Aplicação de Múltiplos HMMs com Refinamento para a Segmentação Automática da Fala

Evandro David S. Paranaguá e Sergio L. Netto

***Resumo*— Este trabalho descreve um sistema de segmentação automática de sinais de fala baseado no uso de múltiplos modelos ocultos de Markov. É investigado o uso de um processo de refinamento, que elimina a tendência do estimador detectada na etapa de treinamento. Duas estimativas para a tendência são testadas, baseadas na média e nas medianas das estimativas de treinamento. Resultados indicam que o refinamento é capaz de reduzir o erro absoluto médio de segmentação em até 38% para uma base de sinais contendo locuções de dígitos concatenados.**

***Palavras-Chave*— Segmentação de fala, HMM, YOHO.**

***Abstract*— This work presents a system for automatic segmentation of speech signals based on multiple hidden Markov models (MHMM). We investigate the effect of a post-processing stage that removes an estimating bias detected in the HMM training stage. Two bias estimates, based on the mean and median metrics, are considered. Results indicate that the post-processing stage may reduce the mean absolute error in about 38% for a signal database containing concatenated digit utterances.**

***Keywords*— Segmentação de fala, HMM, YOHO.**

I. INTRODUÇÃO

A técnica de segmentação é empregada para identificar as fronteiras entre eventos acústicos justapostos, permitindo sua separação, resultando em unidades pré-definidas como palavras, fones, difones, trifones, entre outras. Estas unidades segmentadas são, então, agrupadas, compondo um banco de unidades, que serve de base para os sistemas de síntese concatenativa de sinais de fala. Um sinal de fala sintetizado deve possuir uma naturalidade sonora para não provocar uma sensação de desconforto aos ouvintes, permitindo a contínua comunicação homem-máquina. Neste contexto, a qualidade do sinal resultante é altamente dependente do banco de unidades disponível.

O processo de segmentação pode ser manual, quando realizada por um foneticista, através da análise das características temporais e espectrais do sinal, ou automática, implementando modelos matemáticos que representem adequadamente o sinal da fala. Classifica-se, ainda, o processo nas formas explícita ou implícita [1], de acordo com o conhecimento prévio ou não, respectivamente, do conteúdo fonético do sinal sendo segmentado. Assim, na segmentação explícita, o sistema possui a priori a sequência de fonemas a ser segmentada, evitando a exclusão ou acréscimo de fronteiras.

Na literatura, vários modelos têm sido propostos para a segmentação automática da fala, destacando-se, dentre eles, a técnica de modelo oculto de Markov *(hidden Markov model, HMM)* [2], [3], também amplamente utilizada em reconhecimento da fala [4]. Alguns trabalhos de segmentação [3], [5] realizam a segmentação em duas etapas: na primeira etapa, busca-se uma aproximação da fronteira real entre duas unidades acústicas e a segunda etapa realiza um refinamento para essa fronteira.

Este trabalho apresenta um sistema de segmentação explícita automática em duas etapas seguindo a linha de [3]. Na primeira etapa, é utilizado o algoritmo de Viterbi para obter uma região próxima da fronteira para múltiplos HMMs (MHMM). Numa segunda etapa, de refinamento, a tendência de cada HMM (previamente determinada numa etapa de treinamento) é removida. Por fim, um processo de seleção (por média ou mediana) determina a estimativa da rede como um todo. Os processos de refinamento e de seleção são validados usando-se a base YOHO contendo diversas locuções em inglês de dígitos concatenados.

A estruturação deste trabalho se dá em sete seções: Na Seção II, é descrito o processo geral de segmentação baseado na técnica HMM; A Seção III apresenta a base de dados YOHO; já na Seção IV, são apresentados os critérios para implementação do HMM para a segmentação dos dígitos da base YOHO, incluindo alguns resultados de treinamento do modelo; a Seção V descreve, então, as configurações selecionadas para os MHMMs dos diversos dígitos segmentados, enquanto que a Seção VI inclui os resultados de segmentação obtidos com as técnicas propostas; por fim, a Seção VII conclui o artigo, ressaltando as principais contribuições deste trabalho.

II. SEGMENTAÇÃO UTILIZANDO HMM

Os fenômenos acústicos podem ser representados por sequências de observações extraídas de segmentos de tempo curto do sinal da fala. Desta forma, na modelagem HMM, é percebido que o encadeamento destas observações é modelado por uma sequência de estados (resultados de eventos probabilísticos) que emulam as características estatísticas do sinal de fala ao longo do tempo.

Evandro David S. Paranaguá[1,2] e Sergio L. Netto[2], 1: Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2: Programa de Engenharia Elétrica/COPPE Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil. E-mails: eparanagua@cefet-rj.br, sergioln@lps.ufrj.br.

O processo de modelagem da fala por HMM pode ser dividida em duas etapas: treinamento e reconhecimento. No treinamento, é estimado um conjunto de parâmetros em função da modelagem dos fenômenos acústicos sobre uma sequência de estados. No reconhecimento, identifica-se o modelo, dentre um conjunto pré-estabelecido, que melhor representa uma dada locução.

O conjunto de parâmetros HMM é formado pela matriz de probabilidades de transição entre estados e pelas probabilidades das observações pertencerem a cada estado. Nesta modelagem, é aplicado o alinhamento de Viterbi na estimação e o algoritmo de Baum-Welch na reestimação dos parâmetros. Já na etapa de reconhecimento, somente é aplicado o alinhamento de Viterbi sobre sequências-testes ou conjunto de observações. Nesse algoritmo é realizado o alinhamento temporal do modelo treinado com a sequência-teste, permitindo-se obter as delimitações do evento acústico em questão, que é o objetivo central do processo de segmentação automática.

Para a implementação da técnica HMM neste trabalho, utilizou-se o pacote HTK *(hidden Markov model toolkit)*, que consiste em um conjunto de módulos e ferramentas em C disponibilizados livremente para pesquisa (http://htk.eng.cam.ac.uk) e desenvolvido no Departamento de Engenharia Elétrica da Universidade de Cambridge [6].

III. Base de Dados YOHO

Neste trabalho, a técnica HMM implementada com o pacote HTK foi utilizada na obtenção das fronteiras entre números isolados a partir de sinais de fala da base YOHO contendo sequências de números concatenados. A base YOHO foi criada (pela *ITT Defense Communications Division*) para testar um sistema protótipo de verificação do locutor. São gravações em inglês de sequências de números de dois dígitos, como, por exemplo, 21-35-63 *("twenty one, thirty five, sixty three")*, não existindo regras para a pausa entre eles, a uma taxa de amostragem de 8 kHz e 12 bits por amostras. Os arquivos são gravados em formato WAV, com o nome indicando o conteúdo do mesmo: no exemplo acima, temos o arquivo 21-35-63.wav. A base de dados inclui 33 repetições para cada número de um mesmo locutor.

Para a base utilizada, foram consideradas as unidades fonéticas correspondentes aos seguintes números: 1 *(one)*, 2 *(two)*, 3 *(three)*, 4 *(four)*, 5 *(five)*, 6 *(six)*, 7 *(seven)*, 9 *(nine)*, 20 *(twenty)*, 30 *(thirty)*, 40 *(fourty)*, 50 *(fifty)*, 60 *(sixty)*, 70 *(seventy)*, 80 *(eighty)* e 90 *(ninety)*. Esta lista exclui as unidades 8 *(eight)* e 10 *(ten)*, ausentes da base YOHO.

A base YOHO foi dividida em duas: treinamento e teste. Na parte de treinamento, foram utilizadas 96 locuções, contendo no total 33 repetições para cada unidade em média. Já a parte de testes incluiu 40 locuções, com cerca de 15 repetições para cada unidade.

O desempenho de cada configuração HMM foi avaliado com base na medida do erro absoluto médio *(mean absolute error, MAE)*, definida por

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad (1)$$

onde f_i é a predição, y_i é a marca verdadeira e n é o número de estimativas. Com esta medida, desconsidera-se a direção (se antecipando ou atrasando a colocação da fronteira do segmento) do erro em relação às marcas manualmente colocadas. De modo geral, percebe-se o bom desempenho do sistema, observando-se os baixos valores atingidos para a MAE (na ordem de poucos milissegundos) para grande parte das unidades consideradas. Valores abaixo de 20 ms são considerados adequados para a aplicação de TTS [7].

Para treinamento, a base foi segmentada, manualmente pelo próprio autor, respeitando as variações temporais e espectrais do sinal da fala. As análises temporal e espectral foram realizadas utilizando as ferramentas Sony Sound Forge©versão 7.0 e o sistema Audacity©1.3.6.

IV. Implementação do HMM

Na extração das características para as redes HMMs, os sinais são segmentados por uma janela de Hamming com duração de 10 ms e taxa de superposição de 50%. Para cada janela, o vetor de observação era composto de K coeficientes mel-cepestrais (MFCC), K coeficientes delta mel-cepestrais (D) e K coeficientes delta-delta mel-cepestrais (DD) [5]. O HMM possui a topologia Modelo de Bakis [5] com avanço de $\Delta = 2$ estados. Testes iniciais indicaram um melhor desempenho da topologia com K = 12 na composição do vetores de características usando 1 estado por fonema para cada unidade acústica sendo segmentada, como visto na Tabela I.

Para os demais parâmetros da rede, foram considerados misturas com $M = 2, 3, ..., 10$ gaussianas para cada estado e $N_d = 1, 2, 3$, no cálculo dos parâmetros delta de acordo com [6]:

$$D_t = \frac{\sum_{\theta=1}^{Nd}\theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{Nd}\theta^2} \quad (2)$$

onde D_t são os coeficientes delta no instante de tempo t, aplicados sobre os coeficientes mel-cepestrais $c_{t+\Theta}$ e $c_{t-\Theta}$. Os coeficientes DD_t - da derivada de 2ª ordem - são calculados reaplicando-se a expressão acima nos coeficientes obtidos da derivada de 1ª ordem.

V. Múltiplos HMMs

O sistema proposto é baseado nos testes realizados por [3] utilizando-se um grupo de HMM com diferentes

TABELA I

NÚMERO DE ESTADOS N DO HMM.

Digíto	Representação Fonética	N
one	$/w\hat{a}n/$	3
two	$/tu'/$	2
three	$/\vartheta r\ddot{\imath}'/$	3
four	$/f\acute{o}'e/$	3
five	$/fayv'/$	4
six	$/siks'/$	4
seven	$/se'ven/$	5
nine	$/nayn'/$	4
twenty	$/twen'ti/$	6
thirty	$/\vartheta\mathbf{E}'ti/$	4
forty	$/f\acute{o}'ti/$	4
fifty	$/fif'ti/$	5
sixty	$/siks'ti/$	6
seventy	$/se'venti/$	7
eighty	$/eyt'ti/$	5
ninety	$/nayn'ti/$	6

configurações para representar cada unidade, e treinado a partir da segmentação de um banco de treinamento (BDTR). Com as variações de $M = 2, 3, ..., 10$ e $N_d = 1, 2, 3$, têm-se 27 HMMs distintos que compõem a rede MHMM global para cada unidade acústica. Em geral, estes diferentes modelos produzem resultados de segmentação com valores próximos. Uma técnica de pós-processamento é então utilizada para refinar os limites de segmentação determinados pelos diferentes modelos, com o objetivo de encontrar uma medida mais próxima da referência previamente obtida, conforme ilustrado na Figura 1.

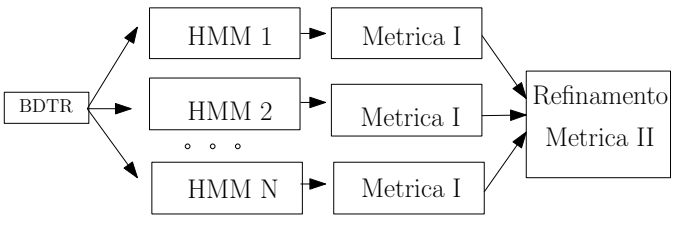

Fig. 1. Estrutura desenvolvida para MHMM com etapa de refinamento.

Na Figura 2 é mostrado um histograma-exemplo com as saídas do sistema para o recorte à esquerda da unidade one, cuja marca de referência, neste caso, se situa em 686 ms. Desta figura, vê-se que grande parte dos modelos gerou uma segmentação adequada, com erro absoluto menor ou igual a 20 ms. Uma pequena parte, porém, gerou resultados extremos

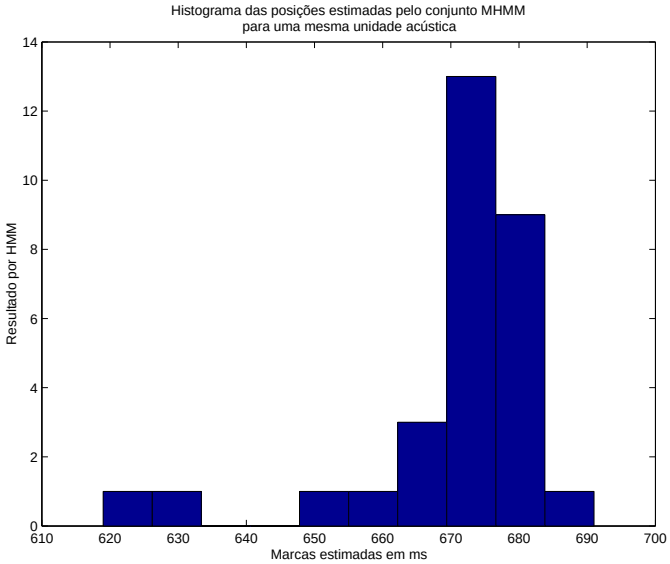

Fig. 2. Variação das posições estimadas para a unidade *one* sem o pós-processamento.

(*outliers*) que devem ser desconsiderados pela etapa de seleção (por média ou mediana).

O refinamento do sistema é implementado através do cálculo de tendência (média ou mediana) do estimador durante a etapa de treinamento. Esta tendência é então subtraída dos HMMs individuais no estágio de refinamento do processo de segmentação. E a marca estimada resultante fica sendo a definida pelo cálculo (média ou mediana) entre todos os HMMs do MHMM.

VI. RESULTADOS FINAIS DA SEGMENTAÇÃO

Neste trabalho, foram consideradas duas técnicas de seleção da estimativa final: por média ou por mediana das 27 estimativas individuais dos HMMs. Assim, há duas configurações distintas do sistema sem refinamento. Os resultados de nossos testes para este tipo de sistema são vistos nas colunas 2 e 3 das Tabelas II e III para as marcações à esquerda e à direita, respectivamente, das unidades selecionadas da base YOHO. Resultados para ambas as marcações indicam um melhor desempenho da mediana na escolha da segmentação final dentre as 27 candidatas.

Na presença do refinamento, a tendência de cada estimador foi determinada pela média ou mediana das estimativas de cada um na etapa de treinamento. Neste caso, então, há quatro configurações do sistema que inclui refinamento, considerando as duas possibilidades distintas no processo de seleção da estimativa final. Os resultados finais para as duas melhores combinações refinamento/seleção (mediana/mediana e mediana/média) são incluídos nas colunas 4 e 5 das Tabelas II e III. As outras duas combinações de refinamento/seleção também foram testadas, mas apresentaram resultados inferiores. Os resultados gerais indicam uma melhora do processo

de segmentação quando do uso do processo de refinamento, em particular para o caso do recorte à esquerda, como visto na Tabela II. Para o recorte à direita, constata-se, a partir da Tabela III, uma melhora, mas não tão significativa, provavelmente devido ao fato da boa segmentação inicialmente obtida.

TABELA II
RESULTADOS EM MILISSEGUNDOS DOS ERROS DE SEGMENTAÇÃO MAE À ESQUERDA DAS UNIDADES ACÚSTICAS.

Unidades	Sem Refinamento		Com Refinamento	
	Média	Mediana	Mediana/Mediana	Mediana/Média
one	8,21	7,30	5,45	6,26
two	45,34	44,13	27,04	26,69
three	25,70	25,56	14,13	12,99
four	29,81	30,89	23,69	22,15
five	23,49	23,35	10,50	10,27
six	10,80	9,96	10,29	10,54
seven	8,27	7,43	4,04	4,31
nine	25,31	24,85	13,61	13,63
twenty	22,64	23,07	21,16	21,11
thirty	9,28	9,45	8,06	7,93
forty	16,20	16,31	15,37	15,75
fifty	19,40	19,27	13,62	13,51
sixty	17,03	15,28	8,44	8,56
seventy	42,99	40,43	28,25	28,82
eighty	12,22	10,53	3,43	3,57
ninety	28,68	9,22	6,21	25,00
TOTAL	**21,59**	**19,81**	**13,33**	**14,44**

De modo geral, os resultados de MAE aqui alcançados, apesar de adequados, apresentam-se ligeiramente superiores aos de [3]. A comparação direta entre ambos trabalhos, porém, é dificultada por se tratar de bases de dados e unidades fonéticas distintas.

VII. CONCLUSÃO

Neste trabalho, foi discutida a segmentação de uma base de dados de números concatenados como estudo inicial sobre o uso de múltiplos modelos ocultos de Markov na automatização do processo. A técnica apresentada é composta de duas etapas, onde uma estimativa inicial das fronteiras desejadas é refinada pela remoção da tendência observada na etapa de treinamento. Um processo de seleção determina a estimativa individual desejada dentre as obtidas pelos múltiplos HMM que compõem o sistema. Resultados indicam que o erro absoluto médio pode ser reduzido pela etapa de refinamento pela mediana utilizando uma seleção também por mediana.

TABELA III
RESULTADOS EM MILISSEGUNDOS DOS ERROS DE SEGMENTAÇÃO MAE À DIREITA DAS UNIDADES ACÚSTICAS.

Unidades	Sem Refinamento		Com Refinamento	
	Média	Mediana	Mediana/Mediana	Mediana/Média
one	10,26	9,67	9,61	9,39
two	9,31	9,20	9,50	9,35
three	37,09	35,78	34,78	34,73
four	4,37	3,81	3,65	3,71
five	16,87	16,46	17,13	17,10
six	26,06	28,30	28,58	25,00
seven	6,69	6,27	6,30	5,78
nine	22,79	20,74	20,13	21,64
twenty	8,40	8,82	7,54	7,17
thirty	26,88	25,91	25,61	24,75
forty	7,28	7,06	6,60	6,39
fifty	23,70	20,77	20,19	21,38
sixty	34,76	29,28	27,59	31,89
seventy	10,35	9,71	9,16	8,97
eighty	15,07	14,91	15,33	14,01
ninety	6,00	5,20	5,36	5,30
TOTAL	**16,62**	**15,74**	**15,44**	**15,41**

REFERÊNCIAS

[1] J. P. van Hemert, "Automatic segmentation of speech,"*IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 1008–1012, Abr. 1991

[2] D. O´Shaughnessy, "Modern methods of speech synthesis,"*IEEE Circuits and Systems Magazine*, vol. 7, no. 3, pp. 06–23, 2007.

[3] S. S. Park e N. S. Kim, ´´On Using Multiple Models for Automatic Speech Segmentation,"*IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, Novembro de 2007.

[4] L. Rabiner e B. H. Juang, *Fundamentals of Speech Recognition*, New Jersey, Prentice Hall, 1993.

[5] A. M. Selmini e F. Violaro, "Segmentação Automática de Fala para o Português Brasileiro," *Anais do Simpósio Brasileiro de Telecomunicações*, Rio de Janeiro, 2008.

[6] Cambridge University Engineering Department. *The HTKBook*, 2002.

[7] D. L. Toledano, L. A. H. Gómez e L. V. Grande, "Trying to mimic human segmentation os speech using HMM and fuzzy logic post-correction rules," *Proc. 3rd International Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 207–212, 1998.

[8] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.