SOBRE A IMPORTÂNCIA DA TRANSCRIÇÃO FONÉTICA EM SISTEMAS DE RECONHECIMENTO DE FALA*

Carlos Alberto Ynoguti e Fábio Violaro

UNICAMP – FEEC - Departamento de Comunicações
Campinas, SP
{ynoguti,fabio}@decom.fee.unicamp.br

Resumo - Nos sistemas de reconhecimento de fala contínua geralmente são utilizados modelos de sub-unidades acústicas para representar as palavras. Deste modo, a transcrição fonética das locuções de treinamento e das palavras que compõem o vocabulário de tais sistemas é fundamental para o desempenho final do reconhecimento. Neste trabalho é investigada a influência da transcrição fonética tanto das locuções de treinamento como das palavras do vocabulário em sistemas de reconhecimento de fala contínua com independência de locutor.

Abstract - In continuous speech recognition systems acoustical sub-units are used to represent word models. For this reason, phonetic transcription of training data and vocabulary words play a fundamental role in recognizers overall performance. In this work we investigate the influence of phonetic transcription effects both in the training data and in the vocabulary words in speaker independent continuous speech recognition systems.

Palavras chave: Reconhecimento de fala contínua, modelos ocultos de Markov.

1. INTRODUÇÃO

Em sistemas de vocabulário pequeno (algumas dezenas de palavras), é comum utilizar-se as palavras como unidades fundamentais. Para um treinamento adequado destes sistemas, deve-se ter um grande número de exemplos de cada palavra. Entretanto, para sistemas com vocabulários maiores, a disponibilidade de um grande número de exemplos de cada palavra torna-se inviável. A utilização de sub-unidades fonéticas, tais como fonemas, sílabas, demi-sílabas, etc., é uma alternativa bastante razoável, pois agora é necessário ter vários exemplos de cada sub-unidade, e não vários exemplos de cada palavra [7]. Na etapa de reconhecimento, os modelos das palavras são formados a partir da concatenação dos modelos destas sub-unidades fonéticas.

Para o treinamento é necessário fornecer uma transcrição fonética das locuções em termos das sub-unidades utilizadas (fones, sílabas, etc.), para que o sistema possa saber quais delas devem ser treinadas. Esta transcrição é uma tarefa bastante penosa e demorada, pois é necessário ouvir com atenção as locuções, e com a ajuda de programas de visualização gráfica da forma de onda e do espectro do sinal, estabelecer exatamente o que foi pronunciado.

Poder-se-ia aliviar a carga de trabalho necessária para a confecção da base de dados se fosse adotada uma transcrição fonética padrão para todas as locuções. Esta poderia ser obtida a partir da transcrição fonética das locuções de um locutor padrão, ou a partir de um léxico.

Espera-se que com este procedimento, o desempenho do sistema caia, pois um fonema poderia estar sendo treinado com a locução de outro. A questão é: quanto? Se a queda verificada no desempenho não for significativa, este procedimento permitiria a construção de bases de dados maiores, o que significa mais exemplos de treinamento e, consequentemente, sub-unidades fonéticas mais bem treinadas. Este compromisso pode fazer com que, mesmo que as transcrições fonéticas padronizadas atrapalhem o treinamento, o maior número de exemplos acabe por compensar a queda no desempenho provocada por este procedimento.

Nesta mesma linha, existe outra questão a ser abordada: a do arquivo de vocabulário do sistema de reconhecimento. De modo a lidar com diferenças de pronúncias entre os locutores, diferenças estas devido a regionalismos, coarticulações e outros fatores, os arquivos de vocabulário costumam trazer várias pronúncias para uma mesma palavra. Isto pode fazer com que o tamanho destes arquivos seja bem maior do que o número de palavras a serem reconhecidas, ocasionando um aumento excessivo no tempo de processamento e na perplexidade da busca. Aplicando-se a mesma idéia de uma locução padronizada, agora para cada palavra do vocabulário, poder-se-ia reduzir o espaço de busca, reduzindo também o tempo de processamento e a perplexidade.

Na verdade, existe um compromisso entre a capacidade de generalização e a perplexidade: um número grande de versões para cada palavra garante que o sistema seja teoricamente capaz de reconhecer locuções de locutores com diferentes sotaques e formas de pronúncia, mas ao mesmo tempo aumenta a perplexidade da busca, derrubando o seu desempenho.

Neste trabalho é investigada a influência da transcrição fonética das locuções de treinamento, e do número de versões de cada palavra na taxa de acertos para sistemas de reconhecimento de fala contínua com independência de locutor, que usam modelos de palavras formados a partir da concatenação de sub-unidades acústicas.

*Este trabalho foi parcialmente financiado pela FAPESP, processo 99/01241-2.

Para os testes e avaliação dos resultados foram utilizados dois sistemas: um baseado em modelos ocultos de Markov discretos [12] e outro baseado em modelos ocultos de Markov contínuos [13].

2. BASE DE DADOS

A base de dados utilizada neste trabalho foi gerada a partir de um trabalho realizado por Alcaim et. al. [1]. Neste, foram criadas 20 listas, cada uma com 10 frases foneticamente balanceadas, segundo o português falado no Rio de Janeiro. Neste conjunto de frases contou-se 694 palavras diferentes, o que caracteriza um vocabulário de tamanho médio.

Para as gravações, foram selecionados 40 locutores adultos, sendo 20 homens e 20 mulheres. Estes foram divididos igualmente em 5 grupos, cada qual com 4 homens e 4 mulheres. Para cada grupo foram designadas 4 das 20 listas da base de dados da seguinte forma: as primeiras 4 listas para o primeiro grupo, as 4 seguintes para o segundo grupo, e assim por diante. Desta forma, cada locutor pronunciou no total 40 frases, e cada frase foi repetida por 8 locutores diferentes.

Um locutor extra do sexo masculino completa a base de dados. Este locutor pronunciou todas as 200 frases, repetindo-as 3 vezes. Daqui para frente, este locutor será referenciado como locutor m01.

As gravações foram feitas em ambiente relativamente silencioso, utilizando um microfone direcional, e uma placa de áudio SoundBlaster AWE 64. A frequência de amostragem utilizada foi 11,025 kHz, com uma resolução de 16 bits.

As locuções foram parametrizadas a partir de 12 coeficientes mel-cepstrais [4], com janelas de 20 ms, atualizadas a cada 10 ms. Antes da parametrização, os sinais de voz passaram por um filtro de pré-ênfase com função de transferência $(1-0{,}95z^{-1})$ e um janelamento de Hamming. De modo a oferecer informações contextuais ao sistema, foram também utilizados os parâmetros delta e delta-delta mel-cepstrais. Aos parâmetros mel-cepstrais foi aplicado o procedimento de remoção da média espectral [6].

Para a versão discreta dos modelos de Markov, os parâmetros foram quantizados a partir de um quantizador vetorial baseado no algoritmo LBG [8]. Foi criado um codebook de 256 vetores código para cada um dos parâmetros (mel-cepstrais, delta mel-cepstrais e delta-delta mel-cepstrais).

Para a versão contínua foi utilizado um modelamento das densidades de emissão através de misturas de gaussianas. Em testes preliminares, o melhor resultado foi obtido utilizando-se 6 gaussianas por mistura, de modo que este valor foi adotado no presente trabalho.

3. SUB-UNIDADES FONÉTICAS

As sub-unidades fonéticas utilizadas foram os fones independentes de contexto. Os modelos das palavras foram gerados a partir da concatenação dos modelos dos fones correspondentes à sua transcrição fonética.

Para cada fone, foi criado um modelo de Markov com arquitetura *left-right* de 3 estados, como mostrado na Figura 1.

Para este trabalho, foram considerados 36 fones independentes de contexto [2] (incluindo o silêncio), e que são mostrados na Tabela 1.

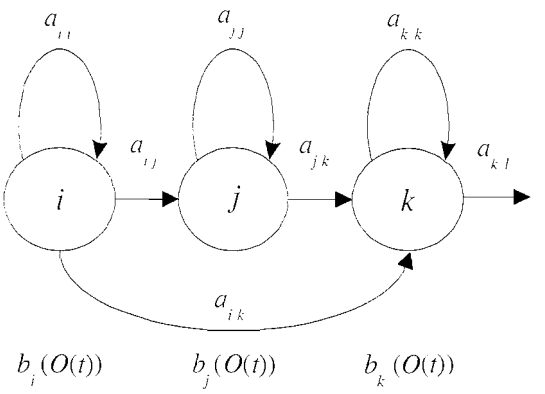

Figura 1. Modelo de Markov para cada um dos fones independentes de contexto. Nesta, a_{ij} é a probabilidade de efetuar uma transição do estado i para o estado j e $b_i(O(t))$ é a probabilidade de emitir o símbolo $O(t)$ no estado i. no instante t.

Fone	Exemplo	Fone	Exemplo
#	silêncio	g	***g*** orila
a	***a*** çafrão	G	***j*** iló
e	***e*** levador	k	***c*** achoeira
ɛ	p ***e*** le	l	***l*** eão
i	s ***i*** no	L	***lh*** ama
j	fu ***i***	m	***m*** ontanha
o	b ***o*** lo	n	***n*** évoa
ɔ	b ***o*** la	N	i ***nh*** ame
u	l ***u*** a	p	***p*** oente
ã	maç ***ã***	r	ce ***r*** a
ẽ	s ***en*** ta	R	ce ***rr*** ado
ĩ	p ***in*** to	R	ca ***r*** ta
õ	s ***om*** bra	s	***s*** apo
ũ	um	t	***t*** empes ***t*** ade
b	***b*** ela	tS	***t*** igela
d	***d*** ádiva	v	***v*** erão
dG	***d*** iferente	S	***ch*** ave
f	***f*** eira	z	***z*** abumba

Tabela 1: Fones independentes de contexto, com seus respectivos exemplos.

4. TREINAMENTO DO SISTEMA

4.1. TREINAMENTO DAS SUB-UNIDADES

O sistema foi treinado usando o algoritmo *FORWARD-BACKWARD* [10]. O procedimento adotado pode ser resumido nos seguintes passos:

- Para os HMM's discretos, as probabilidades de emissão são inicializadas com o valor $1/n$, onde n é o número de vetores utilizados na quantização vetorial dos parâmetros (256 neste caso).

- Para os HMM's contínuos a inicialização é realizada em duas etapas, utilizando-se o algoritmo *Segmental K-Means* [11]. Inicialmente, as locuções de treinamento são divididas em m partes iguais (de mesmo comprimento), sendo m o número de sub-unidades fonéticas da transcrição fonética multiplicada pelo número de estados de cada modelo HMM (3 neste trabalho). É criado um modelo HMM para a locução concatenando-se os modelos HMM das sub-unidades acústicas referentes à sua transcrição fonética. Cada um dos m conjuntos de vetores acústicos deve então ser utilizado para estimar as médias e variâncias de cada uma das gaussianas da mistura. Supondo que temos g gaussianas para cada mistura e n vetores acústicos para estimá-las, as médias são estimadas a partir de um quantizador vetorial de g níveis. Neste trabalho foi utilizado o algoritmo LBG para estimar os vetores código do quantizador vetorial. Uma vez estimadas as médias, faz-se uma quantização dos n vetores nas g médias, e calcula-se a variância correspondente a cada gaussiana. Numa segunda etapa, este procedimento é repetido, só que agora, ao invés de utilizar uma segmentação uniforme, utiliza-se o algoritmo de Viterbi [11] para realizar uma segmentação mais criteriosa das locuções de treinamento.
- As probabilidades de transição são inicializadas segundo o esquema da Figura 2. Os modelos dos trifones são inicializados com os modelos treinados dos fones independentes de contexto.

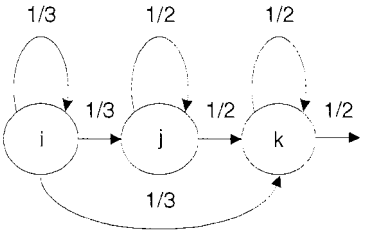

Figura 2: Valores iniciais para as probabilidades de transição para os modelos dos fones independentes de contexto.

- Constrói-se o modelo da sentença (locução) concatenando-se os modelos das sub-unidades fonéticas correspondentes à transcrição fonética da mesma.
- A este modelo são aplicados os parâmetros da locução e realiza-se a contagem das transições e emissões para cada sub-unidade do modelo (as probabilidades de transição e emissão não são atualizadas ainda).
- Depois de cada locução, as contagens para cada fone são acumuladas. É importante lembrar que o mesmo fone pode ocorrer mais de uma vez em cada sentença.
- Este processo é repetido para todas as sentenças do conjunto de treinamento.
- As probabilidades de transição e de emissão são atualizadas para todos os fones utilizando as contagens acumuladas.
- Depois que todas as locuções de treinamento foram apresentadas ao sistema, realiza-se um teste de convergência, que consiste em aplicar a locução ao seu modelo correspondente e verificar a probabilidade de o modelo gerar aquela sequência de observação $P(O/M)$.

Repetindo este procedimento para todas as locuções do conjunto de treinamento, pode-se calcular a $P(O/M)$ média. Este valor médio aumenta a cada época de treinamento até atingir um patamar. Com base nesta quantidade pode-se definir uma medida de distorção:

$$dist = \frac{\overline{P}(O|M)_n - \overline{P}(O|M)_{n-1}}{\overline{P}(O|M)_n} \quad (1)$$

onde:
dist é a distorção;
O é a sequência de observação;
M são os parâmetros do modelo;
$\overline{P}(O|M)_n$ é a verossimilhança média para a época atual;
$\overline{P}(O|M)_{n-1}$ é a verossimilhança média da época anterior.

- Este processo é repetido até que *dist* caia abaixo de um dado limiar ε. Neste trabalho, foi utilizado $\varepsilon = 0.001$.

4.2. TRANSCRIÇÕES FONÉTICAS

Foram utilizadas duas transcrições fonéticas para as locuções de treinamento:

- Uma transcrição fonética criteriosa para cada uma das locuções de treinamento, através de uma audição cuidadosa e o auxílio de programas para visualização gráfica da forma de onda e espectro do sinal
- Uma transcrição fonética padronizada para todos os locutores de treinamento, baseada na transcrição fonética das locuções do locutor m01.

Como uma ilustração do quão diferente pode ser a pronúncia de uma mesma frase por conta do sotaque dos locutores, são mostradas algumas transcrições fonéticas retiradas da base de dados para a frase: "O grêmio ganhou a quadra de esportes":

o g r e m y u g α̃ ◌̯ o u a k u a d r a d e e s p ❷ r t e s

u g r e m y u g α̃ ◌̯ o u a k u a d r a D j s p ❷ r t ♦ j s

u g r e m y u g α̃ ◌̯ o u a k u a d r a D j ♦ p ❷ r̄ t ♦ j ♦

5. SISTEMA DE RECONHECIMENTO

O reconhecimento foi realizado através do algoritmo *One Step* [9], com a utilização do procedimento *Beam Search* [5] para redução do número de cálculos. Ao algoritmo de busca foram adicionados o modelo de duração de palavras e o modelo de linguagem de pares de palavras [3].

Para o modelo de duração foi escolhida a forma proposta por Rabiner [10], que associa à duração d de cada palavra i do vocabulário, uma função densidade de probabilidade gaussiana $f_i(d)$

$$f_i(d) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(d-\bar{d}_i)^2}{2\sigma_i^2}\right) \quad (2)$$

onde $\overline{d_i}$ e σ_i^2 são, respectivamente, a média e a variância da duração da palavra i. Estes valores foram obtidos a partir da segmentação manual das locuções de treinamento do locutor m01.

O procedimento para incorporar o modelo de duração aos algoritmos de busca é o seguinte: a cada instante de tempo t, determina-se a duração da palavra i, através da recuperação do caminho ótimo determinado pelo algoritmo de Viterbi. A verossimilhança acumulada para uma dada palavra é penalizada de acordo com a função densidade de probabilidade gaussiana, com os parâmetros da palavra em análise, no ponto determinado por $d_i(t)$:

$$\hat{P}(O \mid M) = P(O \mid M) + \log_{10}(f_i(d_i(t))) \tag{3}$$

onde:

$P(O \mid M)$ é a verossimilhança acumulada;

$\log_{10}(f_i(d_i(t)))$: é a penalização imposta pelo modelo de duração;

$\hat{P}(O \mid M)$: é a verossimilhança acumulada penalizada pelo modelo de duração de palavra.

O modelo de linguagem foi obtido a partir das 200 frases da base de dados.

No sistema foi utilizado um modelo de linguagem do tipo pares-de-palavras [3], que é uma simplificação do modelo *bigrama*. Esta aproximação pode ser descrita através da expressão:

$$\tilde{P}(W) = \prod_{i=1}^{n} G(w_i \mid w_{i-1}) \tag{4}$$

onde

$$G(w_i \mid w_{i-1}) = \begin{cases} 1, P(w_i \mid w_{i-1}) \neq 0 \\ 0, P(w_i \mid w_{i-1}) = 0 \end{cases} \tag{5}$$

Este modelo de linguagem pode ser visto como uma versão determinística do modelo bigrama. A escolha por este modelo em detrimento do bigrama é devida à limitação da base de dados: como é muito pequena, a utilização das frequências bigrama poderia polarizar o algoritmo de busca em alguns casos. Por exemplo, supondo que a sequência "a casa" tenha ocorrido duas vezes, e a sequência "a taça" apenas uma vez, o sistema poderia reconhecer a locução "a taça" como "a casa", visto que são parecidas, e o modelo de linguagem atribuiria uma probabilidade duas vezes maior para a sequência "a casa".

A incorporação do modelo de linguagem aos algoritmos de busca é trivial: ao início de cada nível, em cada instante t, verifica-se qual a palavra vencedora no nível anterior e, se a palavra sob análise for permitida pelo modelo de linguagem, é expandida.

5.1. ARQUIVOS DE VOCABULÁRIO

Para a determinação do universo de palavras que o sistema pode reconhecer, foram criados dois arquivos de vocabulário contendo todas as palavras contidas nas listas de frases que compõem a base de dados. O primeiro foi construído levando-se em consideração o fato de a mesma palavra poder ter mais de uma pronúncia possível (devido a sotaques de regiões diferentes e/ou efeitos de coarticulação). Algumas palavras tiveram mais de uma variante incluída no arquivo de vocabulário. Deste modo, o espaço de busca passou de 694 para 1633 palavras. Este arquivo será referenciado como vocabulário completo nos testes descritos na seção seguinte. Partindo deste arquivo de vocabulário, foram eliminadas todas as versões alternativas, mantendo apenas uma versão que pode ser considerada como padrão. A escolha desta versão padrão foi feita com base nas locuções da base de dados: foi selecionada a versão mais frequentemente encontrada nas transcrições fonéticas das locuções de treinamento. Este será referenciado como vocabulário simplificado.

6. RESULTADOS EXPERIMENTAIS

6.1. INFLUÊNCIA DA TRANSCRIÇÃO FONÉTICA NO DESEMPENHO DO SISTEMA

Os resultados dos testes utilizando a transcrição criteriosa e a transcrição padrão são mostrados resumidamente na Tabela 2 para o sistema HMM discreto a na Tabela 3 para o sistema HMM contínuo. Nestas, os símbolos D, S e I indicam as porcentagens de erros de deleção, substituição e inserção, respectivamente. Pode-se verificar que a queda de desempenho do sistema é muito pequena quando se adota a transcrição padronizada. Nestes testes foi utilizado o vocabulário simplificado.

Transcrição	D (%)	S (%)	I (%)	Total (%)
original	4,86	12,44	1,78	19,08
padrão	5,31	13,24	2,02	20,57

Tabela 2: Desempenho do sistema em função das transcrições fonéticas das locuções de treinamento para o sistema HMM discreto.

Transcrição	D (%)	S (%)	I (%)	Total (%)
original	4,00	12,14	2,28	18,42
padrão	4,49	12,59	2,55	19,63

Tabela 3: Desempenho do sistema em função das transcrições fonéticas das locuções de treinamento para o sistema HMM contínuo.

6.2. INFLUÊNCIA DO NÚMERO DE VERSÕES DE CADA PALAVRA NO ARQUIVO DE VOCABULÁRIO

Com os resultados obtidos nos testes da seção anterior, pode-se verificar que uma transcrição fonética padronizada para todos os locutores não degrada de forma apreciável o desempenho do sistema. A partir deste resultado, foi investigada a influência que teria a mesma idéia quando aplicada ao arquivo de vocabulário.

Com o sistema treinado a partir das transcrições mais criteriosas, foram realizados testes utilizando o vocabulário completo e o vocabulário simplificado, e o desempenho dos sistemas com os dois arquivos de vocabulário pode ser visto nas Tabelas 4 e 5.

Vocabulário	D (%)	S (%)	I (%)	Total (%)
Completo	4,72	14,00	2,59	21,31
Simplificado	4,86	12,44	1,78	19,08

Tabela 4: Resultados dos testes com vocabulário simplificado (apenas 1 versão de cada palavra) para o sistema HMM discreto.

Vocabulário	D (%)	S (%)	I (%)	Total (%)
Completo	3,77	11,68	2,43	17,88
Simplificado	4,00	12,14	2,28	18,42

Tabela 5: Resultados dos testes com vocabulário simplificado (apenas 1 versão de cada palavra) para o sistema HMM contínuo.

Para o sistema HMM contínuo observou-se que a utilização de um vocabulário simplificado acarreta uma pequena queda no desempenho do sistema, o que não ocorreu para o sistema HMM discreto, onde notou-se um aumento na taxa de acertos com este procedimento.

Como dito anteriormente, existe um compromisso entre a capacidade de generalização e a perplexidade: um número grande de versões para cada palavra garante que o sistema seja teoricamente capaz de reconhecer locuções de locutores com diferentes sotaques e formas de pronúncia, mas ao mesmo tempo aumenta a perplexidade da busca, derrubando o seu desempenho. O sistema HMM contínuo, por ter um desempenho melhor, obtém uma maior vantagem quando se utiliza o vocabulário expandido, pois este permite um casamento mais justo entre os modelos a serem testados e as locuções a serem reconhecidas. Já para o sistema HMM discreto, esta verdade se inverte: por causa da quantização vetorial, seu desempenho é mais baixo, e sua capacidade de discernir um grande número de versões de cada palavra é menor do que o do sistema HMM contínuo, e o resultado final de um vocabulário expandido é principalmente o de aumentar a perplexidade da busca.

Na Tabela 6 tem-se um quadro comparativo do tempo médio de reconhecimento para uma frase com a utilização de cada um dos arquivos de vocabulário. Estes tempos foram obtidos com os sistemas rodando em uma máquina com processador AMD-K6 350 MHz, com 64 MB de memória RAM.

Sistema	Vocabulário completo	Vocabulário simplificado
HMM discreto	05:17	02:10
HMM contínuo	04:59	02:55

Tabela 6: Tempo médio de reconhecimento por frase para os testes com os dois arquivos de vocabulário (em minutos).

Os dados da Tabela 6 mostram que a adoção de um vocabulário simplificado provoca um grande aumento na velocidade de processamento, aproximadamente proporcional ao número de palavras eliminadas. Para estes tempos, cabe um comentário: o sistema foi desenvolvido sobre a plataforma Windows, que por sua vez é um sistema multi-tarefas, o que significa que o processador e demais recursos do sistema são divididos entre outros programas (incluindo o sistema operacional). Desta forma, os tempos de reconhecimento da Tabela 6 devem ser vistos apenas como uma aproximação grosseira dos tempos médios de reconhecimento.

7. CONCLUSÕES

Na construção de bases de dados para o treinamento e avaliação de um sistema de reconhecimento de fala, uma das atividades mais custosas é a da transcrição fonética das locuções. Esta requer uma audição cuidadosa de cada uma das locuções, possivelmente com a ajuda de programas de visualização gráfica da forma de onda e espectro do sinal, para determinar exatamente o que foi pronunciado. A repetição deste processo para centenas ou milhares de locutores fornece uma idéia da dimensão do trabalho a ser realizado. Além disso, é muito comum ocorrerem divergências na transcrição quando feitas por pessoas diferentes.

Neste trabalho foi verificada a influência da precisão da transcrição fonética das locuções de treinamento no desempenho de um sistema de reconhecimento de fala contínua com independência de locutor e vocabulário médio (aproximadamente 700 palavras). Os testes mostraram uma deterioração muito pequena quando se adota uma transcrição fonética padrão para as locuções de treinamento.

A possibilidade de se conseguir bases de dados maiores para o treinamento do sistema sem a preocupação de uma transcrição personalizada para cada locutor é um fator que compensa em excesso a pequena degradação no desempenho provocada pela transcrição fonética padronizada.

Ainda, foi verificada a adoção desta mesma idéia no vocabulário do sistema de reconhecimento. Tradicionalmente, estes possuem mais de uma versão de algumas palavras que são pronunciadas de forma diferente devido a regionalismos e coarticulações. A adoção de uma única versão de cada palavra diminui bastante o espaço de busca, diminuindo também o tempo de processamento. Uma escolha criteriosa da versão de cada palavra a ser adotada no vocabulário pode compensar a falta de generalidade. Outra alternativa viável seria criar um arquivo de vocabulário para cada região do país.

REFERÊNCIAS

[1] ALCAIM, A., SOLEWICZ, J. A., MORAES, J. A., Frequência de ocorrência dos fones e lista de frases foneticamente balanceadas no português falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações*. 7(1):23-41. Dezembro, 1992.

[2] CALLOU, D. e LEITE, Y. *Iniciação à fonética e à fonologia*. Rio de Janeiro : Jorge Zahar. 1995.
[3] COLE, R. A., ed., *Survey of the State of the Art in Human Language Technology*. http://cslu.cse.ogi.edu/publications/index.htm, (26/10/98).
[4] DAVIS, S. & MELMERTSTEIN, P. Comparison of parametric representations for monossylabic word recognition in continously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASP-28(4):357-366. August, 1980.
[5] DELLER Jr., J. R., PROAKIS, J. G., HANSEN, J.H.L. *Discrete time processing of speech signals*. MacMillan Publishing Company. New York. 1993.
[6] HERMANSKY, Hynek. "Exploring temporal domain for robustness in speech recognition". *Proceedings of the International Congress on Acoustics*. Trondheim, Norway, June 26-30, pp 61-64, 1995.
[7] LEE, K. F. Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 38(4):599-609. April, 1990.
[8] LINDE, Y., BUZO, A., GRAY, R. M. An algorithm for vector quantizer design. *IEEE Transactions on Communications*. COM-28(1). January, 1980.
[9] NEY, H. The use of a one-stage dynamic programming algorithm for connected word recognition. . *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2):263-271. April, 1984.
[10] RABINER, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257-286. February, 1989.
[11] RABINER, L. *Fundamentals of speech recognition*. Prentice Hall Press. 1993
[12] YNOGUTI, C. A. Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov. Tese de Doutorado. UNICAMP, Campinas. Maio de 1999.
[13] YNOGUTI, C. A., VIOLARO, F. Um sistema de reconhecimento de fala contínua baseado em modelos ocultos de Markov contínuos. A ser apresentado no XVIII Simpósio Brasileiro de Telecomunicações. 3 a 6 de setembro de 2000. Gramado, RS.

Carlos Alberto Ynoguti nasceu em São Paulo, em 18 de maio de 1967. Formou-se em Engenharia Elétrica pela Escola de Engenharia de São Carlos – USP em 1991. Recebeu o título de Mestre em Engenharia na mesma instituição, no ano de 1994. Em 1999 concluiu o doutoramento pela Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas (UNICAMP). Atualmente desenvolve atividade de Pós-doutoramento na mesma instituição. Suas áreas de interesse são Processamento Digital de Sinais e Reconhecimento de Fala.

Fábio Violaro nasceu em Campinas, São Paulo, em 8 de dezembro de 1950. Graduou-se em Engenharia Elétrica e obteve os títulos de Mestre e Doutor, todos pela Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas (FEEC-UNICAMP) em 1973, 1975 e 1980, respectivamente. Atualmente é professor nível MS-6 do Departamento de Comunicações da FEEC e coordenador do Laboratório de Processamento Digital de Fala. Suas áreas de interesse se concentram em Processamento Digital de Fala: Análise, Codificação, Reconhecimento e Síntese.