Desenvolvimento de um Conjunto de Ferramentas para Pesquisas em Reconhecimento de Fala

Carlos Alberto Ynoguti[1] e Fábio Violaro[2]

[1]Instituto Nacional de Telecomunicações - Santa Rita do Sapucaí - MG - Brasil
[2]DECOM-FEEC-UNICAMP - Campinas - SP - Brasil

***Resumo* — A construção de sistemas de reconhecimento de fala é uma tarefa bastante árdua e complexa. Desta forma, para que os pesquisadores nesta área possam testar suas idéias, seria interessante que já dispusessem de um sistema básico, de modo a economizar tempo de desenvolvimento e pudessem passar rapidamente à fase de implantação e teste de suas próprias idéias. Com este objetivo foi criado um conjunto de ferramentas para pesquisa em reconhecimento de fala contínua baseado em modelos ocultos de Markov contínuos (Hidden Markov Models - HMM). Neste artigo é apresentada uma descrição das ferramentas desenvolvidas e os resultados dos testes iniciais de avaliação do sistema em uma base de dados composta por 40 locutores de ambos os sexos (20 homens e 20 mulheres), com um vocabulário de aproximadamente 700 palavras.**

***Palavras-Chave* — reconhecimento de fala contínua, Modelos Ocultos de Markov.**

***Abstract* — Developing a speech recognition system is a very hard and complex task. So, it would be interesting for the researchers to have a basic system, in order to save developing time and go directly to test his own ideas. With this goal in mind a set of tools was developed for research in continuous speech recognition using Hidden Markov Models (HMMs). In this paper we describe the tools developed and the initial evaluation results on a database consisting of 40 speakers (20 male and 20 female), with a 700 word vocabulary.**

***Index Terms* — continuous speech recognition, Hidden Markov Models.**

I. INTRODUÇÃO

O sonho de todos os pesquisadores na área de reconhecimento de fala é a criação de sistemas em tempo real, com alta taxa de acertos de palavras, com três características principais: independência de locutor, fala contínua e vocabulários extensos. Cada uma destas características dificulta enormemente a tarefa de reconhecimento, e várias pesquisas tem sido feitas no sentido de melhorar o desempenho destes sistemas [2].

Com a consciência de que este objetivo só poderá ser alcançado através do esforço conjunto de vários pesquisadores, é necessário que haja uma base em comum para que não se tenha que começar do zero, e que cada um possa trabalhar apenas na contribuição específica, e que esta possa também ser incorporada ao sistema base.

Perseguindo este objetivo, foi construído um sistema de reconhecimento de fala contínua, que tem a pretensão de servir de base para que futuros pesquisadores possam testar suas idéias rapidamente, diminuindo o tempo entre a concepção e o teste de novas idéias. Ainda, as idéias que realmente representarem um ganho tanto em termos de processamento como em taxa de acertos de palavra poderão ser definitivamente incorporadas ao sistema base.

O modelamento das palavras do vocabulário é feito por sub-unidades fonéticas. Embora os resultados aqui apresentados sejam correspondentes a testes realizados utilizando-se fones independentes de contexto, o sistema suporta qualquer outro tipo de sub-unidade, bastando para isto fornecer a listagem daquelas utilizadas para a transcrição fonética das locuções.

Para a avaliação deste sistema foi utilizada uma base de dados desenvolvida localmente [12], que consta de 40 locutores adultos (20 homens e 20 mulheres), que pronunciaram 200 frases retiradas de um trabalho realizado por Alcaim et. al. [1]. Nestas frases foram detectadas 694 palavras diferentes, o que corresponde a um vocabulário de porte médio.

Foram realizados testes com independência de locutor, somente com locutores masculinos, somente com locutores femininos, e com dependência de locutor, sendo que para este último teste foi coletado material de um locutor extra do sexo masculino.

Este sistema está sendo continuamente melhorado, e já está sendo utilizado por outros pesquisadores tanto no LPDF-UNICAMP como INATEL, no desenvolvimento de seus trabalhos em reconhecimento de fala, já tendo sido inclusive empregado no desenvolvimento de uma tese de mestrado [5].

II. DESCRIÇÃO DO SISTEMA

O sistema desenvolvido é formado por quatro módulos principais:

1. Módulo de extração de parâmetros
2. Módulo de treinamento
3. Módulo de geração de modelo de linguagem
4. Módulo de reconhecimento

A seguir, cada um destes módulos será descrito com maiores detalhes.

II.1 MÓDULO DE EXTRAÇÃO DE PARÂMETROS

Este módulo tem por entrada um sinal de voz em formato WAV ou binário puro (sem cabeçalho), e calcula parâmetros da locução. Atualmente estão disponíveis os parâmetros mel-cepstrais [3] e log-energia normalizada, bem como seus respectivos parâmetros delta e delta-delta, nas freqüências de amostragem de 8, 11,025 e 16 kHz, tanto para 8 quanto para 16 bits de resolução.

Os parâmetros são calculados utilizando-se janelas de 20 ms, atualizadas a cada 10 ms. Antes da extração, o sinal é submetido a alguns pré-processamentos: retirada do nível DC, pré-ênfase com um filtro passa altas (1-0,95 z^{-1}), e janelamento através de uma janela de Hamming [11].

Os parâmetros log-energia foram normalizados tomando como referência o quadro de maior energia em toda a locução sob análise.

Os parâmetros delta foram calculados segundo a expressão:

$$\Delta_i(n) = \frac{1}{2K+1} \sum_{k=-K}^{K} k y_{i-k}(n) \qquad (1)$$

onde:

$y_i(n)$: é o n-ésimo elemento do vetor de parâmetros y_i;

$\Delta_i(n)$: é o n-ésimo elemento do vetor delta correspondente ao vetor de parâmetros y_i;

K : é o número de quadros adjacentes de vetores de parâmetros a serem considerados no cálculo dos parâmetros delta. Neste trabalho foi utilizado $K = 1$ tanto para o cálculo dos parâmetros delta como para os delta-delta.

II.2 MÓDULO DE TREINAMENTO

Este programa tem por função treinar os modelos HMM das sub-unidades acústicas a partir de locuções de treinamento e das respectivas transcrições fonéticas.

O modelo HMM é caracterizado por uma matriz de transição $[a_{ij}]$ e por uma densidade de emissão de símbolos associada a cada estado $b_i[\ ]$. Essa densidade pode ser discreta (HMM discreto) ou contínua (HMM contínuo). No caso do sistema em consideração, são empregados modelos HMM contínuos.

Os símbolos correspondem aos vetores de parâmetros acústicos. No caso de se empregar, por exemplo, os coeficientes mel-cepstrais, delta-mel e delta-delta-mel, tem-se 3 densidades de emissão de símbolos associadas a cada estado, uma para cada vetor de parâmetros, que são pois considerados independentes entre si.

No HMM contínuo as densidades de emissão são modeladas por misturas de gaussianas multidimensionais. Assim, a densidade de emissão dos coeficientes mel-cepstrais é modelada por uma mistura de g gaussianas multidimensionais (dimensão 12).

Adicionalmente, nesse modelamento, são empregadas matrizes de covariância diagonais, que correspondem a considerar as componentes de cada vetor de parâmetros independentes entre si. Embora isto não ocorra na realidade, é uma aproximação bastante razoável. Ainda, com uma matriz diagonal, temos menos parâmetros a modelar, o que significa que com um mesmo material de treinamento, estes são estimados com maior confiabilidade.

O algoritmo de treinamento utilizado é o Baum-Welch. Uma visão geral da arquitetura deste programa é fornecida na Figura 1.

Como mostrado na Figura 1, é necessário fornecer ao programa de treinamento as seguintes informações:

- Sub-unidades acústicas a serem utilizadas na transcrição fonética das locuções de treinamento (fones independentes de contexto).
- Transcrição das locuções utilizando estas sub-unidades fonéticas.
- Locuções de treinamento (em formato WAV).

O procedimento adotado para o treinamento das sub-unidades é o seguinte: inicialmente são criados modelos HMM para cada uma das sub-unidades acústicas. A arquitetura utilizada para os HMMs é mostrada na Figura 2.

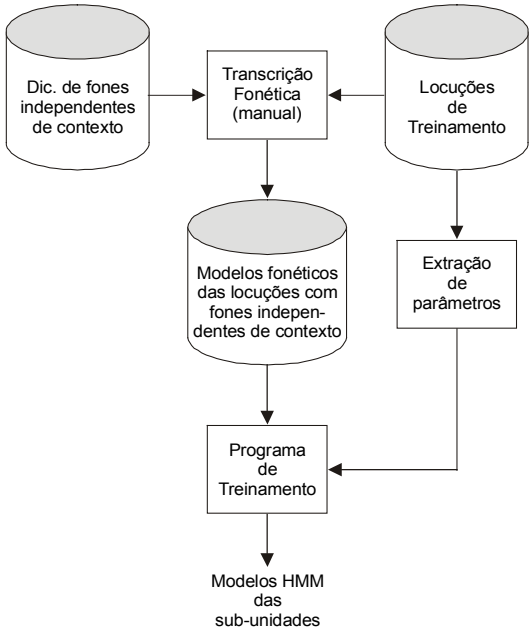

Figura 1: Esquema de funcionamento do programa de treinamento das sub-unidades com indicação das informações a serem fornecidas ao sistema.

Estes modelos são inicializados pelo algoritmo *Segmental K-Means* [11] e depois treinados via algoritmo *Baum Welch* [4]. Nas seções seguintes o treinamento dos modelos é explicado com maiores detalhes.

II.2.1 INICIALIZAÇÃO DOS MODELOS VIA SEGMENTAL K-MEANS

Para a inicialização destes modelos foi utilizado o algoritmo *Segmental K-Means*. As probabilidades de transição são inicializadas como sendo equiprováveis, como mostrado na Figura 3.

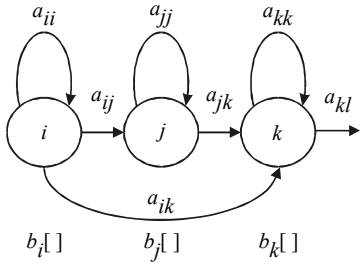

Figura 2: Modelo HMM utilizado para cada uma das sub-unidades fonéticas. A probabilidade de transição akl indica a probabilidade de fazer uma transição para a sub-unidade seguinte.

O método via *Segmental K-Means* é dividido em duas partes: segmentação uniforme e segmentação via algoritmo de Viterbi.

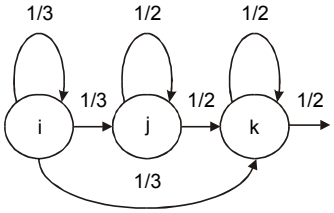

Figura 3: Valores iniciais para as probabilidades de transição dos modelos dos fones para inicialização com distribuição uniforme.

II.2.2 SEGMENTAÇÃO UNIFORME

Na etapa de inicialização, as locuções de treinamento são divididas em *m* partes iguais (de mesmo comprimento), sendo *m* o número de sub-unidades fonéticas da transcrição fonética multiplicado 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 [9] para estimar os vetores código do quantizador vetorial. Uma vez estimadas as médias, realiza-se um agrupamento (clustering) dos *n* vetores em torno destas *g* médias e calcula-se a variância correspondente a cada gaussiana.

Existe um inconveniente com este método: as sub-unidades fonéticas apresentam durações diferentes, e a segmentação uniforme pode fazer com que o modelo de um fone seja inicializado com uma porção da locução correspondente a outro fone. Isto poderia ser evitado se as locuções de treinamento fossem segmentadas, e desta forma, poderia-se inicializar os modelos de cada fone com a porção da locução a ele correspondente. Entretanto, como não se dispõe de uma base de dados segmentada, esta opção não foi implementada. Uma forma de minimizar este efeito seria fazer a inicialização com locuções mais curtas, possivelmente contendo apenas uma palavra. Desta forma, o erro introduzido pela duração não uniforme dos fones seria menor.

Um exemplo ajuda a compreender melhor o procedimento empregado. Seja uma locução de treinamento consistindo de um único fone (por exemplo o fone 'a'), e o modelo HMM correspondente cujos parâmetros desejamos estimar. De acordo com o modelo da Figura 2, o modelo HMM é constituído de 3 estados. Suponha ainda que a locução de treinamento tenha sido parametrizada com 240 quadros. Desta forma teríamos 80 quadros para cada estado. Os vetores de coeficientes associados aos 80 primeiros quadros (símbolos) irão inicializar o primeiro estado, os 80 seguintes o segundo, e os 80 símbolos finais, o último estado do modelo. Para cada estado os 80 símbolos correspondentes irão ser utilizados para estimar *g* médias pelo algoritmo LBG, e depois utilizados para o cálculo das variâncias.

Na prática, entretanto, a situação é um pouco mais complicada. Seja por exemplo uma locução consistindo apenas da palavra 'banana'. Incluindo os silêncios inicial e final, a transcrição para esta locução é dada por:

b a n an n a

Se utilizarmos o modelo HMM da Figura 2 para cada um dos fones, teremos um modelo resultante da concatenação de 8 modelos básicos, gerando um modelo composto por 24 estados. Supondo novamente uma locução parametrizada com 240 quadros, teremos 10 símbolos para inicializar as misturas de cada um dos estados do modelo. Entretanto, verifica-se no exemplo acima que existem duas ocorrências dos fones [#], [a] e [n]. Como representam o mesmo modelo, ao final do treinamento, devem apresentar os mesmos parâmetros. Neste caso, as médias e variâncias devem ser inicializadas com os símbolos correspondentes a todas as ocorrências de cada um dos fones.

Vamos analisar o caso do fone [#]: este fone ocorre no início e no final da locução. Desta forma, para estimar as médias e variâncias para este modelo, usamos os 10 primeiros e 10 últimos símbolos da locução.

Outro fato a ser considerado é que o treinamento é realizado com várias locuções. Desta forma devemos considerar para efeito de cálculo das médias e gaussianas das misturas todos os símbolos provenientes de todas as locuções de treinamento.

II.2.3 SEGMENTAÇÃO VIA ALGORITMO DE VITERBI

Nesta etapa o procedimento é parecido com o da inicialização descrito acima, com a diferença de que agora a segmentação não é uniforme, ou seja, a cada um dos estados são associados mais ou menos quadros dependendo do caminho escolhido pelo algoritmo de Viterbi [6], como mostrado na Figura 4. As probabilidades de emissão são atualizadas, como no caso anterior, pelo processamento (cálculo de covariância) dos símbolos emitidos em cada estado, e

as de transição, pelo número de quadros associados a cada estado.

II.2.4 TREINAMENTO DOS MODELOS VIA ALGORITMO BAUM-WELCH

Após a inicialização vem o treinamento propriamente dito, onde é utilizado o algoritmo *Baum-Welch*. O procedimento é similar ao da inicialização utilizando o método *Segmental K-Means*: para cada locução de treinamento é gerado um modelo HMM através da concatenação dos modelos referentes às sub-unidades acústicas da sua transcrição fonética. Este modelo composto pode então ser tratado como uma única palavra, e a locução da frase, a palavra correspondente a este modelo composto. Desta forma o algoritmo de treinamento maximiza a probabilidade de o modelo composto gerar a locução correspondente. Depois disso, os modelos individuais das sub-unidades fonéticas são separados, e as contagens geradas pelo algoritmo *Baum-Welch* são acumuladas durante todo o processo de treinamento, e somente após serem processadas todas as locuções de treinamento (uma época de treinamento), são transformadas em medidas de probabilidade.

Após cada época de treinamento, faz-se uma verificação da convergência da seguinte maneira: para cada locução de treinamento monta-se o modelo

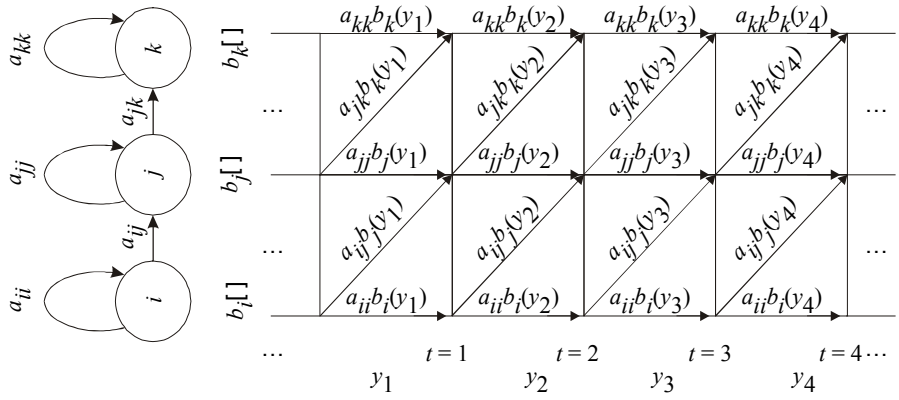

Figura 4. Exemplo de funcionamento do algoritmo de Viterbi

HMM correspondente através da concatenação dos modelos das sub-unidades fonéticas e aplica-se o algoritmo de Viterbi para calcular a probabilidade de o modelo gerar a locução correspondente. Repetindo este procedimento para todas as locuções de treinamento, pode-se calcular uma 'probabilidade média' de os modelos gerarem as seqüências de vetores acústicos (símbolos) correspondentes às locuções de treinamento. A cada época esta probabilidade média cresce até que um patamar é atingido. O treinamento é realizado até que a probabilidade média pare de crescer.

O cálculo da probabilidade média para todas as locuções demora quase tanto tempo quanto o treinamento propriamente dito. Para diminuir este tempo, a verificação da convergência foi feita apenas sobre 10 % das locuções. Testes realizados mostraram que este procedimento não modifica o número de épocas determinado por uma verificação sobre todo o conjunto de treinamento. Estas conclusões foram tiradas tendo por base conjuntos de treinamento de 600 a 1200 locuções.

II.3 MÓDULO DE RECONHECIMENTO

O módulo de reconhecimento é o responsável pelo mapeamento dos parâmetros acústicos correspondentes à locução de entrada em sua transcrição ortográfica. Foram implementados dois algoritmos de busca para o reconhecimento de fala contínua: o *Level Building* [11] e o *One Step* [8][10]. Embora ambos apresentem exatamente os mesmos resultados, o One Step tem a vantagem de ir efetuando os cálculos à medida que os parâmetros acústico da locução são calculados, ao passo que o Level Building exige que sejam calculados todos os parâmetros acústicos da locução para que possa ser iniciado o cálculo das verossimilhanças.

Para melhorar o desempenho do sistema em termos de taxa de acertos foram incluídos o modelo de duração de palavras [11] e o modelo de linguagem bigram [6]. Para o algoritmo *One Step* foi ainda implementada a estratégia Viterbi *Beam Search* [4] de poda de caminhos de baixa probabilidade. Um diagrama de blocos para este sistema é mostrado na Figura 5.

Devem ser fornecidos ao sistema, além da locução a ser reconhecida, o modelo de linguagem a ser utilizado (gramática bigram), os modelos HMM das sub-unidades acústicas, o vocabulário com o universo das palavras que podem ser reconhecidas, e o modelo de duração para cada uma das palavras. O modelo de linguagem utilizado é do tipo bigram, construído a partir das 200 frases que formam a base de dados para este trabalho. Este modelo de linguagem é bastante restritivo, pois não permite seqüências de palavras que não aquelas presentes no treinamento. Neste caso isto foi possível pois as frases que constituem o material de treinamento são as mesmas dos testes. O que muda de um conjunto para outro são apenas os locutores envolvidos. Os modelos HMM das sub-unidades são gerados pelo módulo de treinamento a partir das locuções de treinamento.

O vocabulário é armazenado em arquivo texto e contém as seguintes informações: listagem das sub-

unidades fonéticas, listagem das palavras que compõem o universo das palavras que podem ser

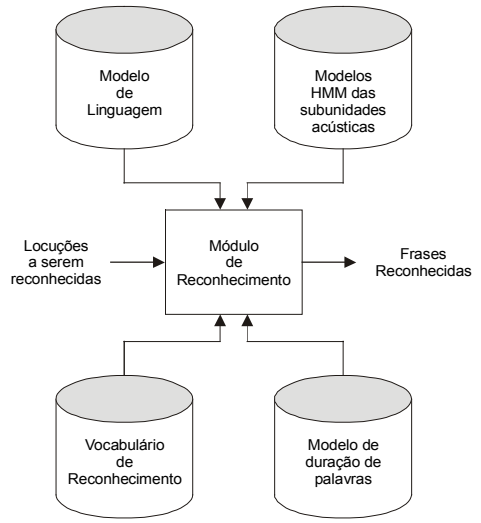

Figura 5: Diagrama de blocos do módulo de reconhecimento.

reconhecidas, juntamente com a transcrição fonética e o modelo de duração.

Para cada palavra do vocabulário foi considerada apenas uma transcrição fonética, ignorando assim os efeitos de coarticulação e o regionalismo dos locutores. Como algumas palavras apresentam várias variantes de pronúncia, foi selecionada a versão encontrada com mais frequência quando da construção da base de dados (Seção 3).

O modelo de duração foi levantado a partir das locuções do locutor correspondente aos testes com dependência de locutor. Para este fim adotou-se um modelo paramétrico, proposto por Rabiner [11]. Este associa à duração *d* de cada palavra *i* do vocabulário uma função densidade de probabilidade gaussiana $f_i(d)$

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

onde $\bar{d}_i$ e σ_i^2 são, respectivamente, a média e a variância da duração da palavra *i*. Estes valores são obtidos a partir da segmentação das locuções de treinamento.

No presente trabalho, o modelo de duração de palavras foi levantado manualmente a partir das locuções de um único locutor (o mesmo utilizado nos testes com dependência de locutor). Nos testes com independência de locutor, sempre haverá casos em que as durações das palavras nas locuções de teste estejam significativamente distantes daquelas armazenadas no modelo de duração. Isto pode fazer com que o reconhecimento seja prejudicado nestes casos, mas algum procedimento de adaptação poderia minimizar este problema.

III. BASE DE DADOS

As frases foram escolhidas segundo o trabalho realizado por Alcaim et. al. [1]. Neste, foram criadas 20 listas de 10 frases foneticamente balanceadas, segundo o português falado no Rio de Janeiro. Nestas listas, contou-se 694 palavras distintas.

Para as gravações foram selecionados 40 locutores adultos, sendo 20 homens e 20 mulheres. Para a formação do subconjunto de teste foram escolhidos, de forma aleatória, cinco locutores do sexo masculino e cinco do sexo feminino. Os demais locutores, 15 masculinos e 15 femininos, formam o subconjunto de treinamento. Nos testes com dependência de sexo, os locutores de treinamento e teste são extraídos dos subconjuntos anteriores, resultando em 5 locutores de teste e 15 de treinamento.

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

Tabela 1: Sub-unidades acústicas utilizadas na transcrição fonética das locuções.

Um locutor extra do sexo masculino completa a base de dados. Este locutor pronunciou todas as 200 frases, repetindo-as 4 vezes. Três repetições foram utilizadas para treinar o sistema e a última serviu como material de testes. Estas locuções foram utilizadas para testes com dependência de locutor.

As gravações foram realizadas em ambiente relativamente silencioso, com um microfone direcional de boa qualidade, utilizando uma placa de som SoundBlaster AWE 64. A taxa de amostragem utilizada foi de 11,025 kHz, e resolução de 16 bits. Os dados foram armazenados em formato Windows PCM (WAV).

A transcrição fonética foi feita manualmente para cada locução, utilizando programa de visualização gráfica do espectrograma e forma de onda do sinal, e fones de ouvido para audição da mesma. As sub-unidades utilizadas nesta tarefa (36 ao todo) são mostradas na Tabela .

IV. TESTES E ANÁLISE DOS RESULTADOS.

A seguir são apresentados os resultados dos testes realizados com o intuito de encontrar a configuração ótima e o desempenho final do sistema. Este desempenho foi medido levando-se em consideração

dois fatores: taxa de acertos de palavra e tempo médio de reconhecimento para uma locução.

Para todos os testes foi utilizado o algoritmo *One Step* com 15 níveis e gramática bigrama. A escolha deste número de níveis está relacionada às frases da base de dados: a frase mais longa tem 11 palavras, e contando os silêncios inicial e final, temos 13 palavras. Com 15 níveis é possível reconhecer todas as frases, e ainda verificar se ocorrem erros de inclusão, mesmo nas frases mais longas.

Foram realizados testes com e sem modelo de duração de palavras para verificar a influência desta no desempenho do sistema. Também foi variado o número de gaussianas nas misturas para verificar o detalhamento necessário nas funções densidade de probabilidade de emissão de símbolos.

Em testes preliminares foram utilizadas várias combinações de parâmetros, e os melhores resultados foram obtidos utilizando-se os parâmetros mel-cepstrais, delta-mel cepstrais e delta-delta mel-cepstrais. Para o cálculo dos parâmetros delta utilizou-se uma janela à esquerda e uma à direita. A cada um dos parâmetros foi associada uma densidade de probabilidade modelada como uma mistura de g gaussianas. Como se tem 36 sub-unidades com 3 estados cada uma e foram considerados 3 parâmetros acústicos, resultam 324 densidades de emissão, cada uma modelada como uma mistura de g gaussianas.

Foram realizados quatro testes, variando-se os locutores envolvidos:

- Teste com independência de locutor (Tabela 2)
- Teste com locutores masculinos (Tabela 3)
- Teste com locutores femininos (Tabela 4)
- Teste com dependência de locutor (Tabela 5)

# gauss	mod dur	D (%)	S (%)	I (%)	total (%)	tempo (min.)
2	não	4,19	37,52	18,26	59,97	01:48
	sim	10,65	37,60	7,69	55,94	02:46
3	não	2,40	13,47	7,61	23,48	03:01
	sim	4,72	14,57	3,39	22,68	01:58
4	não	2,09	12,98	7,46	22,52	02:10
	sim	4,68	12,60	2,70	19,98	03:09
5	não	2,17	10,73	5,90	18,80	02:49
	sim	3,88	11,83	2,74	18,46	03:14
6	não	2,02	11,45	6,81	20,28	02:12
	sim	4,03	12,06	2,32	18,42	02:05

Tabela 2. Resultados para testes independentes de locutor

Estes testes visam estabelecer a melhor configuração para o sistema e um desempenho de referência, tanto em termos de taxa de acerto de palavras como de tempo de processamento. Nas tabelas, as configurações que proporcionaram o melhor desempenho aparecem em destaque. A variação dos locutores envolvidos visa mostrar como o desempenho do sistema é degradado ao aumentarmos a faixa de variação das características dos locutores.

Nas tabelas, a primeira coluna indica o número de gaussianas por mistura, e a segunda, se o teste foi feito utilizando o modelo de duração ou não. Os símbolos 'D', 'S' e 'I' correspondem às porcentagens de erros de deleção, substituição e inserção de palavras, respectivamente. Os tempos de reconhecimento foram obtidos utilizando-se uma máquina com processador AMD-K6 de 350 MHz e 64 MB de memória RAM.

# gauss	mod dur	D (%)	S (%)	I (%)	total (%)	tempo (min.)
2	não	1,75	12,94	6,93	21,61	01:44
2	sim	4,49	13,01	2,89	20,40	01:56
3	não	1,98	9,51	5,48	16,97	02:31
3	sim	4,03	10,20	2,28	16,51	02:58
4	não	1,83	8,30	5,18	15,30	01:52
4	sim	4,26	8,52	1,52	14,31	01:51
5	não	1,60	7,61	5,10	14,31	03:03
5	sim	3,81	7,91	2,21	13,93	02:53
6	não	1,83	8,30	4,26	14,38	01:56
6	sim	3,96	8,98	1,67	14,61	02:00

Tabela 3. Resultados para testes com locutores masculinos

# gauss	mod dur	D (%)	S (%)	I (%)	total (%)	tempo (min.)
2	não	2,51	15,83	8,98	27,32	02:09
	sim	4,41	17,58	4,34	26,33	01:57
3	não	1,14	11,95	8,14	21,23	02:21
	sim	2,89	13,39	3,50	19,79	02:40
4	não	1,83	10,73	6,85	19,41	02:04
	sim	3,96	12,02	3,27	19,25	02:11
5	não	1,45	10,43	6,32	18,19	03:03
	sim	2,66	10,65	3,04	16,36	02:10
6	não	1,22	9,82	7,15	18,19	02:28
	sim	2,89	10,65	2,89	16,44	02:14

Tabela 4. Resultados para testes com locutores femininos.

IV.1 AVALIAÇÃO DOS RESULTADOS

Os resultados mostram que um número conveniente de gaussianas seria 5 para os sistemas dependente de locutor e dependente de sexo, e 6 para o sistema independente de locutor. O fato de o sistema independente de locutor necessitar de mais gaussianas para atingir o melhor desempenho parece ser razoável, visto que as variações a serem absorvidas (devido ao sexo, sotaque, etc.) são maiores. Ainda, como era de se esperar, o desempenho do sistema é tanto melhor quanto menores forem estas variações.

Excetuando-se os testes com dependência de locutor, o modelo de duração de palavras teve influência positiva no desempenho do sistema. Isto parece indicar que para um sistema mais discriminativo em termos acústicos, este modelo de duração de palavras possa talvez ser descartado.

Em relação aos tempos de processamento, o tempo médio foi em torno de 2 a 3 minutos para a decodificação de uma frase.

V. PRÓXIMAS ETAPAS

Como dito anteriormente, este sistema encontra-se em desenvolvimento permanente, e o objetivo é que cada pesquisador que venha a utilizá-lo deixe a sua contribuição para a melhoria de seu desempenho.

Os planos para o futuro envolvem a organização do léxico em árvore para diminuição do tempo de processamento [2], uso de fones dependentes de contexto [6], redução da dimensão dos vetores de

# gauss	mod dur	D (%)	S (%)	I (%)	total (%)	tempo (min.)
2	não	1,67	4,87	3,27	9,82	02:37
	sim	2,74	5,18	1,90	9,82	02:52
3	não	1,52	3,42	2,59	7,53	02:46
	sim	2,21	4,11	1,29	7,61	02:53
4	não	0,68	2,59	2,44	5,71	02:52
	sim	2,05	3,73	1,22	7,00	03:10
5	não	0,84	2,51	2,05	5,40	02:44
	sim	1,67	2,97	1,07	5,71	02:55
6	não	0,99	2,13	2,28	5,40	03:14
	sim	1,83	2,66	1,41	5,63	02:59

Tabela 5. Resultados para testes com dependência de locutor.

parâmetros via Análise de Componente Principal [7], e a expansão do modelo de linguagem de bigrama para trigrama [6], e determinação automática do número de gaussianas por mistura em cada estado, que na versão atual do sistema é mantido fixo para todos os estados.

VI. CONCLUSÕES

Foi construído e testado um sistema de reconhecimento de fala contínua baseado em modelos ocultos de Markov contínuos. Este sistema utiliza sub-unidades fonéticas como unidades fundamentais. Neste trabalho foram utilizados fones independentes de contexto, mas outras sub-unidades podem ser utilizadas, bastando para isto fornecer a transcrição fonética das locuções em termos das sub-unidades escolhidas. Os parâmetros disponíveis até o momento são os mel-cepstrais e log-energia normalizada, com suas respectivas derivadas primeira e segunda (parâmetros delta e delta-delta). Para os parâmetros delta e delta-delta pode-se também escolher o número de quadros adjacentes a serem considerados.

No modo independente de locutor, e para um vocabulário de aproximadamente 700 palavras, a taxa de erros de palavra obtida foi de 18,42 %, com um tempo médio de reconhecimento por volta de 2 minutos em uma máquina com processador AMD-K6 de 350 MHz e 64 MB de memória RAM, rodando sob a plataforma Windows. Este desempenho pode ser aumentado consideravelmente se os locutores forem separados por sexo, atingindo assim uma taxa de erros de palavra de 13,93 % para os locutores masculinos e 16,36 % para os locutores do sexo feminino.

O modelo de duração de palavras proporcionou uma ligeira melhora no desempenho do sistema em todos os testes, exceto no dependente de locutor. Uma das razões possíveis para este efeito é que quando se tem um sistema cuja decodificação acústica é eficiente, o modelo de duração não é necessário. Este é um bom resultado, visto que o modelo de duração impõe uma carga adicional no processamento. Além disso, o levantamento do modelo de duração para vocabulários extensos pode ser bastante trabalhoso.

Este sistema deve servir de base para pesquisas futuras em reconhecimento de fala, tendo já sido inclusive empregado para o desenvolvimento de uma tese de mestrado [5]. Os resultados das pesquisas desenvolvidas deverão ser incorporadas a este sistema base, de modo que o mesmo estará em constante desenvolvimento.

Agradecimentos: Este trabalho foi parcialmente financiado pela FAPESP, através de Bolsa de Pós-Doutoramento, processo 99/01241-2.

VII. REFERÊNCIAS

[1] Alcaim, A., Solewicz, J. A., Moraes, J. A. "Freqüência 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, 1992.

[2] Aubert,X., Dugast, C., Ney, H., and Steinbiss, V. "Large vocabulary continuous speech recognition on Wall Street Journal data". *Proceedings of the 1994 International Conference on Acoustics, Speech and Signal Processing*, Adelaide, Australia, 1994, volume 2, pages 129-132.

[3] 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.

[4] Deller Jr., J. R., Proakis, J. G., Hansen, J.H.L. *Discrete time processing of speech signals*. MacMillan Publishing Company, New York, 1993.

[5] Dias, R. S. F. "Normalização de locutor em sistema de reconhecimento de fala". *Tese de Mestrado*. UNICAMP. Campinas. 2000.

[6] Jelinek, F. *Statistical methods for speech recognition*. MIT Press, London, 1998.

[7] Johnson, R. A., Wichern, D. W. *Applied multivariate statistical analysis*. Prentice Hall, New Jersey, 1992.

[8] Lee, C. H., Rabiner, L. "A frame-synchronous network search algorithm for connetcted word recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11), november, 1989.

[9] Linde, Y., Buzo, A., Gray, R. M. "An algorithm for vector quantizer design". *IEEE Transactions on Communications*, COM-28(1), january, 1980.

[10] Ney, H. "The use of a one-stage dynamic programming algorithm for connected word recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-32(2), april, 1984.

[11] Rabiner, L. *Fundamentals of speech recognition*. Prentice Hall Press, 1993

[12] Ynoguti, Carlos Alberto. "Reconhecimento de fala contínua usando modelos ocultos de Markov". *Tese de Doutorado*, UNICAMP, Campinas, 1999.

SOBRE OS AUTORES

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). No período de 1999 a 2000 desenvolveu atividade de Pós-doutoramento na mesma instituição. Atualmente é professor junto ao Instituto Nacional de Telecomunicações – INATEL. Suas áreas de interesse são Processamento Digital de Sinais e Reconhecimento de Fala.

E-mail: ynoguti@inatel.br

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.

E-mail: fabio@decom.fee.unicamp.br