Segmentação Automática de Fala para o Português Brasileiro

Antonio Marcos Selmini e Fábio Violaro

***Resumo* – Este trabalho descreve o projeto e avaliação de um sistema para segmentação automática de fala realizada pelo algoritmo de Viterbi e seguida por um processo de refinamento utilizando regras fonético-acústicas. As regras empregadas para refinar cada fronteira da locução são dependentes da identidade dos fones do lado esquerdo e direito da fronteira em análise. O sistema proposto foi avaliado usando duas bases de fala dependentes de locutor do Português do Brasil (uma com locutor masculino e a outra com locutor feminino) e uma base independente de locutor (TIMIT). Um ganho de aproximadamente 29% na percentagem de erros de segmentação abaixo de 20 ms foi obtido após o refinamento da base dependente de locutor com locutor masculino.**

***Palavras-Chave* – Segmentação automática de fala, Refinamento da segmentação automática de fala, Características fonético-acústicas.**

***Abstract* – This paper describes the design and evaluation of a system for automatic speech segmentation using Viterbi's algorithm and a refinement process using acoustic-phonetic rules. Acoustic-phonetic rules used to refine each boundary of the utterance are dependent of the identity of the phones on the left and right side of the boundary under analysis. The proposed system was evaluated using two Brazilian Portuguese speaker dependent databases (one with a male speaker and another with a female speaker) and a speaker independent database (TIMIT). After the refinement process, an improvement of 29% in the percentage of segmentation errors below 20 ms was achieved in the male speaker Brazilian Portuguese database.**

***Keywords* – Automatic speech segmentation, Refining automatic speech segmentation, Acoustic-phonetic features.**

I. INTRODUÇÃO

A segmentação automática de fala tornou-se um processo importante e indispensável em diversos sistemas que usam a fala para uma interface homem-máquina. Dentre esses sistemas destacam-se o reconhecimento automático de fala, a conversão texto-fala e sistemas de animação facial sincronizada com a fala.

Independente da aplicação, uma base de fala segmentada com qualidade é altamente desejada. A segmentação manual é uma tarefa extremamente cansativa, que consome muito tempo, além de ser subjetiva.

Em virtude dos problemas decorrentes da segmentação manual e da crescente necessidade de grandes bases de fala segmentada, este trabalho descreve o projeto de um sistema de segmentação automática de fala baseado no algoritmo de Viterbi. O algoritmo de Viterbi é empregado para determinar as fronteiras entre os fones que compõem a locução. Em seguida, um ajuste fino é aplicado a cada fronteira previamente determinada, usando regras fonético-acústicas dependentes das transições entre as classes fonéticas consideradas.

O sistema desenvolvido foi avaliado em três bases de fala diferentes: duas bases dependentes de locutor em Português do Brasil (PB), uma para locutor masculino (BDM) e outra para locutor feminino (BDF), e uma base independente de locutor em Inglês Americano (TIMIT).

Este artigo está dividido em sete seções. Na seção II são analisados os fundamentos da segmentação automática de fala e, na seção III, os fundamentos sobre refinamento da segmentação automática. O sistema proposto é descrito na seção IV, seguido pela caracterização das classes fonéticas na seção V. Os resultados são apresentados na seção VI e as conclusões na seção VII.

II. FUNDAMENTOS DE SEGMENTAÇÃO AUTOMÁTICA DE FALA

Diversas técnicas foram propostas na literatura para resolver o problema da segmentação automática de fala. Dentre as principais destacam-se o uso de HMMs (*Hidden Markov Models*), redes neurais artificiais, regras baseadas em lógica fuzzy, função de variação espectral do sinal de fala, dentre outras.

As técnicas de segmentação são divididas em dois grupos: segmentação implícita ou lingüisticamente irrestrita e segmentação explícita ou lingüisticamente restrita. O critério adotado para a classificação depende do tipo de informação que as técnicas utilizam para segmentar a locução [1] [2] [3].

Na segmentação implícita, toda a informação necessária para a segmentação é extraída a partir da locução, ou seja, a transcrição fonética da locução não está disponível. A vantagem desta técnica é que não é necessário gerar a transcrição fonética das locuções para a segmentação. Por outro lado, a grande desvantagem é que podem ocorrer inserções e até mesmo deleções de fronteiras.

Para a segmentação explícita, as fronteiras são determinadas de acordo com o número de símbolos presentes na transcrição fonética das locuções que serão segmentadas. A desvantagem é que a transcrição fonética das locuções deve ser gerada antes da segmentação. A vantagem é que não ocorrem inserções e ou deleções, mas as fronteiras determinadas podem estar um pouco distante das fronteiras de referência (segmentação manual), o que pode ser corrigido através das técnicas de refinamento.

Entre as diversas técnicas que podem ser empregadas para a segmentação, o uso de HMMs juntamente com o algoritmo de Viterbi é o mais difundido.

Antonio Marcos Selmini e Fábio Violaro, Departamento de Comunicações, Faculdade de Engenharia Elétrica e Computação, Universidade Estadual de Campinas, Campinas, Brasil, Emails: {selmini, fabio}@decom.fee.unicamp.br

A. Algoritmo de Viterbi para Segmentação Automática de Fala

Os HMMs são largamente empregados em sistema de reconhecimento automático de fala para representar as subunidades fonéticas devido à sua capacidade de modelar a dinâmica das variações temporais no sinal de fala [4].

Em segmentação de fala os HMMs também têm forte aplicação motivada pelos excelentes resultados em reconhecimento automático de fala.

O alinhamento de Viterbi é utilizado para determinar a seqüência ótima de estados a partir dos parâmetros do sinal de fala associados a uma locução de entrada. Estes parâmetros de entrada são alinhados com os modelos dos fones, gerando dessa forma uma estimativa das fronteiras.

Como a finalidade do algoritmo de Viterbi não é atuar como um segmentador de fala, os resultados obtidos devem ser refinados de forma a aproximá-los da segmentação manual produzida por um especialista.

III. Refinamento da Segmentação Automática de Fala

O refinamento da segmentação automática de fala consiste em realizar um processamento automático na locução previamente segmentada, com o objetivo de aproximar as fronteiras das que seriam obtidas através de uma segmentação manual.

Diversas técnicas têm sido reportadas para o refinamento. Dentre elas destacam-se: variação espectral de energia [5], regras fuzzy, uso de Máquinas de Vetor de Suporte (SVM – *Support Vector Machine*) [6], dentre outras.

No sistema proposto, o processo de refinamento foi realizado usando características fonético-acústicas específicas de cada classe de fones. Inicialmente todos os fones utilizados na transcrição fonética foram agrupados em classes. Em seguida, uma exaustiva pesquisa foi realizada de forma a descobrir os principais parâmetros que pudessem ser utilizados para cada classe durante o processo de refinamento.

IV. Sistema Baseado em Regras para o Refinamento da Segmentação Automática de Fala

O sistema proposto tem inspiração no trabalho de Amit Juneja [7], em que características acústicas dos fones foram utilizadas para realizar classificação fonética.

Nenhum trabalho explorando as características acústicas de cada classe fonética para o refinamento das marcas de segmentação foi encontrado. A maioria dos trabalhos utiliza técnicas que normalmente apresentam uma complexidade computacional alta ou necessitam de grande quantidade de material de treinamento.

O sistema proposto é dividido em duas partes [8] [9]. A primeira parte é composta por dois módulos: módulo de treinamento dos HMMs associados às unidades acústicas e módulo de segmentação das locuções. A segunda parte, por sua vez, é composta por um único módulo que é responsável por refinar cada fronteira previamente determinada pelo módulo de segmentação. O módulo de treinamento executa o algoritmo de Baum-Welch e o módulo de segmentação o alinhamento forçado de Viterbi.

A. Módulo de Treinamento

A Figura 1 descreve os módulos de treinamento e segmentação.

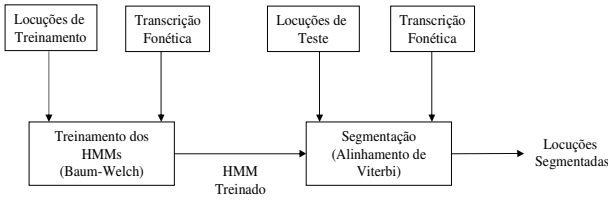

Fig. 1. Módulos de treinamento e segmentação.

Para o treinamento dos HMMs associados às unidades acústicas são necessárias as locuções de treinamento e as respectivas transcrições fonéticas. A partir da transcrição fonética são gerados os modelos acústicos da locução completa através da concatenação dos modelos de cada fone constituinte. Cada unidade fonética foi representada por um HMM contínuo de três estados com topologia *left-right*.

A função densidade de probabilidade para emissão dos símbolos foi modelada através de uma mistura de apenas duas Gaussianas. Este valor foi determinado através de testes, em que o número de Gaussianas foi variado de 1 até 20. O sistema utiliza uma matriz de covariância diagonal com componente independentes.

Antes do treinamento, cada locução passa por uma fase de pré-processamento. Primeiro o nível DC do sinal é removido e em seguida o sinal passa por um filtro passa-altas de pré-ênfase ($1\text{-}az^{-1}$). Nas simulações usando a base de fala dependente de locutor foi utilizado o coeficiente a = 0,95 e, nas simulações usando a TIMIT, foi utilizado a = 0,97 (conforme indicado na literatura).

As locuções são janeladas através de janela de Hamming com duração de 20 ms e, a cada 10 ms, um novo conjunto de parâmetros é calculado. Os parâmetros empregados no treinamento são: 12 coeficientes mel-cepstrais, 1 parâmetro log-energia normalizado e suas derivadas de primeira e segunda ordem (utilizando uma janela à direita e uma janela à esquerda). Os parâmetros são agrupados em um vetor de dimensão 39.

Para o cálculo dos parâmetros delta foi utilizada a expressão definida no HTKBook [10]:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \qquad (1)$$

onde Θ é o número total de janelas adjacentes (Θ janelas à esquerda e Θ janelas à direita) e c_t são os coeficientes no instante de tempo t em que se deseja calcular o parâmetro delta.

B. Módulo de Segmentação

A segmentação das locuções é realizada pelo algoritmo de Viterbi, que utiliza os HMMs treinados no módulo de treinamento.

Para segmentar uma determinada locução, inicialmente é gerado o modelo da locução com base na transcrição

fonética da mesma. Em seguida, é realizada a fase de pré-processamento como descrito no módulo de treinamento e, a cada instante de tempo *t*, os parâmetros acústicos de cada janela da locução são apresentados ao algoritmo de Viterbi, que calcula a probabilidade do modelo emitir os símbolos acústicos.

A seqüência de parâmetros que compõem a locução é alinhada com o HMM correspondente. O algoritmo de Viterbi escolhe o melhor caminho (entre todas as possibilidades) que maximiza a verossimilhança do modelo emitir os símbolos de entrada. A partir do caminho ótimo de Viterbi, o número de janelas associadas a cada fone é determinado. A partir do número de janelas é possível calcular o número de amostras associadas a cada fone e, conseqüentemente, estimar as fronteiras entre os fones adjacentes.

C. Módulo de Refinamento

A Figura 2 exemplifica o módulo de refinamento.

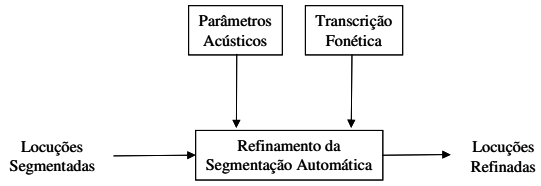

Fig. 2. Módulo de refinamento.

O módulo de refinamento utiliza três informações: a locução segmentada (resultado do módulo de segmentação), os parâmetros acústicos das classes fonéticas e a transcrição fonética já disponível.

Como o refinamento é baseado nas características acústicas dos fones, primeiro os 38 fones utilizados para a base dependente de locutor foram agrupados em 15 classes fonéticas, e os 48 fones usados para a base independente de locutor foram agrupados em 13 classes.

Para a base dependente de locutor as classes fonéticas definidas foram: silêncio, fricativas surdas e sonoras, plosivas surdas e sonoras, vogal anterior, vogal média, vogal posterior, vogal nasal, consoantes nasais, laterais, vibrantes, africadas, *stop* e *voiced closures*. Para a base independente de locutor só não foram definidas classes para as vogais nasais, laterais e vibrantes. As laterais e vibrantes foram agrupadas em uma única classe que foi denominada semivogal. Uma vogal foi considerada nasal na TIMIT quando seguida por uma consoante nasal [11].

O processo de refinamento proposto é baseado em um conjunto de regras. Cada regra por sua vez é formada por alguns parâmetros, que são os mais representativos de cada classe fonética. Alguns parâmetros utilizam um limiar para determinar a nova posição da marca de segmentação e outros são baseados na detecção de picos (*pick-peaking*). O número de parâmetros para cada classe é variável.

Os limiares de cada parâmetro foram determinados a partir de uma base de fala dependente de locutor do PB segmentada manualmente. Este processo foi realizado em duas etapas. Primeiro todos os valores de cada parâmetro de cada fone foram calculados. Em seguida os valores foram distribuídos em histogramas e uma análise detalhada foi realizada com o objetivo de determinar os melhores valores para os limiares.

O processo de refinamento leva em consideração a classe fonética do lado direito e do lado esquerdo da fronteira em análise, ou seja, as regras são dependentes do tipo de transição. Um levantamento de todas as possíveis transições foi realizado com o objetivo de determinar qual ou quais os parâmetros que melhor se aplicam a cada tipo de transição e dessa forma construir as regras. Essa estratégia foi adotada com o objetivo de tentar reduzir o número de possíveis combinações de parâmetros nas regras e também poder trabalhar com os melhores parâmetros para cada tipo de transição.

Durante o processo de refinamento, cada marca de segmentação é analisada separadamente e, através da transcrição fonética, sabe-se quais são os fones que cada marca separa. Com base nos fones determina-se o tipo de transição e, conseqüentemente, quais parâmetros acústicos devem ser empregados. Por exemplo, se uma marca de segmentação separa o silêncio inicial de uma consoante fricativa, sabe-se que na transição para a consoante fricativa, o valor da energia total vai aumentar até ultrapassar um determinado limiar. O instante de tempo em que o limiar é atingido é definido como a fronteira entre os dois fones.

Para o cálculo dos parâmetros durante o processo de refinamento, é definido um intervalo de refinamento. O intervalo de refinamento tem início no ponto médio entre a fronteira imediatamente anterior à fronteira que está sendo analisada e a fronteira em análise. O final é definido na fronteira posterior à fronteira em análise. Os parâmetros são calculados apenas neste intervalo e, conseqüentemente, a fronteira em análise também só poderá ser movida neste intervalo. Os parâmetros especificados pelas regras são calculados usando janelas de análise de 20 ms, mas com deslocamento de apenas 1 ms nesta fase.

V. CARACTERIZAÇÃO ACÚSTICA DAS CLASSES FONÉTICAS

Como apresentado nas seções anteriores, o refinamento das fronteiras de segmentação é realizado com base nas características acústicas das classes fonéticas, mais especificamente com base no tipo de transição entre as classes. Os principais parâmetros acústicos para cada classe fonética são apresentados nesta seção.

Para evitar problemas com o nível de gravação, as locuções são normalizadas em relação ao valor do maior pico da amplitude.

A. Silêncio

Esta classe representa o silêncio no início e no final de cada locução, e também as possíveis pausas entre as palavras. O melhor parâmetro para caracterizar o intervalo em que ocorre o silêncio é a energia total da janela de análise.

O limiar definido para a energia foi de -60 dB, ou seja, no intervalo de refinamento o centro da janela de análise que cruzar o limiar estabelecido representa a fronteira de separação entre o silêncio e outra classe fonética.

B. Fricativas

Dentre os vários parâmetros que podem ser empregados para caracterizar as fricativas (surdas e sonoras), apenas dois foram utilizados: a taxa de cruzamentos por zero e o centro de gravidade espectral.

A taxa de cruzamentos por zero foi determinada usando o algoritmo clássico apresentado em [12], sem alterações. O centro de gravidade espectral representa a freqüência abaixo da qual 50% da energia total do sinal janelado está concentrada. Como são utilizadas duas regras, a fronteira é estabelecida no centro da janela de análise em que os dois parâmetros cruzaram os limiares estabelecidos.

As fricativas surdas apresentam um maior valor para a taxa de cruzamentos por zero em relação às fricativas sonoras. Para as fricativas surdas o limiar estabelecido foi de 0,52 e, para as fricativas sonoras, 0,28. Para o centro de gravidade espectral, o limiar foi estabelecido em 2500 Hz para as fricativas surdas e sonoras.

C. Consoantes Laterais e Vibrantes

A transição entre as consoantes laterais e as demais classes fonéticas (normalmente as vogais no PB) é marcada por uma leve variação de energia, uma vez que esses sons são muito parecidos com as vogais. Por outro lado, as consoantes vibrantes apresentam maior variação de energia.

Para o refinamento foram utilizados cinco bandas de freqüência e os limites de cada banda foram calculados de forma a conter determinadas freqüências formantes (regiões caracterizadas por alta concentração de energia).

A primeira banda corresponde à banda total de energia. A segunda banda (0 - 500 Hz) corresponde à região de ocorrência do primeiro formante, a terceira banda (500 - 1500 Hz) corresponde à região de ocorrência do segundo formante, a quarta banda (1500 – 2400 Hz) corresponde à região de ocorrência do terceiro formante e a quinta e última banda (2400–fs/2 Hz), onde fs representa a freqüência de amostragem do sinal, que para a base dependente de locutor é 22,05 kHz e para a base independente é 16 kHz. Os limites de cada banda foram determinados através de testes.

A energia espectral em cada banda de freqüência foi calculada via DFT com 1024 pontos, a partir de janelas de análise de 20 ms, deslocadas a cada 1 ms, ponderadas com a janela de Hamming. A derivada da energia espectral foi calculada para cada banda de freqüência utilizando cinco janelas adjacentes de cada lado utilizando a eq. (1).

A combinação da variação da energia espectral de todas as bandas é feita através da soma de seus valores a cada instante de tempo. Esse procedimento adotado permite realçar os picos da variação de energia e, dessa forma, facilitar a localização das fronteiras corretamente.

D. Consoantes Nasais

As consoantes nasais apresentam características semelhantes às vogais: são sonoras, possuem uma estrutura de formantes bem definida e apresentam valor de F1 baixo (o que normalmente pode ser confundido com as vogais anteriores e posteriores). Como as três consoantes nasais do PB são sempre seguidas pelas vogais, esses parâmetros não são adequados para a detecção eficiente das fronteiras.

Uma outra característica importante é a variação da energia espectral. Para as consoantes nasais existe uma concentração de energia nas baixas freqüências, uma vez que essas consoantes apresentam valor de F1 abaixo de 300 Hz.

Para a detecção das fronteiras foi utilizada a energia em duas bandas de freqüência: [0 – 358 Hz] e [358 – 5378 Hz] [7]. A primeira banda está relacionada com a concentração da energia nas baixas freqüências, característica marcante das consoantes nasais. A segunda banda está relacionada com a concentração da energia nas altas freqüências (características das vogais). A variação da energia em cada banda de freqüência é determinada e, em seguida, os valores de cada banda são somados de forma a destacar os picos de variação. A energia espectral foi calculada via DFT.

E. Plosivas

São caracterizadas por um longo período de silêncio e por uma "explosão" que corresponde à liberação do ar. O período que constitui a explosão é muito curto, o que dificulta o seu processamento e, conseqüentemente, a localização do instante em que ocorre a transição para o fone seguinte. Para facilitar o refinamento, as plosivas foram tratadas como dois fones diferentes: período de constrição (silêncio que antecede a plosiva) e a explosão.

O refinamento das plosivas é realizado em duas etapas: primeiro é determinado o instante da liberação do ar e, em seguida, é determinada a fronteira entre a plosiva e o fone adjacente seguinte. Para a determinação do início da explosão é utilizada apenas a derivada da energia, que por sua vez exibe um pico no início da explosão. A energia foi calculada a partir de janelas de análise de 10 ms, deslocadas a cada 1 ms.

Para a determinação da fronteira entre a plosiva e o fone seguinte, o intervalo de refinamento começa no início da explosão e termina no final do fone adjacente à direita. Outra alteração é com relação ao tamanho da janela de análise que passa a ser de 10 ms ao invés de 20 ms. Essa modificação fez-se necessária em virtude da curta duração das plosivas.

Dois parâmetros foram utilizados no refinamento: derivada da energia espectral na banda [0 – F3 Hz] e [F3 – fs/2 Hz], onde F3 é a terceira freqüência formante. A derivada da energia das duas bandas é somada de forma a realçar a variação da energia. A fronteira é definida no ponto máximo de variação da energia.

F. Africadas

Devido às semelhanças acústicas com as consoantes fricativas e com as plosivas, as regras de refinamento para as consoantes africadas seguem o mesmo padrão já definido para as fricativas e plosivas.

As africadas são caracterizadas por um período de baixa energia espectral nas baixas freqüências (período de constrição), seguida por uma região que apresenta uma alta taxa de cruzamentos por zero e também centro de gravidade espectral acima de 4 kHz. A combinação dessas características acústicas é suficiente para determinar com precisão a fronteira entre uma consoante africativa e a vogal /i/ (vogal que sempre segue uma consoante africativa no PB).

A transição entre qualquer classe fonética e uma consoante africativa é definida na região em que ocorre uma queda abrupta de energia. Essa queda abrupta é determinada através do pico da derivada primeira da energia.

O refinamento (transição entre a africativa e a vogal /i/) ocorre em duas etapas. Na primeira é determinado o instante em que ocorre a liberação do ar e, na segunda, a transição propriamente dita. O intervalo de refinamento para essa classe fonética segue o mesmo padrão das outras classes. A fronteira é definida no centro da janela de análise em que a

taxa de cruzamentos por zero é maior que 0,4 e o centro de gravidade espectral maior que 4 kHz.

G. Vogais e Vogais Nasais

As vogais representam o coração das sílabas no PB e, portanto, podem estar "ligadas" a todas as outras classes de fones. É muito comum no PB a ocorrência de ditongos que, juntamente com as plosivas, representam as classes mais difíceis de serem refinadas. As vogais também apresentam características acústicas que estão presentes nas outras classes fonéticas, tais como: estrutura de formantes bem definida, alta concentração de energia, etc. A transição entre uma vogal e qualquer outra classe fonética é determinada empregando-se parâmetros que melhor caracterizam o tipo de transição (o parâmetro depende da classe seguinte).

Para as vogais presentes em ditongos, três tipos de transições podem ocorrer: i) transição entre as vogais anteriores ou posteriores e a vogal média, ii) transição entre vogais anteriores e vogais posteriores e iii) transição entre as vogais da mesma classe.

Para o primeiro tipo de transição foi utilizada uma análise da primeira e da segunda freqüência formante. A determinação dos formantes em cada janela de análise foi realizada através da análise LPC com ordem 14. Os coeficientes do preditor linear foram calculados aplicando o algoritmo de Levinson-Durbin. Tendo os coeficientes do preditor, a DFT com 1024 pontos é empregada para calcular a resposta em freqüência do filtro de síntese. Os picos resultantes da análise LPC representam as freqüências formantes.

Os limiares definidos para as vogais foram: para as vogais anteriores e posteriores, o valor do primeiro formante é menor que 450 Hz e, para a vogal média, acima de 450 Hz. Para o segundo formante cada classe apresenta valores diferentes. O valor do segundo formante para as vogais anteriores está acima de 1845 Hz e, para as vogais posteriores, abaixo de 1135 Hz. Já a vogal média apresenta valores intermediários (maior que 1135 Hz e menor que 1845 Hz). A combinação desses valores foi utilizada para refinar a ocorrência da vogal média com as vogais anteriores e posteriores.

O segundo tipo de transição em um ditongo ocorre quando se tem uma vogal anterior e uma vogal posterior ou vice-versa. Neste caso, dois parâmetros foram utilizados: a variação do segundo formante e o perfil energia [8]. Para o perfil energia foi utilizada a taxa de 75%, que é suficiente para discriminar entre as duas classes de vogais. O limiar foi estabelecido em 1550 Hz.

O valor do segundo formante tende a ser menor para as vogais posteriores em relação às vogais anteriores. O limiar para o segundo formante foi estabelecido em 1490 Hz.

A grande dificuldade no refinamento dos ditongos está na ocorrência de vogais que pertencem à mesma classe fonética. A estratégia então adotada é o uso do critério de informação Bayesiana (BIC – *Bayesian Information Criterion*) [13].

O BIC é largamente utilizado em modelagem estatística, e também pode ser usado para detectar pontos de mudança acústica em um sinal de fala. Segmentos adjacentes são modelados usando diferentes distribuições de Gaussianas. A concatenação desses segmentos obedece a uma terceira distribuição.

Considere que H_0:(c_1, c_2, ..., c_n) ~ $N(\mu_0, \Sigma_0)$ seja a seqüência de vetores de características acústicas para o segmento maior e, H_1:(c_1, c_2, ..., c_m) ~ $N(\mu_1, \Sigma_1)$ e H_2:(c_{m+1}, c_{m+2}, ..., c_n) ~ $N(\mu_2, \Sigma_2)$ a seqüência de vetores de características acústicas do primeiro e do segundo segmento respectivamente. Os vetores c_i de dimensão Q podem representar os coeficientes mel-cepstrais obtidos a partir de cada segmento, Σ_0, Σ_1 e Σ_2 são as matrizes de covariância completas para cada segmento. A variação do valor do BIC entre os modelos é dada por:

$$BIC(i) = R(i) - \lambda P_0 \tag{2}$$

onde $R(i)$ a razão de verossimilhança calculada por:

$$R(i) = n\log|\Sigma_0| - m\log|\Sigma_1| - (n-m)\log|\Sigma_2| \tag{3}$$

O parâmetro P_0 é o fator de penalização para a complexidade do modelo, e seu valor é calculado usando a eq. (4):

$$P_0 = \frac{1}{2}\left(Q + \frac{Q(Q+1)}{2}\right)\log n \tag{4}$$

Na eq. (4), o parâmetro λ representa o peso para o fator de penalização. Seu valor pré-definido é 1. O ponto de mudança acústica calculado através do BIC ocorre no centro da janela de análise em que o valor de $BIC(i)$ é máximo, para todos os valores de i.

VI. RESULTADOS E DISCUSSÃO

O sistema proposto foi avaliado usando três bases de fala, das quais duas são dependentes de locutor do PB e uma é independente de locutor (TIMIT). Tanto o treinamento dos HMMs quanto o alinhamento de Viterbi foi feito no HTK. A transcrição fonética da base masculina foi feita pelo próprio autor, enquanto que da base feminina foi feita por profissionais habilitados.

Os 38 modelos dos fones dependentes de locutor foram treinados a partir de 1026 locuções gravadas por um único locutor paulista do sexo masculino. As locuções foram amostradas a 22,05 kHz e quantizadas com 16 bits/amostra. Para o teste de segmentação foram utilizadas 200 locuções diferentes das locuções da base de treinamento. Um segundo teste foi realizado com 100 locuções pronunciadas por uma locutora do sexo feminino, paulista e musicista. A base de fala dependente de locutor do sexo feminino foi gentilmente cedida pela empresa Vocalize – Soluções em Tecnologias da Fala e da Linguagem Ltda, com sede em Campinas. A segmentação inicial desta base pelo algoritmo de Viterbi foi feita a partir do HMM treinado com a voz do locutor masculino (BDM).

Para a base independente de locutor (TIMIT), foram considerados 48 fones e empregadas 3696 locuções, que por sua vez são amostradas a 16 kHz e quantizadas com 16 bits/amostra. Para a segmentação foram utilizadas 1344 locuções diferentes das locuções de treinamento.

Independente da base de treinamento, cada fone foi representado por um HMM dependente de contexto, com

três estados emissores e apenas duas Gaussianas por mistura. A modelagem com fones dependentes de contexto resultou na melhor segmentação em relação à modelagem sobre fones independentes de contexto.

Tanto a avaliação da segmentação automática quanto do refinamento foi realizada através de comparação com resultados de segmentação manual. Os resultados são expressos em porcentagens, refletindo o total de fronteiras que apresentam erro abaixo de um determinado limiar previamente definido.

A Tabela I mostra os resultados da segmentação automática fornecida pelo alinhamento forçado de Viterbi antes do processo de refinamento para as três bases.

Os resultados obtidos para a base independente de locutor (BI) são melhores do que os resultados obtidos para a base dependente de locutor masculino (BDM) e dependente de locutor feminino (BDF). A principal razão está no número de locuções disponíveis na TIMIT para treinamento e, conseqüentemente, geração dos fones dependentes de contexto. A tabela II mostra os resultados após o refinamento.

TABELA I
RESULTADOS DA SEGMENTAÇÃO AUTOMÁTICA DE FALA FORNECIDA PELO ALINHAMENTO DE VITERBI ANTES DO REFINAMENTO.

Limiar (ms)	Porcentagem de Erro		
	BDM	BDF	BI
<= 5	21,98	17,86	26,98
<= 10	40,00	32,09	51,15
<= 20	66,49	55,13	81,01
<= 30	83,24	70,59	90,87
<= 40	89,77	81,78	95,09
<= 50	93,48	89,23	97,06
<= 100	99,18	95,12	99,63

TABELA II
RESULTADOS DA SEGMENTAÇÃO AUTOMÁTICA DE FALA APÓS O REFINAMENTO.

Limiar (ms)	Porcentagem de Erro		
	BDM	BDF	BI
<= 5	61,00	35,20	57,10
<= 10	73,00	50,50	68,69
<= 20	95,55	78,02	91,97
<= 30	96,98	82,67	94,50
<= 40	98,23	92,14	96,32
<= 50	98,50	95,92	97,70
<= 100	100	98,10	100

Após o refinamento a base de fala masculina (BDM) apresentou os melhores resultados quando comparada com as outras duas bases. Como já mencionado, os limiares utilizados para as três bases foram determinados a partir da base de fala masculina, o que pode justificar essa vantagem.

A base de fala feminina não apresentou bons resultados em virtude de apresentar características acústicas e fonéticas diferentes da fala masculina, com a qual o HMM empregado na segmentação inicial foi treinado. Em [9] também foi utilizado um módulo para extração de erros sistemáticos, cujos resultados aqui não foram reportados uma vez que o refinamento já é suficiente para aproximar a segmentação automática da segmentação manual.

VII. CONCLUSÕES

Neste artigo foram descritos o projeto e avaliação de um sistema que realiza segmentação automática de fala seguida por um processo de refinamento, com foco no PB. A segmentação é realizada pelo algoritmo de Viterbi.

A estratégia adotada para o refinamento leva em consideração o tipo de transição, ou seja, os fones presentes no lado direito e esquerdo da fronteira que está sendo refinada. As características dependentes dos fones e, conseqüentemente, do tipo de transição, representam uma boa solução para o problema do refinamento da segmentação.

Como trabalhos futuros podemos citar o uso de treinamento discriminativo de forma a reduzir erros de segmentação e o emprego das regras acústicas para segmentar as locuções ao invés de apenas refinar.

REFERÊNCIAS

[1] E. Vidal e A. Marzal, A review and new approaches for automatic segmentation of speech signals, Signal Processing V: Theories and Applications, L. Torres, E. Masgrau and M. A. Lagunas (eds.), Elsevier Science Publishers B. V., pp. 43-53, 1990.

[2] T. Svendsen e F. K. Soong, On the automatic segmentation of speech signals. Proceedings on the International Conference on Acoustic, Speech and Signal Processing, pp. 77-89, 1987.

[3] J. P. van Hemert, Automatic Segmentation of Speech, IEEE Transactions on Signal Processing, vol. 39, issue 4, pp. 1008-1012. April 1991.

[4] D. Toledano, L. A. Gómez e L. V. Grande, Automatic Phonetic Segmentation, IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6. November 2003.

[5] L. Golipour e D. O'Shaughnessy, A New Approach for Phoneme Segmentation of Speech Signals. INTERSPEECH, Antwerp, Belgium, August, 2007.

[6] H. Y. Lo e H. M. Wang, Phonetic Boundary Refinement Using Support Vector Machine. Proceeding of the International Conference on Acoustic, Speech and Signal Processing, Honolulu, Havaí, USA, April, 2007.

[7] A. Juneja, Speech Recognition Based on Phonetic Features and Acoustic Landmarks. PhD Thesis, University of Maryland, College Park, USA, 2004.

[8] A. M. Selmini e F. Violaro, Improving the Explicit Automatic Speech Segmentation Provided by HMMs, IWT2007, Santa Rita do Sapucaí, MG, Brasil, Fevereiro, 2007.

[9] A. M. Selmini e F. Violaro, Acoustic-Phonetic Features for Refining the Explicit Speech Segmentation, INTERSPEECH, Antwerp, Belgium, August, 2007.

[10] Cambridge University Engineering Department. The HTK Book, 2002.

[11] T. Pruthi e C. Y. Espy-Wilson, Acoustic Parameters for the Automatic Detection of Vowel Nasalization. INTERSPEECH, Antwerp, Belgium, August, 2007.

[12] L. R. Rabiner, e R. W. Schafer, Digital Processing of Speech Signals. Prentice Hall, New Jersey, 1978.

[13] G. Almpanidis e C. Kotropoulos, Phonemic segmentation using the generalized Gamma distribution and small sample Bayesian information criterion. Speech Communication, Vol. 50, pp. 35-55, January, 2008.