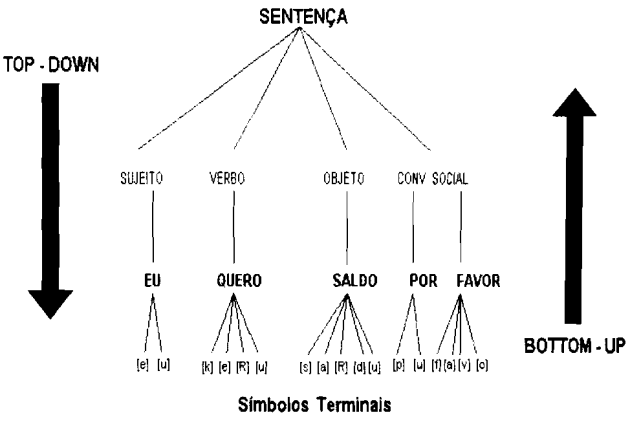


Figura 8. Exemplo das abordagens top-down e bottom-up.

O processo de busca executado pelo reconhecedor pode proceder a identificação acústica dos padrões, relacioná-los às palavras segundo um modelo de linguagem ou rede sintática, e ao final combinar e montar a sentença. Deste modo, o reconhecimento começa pelos símbolos terminais da linguagem, ou seja os fonemas, ou ainda quaisquer sub-unidades acústicas utilizadas pelo reconhecimento acústico, e ao final monta a seqüência de palavras, ou ainda a sentença que apresenta maior semelhança com a locução de entrada.

Este tipo de procedimento é chamado de *bottom-up*, uma vez que realiza o caminho **símbolos terminais → sentença**, partindo de sub-unidades acústicas até o reconhecimento final da sentença.

Os sistemas que utilizam abordagem *top-down*, por outro lado, realizam o caminho **sentença → símbolos terminais**. Neste procedimento, o algoritmo de busca gera hipóteses das sentenças aceitáveis, de acordo com um dado modelo de linguagem e a seguir realiza o reconhecimento de padrões acústicos para validar aquela hipótese que melhor estima a sentença contida na locução de fala.

O diagrama de blocos de um sistema de reconhecimento de fala que emprega modelagem fonética pode ser visto a seguir.

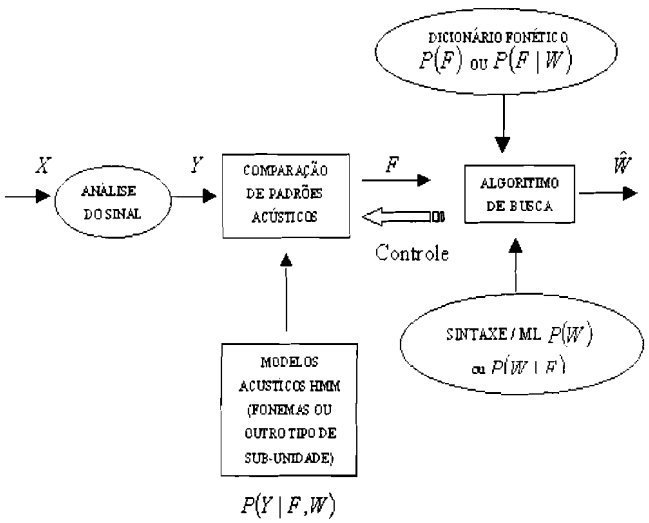


Figura 9. Diagrama de blocos de sistemas de reconhecimento de fala que utilizam representação fonética.

Neste caso, o processo de reconhecimento pode ser descrito por uma das duas expressões [23]:

$$\hat{W} = \arg\max_{W} P(Y \mid F,W)P(W \mid F)P(F) \qquad (9.I)$$

ou

$$\hat{W} = \arg\max_{W} P(Y \mid F,W)P(F \mid W)P(W) \qquad (9.II)$$

As expressões (9.I) ou (9.II) serão válidas, dependendo do processo executado pelo algoritmo de busca.

Se o algoritmo de busca efetua hipóteses de seqüência de palavras, antes de proceder a busca na rede fonética (como visto em sistemas *top-down*), então W é gerado antes de F e a expressão (9.II) deve ser usada, dado que o modelo de linguagem $P(W)$ orientará a formação da seqüência de palavras W. O termo $P(F \mid W)$ representa um modelo de relacionamento ou ainda uma estrutura fonético-fonológica, que relaciona as possíveis seqüências de fonemas F (ou, de outro modo, as possíveis formas de pronúncia existentes para as palavras que compõem a seqüência W) com a seqüência de palavras W do vocabulário do sistema. Usualmente esta estrutura é o dicionário fonético, apresentando a descrição fonética de cada palavra do vocabulário.

No entanto, se o algoritmo de busca efetua a análise da locução de entrada identificando os padrões acústicos existentes, e a partir dos fonemas identificados monta as palavras da seqüência W, isto é, o sistema determina F antes de W, (como visto em sistemas *bottom-up*), então a expressão (9.I) deverá ser usada e $P(W \mid F)$ orientará a construção da seqüência de palavras W. $P(F)$ será uma estrutura em árvore como vista na Figura 4 e que identificará as palavras a partir da identificação seqüencial dos fonemas.

Pode-se então perceber que, ao impor qualquer tipo de estrutura visando organizar segundo algum critério ou conjunto de regras, ou ainda representar o conhecimento léxico, isto é, relacionando palavras e fonemas, estamos na verdade introduzindo (ainda que até aqui de forma não clara) o termo $P(F)$ ou $P(F \mid W)$, que como demonstrado, está presente e é necessário. Deve-se também notar que as formas de implementação de $P(F)$ ou $P(F \mid W)$, como os dicionários fonéticos e os grafos em árvore, são usualmente estruturas fixas ou ainda estáticas, similares às estruturas sintáticas empregadas ao nível das palavras.

8. MODELAGEM FONÉTICA ESTATÍSTICA

Como visto nos itens anteriores, a abordagem utilizada para representação das estruturas fonético-fonológicas atende à filosofia lingüística gerativista, isto é, organiza-se as estruturas de relacionamento fonético-fonológico, segundo um conjunto de regras fixas tais como os dicionários fonéticos (que possuem as transcrições fonéticas de cada palavra do vocabulário) e os grafos fonéticos. No entanto, cabe aqui propor uma nova forma de representação, baseada nas formas já empregadas de representação de estruturas de linguagem, notadamente os modelos de linguagem estatísticos apresentados anteriormente. Deste modo utiliza-se, ao nível fonético, a abordagem filosófica lingüística empirista, introduzindo-se elementos probabilísticos que traduzem melhor as variações e diferenças existentes entre as formas estáticas de pronúncias ditadas pelos dicionários fonéticos, e as formas de pronúncia encontradas na execução livre da língua.

Propõe-se então um modelo de linguagem fonético estatístico, onde a freqüência das seqüências de fonemas seriam caracterizadas agora por bigramas e trigramas fonéticos. Esta nova estrutura estatística é que determinaria a organização, ou ainda o modelo fonético do reconhecedor, orientando assim, o algoritmo de busca, durante o processo de reconhecimento. Nesta nova abordagem, o termo $P(F)$ ou $P(F \mid W)$ teria uma expressão estatística, a ser apresentada a seguir:

Tomando-se uma seqüência de fonemas

$$F = f_1, f_2, \ldots, f_n ,$$

o cálculo de um bigrama fonético seria da forma:

bigramas fonéticos:

$$P(f_n \mid f_1 f_2 f_3 \ldots f_{n-1}) \cong P(f_n \mid f_{n-1}) \qquad (10)$$

sendo o cálculo de $P(f_n \mid f_{n-1})$ dado por

$$P(f_2 \mid f_1) = \frac{C(f_2 f_1)}{C(f_1)} \qquad (11)$$

em que $C(\bullet)$ é a contagem do número de vezes que a seqüência de fonemas apareceu durante as transcrições fonéticas dos textos usados na etapa de treinamento.

De forma similar os trigramas fonéticos são da forma:

trigramas fonéticos:

$$P(f_n \mid f_1 f_2 f_3 \ldots f_{n-1}) \cong P(f_n \mid f_{n-2} f_{n-1}) \qquad (12)$$

e $P(f_3 \mid f_2 f_1)$ é calculado por:

$$P(f_3 \mid f_2 f_1) = \frac{C(f_3 f_2 f_1)}{C(f_1 f_2)} . \qquad (13)$$

Visto que a quantidade de fonemas é pequena em comparação com a quantidade de palavras empregadas nos modelos de linguagem, um *corpus* de dados suficientemente grande fornecerá os elementos necessários para o cálculo dos bigramas e trigramas fonéticos, sem necessidade de interpolações. Mais ainda, numa primeira etapa, durante o treinamento, a construção de um segmentador fonético e sua posterior utilização, poderia fornecer as seqüências realmente encontradas nas locuções utilizadas no treinamento, revelando assim as freqüências das seqüências fonéticas realmente existentes durante a fala e não as fornecidas por um dicionário fonético.

Além disso, as poucas seqüências de fonemas não encontradas durante o cálculo, denotam seqüências de fonemas não existentes na língua e, portanto, não devem ser consideradas na estrutura fonética. Nestes casos, deve-se impor manualmente valores próximos de zero às probabilidades destas seqüências, forçando posteriormente o algoritmo de busca a abandonar esta seqüência durante o processo de reconhecimento, pois, de resto é uma seqüência errada, isto é, tal seqüência de fonemas não ocorre no idioma do sistema. Em outras palavras, esta estratégia garante uma maior robustez durante o processo de reconhecimento.

8.1. APLICAÇÃO DO MODELO FONÉTICO ESTATÍSTICO

Durante a exposição anterior, uma pergunta pode ser apresentada: qual a finalidade de impor um modelo estatístico no plano fonético, quando a própria seqüência fonética definida no dicionário já conduz o reconhecedor na identificação das palavras? Ou seja, uma vez que durante o processo de reconhecimento, o algoritmo mantém as palavras que possuem máxima similaridade com a sentença acústica de entrada, esta mesma palavra possui, pelo dicionário, uma seqüência fonética definida, tornando o uso de uma nova estrutura fonética, em princípio, redundante.

Todavia, se isto fosse realmente verdade, não ocorreriam substituições ou inserções[8] de palavras no reconhecimento, pois a identificação da seqüência fonética conduziria à identificação das palavras da sentença de forma inequívoca.

Assim, ao que parece, a identificação de uma seqüência fonética não implica no reconhecimento correto de todas as palavras do vocabulário. Isto ocorre devido aos seguintes pontos:

- Como já discutido, as palavras possuem seqüências fonéticas comuns, o que dificulta, para um dado vocabulário extenso, a seleção da palavra que mantém o escore de máxima similaridade. A utilização de uma árvore léxica, tende a reduzir o problema. No entanto, uma árvore léxica, como apresentada no sistema Herman Ney, é constituída apenas com seqüências comuns de um ou dois fonemas, mas é frequente existirem mais de dois fonemas comuns, (exigindo talvez o emprego de seqüências maiores), e que estarão no restante da seqüência fonética referenciada à palavra.

-Os parâmetros dos modelos fonéticos HMM são dependentes das amostras utilizadas durante o treinamento. Assim, mesmo empregando técnicas de normalização dos escores de estimação calculados, existirão diferenças nos parâmetros HMM que determinarão uma tendência durante as características das amostras utilizadas para treinamento, bem como a disponibilidade de amostras para treinar cada modelo fonético, os modelos de HMM apresentarão diferentes escores de comparação, onde tais diferenças são independentes do padrão acústico, e diretamente relacionadas com a etapa de estimação.

Para exemplificar estes pontos, considere a frase a seguir:

EU QUERO CARNE, TÁ?

Para este exemplo, vamos avaliar o instante em que o reconhecedor está processando o trecho de sentença de teste relativo à palavra CARNE. Uma representação fonética possível no dicionário fonético seria:

CARNE
CANETA

Ou ainda segundo uma árvore léxica com uma seqüência comum de dois fonemas:

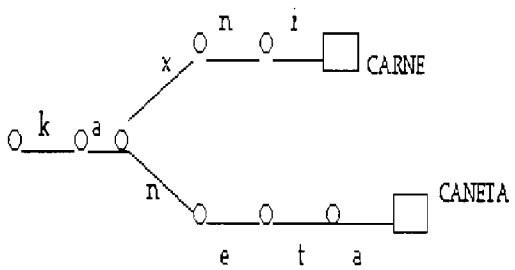


Figura 10. Representação por árvore léxica.

Assim, para uma correta identificação da sentença, o fonema / / deverá apresentar um escore, acumulado durante o reconhecimento, maior do que o fonema / /ẽ No entanto, devido às características intrínsecas do padrão acústico, bem como aos problemas devidos à estimação já discutidos, o fonema /n/ poderá apresentar um escore de similaridade acumulado maior, conduzindo o reconhecedor para a sen-tença.

EU QUERO CANETA.

ocorrendo um erro de substituição.

Outro ponto importante a se considerar aqui é o de que mesmo o emprego de uma representação fonética dependente de contexto, isto é, com o uso de bifones ou trifones, não resolveria o problema. Isto porque tal abordagem permite resolver problemas de co-articulação, mas devido à escassez de dados, os problemas de estimação dos parâmetros HMM dos modelos bifonéticos ou trifonéticos permaneceria.

Além disso, pode-se considerar este tipo de representação fonética contextual como uma descrição "estática", visto que representa um dado fonema central com os fonemas

[8] Dois tipos de erros de ocorrência freqüente. Um erro de inserção é dito quando o reconhecedor inclui palavras não presentes na sentença original. Um erro de substituição é dito quando o reconhecedor troca uma palavra por outra.

vizinhos, mas não considera a seqüência de ocorrência, no idioma, de tal combinação.

Deste modo, a principal aplicação de um modelo fonético estatístico é introduzir uma informação adicional sobre o comportamento fonético da língua. Este novo fator adicional contribuiria para melhor orientar a decisão durante o processo de reconhecimento, visto que tal informação seria independente de uma palavra específica do dicionário, sendo resultado direto da avaliação fonética da língua como um todo. Em outras palavras, a probabilidade fonética condicional apresentaria as condições fonéticas freqüentes na língua e, quando empregadas no processo de reconhecimento, conduziriam ao "desempate" entre duas palavras com seqüências fonéticas similares. No exemplo, as probabilidades condicionais fonéticas *P*(/ /I/ /) e *P*(/ /I/ /), empregadas adicionalmente no cálculo do escore acumulado, poderiam contribuir para o desempate entre as palavras CARNE e CANETA, resultando no reconhecimento correto da sentença.

Cabe lembrar aqui que em nenhum momento se despreza ou ainda se desconsidera as probabilidades condicionais das palavras, ou seja, o modelo de linguagem. Como visto na expressão (9.I e 9.II) uma estrutura sintática, ou modelo de linguagem sempre estará presente, assim como estrutura fonético-fonológica que agora apresentará um detalhamento estatístico.

Ainda que o exemplo apresentado não tenha sido suficientemente completo para demonstrar a utilidade de um modelamento fonético estatístico, pode-se perceber a contribuição desta abordagem no desempenho final de um reconhecedor de fala de vocabulário extenso.

9. METODOLOGIA

O presente trabalho foi desenvolvido nas instalações e máquinas do grupo de pesquisa GREP (Groupe de Reconaissance Automatique de la Parole, ou ainda Speech Recognition and Understanding Group) do CRIM - Centre de Recherche Informatique de Montreal, Canadá, e no Laboratório de Processamento de Sinais da Escola Politécnica da Universidade de São Paulo.

Para implementação e testes dos algoritmos, bem como treinamento do sistema, foram utilizadas as seguintes máquinas: estações de trabalho SUN SPARC 10, SPARC 20 e ULTRA 1, sobre sistema operacional UNIX, SUN e SOLARIS.

Os recursos de *software* utilizados para programação e treinamento foram o *kit* de desenvolvimento *HTK* (*HMM ToolKit*) da empresa Entropics Inc. e a biblioteca de programação de interfaces para aplicações, *HAPI* (*HTK Application Programming Interface*) também distribuído pela Entropics como parte da versão 2.1 do HTK.

Para treinamento e testes do sistema foi utilizado o *corpus* de língua inglesa *SWITCHBOARD* distribuído pelo LDC - *Linguistic Data Consortium*.

Os tópicos a seguir apresentarão em maior profundidade estes recursos, visando a obtenção de um sistema de reconhecimento de fala de linguagem contínua, para vocabulário grande. Tal sistema emprega representação do vocabulário por meio de uma árvore léxica, algoritmo Herman Ney para execução do reconhecimento das palavras da sentença, e ainda um modelo fonético estatístico, proposta deste trabalho, para aumentar o desempenho do reconhecedor.

9.1. FERRAMENTA DE DESENVOLVIMENTO HTK

O *kit* HTK é um conjunto de módulos projetados para o desenvolvimento de um sistema de reconhecimento de fala que empregue modelos HMM. Tais módulos são independentes e podem ser utilizados tanto para as tarefas de treinamento quanto para as de teste e avaliação. Além disso, o HTK pode ser utilizado para o desenvolvimento de diversos tipos de sistemas de reconhecimento de fala, como os de palavras isoladas, linguagem contínua, com sintaxe fixa, com modelos de linguagem, com modelagem fonética dependente de contexto ou ainda sistemas de identificação de locutor.

O *kit* possui também recursos de leitura de arquivos de sinais em vários formatos diferentes, e ainda visualização e edição gráfica dos sinais de fala. Assim, mesmo que o *kit* HTK não seja uma ferramenta computacional do tipo "amigável", exigindo que o usuário dedique algum tempo para o aprendizado de sua utilização, este *kit* é um recurso completo para o desenvolvimento de sistemas de reconhecimento de fala.

9.1.1. ARQUIVOS DE ENTRADA E DE CONFIGURAÇÃO

O HTK utiliza diferentes tipos de arquivos de entrada para cada módulo do *kit*. Tais arquivos de entrada podem, com freqüência, ser editados pelos próprios módulos do HTK, na etapa de preparação de dados.

Buscando uma implementação integrada, o HTK utiliza um único arquivo de configuração (usualmente de nome HConfig), que descreverá as informações pertinentes de cada etapa, tanto no treinamento quanto no teste e avaliação para todos os módulos do sistema. Deste modo, uma vez definidas as características gerais do sistema a ser implementado no arquivo de configuração, os resultados obtidos por cada módulo do sistema podem ser intercambiados, como arquivo de entrada de algum outro módulo, sem problemas de compatibilidade.

9.2. CONFIGURAÇÃO ADOTADA

As especificações definidas para o sistema podem ser resumidas abaixo:

Freqüência de amostragem[9] $f_S = 16kHz$

-Vetor de 12 Coeficientes Mel-Cepstrais, com adição de um vetor de coeficientes delta (12) e com o vetor de coeficientes delta-delta (12), totalizando um vetor de 36 elementos. Além disso, os coeficientes Mel-Cepstrais são normalizados pela média calculada dos coeficientes.

-Coeficiente de pré-ênfase igual a 0,97

-Enjanelamento (*Windowing*) utilizando janela Hamming.

-Tamanho da janela de 25ms, com deslocamento de *frame* de 10ms.

[9] Freqüência empregada para gravação de toda a base de dados *SWITCHBOARD*.

9.3. MEDIDAS DE DESEMPENHO

Hresults é o módulo do HTK responsável pela computação do desempenho do sistema, fazendo uma comparação das seqüências de palavras identificadas em cada sentença contra o respectivo arquivo de transcrição de texto, executando um alinhamento entre a sentença reconhecida e a transcrição de texto correspondente[10].

A seguir o módulo calcula as taxas de Acerto e de Precisão [11], segundo expressões:

$$H = N - D - S$$

$$Acerto\% = \frac{H}{N} \times 100$$

$$Precisão\% = \frac{H - I}{N} \times 100$$

sendo:

N número total de palavras existentes nas transcrições.

D número de omissões (*"deletions"*) realizadas.

S: número de substituições realizadas.

I: número de inserções realizadas.

É interessante notar que para determinar substituições, inserções e omissões, o módulo HResults realiza um procedimento de alinhamento segundo algoritmo de programação dinâmica. Para tal, inicialmente são definidos escores de acordo com as comparações feitas entre a seqüência reconhecida e a respectiva transcrição. Palavras que estiverem alinhadas perfeitamente entre a seqüência reconhecida e a transcrição recebem escore 0, palavras inseridas ou omitidas recebem escore 7 e palavras que foram substituídas por outras recebem escore 10. Assim, o algoritmo procura reduzir o escore obtido até o mínimo valor possível.

Além disso, deve-se considerar que as taxas de acerto e precisão devem ser estudadas em conjunto para uma correta avaliação do sistema obtido.

9.4. BASE DE DADOS *SWITCHBOARD*

A base de dados *SWITCHBOARD* (SWB) foi originalmente criada pela Texas Instruments com o apoio do DARPA e hoje é distribuída pelo *Linguistic Data Consortium* (LDC)[12]. O conjunto completo de 28 CD's inclui cerca de 2.430 conversações, com duração média de seis minutos cada. Em outras palavras, existem mais de 240 horas de conversações **espontâneas** gravadas, e cerca de 3 milhões de palavras de texto faladas por mais de 500 falantes de ambos os sexos, em cada um dos dialetos mais significativos do inglês americano.

Além disso, o SWB possui características únicas em sua construção, como por exemplo o desenvolvimento de uma tecnologia de coleta das amostras de fala por via telefônica, bem como serve como referência básica para pesquisas que trabalham com **fala espontânea**. Dentre essas características pode-se destacar:

- A base de dados *SWITCHBOARD* foi coletada sem intervenção humana, isto é, totalmente controlada por computador. Deste modo a interação com o sistema era via teclado telefônico e por instruções pré-gravadas, e os participantes uma vez conectados poderiam ter até um tempo de aquecimento antes de começar a gravação.
- As linhas telefônicas utilizadas foram linhas digitais do tipo T1, providas por um *software* automático de chaveamento, tornando possível a coletagem em versão digital dos sinais de fala diretamente da rede telefônica e também o isolamento dos dois lados das conversações. Assim, permitindo a separação dos interlocutores, o material do *SWITCHBOARD* permite que os pesquisadores possam treinar os seus sistemas com a fala de cada locutor separadamente, ou ainda na fase de teste utilizar cada interlocutor, ou ambos, em qualquer conversação.

A base de dados *SWITCHBOARD* possui seus arquivos de sinal transcritos em sua totalidade em arquivos de texto (*label files*), sendo que cada transcrição foi realizada de acordo com convenções previamente definidas e documentadas. Assim, relatores especialmente treinados produziram a maior parte das transcrições verbais, seguindo regras de transcrição de um manual preparado especificamente para o projeto. Além disso, o trabalho desses relatores foi conferido por meio de um *script* em *awk*[13] desempenhando esta revisão com o dobro da qualidade realizada por inspeções feitas por humanos.

Os arquivos de transcrições de texto são também acompanhados por um arquivo de alinhamento temporal, o qual estima o tempo inicial e a duração de cada palavra contida nas transcrições em centisegundos. Este *corpus* então é capaz de dar suporte não somente a uma abordagem independente de texto, e própria para a identificação de locutores, como também pode ser empregado para o reconhecimento de fala, realizando a identificação de texto, inclusive com o emprego de modelagem fonética. Além disso, o *SWITCHBOARD* deve facilitar os estudos das características fonéticas da fala espontânea numa escala até então impossível.

As informações dos locutores da montagem deste *corpus*, bem como datas, tempos de aquisição e outras informações pertinentes às ligações estão gravadas e representadas em um banco de dados relacional. Assim, com exceção de informações de foro pessoal, os voluntários que participaram do projeto forneceram informações importantes para o estudo de fala, tais como dialeto, idade, sexo, grau de instrução, residência atual e locais onde residiu durante os anos de formação. Além disso, o código de área e o exato tempo de cada ligação é fornecido, bem como um registro de que um dado falante possa ter fornecido amostras de diferentes telefones, pois muitos dos falantes fizeram chamadas de vários aparelhos telefônicos diferentes de forma a facilitar o estudo deste tipo de variação em reconhecimento de fala.

9.5. MODELO FONÉTICO ESTATÍSTICO

Os modelos fonéticos estatísticos são a contribuição deste trabalho. Tais modelos foram treinados com os arquivos

[10] Este alinhamento pode ser apresentado na saída de HResults com o parâmetro -t.

[11] *Percent Correct e Percent Accuracy*.

[12] O LDC pode ser encontrado no endereço internet: http://www.ldc.upenn.edu.

[13] Utilitário para manipulação e edição de arquivos do ambiente UNIX. O nome *awk* é uma sigla formada dos nomes dos autores do programa.

de transcrição fonética das sentenças de treinamento e uma lista de símbolos fonéticos empregados pela base de dados *SWITCHBOARD*, como arquivos de entrada do módulo HLStats. Deste modo, os modelos fonéticos estatísticos são, essencialmente, estruturas de relacionamento estatístico tomadas agora ao nível fonético-fonológico, a exemplo dos modelos de linguagem estatísticos empregados com palavras. Foram computadas as probabilidades condicionais de segunda ordem, isto é os bigramas, gerando-se ao final o modelo fonético estatístico desejado.

O arquivo modelo fonético estatístico foi utilizado no algoritmo de pesquisa proposto por Herman Ney, adicionando-se, durante o processo de busca, as probabilidades condicionais dos modelos fonéticos acústicos identificados pelo sistema.

Assim, o novo escore obtido influencia a decisão tomada na busca, devido ao limiar de corte imposto ao longo desta. Os resultados obtidos por esta técnica demonstram uma melhoria no desempenho geral do sistema, como aliás esperado.

9.5.1. PARÂMETRO DE CONTRIBUIÇÃO α

O processamento do reconhecedor, usando o módulo HVite, ou o módulo HNey, (que processa o sistema proposto por Herman Ney), executa o cálculo da expressão 9.I ou 9.II pelos valores logaritmos dos termos destas expressões.

Deste modo, chamando os termos $P(W \mid F)$ e $P(W)$ das expressões (9.I e 9.II) de $P(\omega)$ [14], e o termo $P(F)$ e $P(F \mid W)$ das expressões (9.I e 9.II) de $P(f)$ [15], pode-se chegar a uma expressão mais geral envolvendo ambas as expressões de forma:

$$\hat{W} = \arg\max_{W} P(Y \mid F,W)P(\omega)P(f) \qquad (14)$$

que será calculada pelo algoritmo de busca na forma:

$$\hat{W} = \arg\max_{W}\{\log[P(Y \mid F,W)] + \log[P(\omega)] + \log[P(f)]\} \qquad (15)$$

Todavia, os termos $\log[P(\omega)]$ e $\log[P(f)]$ são de mesma ordem de grandeza, o que significa que ambos possuem, em módulo, valores próximos, alterando o cálculo do escore acumulado final. De fato, será necessário pon-derar a contribuição do termo $\log[P(f)]$ no processo de reconhecimento visto que sua influência não deve ser determinante no processo de identificação das palavras da sentença, mas sim, deve portar-se como valor adicional para ajuste dos escores obtidos durante o reconhecimento.

Sendo assim, para implementação do módulo da expressão (15) será adotado um parâmetro de contribuição do modelo fonético estatístico no escore acumulado durante o processo de busca. A expressão (15) torna-se então:

$$\hat{W} = \arg\max_{W}\{(1-\alpha)(\log[P(Y \mid F,W)] + \log[P(\omega)]) + \alpha(\log[P(f)])\} \qquad (16)$$

Deste modo, o parâmetro α traduzirá o percentual de participação do modelo fonético estatístico, no cálculo do escore acumulado durante o reconhecimento das palavras da sentença.

O valor ótimo de α deverá ser empiricamente determinado, variando-se α até a obtenção do valor de máximo desempenho do reconhecedor. A faixa de variação de α ficará, de acordo com a expressão (16) no intervalo entre 0 e 1. No entanto, valores entre 0,01 até 0,5 serão experimentados, para melhor ajuste do sistema.

10. RESULTADOS

Para preparação do sistema na etapa de treinamento foram utilizadas 3943 sentenças de treinamento, aleatoriamente escolhidas do Banco de dados *SWITCHBOARD*, constituindo assim um sub-lote de sentenças de treinamento. Os modelos de Markov HMM gerados neste treinamento são modelos fonéticos do tipo monofones, de topologia *left-right* e com 5 estados. Além disso, foram utilizados dois coeficientes *mixtures* por estado, num primeiro conjunto de modelos HMM e depois foi gerado um novo conjunto de modelos HMM com 4 *mixtures* por estado. O conjunto com 4 *mixtures* apresentou um desempenho comparativamente maior e foi o escolhido para avaliação.

Na etapa de avaliação foi utilizado um outro lote de dados, diferente do lote empregado no treinamento, com 315 sentenças aleatoriamente escolhidas.

Foi criado o modelo fonético estatístico, empregando agora o módulo HLStats com a transcrição fonética dos *labels* [16] dos arquivos de treinamento.

Assim, após a preparação destes arquivos de dados, foram elaborados os seguintes sistemas de reconhecimento a saber:

-Reconhecedor padrão (Viterbi)
-Reconhecedor padrão com Modelo Fonético Estatístico
-Reconhecedor Herman Ney
-Reconhecedor Herman Ney com Modelo Fonético Estatístico

O reconhecedor padrão emprega o algoritmo de Viterbi padrão, sendo então o mesmo reconhecedor utilizado para reconhecimento de fala de linguagem contínua. No presente trabalho, os sistemas empregados foram o Hvite e o HAPIVite, um módulo reconhecedor funcionalmente idêntico ao HVite, e todo construído com funções da Biblioteca HAPI.

O reconhecedor padrão com modelo fonético estatístico é uma nova versão do HAPIVite, alterada para executar a inclusão adicional do modelo fonético estatístico no processamento do reconhecedor.

O reconhecedor Herman Ney é o reconhecedor que emprega os algoritmos e a estrutura em árvore léxica do sistema proposto por Herman Ney. Deste modo, os arquivos de

[14] visto que tais elementos são essencialmente iguais dependendo da abordagem *top-down* ou *botton-up* empregada.

[15] de forma análoga, tais termos, como visto em 3.2 são funcionalmente semelhantes.

[16] *Label* é o nome empregado para designar o arquivo de texto associado ao arquivo de sinal, que contém a sentença escrita.

dados referentes ao vocabulário proposto foram anteriormente gerados pelo módulo HLextree[17], como visto no tópico anterior e que são, respectivamente, o dicionário fonético modificado e a estrutura sintática relacionando as palavras do dicionário.

O reconhecedor Herman Ney com modelo fonético estatístico emprega adicionalmente esta estrutura no processo de reconhecimento.

Para avaliação do sistema, inicialmente foi utilizado um dicionário fonético de aproximadamente **18000 palavras,** e modelo de linguagem já treinado[18], bem como uma árvore léxica gerada a partir deste dicionário fonético, com o uso do módulo HLextree.

10.1. RECONHECEDOR VITERBI

A Tabela 3 apresenta os desempenhos apresentados pelo reconhecedor Viterbi padrão e pelo mesmo reconhecedor empregando o modelo fonético estatístico. Pode-se notar nesta tabela que o modelo fonético estatís-tico aumentou o desempenho do sistema, reduzindo o nú-mero de inserções, aumentando perceptivelmente a taxa de precisão percentual.

Tipo do reconhecedor	%Acerto	%Precisão
Reconhecedor Padrão	53,29%	13,31%
Reconhecedor com Modelo Fonético Estatístico	54,73%	17,31%

Tabela 3. Avaliação do reconhecedor padrão com Modelo fonético estatístico.

10.2. RECONHECEDOR HERMAN NEY

De forma análoga ao item anterior, as tabelas 4, 5 e 6 apresentarão os totais obtidos com os reconhecedores empregando o sistema Herman Ney padrão e com modelo fonético estatístico.

10.3. ÁRVORE LÉXICA: DADOS COMPARATIVOS

Foram preparadas para avaliação duas árvores léxicas: uma com apenas um fonema na seqüência comum, e outra com dois fonemas na seqüência comum. A Tabela 4 apresenta o número de arcos obtidos para as árvores léxicas utilizadas para avaliação. Como arco deve-se entender os ramos da seqüência comum conectando o início da árvore, ou ainda nó raiz, até o início de cada palavra do vocabulário. Assim, para um conjunto de 48 símbolos fonéticos empregados no presente sistema, uma árvore léxica de apenas um fonema, o inicial, na seqüência comum apresentou 42 arcos, isto é, 42 fonemas, dos 48 que são encontrados no início das palavras do vocabulário. Este resultado é totalmente esperado, visto que nem todos os fonemas da língua são empregados no início de um palavra. Outro aspecto interessante é notar que, para uma seqüência de dois fonemas iniciais, poder-se-ia esperar $\approx 42^2 = 1764$ arcos. No entanto, para dois fonemas na seqüência comum obtém-se 491 arcos, evidenciando uma perplexidade bem menor do que 42.

	Um fonema	Dois fonemas
número de arcos	42	491

Tabela 4. Número de arcos obtidos para cada árvore léxica

10.4. RESULTADOS COM UM FONEMA NA SEQÜÊNCIA COMUM

A Tabela 5 apresenta o resultado obtido com o emprego de uma árvore léxica de apenas um fonema na seqüência comum. Pelos valores obtidos, pode-se perceber que o modelo fo-nético estatístico produziu uma diferença sensível (toman-do-se os valores 15.30% e 19.77%) na precisão do sistema, isto é, reduzindo o número de inserções durante o reconhe-cimento. Por outro lado, não há melhora perceptível na taxa de acertos, visto que a diferença obtida é inferior a 10%.

Tipo do reconhecedor com 1 fonema na seqüência comum	%Acerto	%Precisão
Reconhecedor Herman Ney Padrão	56,27%	15,30%
Reconhecedor Herman Ney com Modelo Fonético Estatístico	59,50%	19,77%

Tabela 5. Resultados com um fonema na seqüência comum.

10.5. RESULTADOS COM DOIS FONEMAS NA SEQÜÊNCIA COMUM

Novamente pode-se perceber uma melhora na precisão do sistema, com um escore final de 22,17%. Comparativamente aos resultados obtidos na Tabela 6 pode-se perceber que o número de fonemas na seqüência comum influência o resultado final. Juntamente com o emprego do modelo fonético estatístico, a diferença entre o Herman Ney padrão com um fonema na seqüência comum, e o Herman Ney com dois fonemas na seqüência comum usando modelo fonético estatístico é de 6,87% na taxa de precisão e de 4,87% na taxa de acerto, o que representa uma melhora expressiva no desempenho do sistema. Percebe-se também que o modelo fonético apresenta melhorias de desempenho também na taxa percentual de acertos, mas a influência na precisão é maior.

Assim, em todos os casos pode-se notar que a precisão do sistema apresenta uma expressiva melhora em comparação com o algoritmo Viterbi padrão, tomado como referência. Cabe também ainda ressaltar que os resultados obtidos devem ser avaliados à luz dos sistemas de reconhecimento de **fala espontânea**, que são ainda um

[17] HLextree é um módulo desenvolvido para gerar a árvore léxica e especificamente criado em [23].

[18] Estes arquivos foram obtidos no site da John Hopkins University, direto do CLSP (Center for Language and Speech Processing): http://www.clsp.jhu.edu.

grande desafio para as pesquisas em processamento de fala, e onde o aumento na taxa de precisão constitui importante contribuição para a melhoria do desempenho final.

Tipo do reconhecedor com 2 fonemas na seqüência comum	%Acerto	%Precisão
Reconhecedor Herman Ney Padrão	57,27%	18,50%
Reconhecedor Herman Ney com Modelo Fonético Estatístico	61,14%	22,17%

Tabela 6. Resultado com dois fonemas na seqüência comum.

REFERÊNCIAS

[1] E. Morin, "O Enigma do Homem Para uma Nova Antropologia", Zahar , Rio de Janeiro, 1979.

[2] J.J. Wolf, Speech recognition and understanding. In: K.S. FU et al, "Digital pattern recognition", Spriger, Berlin, 1980. pp.167-203.

[3] Departamento de Lingüística, "Noções de fonética para o estudo do processamento do som da fala", Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, São Paulo: 1992.

[4] D .O'Shaughnessy, "Speech Communication Human and machine", Addison-Wesley, Massachussets, 1987.

[5] J. Deller, J. G. Proakis, J. H. L. Hansen, "Discret-time processing of speech signals,", Macmillan, New York, 1993.

[6] J. Allen, Overview of text-to-speech systems. In: S. Furui, M.M. Sandhi, "Advances in Speech Signal Processing", Marcel Dekker, New York, 1992, pp.741-790.

[7] R.D.R. Fagundes, "Reconhecimento de Fala, Linguagem Contínua, Usando Modelos de Markov", Dissertação (mestrado) - Escola Politécnica, Universidade de São Paulo. São Paulo, 1993.

[8] L.R. Rabiner; S.E. Levinson, A speaker-independent, syntax-directed, connected word recognition system based on hidden markov models and level building, "IEEE Transactions on Acoustics, Speech, and Signal Processing", v.33, pp.561-73, June 1985.

[9] L.R. Rabiner, B.H. Juang, An introduction to Hiden Markov Models, "IEEE Acoustics, Speech, and Signal Processing",pp. 4-16, january 1986.

[10] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. "Proceedings of the IEEE", v.77, pp. 257-86, February. 1989.

[11] X.D. Huang, Y. Ariki; M.A. Jack, "Hidden Markov models for speech recognition", Edinburgh University Press, Edinburgh, 1990..

[12] K.F. Lee, H.W. Hon, R. Reddy, An overview of the SPHINX speech recognition system, "IEEE Transactions on Acoustics, Speech, and Signal Processing", v.38, pp.35-45, January 1990.

[13] E.J. Casaes, "Descrição acústico-articulatória dos sons da fala.", Tese de Doutorado - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo 1990.

[14] H. Ney et alii,. Improvements in beam search for 10000-word continuous speech recognition, "IEEE International Conference on Acoustic Speech and Signal Processing" pp.I-9- I-12, 1992.

[15] P. Kenny, R. Hollan, V.N. Gupta et alii, A* - Admissible Heuristics for Rapid Lexical Access, "IEEE Transactions on Speech and Audio Proceedings", vol.1, pp.49-58, January 1993.

[16] P. Kenny, Z. Li, D. O'Shaughnessy, Searching with a transcription graph, "IEEE International Conference on Acoustic Speech and Signal Processing", pp. 564-567, 1995.

[17] S. E. Levinson, The effects of syntactic analysis on word recognition accuracy, "The Bell System Technical Journal", v.57, pp.1627-44, May 1978.

[18] S.E. Levinson, K.L. Shipley, A conversational-mode airline information and reservation system using speech input and output, "Bell System Technical Journal", v.59, pp.119-37, January 1980.

[19] F. Jelinek, L.R. Bahl, R.L. Mercer, Design of a linguistic statistical decoder for the recognition of continuous speech, "IEEE Transactions on Information Theory", v.IT-21, pp. 250-256, may 1975.

[20] S. Roukos, Language representation, In: R. Cole (ed.) "Survey of the state of the art in humam language technology",pp.35-41,1995. http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html.

[21] R. Lacouture, "Au sujet des algorithmes de recherche des systèmes de reconnaissance de la parole à grands vocabulaires", Thèse de Doctorat, McGill University, Montreal, 1995.

[22] H.F. Silverman, D.P. Morgan, The Application of Dynamic Programming to Connnected Speech Recognition, "IEEE Acustic Speech and Signal Processing Magazine", pp. 6-25, July 1990.

[23] R. D. R. Fagundes, "Abordagem Fonético-Fonológica em Sistemas de Reconhecimento de Fala de Linguagem Contínua", Tese de Doutorado - Escola Politécnica, Universi-dade de São Paulo, São Paulo, 1998.

[24] H. Ney, The use of a one-stage dynamic programming algorithm for connected word recognition, "IEEE Transactions on acoustics, speech and signal processing", v.ASSP-32, pp.263-271, april 1994.

[25] H. Ney, X. Aubert, Dinamic programming search strategies: from digit strings to large vocabulary word graphs, In.: C.H. Lee et alii (eds.), "Automatic speech and speaker recognition". Kluwer Academic Publishers, Boston, 1996.

[26] H. Niemann, G.F. Sagerer, S. Schroder, F. Kummert, Ernest: A Semantic Network System for Pattern Understanding, "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol.12, pp.883-905, Sepetember 1990.

[27] E. Rich, K. Knight, "Inteligência Artificial", Makron Books do Brasil, São Paulo, 1993.

[28] S. Russell, P. Norving, "Artificial intelligence: a modern approach", Prentice Hall, New Jersey, 1995.

[29] J. Bresnan, R.M. Kaplan, Introduction: grammars as mental representantions of language, In: J. Bresnan (org.) "The mental represntation of gramatical relations", MIT Press, Massachusetts, 1982.

[30] F. Pereira, Sentence modeling and parsing, In: R. Cole (ed.), "Survey of the state of the art in humam language technology",pp.130-140,1995. http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html

Rubem Dutra Ribeiro Fagundes graduou-se em Engenharia Elétrica em 1989, pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS). Em 1993 obteve o título de Mestre em Engenharia Elétrica pela Escola Politécnica Universidade de São Paulo. Obteve o título de Doutor em Engenharia Elétrica em 1998, pela Escola Politécnica da Universidade de São Paulo, tendo desenvolvido seu projeto de doutoramento no Centre de Recherche Informatique de Montreal (CRIM), em Montreal – Canadá. Desde 1990 é professor do Departamento de Engenharia Elétrica da PUCRS, ministrando aulas tanto na graduação quanto na pós-graduação. É Coordenador do grupo SISC (Sistemas, Sinais e Computação) na mesma Universidade.

Ivandro Sanches graduou-se em Engenharia Eletrônica e obteve o título de Mestre em Engenharia pela Escola Politécnica da USP (EPUSP) em 1987 e 1989, respectivamente, e obteve o título de Doutor pelo Imperial College of Science, Technology and Medicine da Universidade de Londres em 1994. Foi professor do Departamento de Engenharia de Sistemas Eletrônicos da EPUSP, onde desenvolveu atividades didáticas e de pesquisa nas especialidades de Processamento Digital de Sinais e Reconhecimento Automático de Voz. Atualmente é pesquisador do Genius Instituto de Tecnologia.