DETERMINAÇÃO AUTOMÁTICA DE NÚMERO DE ESTADOS PARA *CDHMM*

M. A. R. ANDRADE E S.C.B. SANTOS

Ministério da Defesa – Exército Brasileiro
Centro de Desenvolvimento de Sistemas – CDS
Esplanada dos Ministérios – Bloco O – 9º andar
Brasília – DF – Brasil, CEP: 70065-900
Tel: (61)9958-5426, rocca@cds.eb.mil.br

Sistema de Proteção da Amazônia – SIPAM
Secretaria Executiva do SIPAM – CONSIPAM
SPO, Área 5, Quadra 3, Bloco J
Brasília – DF – Brasil, CEP: 70610-200
Tel: (61)411-5282 sidney@defesa.gov.br

SUMÁRIO

É apresentado um sistema para determinação automática de número de estados de *CDHMM* para treinamento de modelos de palavras, dependente de locutor. São usados coeficientes PLP-Ceps e energia de curto período, como atributos da voz. É utilizado um vocabulário de 45 palavras para treinamento e testes relacionadas com deslocamentos de objetos no espaço. É apresentado o método de determinação de limiar para contagem de estados necessários e os resultados atingidos. São comparadas as verossimilhanças de variantes do sistema proposto. O sistema proposto é comparável à determinação manual.

1. INTRODUÇÃO

Este trabalho tem por objetivo: propor, implementar e avaliar o desempenho de um sistema para determinação automática do número de estados de *CDHMM (Continuous Density Hidden Markov Model)* [1,2,3,4,5,6,7,8] necessários para modelagem acústica de conjuntos de comandos vocais conectados, em língua portuguesa, voltado para o controle de deslocamentos de objetos genéricos no espaço e acionamentos eletromecânicos [8], com emprego de coeficientes PLP-Ceps como atributos da voz [9].

A seção 2 apresenta as etapas de pré-processamento da voz para tornar o sinal tratável pela parte de determinação de limiares e treinamento do sistema. Apresenta também os atributos de voz empregados no restante do trabalho. A seção 3 apresenta o vocabulário específico para a tarefa de deslocamento e acionamento de dispositivos eletromecânicos. Apresenta também a forma de determinação não automática do número de estados de *CDHMM* necessários à correta modelagem de palavras e os parâmetros de modelagem. A seção 4 descreve o método e suas variantes na determinação automática de limiares de decisão propostas, a subseqüente determinação do número de estados e modificações nos parâmetros iniciais de modelagem usados na seção 3. A seção 5 apresenta os resultados obtidos, avalia o desempenho do sistema e expõe as conclusões atingidas.

2. PRÉ-PROCESSAMENTO DA VOZ

Todos os sinais foram adquiridos com uma placa de som da marca *Sound Blaster*, com taxa de amostragem de 11.025 Hz, 16 *bits* por amostra, em mono-canal, com o controle automático de ganho incorporado ativado, e com microfone acoplado mecanicamente a fones de ouvido da marca *Boeder*. As amostras de sinais de voz foram agrupadas em quadros de 10 milisegundos e janelado com superposição de 50% entre quadros adjacentes (20 milisegundos por janela).

Para o presente trabalho, com base na literatura de referência[5,6,8,9], optou-se por fazer uso de coeficientes *PLP-Cepstrum* oriundos da análise de *predição linear perceptiva (Perceptual Linear Predictive PLP)*[9], tendo em vista os resultados atingidos, em tarefas de reconhecimento, superiores a outros atributos [5,9]. Além dos 5 primeiros coeficientes PLP-Cepstrum, fez-se uso do atributo energia de tempo curto e das 1ªs e 2ªs derivadas, totalizando 18 atributos para cada janela de sinal de voz. O método para determinação de pontos terminais (*end-points*), os ajustes do método de extração de coeficientes *PLP-Ceps (Perceptual Linear Prediction - Cepstrum)* [9] e o tratamento para redução de erros de quantização [10] estão explicados em [11].

3. PROCEDIMENTO NÃO AUTOMÁTICO

A tarefa para a qual destina-se o modelo de CDHMM a serem treinados é a de realizar controle de deslocamentos de objetos genéricos no espaço e acionamentos eletromecânicos [8]. Para tal tarefa, o vocabulário contém os dígitos para designação de objetos e determinação de quantidades de deslocamento. O vocabulário contém também palavras que ajudem na designação de objetos genéricos. Foram incluídas também as direções, unidades de deslocamento e ações possíveis.

Uma das primeiras etapas da modelagem de palavras por *CDHMM* é a determinação de quantos estados serão usados para representar cada palavra [5,6,7,8,12,13]. A existência de palavras de vários comprimentos dificulta o uso de um número fixo de estados para todas as palavras. Como exemplo, o número de estados adotados para a palavra ‘a’ foi 3, enquanto para a palavra ‘centímetros’ adotou-se 15 estados. O número de estados para cada palavra do vocabulário foi fixado com base nos fonemas presentes, usando-se em geral 3 estados por sílaba da língua portuguesa [5,14].

Para várias palavras foram acrescidos alguns estados para abrangerem diferentes pronúncias da letra ‘r’ e de encontros vocálicos, visando tornar o modelo menos sensível às variações de pronúncia. A distribuição de estados apresentados na Tabela 1 apresenta-se de acordo com o encontrado na literatura de referência [5,8]. Para efeito de simplificação de modelos, o par

'para_a' é considerado como palavra única [8]. A Tabela 1 apresenta as 45 palavras do vocabulário de trabalho.

Tabela 1. Vocabulário [8].

Nº	Palavra	Nº	Palavra	Nº	Palavra
1	zero	16	circuito	31	passo
2	um	17	dispositivo	32	grau
3	uno	18	ande	33	metros
4	dois	19	mova	34	centímetros
5	três	20	gire	35	passos
6	quatro	21	vire	36	graus
7	cinco	22	rode	37	para_a
8	seis	23	ligue	38	à
9	meia	24	desligue	39	pra
10	sete	25	pare	40	esquerda
11	oito	26	abra	41	direita
12	nove	27	feche	42	cima
13	motor	28	volte	43	baixo
14	unidade	29	metro	44	frente
15	sistema	30	centímetro	45	trás

Pode-se notar neste ponto a quantidade de operações em que a subjetividade humana é empregada para tentar incorporar ao modelo o conhecimento da fala e facilitar a tarefa de decodificação a ser realizada por uma máquina.

Para o locutor único, foram realizadas gravações de 30 repetições de cada palavra do vocabulário, de forma isolada, sendo 25 para treinamento e 5 para teste dos modelos. Também foram gravadas 131 frases de treinamento contendo números variáveis de palavras do vocabulário formando sentenças válidas de comandos conectados [8].

Para a modelagem e os testes foram adotados o modelo de Bakis [1,2], 5 gaussianas por estado, 18 atributos da voz (conforme Seção 2), 10 interações máximas para o algoritmo *Segmental k-means* [2,5,6] e um fator de convergência de 0,1% para o término dos treinamentos dos modelos de *CDHMM*. O conjunto de modelos gerados com estes parâmetros foi chamado de *fon* (modelo originado por análise fonética).

4. PROCEDIMENTO AUTOMÁTICO

A idéia de implementar um procedimento automático de determinação do número de estados a serem modelados em *CDHMM* tem como objetivo retirar a interferência de mais um operador no processo de treinamento de sistemas para reconhecimento de voz orientado para tarefas. Visa também facilitar modificações de vocabulário feitas pelo próprio usuário do sistema e permitir o uso das mesmas elocuções de treinamento como objeto de estudo e obtenção de limiares de decisão.

Na determinação usual do número de estados é necessário um estudo fonético, atribuindo-se geralmente 3 estados para cada fonema ou sílaba (início–meio–fim) [2,5,14].

Na determinação automática proposta, o próprio sistema faz seu estudo das variações acústicas de modo semelhante ao estudo fonético. É feita uma análise das janelas de atributos das elocuções a serem modeladas tentando-se determinar onde ocorrem as transições entre fonemas e entre partes de um mesmo fonema. Várias repetições de um vocábulo são usadas para tornar o limiar tolerante à variação de pronúncia do locutor. De modo semelhante a um método de determinação de *end-points* [11], é usado como limiar de decisão o valor de 2 vezes o desvio padrão (*dp*). Este *dp* é obtido do conjunto de diferenças de distâncias euclidianas entre vetores de atributos adjacentes no tempo das elocuções de treinamento. Exemplo: seja uma elocução com 43 quadros de 10 ms; extrai-se 43 vetores de atributos com 18 elementos cada; monta-se, então, um vetor de 42 elementos com as distâncias euclidianas entre vetores de atributos adjacentes dois a dois; depois é determinado o desvio padrão dessas distâncias vezes 2 para definição de um limiar; esse limiar é usado no processo de contagem de estados necessários para modelar essa elocução em particular (Fig. 1).

São testados três modos para definir os limiares de determinação do número de estados de cada palavra a ser modelada:

- Limiar individual (*li*): onde cada conjunto de 25 elocuções de uma mesma palavra de treinamento são analisados para obtenção de um limiar diferente para cada vocábulo;
- Limiar geral (*lg*): onde todas as 25 repetições das 45 palavras são analisadas juntas e fornecem um limiar único para todas as palavras;
- Limiar de frases (*lf*): onde as frases de treinamento são analisadas para gerar um limiar único para determinação das transições acústicas das palavras isoladas a serem modeladas.

Após a determinação do limiar, analisam-se as elocuções de treinamento e verificam-se quantas vezes a distância euclidiana entre janelas adjacentes superou o valor do limiar. Distâncias abaixo do limiar representam poucas variações acústicas entre os atributos das janelas verificadas e indica possivelmente a continuidade de um mesmo fonema. Caso o limiar seja atingido ou superado, isto indica uma diferença acústica maior entre as janelas e possivelmente uma transição entre fonemas ou entre partes de um fonema, sendo então incrementada a contagem de estados de *CDHMM* necessária para representar essa elocução. Caso ocorram passagens seguidas pelo valor de limiar de distâncias entre janelas adjacentes, estas transições são consideradas como parte de algum fonema com fortes variações acústicas de pequena duração (apenas uma janela) e não são consideradas para efeito de contagem de nº de estados. Após encontrar a próxima diferença abaixo do limiar, a contagem de estados necessários é incrementada apenas uma vez. De posse dos números de estados necessários para cada elocução, pode-se extrair a média do número de estados para o vocábulo correspondente às elocuções e também o respectivo desvio padrão.

Estuda-se neste trabalho as seguintes variantes do procedimento automático:

- Os números de estados serão às médias dos números de estados atribuídos a cada elocução do respectivo vocábulo utilizando cada um dos 3 modos de obtenção de limiares de decisão;
- Os números de estados para modelagem serão os valores das médias mais os desvios padrão (*dp)* dos números de estados obtidos das elocuções dos respectivos vocábulos com uso dos 3 modos de determinação de limiar;
- Os números de estados serão os mesmos do parágrafo anterior, mas com uma modificação nos parâmetros de modelagem: o modelo de Bakis (com transição esquerda-

direita até o 2º estado) será substituído por um modelo em que seja possível a transição até o 3º estado seguinte no *CDHMM*.

Os acréscimos dos *dp* às médias e dos aumentos de possibilidades de transições visam permitir que os modelos a serem gerados, mais longos que a média, possuam características mais robustas às variações de pronúncias de um mesmo vocábulo, já visando a aplicação do sistema no modo independente do locutor. Assim, como exemplo, partes de fonemas sibilantes, omitidas durante a pronúncia em uma eventual elocução, poderiam ser mais bem representadas nestes modelos. Fato semelhante pode ocorrer com diferentes articulações da pronúncia do fonema /r/.

Ao todo, 9 variantes de determinação automática de estados são testadas. Seus conjuntos de valores são designados a seguir:
- *mi* : médias obtidas pelo limiar *li* e modelo de Bakis;
- *mg* : médias obtidas pelo limiar *lg* e modelo de Bakis;
- *mf* : médias obtidas pelo limiar *lf* e modelo de Bakis;
- *mdi* : média + *dp* pelo limiar *li* e modelo de Bakis;
- *mdg* : média + *dp* pelo limiar *lg* e modelo de Bakis;
- *mdf* : média + *dp* pelo limiar *lf* e modelo de Bakis;
- *mdi3*: conforme *mdi*, porém com transição até 3 estados;
- *mdg3*: conforme *mdg*, porém com transição até 3 estados;
- *mdf3*: conforme *mdf*, porém com transição até 3 estados.

Todas as 9 variações são comparadas e normalizadas pelas verossimilhanças obtidas pelo procedimento não automático do item 3, denominado por *fon*.

5. RESULTADOS E CONCLUSÕES

O procedimento geral de obtenção de resultados para cada variante foi o seguinte:
- Determinação do limiar com o estudo das elocuções de treinamento (palavras ou frases, dependendo da variante);
- Determinação do número de estados para cada vocábulo com o uso do limiar;
- Treinamento dos 45 modelos de *CDHMM*;
- Obtenção das verossimilhanças das elocuções de treinamento;
- Obtenção da verossimilhança total de final de treinamento;
- Obtenção da taxa total de acerto com uso das 225 elocuções de teste.
- Obtenção das verossimilhanças das elocuções de teste;
- Obtenção da verossimilhança total de final de teste.

As Tabelas 2 e 3 apresentam os limiares *li*, *lg* e *lf* obtidos das elocuções de treinamento. A Tabela 4 consolida todos os valores dos conjuntos de números de estados a serem modelados nas variantes de determinação automática e no procedimento não automático. A Tabela 5 apresenta as verossimilhanças globais de *CDHMM* de treinamento das 9 variantes analisadas, normalizadas pela verossimilhança do procedimento não automático *fon*. Já a Tabela 6 apresenta a verossimilhança de treinamento de cada palavra, normalizada pela verossimilhança atingida pelo mesmo vocábulo na variante *fon*. A Tabela 7 apresenta a classificação de cada variante de determinação de número de estados utilizando como parâmetro o valor da verossimilhança de treinamento, incluindo o procedimento não automático *fon*.

Por meio dos dados da Tabela 6 é possível contar o número de vezes que cada variação obteve o melhor resultado (maior verossimilhança) para cada vocábulo, o número de segundos lugares e assim sucessivamente, até o número de décimas colocações atingidas.

Tabela 2. Limiares de decisão *li* (por vocábulo).

Pal	Limiar	Pal	Limiar	Pal	Limiar	Pal	Limiar
1	0,3088	13	0,3674	25	0,4347	37	0,4530
2	0,4050	14	0,3108	26	0,4398	38	0,4582
3	0,3335	15	0,3277	27	0,3197	39	0,5675
4	0,2753	16	0,3553	28	0,2718	40	0,3369
5	0,4965	17	0,2937	29	0,4117	41	0,3832
6	0,4399	18	0,3703	30	0,3394	42	0,2729
7	0,4208	19	0,2178	31	0,3634	43	0,3081
8	0,2938	20	0,2902	32	0,3251	44	0,4205
9	0,2454	21	0,3365	33	0,3857	45	0,4890
10	0,3667	22	0,2058	34	0,3534		
11	0,3807	23	0,2481	35	0,3641		
12	0,2237	24	0,2227	36	0,2769		

Tabela 3. Limiares de decisão *lg* e *lf*.

Grupo	*lg*	*lf*
Limiar	0,3592	0,2152

Cada conjunto de modelos foi usado para reconhecer as elocuções de teste. Somente o conjunto obtido com a variação *mdi3* apresentou erro de reconhecimento, reduzindo a taxa de acerto para 99,56%. As demais variações, incluindo *fon*, apresentaram 100% de acerto para o conjunto de 255 elocuções de teste.

As variações *mi*, *mf* e *mdf3* apresentaram verossimilhanças globais inferiores a variante *fon*. A variação *mf* apresentou o pior desempenho obtendo a última colocação na comparação das verossimilhanças de fim de treinamento por vocábulo em 80% das 45 possíveis (Tabela 7). As variações *mi*, *mg* e *mdf3* não obtiveram bons resultados por estarem mais presentes nas 8ª, 7ª, e 9ª posições respectivamente. Os valores obtidos com as elocuções de testes e apresentados nas Tabelas 8, 9 e 10 confirmam esse resultados. Por isso, estas 4 variantes, juntamente com *mdi3*, são excluídas das próximas análises.

A Tabela 11 apresenta a classificação obtida pela contagem do número de vezes que cada variante selecionada atingiu o maior valor, o segundo maior, até o número de vezes que atingiu a quinta colocação de verossimilhança para cada vocábulo. As variantes *mdf* e *mdg3* alcançaram um número maior de primeiras colocações, porém também apresentaram uma quantidade significativa de quartos lugares. A variante *mdi* apresenta uma classificação mais uniforme entre as colocações a partir da segunda colocação. A variante *mdg* obteve valores elevados de segundas e terceiras colocações e a mais baixa ocorrência de quintos lugares. A variante *fon* obteve o pior resultado.

Devido ao maior valor de verossimilhança de final de treinamento da Tabela 5, e aos resultados da Tabela 11 e a taxa de acerto atingida com as elocuções de teste, a variante *mdg* é a mais indicada para um procedimento automático de determinação do número de estados necessários para modelagem de *CDHMM* em comandos conectados dependente do locutor.

Tabela 4. Nº de estados por vocábulo por variante do método.

	Variantes						
Palavra	***fon***	***Mi***	***mg***	***mf***	***mdi / mdi3***	***mdg / mdg3***	***mdf / mdf3***
1	9	7	7	5	9	9	7
2	3	3	3	4	4	4	6
3	6	7	8	4	9	10	7
4	6	8	9	6	10	11	8
5	9	6	7	6	7	8	7
6	9	7	7	5	9	8	7
7	6	8	8	4	10	10	6
8	6	9	9	6	12	11	8
9	9	7	8	7	11	11	10
10	6	9	9	4	12	12	5
11	9	8	8	5	10	10	6
12	6	7	9	6	9	12	9
13	9	10	10	3	14	13	4
14	12	7	8	3	9	11	5
15	12	11	12	5	15	14	7
16	15	10	11	4	13	13	6
17	15	6	9	4	9	12	5
18	6	5	5	4	7	7	5
19	6	5	10	5	7	13	7
20	6	8	8	5	9	10	6
21	6	6	7	5	8	8	7
22	6	3	12	4	5	15	5
23	6	5	8	3	6	11	5
24	9	5	11	5	7	14	7
25	9	5	5	4	7	7	5
26	9	7	6	5	9	9	7
27	6	8	9	5	10	11	7
28	6	7	10	6	9	13	8
29	9	8	9	6	10	11	8
30	15	10	10	4	14	14	7
31	6	7	7	5	9	8	7
32	6	6	6	4	8	8	6
33	9	11	11	6	13	13	8
34	15	12	12	4	16	16	6
35	6	10	10	5	12	12	7
36	6	10	10	9	12	12	11
37	9	5	5	6	8	7	8
38	3	4	4	4	5	6	6
39	6	4	5	3	5	6	4
40	12	10	11	2	13	14	4
41	12	9	8	5	12	10	7
42	6	7	10	4	10	12	7
43	9	7	9	4	9	11	5
44	9	7	7	4	9	9	6
45	6	6	7	5	7	9	7

Analisando o número de estados obtidos com *mdg* na Tabela 4 e comparando valores da Tabela 6, conclui-se que em apenas 9% dos casos o número de estados obtidos é igual ao sugerido pela variante não automática *fon*. Em 20% dos casos o valor de *mdg* é menor que o sugerido em *fon*, em 71% dos casos é maior, sendo que em 20% das vezes o valor encontrado é igual ou superior ao dobro do sugerido. De uma maneira geral, as distribuições de classificação da variante *mdg* em relação as demais se mantiveram proporcionais aos valores da Tabela 11 quando avaliado somente os vocábulos que apresentaram número de estados menor, igual ou maior que o sugerido por *fon*. Verifica-se com a Tabela 6 que na maioria dos casos onde o valor do número de estados encontrado é menor que o sugerido por *fon*, esta é a que possui melhores verossimilhanças para os vocábulos treinados. Pode-se verificar que as duas ocorrências de quintos lugares da variante *mdg* ocorrem com vocábulos curtos e de terminações semelhantes contendo a letra 'r' em 'zero' e 'quatro', indicando uma tendência da variante que necessita de investigação mais detalhada.

Tabela 5. Verossimilhança normalizada de fim de treinamento.

Variantes									
fon	***mi***	***mg***	***mf***	***mdi***	***mdg***	***mdf***	***mdi3***	***mdg3***	***mdf3***
1	0,981	1,006	0,896	1,020	1,035	1,021	1,019	1,034	0,960

Tabela 6. Verossimilhanças de fim de treinamento / vocábulos.

	Variantes									
Pal.	***fon***	***mi***	***mg***	***mf***	***mdi***	***mdg***	***mdf***	***mdi3***	***mdg 3***	***mdf3***
1	1	0.968	0.962	0.923	0.997	0.994	0.995	0.996	0.996	0.964
2	1	1.010	1.000	1.046	1.067	1.059	1.047	1.052	1.067	1.129
3	1	1.020	1.035	0.930	1.031	1.074	1.030	1.040	1.058	1.013
4	1	1.023	1.032	0.975	1.031	1.053	1.042	1.043	1.055	1.026
5	1	0.930	0.967	0.956	0.965	0.993	0.960	0.960	0.986	0.969
6	1	0.957	0.959	0.900	1.001	0.992	1.003	1.004	0.994	0.960
7	1	1.054	1.058	0.891	1.085	1.082	1.098	1.098	1.093	1.004
8	1	1.031	1.046	1.006	1.053	1.036	1.062	1.062	1.043	1.016
9	1	0.981	0.990	0.975	1.029	1.026	1.027	1.029	1.020	1.019
10	1	1.097	1.065	0.927	1.095	1.087	1.140	1.110	1.111	0.993
11	1	1.035	1.026	0.902	1.038	1.032	1.043	1.036	1.042	0.951
12	1	1.034	1.059	1.005	1.059	1.096	1.065	1.062	1.098	1.059
13	1	1.020	1.040	0.801	1.070	1.070	1.036	1.075	1.052	0.825
14	1	0.924	0.933	0.702	0.955	0.977	0.965	0.967	0.980	0.865
15	1	0.990	1.009	0.822	1.027	1.030	1.032	1.032	1.024	0.905
16	1	0.935	0.980	0.740	0.983	0.972	0.970	0.974	0.969	0.867
17	1	0.845	0.928	0.750	0.930	0.983	0.925	0.926	0.958	0.808
18	1	0.976	0.977	0.946	1.029	1.024	1.026	1.024	1.015	0.977
19	1	0.993	1.089	0.990	1.051	1.125	1.047	1.049	1.114	1.041
20	1	1.059	1.057	0.986	1.071	1.072	1.073	1.066	1.075	1.001
21	1	1.001	1.029	0.984	1.058	1.056	1.054	1.056	1.073	1.039
22	1	0.877	1.124	0.892	0.968	1.123	0.977	0.968	1.201	0.988
23	1	0.988	1.046	0.866	0.999	1.087	0.992	0.998	1.095	0.984
24	1	0.865	0.991	0.868	0.924	1.037	0.929	0.926	1.037	0.924
25	1	0.937	0.932	0.900	0.968	0.968	0.966	0.966	0.966	0.935
26	1	0.952	0.962	0.929	1.000	1.000	0.992	1.001	0.997	0.967
27	1	1.024	1.029	0.982	1.047	1.071	1.034	1.044	1.066	1.024
28	1	0.999	1.057	0.986	1.047	1.081	1.047	1.045	1.087	1.053
29	1	0.990	1.002	0.944	1.004	1.041	1.003	0.994	0.997	1.003
30	1	0.930	0.940	0.686	1.002	1.003	1.021	1.001	0.997	0.862
31	1	1.016	1.017	0.950	1.041	1.037	1.048	1.033	1.033	1.022
32	1	0.974	0.986	0.886	1.043	1.051	1.042	1.047	1.033	0.971
33	1	1.015	1.017	0.942	1.029	1.025	1.019	1.028	1.011	0.979
34	1	0.960	0.968	0.691	1.009	0.998	0.997	0.994	1.016	0.774
35	1	1.067	1.065	0.986	1.086	1.084	1.073	1.086	1.086	1.020
36	1	1.083	1.070	1.059	1.118	1.113	1.115	1.105	1.110	1.080
37	1	0.947	0.952	0.960	0.994	0.981	0.996	0.996	0.979	0.996
38	1	1.031	1.034	1.036	1.049	1.068	1.050	1.050	1.070	1.070
39	1	0.954	0.983	0.905	0.976	0.996	0.988	0.982	0.993	0.954
40	1	0.975	0.997	0.682	1.015	1.029	1.024	1.017	1.023	0.773
41	1	0.930	0.929	0.856	0.968	0.941	1.004	0.970	0.937	0.905
42	1	1.015	1.079	0.933	1.060	1.076	1.070	1.076	1.081	1.015
43	1	0.968	1.004	0.824	0.996	1.001	1.005	0.999	0.977	0.878
44	1	0.918	0.909	0.843	1.017	1.014	0.994	0.993	0.994	0.908
45	1	0.992	1.021	0.965	1.021	1.041	0.996	0.998	1.045	1.022

Tabela 7. Classificação de treinamento (%) das 10 variantes

Class	**1os**	**2os**	**3os**	**4os**	**5os**	**6os**	**7os**	**8os**	**9os**	**10os**
fon	18	2	7	11	2	11	4	13	20	11
mi	0	0	0	2	2	9	20	38	20	9
mg	0	7	9	7	18	7	27	18	9	0
mf	0	0	0	0	0	0	4	2	13	80
mdi	11	22	16	16	11	18	7	0	0	0
mdg	13	24	33	9	7	13	0	0	0	0
mdf	16	16	9	9	27	16	7	2	0	0
mdi3	13	11	11	27	11	15	4	7	0	0
mdg3	24	18	13	11	18	4	11	0	0	0
mdf3	5	0	2	8	4	7	16	20	38	0

Conclui-se que é possível implementar um sistema automático de determinação de número de estados necessários para modelagem de palavras com *CDHMM* dependente do locutor que apresente resultados gerais e por vocábulos semelhantes ou superiores (em verossimilhança de treinamento) aos obtidos em procedimento não automático usual e que faça uso das mesmas elocuções que serão usadas para treinamento.

Tabela 8. Verossimilhanças de fim de teste por vocábulos.

	Variantes									
Pal.	***fon***	***mi***	***mg***	***Mf***	***mdi***	***mdg***	***mdf***	***mdi3***	***mdg 3***	***mdf3***
1	1	1.000	1.001	0.978	1.002	1.014	1.001	1.003	1.008	0.998
2	1	1.009	1.000	1.049	1.069	1.064	1.039	1.049	1.071	1.109
3	1	0.989	1.003	0.966	1.020	1.042	1.024	1.019	1.030	0.991
4	1	1.006	1.010	0.964	1.007	1.010	1.017	1.013	1.022	1.012
5	1	0.947	0.977	0.970	0.976	0.992	0.980	0.980	0.990	0.983
6	1	0.958	0.974	0.919	1.009	0.998	1.006	1.001	0.993	0.963
7	1	1.025	1.029	0.907	1.072	1.074	1.077	1.075	1.073	0.990
8	1	1.024	1.017	0.995	1.022	1.004	1.017	1.029	1.018	1.011
9	1	0.989	0.995	0.992	1.010	1.007	1.010	0.994	1.009	0.987
10	1	1.016	1.030	0.953	1.021	1.022	1.032	1.012	1.032	0.995
11	1	0.993	0.993	0.931	1.017	1.016	1.023	1.022	1.022	0.954
12	1	1.030	1.036	1.004	1.050	1.045	1.041	1.045	1.054	1.030
13	1	1.033	1.036	0.835	1.042	1.043	1.018	1.060	1.019	0.855
14	1	0.967	0.956	0.761	0.973	0.976	0.985	0.984	0.977	0.922
15	1	0.987	1.000	0.848	1.016	1.004	1.011	1.007	1.000	0.946
16	1	0.961	0.967	0.782	0.980	0.977	0.982	0.982	0.977	0.898
17	1	0.885	0.950	0.789	0.952	1.001	0.948	0.948	0.972	0.853
18	1	0.990	0.988	0.950	1.017	1.024	1.026	1.020	1.014	0.992
19	1	0.980	1.042	0.982	1.002	1.037	1.008	1.003	1.022	1.003
20	1	1.047	1.053	1.008	1.057	1.063	1.060	1.055	1.067	1.011
21	1	1.005	1.000	0.977	1.027	1.026	1.019	1.021	1.016	0.995
22	1	0.947	1.039	0.938	1.023	1.017	1.008	1.022	0.983	1.033
23	1	1.024	1.027	0.958	1.001	0.998	0.998	1.000	0.992	1.027
24	1	0.951	1.039	0.973	0.987	1.059	0.992	0.990	1.042	0.984
25	1	1.008	1.003	0.973	0.995	0.986	0.985	0.986	1.001	0.998
26	1	0.971	0.977	0.959	0.994	0.996	0.992	0.997	0.994	0.982
27	1	1.033	1.046	0.988	1.041	1.050	1.033	1.038	1.040	1.041
28	1	0.974	1.043	0.982	1.033	1.050	1.028	1.021	1.053	1.040
29	1	1.014	0.998	0.969	1.004	1.037	1.018	1.000	0.991	0.997
30	1	0.943	0.954	0.725	0.990	0.990	0.988	0.983	0.988	0.889
31	1	1.017	1.015	0.976	1.024	1.018	1.015	1.021	1.023	1.018
32	1	0.998	0.990	0.935	1.003	1.006	1.004	1.006	0.997	0.994
33	1	1.014	1.015	0.955	1.023	1.014	1.017	1.020	0.997	0.994
34	1	0.972	0.986	0.708	0.984	0.973	0.988	0.976	0.985	0.793
35	1	1.034	1.039	0.980	1.042	1.049	1.036	1.043	1.043	1.008
36	1	1.043	1.035	1.037	1.047	1.055	1.045	1.038	1.049	1.026
37	1	0.971	0.972	0.984	0.994	0.981	0.990	0.991	0.981	0.982
38	1	1.013	1.015	1.011	1.024	1.038	1.023	1.025	1.035	1.046
39	1	0.978	0.990	0.939	0.985	1.010	0.994	0.985	1.006	0.978
40	1	0.993	0.983	0.726	1.007	1.027	1.037	1.021	1.004	0.812
41	1	0.989	0.974	0.955	1.002	0.982	0.998	1.000	0.990	0.979
42	1	0.969	0.968	0.918	0.951	0.939	0.977	0.986	0.938	0.984
43	1	0.964	1.002	0.846	0.999	0.969	1.002	1.005	0.938	0.886
44	1	0.948	0.953	0.894	1.020	1.005	0.995	1.008	1.010	0.930
45	1	1.007	1.033	0.988	1.012	1.047	0.997	1.009	1.038	1.018

Tabela 9. Verossimilhança normalizada de fim de teste.

Variantes									
fon	***mi***	***mg***	***mf***	***mdi***	***mdg***	***mdf***	***mdi3***	***mdg3***	***mdf3***
1	0,992	1,005	0,928	1,013	1,018	1,010	1,012	1,013	0,976

Tabela 10. Classificação de teste (%) das 10 variantes.

Class	**1os**	**2os**	**3os**	**4os**	**5os**	**6os**	**7os**	**8os**	**9os**	**10os**
fon	18	2	4	10	11	10	10	16	11	11
mi	2	2	4	0	7	13	22	27	11	11
mg	7	4	13	4	9	20	18	13	11	0
mf	0	0	0	0	4	0	2	2	16	76
mdi	18	7	13	29	16	7	9	2	0	0
mdg	22	22	7	16	13	4	4	11	0	0
mdf	9	22	9	16	11	18	9	2	4	0
mdi3	9	16	29	4	7	22	13	0	0	0
mdg3	11	20	18	11	16	4	7	7	7	0
mdf3	4	4	2	11	7	2	7	20	40	2

Tabela 11. Classificação de treinamento (%) das selecionadas.

Class.	**1os**	**2os**	**3os**	**4os**	**5os**
fon	17,8	2,2	17,8	6,7	55,6
mdi	17,8	24,4	24,4	24,4	8,9
mdg	13,3	42,2	24,4	15,6	4,4
mdf	22,2	13,4	17,8	28,9	17,8
mdg3	28,9	17,8	15,6	24,4	13,3

Variantes do método proposto que apresentaram desempenho inferior ao atingido por *mdg* podem-se revelar mais úteis em versões independente de locutor do sistema caso possibilitem um maior grau de adaptação dos modelos às diferentes pronúncias.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Deller Jr., J. R. et alii; *Discrete-Time Processing of Speech Signals*, Macmillian Publishing Company, Nova Iorque, 1993.

[2] Rabiner L. R. e Juang B. H.; *Fundamentals of Speech Recognition*. Prentice Hall, USA, 1993.

[3] Rabiner, L. R., Junag B. H., Levinson, S. E. e Sondhi M.M.; *Recognition of Isolated Digitis Using Hidden Markov Models with Continuos Mixture Densities*, AT&T Technical Journal, Vol 64, No. 6, July-August 1985.

[4] Rabiner L. R., Wilpon, J. G., Soong F. K.; *High Performace Connected Digit Recognition Using Hidden Markov Models*, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, No. 8, August 1989.

[5] Santos S. C. B.; Reconhecimento de Voz Contínua para o Português Utilizando Modelos de Markov Escondidos, Tese de Doutorado, Departamento de Engenharia Elétrica, Pontifícia Universidade do Rio de Janeiro, 1997.

[6] Santos S. C. B. e Alcaim A.; Fundamentos de Reconhecimento de Voz, Centro de Estudos em Telecomunicações da Pontifícia Universidade Católica do Rio de Janeiro, CETUC-DID-01/95, setembro de 1995.

[7] Paranaguá E. D. S.; Reconhecimento de Locutores Utilizando Modelos de Markov Escondidos Contínuos. Dissertação de Mestrado, IME, 1997.

[8] Andrade M. A. R.; Reconhecimento de Comandos Conectados a Voz. Dissertação de Mestrado, IME, 1999.

[9] Hermansky H.; *Perceptual predictive (PLP) analysis of speech*, J. Acoust. Soc. Am. 87 (4), April 1990.

[10] Oppenheim A. V. & Schafer R. W.; *Digital Signal Processing*, Prentice-Hall, 1975.

[11] Andrade M. A. R. e Santos S. C. B.; Determinação de Pontos Terminais (*End-Points*) Baseada no Banco de Filtros de Coeficientes *PLP*. XVIII Simpósio Brasileiro de Telecomunicações, setembro de 2000.

[12] Schwartz R., Chow Y., Kimball O., Roucos S., Krasner M. e Marhoul J.; *Context-dependent Modeling for Acoustic-phonetic Recognition of Continuos Speech*, IEEE, 1985.

[13] Brown M.K., McGee M. A., Rabiner L.R., Wilpon J.G.; *Training Set Design for Connected Speech Recognition*, , IEEE Transactions on Signal Processing, vol. 39, No. 6, August 1991.

[14] Silveira R.C.P.; Estudo de Fonologia Portuguesa, Cortez Editora, São Paulo, 1986.

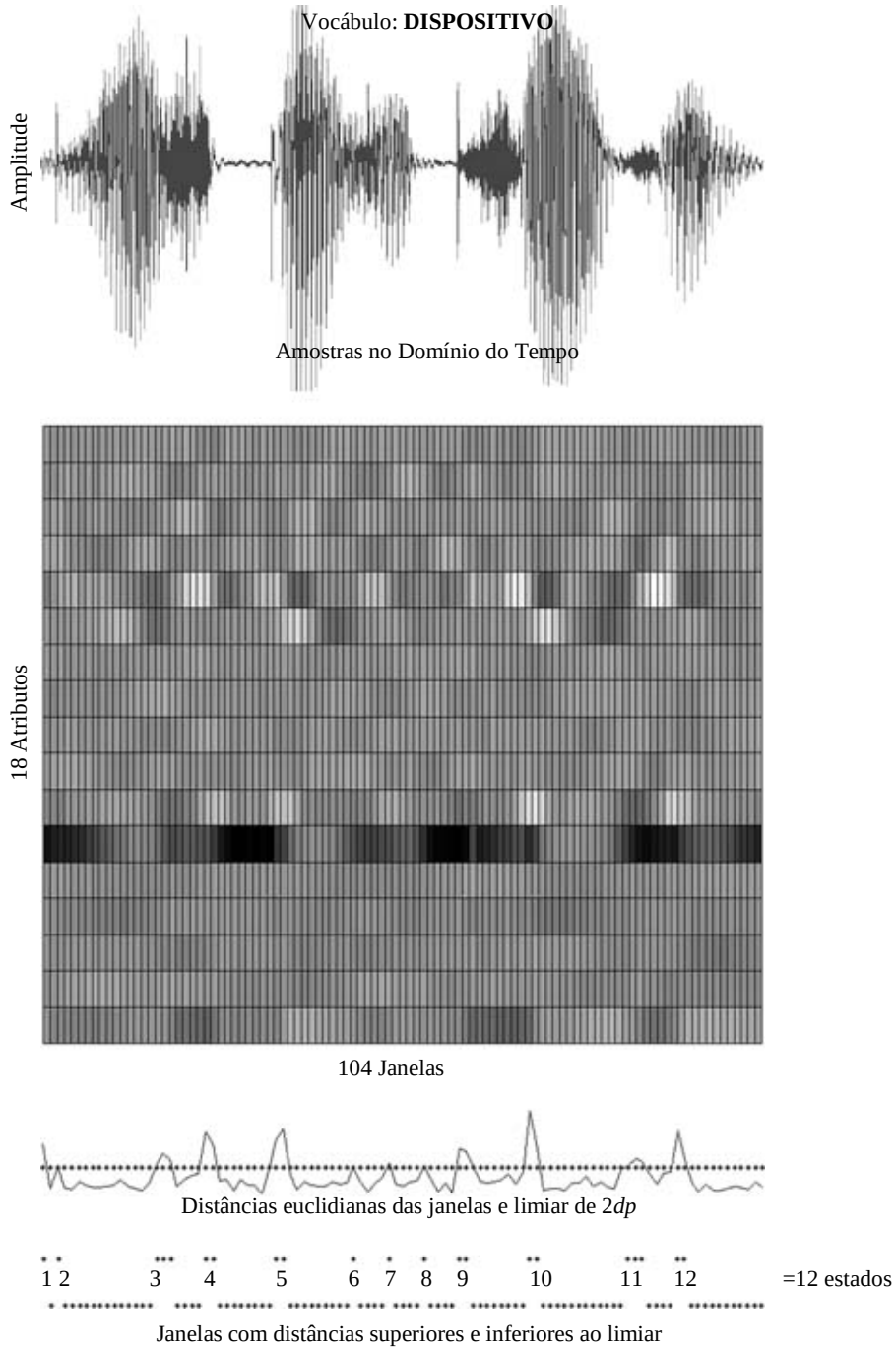

Figura 1. Método aplicado a uma elocução