SEGMENTAÇÃO AUTOMÁTICA DO SINAL DE VOZ PARA SISTEMAS DE CONVERSÃO TEXTO-FALA

Evandro David Silva Paranaguá

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Sergio Lima Netto

Rio de Janeiro
Março de 2012

SEGMENTAÇÃO AUTOMÁTICA DO SINAL DE VOZ PARA SISTEMAS DE CONVERSÃO TEXTO-FALA

Evandro David Silva Paranaguá

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Sergio Lima Netto, Ph.D.

Prof. Gelson Vieira Mendonça, Ph.D.

Prof. Amaro Azevedo de Lima, Ph.D.

Prof. Fábio Violaro, D.Sc.

Prof. Sidney Cerqueira Bispo dos Santos, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2012

Silva Paranaguá, Evandro David

Segmentação Automática do Sinal de Voz para Sistemas de Conversão Texto-Fala/Evandro David Silva Paranaguá. – Rio de Janeiro: UFRJ/COPPE, 2012.

X, 90 p.: il.; 29, 7cm.

Orientador: Sergio Lima Netto

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2012.

Referências Bibliográficas: p. 86 – 90.

1. Hidden Markov Models. 2. Segmentação. 3. Regras Fonéticas. I. Lima Netto, Sergio. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Em primeiro lugar, dedico esta tese à minha esposa, amiga e companheira inseparável em todas as minhas atividades. Aos meus pais, aos meus irmãos, ao meu Bruninho e aos meus amigos que, mesmo em tantos momentos de ausência, estiveram ao meu lado e me apoiaram.

Agradecimentos

Meus sinceros agradecimentos:

- a Deus que me norteia em todos os meus momentos;

- ao Professor Sergio Lima Netto pela paciência e incentivo nas horas difíceis e, ainda por sua orientação precisa e segura, que contribuiu e foi fundamental para que esta tese se concretizasse;

- ao Professor Fábio Violaro pelo seu auxílio em fornecer material indispensável e importante para a elaboração desta tese;

- aos colegas do CEFET pela amizade e compreensão, que na parte final da tese, permitiram que eu me dedicasse integralmente a sua elaboração;

- ao amigo Dirceu Gonzaga da Silva pelo apoio e pela troca de ideias, sempre de grande valor para mim;

- aos colegas e funcionários da UFRJ pelo carinho e pela atenção dispensados durante todo esse período de pesquisa.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

SEGMENTAÇÃO AUTOMÁTICA DO SINAL DE VOZ PARA SISTEMAS DE CONVERSÃO TEXTO-FALA

Evandro David Silva Paranaguá

Março/2012

Orientador: Sergio Lima Netto

Programa: Engenharia Elétrica

Nesta tese abordamos o problema de segmentação automática de sinais de fala com o intuito de organizar um banco de unidades fonéticas para conversores texto-fala (TTS, do inglês *text-to-speech*) concatenativos.

O sistema básico de segmentação utiliza modelos ocultos de Markov (HMM, do inglês *hidden Markov model*) que faz a análise estatística do sinal de fala, gerando um modelo para cada unidade fonética. Como subproduto desta modelagem surgem as fronteiras fonéticas que são o objetivo do sistema de segmentação.

Nesta tese são consideradas ainda duas técnicas de refinamento do processo de segmentação. A primeira considera o uso de múltiplos HMMs, que geram diferentes estimativas da fronteira fonética. Uma análise estatística posterior, por medidas simples como média ou mediana, combina estas estimativas parciais para gerar a estimativa final. O segundo método de refinamento considera uma análise determinística das transições entre os diferentes fonemas de uma língua. Neste processo, cada tipo de transição é caracterizada por algum fenômeno acústico modelado por regras fonéticas. Com este processo, uma primeira estimativa é refinada pelas regras fonéticas gerando a estimativa final.

Esta tese considera os dois tipos de refinamento e procura combiná-los de modo a gerar um sistema final que combina as qualidades de cada tipo de técnica para obter o melhor desempenho final de segmentação. O resultado é um sistema com o menor erro absoluto (MAE, do inglês *mean absolute error*) em relação a uma segmentação feita por um foneticista.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

AUTOMATIC SEGMENTATION OF SPEECH SIGNAL TO SYSTEMS FOR TEXT-SPEECH

Evandro David Silva Paranaguá

March/2012

Advisor: Sergio Lima Netto

Department: Electrical Engineering

This thesis presents a new system for automatic segmentation of speech signals for text-to-speech (TTS) synthesis.

The traditional segmentation technique is based on the hidden Markov model (HMM) which performs a statistical model of the speech process for each phonetic unit. As a result of this modeling the phonetic frontiers are also obtained.

In this thesis, we consider two refining techniques for the initial segmentation yielded by the HMM: the first approach uses multiple HMMs and combines all the individual estimates in a subsequent statistical analysis; the second approach performs a phonetic analysis of the speech signal to generate phonetic rules to separate different phonemes.

The thesis considers the performance of the two refining techniques in separate and also considers the combination of both techniques to achieve the best segmentation performance in terms of the mean absolute error (MAE) with respect to a given set of frontiers manually obtained by a professional phoneticist.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Considerações Iniciais

As técnicas de segmentação de um banco de fala contínua em unidades fonéticas têm se aprimorado em função do aumento da interação homem-máquina, seja pelo reconhecimento por parte da máquina dos comandos por voz, ou pela geração de fala sintetizada. Exemplificando, podem ser citadas aplicações como conversão texto-fala dos e-mails; das mensagens enviadas nos celulares, via serviço SMS; o apoio na alfabetização de línguas e a difusão de literaturas através de mídia sonora e dos sistemas de animação facial sincronizada com a fala.

O processo de segmentação pode ser manual ou automático. Ocorre o primeiro quando um foneticista profissional, por meio de análises das características temporais e espectrais do sinal da fala, define as fronteiras das unidades acústicas, e o segundo, quando determinadas fronteiras são definidas por técnicas implementadas em computadores que evidenciam os limites das características acústicas em estudo.

Das técnicas apresentadas na literatura científica para a segmentação automática, a mais utilizada é a do modelo oculto de Markov (HMM, do inglês *hidden Markov models*), que busca modelar as características temporais do sinal em função das propriedades estocásticas da acústica da fala [5], [6].

Este presente trabalho propõe o estudo da fusão de múltiplos HMMs (MHMMs) com outra técnica na fase de refinamento para aplicação na segmentação de banco de unidades para conversão texto-fala. Neste sentido, são apresentados resultados parciais do estudo da parametrização da técnica de HMM aplicada em um sistema de segmentação de dígitos concatenados. Em seguida, considera-se o uso de regras fonéticas, que usam uma modelagem determinística do sinal acústico característico de cada conjunto (ou classe) de fonemas de uma dada língua. A proposta central do presente trabalho é avaliar a combinação das técnicas estocástica (MHMM) e determinística (regra fonética) no processo de segmentação do sinal de fala.

1.2 Organização da Tese

A organização desta proposta se dá em seis capítulos.

No capítulo 2 são demonstradas as principais etapas envolvidas no sistema de conversão texto-fala e a técnica de síntese concatenativa.

O capítulo 3 apresenta os conceitos e algoritmos aplicados à técnica de HMM, que modela o processo da fala por um conjunto de estados. Os estados, associados a início, meio e final de cada fonema, contém os parâmetros acústicos associados a esses segmentos de cada fone. Deste modo serão verificadas as influências das transições entre os estados e a distribuição dos vetores de observações nas misturas por estado.

No capítulo 4 são apresentadas as técnicas de múltiplos HMMs [6] e de regras fonéticas [1] para se fazer um refinamento da estimativa da segmentação. A técnica de MHMM gera várias estimativas dos limiares que são, posteriormente, processadas para se gerar a estimativa final. As regras fonéticas procuram modelar cada tipo de fonema de uma língua a fim de detectar a transição de um fonema para outro. Esta técnica define então regras de caracterização de cada fonema e são, por isto mesmo, específicas para uma dada língua.

No capítulo 5, apresentamos os resultados gerados por cada etapa do sistema: com a modelagem HMM simples; com o uso de MHMMs; com a incorporação das regras fonéticas; e, por fim, com o sistema operando num modo misto, ativando ou não as regras fonéticas, quando achar que isto é vantajoso.

O capítulo 6 fecha a tese apresentando as principais contribuições do trabalho e apontando possíveis propostas para sua continuação.

Capítulo 2

Conversão Texto-Fala

2.1 Introdução

Na conversão texto-fala, as unidades acústicas segmentadas são concatenadas conforme a transcrição fonética (e possivelmente prosódica) do texto que se deseja sintetizar, gerando o sinal acústico daquela sentença. Para isso, é verificada a necessidade do conhecimento da sequência ortográfico-fonética, para a associação com unidades acústicas, e da técnica de síntese que realiza o encadeamento das unidades acústicas, que gerará o sinal de fala sintetizada com a entonação desejada. Neste capítulo são descritas as técnicas utilizadas atualmente para a implementação de um sintetizador de voz baseado na concatenação de unidades. Assim, na seção 2.2 são apresentados os módulos básicos de um sistema TTS (do inglês *text-to-speech*) como as etapas de normalização de texto, conversão ortográfico-fonética, concatenação de unidades e inserção de prosódia. Na seção 2.3 é feita uma análise do processo de geração do sinal da fala e dos possíveis efeitos de coarticulação entre unidades acústicas. Por fim, a seção 2.4 conclui o capítulo, ressaltando seus principais objetivos.

2.2 Teoria da Conversão Texto-Fala

A conversão texto-fala é uma mudança de domínio da representação da informação da forma escrita para a forma falada, em que se procura impor também as características da musicalidade da fala que o texto quer exprimir. Em geral, estas características são modeladas por meio de curvas de entonação, como entonação declarativa, afirmativa e interrogativa, assegurando-se à fala sintetizada um ritmo natural. De forma simplificada, todo o processo de conversão TTS pode ser dividido em duas etapas, análise do texto e síntese do sinal, como representado na figura 2.1. Na primeira etapa, o texto de entrada é normalizado e transcrito da sequência ortográfica para a sequência fonológica, procurando-se incorporar ainda

elementos determinantes da prosódia (possivelmente definidos pela pontuação original, pelo contexto ou mesmo por algum meta-comando de entrada pré-definido). Já na etapa de síntese, realiza-se a concatenação das unidades fonológicas e impõe-se um contorno prosódico adequado ao conteúdo do texto original. Pela importância do tema, o detalhamento de cada uma destas duas etapas é feito nas subseções a seguir.

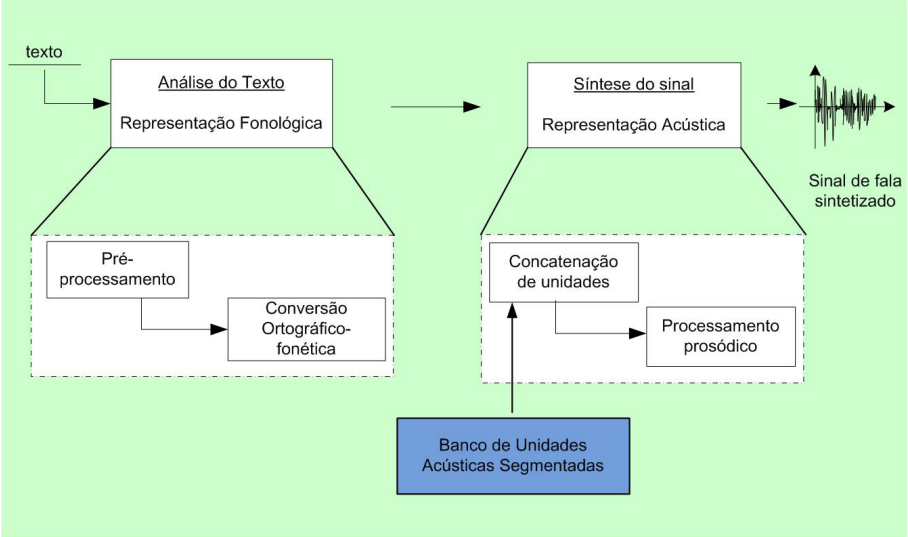


Figura 2.1: Etapas na geração de um sinal da fala sintetizado.

2.2.1 Etapa da Análise do Texto

A função desta etapa, dividida em pré-processamento e conversão ortográfico-fonética, é a extração da representação fonológica do texto escrito a ser entregue para a etapa seguinte. Em geral, num primeiro estágio, ocorre a normalização da sequência escrita de entrada, com a substituição de elementos de texto (incluindo, por exemplo, numerais, abreviaturas, siglas, sinais de pontuação etc.) pelo registro de palavras ou sequência de palavras por extenso, como exemplificado na tabela 2.1.

Já na etapa de conversão ortográfico-fonética, é obtida a sequência de fones que representa cada sequência de grafemas. Nessa conversão, um transcritor aplica regras que reescrevem para a representação fonética os caracteres ortográficos e marca outras informações relevantes, tais como acentuação lexical e fronteiras silábicas. A análise da sentença para a transcrição fonética é sempre dependente da língua com que se está trabalhando e até mesmo de sotaques locais. Alguns exemplos são mostrados na tabela 2.2 para o português brasileiro.[1]

[1]* Transcrição Biunívoca Brasileira - TBB.** Fonte de consulta: site http://www.radames.manosso.nom.br/

Tabela 2.1: Exemplos de normalização do texto de entrada.

Texto Original	**Texto Processado**
9	nove
10	dez
103,7	cento e três vírgula sete
23/12	vinte e três dividido por doze
23/12	dia vinte e três de dezembro
269-8050	dois meia nove oito zero cinco zero
2000 pessoas	duas mil pessoas
3ª série	terceira série
2º grau	segundo grau
Av.	Avenida
km^2	quilômetro quadrado
Cap. 1	capítulo um
Cap. Amâncio	capitão Amâncio
“	abre aspas
”	fecha aspas

A transcrição ortográfico-fonética é realizada por meio de um **Dicionário de Pronúncias** ou **Regras de Transcrição**. O dicionário de pronúncias contém a representação fonética e o padrão de acentuação de cada uma das palavras da língua, enquanto as regras de transcrição tratam das correspondências regulares entre letras e sons, como

- letra “**c**” sucedida pelas vogais “**e**” ou “**i**” deve ser transformada no fonema /s/;

A dificuldade do uso do dicionário é o seu tamanho e, em consequência, a localização da palavra por seu algoritmo [7]. Já as regras pecam por sua falta de universalidade. Certas palavras no português são homógrafas e não homófonas, ou seja, ortograficamente equivalentes mas com pronúncias diferentes, e a sua distinção depende de uma análise semântica (linguística). Por exemplo, a palavra **sede**, pertencente à classe gramatical substantivo feminino, pode ser utilizada como:

- A **sede** da ONU fica em Nova York.
- Obedeça sua **sede**, beba água.

Para esses casos, o sistema conversor texto-fala desenvolvido na UNICAMP [7] utilizou um dicionário de exceções, constituído pelos 1.383 verbetes do Minidicionário Aurélio [8].

Tabela 2.2: Exemplos de conversão ortográfico-fonética.

Fonema	**Grafema**
/ s /	Grafema s: *s*ela
	Grafema c: *c*edo
	Grafema x: pró*x*imo
/ j /	Grafema g: *g*ente
/ g /	Grafema g: *g*ato
/ R /	Grafemas rr: Ca*rr*o
/ λ /	Grafemas lh: A*lh*o
/ x /	Grafemas ch: Fi*ch*a
/ k / + / s /	Grafema x: Tó*x*ico
Grafema não mapeado para nenhuma unidade fonológica	
Grafema h da palavra *h*omem	

2.2.2 Etapa da Síntese do Sinal

A segunda etapa da conversão texto-fala consiste na geração de um sinal acústico correspondente ao texto de entrada. Neste caso, recebe-se a sequência de fonemas determinada pelo módulo de transcrição ortográfico-fonética e produz-se um sinal de conteúdo sonoro correspondente, inclusive com o prosódico desejado.

Diferentes estratégias para gerar um sinal acústico sintético são apresentadas na literatura, como, por exemplo: síntese por regras, síntese articulatória, síntese baseada em cadeias de Markov e síntese concatenativa. Por ser muito utilizada [1], [9], a síntese concatenativa é abordada nesta tese. Nesta técnica, ocorre uma seleção das unidades acústicas (comumente, difone ou trifone) de um banco de fala (dicionário) conforme definido pela etapa de transcrição ortográfico-fonética. Essas unidades são concatenadas e a elas são impostos contornos prosódicos (duração, *pitch* e energia) adequados para se obter a entonação desejada, como ilustrado na figura 2.2. Na síntese PSOLA, a concatenação é feita síncrona com as marcas de *pitch*. Quando concatenamos difones ou trifones, a junção é feita na posição central dos fones extremos. Na junção de dois segmentos, usa-se a superposição de 1 período de pitch.

Para que a concatenação seja suave, as unidades acústicas não podem ser simplesmente justapostas, pois isso, geralmente, resulta em ruído audível (estalido) no momento da junção, devido a possíveis diferenças entre a última amostra do primeiro segmento e a primeira amostra do segmento seguinte, conforme se mostra na figura 2.3. Para evitar este efeito indesejado, é necessário garantir uma transição gradual e suave entre os dois segmentos, como indicado na figura 2.4. Para isto, a solução mais simples é calcular uma combinação linear dos dois sinais, alinhados de modo que eles se sobreponham por certo número m de amostras. Nessa com-

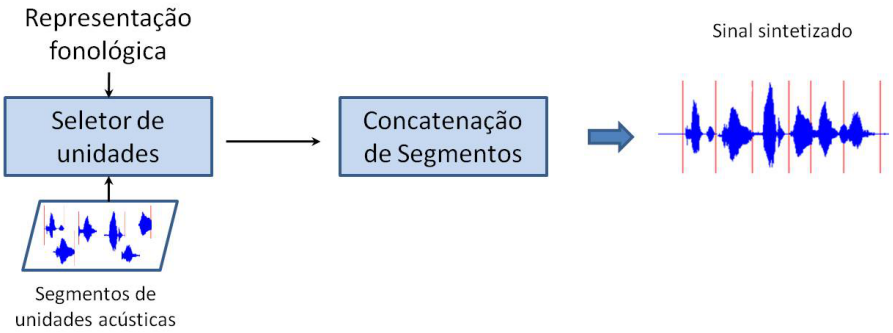


Figura 2.2: Representação simplificada do método de síntese concatenativa.

binação, o peso do primeiro segmento cai de 1 para 0 no decorrer dessas m amostras, ao mesmo tempo que o peso do segundo aumenta de 0 para 1.

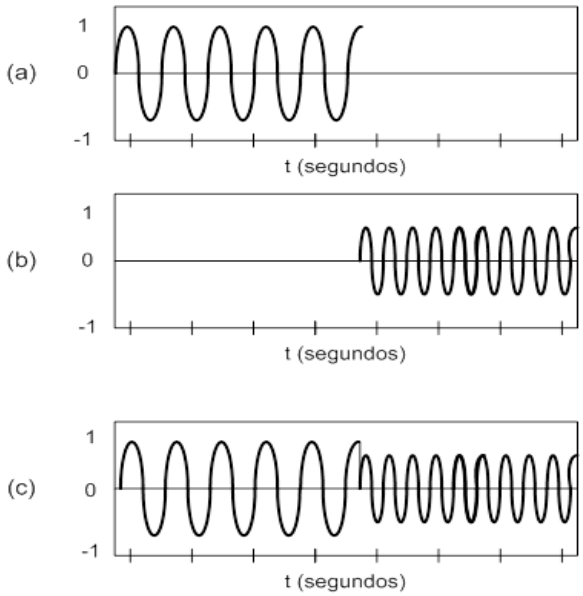


Figura 2.3: Concatenação abrupta de duas unidades de fala por simples justaposição: (a,b) As unidades separadas; (c) Resultado da concatenação com descontinuidade.

A prosódia confere uma estrutura oral à sentença, permitindo a quebra mental da mensagem pelo ouvinte e facilitando a sua compreensão. Além disto, a prosódia acrescenta uma individualidade ao falante, quer seja através da identificação do sexo, como também traços da personalidade quanto à arrogância ou timidez, e emoções, como alegria e tristeza.

Os parâmetros utilizados para obter a prosódia da sentença são a frequência fundamental, a intensidade e a duração [10]. A frequência fundamental dos sinais sonoros é a frequência de vibração das cordas vocais, que é determinada pelo comprimento e tensão das cordas: quanto mais tensas, maior a frequência e mais agudo o som; quanto maiores, menor a frequência e mais grave o som. Em particular,

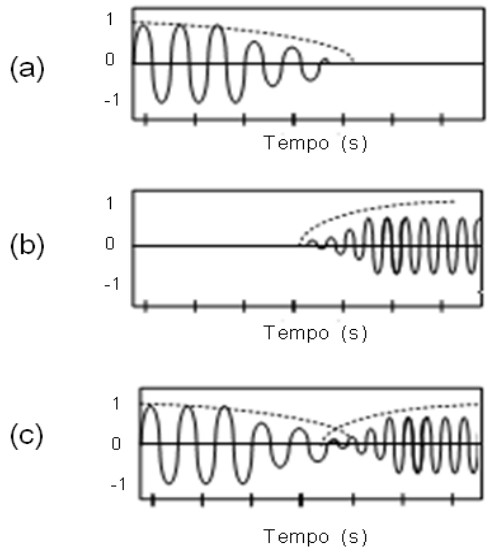


Figura 2.4: Concatenação suave de duas unidades de fala: (a,b) Unidades recortadas através de função de janelamento; (c) Resultado da combinação linear com sobreposição.

o comprimento das cordas vocais é a principal causa da diferença entre as alturas naturais da voz infantil, voz feminina e voz masculina. Nos segmentos de fala não sonoros (surdos), não há sentido falar em frequência fundamental. Outro parâmetro prosódico, a intensidade, possui uma função de contraste menos significativa em relação à frequência fundamental e à duração [11]. De modo geral, mas não de forma unívoca, a intensidade está diretamente relacionada ao aspecto de tonicidade de uma sílaba. Já o parâmetro de duração, que quantifica a diferença de tempo entre dois eventos, imprime o ritmo e a pausa na frase, determinando, de acordo com Barbosa [10], contornos que identificam acentos lexicais.

Talvez a solução mais simples e popular para a inserção da prosódia no sinal concatenado, possuindo por hipótese uma prosódia considerada neutra, é o método TD-PSOLA (*Time Domain - Picth Syncronous Overlap and Add*) [12]. O algoritmo PSOLA trabalha de forma síncrona com o período de pitch do sinal, assim, a qualidade do sinal produzido dependente de algoritmo de marcação de pitch eficiente [9], essa marcação, efetuada uma única vez, são posicionadas nos picos do sinal de cada período de pitch e chamada por marcas de pitch. O algoritmo TD-PSOLA pode ser dividido em três passos, a saber:

1. O sinal original difone ou trifone é decomposto em uma sequência de sinais menores e parcialmente sobrepostos, denominados de sinais elementares ou elementos, cuja soma resulta no sinal original (Figura 2.5). Para os sons sonoros, periódicos ou quase-periódicos, a duração dos sinais elementares equivale

a um período fundamental. Neste processo de segmentação, é utilizada uma função de janelamento para cada sinal elementar, evitando-se assim transições abruptas no início e no fim das unidades. Para trechos surdos, costuma-se utilizar uma duração padrão ou o último valor determinado em um trecho sonoro.

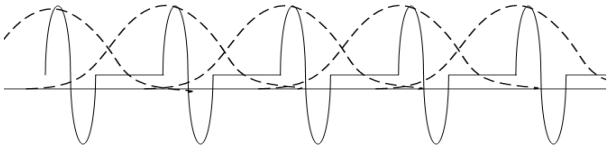


Figura 2.5: Decomposição de um sinal de voz em sinais elementares de forma síncrona com a frequencia fundamental para o método TD-PSOLA.

2. O algoritmo TD-PSOLA permite a alteração da frequência fundamental do sinal sintetizado, aproximando-se ou afastando-se os sinais elementares de modo a aumentar ou diminuir a frequência fundamental, respectivamente, como ilustrado na figura 2.6. Este processo altera a duração do sinal e, portanto, geralmente exige um ajuste apropriado da mesma, com a omissão de alguns de seus elementos, para reduzi-la, ou a duplicação de alguns deles, para aumentá-la, conforme ilustrado na figura 2.7.

   Em trechos surdos, o método TD-PSOLA pode não produzir bons resultados, pois a duplicação de elementos torna o sinal sintetizado quase-periódico, podendo produzir sons metálicos [13].

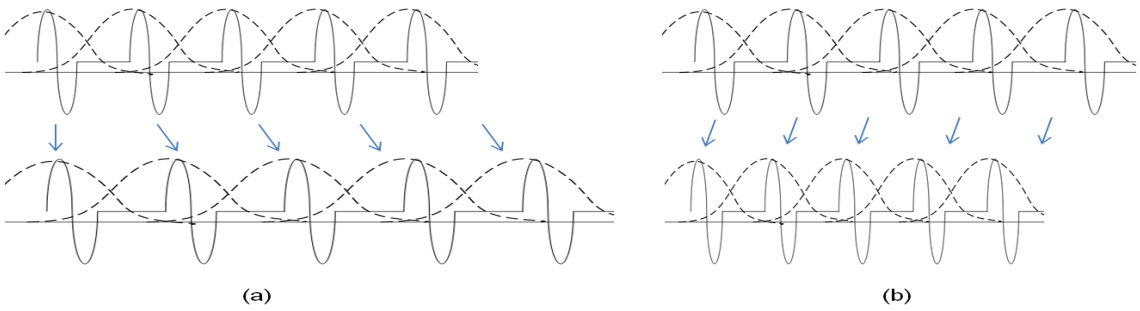


Figura 2.6: (a) Aumento do período provocando a diminuição da frequência fundamental de um sinal. (b) Redução do período provocando o aumento da frequência fundamental de um sinal.

3. O terceiro e último passo consiste na adição dos sinais elementares, já devidamente alinhados no tempo de acordo a frequência fundamental desejada, para se obter o sinal sintetizado com a devida entonação.

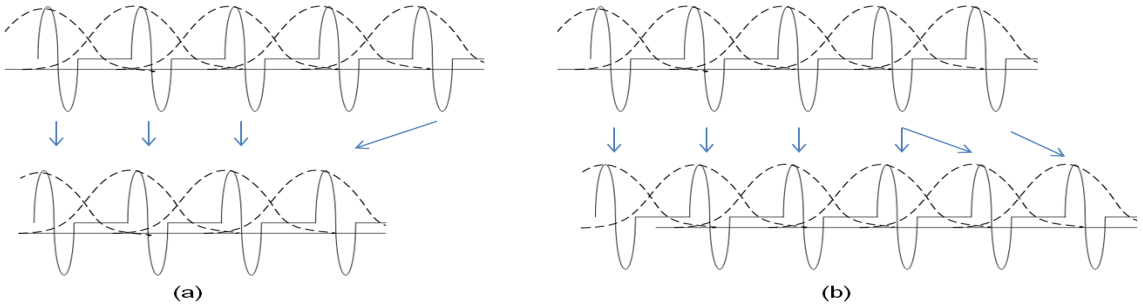


Figura 2.7: (a)Redução da duração de um sinal de voz por omissão de sinais elementares. (b) Aumento da duração de um sinal de voz por duplicação de sinais elementares.

2.3 Banco de Unidades

É de se observar que, durante a fala, o trato vocal muda relativamente devagar, comparado com as vibrações das pregas vocais. A duração mínima de um fone é determinada pelo tempo necessário para que os nervos e músculos consigam modificar a articulação, que é da ordem de 50 milissegundos, o que corresponde a uma taxa máxima de emissão de 20 fones por segundo [5].

Como os sons não são pronunciados discretamente (pausadamente), mas juntando ou subtraindo fonemas adjacentes e passando para o próximo fonema durante a fala, isto gera coarticulação entre unidades consecutivas, resultante da economia do esforço na produção contínua da fala. Este fenômeno de coarticulação deve ser, então, representado nas unidades fonéticas utilizadas na síntese concatenativa [14], [15]. Um dos possíveis tipos de unidade é o difone (junção de dois fones), que é capaz de capturar grande parte do fenômeno coarticulatório que ocorre entre os segmentos subjacentes e de minimizar as descontinuidades, quando o processo de concatenação ocorre em trechos espectralmente estáveis. Usando difones, por exemplo, a palavra *casa* é composta da forma $/\#k/, /ka/, /az/, /zA/e/A\#/$, onde $/\#/$ indica uma pausa entre palavras. Neste tipo de sistema, o dicionário é de tamanho médio (em torno de um a dois milhares de unidades), pois é guardada apenas uma cópia de cada par de fones que ocorre na língua. Em diversas línguas, alguns sons podem exigir trifones (triplas de fones) para uma melhor representação sonoro, como em /tra/ e /pra/ no Língua Portuguesa. A síntese concatenativa por difones e trifones pode produzir som de qualidade aceitável ou obter melhores resultados com o uso de unidade maiores (polifones), agregando mais coarticulação. Neste caso, o tamanho do dicionário torna-se desvantajoso [10], já que a coarticulação e a prosódia possuem efeito sobre sequências de 10 ou mais fonemas adjacentes, tornando o número de contextos relevantes extremamente grande [5].

Uma das técnicas de construção de banco de unidades que facilita a obtenção de sentenças sintéticas próximas do natural é a seleção de unidades fonéticas monótonas

(isto é, de prosódia neutra) em frases-veículos [9]. O algoritmo para extração dessas unidades segue os passos:

- Criar frase-veículo para inserir logotoma, isto é, uma unidade isolada sem significado dentro da frase. O objetivo da frase-veículo é prover um ambiente prosodicamente neutro. É importante, neste processo, considerar os casos em que o logotoma é precedido e/ou sucedido por um trecho de silêncio.
- Gravar a frase com logotoma, evitando a geração da prosódia, ou seja, mantendo a mesma taxa de elocução e entonação regular ao longo de todo o processo de gravação;
- Extrair a unidade logotoma da frase-veículo;
- Segmentar e etiquetar as diversas sub-unidades (fones, difones, trifones etc.) que formam o logotoma em questão, tarefas estas realizadas apenas uma vez, na criação do banco de unidades.

2.4 Conclusão

Foi apresentada neste capítulo a base da implementação de um sistema de conversão texto-fala, verificando-se que o texto a ser convertido para sinal de fala deverá ser normalizado, reescritas todas as abreviaturas, pontuação, símbolos especiais e numeração para o extenso. Desta sequência de grafemas, obtém-se a sequência fonológica e, a partir desta, realiza-se a conversão do texto-fala, através de alguma técnica de síntese. Finalmente, aplica-se o método TD-PSOLA, que permite o ajuste da prosódia da sentença, alterando duração e frequência.

Foi verificado que o banco de unidades deve ser constituído por unidades com comprimentos variados, abrangendo o máximo dos efeitos da coarticulação facilitando-se, assim, ao sistema TTS realizar a concatenação dessas unidades acústicas.

No capítulo seguinte, serão apresentados conceitos de algumas técnicas que realizam a segmentação do sinal da fala em unidades fonéticas, utilizando fusões de algoritmo e dividindo a estimação do limiar em etapas.

Capítulo 3

Acerca da Segmentação de Sinais de Voz

3.1 Introdução

Atualmente, há uma divulgação crescente, pela comunidade científica, de pesquisas que realizam a fusão de diferentes técnicas para a estimação dos limiares ou fronteiras dos fones, segmentando o sinal da fala. Em geral, este processo de segmentação é dividido em etapas, nas quais, primeiramente, obtém-se um limiar "provisório" e, posteriormente, é feito um refinamento, onde se encontra para o limiar um valor mais próximo do obtido por um foneticista profissional.

Neste capítulo, faz-se um estudo de algumas técnicas de segmentação de sinais de voz. As técnicas aqui apresentadas possuem na sua primeira etapa, o uso de modelos ocultos de Markov (HMM, do inglês *hidden Markov models*), que estimam o limiar entre os fones por meio do alinhamento forçado do algoritmo de Viterbi. Em particular, vamos considerar o uso de um conjunto de modelos HMM, denominado por múltiplos HMM (MHMM), que através de uma métrica, obtém uma tendência (ou *bias*) que permite a redução do erro de uma primeira estimativa do limiar. Para a segunda etapa, a do refinamento, outras técnicas realizam o ajuste das medidas encontradas, como por exemplo o uso de regras fonéticas, que procuram características específicas de grupos fonéticos que facilitam a sua delimitação.

Assim, este capítulo estrutura-se da seguinte forma: na seção 3.2 serão apresentadas as motivações para o desenvolvimento de fusões de algoritmos, como a eliminação do erro sistemático e o algoritmo desenvolvido por Jarifi [16], que realiza fusão de três técnicas obtendo resultado de 95,73% de acerto na segmentação do sinal da fala; na seção 3.3 será discutido o algoritmo baseado em MHMMs apresentado por Park [6], com sua aplicação na base YOHO, permitindo-se um estudo de critérios de treinamento; na seção 3.4, serão mostradas as regras definidas por Selmini [1],

para o refinamento dos fones do português brasileiro. Estas técnicas serão posteriormente combinadas no âmbito do trabalho desta tese de doutorado procurando-se atingir um desempenho superior aos desempenhos obtidos individualmente por estes mesmos métodos.

3.2 Conceitos Preliminares

De modo geral, podemos considerar que a fronteira entre dois fones "corresponde a um instante de tempo em que as características acústicas ou fonéticas, de um determinado fone, se tornam menos perceptíveis e as características do fone seguinte se tornam mais perceptíveis à medida que o tempo passa" [1].

O HMM é constituído de um conjunto de N estados, cujas transições são governadas por uma dada distribuição de probabilidades. Associado a cada estado, há um conjunto de observações dividido em M distribuições gaussianas [2], com vetor média e matriz covariância. Dessa forma, para cada mistura associada a um estado, existe uma função de densidade de probabilidade que define a probabilidade da observação pertencer àquela mistura e àquele estado em conjunto, como indicado na figura 3.1. Assim, tem-se um par de processos estocásticos (X,Y), onde o processo X é uma cadeia de Markov de primeira ordem ou a sequência de estados, não sendo diretamente observável; e o processo Y, uma sequência de variáveis aleatórias no espaço dos parâmetros acústicos (observações).

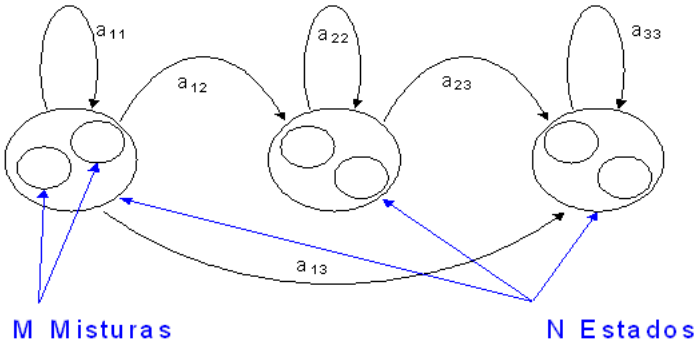


Figura 3.1: Diagrama de modelo oculto de Markov (HMM) do tipo Bakis com N estados, M misturas e as respectivas probabilidades de transição de estados.

Para a caracterização de um HMM para uma determinada unidade fonética é feito um treinamento utilizando o algoritmo de Baum-Welch e, para o reconhecimento correspondente usa-se o alinhamento de Viterbi [2]. O procedimento iterativo chamado de método de Baum-Welch, conhecido também como reestimação *forward-backward*, realiza a estimação das matrizes de probabilidades de transição entre os estados e a matriz de distribuição da probabilidade de observação em cada estado,

ajustando-as de forma a maximizar a probabilidade da seqüência de observações, dado o modelo $P(O|\lambda)$. Como a quantidade de seqüências de treinamento é finita, não existe solução ótima para essa estimativa. Pode-se, entretanto, estimar as matrizes de modo que $P(O|\lambda)$ seja localmente maximizado.

Já no alinhamento, é gerada uma estimativa inicial das fronteiras de segmentação que serão posteriormente refinadas, como indicado na figura 3.2. Neste caso, para cada coluna da figura 3.2 são armazenados os valores das verossimilhanças acumuladas em cada estado do modelo HMM, para todos os instantes de tempo considerados. Cada intervalo de tempo entre as colunas corresponde a uma janela de análise do sinal que está sendo processada pelo algoritmo. Ao final, é encontrada pelo alinhamento de Viterbi uma seqüência ótima de estados q_t*, dentre todas as possíveis sequências de estados q, relacionada com a sequência de observação $\boldsymbol{O}$ aplicada.

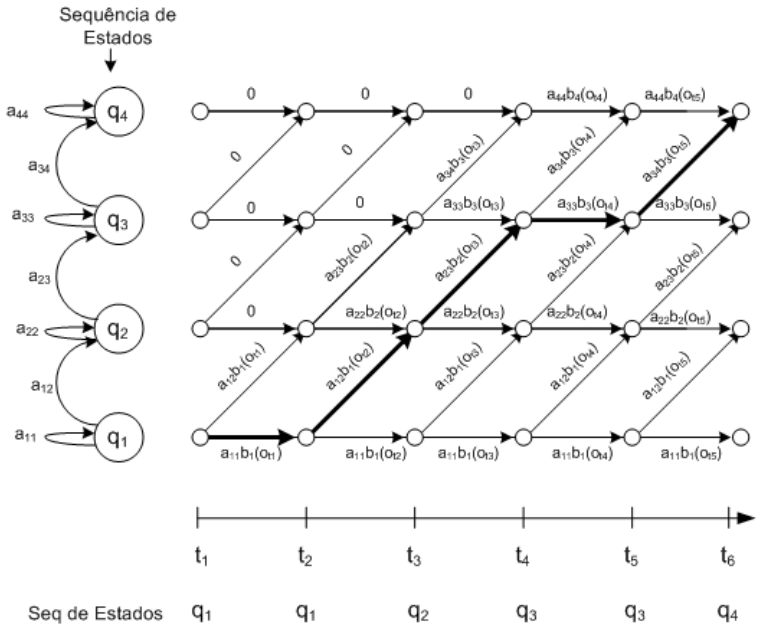


Figura 3.2: Uso do alinhamento forçado de Viterbi para definir as fronteiras das unidades acústicas no reconhecimento: para cada nó (estado) é acumulado o valor da verossimilhança, probabilidade de transição de um estado para o outro, em conjunto com a ocorrência do vetor de observação o_t. A sequência "ótima", que aparece em negrito, maximiza esta verossimilhança. [1].

Para a geração dos modelos dos fones foi utilizada a ferramenta computacional HTK (do inglês *hidden Markov model toolkit*), desenvolvido no *Cambridge University Engineering Department* [17] e amplamente usado para reconhecimento da fala, permitindo a construção de modelos HMM e a manipulação de seus parâmetros de forma simples e robusta. O HTK é composto por quatro blocos, que são:

- **Preparação dos Dados**: A partir de um segmento do sinal de fala, gera os vetores de características tais como LPC, MFCC, coeficientes delta, energia.

- **Treino**: Permite estimar HMMs referentes a cada unidade e treiná-los com os vetores de características das locuções.

- **Teste**: Permite testar a capacidade de reconhecimento (e consequente segmentação) das HMMs com as locuções de teste.

- **Análise dos resultados**: Obtém as estatísticas de desempenho do reconhecimento (e segmentação) das locuções de teste, como percentual e número absoluto de acertos.

Neste trabalho o pacote HTK foi utilizado nas etapas de preparação dos dados, treino e testes, enquanto que a análise dos resultados foi feita com as ferramentas Sound Forge v7.0 (www.sonycreativesoftware.com/soundforge), Audacity (audacity.sourceforge.net/) e Praat (www.praat.org/) para uma melhor visualização de todo o processo de segmentação. Na implementação do HTK, foram observadas as seguintes considerações:

- Para a implementação dos modelos no HTK devem ser acrescentados dois estados (de entrada e saída) no número total de estados, que servirão para conexões entre os HMMs, indicados na figura 3.3 como estados $S1$ e o $S5$.

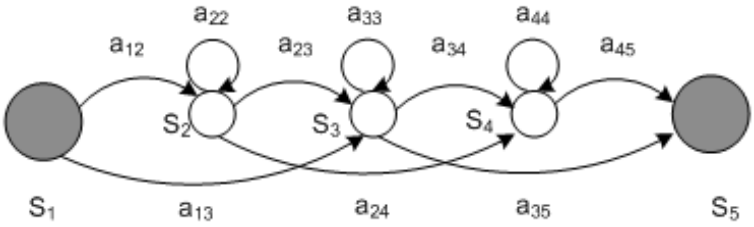


Figura 3.3: - Representação dos estados no HTK explicitando o estado de entrada $S1$ e de saída $S5$.

- Para a fase de treinamento *embedded* e reconhecimento no HTK, deve ser implementada uma rede dinâmica, isto é, uma sequência concatenada de HMMs que o HTK gera logo após receber a sequência de fones para reconhecimento, como indicado na figura 3.4 para a palavra [k][a][z][a].

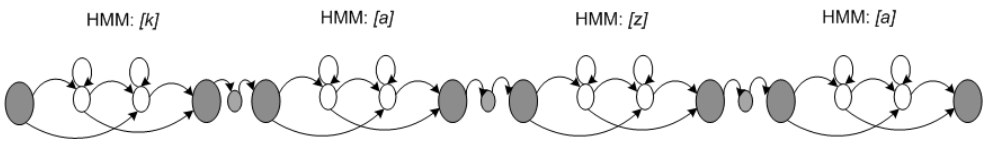


Figura 3.4: Rede dinâmica da locução "casa" criada para treinamento das unidades acústicas isoladas [k], [a], [z] e [a].

A utilização de n modelos HMM para a segmentação do sinal da fala em unidades fonéticas gera, em sua saída, n limiares estimados para a fronteira real entre os fones da locução. Assumindo-se que a fronteira de um determinado fone é afetada pelos fones adjacentes (à direita ou à esquerda), que compõem a locução [18], [1], e, sabendo-se que, a técnica HMM é modelada por funções probabilísticas, é razoável supor que os valores das fronteiras, quando devidamente estimadas, situam-se em uma região em torno da fronteira de referência (obtida, por exemplo, pela marcação manual por um foneticista).

A média da diferença entre as fronteiras estimadas e a de referência é denominada erro sistemático, tendência ou viés [19], [20]. A redução ou eliminação do erro sistemático é objeto de várias pesquisas dentro da área de processamento de voz. Algumas destas pesquisas aplicam uma etapa de pós-processamento para refinar os resultados, tais como em [6], [1], [21], [22], e [23]. A técnica de modelagem por HMM tem sido utilizada para estimar uma região próxima a fronteira e, em conjunto com outras técnicas na etapa de pós-processamento, tem obtido estimativas aceitáveis da fronteira entre fones com erro abaixo de 20 ms na fase de refinamento [6], [1], [16], [20]-[24].

No trabalho apresentado por Jarifi [16], por exemplo, é estimada, em uma primeira fase, uma região em torno do limiar obtido usando a técnica HMM. Na fase seguinte, de pós-processamento, são aplicados dois outros algoritmos para refinamento do limiar. Para o primeiro algoritmo, são utilizados modelos de misturas gaussianas (GMM, do inglês *Gaussian mixture model*) para a modelagem das unidades fonéticas. Para aumentar a precisão na identificação do limiar, cada GMM é alimentado por um pequeno banco de dados, rotulado e segmentado manualmente, e em sua saída é estimado um novo limiar refinando a região em torno da marca estimada grosseiramente, dada pelo HMM. Em seguida, é aplicado um segundo algoritmo de refinamento, o método GLR Brandt (do inglês *Brandt's generalized likelihood ratio*). O objetivo deste método é detectar descontinuidades do sinal da fala no intervalo em torno da marca reestimada, definida pelas técnicas anteriores. Neste trabalho, foram obtidas taxas de 95,73% para limiares com erro abaixo de 20 ms em relação ao limiar de referência.

A seguir, nas seções 3.3 e 3.4 estudaremos duas técnicas de refinamento que têm, com sucesso, conseguido reduzir o erro sistemático de sistemas de segmentação automática para níveis bastante razoáveis (abaixo de 20 ms). Posteriormente, procuraremos combinar estas duas técnicas para atingir um desempenho ainda superior aos desempenhos obtidos pelas técnicas individualmente.

3.3 Refinamento por Múltiplos HMMs

É fato que devido à natureza estocástica dos HMMs os vetores do sinal da fala (observações) distribuem-se nas M misturas de cada um dos N estados em função da distribuição de probabilidade das transições entre os estados e, em cada estado, em função dos pesos para cada gaussiana. Desta forma, se usamos n diferentes HMMs (com diferentes valores de M ou N), obtemos n diferentes resultados para um processo de segmentação, possivelmente com valores próximos entre si quando os n modelos são treinados com a mesma base de dados. Este processo é ilustrado na figura 3.5.

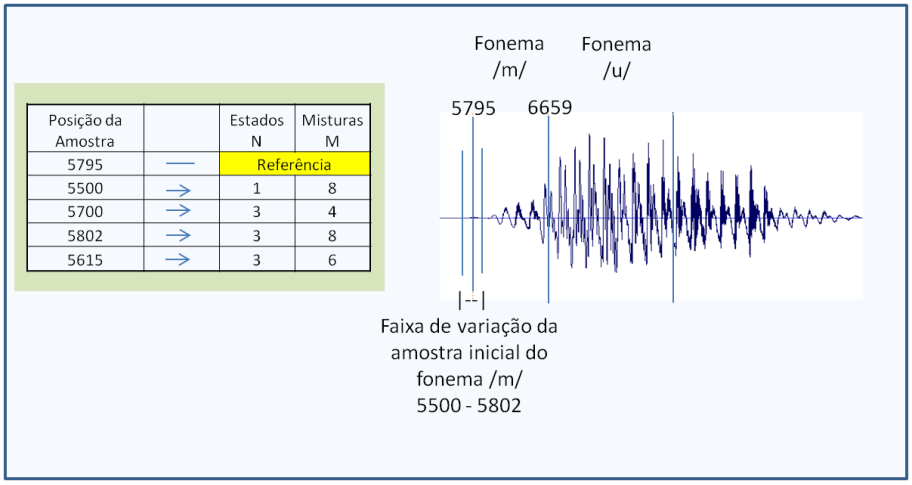


Figura 3.5: Variação das posições estimadas por HMM treinados por diferentes parâmetros de estados e gaussianas.

Neste sentido, uma fusão de múltiplos HMMs com uma técnica de otimização foi proposta por Park [6]. Nesse trabalho, cada HMM é visto como uma "máquina de segmentação automática (ASM, do inglês, *automatic segmentation machine*). São usadas, então, 33 ASMs com diferentes parâmetros, ou múltiplas ASM independentes, que estimam, dadas as locuções e suas correspondentes transcrições fonéticas, uma sequência de 33 fronteiras possivelmente distintas para cada fronteira presente na locução. Numa etapa de treinamento, é calculado o erro sistemático deste conjunto de estimativas e aplica-se a técnica de otimização pelo método projeção de gradientes para se determinar o peso de cada estimativa individual na composição da estimativa final de cada fronteira. Neste processo de otimização, o objetivo é se ajustar o peso de cada estimador de modo a minimizar a distância entre as fronteiras de referência e estimada, como ilustrado na figura 3.6.

Com este sistema [6], Park obteve uma taxa de reconhecimento de 97,05% para a base de dados coreana, distribuída da seguinte forma: 1600 locuções segmentadas manualmente para treinamento dos modelos isolados; 5000 locuções não segmenta-

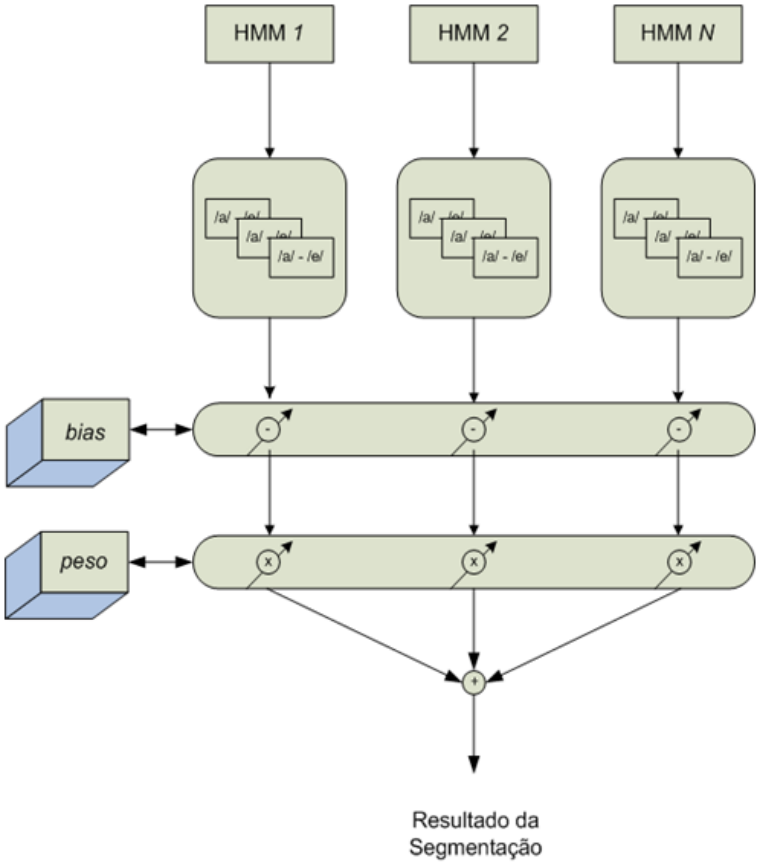


Figura 3.6: Diagrama de blocos do algoritmo de Park com múltiplos HMMs para segmentação automática de voz. Neste algoritmo, diversas estimativas individuais são combinadas, de forma ponderada, para se determinar a estimativa final de uma fronteira de segmentação.

das na fase de treinamento *embedded* modelos concatenados(cálculo dos pesos); e 400 locuções para avaliação do sistema.

3.3.1 Múltiplas ASM Aplicadas à Base YOHO para Segmentação de Dígitos

Nesta subseção descrevemos o uso da técnica de MHMMs de Park na segmentação de palavras (dígitos) da base YOHO [25].

A base de fala YOHO foi criada nos Estados Unidos, pela *International Telephone and Telegraph Corporation* (ITT), para treinamento e testes de sistemas protótipos de verificação do locutor. A base consiste de gravações de sequências acústicas em inglês de três dezenas concatenadas, como, por exemplo, 21-35-63 ("twenty one, thirty five, sixty three"), sem regras para a pausa entre os números, e gravadas em ambiente de escritório a uma taxa de amostragem de 8000 Hz. As locuções são comumente divididas em dois grupos: o primeiro, *enroll*, com 96 locuções por locutor

e, o segundo, *verify*, com 40 locuções por locutor. Nesta base foram utilizados 138 locutores, sendo 108 masculinos e 30 femininos.

A figura 3.7 ilustra, nos domínios do tempo e da frequência, o sinal 93_39_76.wav segmentado manualmente entre os limites dos seus dígitos. Os trechos de silêncio são utilizados para criar o HMM do silêncio.

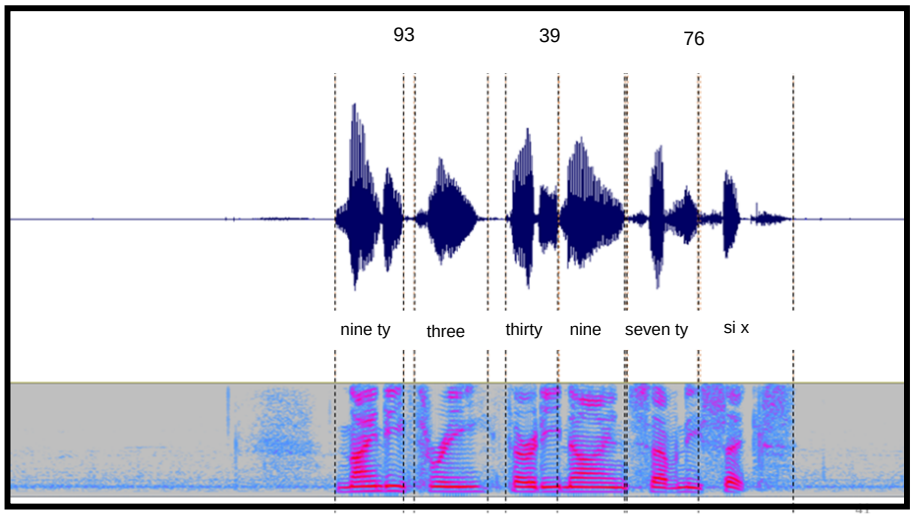


Figura 3.7: Sequência de dígitos, com a devida segmentação manual, correspondente ao sinal 93_39_76.wav da base YOHO.

Para efeito de demonstração do algoritmo de Park, de forma aleatória, foi escolhido o locutor da base YOHO de número 101 para treinamento e teste. Neste sentido, todas as locuções deste locutor, tanto do conjunto de treinamento quanto do conjunto de testes, foram manualmente segmentadas pelo próprio autor desta tese.

Quanto ao modelo HMM utilizou-se o modelo *left-right* ou *Bakis*, considerado como o mais apropriado para a representação do sinal da fala, visto que nele se tem a sequência de ocorrência sempre crescente dos estados associada de forma direta à linha do tempo, conforme sugestão de Rabiner [2].

As locuções foram segmentadas em blocos de 20 ms, usando-se a janela de Hamming, com justaposição de 50% entre janelas consecutivas, isto é, o início de cada segmento ocorria a cada 10 ms. Para cada segmento um conjunto de características era obtido, incluindo: C coeficientes mel-cepstrais c_t (MFCC, do inglês *mel-frequency cepstral coefficient*), o logaritmo da energia de cada segmento, e suas respectivas derivadas D_t de primeira e DD_t de segunda ordens no instante t, totalizando $3(C+1)$ coeficientes. Em nossos testes, consideramos C igual 12, 14 e 18, correspondendo a vetores com 39, 45 e 57 coeficientes, respectivamente.

Os coeficientes mel-cepstrais são determinados por [2], [26]:

$$MFCC_i = \sum_{m=1}^{M} \left\{ E_m \cdot \cos\left[i \left(m - \frac{1}{2} \right) \frac{\pi}{M} \right] \right\}, \quad i = 1, 2, ..., C \tag{3.1}$$

onde $M = 22$ é o número total de filtros do banco e E_m é o logaritmo da energia calculada na saída do m-ésimo filtro.

No caso, as derivadas são utilizadas para se detectar alguma variação brusca dentro do espectro da voz [27], aumentando a robustez para a identificação da unidade acústica, e para os MFCCs são calculadas como [2]

$$D_t = \frac{\sum_{k=1}^{N_d} k(c_{t+k} - c_{t-k})}{2\sum_{k=1}^{N_d} k^2}, \quad DD_t = \frac{\sum_{k=1}^{N_d} k(D_{t+k} - D_{t-k})}{2\sum_{k=1}^{N_d} k^2}. \tag{3.2}$$

onde N_d é a distância em janelas para qual se quer calcular a diferença. Além disto, o logaritmo da energia do sinal x é ainda calculado como

$$E = 10 \log \left(\sum_{i=0}^{N-1} x^2(i) \right), \tag{3.3}$$

onde N representa o número de amostras que compõem um segmento do sinal. As derivadas da energia são determinadas de forma análoga às derivadas obtidas para os MFCCs.

Para treinamento dos HMMs que representam determinado sinal de fala, são utilizados dois critérios de busca: treinamento exaustivo e de fones por estado.

Para o *treinamento exaustivo dos modelos*, variam-se o número N de estados, o número M de gaussianas e o número C de MFCCs. O resultado é analisado, verificando-se quais os conjuntos de parâmetros que apresentam menores taxas de erro na segmentação. A faixa de variação dos parâmetros para a modelagem dos HMMs dos dígitos baseou-se no trabalho realizado por Park [6]. Assim, consideramos $1 \leq N \leq 5$ estados/dígito, $1 \leq M \leq 8$ gaussianas e C igual a 12, 14 ou 18 MFCCs para cada segmento. Para o modelo HMM do silêncio definiu-se ter $N = 1$ estado e $M = 1$ gaussiana. Deve-se considerar que o banco de dados para treinamento é pequeno, restringindo, assim, o uso de maiores quantidades de estados ou gaussianas, como é sugerido no trabalho de Park [6]. Os parâmetros relacionados com o vetor de características seguem novamente as sugestões de Park.

Para o *critério de fones por estados* são aplicadas as sugestões de Rabiner [2]. Neste sentido, o número N de estados é um parâmetro empírico e a sua escolha

deve ser associada ao número de eventos acústicos (fones) que existem dentro de uma unidade fonética (dígito), atribuindo-se um estado por fone, como apresentado na tabela 3.1. Quanto ao número de gaussianas por misturas, utilizou-se como referência o trabalho realizado por Selmini [1], no qual o número de gaussianas variou de 1 a 20, e neste trabalho, considerando o reduzido tamanho da base de dados utilizada, definiu-se a faixa $1 \leq M \leq 10$. O número de coeficientes C é referente ao modelo que apresentou melhor resultado para o teste exaustivo. Para o cálculo das derivadas, foi utilizado o valor $N_d = 1$ sugerido por Selmini [1] e os valores $N_d = 2$ e 4 utilizados por Park [6].

Tabela 3.1: Representação fonêmica dos dígitos conforme padrão inglês (HOUAISS, 1984) e o número de estados atribuído a cada dígito, em função do número de fones, de acordo com o critério de fones por estados.

Dígito	Representação Fonética	Estados N	Gaussianas M
one	/*wân*/	3	
two	/*tu′*/	2	
three	/*ϑrï′*/	3	
four	/*fó′ɛ*/	3	
five	/*fayv′*/	4	Faixa de valores
six	/*siks′*/	4	utilizados
seven	/*se′vɛn*/	5	2 a 10
nine	/*nayn′*/	4	
twenty	/*twen′tï*/	6	
thirty	/*ϑ***E***′tï*/	4	
forty	/*fó′tï*/	4	
fifty	/*fif′tï*/	5	
sixty	/*siks′tï*/	6	
seventy	/*se′vɛntï*/	7	
eighty	/*eyt′tï*/	5	
ninety	/*nayn′tï*/	6	

O desempenho avaliado pela métrica de erro absoluto médio (MAE, do inglês *mean absolute error*) dos HMMs caracterizados pelos critérios exaustivo e de fones para as locuções do locutor 101 da base YOHO é indicado na tabela 3.2. Desta tabela, percebe-se que ambas as técnicas obtiveram um MAE abaixo de 20 ms para quase todas as unidades consideradas, com um melhor desempenho global da técnica de critério de fones, que apresentou uma MAE geral de 7,7 ms contra 9,7 ms para a técnica exaustiva.

Com as variações de $2 \leq M \leq 10$ e $N_d = 1, 2, 4$ têm-se 27 HMMs distintos para

Tabela 3.2: Desempenho MAE da segmentação inicial e final para as locuções da base YOHO com sistemas HMMs usando treinamento exaustivo ou por critério de fones/dígito.

Modelos	Método Exaustivo						Método Fonemas					
	Início (ms)	Fim (ms)	N	M	N_d	C	Início (ms)	Fim (ms)	N	M	N_d	C
one	13,31	10,10	2	8	2	12	11,5	9,8	3	3	2	12
two	4,61	8,52	3	8	2	12	4,8	6,9	2	8	2	12
three	7,17	20,14	3	5	2	12	5,6	14	3	4	2	12
four	10,71	7,73	3	8	4	14	10,3	7,5	3	4	2	12
five	12,22	22,97	4	7	2	12	10,7	12,85	4	6	4	12
six	8,87	6,88	3	6	2	12	8,6	9,34	4	9	4	12
seven	5,10	17,55	5	1	2	12	5,7	11,7	5	8	1	12
nine	9,24	12,12	5	4	2	18	5,3	9,45	4	7	1	12
twenty	4,45	2,85	4	7	4	12	3,3	2,9	6	8	4	12
thirty	16,52	9,55	1	4	4	14	18,7	4,16	4	10	2	12
forty	13,01	12,40	1	4	4	12	14,1	2,8	4	5	1	12
fifty	17,59	11,57	3	7	4	14	15,4	4,6	5	10	4	12
sixty	10,75	4,72	5	7	4	14	6,2	3,3	6	4	1	12
seventy	7,95	4,41	4	2	2	12	9	4	7	10	1	12
eighty	4,10	4,56	3	7	2	12	2,7	3,42	5	10	1	12
ninity	4,61	3,42	4	5	2	12	3,43	3,1	6	10	1	12
MAE	9,7 ms						7,7 ms					

formar uma rede MHMM para cada unidade acústica. Em geral, estes diferentes modelos produzem resultados de segmentação com valores próximos e uma simples análise estatística pode ser usada 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 3.8.

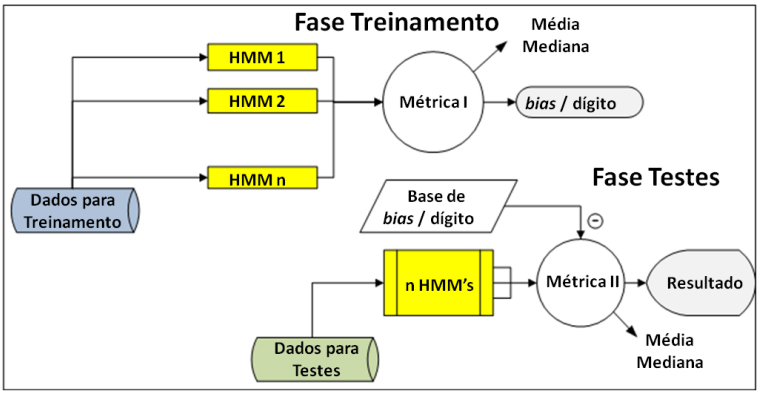


Figura 3.8: Estrutura desenvolvida para o MHMM apresentando as aplicações das métricas nas duas etapas. No treinamento, para cada fone é calculado o viés que será subtraído do valor estimado final para cada limiar.

Neste trabalho, foram consideradas duas técnicas de seleção da estimativa final

ainda para o sistema sem refinamento: por média ou por mediana das 27 estimativas dos HMM individuais.

O procedimento utilizado para refinamento das estimativas obtidas para cada dígito segue conforme algoritmo:

1. Fase Treinamento:
    - treinam-se os 27 HMMs com a base de treinamento;
    - calcula-se o $bias_{TR}$ gerado por cada um dos 27 modelos durante a fase de treinamento. Para isso, utiliza-se o alinhamento forçado de Viterbi com a base de treinamento (segmentada manualmente) e, obtém-se 27 estimativas de limiares da esquerda e 27 para a direita;
    - aplica-se a métrica média ou a mediana sobre estas 27 estimativas para cada lado do dígito, obtendo-se o $\overline{bias_{TR}}$ específico por dígito;
2. Fase Teste - Estimativa inicial:
    - estimam-se os limiares dos dígitos da base de teste, lado esquerdo e direito, aplicando o alinhamento forçado de Viterbi nos 27 modelos HMMs treinados;
3. Fase Teste - Refinamento:
    - retira-se o $\overline{bias_{TR}}$ dos 27 modelos;
    - aplica-se a métrica média ou a mediana sobre as 27 estimativas por limiar/dígito;

Os resultados dos testes para este tipo de sistema são vistos nas colunas 2 e 3 das tabelas 3.3 e 3.4 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 [25].

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 são incluídos nas colunas correspondentes das tabelas 3.4 e 3.3 para as marcações à esquerda e à direita, respectivamente, das unidades selecionadas da base YOHO. 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 no cálculo (média ou mediana) entre todos os HMMs do MHMM.

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 3.3. Para o recorte à direita, constata-se, a partir da tabela 3.4, uma melhora, mas não tão significativa, provavelmente devido ao fato da boa segmentação inicialmente obtida.

Tabela 3.3: Resultados em milissegundos dos erros de segmentação MAE à esquerda das unidades acústicas. Nas colunas referentes a "Sem Refinamento" a média e a mediana são calculadas sobre os resultados dos 27 HMMs/dígito e nas colunas "Com Refinamento" aplica-se o procedimento de refinamento, visto anteriormente

Unidades	**Sem Refinamento**		**Com Refinamento**			
	Média	Mediana	Média/ Média	Média/ Mediana	Mediana/ Média	Mediana/ Mediana
one	25,38	25,14	13,58	13,39	13,61	13,56
two	13,48	10,69	5,90	4,38	4,95	3,83
three	19,80	15,13	7,56	8,66	7,30	7,92
four	25,61	23,73	10,62	11,10	10,78	10,82
five	26,55	25,56	12,16	13,36	12,37	13,56
six	19,23	19,27	12,67	12,69	13,35	13,34
seven	9,22	9,61	8,54	8,50	7,58	7,64
nine	10,78	10,22	13,16	13,14	10,22	10,25
twenty	30,25	9,50	56,99	23,39	25,88	6,22
thirty	46,28	41,77	70,09	71,12	29,24	27,75
forty	31,44	30,39	61,94	63,01	21,42	22,09
fifty	45,45	44,34	91,74	97,43	24,86	26,62
sixty	16,05	16,31	70,44	79,16	14,59	14,59
seventy	23,74	23,25	22,19	21,79	21,52	21,20
eighty	9,60	7,30	7,54	5,47	7,64	5,44
ninety	8,67	7,79	6,84	6,81	3,94	4,31
Média Global	22,60	20,00	29,50	28,34	14,33	13,07

Os valores da tabela 3.4 e 3.3 apresentam uma situação bastante confortável para um sistema segmentador automático, visto que a medida utilizada como parâmetro que classifica o quão bom é o segmentador é uma taxa MAE abaixo de 20 ms. Resultados indicam que o erro absoluto médio pode ser reduzido pela etapa de refinamento pela mediana utilizando uma seleção também por mediana. A segmentação do recorte esquerdo apresentou-se, em relação ao recorte direito, com uma variação significativa do erro MAE medido. Verificou-se que alguns HMMs deste conjunto geravam valores *outliers* provocando deslocamento da medida final, situação esta

Tabela 3.4: Resultados em milissegundos dos erros de segmentação MAE à direita das unidades acústicas. Nas colunas referentes a "Sem Refinamento" a média e a mediana são calculadas sobre os resultados dos 27 HMMs/dígito e nas colunas "Com Refinamento" aplica-se o procedimento de refinamento, visto anteriormente

Unidades	**Sem Refinamento**		**Com Refinamento**			
	Média	Mediana	Média/ Média	Média/ Mediana	Mediana/ Média	Mediana/ Mediana
one	23,79	20,32	22,89	21,46	22,32	19,92
two	16,97	16,13	14,49	15,17	14,35	15,17
three	35,24	30,53	31,46	28,64	30,91	27,46
four	17,58	16,46	18,92	18,74	17,69	17,47
five	35,11	33,14	33,20	33,74	31,88	32,87
six	26,02	21,31	22,55	20,52	22,51	20,38
seven	28,17	27,28	26,38	26,87	25,61	27,25
nine	26,52	28,00	24,97	28,12	25,02	27,94
twenty	6,70	5,88	4,91	4,96	4,87	4,92
thirty	10,81	10,11	8,70	9,34	8,65	8,92
forty	5,67	4,48	4,00	3,63	4,00	3,90
fifty	9,84	9,48	9,16	9,02	9,15	8,97
sixty	8,55	7,38	7,01	6,75	7,08	6,80
seventy	9,21	9,06	7,91	7,96	7,92	8,05
eighty	9,68	10,13	9,09	9,32	8,94	9,18
ninety	7,77	6,91	6,71	7,14	5,56	5,51
Média Global	17,35	16,04	15,77	15,71	15,40	15,30

verificada com os números *thirty, forty, fifty e sixty.*

3.4 Refinamento das Fronteiras dos Fones por Regras Fonéticas

Em trabalhos recentes, tem sido proposto o método de refinamento por meio de regras fonéticas, para ajustes do limiar estimado dos fones do sinal da fala, na etapa de pós-processamento. Dentre os propostos, podemos citar os de Juneja [28], [29], Hosom [23], Toledano [20] e Selmini [1], este último para o português brasileiro. Nestas técnicas de refinamento da fronteira estimada, ou região de transição, têm sido utilizados algoritmos como o DTW (do inglês *dynamic time warping*), as redes neurais ou HMM para fornecer uma estimativa inicial.

No caso do HMM, após a etapa de treinamento dos modelos de cada fone, aplica-se o algoritmo de alinhamento forçado de Viterbi que gerará estimativas dos limiares

de início e fim para cada fone. Estes limiares, provavelmente, estarão próximos aos obtidos por um foneticista, visto que, o algoritmo de Viterbi não estima as fronteiras através da análise por descontinuidades acústicas, mas procura por sequências mais prováveis de fones a partir da locução e da transcrição fonética (segmentação explícita). Com a definição da sequência mais provável por Viterbi, ou seja, da sequência de estimativas grosseiras entre fones, aplica-se o processo de refinamento por meio de regras fonéticas baseadas nas características acústicas dos fones adjacentes à fronteira.

As pesquisas realizadas por Selmini [1] foram baseadas nos estudos iniciais sobre fones do português do Brasil realizada por Ynoguti [30], o qual seguiu a classificação fonética acústica dos estudos de Callou [4], identificando-se 35 fones e 10 classes fonéticas, como apresentado na tabela 3.5.

Tabela 3.5: Divisão dos fones e das classes fonéticas de acordo com [4].

Classificação dos Fones do Português Brasil		
Consoantes	Fricativas	[f], [s], [x], [z], [v], [j]
	Plosivas	[p], [t], [k], [b], [d], [g]
	Africadas	[D], [T]
	Laterais	[l], [L]
	Vibrantes	[r], [rr], [R]
	Nasais	[m], [n], [N]
Vogais	Anteriores	[e], [E], [i], [y]
	Média	[a]
	Posteriores	[o], [O], [u]
	Nasais	[an], [en], [in], [on], [un]

A classificação dos sons baseia-se essencialmente no modo como os sons são produzidos, ou seja, em função da sua articulação. De modo geral, os sons, de acordo com a sua excitação, podem ser de dois tipos: sonoro e surdo. A geração do som vem a partir do ar sendo expelido pelos pulmões, passando pela laringe e pregas vocais, encontrando duas situações: a primeira, as pregas retesadas vibram, quando se conservam próximas, enquanto se expulsa o ar, produzindo desta forma após o trato vocal, um som sonoro como o fone [z] da palavra ca*S*a; a segunda, quando as pregas estão relaxadas, elas não vibram com a passagem do ar, gerando um som chamado de surdo, como o fone [s] da palavra *S*apo. Na sequência, é apresentada a classificação das vogais e consoantes, de acordo com os grupos listados na tabela 3.5.

Para as vogais consideram-se quatro classes: anterior, média e posterior, definidas em função da zona de articulação que está relacionada com a região da boca onde as vogais são geradas, e nasal, gerada pela corrente de ar vibrante que passa pelas cavidades bucal e nasal [31], [32], [33]:

- **anterior ou palatal**: a vogal é articulada com a língua elevada em direção ao palato duro, próximo ao dentes. Fones [e] - d*e*do, [E] - p*e*le, [i] - bot*i*na;
- **média ou central**: a vogal é articulada com a língua abaixada, quase em repouso. Fone [a] p*á*, *á*tomo;
- **posterior ou velar**: a vogal é articulada quando a língua se dirige ao palato mole. Fones: [o] - b*o*lo, [O] - p*ó* e [u] - l*u*a;
- **nasal**: a vogal é formada pela corrente de ar vibrante que passa pelas cavidades bucal e nasal, compondo os fones [an] - maç*ã*, [en] - s*e*nta, [in] - s*i*nto, [on] - s*o*mbra e [un] - *u*m.

Para as consoantes, classificadas em seis classes, os fones são produzidos com algum tipo de obstrução no trato vocal de forma que há impedimento total ou parcial da passagem de ar [31], [32], [33]. As classes formadas pelos tipos plosiva (ou oclusiva) e constritiva são classificadas em função do modo de articulação. A classe plosiva consiste naqueles fones gerados quando o ar que vem dos pulmões encontra os lábios fechados, abrindo repentinamente, enquanto que a classe constritiva inclui os sons que têm um bloqueio parcial do ar, ou seja, um estreitamento da passagem do ar entre os lábios. Assim como as vogais, as consoantes também possuem uma classe nasal, e aparece ainda a classe das africadas, como detalhado a seguir:

- **Classe plosiva**: Inclui os fones: [p] - *p*ai, [t] - *t*io, [b] - *b*ola, [k] - *c*asa, [d] - *d*e*d*o e [g] *g*ude;
- **Classe constritiva fricativa**: Ocorre fricção do ar através de uma fenda no meio da boca. Fones: [f] - *f*aca, [v] - *v*aca, [s] - *s*eda, [z] - *z*oo, [x] - *x*is e [j] - *g*iz;
- **Classe constritiva lateral**: Quando a língua permite a passagem do ar somente pelas laterais, têm-se dois fones consonantais laterais. Fones: [l] - *l*ua e [L] - ca*lh*a;
- **Classe constritiva vibrante**: Quando o som é produzido através da vibração da língua ou do véu palatal. Fones: [r] - ce*r*a, [rr] - ce*rr*ado e [R] - ca*r*ta;
- **Classe nasal**: Quando o ar se desloca também pelo nariz, por encontrar obstáculo pela cavidade bucal. Inclui os fones: [n] - *n*évoa, [m] - *m*ontanha e [N] - i*nh*ame;
- **Classe africada**: Que corresponde primeiro ao bloqueamento completo do trato vocal, seguido de uma pequena abertura que produz um ruído de fricção, neste caso, combinando o som de uma oclusiva com o de uma fricativa. Inclui os fones: [D] - *d*ia e [T] - *t*ia.

Para cada classe fonética, define-se uma métrica ou um conjunto de métricas que identifica, dentro do intervalo de refinamento, o limiar de separação entre pares de fones e, possivelmente, entre fones de diferentes classes [1].

Classes das Vogais

Em particular, as métricas definidas para a classe das vogais baseiam-se nas ocorrências de zonas de ressonância do trato vocal durante o processo de produção do sinal vozeado. Neste processo de produção dos sons das vogais, ocorrem constrições no trato vocal em função do posicionamento da língua, provocando alterações na sua acústica e, portanto, nos valores das frequências ressonantes. Estas alterações podem ser verificadas na figura 3.9, que apresenta a ocorrência das vogais em relação à variação do primeiro *F1* e segundo *F2* formantes.

Os formantes foram determinados a partir dos coeficeintes LPC (algoritmo de Levinson-Durbin [2]), aplicado DFT com 1024 pontos. A janela de análise de 10 ms.

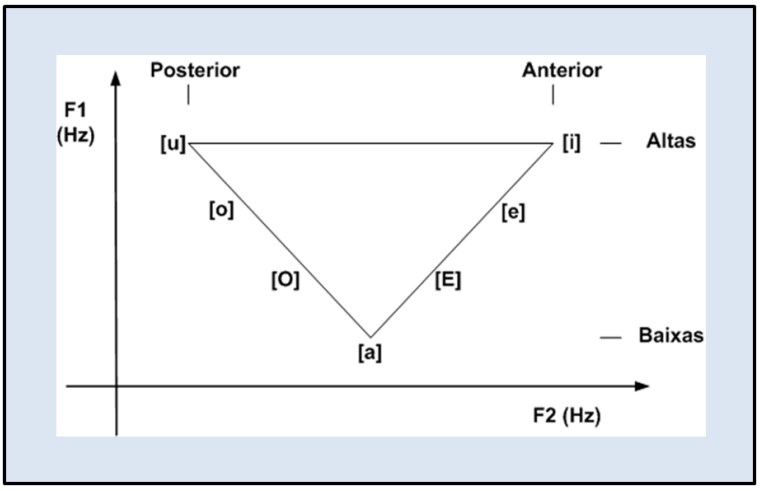


Figura 3.9: Comportamento das formantes $F1$ e $F2$ para as vogais quanto à altura (alta ou baixa) e posição (anterior ou posterior) da língua no trato vocal.

Na figura 3.10 é ilustrada a variação dos quatro primeiros formantes $F1$, $F2$, $F3$ e $F4$ para os ditongos das palavras "lixeiro" e "lixou", onde pode ser verificado que, conforme Rabiner [2], a distribuição das energias das vogais acompanham a distribuição das frequências formantes. De fato, para as vogais posteriores (fones [O], [o] e [u]), a ocorrência de ressonâncias nas baixas frequências concentra a energia do sinal também nas baixas frequências, enquanto que para as vogais anteriores (fones [E], [e] e [i]), as ressonâncias em frequências elevadas provocam uma distribuição da energia ao longo de uma banda mais larga em frequências.

As métricas para refinamento [1] desta classe foram divididas nas seguintes transições entre fones, como descrito a seguir e sumarizado na tabela 3.6:

- classe vogal com outras classes: uso das regras definidas para a classe adjacente.

- classe vogal anterior ou posterior com vogal média: entre as três classes foram utilizados valores empíricos e baseados nas faixas de variações dos formantes $F1$ e $F2$, conforme figura 3.9.

- classe vogal anterior com vogal posterior: uso de duas características, frequência $F2$ média de 1490 Hz e o uso de perfil de energia, sugerida por Araújo [34], que representa a frequência abaixo da qual está contida uma determinada porcentagem β da energia total calculada sobre o espectro de frequência, isto é,

$$F_\beta = k_\beta \frac{f_a}{M_{fft}}, \tag{3.4}$$

  onde f_a é a frequência de amostragem, k_β é o índice da FFT do sinal $x(n)$ e M_{fft} é o número de pontos utilizados no cálculo desta FFT.

- transição entre vogais de mesmas classe: uso do critério de informação Bayesiana (BIC, do inglês *Bayesian information criterion*), que estima pontos de variação acústica entre dois segmentos adjacentes analisados. De fato, o BIC é caracterizado por [1]:

$$BIC(i) = R(i) - \lambda P_0 \tag{3.5}$$

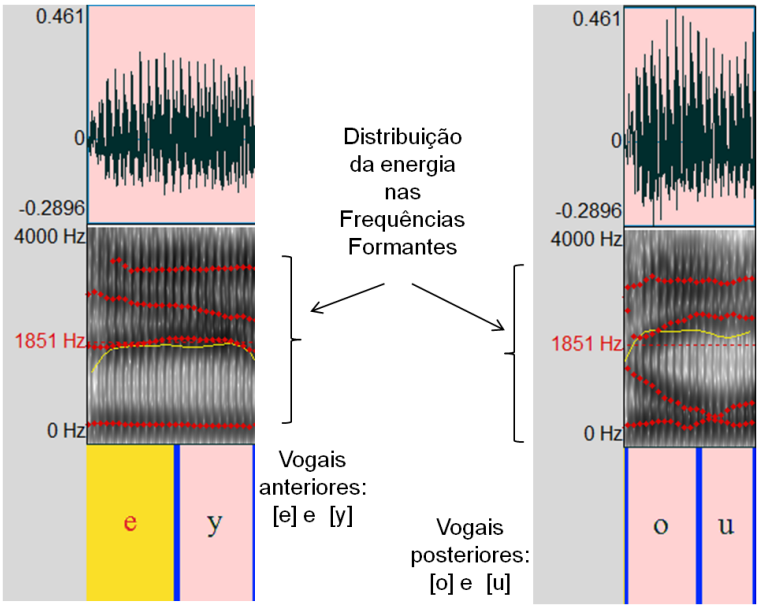


Figura 3.10: Variação dos formantes nos ditongos “ei” e “ou”. Na figura percebe-se aumento da energia nas frequências mais altas do ditongo “ei” em relação ao ditongo “ou”, enquanto há uma redução da energia nas frequências mais baixas.

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

$$R(i) = n \log |\Sigma_0| - m \log |\Sigma_1| - (n - m) \log |\Sigma_2|, \tag{3.6}$$

onde Σ_0 a matriz covariança do segmento completo, Σ_1 e Σ_2 são as matrizes covariança dos dois segmentos adjacentes. O parâmetro P_0 é o fator de penalização para a complexidade do modelo, calculado pela expressão

$$P_0 = \frac{1}{2}\left(p + \frac{1}{2}p(p+1)\right) \log(n), \tag{3.7}$$

e o parâmetro λ representa o peso para o fator de penalização, em geral definido como 1. A fronteira de segmentação será quando o valor do BIC é máximo para todos os valores de i, indicando a ocorrência de mudança acústica entre os segmentos analisados.

Tabela 3.6: Regras fonéticas para segmentação da classe das vogais [1].

Regras
1 - Transição entre vogal com outras classes fonéticas
Regra de decisão: Usar métrica definida para a classe seguinte
2 - Transição entre vogal anterior ou posterior com vogal média
Regra de decisão: F1 / F2 Anterior: $< 450Hz$ / $> 1845Hz$ Posterior: $< 450Hz$ / $< 1135Hz$ Média: $> 450Hz$ / $1135Hz < X < 1845Hz$
3 - Transição entre vogal anterior e vogal posterior
Regra de decisão: F2: Vogal Posterior $< (F2 = 1490Hz) < VogalAnterior$ Perfil de Energia: $Beta = 75\%$ na frequência $F = 1550Hz$
4 - Transição entre vogais de mesma classe fonética
Regra de decisão: Critério de informação bayesiana (BIC)

Classe das Consoantes Plosivas

As consoantes plosivas possuem um longo período de oclusão seguido pela abertura dos lábios quando ocorre uma "explosão" na liberação do ar. De modo geral, esta classe pode ser dividida em plosiva surda, quando há um longo período de silêncio (por exemplo, [p]) ou plosiva sonora, quando for um longo período com baixa energia espectral (por exemplo, [g]). Nos testes realizados em [1], o limiar de separação entre o silêncio e a plosiva surda ocorrerá quando a energia total alcançar o limiar de -60 dB, já para as plosivas sonoras o corte deverá anteceder o início do som vozeado do fone plosivo, definindo o limiar de -70 dB. De outras classes para a classe plosiva é utilizada a variação abrupta da energia espectral, detectada através de um pico na derivada da energia no intervalo de refinamento. E, da classe plosiva para outras classes, verificou-se no estudo da variação da derivada da energia a ocorrência de dois picos, o primeiro quando ocorre na "explosão" do fone e o segundo, menor que o primeiro, quando na transição para o fone adjacente. Estas regras são sumarizadas na tabela 3.7.

Classes das Consoantes Laterais e Vibrantes

Para a transição destas classes para as demais classes é utilizado o cálculo da derivada da energia espectral com três janelas adjacentes, isto é a derivada da soma das energias espectrais das bandas dos formantes: 0 a 500Hz ($F1$), 500Hz a 1500Hz ($F2$) e 1500Hz a 2400Hz ($F3$); mais a energia total. E a maior variação de energia obtida no intervalo é identificada como a transição de uma classe para outra. Esta regra também é válida quando for a transição das demais classes para as classes das consoantes laterais e vibrantes, como apresentado na tabela 3.8.

Classe das Consoantes Nasais

Da mesma forma como nas classes das consoantes laterais e vibrantes, na transição entre a classe nasal para as demais classes (e vice-versa) utiliza-se o cálculo da derivada da energia espectral com 3 janelas adjacentes, DFT de 1024 pontos, no caso, a soma das energias espectrais em duas bandas nas frequências 0-358 Hz e 358-5378 Hz [29], e a maior variação de energia obtida no intervalo é identificado como a transição de uma classe para outra, como sumarizado na tabela 3.9.

Classe das Consoantes Fricativas

As consoantes fricativas são geradas pela passagem forçada do ar por alguma constrição criada pelos articuladores constituintes do trato vocal. Domingos Cegalla [35] define que a turbulência é gerada quando o ar sai roçando ruidosamente as paredes

Tabela 3.7: Regras fonéticas para segmentação da classe das consoantes plosivas [1].

Regras para Plosivas [p][t][b][k][d][g]
1 - Silêncio - Plosivas
Regra de decisão: 1 - Surda - limiar acima de -60 dB 2 - Sonora - limiar acima de -70 dB
2 - Outras classes - Plosivas
Regra de decisão: 1 - Detecção, no intervalo de refinamento, do maior pico na derivada da energia
3 - Plosivas - outras classes
Regra de decisão: 1 - Detecção, no intervalo de refinamento, do segundo maior pico na derivada da energia Bandas passantes (BW, do inglês Bandwidth) para cálculo da energia espectral: 1ª BW: 0 - F3 2ª BW: F3 - $\frac{f_a}{2}$ f_a - frequência de amostragem

da boca estreitada, e será sonora quando o ar põe as cordas vocais em vibração, caso contrário será surda.

Assim como as plosivas, a classe das consoantes fricativas também é dividida em fricativas sonoras, representada pelos fones [v], [z] e [j], e fricativas surdas, pelos fones [x], [s] e [f]. Para exemplificar, as figuras 3.11 e 3.12 mostram diferenças entre fricativas surda e sonora com relação aos formantes do sinal. Na figura 3.11 é mostrada a forma de onda do sinal "tas", onde se observa que o fone [s], por ser gerado sem a vibração das pregas vocais, não possui bem definidas as frequências dos formantes. No caso da figura 3.12 existe uma fricativa, fone [j], entre duas vogais "a"na trecho "a jaca"em que as frequências formantes variam com uma certa continuidade, característica do som vozeado.

Para a classe fricativa, o limiar é obtido a partir do uso de duas funções: a taxa de cruzamento de zeros e centro de gravidade espectral (CGE). Na análise experimental

Tabela 3.8: Regras fonéticas para segmentação da classe das consoantes laterais e vibrantes [1].

Regras para Consoantes Laterais [I][L] e Vibrantes [r][rr][R]	
Regra de decisão: Somatório das variações das energias das bandas a cada instante de tempo, a fronteira será o pico máximo da variação. Cálculo é executado considerando a derivada com 3 janelas para cada lado da janela em questão. Fronteira de separação: ponto de maior pico do intervalo analisado.	
Energia total da janela;	
1ª BW - 0 a 500Hz	1º formante;
2ª BW - 500Hz a 1500Hz	2ª formante;
3ª BW - 1500Hz a 2400Hz	3º formante;
4ª BW - 2400Hz a $\frac{f_a}{2}$	

Tabela 3.9: Regras fonéticas para segmentação da classe das consoantes nasais [1].

Regras para Consoantes Nasais [n][m][N]	
Regra de decisão: Somatório das variações das energias das bandas a cada instante de tempo, e a fronteira será o pico máximo da variação. Cálculo é executado considerando a derivada com 3 janelas para cada lado da janela em questão.	
Concentração da energia nas baixas frequências - característica das consoantes nasais;	1ª BW - 0 a 358Hz
Concentração da energia nas altas frequências - característica das vogais;	2ª BW - 358Hz a 5378Hz
Fronteira de separação: ponto de maior pico do intervalo analisado.	

da referência [1], os valores utilizados para a separação entre classes foram de 0,52 para taxa de cruzamento de zeros quando fricativa surda e 0,28 para fricativa sonora, e para o CGE obteve-se o limiar de 2500 Hz. Dentro da região de refinamento, o ponto de segmentação será onde os valores dos limiares estabelecidos, para os dois

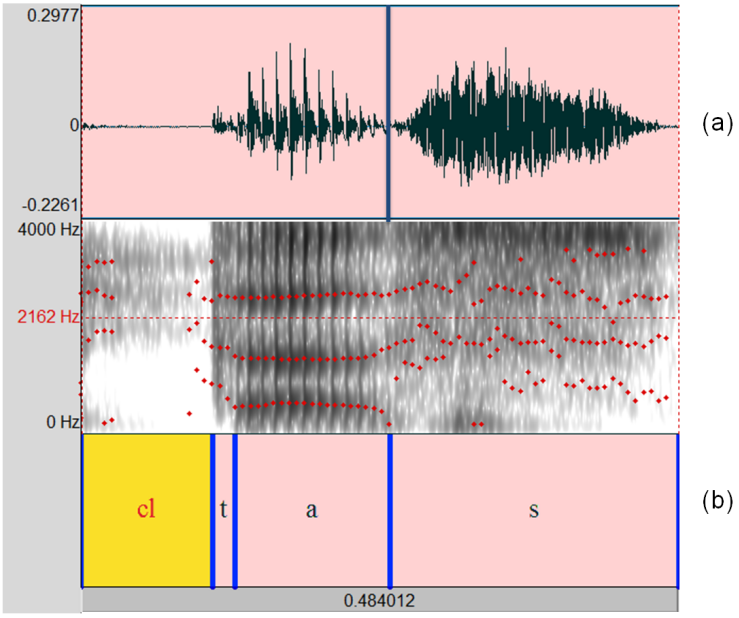


Figura 3.11: Trecho "tas"retirado da palavra "colheitas". (a) Forma de onda. (b) Espectrograma, onde as linhas pontilhadas representam os formantes $F1$, $F2$ e $F3$ (respectivamente, de baixo para cima).

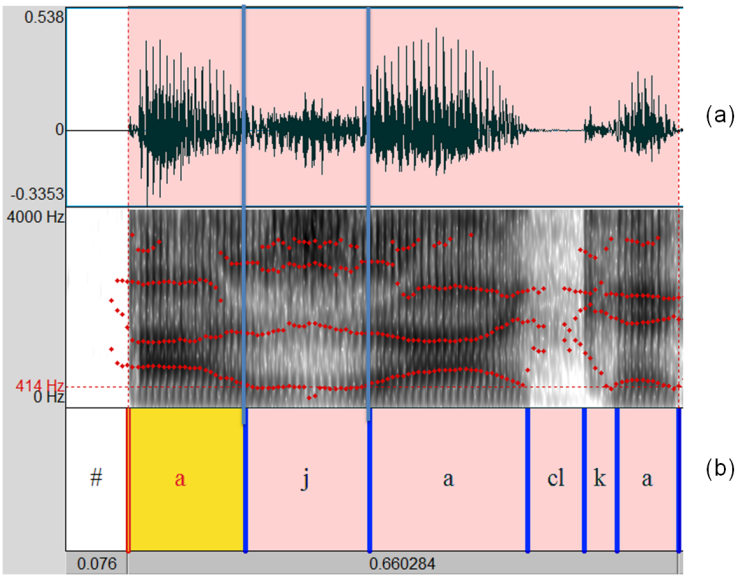


Figura 3.12: Palavra "jaca". (a) Forma de onda. (b) Espectrograma, onde as linhas pontilhadas representam os formantes $F1$, $F2$, $F3$ e $F4$ (respectivamente, de baixo para cima). Neste caso, Por ser um fricativo sonoro, os formantes são bem definidos, ou seja, existe a vibração das cordas vocais ao pronunciá-lo.

parâmetros em análise, estão abaixo dos valores definidos, conforme resumido na tabela 3.10.

Tabela 3.10: Regras fonéticas para segmentação da classe das consoantes fricativas [1].

Regras para Fricativas [f][v][s][z][x][j]
Regra de decisão: 1 - Taxa de Cruzamento por zeros a. Fricativa surda - 0.52 b. Fricativa Sonora - 0.28 2 - Centro de Gravidade Espectral a. Limiares: CGE - 2500 Hz.
Fronteira de separação: ponto de cruzamento entre os limiares estabelecidos dos dois parâmetros.

Classe das Consoantes Africadas

A transição de outras classes para a classe das consoantes africadas ocorre no momento de queda abrupta da energia espectral. Isto porque as consoantes africadas são caracterizadas por um longo período de baixa energia espectral nas baixas frequências, justamente por ser o período de constrição [1]. Quanto a transição para o fone [i], identificou-se que somente a combinação das características acústicas taxa de cruzamento de zeros e centro de gravidade espectral apresentaram-se suficientes para a obtenção do limiar, como sumarizado na tabela 3.11.

3.4.1 Resultados Experimentais

Na etapa de refinamento utilizando regras fonéticas foram obtidas taxas de reconhecimento para o português brasileiro de 95,55% para até 20 ms, utilizando uma base de fala pronunciada por um locutor paulista do sexo masculino [1]. Nesse trabalho, foram usadas duas bases: a primeira base, para treinamento, com 1026 locuções, e a segunda base, para realização dos testes, com um total de 200 locuções.

Tabela 3.11: Regras fonéticas para segmentação da classe das consoantes africadas [1].

Regras para Consoantes Africadas [D][T]
Outras classes - Africada
Regra de decisão: Derivada 1ª da energia
Africadas - fone [i]
Regra de decisão: Taxa de Cruzamento por zeros - limiar: TZC - 0.4 Centro de Gravidade Espectral - limiar: CGE - 4000 Hz
Fronteira de separação: ponto de cruzamento entre os limiares estabelecidos dos dois parâmetros.

3.5 Conclusão

Neste capítulo, alguns exemplos de técnicas de segmentação de sinais de voz foram apresentadas. Todas as técnicas aqui consideradas utilizam em uma primeira etapa HMMs para a primeira segmentação dos fones, sendo esta considerada como uma segmentação com sequência de limiares grosseiros e, numa segunda etapa, chamada por etapa de refinamento, onde a sequência de limiares são ajustados para o limiar mais próximo do real, idealmente dentro de um intervalo de erro absoluto da ordem de 20 ms.

Nas técnicas de refinamento utilizando múltiplos HMMs[6], combinam-se os resultados de n sistemas HMMs para se gerar um limiar geral de segmentação. Em particular, dos n resultados parciais, pode-se extrair o erro sistemático obtido para cada marca, para se gerar a marca final. No trabalho original apresentando o funcionamento desta técnica, foi utilizada uma base de dados coreana, obtendo uma taxa de acerto de 97,07% (erro absoluto menor do que 20 ms), representando um ganho significativo em relação ao experimento utilizando somente HMMs, onde a taxa de acerto era de 95,06% [6].

A aplicação desta técnica no reconhecimento de marcas entre dígitos também foi apresentada, onde se observou um desempenho MAE, incluindo as marcas à es-

querda e à direita, de 14,2 ms na etapa de treinamento, valor este abaixo dos 20 ms desejados. Neste experimento foi utilizada a base YOHO para a determinação das marcas dos dígitos pronunciados por um locutor, falado em inglês. Considerando um conjunto de 27 HMMs com diferentes configurações, foi possível extrair as média e mediana dos erros de estimativa para uma posterior correção de cada viés por unidade de segmentação. Na etapa de testes, a etapa de refinamento foi implementada removendo-se o viés previamente detectado, o que ocasionou uma melhora no erro absoluto médio do limiar estimado para o correto de 22,60 ms para 13,07 ms após o refinamento.

Por fim, uma outra técnica aplicada na etapa de refinamento foi apresentada. Esta baseia-se nas regras fonéticas para o português brasileiro, observando o tipo de transição entre fones do lado esquerdo com o lado direito da marca estimada por um segmentador HMM. Na primeira fase, as marcas são obtidas através do alinhamento forçado de Viterbi, que gera uma sequência de marcas de fones para a frase-teste apresentada. Define-se, então, uma região de refinamento, que compreende as marcas imediatamente anterior e posterior à marca que está sendo analisada, extrai desta região as características acústicos-fonéticas já pré-definidas conforme a dupla de sons ou classes fonéticas em análise, características estas que podem ser: energia, correlação entre os logaritmos das energias dos lados esquerdo e direito da marca que está sendo refinada, taxa de cruzamentos por zero (TxZ) ou derivada da energia. Nesse trabalho apresentado por Selmini, foram definidas 10 classes fonéticas entre os sons das consoantes e vogais, e para cada combinação inter ou intra-classes foram estudados os parâmetros acústicos que melhor definem a marca final, levando a uma taxa de acertos (erro absoluto abaixo de 20 ms) das marcas de 95,55%.

No capítulo a seguir as técnicas apresentadas no presente capítulo serão norteadoras para a fusão de técnicas com o objetivo de alcançar um sistema que realize a segmentação do português brasileiro, aplicando técnicas de refinamento por múltiplos HMMs e por meio de regras fonéticas de forma combinada.

Capítulo 4

Sistema Proposto de Segmentação

4.1 Introdução

No presente capítulo, será apresentado o sistema desenvolvido para a segmentação de um banco de fala do português brasileiro combinando as duas propostas, no caso as das referências [6] e [1], apresentadas no capítulo 3. Esses trabalhos anteriores apresentam dois métodos distintos de segmentação, sendo o primeiro baseado no conceito de múltiplos HMMs e a eliminação do viés correspondente e o segundo baseado em regras acústico-fonéticas, aplicadas na região de refinamento, lado esquerdo e direito da fronteira em análise. As técnicas avaliadas e implementadas permitirão desenvolver métricas para validação do desempenho e na solução de problemas verificados. Dentre esses problemas, inclui-se até mesmo a primeira etapa uma segmentação via HMM, que deve estimar um limiar que permitirá, na fase seguinte, a obtenção da fronteira a partir de uma região de refinamento. A ideia central é combinar os aspectos positivos de cada método anteriormente proposto na literatura para gerar um método mais robusto e eficiente de segmentação automática de sinais de voz.

Desta forma, este capítulo está organizado como segue: na seção 4.2, apresentamos a base de dados utilizada ao longo deste trabalho, a qual foi a mesma utilizada em [1], com alguns ajustes na marcação; a seção 4.3 descreve o conceito geral do sistema proposto combinando as técnicas de Park e de Selmini, enquanto que as seções 4.4 e 4.5 descrevem em detalhes as duas etapas do novo algoritmo: a de segmentação inicial e a de refinamento; a seção 4.6 conclui o capítulo enfatizando os principais aspectos de seu conteúdo.

4.2 Base de Fala

Para o desenvolvimento deste trabalho optou-se pela utilização de uma base do português brasileiro já existente, e, como este trabalho utilizará como base a técnica de refinamento apresentada por Selmini [1], em seu doutorado, a sua escolha foi natural. A base de fala, gravada em ambiente de laboratório pelo próprio autor da referência [1], está dividida em duas partes: treinamento e testes. A base usada para o treinamento do sistema possui 1026 locuções e a utilizada para testes possui 200 locuções, totalizando 1226 locuçoes amostradas à taxa de 22050 Hz e quantizadas em 16 bits/amostra, com o locutor masculino apresentando um sotaque do interior de São Paulo. A transcrição completa destas 1226 locuções pode ser encontrada no apêndice A de [1]. A parte de teste foi subdividida em duas partes: 50 locuções foram usadas na geração do HMM inicial (semente) e as demais 150 para teste efetivo de desempenho do sistema.

As locuções da base de testes foram segmentadas manualmente, pelo próprio autor da base, gerando a base de referência para validação do sistema de segmentação automática. Para esta segmentação manual, utilizou-se o *software* livre Praat (www.praat.org), que permite a análise do som nos domínios do tempo e da frequência de forma simultânea. Os parâmetros observados para a determinação das marcas foram: forma de onda do sinal, o espectrograma, curvas de energia e trajetória de formantes [1].

Todas as frases foram transcritas utilizando uma notação própria para a representação dos fonemas do português brasileiro. A tabela 4.1 mostra a relação dos símbolos com os fones e a suas classes fonéticas utilizadas neste processo.

No total são 35 símbolos distribuídos entre as classes fonéticas, 2 símbolos que representam o tempo de início da consoante até a explosão do ar (barra de voz): cl - *unvoiced closure* e vcl - *voiced closure*; e mais 2 símbolos que representam o silêncio e a pausa entre palavras. A distinção de *cl* (surda) e *vcl* (sonora), mostrada na figura 4.1, pode ser claramente percebida na pronúncia de pares de consoantes da classe plosiva, [p] [b], [t] [d], [k] [g]. E, em função da sonoridade, os fones [p,t,k] são classificados como não vozeados (ou surdos), enquanto que os fones [b,d,g] são classificados como vozeados (ou sonoros), por terem vibração significativa das cordas vocais durante a fase de acúmulo de ar [36].

4.2.1 Ajustes na Marcação da Base de Fala

Como contribuição ao trabalho iniciado e desenvolvido por Selmini [1] na construção de um banco de fala, foi contratado um foneticista profissional para avaliar e, caso necessário, corrigir a transcrição fonética das 200 locuções da base de teste. Nesta nova marcação, foi mantida a mesma notação originalmente adotada.

Tabela 4.1: Exemplos da notação utilizada na transcrição da base de fala e as respectivas classes fonéticas [1].

Notação utilizada na transcrição da base de fala			
Classe Fonética	**Simbologia**	**Exemplo**	**Transcrição**
Classe Vogal Média	a	bola	vcl b O l *a*
Classe Vogal Anterior	e	dedo	vcl d *e* vcl d u
	E	pele	cl p *E* l y
	i	botina	vcl b o cl t *i* n a
	y	pele	cl p E l y
Classe Vogal Posterior	o	bolo	vcl b *o* l u
	O	pó	cl p *O*
	u	lua	l *u* a
Classe Nasal	an	maçã	m a s *an*
	en	senta	s *en* cl t a
	in	sinto	s *in* cl t u
	on	sombra	s *on* vcl b r a
	un	um	*un*
Classe Plosiva	p	pai	cl *p* a i
	t	tempo	cl *t* en cl p u
	b	bola	vcl *b* O l a
	k	casa	cl *k* a z a
	d	dedo	vcl *d* e vcl *d* u
	g	galeto	vcl *g* a l e cl t u
Classe Fricativa	f	faca	*f* a cl k a
	v	vaca	*v* a cl k a
	s	seda	*s* e vcl d a
	z	zona	*z* o n a
	x	chave	*x* a v y
	j	giz	*j* i s
Classe Laterais	l	lua	*l* u a
	L	calha	cl k a *L* a
Classe Vibrante	r	cera	s e *r* a
	rr	cerrado	s e *rr* a vcl d o
	R	carta	cl k a *R* cl t a
Classe Consoante Nasal	n	névoa	*n* E v o a
	m	montanha	*m* on cl t an N a
	N	inhame	i *N* an m y
Classe Africada	D	dia	vcl *D* i a (djia)
	T	tia	cl *T* i a (tchia)
Período de oclusão para as plosivas surdas			cl
Período de oclusão para as plosivas sonora			vcl
Pausa entre as palavras			sp
Silêncio presente no início e no fim das locuções			#

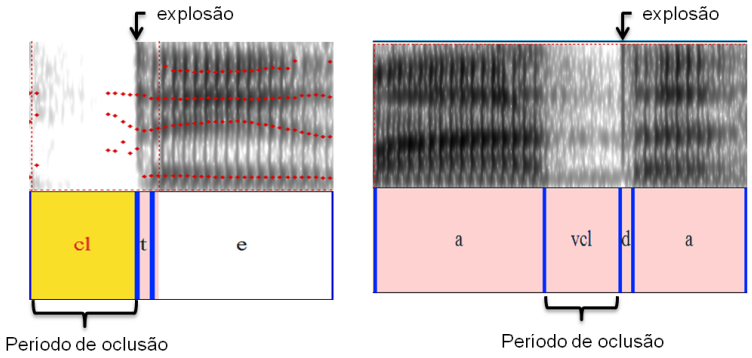


Figura 4.1: Trechos das palavras ***te***mpo e anim***ada***, destacando a geração dos períodos anteriores a explosão do som, observado na pronúncia do par [t] e [d].

Dentre as observações e sugestões elaboradas pelo foneticista, duas serão tratadas e discutidas nesta subseção:

- Quanto à atribuição do som do fone [z] em razão da harmonização vocálica dada a coarticulação do fone em destaque entre sons vozeados ([a][z][a]). A figura 4.2 mostra uma comparação das características acústicas do som [z] e do som [s]. Pode-se verificar que as frequências formantes do som ([a][z][a]) mantêm uma continuidade entre as vogais *a*, percebendo um som vozeado. Esta disposição é similar aos exemplos de me***Z***a, ca***Z***a.

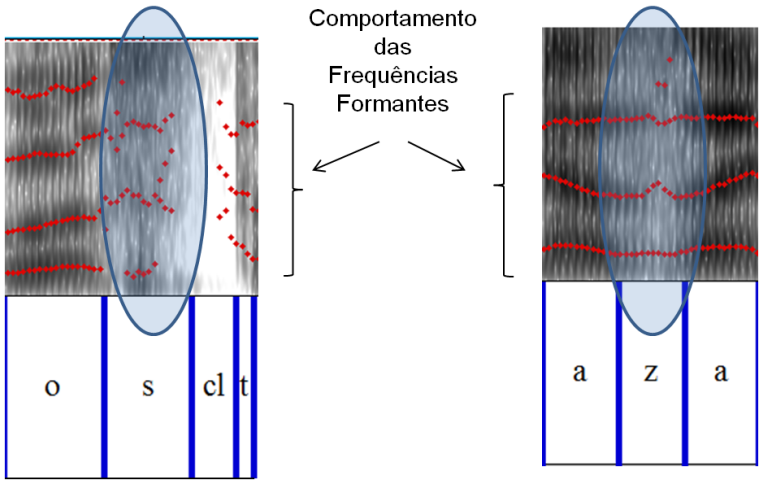


Figura 4.2: Trechos das palavras go***s***tava e música***s***_animadas.

- Quanto ao tratamento das vogais epentéticas. O aparecimento de vogais epentéticas em encontros consonantais CCV (C - consoante; V - vogal), casos como a palavra construção - [k on s cl t r u s an u], que surge um elemento vocálico entre a soltura do [t] e a consoante [r]. Nestes casos, a vogal epentética será admitida como pertencente ao vibrante simples [r]. No trabalho apresen-

tado por Rui Seara [37], considera-se que a transcrição dessa vogal epentética leva a uma seleção de unidades acústicas possivelmente inadequada para um sistema de fala sintética, pois, em sendo sintetizada ela não seria apropriadamente curta para dar naturalidade à voz resultante. A figura 4.3 mostra a percepção do som epentético [u] por meio de análise criteriosa, entretanto, não percebida claramente em uma fala natural. Desta forma, mesmo sendo notada por meio de *software* de análise, foi mantida a desconsideração no processo de segmentação.

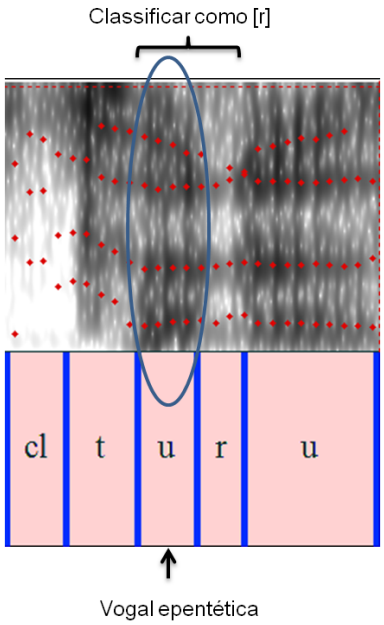


Figura 4.3: Trecho da palavra cons***tru***ção enfatizando a existência, mas de baixa percepção, de vogal epentética.

4.3 Sistema Proposto de Segmentação

A base do sistema que será discutido ao longo deste capítulo é apresentado na figura 4.4. Conforme colocado no capítulo anterior, numa primeira etapa do sistema será aplicada a técnica HMM, que tem se mostrado eficiente na segmentação de unidades acústicas [20], através do alinhamento forçado do algoritmo de Viterbi fornecendo os limites entre as unidades fonéticas. E, numa segunda etapa serão utilizadas diferentes técnicas de refinamento combinando o uso de múltiplos HMMs e regras fonéticas para o português brasileiro.

A discussão neste ponto é avaliar a qualidade da segmentação para um grupo de parâmetros, desde aqueles relacionados com o modelo como número de estados ou gaussianas por misturas, quanto aos relacionados com a extração das características. Por exemplo, em [1], foi usado o tempo de janelas de análise para a extração das

características, com duração de 20 ms e deslocamento a cada 10 ms ou 1 ms para o caso específico da energia; já em [6], é utilizado um tempo de 24 ms de duração de janela e 10 ms entre janelas. No caso de [26], é sugerido um tamanho para duração da janela entre 10 e 20 ms e um passo entre janelas de 5 ms. Já [20] utilizou janelas com duração de 24 ms e deslocamentos de 3 ms. Dessa forma, não existindo um valor apontado na literatura como o ideal, o que se verifica é a existência de conjuntos de parâmetros que melhor se ajustam a uma determinada base de unidades.

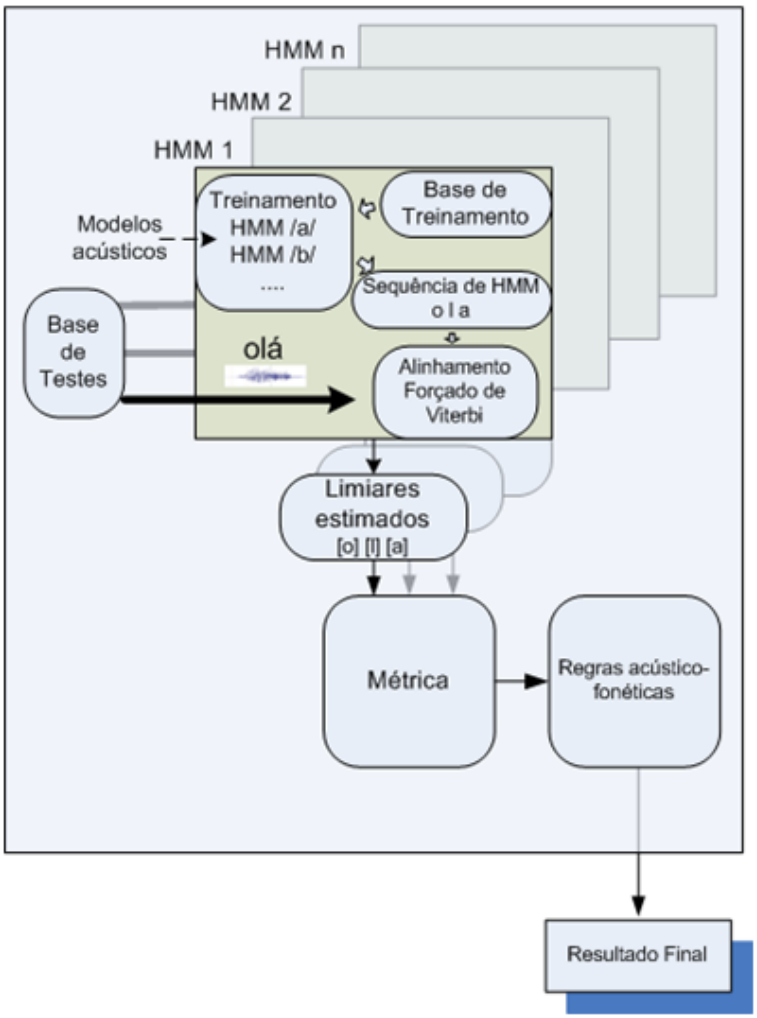


Figura 4.4: Sistema proposto para segmentação automática utilizando múltiplos HMMs e regras fonéticas.

A mesma discussão será direcionada à segunda etapa do sistema, onde regras fonéticas são aplicadas sobre o sinal na região de refinamento, e que, após análise, obtêm a fronteira de separação entre fones adjacentes. Esta análise aborda o modelo de produção dos sons da fala, sua articulação, as relações entre classes fonéticas, entre outros, definindo a parametrização adequada para o processo de refinamento dos limiares dos fonemas.

Em algumas classes fonéticas, como por exemplo, as classes fricativas e vogais,

são utilizados diferentes parâmetros acústicos para identificar o limiar de separação para outras classes ou, entre as mesmas. O objetivo é encontrar um ponto que seja ótimo para todos os parâmetros e que identifique a marca com menor erro. No caso da classe das fricativas para as demais classes, são usados os parâmetros taxa de cruzamento de zeros e centro de gravidade espectral, e a métrica que define a marca de segmentação é o ponto onde os dois parâmetros medidos estão abaixo do limiar estabelecido [1]. Na transição entre as classes vogal anterior com a vogal posterior, os parâmetros definidos são a variação do formante $F2$ e o perfil de energia. Nestes exemplos citados, a métrica estabelecida definiu uma marca final a partir dos parâmetros acústicos que é dependente das características momentâneas do falante, e por isso, possibilita a existência de uma possível região de variação da posição da marca final. Nas seções 4.4 e 4.5, a seguir, serão discutidos e apresentados técnicas com a finalidade de aumentar a robustez do sistema no refinamento.

4.4 Primeira Etapa: Múltiplos HMMs

Por ser a técnica de HMM gerada por funções estatísticas, a sua aplicabilidade no processo da segmentação explícita é reconhecer e delimitar padrões. No caso, os padrões são de segmentos acústicos, quando em sua entrada é apresentado um banco de sinais de fala e as respectivas transcrições fonéticas, e, em particular, quando se utiliza a topologia do modelo Bakis, que permite modelagem temporal do sinal do processo sendo analisado. As fronteiras geradas pelos HMMs servem como ponto de partida para ajustes finos das reais fronteiras das sub-unidades, numa possível etapa posterior ao reconhecimento inicial. Neste caso, o reconhecimento das unidades acústicas é dito explícito, pois o total de segmentos selecionados (isto é, o número de fronteiras) é restrito ao número de símbolos presentes na transcrição fonética da locução previamente conhecido e fornecido ao sistema. Por exemplo, na separação de fones concatenados, onde previamente existe a informação de quais e quantos são os fones, cabe ao sistema identificar somente a localização das fronteiras em número predeterminado.

Uma das principais vantagens em se utilizar a técnica de segmentação de fonemas HMM é a vasta quantidade de conhecimentos e infra-estrutura disponível no campo de reconhecimento da fala aplicável a HMM, apesar das distinções existentes entre os processos de segmentação fonética e reconhecimento da fala [20].

No modelo HMM, uma boa estimativa dos parâmetros iniciais é essencial para uma rápida convergência do algoritmo, principalmente com relação aos valores das densidade de probabilidade de emissão dos símbolos ($b_j(O_t)$) que são bastantes sensíveis a este aspecto, podendo conduzir a resultados não satisfatórios no reconhecimento das unidades fonéticas [2]. Uma forma de implementar o modelo e

obter uma rápida convergência no treinamento é a utilização do algoritmo *segmental k-means* [38], [39], que estima iterativamente o conjunto dos parâmetros iniciais, gerando "sementes" para os HMMs finais, conforme o procedimento:

1. Para iniciar o algoritmo, usam-se modelos iguais para todos os fonemas, com média 0, variância 1 e matriz de transição entre estados **A** uniforme, criando um conjunto de modelos das unidades fonéticas a partir de uma base de 50 locuções;

2. Para a estimação dos parâmetros, aplica-se o algoritmo de Viterbi no conjunto das unidades fonéticas segmentadas manualmente. Para cada modelo de unidade fonética, o alinhamento de Viterbi segmenta os vetores de observações em N estados. Em cada estado, aplica-se o algoritmo *k-means* modificado, detalhado na figura 4.5, estimando os parâmetros (média e variância) das respectivas distribuições gaussianas, obtendo-se um novo conjunto de parâmetros para cada modelo;

3. Reestimam-se todos os parâmetros com o uso da reestimação de Baum-Welch;

4. Comparam-se os novos parâmetros reestimados com os anteriores, por meio de uma distância de similaridade estatística entre os HMMs. Se o novo modelo for qualificado como melhor, substituem-se os parâmetros estimados pelos re-estimados; caso contrário, realiza-se nova re-estimação.

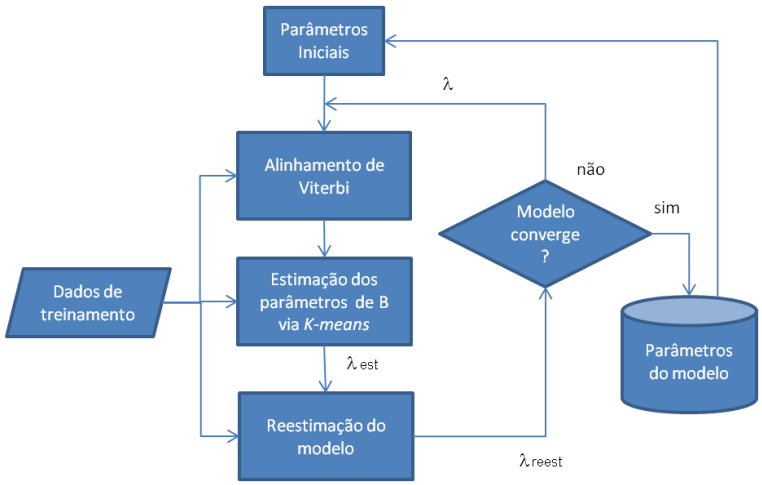


Figura 4.5: Detalhamento do procedimento de treinamento *segmental k-means* [2].

Após esta primeira rodada, aplica-se o procedimento *embedded* treinando com 1026 locuções sem marcação. Neste treinamento, os modelos HMMs sementes são conectados conforme a sequência da transcrição fonética para cada locução, e realizada a reestimação de todos os modelos em um mesmo ciclo de treinamento. Em [2],

foi verificado que a sequência de unidades fonéticas desconectadas não é natural e grosseiramente ineficiente para reconhecimento da fala contínua, e em [6] na maioria dos testes onde se aplicou o procedimento *embedded* houve ganho na taxa de acertos das marcações entre unidades em relação ao treinamento com unidades isoladas. Para este último caso, o autor destaca que se a base que foi usada para treinamento das unidades isoladas for pequena, por exemplo com até 100 locuções, existe ganho quando se aplica o procedimento *embedded*; entretanto, se a base dados para treinamento das unidades isoladas for grande, o uso adicional do procedimento *embedded* pode até mesmo degradar o resultado, o que também foi verificado em [19]. No presente estudo desta tese, a base para treinamento possui 1026 locuções não segmentadas manualmente e 200 locuções segmentadas para testes. Sendo assim, optou-se por dividir a base de testes em duas: uma com 50 locuções para criação dos HMMs "sementes" e, o restante, 150 para teste de desempenho do sistema.

A partir dos modelos-sementes, o treinamento prossegue com as seguinte etapas:

1. Para cada locução do banco de treinamento com 1026 locuções, ou seja, para cada sequência de unidades acústicas concatenadas, usa-se a concatenação dos HMMs "sementes"das unidades conforme a sua transcrição fonética, concatenação esta chamada de rede dinâmica de HMM;

2. Aplica-se uma re-estimação no conjunto de modelos HMMs concatenados utilizando o algoritmo de Baum-Welch, estimando novas marcações;

Este procedimento pode ser aplicado quantas vezes se achar necessário. Em testes realizados no âmbito deste trabalho, verificou-se que o ganho após algumas rodadas do procedimento *embedded* é estabilizado. Como exemplo, foi realizado um treinamento com 200 iterações. Na primeira iteração, houve uma variação na taxa de erro absoluto médio de 46,7 para 29,5 ms; já para as demais iterações, houve pequena variação em torno da média de 27,3 ms, com leve tendência para baixo, como pode ser verificado na figura 4.6, o que pode indicar uma estabilização em torno de um mínimo local, cuja possível solução envolveria um acréscimo de gaussianas nos estados, conforme [17].

Com os HMMs treinados e selecionados, foram consideradas quatro configurações para o sistema MHMM proposto:

1. **MHMM 1**: Sistema utilizando o mesmo HMM (que obteve o menor MAE global) para todas as unidades fonéticas;

2. **MHMM 2**: Sistema com um HMM distinto para cada unidade fonética;

3. **MHMM 3**: Sistema utilizando os n melhores HMMs (de menores MAEs globais) para todas as unidades fonéticas;

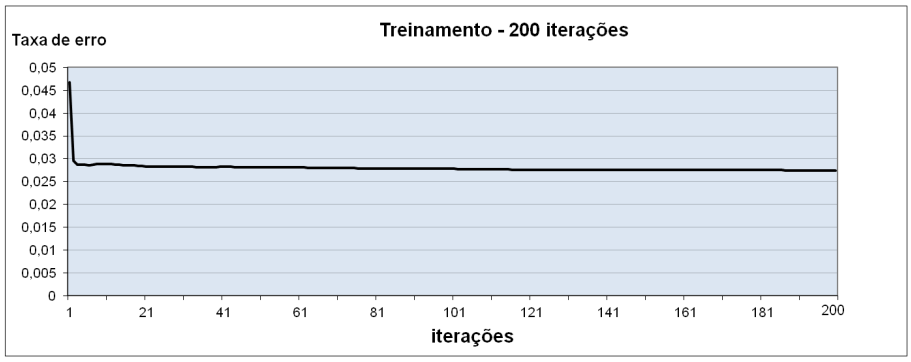


Figura 4.6: Desempenho ao longo do processo de treinamento usando o procedimento *embedded* com 200 iterações.

4. **MHMM 4**: Sistema com n HMMs para cada unidade fonética.

Estas configurações podem ser compreendidas com o apoio da figura 4.7 que ilustra um exemplo fictício, no qual existem $n = 3$ configurações de HMM e seus respectivos MAEs. No sistema MHMM-1, a escolha do modelo global (isto é, para todas as unidades fonéticas) é a configuração com o menor MAE. No sistema MHMM-2, a escolha se baseia nos MAEs individuais para cada unidade; assim, para os fones [a], [b] e [t], são selecionados os HMMs 3, 2 e 1, respectivamente. Nas configurações MHMM-3 e MHMM-4, realiza-se o mesmo procedimento escrito para MHMM-1 e MHMM-2, respectivamente, só que selecionando n modelos em cada caso.

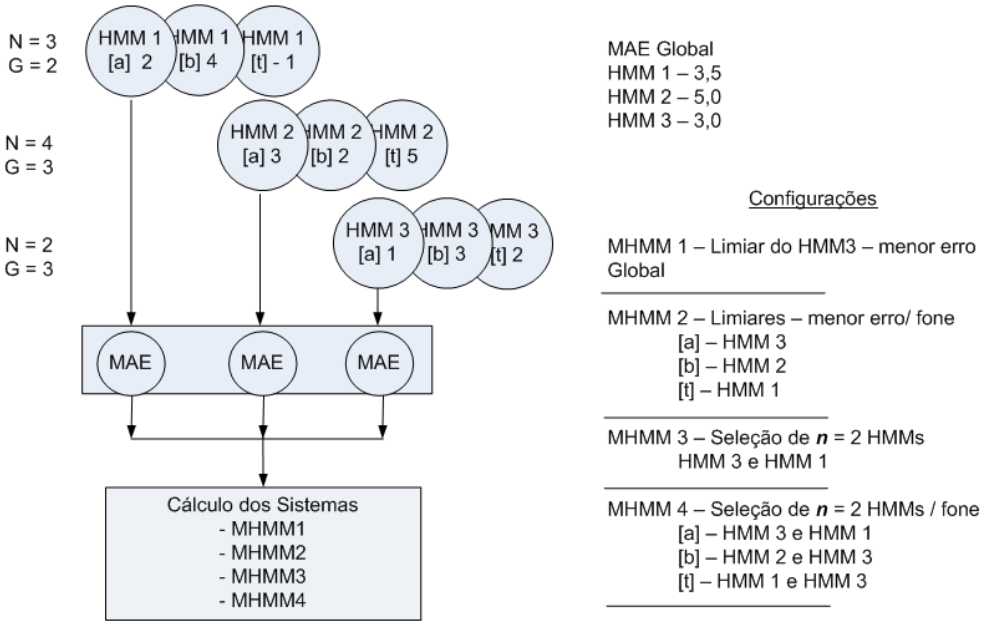


Figura 4.7: Ilustração do procedimento para seleção dos modelos em cada configuração do sistema MHMM.

4.5 Segunda Etapa: Análise de Métricas Fonéticas

Em [5] é citado que existe, para cada linguagem, uma média de 30 a 50 sons básicos chamados por fones, formados pelos sons das vogais e consoantes. No português brasileiro, seguindo o trabalho de [1], podemos considerar um total de 35 fones agrupados em 10 classes, conforme visto na seção 3.4.

Para a base de dados utilizada neste trabalho, gravada por um único locutor, a duração média dos fones foi de 79,2 ms, sendo que, o menor valor de média individual foi para o fone [b] (de ***b*ola**) com uma duração média de 8,4 ms, enquanto a máxima duração foi do fone [O] (como em **t*o*ca**) com duração média de 173,1 ms. Quanto ao desvio padrão, também existe uma ampla faixa de variação em seus valores, do menor para o fone [b], na faixa de 2,5 ms, ao maior, obtido para o fone [E] (como em **p*e*le**), na faixa de 51,2 ms. Observa-se, então, pelo desvio padrão apresentado, que a pronúncia de uma dada locução, ainda que feita pelo mesmo locutor, sempre possui alterações dinâmicas significativas nos parâmetros que a representam.

O desvio padrão obtido tanto pode ser devido às adversidades nas condições de gravação, tais como a proximidade do microfone e níveis de ruído do ambiente, ou quanto à variabilidade das características da fala devido às condições físico-emocionais do locutor. Um outro fenômeno que ocorre na geração do sinal da fala é a coarticulação entre fones adjacentes, inclusive de palavras distintas [5]. Ao pronunciar sequências de palavras não separando-as por pausas, como "maizuma-cerveja", suprimindo finais de palavras, como "ligoudinov", ou seja, os locutores não pronunciam a sequência de sons através da pronúncia discretas de cada som, e as informações sobre qualquer fonema individual fica espalhada sobre um período que excede a duração do fonema em questão. A duração dos sons ainda é afetada pela entonação que se deseja dar a uma expressão, verificado por Barbosa em [40], que os acentos lexicais alongam tanto a vogal quanto, também, a consoante da sílaba acentuada, enquanto os acentos frasais aumentam a duração também da consoante seguinte à vogal acentuada.

A adoção de definições para a segmentação do sinal da fala deve ser bastante cuidadosa, pois deve-se buscar ferramentas que apresentem algum mecanismo dinâmico na busca das fronteiras dos fones. Em um estudo para a segmentação de sentenças, realizado por Wang [3], foi apresentada uma metodologia que ponderava a dinâmica da fala do locutor, utilizando a duração da pausa e a velocidade de emissão das vogais. Neste método de segmentação de sentenças (para uma base de dados em inglês) foram utilizadas as seguintes características da voz: duração da pausa, duração do fonema e as regras de prosódia. Após a aplicação de um método de classificação adaptativa vogal-consoante-pausa (VCP), também proposto por Wang, processa-se

o sinal em três algoritmos, como indicado na figura 4.8: (1) inicialmente, obtêm-se as fronteiras das sentenças baseadas nas ocorrências de pausas; (2) realiza-se, então, uma estimação da taxa de voz (ROS, do inglês *rate of speech*); e finalmente, (3) aplicam-se conceitos das curvas da prosódia, obtendo uma fronteira candidata. Para o primeiro passo, o autor avalia que a duração da pausa é afetada pela velocidade da fala, sendo curta para falantes com fala rápida, sendo esta uma característica natural do falante ou atribuído à emoção que se deseja expressar. Esta variação dinâmica da duração da pausa é compensada quando combinada com a característica taxa da voz, aumentando a robustez do sistema. O sistema finaliza-se com o uso do classificador AdaBoost [41] para estimar a fronteira da sentença. Este estimador obtém a fronteira final após a soma ponderada das entradas chamadas por classificadores fracos, onde os pesos são atualizados dinamicamente de acordo com os erros na aprendizagem anterior. Usando este procedimento a taxa de reconhecimento obtida das fronteiras entre as palavras para a base completa foi de 82,3%.

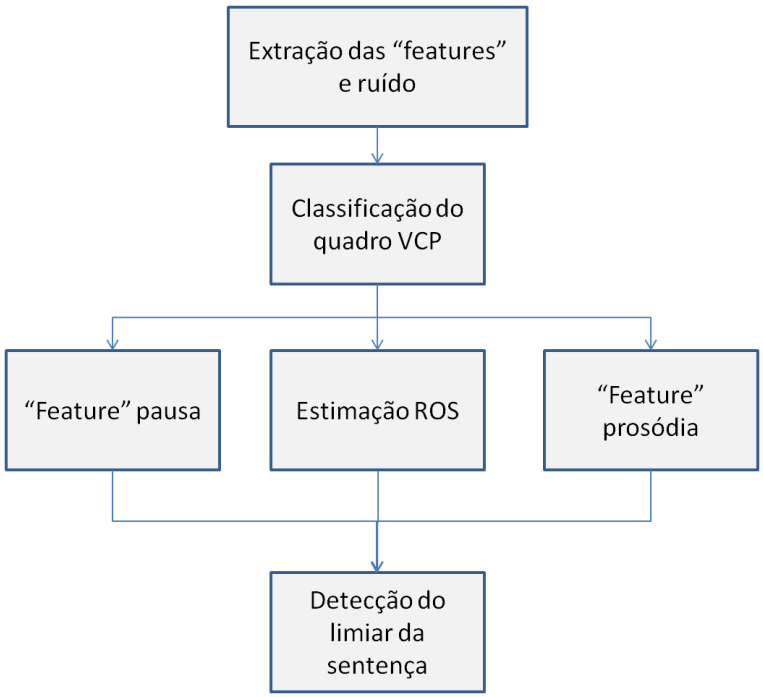


Figura 4.8: Ilustração do sistema de segmentação de sentenças [3].

Desta forma, um dos objetivos deste trabalho é proporcionar ao algoritmo proposto ferramentas capazes de monitorar variações e, sendo possível, compensá-las diminuindo a taxa de erro. Para tanto, serão analisadas nesta seção algumas métricas baseadas nas características fonéticas, medidas de distâncias, variações da região de refinamento e a duração dos fones.

4.5.1 Métricas Fonéticas

Medida de Distância

A região correspondente ao término de um fone e o início do fone seguinte, devido à coarticulação, possui suas características correlacionadas [5], dificultando a estimação/seleção de um ponto específico que separe os dois fones, discriminando duas regiões (ou dois conjuntos de atributos) distinta(o)s. Esta discriminação pode ser analisada sob a ótica das técnicas de reconhecimento de padrões, onde se pretende analisar o conjunto de características acústicas dos dois fones e encontrar uma distância de dissimilaridade entre dois segmentos de voz.

Campbell em [42] apresenta um trabalho sobre reconhecimento automático de locutor, utilizando técnicas de identificação de padrões aplicadas em conjunto de vetores de características da fala. O objetivo é identificar ou verificar o falante em relação a um banco de amostras armazenadas. Para a atividade de verificação do locutor existem três possíveis conjuntos de amostras: um conjunto de características acústicas do locutor pretendente, n conjuntos de características acústicas da fala de n locutores treinados; e mais uma possibilidade de classificação como locutor desconhecido. A tarefa do sistema consiste em verificar, através de medidas de distâncias, qual dos n locutores é o locutor pretendente ou se esse será classificado como desconhecido. As distâncias utilizadas por Campbell foram a distância euclideana, a distância de Mahalanobis e a distância de Bhattacharyya, apresentando-se os respectivos resultados de taxa de falsa aceitação 1.96%, 1.08% e 0.21%. Como visto, o melhor resultado, obtido com o uso da Distância de Bhattacharyya, foi considerado por Campbell como satisfatório em comparação a outras medidas.

Considerando ainda a definição dada por Fukunaga em seu livro [43], que define a equação da distância de Bhattacharyya(equação 4.1) como aplicável para distribuições normais e não-normais, para uma avaliação adequada da separabilidade de classes, a distância de Bhattacharyya foi escolhida, então, como medida de descontinuidade entre as distribuições estatísticas, para cada dupla de fones adjacentes neste trabalho, junto com a medida de BIC definida na Subseção 3.4. Para as classes k e x, a distância de Bhattacharyya é dada pela expressão

$$B_{xk} = \frac{1}{8}(\mu_k - \mu_x)^T \left(\frac{\Sigma_k + \Sigma_x}{2}\right)^{-1} (\mu_k - \mu_x) + \frac{1}{2} \ln \frac{\left|\frac{\Sigma_k + \Sigma_x}{2}\right|}{\sqrt{|\Sigma_k||\Sigma_x|}} \qquad (4.1)$$

Para ilustrar o uso destas duas técnicas de medidas segue a distribuição dos parâmetros acústico dos fones [a] e [s], como apresentada na figura 4.9. No eixo das ordenadas têm-se os valores de taxa de cruzamento de zeros (TxZ) e no eixo das abscissas os valores de centro de gravidade espectral (CGE), extraídos do segmento de sinal [a][s], da palavra "muit*as*" presente na locução L001, segmentados com

uma janela de Hamming de 20 ms.[1] Por este gráfico percebe-se que existe um espalhamento das características do fone [s], fonema fricativo surdo, com uma pequena concentração para a TxZ em torno de 0.5, conforme sugerido como ponto de corte por [1].

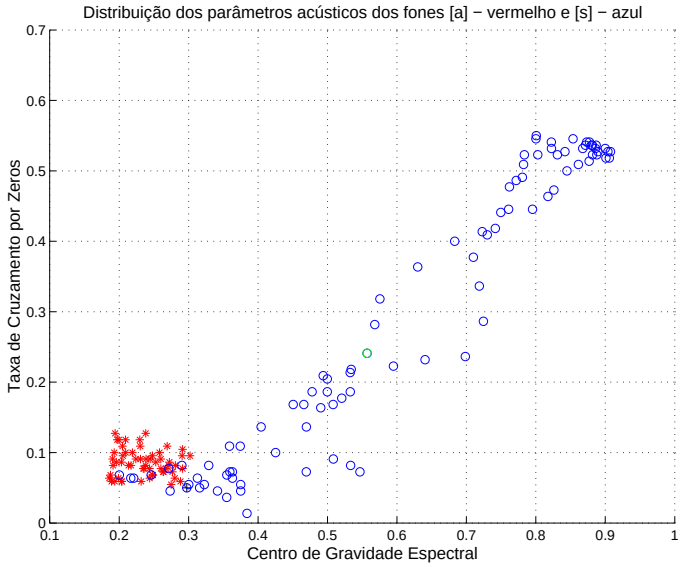


Figura 4.9: Relação entre taxa de cruzamento por zero (TxZ) e centro de gravidade espectral (CGE) para os dois fones [a] (representado por asterísticos) e [s] (representado por círculos).

Para a distância de Bhattacharyya, a figura 4.10 mostra a variação dos resultados da distância entre as características TxZ e CGE, para a mesma dupla de fones [a][s], enquanto que a distância BIC dos mesmos parâmetros acústicos é mostrada na figura 4.10.

Na figura 4.11 os limiares resultantes das distâncias BIC e Bhattacharyya são comparados com o limiar de referência obtido por um foneticista, indicando a grande similaridade das marcações obtidas por estas métricas objetivas, tanto entre si como quando comparadas com a marca dada pelo profissional.

Duração Média dos Fones e Região de Refinamento

Como visto anteriormente, o fenômeno da coarticulação entre fones adjacentes dificulta a definição da fronteira acústica real que os separa. Isto porque passa a existir uma variabilidade acústica, dependente do contexto, tal que a correspondência entre fones pronunciados e o acontecimento acústico não são separados em fronteiras bem definidas.

Como exemplo desta situação, a palavra "pessoas", disposta em diferentes posições e contextos prosódicos dentro das frases L001 e L097 foram geradas com di-

[1]* L001 - Muitas pessoas participam da construção de um texto.

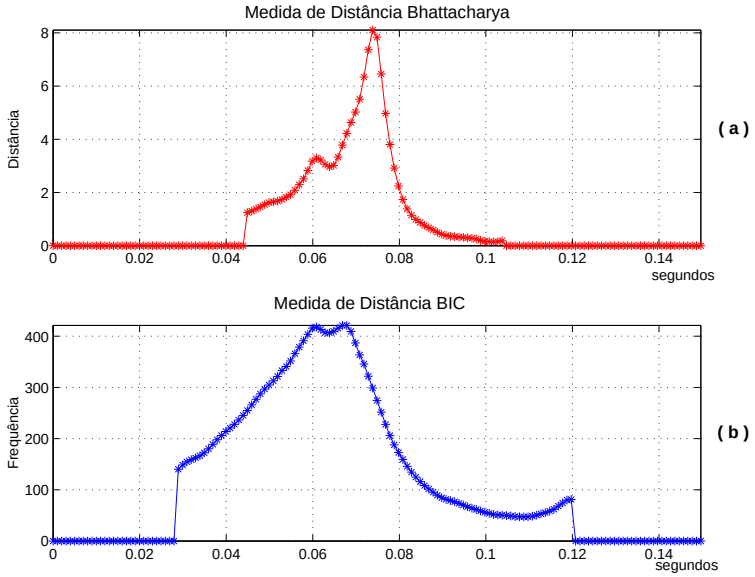


Figura 4.10: Distâncias de (a) Bhattacharyya e (b) BIC aplicadas sobre as representações paramétricas da figura 4.9 para os fones [a] e [s].

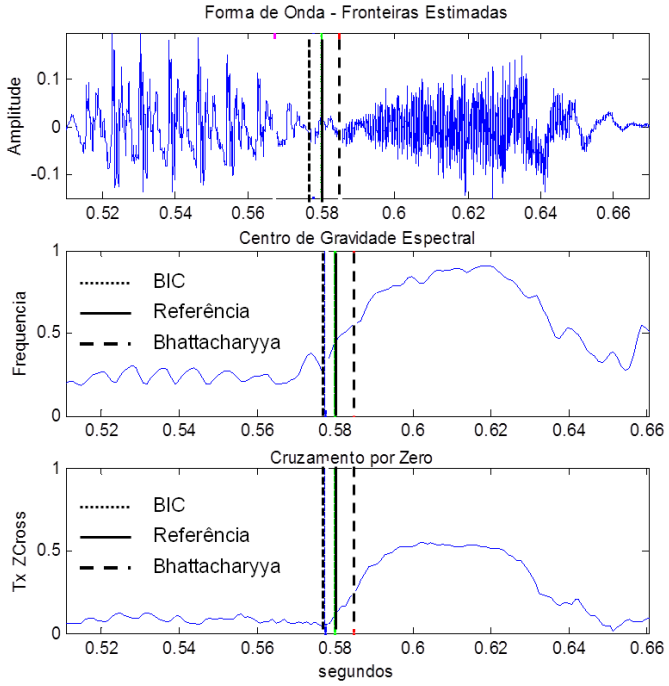


Figura 4.11: (a) Forma de onda dos fones [a][s] da palavra "muitas" presente na locução L001. Estão colocados os limiares referência obtidos por foneticista, a partir das distâncias BIC e de Bhattacharyya; (b) Variação temporal do CGE com respectivos limiares de marcação; (c) Variação temporal da TxZ com respectivos limiares de marcação.

ferentes entonações pelo mesmo locutor, conforme indicado nas figuras 4.12 e 4.13.[2]

[2]* L097 - Pessoas comuns pagam muitos impostos

No caso, as figuras 4.12(a) e 4.13(a), possuem a sequência de fones [o]+[a]+[s] separadas por limiares de referência ajustados por um foneticista, enquanto as figuras 4.12(b) e 4.13(b) mostram as sequências de fones [a]+[s] com limiares definidos pelo estimador HMM.

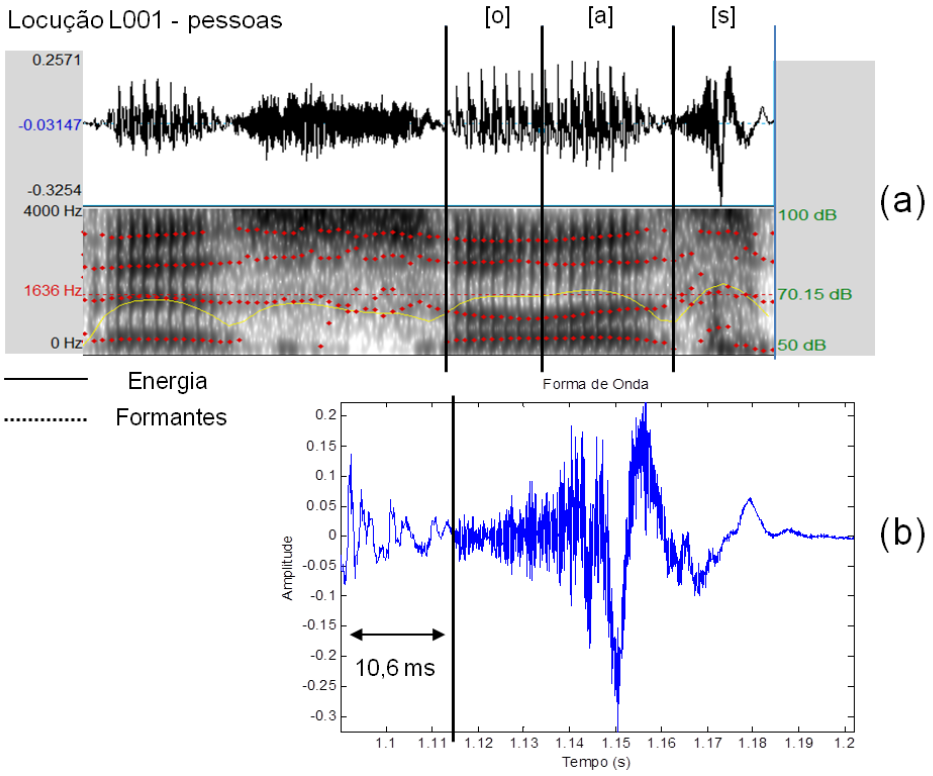


Figura 4.12: Forma de onda e espectrograma da palavra ´´pessoas"na sentença L001. (a) sequência de fones [o][a][s] com suas fronteiras fornecidas por um foneticista. (b) forma de onda da sequência de fones [a][s] com limiar estimado pelo HMM. Note neste caso que a duração do fone [a] estimado pelo HMM é de 10,6 ms, enquanto o fone [a] definido pelo foneticista possui 86,3 ms.

Para a sequência [o][a][s] da locução L001 na figura 4.12, percebe-se, ao escutar a sequência, que características acústicas do fone [o] estende-se para o fone [a], justificando o limiar estimado pelo HMM para a fronteira entre os fones [o][a], e resultando na duração do fone [a] em 10,6 ms, enquanto que deveria ser aproximadamente 86,3 ms, segundo o foneticista. É importante ressaltar que na verificação visual da figura 4.12(a), observa-se que existe uma mudança de tendência de redução para aumento da formante $F2$ e um ponto de quebra da energia , caracterizando pontos de mudança dos fones [o] para [a], sendo determinado neste ponto o limiar de separação por um foneticista.

Já na figura 4.13, a palavra “pessoas”, que inicia a frase, possui as características acústicas dos fones bem mais definidas, de modo que o estimador por HMM gera limiares mais próximos da referência. A duração deste fone [a] foi de 65,1 ms e,

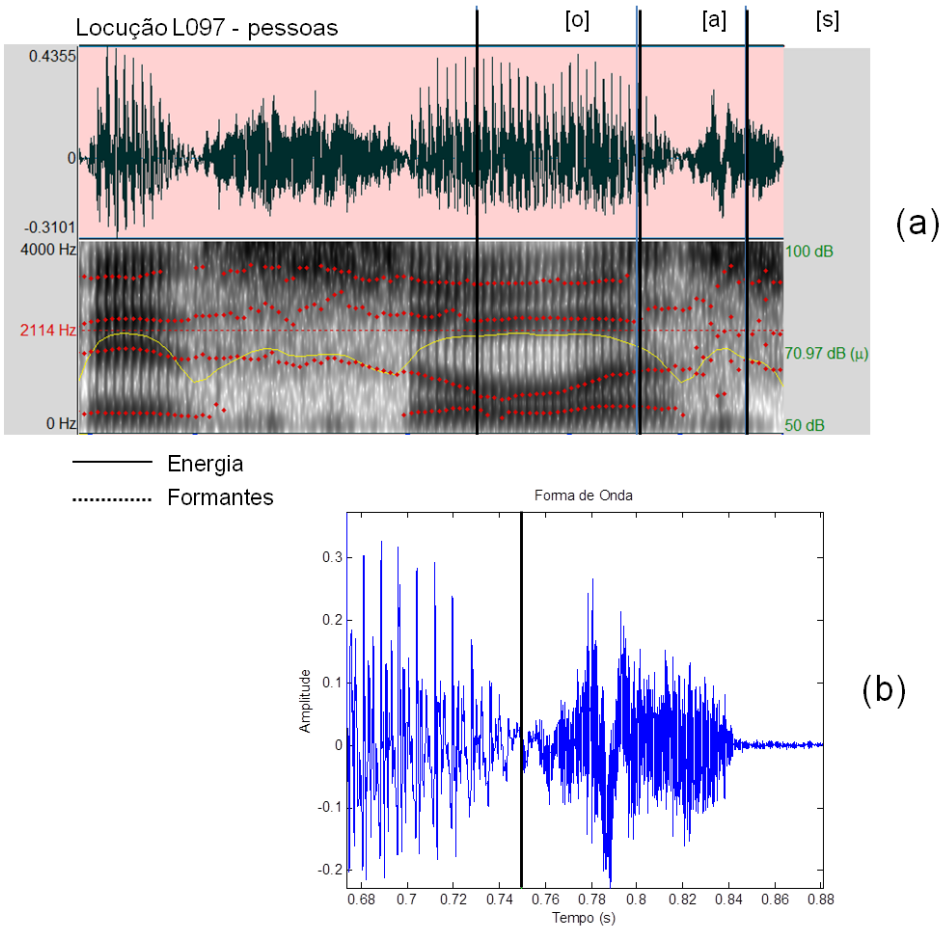


Figura 4.13: Forma de onda e espectrograma da palavra "pessoas" na sentença L097. (a) sequência de fones [o][a][s] com suas fronteiras fornecidas por um foneticista. (b) forma de onda da sequência de fones [a][s] com limiares estimados pelo HMM. Neste caso houve uma correta estimação da fronteira entre fones [a] e [s] pelo HMM.

entre os limiares de referência na figura 4.13, de 99 ms.

Para a técnica de refinamento por regras fonéticas, a duração da sequência de fones adjacentes é uma função dos limiares estimados pelo HMM na primeira etapa do sistema segmentador. Desta forma, para a sequência de fones [a][s] da figura 4.12, a duração total será entre os limiares de início do fone [a] até o limiar final do fone [s]. Entretanto, para a sequência de fones [a][s] na locução L001, o fone [a], numa primeira estimativa gerada pelo HMM, possui uma duração de apenas 10,6 ms, dificultando o processo de refinamento subsequente. Para evitar situações onde venha ocorrer fones com pobreza de características acústicas, como indicado anteriormente, utiliza-se a regra de duração média dos fones. Com esta regra somente os fones que estiverem com a sua duração dentro de uma faixa de duração esperada, definida a seguir, serão considerados para o cálculo das suas fronteiras, e o processo de refinamento por regras fonéticas segue o procedimento:

- Identificação dos fones: A partir da transcrição das frases analisadas, tem-se a sequência de fones. Define-se o fone da esquerda do par de fones em análise;
- Busca do tempo médio e do desvio padrão: Busca-se na base a duração média $\overline{t_{m[fone]}}$ e o desvio padrão $\delta_{[fone]}$ do fone da esquerda;
- Definição da faixa de busca do novo ponto de segmentação: Sejam os limites desta faixa de busca dados por

$$limiar_{[inferior]} = \overline{t_{m[fone]}} - \delta_{[fone]} - \sigma, \tag{4.2}$$

$$limiar_{[superior]} = \overline{t_{m[fone]}} + \delta_{[fone]} + \sigma, \tag{4.3}$$

onde σ é um ajuste empírico, a nova faixa de busca é dada por

$$\Delta_{[fone]} = limiar_{[superior]} - limiar_{[inferior]}. \tag{4.4}$$

- Verificação do limiar estimado: Identifica nesta nova faixa de busca o limiar estimado $limiar_{estimado}$ na primeira etapa da segmentação;
- Definição da região de refinamento: Aplica-se em torno deste limiar o intervalo

$$\Theta_{[fone]} = limiar_{estimado} + 2\theta_{[fone]}, \tag{4.5}$$

onde θ é um ajuste empírico, definindo-se a região em que será feito o refinamento, conforme ilustrado na figura 4.14.

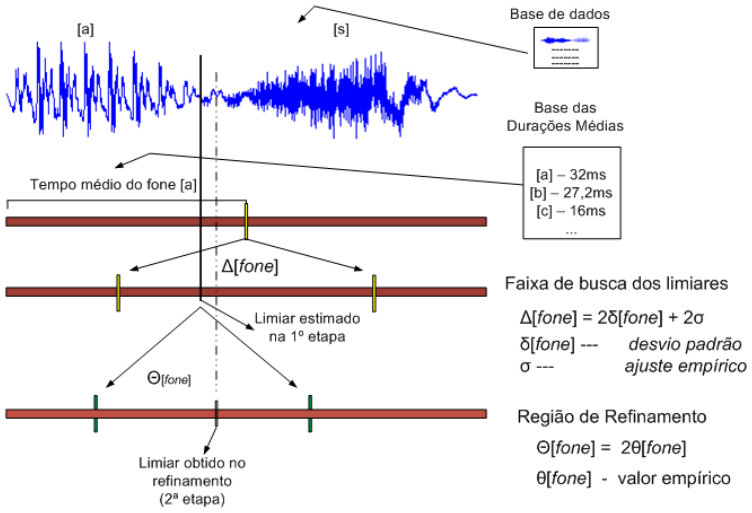


Figura 4.14: Exemplo de definição da faixa de busca e da região de refinamento por regras fonéticas.

- Extração dos parâmetros: Após a definição da faixa de refinamento, esta será janelada a cada 20 ms, com passos de 1 ms ou 10 ms, conforme o fone. Para cada janela, são extraídos os parâmetros acústicos adequados para o processamento das regras acústicas-fonéticas, como por exemplo: energia, taxa de cruzamento de zeros e centro de gravidade espectral.

- Definição da nova fronteira: Esta etapa é feita em cima dos parâmetros anteriormente extraídos aplicando-se as regras de segmentação para cada classe de unidade acústica, conforme discutido no capítulo 3.

4.5.2 Classes Fonéticas

Nesta subseção, serão relacionadas as classes com as respectivas métricas utilizadas para segmentação das unidades fonéticas.

Classe das Consoantes Fricativas

Para a definição das fronteiras na classe fricativa foram utilizadas a energia do sinal e as distâncias de Bhattacharyya e BIC, aplicadas sobre as características de TxZ e CGE. Os parâmetros acústicos foram calculados para janelas de análises com duração de 10 ms e deslocamento de 1 ms, de acordo com os testes. A energia foi calculada com o sinal segmentado com a janela de Hamming, considerando-se cinco janelas adjacentes (janela atual, duas janelas à esquerda e duas à direita) e calculado para uma banda de 0 a 4000 Hz.

Para ilustrar, nas figuras 4.15 e 4.16 é mostrado um exemplo da estimativa de fronteira entre dois fones [u] e [f] de um trecho da locução L003. [3] Pela distribuição dos parâmetros acústicos vista na figura 4.15, percebe-se a existência de duas regiões distintas, que, após aplicar as distâncias BIC ou Bhattacharyya (mostradas na figura 4.16), seus limiares são gerados muito próximos do limiar de referência.

Classe das Consoantes Plosivas

As consoantes classificadas como plosivas sonoras são (/k/, /d/ e /g/) e como plosivas surdas (/p/, /t/ e /b/).

Para a classe das consoantes plosivas podem ocorrer as seguintes situações de transições entre fones:

- Transição entre silêncio e início da classe plosiva. Período definido por uma região com pouca ou nenhuma energia seguida por uma explosão. Neste cenário pode ser aplicado a regra de transição entre o silêncio e uma classe

[3]* L003 - Cada aluno fez a sua avaliação.

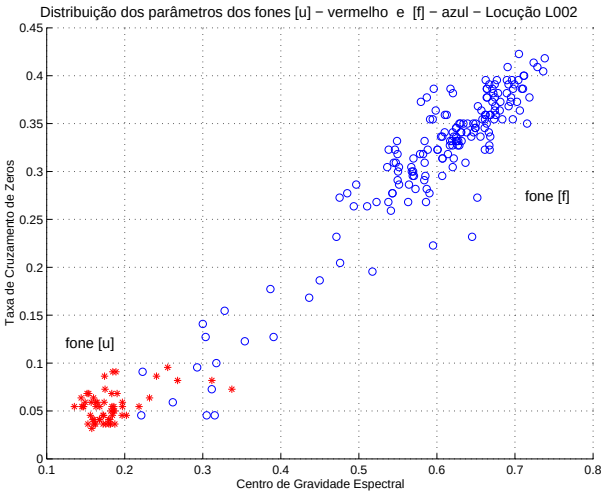
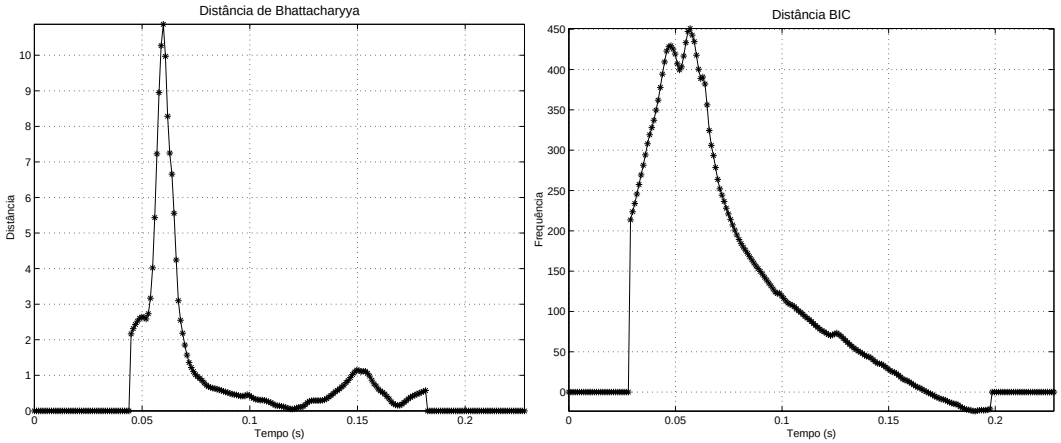


Figura 4.15: Comportamento dos parâmetros acústicos da sequência de fones [u] (asterísticos) e [f] (círculos).



(a) Distância de Bhattacharyya (b) Distância BIC

Figura 4.16: Distâncias de Bhattacharyya e BIC para a distribuição dos parâmetros TxZ e CGE da figura 4.15.

fonética, qual seja, uso do limiar de - 60 dB entre silêncio e plosivas sonoras; e -70 dB entre silêncio e plosivas surdas como sugerido por Selmini [1], em sua tese de doutorado;

- Transição entre outras classes para a classe plosivas. Ocorre quando a consoante plosiva não se encontra no início da locução. Neste caso, a classe que precede a plosiva faz a transição para uma região que está ocorrendo a oclusão, ou seja, uma região de natureza dinâmica e que apresenta informações acústicas de curta duração. Bonatt [44] destaca em seu trabalho características de produção de fala de sons plosivos, tais como, qualidade do *burst* (ruído que ocorre na sequência à fase de oclusão, na liberação repentina da corrente de ar, visto no espectrograma como uma faixa vertical de energia, como visto

na figura 4.17)[4] e a medida de duração do *voice onset time* (VOT), que corresponde ao intervalo de tempo entre a liberação da oclusão e o início do vozeamento. Para [1] a transição de uma classe fonética para uma consoante plosiva é marcada por uma variação abrupta da energia espectral e, no trabalho apresentado por Juneja [29] são sugeridos dois parâmetros para cálculo do limiar de transição: as respectivas energias nas bandas 0-$F3$ e $F3$-$f_a/2$.

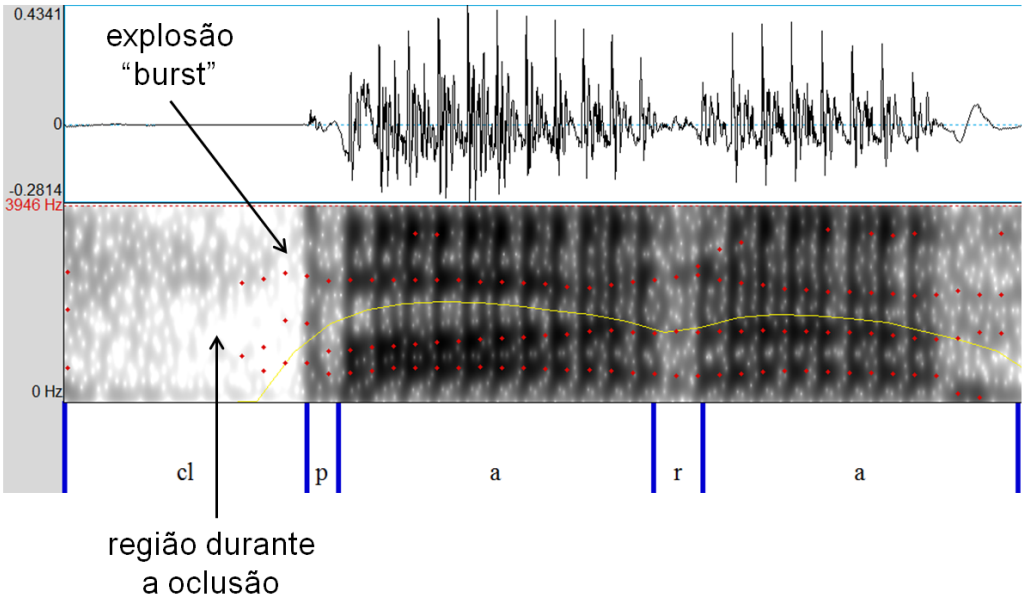


Figura 4.17: (a) Forma de onda e (b) espectrograma da palavra "para"retirada da locução L019 (base teste), onde se vê a região da plosiva surda [p], com destaque para a ocorrência do "burst"(barra vertical de energia no espectograma) verificado antes da explosão do som.

- Transição da classe plosiva para outras classes. Neste tipo de transição, quase sempre de uma plosiva para uma vogal, existem características acústicas que contribuem para a identificação de fronteiras. Em [1] foi verificada a existência de dois picos, identificados pela derivada da energia, que correspondem a ocorrência do *burst*, explosão na liberação do ar, e outro pico, gerado pelo aumento de energia devido ao início do fone seguinte. O segundo pico torna-se um candidato ao limiar refinado. Na figura 4.18, são mostradas as transições das plosivas surdas [t] e [k] para o fone [a] da classe das vogais. Para estas transições, são extraídos os vetores de características acústicas TxZ e CGE para segmentos de 10 ms obtidos com a janela de Hamming e deslocados de 1 ms entre si, como visto nas figuras 4.18a e 4.18b. Estas figuras mostram distribuições concentradas dos parâmetros do fone [a] e distribuições dispersas para

[4]* L019 - Todos correram para pegar o cachorro.

as consoantes, não havendo, de modo geral, uma mistura de características acústicas das duas classes. Na análise seguinte, para as plosivas sonoras, cujos parâmetros acústicos são representados na figura 4.19, percebe-se uma dificuldade na separação dos grupos em duas classes distintas.

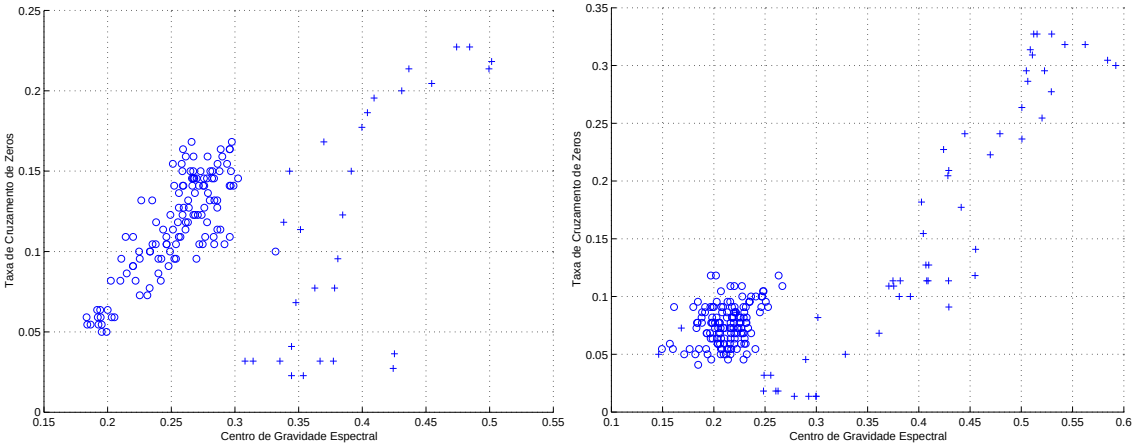


Figura 4.18: Distribuição dos parâmetros acústicos TxZ e CGE para transições da classe plosiva surda para vogal: (a) fone [t] "+" para [e] "o"; (b) fone [k] "+" para [a] "o".

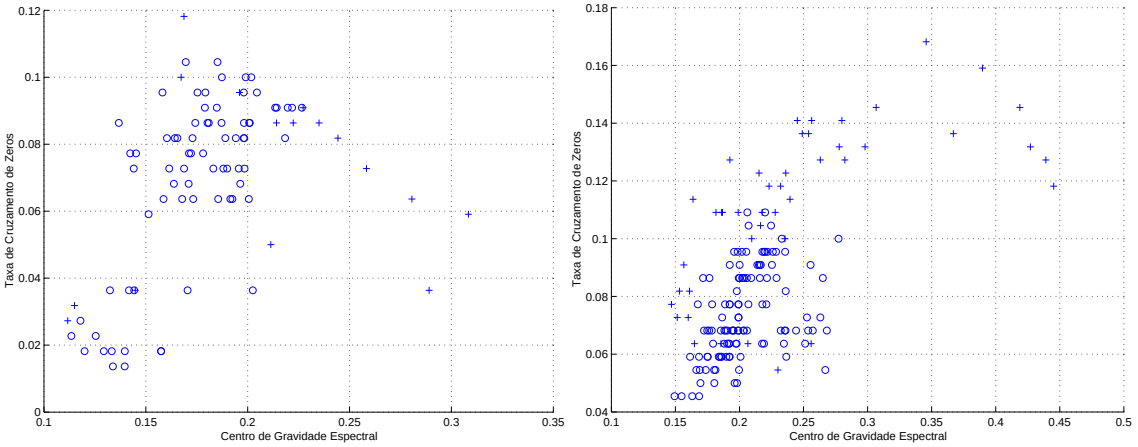


Figura 4.19: Distribuição dos parâmetros acústicos TxZ e CGE para transições da classe plosiva sonora para vogal: (a) fone [d] "+" para [a] "o"; (b) fone [g] "+" para [a] "o".

A forma sugerida [1] para contornar este problema é o uso do mesmo procedimento apresentado no item anterior. Isto porque, neste caso, a consoante plosiva se torna semelhante a uma consoante fricativa, com o fluxo de ar adquirindo características turbulentas ao passar pelo estreitamento do trato vocal.

Devido à duração dos fones, as consoantes plosivas podem não ser devidamente produzidas quando ocorrem no final de uma sílaba, onde a pressão pulmonar é inferior, o que reduz ainda mais a pressão na obstrução do trato vocal que é necessária para a produção da plosiva. Desta forma, a análise da energia pode ser aplicada

às plosivas sonoras e a métrica de distância às plosivas surdas, como analisado com mais detalhes no capítulo 5.

Classe das Consoantes Africadas

A classe das consoantes africadas são caracterizadas por uma transição entre uma consoante plosiva e uma consoante fricativa. Ocorre primeiro o bloqueamento completo do trato vocal, seguido de um pequeno estreitamento da passagem do ar entre os lábios. Ocorrendo uma fricção do ar através de uma fenda no meio da boca. As consoantes africadas do português brasileiro são [T] e [D] seguidas pela vogal [i].

As possibilidades de transição desta classe são:

- Transição entre outras classes e a classe africada. Para esta classe é interessante a aplicação das regras das classes plosivas e fricativas, dado o comportamento do fone em questão.

- Transição entre a classe africada com a vogal "i" . Nesta situação, em que existe uma passagem de ar com fricção pelas paredes labiais, há uma alta taxa de cruzamento de zeros e o centro de gravidade espectral abaixo de 2 kHz [1]. Isto permite a aplicação dos mesmos conceitos da transição de outras classes para a classe fricativa. De fato, a figura 4.20 mostra a distribuição dos parâmetros acústicos para as sequências [T][i] e [D][i], retirado de um trecho de locuções do banco de testes, indicando uma grande diferença entre as duas classes de fones.

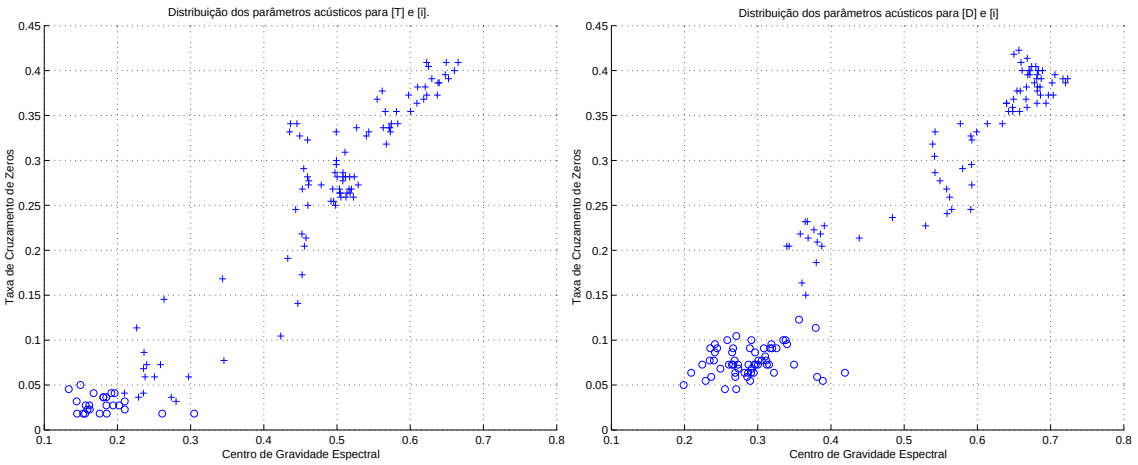


Figura 4.20: Distribuição dos parâmetros acústicos TxZ e CGE para transições da classe africada para vogal: (a) fone [T] "+" africado para [i] "o"; (b) fone [D] "+" africado para [i] "o".

Classe das Consoantes Laterais e Vibrantes

Conforme visto anteriormente, os fones da classe lateral são gerados quando a passagem da corrente de ar é permitida pela língua somente pelas laterais. Na língua portuguesa, são dois os fones consonantais laterais, fones [l] e [lh]. A classe vibrante representa os fones gerados pela vibração da ponta da língua em contato intermitente com os alvéolos, abrindo e fechando a passagem à corrente respiratória. Os fones gerados pertencentes à classe vibrantes são [r] e [R]. Este trabalho segue a sugestão de [1] classificando o fone [rr], pertencente a fricativa posterior, na classe vibrante devido ao desempenho no refinamento.

Para estas classes foi observado que a variação da energia espectral tem sido sugerida por algumas pesquisas como uma alternativa de analisar os limiares de fronteira. Sendo assim, mostra-se que a distribuição de energia da classe lateral é próxima à da classe das vogais, ocorrendo abaixo de 4 kHz, enquanto que para as vibrantes, ocorre abaixo de 1 kHz. A variação da energia espectral em diferentes bandas de frequência motivou que Golipour [45] propusesse um método de cálculo da variação da energia em diferentes bandas, localizando as variações de energia nos picos, definindo como as fronteiras de segmentação, como indicado na tabela 4.2. Neste caso, o cálculo da variação da energia espectral com 7 janelas adjacentes (a central mais 3 para cada lado) é dado por

$$\Delta Eb_i(n) = \frac{\left\| \sum_{\theta=-3}^{3} \theta \cdot Eb_i(n+\theta) \right\|}{7}, \tag{4.6}$$

onde $\Delta Eb_i(n)$ é a variação da energia espectral na banda de freqüência i para a janela de análise n. E para a variação total por janela (ΔE), somam-se as 5 bandas, conforme

$$\Delta E(n) = \sum_{\imath=1}^{5} \Delta Eb_i(n). \tag{4.7}$$

Tabela 4.2: Bandas espectrais de caracterização das consoantes laterais e vibrantes.

Bandas	Variação
B1	0 - $f_a/2$
B2	0 - 500 Hz
B3	500 - 1500 Hz
B4	1500 - 2400 Hz
B5	2400 - $f_a/2$

Classe das Vogais e Consoantes Nasais

As definições das transições para as classes fonéticas vistas até o momento já incluem a transição das outras classes para a classe das vogais. Os fones desta classe são sons produzidos pela passagem do ar pelas cordas vocais, e que passam livremente pelo pelo trato vocal e nasal sem a ocorrência de nenhum tipo de constrição. Como visto anteriormente, as vogais são divididas, de acordo com a zona de articulação em:

- **Classe anterior ou palatal** - Fones [e] - d*e*do, [E] - p*e*le, [i] - bot*i*na;
- **Classe média ou central** - Fone [a] p*á*, *á*tomo;
- **Classe posterior ou velar** - Fones: [o] - b*o*lo, [O] - p*ó* e [u] - l*u*a;

As transições analisadas para as vogais são:

- Para a transição entre a vogal média [a] com as vogais anteriores ([e], [E], [i] e [y]) ou posteriores ([o], [O] e [u]), são analisadas variações dos formantes $F1$ e $F2$;
- Para a transição entre vogal anterior e uma vogal posterior ou vice-versa, usa-se o perfil de energia e faz-se a análise do formante $F2$;
- Para a transição entre vogais de mesma classe, usa-se a distância BIC.

Para a classe das consoantes nasais, o cálculo da transição sugerida por Juneja [29] utiliza a energia das bandas, conforme colocado na tabela 4.3, para a detecção dos limiares na transição entre as consoantes nasais e as vogais.

Tabela 4.3: Caracterização da transição das consoantes nasais.

Bandas	Variação	Motivo
$B1$	0 - 358	concentração de energia nas baixas frequências
$B2$	358 - 5378 Hz	concentração da energia das vogais

4.6 Conclusão

Neste capítulo foram detalhadas as fases do processo de segmentação de fones através da etapa de estimar uma fronteira e, a partir desta, na etapa seguinte, obter a nova fronteira analisando a região entre fones adjacentes. Para a base de fala utilizou-se, nesta tese, a base desenvolvida durante o trabalho de doutorado de Selmini [1],

proporcionando uma referência para resultados dos métodos aplicados nesta tese. Como pequena contribuição, foi contratado um foneticista profissional para validar as fronteiras estimadas, que realizou pequenos ajustes no posicionamento das fronteiras quando necessário.

Foi visto que na primeira etapa de segmentação, a estimação dos limiares entre fones será por meio de um conjunto de n estimadores HMM, parametrizados independentemente, que gerarão n estimativas. Após a aplicação de uma métrica, será definida a fronteira estimada desta etapa. As métricas avaliadas foram a média e a mediana, e seus resultados serão apresentados no próximo capítulo.

Para a segunda etapa do segmentador, foram mostradas as métricas utilizadas para estimar os limiares. Foi colocada a questão da variabilidade das diferentes pronúncias de uma determinada locução feita pelo mesmo locutor, o que nos levou a definir conceitos como duração média (e desvio padrão correspondente) para cada fone. Estes conceitos são utilizados então para definir a região de refinamento do limiar de segmentação. O problema da segmentação por classes fonéticas foi revisitado com detalhes dos comportamentos das métricas utilizadas para todas as possíveis transições.

Capítulo 5

Resultados Experimentais

5.1 Introdução

Neste capítulo, discutimos o desempenho das técnicas de refinamento da segmentação. Com este intuito, o capítulo está organizado da seguinte forma: Na seção 5.1 apresentamos uma sinopse do sistema proposto, considerando as duas técnicas de refinamento anteriormente apresentadas; na seção 5.2 apresentamos quatro configurações do sistema MHMM com os respectivos desempenhos enquanto que na seção 5.3 discutimos os resultados obtidos com as regras fonéticas para classe de fones. Por fim, sintetizamos os resultados com a combinação de ambas as técnicas na seção 5.4, procurando explorar a vantagem de cada uma para gerar o sistema final de segmentação.

5.2 Sistema Proposto

Nos capítulos 3 e 4 foram apresentados os conceitos teóricos acerca da modelagem MHMM estatística do sinal da fala e da análise determinística por meio de regras acústico-fonéticas. Esses estudos compõem a proposta de um sistema de estimação das fronteiras, ilustrado na figura 5.1, cuja ideia básica é a combinação das modelagens estatística e determinística, usando-se três sistemas interligados.

O primeiro sistema é formado por 34 HMMs, treinados com diferentes parametrizações usando-se uma base de fala não segmentada manualmente. Nesta etapa, quatro tipos de limiar para a primeira estimativa L_1 das fronteiras são considerados: (i) limiar estimado pelo modelo com a menor taxa de erro para todos os tipos de fonemas; (ii) limiar estimado pelo melhor HMM para cada tipo de fonema; (iii) limiar estimado pela média ou mediana de um subconjunto dos HMMs; (iv) a mesma ideia anterior, entretanto, usando um subconjunto de HMMs para cada tipo de fonema. A segunda etapa usa regras fonéticas, específicas para cada classe de fonema, para

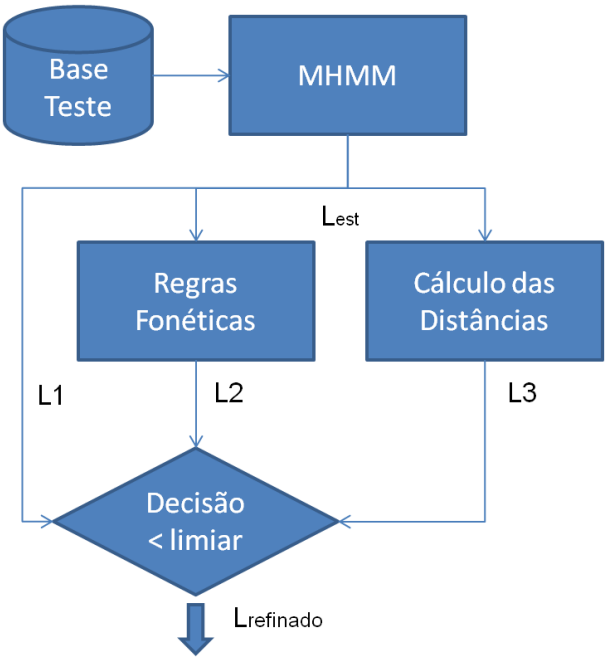


Figura 5.1: Sistema MHMM fornece o limiar estimado para refinamento e para comparação com os limiares já refinados.

fazer um refinamento L_2 do limiar L_1. Por fim, a terceira etapa gera o limiar L_3 por meio da identificação de pontos que separam eventos distintos, usando-se as distâncias de Battacharyya e BIC.

Nos experimentos descritos no presente capítulo, verificamos os desempenhos dos limiares obtidos em cada etapa, para todas as transições entre classes fonéticas, permitindo uma definição final $L_{refinado}$ de cada tipo de transição, como indicado na figura 5.1. Esta possibilidade de alternância das técnicas de segmentação para cada tipo de classe fonética é motivada pelo fato de todo sinal possuir uma elevada variabilidade de geração [42], [2], [5], o que pode dificultar o uso das regras fonéticas para algumas classes específicas, como, por exemplo, a das consoantes laterais na língua portuguesa [1].

Para a definição dos melhores HMMs que representem as classes fonéticas, são necessários vários testes de configurações e análise de seus resultados. Assim, a implementação computacional desta proposta de trabalho se desenvolveu em três etapas: (i) Analisar os parâmetros que configuram as características do sinal da fala; (ii) Identificar os parâmetros que melhor configuram os HMMs; (iii) Avaliar os limiares entre fones adjacentes, identificar o melhor sistema e a melhor métrica para a classe em questão.

5.3 Etapa MHMM

5.3.1 Parâmetros Acústicos

Na fase de pré-processamento, o sinal é submetido a uma filtragem de pré-ênfase, com filtro passa-altas $(1 - cz^{-1})$, com $c = 0{,}97$. Usando-se a função-janela de Hamming, obtém-se um segmento de comprimento J, que se desloca por um espaço de tempo Q, chamado de quadro. As características ou parâmetros são extraídas(os) destes segmentos $x(i)$ de curto tempo.

Para a definição dos parâmetros de extração das características do sinal da fala, foram realizados testes para identificar quais faixas de valores poderiam ser utilizadas no treinamento dos modelos HMMs. Seguem as definições e conclusões de valores para estas faixas:

- **Números de janelas adjacentes N_d para cálculo da derivada (ver equação (3.2))**: Ambikairajah [46], em seu tutorial sobre identificação de linguagem, mostra que o uso de $N_d = 2$ ou $N_d = 4$ janelas adjacentes tem sido amplamente aceito na literatura. Park [6], em seu trabalho, mostrou que o uso de $N_d = 4$ janelas adjacentes apresentou melhores resultados que $N_d = 2$. Por outro lado, Picone [27] acrescenta que há aplicações com $5 \leq N_d \leq 7$, utilizando um quadro de $8ms \leq Q \leq 10ms$, enquanto que Selmini [1] sugere $N_d = 1$.

  A figura 5.2 mostra um teste comparativo de HMMs com misturas de $M = 1, 2, 3$ gaussianas e N_d variando de 1 a 5. Neste teste, utilizou um sinal janelado a 20 ms e quadros entre janelas de 1 ms. Neste teste, os valores $N_d = 1, 2$ e 4 geraram os menores erros, sendo definidos como os valores de N_d que serão aplicados adiante na extração de características.

- **Durações da janela J e do quadro Q**: Os valores de J e Q usados na literatura possuem grande variabilidade. Em particular, podemos citar os seguintes trabalhos: Park [6] $J = 24$ ms e $Q = 5$ ms; Toledano [20] ($J = 24$ ms e $Q = 3$ ms); Charonnat [47] ($J = 32$ ms e $Q = 10$ ms); Mporas [48] ($J = 16$ ms e $Q = 5$ ms); Ting [49] ($J = 15$ ms e $Q = 5$ ms); Selmini [1] ($J = 20$ ms e $Q = 5$ ms). Com isto, observa-se uma variação da duração da janela na faixa de 15 a 32 ms com deslocamento de 1 a 10 ms.

  No presente trabalho, consideramos inicialmente um HMM de 3 estados e 2 gaussianas por mistura. Inicialmente, mantendo-se $Q = 1$ ms, variamos J de 15 a 24 ms, obtendo-se os resultados indicados na figura 5.3a. Neste experimento, obtivemos pouca variação no erro médio global para J entre 15 e 20 ms, com o menor valor obtido para $J = 17$ ms, e altos níveis de erro

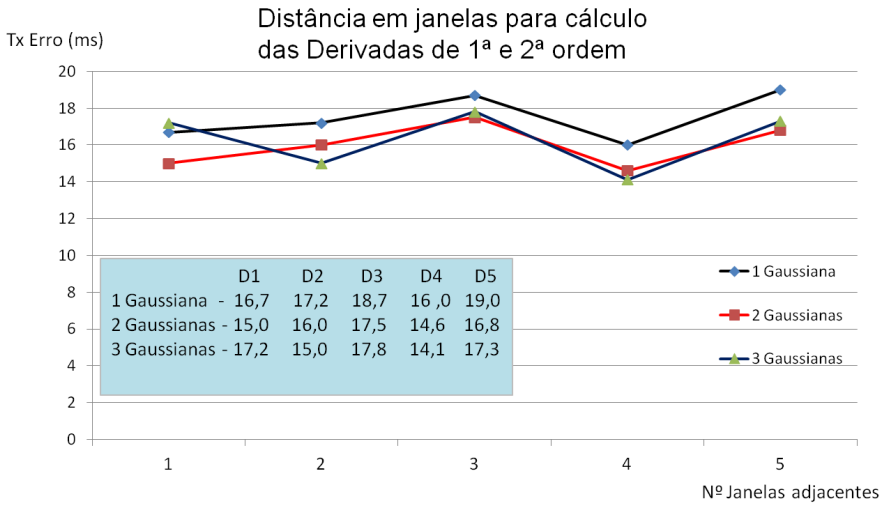
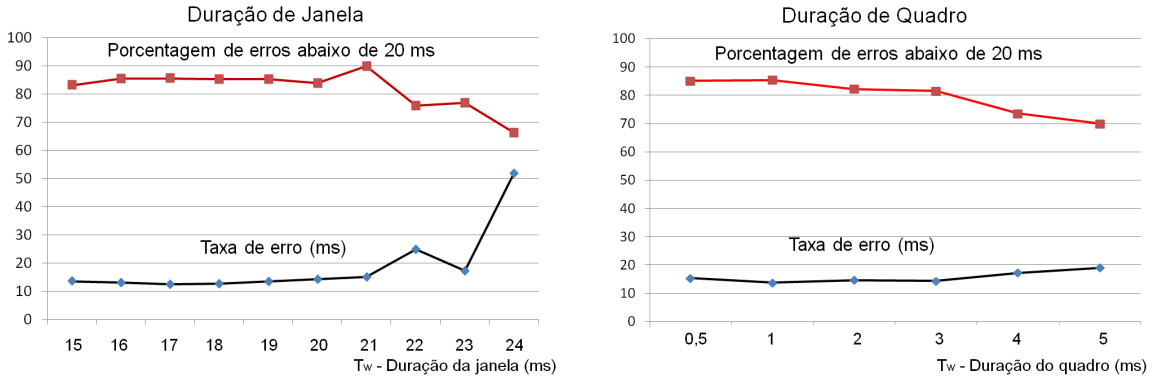


Figura 5.2: Erro absoluto médio para HMMs com diferentes valores de N_d.

para J acima de 20 ms. Numa segunda etapa, fixando $J = 17$ ms, variamos o deslocamento na faixa de 0,5 a 5 ms, gerando os resultados mostrados na figura 5.3b. Neste caso, foi verificado que do início da faixa até os 3 ms, os valores de erro médio permanecem relativamente próximos e baixos, com uma clara tendência de alta a partir de 4 ms.

Neste trabalho, os modelos HMMs "sementes"são treinados com uma pequena base de fones (usando apenas 50 locuções). Sendo assim, considerando que os parâmetros iniciais do HMM são fundamentais para a convergência do modelo [2], [50], optou-se por restringir os valores de J no intervalo de 15 a 20 ms e os valores de Q na faixa de 0,5 a 3 ms.



(a) (b)

Figura 5.3: Desempenho de sistema HMM para: (a) Diferentes durações J da janela (N3G2N_d2Q1); (b) Diferentes deslocamentos Q da janela (N3G2N_d2J17).

5.3.2 Parâmetros do HMM

O sistema proposto para esta tese inclui n HMMs treinados com diferentes parametrizações, gerando n estimativas das fronteiras que possam, após uma análise estatística adequada, gerar uma estimativa robusta do limiar desejado. Neste trabalho, os parâmetros HMM analisados são:

- **Número de estados** N: A escolha da faixa de variação do número de estados leva em consideração a quantidade de amostras que se tem para o treinamento do modelo. No caso, foi considerada a faixa de 1 a 5 estados para cada unidade fonética, cujos desempenhos (usando-se 2 misturas) é ilustrado na figura 5.4.

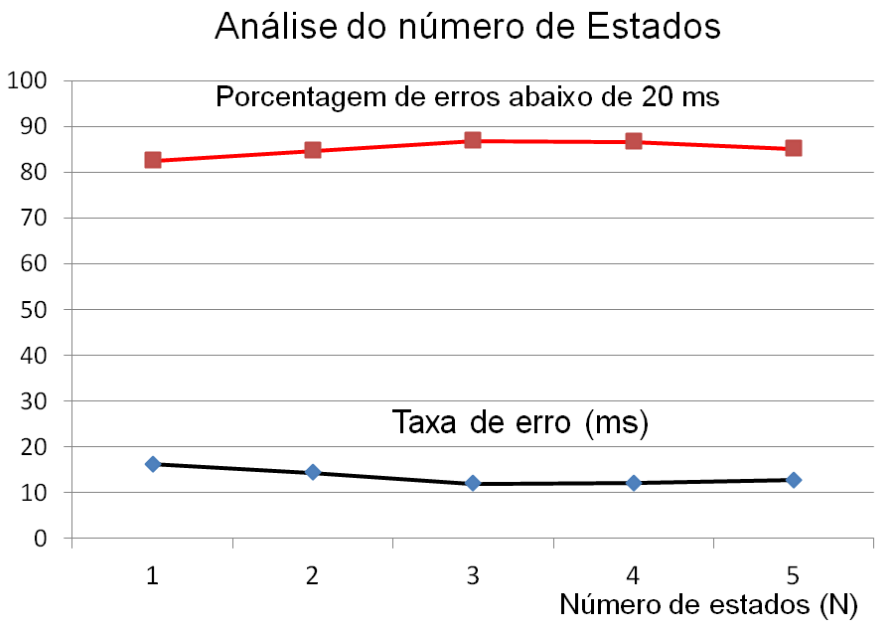


Figura 5.4: Desempenho do sistema HMM em função do número N de estados (G2N_d2Q1J15).

- **Número M de gaussianas**: O número de gaussianas em uma mistura também atende à condição especificada no item anterior. Neste trabalho, após vários testes, foi selecionada a faixa de 1 a 8 gaussianas por mistura, cujo desempenho é indicado na figura 5.5 (usando-se $N = 2$ estados). Desta figura, percebe-se um leve incremento do erro com o aumento do número das gaussianas, o que pode ser explicado pela diminuição no número de vetores de observações para cada gaussiana, o que empobrece a modelagem de seus parâmetros.

  Este problema é particularmente perceptível para fones de curta duração (em geral, as consoantes plosivas de uma língua) quando se usam muitos estados (no caso $N > 2$) e um número elevado de gaussianas por mistura. Para ilustrar isto, considere a modelagem dos fones plosivos usando $N = 2$ estados

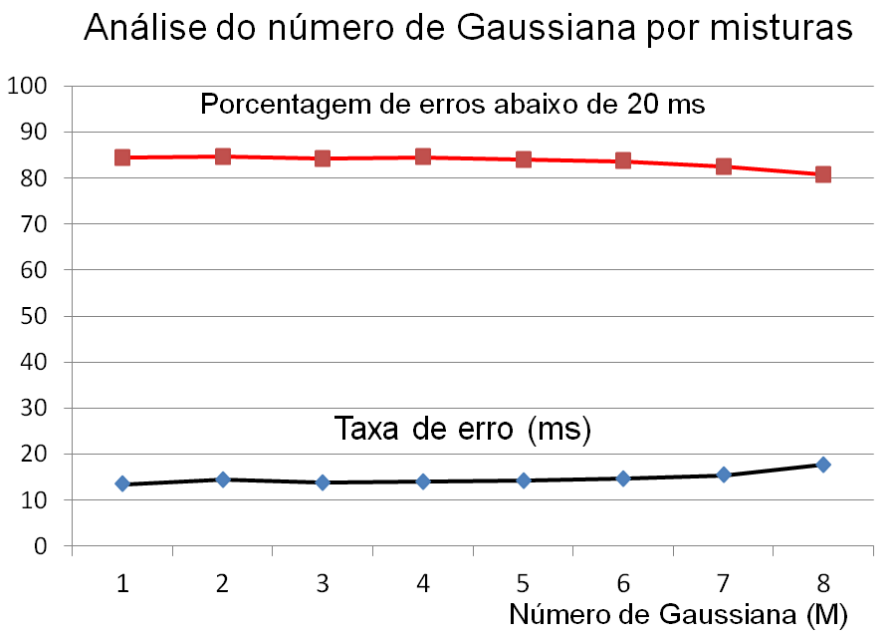


Figura 5.5: Desempenho do sistema HMM em função do número M de misturas.

e $1 \leq M \leq 8$ misturas. Neste processo, nota-se uma maior ocorrência de erros de modelagem de mais tipos de fonemas conforme M aumenta, como ilustrado na tabela 5.1. Durante a fase de treinamento dos modelos HMMs, foram observados erros de falta de amostras de observações durante a execução do algoritmo *k-means* em determinadas configurações. Assim, para os testes subsequentes, os valores dos parâmetros de modelagem do HMM foram: $N = 1$ e 2, com M variando de 1 a 8; para $N = 3$ com M de 1 a 3, e passo (Q) de $0,5$ ms e 1 ms; e, para $N = 4$ e 5, M varia apenas em 1 e 2. Demais parâmetros seguem definições anteriores.

Tabela 5.1: Erros de modelagem dos fones plosivos com o aumento do número de gaussianas.

M	Fones problemáticos
1	
2	
3	[p]
4	[p], [g]
5	[p], [g], [d]
6	[p], [g], [d]
7	[p], [g], [d], [b]
8	[p], [g], [d], [b], [t]

5.3.3 Treinamento e Seleção dos HMMs

O procedimento de treinamento *embedded*, visto no capítulo 4, realiza o treinamento de todos os HMMs em bloco. Assim, dada uma locução e sua transcrição fonética, cria-se uma estrutura sequencial de HMMs conectados, e a cada nova locução da base de treinamento, um novo HMM é obtido para cada unidade fonética especificada, conforme a transcrição correspondente. Neste processo, tem-se a definição das fronteiras desejadas, ou das coarticulações entre fones.

Como colocado anteriormente, 50 locuções são usadas para se determinar o HMM semente de cada unidade fonética. Em seguida, as 1026 locuções da base de treinamento são utilizadas para retreinar estes modelos pelo procedimento *embedded*. Ao final, testam-se estes modelos com 150 locuções da base de teste.

Para selecionar os HMMs que irão compor o sistema MHMM dentro das muitas combinações dos diferentes parâmetros acústicos ou da estrutura do HMM, selecionamos aquelas configurações que obtinham erro absoluto médio abaixo ou igual a 20 ms (valor aceitável para a segmentação automática), considerando todas as unidades, como indicado na figura 5.6. Na tabela 5.2, estão descritas as 34 configurações de HMM que obtiveram erro absoluto médio inferior a 20 ms.

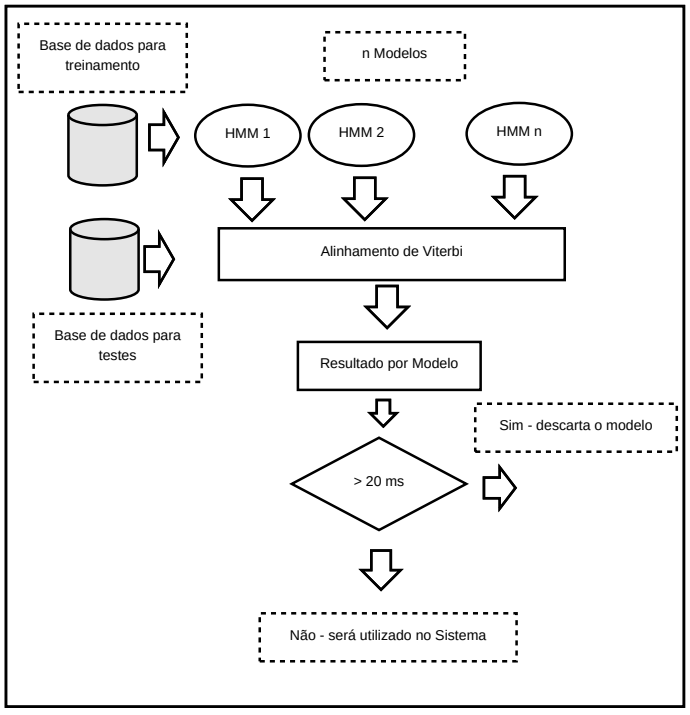


Figura 5.6: Ilustração do procedimento de verificação do limiar do erro na seleção dos HMMs individuais que irá compor o sistema MHMM proposto.

Tabela 5.2: Configurações de HMM com menor taxa de erro selecionadas para compor o sistema MHMM.

Configuração MHMM				
Estados (N)	Gaussianas (M)	Delta (N_d)	Janela (J)	Quadro (Q)
3	1	2	15	0.5
	1	2	16	0.5
	1	2	17	0.5
	2	2	17	0.5
	3	2	17	0.5
	1	2	17	1
	2	2	17	1
	3	2	17	1
	1	4	17	0.5
	2	4	17	0.5
	3	4	17	0.5
	1	4	17	1
	2	4	17	1
	3	4	17	1
	1	2	18	0.5
	2	2	18	0.5
	3	2	18	0.5
	1	2	18	1
	2	2	18	1
	3	2	18	1
	1	4	18	0.5
	2	4	18	0.5
	3	4	18	0.5
	1	4	18	1
	2	4	18	1
	1	2	20	0.5
	2	2	20	0.5
	3	2	20	0.5
	1	2	20	1
	1	4	20	0.5
	2	4	20	0.5
	3	4	20	0.5
	1	4	20	1
4	2	2	18	0.5

5.3.4 Caracterização dos Sistemas MHMM

Na sequência, os desempenhos das quatro configurações MHMM, descritas anteriormente, são apresentados com os respectivos HMMs selecionados em cada caso.

1. Para o sistema MHMM-1, o HMM descrito na tabela 5.3 obteve a menor MAE global, igual a 13,8 ms, dentre os 34 modelos selecionados anteriormente.

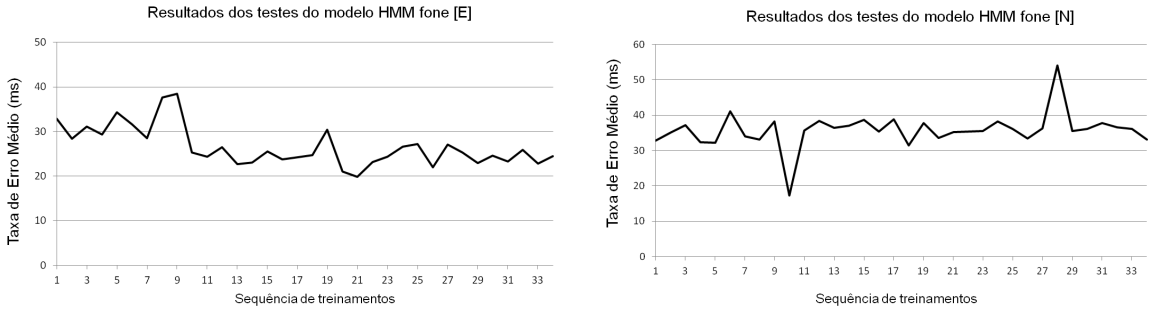
Tabela 5.3: Sistema MHMM 1 - Seleção do modelo com menor MAE global.

MHMM 1					
Estados (N)	Gaussianas (M)	Delta (N_d)	Janela (J)	Quadro (Q)	MAE
3	2	2	17	1	13,8 ms

2. O sistema MHMM-2 é formado por 38 HMMs independentes (incluído o silêncio), que obtiveram o menor MAE para cada unidade fonética, conforme descrito na tabela 5.4.

   Neste sistema deve-se obter um único limiar entre fones adjacentes a partir de uma dupla de limiares. Por exemplo, para a palavra *casa* serão 3 HMMs para estimar as fronteiras. O primeiro HMM_k estima as duas fronteiras da vogal "c", o segundo HMM_a estima 4 fronteiras, duas para cada vogal "a"e o último, o HMM_z, mais duas fronteiras da consoante "z". Desta forma existem duas fronteiras para uma mesma marcação. Aplica-se a média a cada dois limiares consecutivos obtendo a fronteira estimada.

   É interessante observar que alguns fones não são bem modelados pelo respectivo HMM. O fone [E] (presente na palavra t**E**la) e o fone [N] (na palavra tama**nh**a), por exemplo, apresentaram MAEs iguais a 19,87 ms e 17,35 ms, respectivamente. Para estes fones, o treinamento não apresentou uma tendência de diminuição da MAE, como mostram as figuras 5.7(a) e 5.7(b).



(a) (b)

Figura 5.7: MAE para os 34 HMMs treinados: (a) Fone [E]; (b) Fone [N].

Tabela 5.4: Sistema MHMM-2 usando HMMs ótimos para cada unidade fonética.

MHMM 2						
Fone	Estados (N)	Gaussianas (M)	Delta (N_d)	Janela (J)	Quadro (Q)	MAE (ms)
[a]	3	3	2	17	0,5	12,86
[e]	3	3	2	17	0,5	13,85
[E]	3	1	2	16	0,5	19,87
[i]	3	1	2	15	0,5	14,72
[y]	3	1	2	16	0,5	16,85
[o]	3	1	4	20	0,5	13,87
[O]	3	2	4	17	0,5	11,26
[u]	3	3	2	17	0,5	16,12
[an]	3	3	4	17	0,5	12,94
[en]	3	2	4	17	1	10,64
[in]	3	1	2	15	0,5	12,95
[on]	3	1	2	17	1	7,36
[un]	3	3	4	17	0,5	18,76
[b]	3	3	4	17	0,5	6,74
[d]	3	1	2	15	0,5	4,38
[D]	3	2	4	20	0,5	14,78
[f]	3	3	2	17	0,5	6,39
[g]	3	3	2	17	0,5	11,46
[j]	3	1	2	18	0,5	19,19
[k]	3	2	4	20	0,5	8,03
[l]	3	1	4	20	0,5	12,42
[L]	3	3	2	20	0,5	12,76
[m]	3	2	2	17	0,5	9,55
[n]	3	3	2	17	0,5	9,64
[N]	3	1	2	17	0,5	17,35
[p]	3	2	4	20	0,5	5,71
[r]	3	2	2	17	0,5	11,45
[rr]	3	3	2	18	0,5	16,53
[R]	3	3	2	17	0,5	10,91
[s]	3	2	4	18	0,5	13,11
[t]	3	2	4	20	0,5	7,36
[T]	3	2	4	20	0,5	11,96
[v]	3	1	2	20	1	10,72
[x]	3	1	2	18	0,5	6,96
[z]	3	1	2	18	0,5	9,24
[cl]	3	2	4	20	0,5	12,34
[vcl]	3	3	2	17	0,5	10,03

3. Para o sistema MHMM-3, são selecionados os $n = 7$ melhores HMMs (com menor MAE global), caracterizados na tabela 5.5, para compor o MHMM. Para as $n = 7$ estimativas de limiar, aplica-se uma análise estatística simples (por média ou mediana) para se obter o limiar final de segmentação.

Tabela 5.5: Sistema MHMM-3 com $n = 7$ HMMs.

MHMM 3					
Estados (N)	Gaussianas (M)	Delta (N_d)	Janela (J)	Quadro (Q)	MAE
3	2	2	17	1	13,8 ms
3	1	2	17	1	14,5 ms
3	2	4	17	1	14,60 ms
3	1	4	17	1	14,61 ms
3	1	2	18	1	14,63 ms
3	2	2	18	1	14,8 ms
3	1	4	17	0.5	15,1 ms

4. E, no sistema MHMM-4, assim como no MHMM-3, são usados $n = 7$ HMMs. Neste caso, porém, o conjunto de $n = 7$ HMMs é distinto para cada classe de unidade fonética, cujos desempenhos em termos da taxa de erro absoluto médio são indicados nas figuras 5.8 e 5.9.

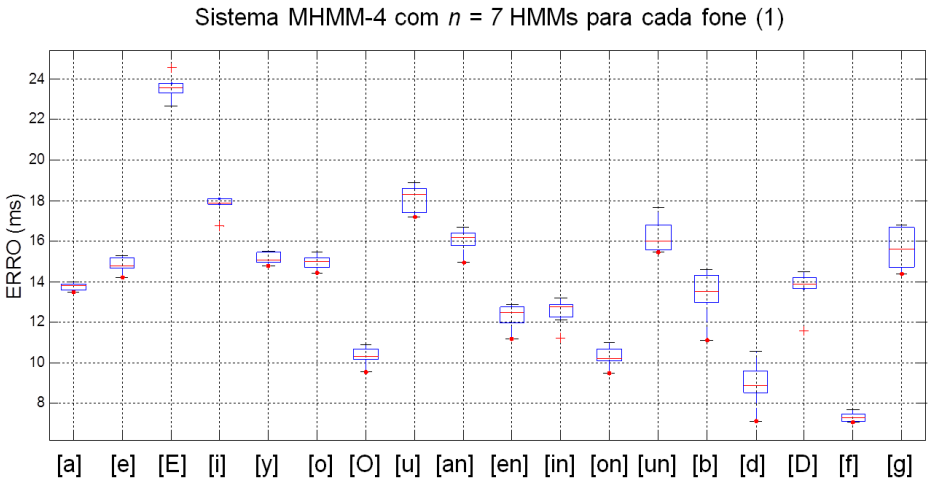


Figura 5.8: Valor de MAE do sistema MHMM-3 para cada fone (parte 1).

5.3.5 Desempenhos dos Sistemas MHMM

O desempenho médio dos diferentes sistemas MHMM para a base de teste de 150 locuções é sumarizado na coluna "Estimado" da tabela 5.6. Neste teste, a configuração MHMM-4 apresentou o melhor resultado médio MAE, com vantagem do

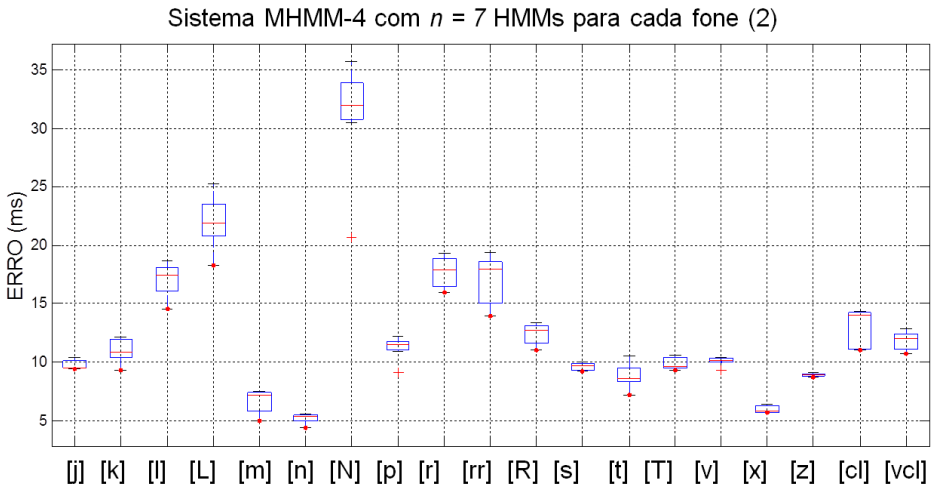


Figura 5.9: Valor de MAE do sistema MHMM-3 para cada fone (parte 2).

uso do operador *média* sobre o operador *mediana* no processamento das $n = 7$ marcações individuais.

Tabela 5.6: Avaliação dos Sistemas MHMM, para a base de testes, nas etapas estimação e refinamento com regras fonéticas. Valores em MAE (ms).

Análise do Estimador MHMM		
Sistemas MHMM	Estimado	Refinado
MHMM-1	17,07	15,39
MHMM-2	15,85	14,85
MHMM-3 (Média)	15,98	14,27
MHMM-3 (Mediana)	15,79	14,28
MHMM-4 (Média)	15,05	10,90
MHMM-4 (Mediana)	15,21	13,84

5.4 Desempenho do Refinamento por Regras Fonéticas

- Para a classe fricativa surda, o melhor resultado sem refinamento foi obtido para a configuração MHMM-4, com MAE de 8,10 ms, enquanto o erro com o uso do refinamento foi de 13,8 ms, conforme a tabela 5.7. O bom resultado do sistema MHMM-4 pode ser creditado à alta duração média e à relativa estacionaridade característica da classe fricativa surda, o que propicia uma melhor modelagem estatística de suas unidades.

  Para a classe fricativa sonora, o melhor resultado foi obtido pela média entre o ponto de separação dado pela distância BIC com o limiar dado pelo MHMM. Com essa fusão o MAE foi de 9,20 ms, enquanto o MAE pelas regras fonéticas foi de 14,20 ms. Para comparação os valores individuais obtidos por MHMM e distância BIC foram, respectivamente, de 9,8 ms e 10 ms. Como os fones da classe fricativa surda, as fricativas sonoras também possuem longa duração e relativa estacionaridade, facilitando a modelagem por HMM.

Tabela 5.7: Comparação entre métricas utilizadas no refinamento (erro ms).

Classe Fonética: Consoantes Fricativas		
Métricas	Surda	Sonora
MHMM-4	**8,10**	9,80
Regras	13,80	14,20
Fusão BIC / MHMM-4	8,30	**9,20**

Pelas tabelas 5.8 e 5.9 as porcentagens de erro para valores abaixo de 20 ms desta classe, tem apresentado resultados satisfatórios com 93,89 % para surda e 91,74% para as fricativas sonora.

Tabela 5.8: Distribuição dos erros para a classe fricativa.

Classe Fonética: Consoantes Fricativas Surda							
Tipo	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
Surda	46,87	75,84	**93,89**	97,53	97,82	98,54	99,56

- No caso das consoantes africadas, conforme indicado na tabela 5.10, o resultado obtido pelo sistema MHMM-4 foi de 14,20 ms e taxa de erro abaixo de 20 ms de 86,75 % enquanto pelas regras fonéticas foi melhor com valor de 12,5 ms. Para esta classe um teste realizado aplicando a média entre os limiares obtidos pela taxa de cruzamento de zeros e centróide, foi obtido 11,4 ms de erro.

Tabela 5.9: Distribuição dos erros para a classe fricativa.

Classe Fonética: Consoantes Fricativas Sonora							
Tipo	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
Sonora	43,75	69,20	**91,74**	96,43	97,54	98,66	99,78

O melhor resultado, foi obtido usando-se a distância BIC em conjunto com o MHMM-4, obtendo uma taxa de erro de 12 ms, o qual foi muito próximo do obtido pelas regras, de 12,5 ms. A tabela 5.11 compara entre as métricas utilizadas para esta classe. Somente a distância BIC o resultado foi de 15,3 ms.

Tabela 5.10: Distribuição dos erros para a classe das consoantes africadas.

Classe Fonética: Consoante Africada							
Métricas	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
MHMM-4	26,51	50,20	**86,75**	89,96	93,57	95,58	99,20
Regras	34,14	72,29	**86,35**	91,97	93,17	95,18	99,20
Mista (Fusão BIC+MHMM-4)	36,55	70,28	**86,35**	91,57	94,78	96,79	99,20

Tabela 5.11: Comparação entre métricas utilizadas no refinamento (erro ms).

Classe Fonética: Consoantes Africadas	
Métricas	Africadas
MHMM-4	14,20
Regras	12,50
Mista (Fusão BIC / MHMM-4)	**12,00**

- No refinamento para a classe das consoantes laterais e vibrantes, o uso das regras fonéticas levou a um MAE de 17,00 ms para consoantes laterais e 15,30 ms para vibrantes, enquanto que para o sistema MHMM-4 atinge-se o valor de 16,10 ms e 11,20 ms com taxa de erro abaixo de 20 ms de 80,81 % e 87,69 %, respectivamente, conforme detalhado nas tabelas 5.12 e 5.13.

- Para a classe das consoantes nasais, o MHMM-4 consegue modelar o sinal com MAE de 11,2 ms, conforme detalhado na tabela 5.14, o que representa um desempenho melhor do que o obtido pelas regras.

- No caso das transições de outras classes para uma consoante plosiva, os limiares gerados pelas regras fonéticas ficaram com MAE de 13,8 ms, melhor resultado

Tabela 5.12: Comparação entre métricas utilizadas no refinamento (erro ms).

Classe Fonética: Laterais e Vibrantes		
Métricas	Laterais	Vibrantes
MHMM-4	**16,10**	**11,20**
Regras	17,00	15,30

Tabela 5.13: Distribuição dos erros para a classe das consoantes laterais e vibrantes.

Classe Fonética: Laterais e Vibrantes							
Classe	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
Laterais (**16,10 ms**)	32,83	58,08	**80,81**	87,88	91,92	92,93	97,47
Vibrantes (**11,20 ms**)	42,09	68,84	**87,60**	94,45	96,57	97,88	98,86

Tabela 5.14: Distribuição dos erros para a classe das consoantes nasais.

Classe Fonética: Consoante Nasal							
Métricas	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
MHMM-4 (**11,2 ms**)	51,04	75,07	**88,12**	91,38	93,73	94,65	98,96
Regras (**14,3 ms**)	31,33	56,01	**80,03**	90,73	93,99	95,17	98,83

obtido do que pelo MHMM-4, cujo desempenho foi na faixa de 9,50 ms.

Considerando todas as transições de/para a classe plosiva, as regras fonéticas foram capazes de definir a fronteira muito melhor do que MHMM-4 para a plosiva surda. No caso, para plosiva sonora, os resultados foram próximos, 7,5 ms para MHMM-4 e 7,6 ms para regras.

Os resultados gerais para a classe das plosivas é detalhado na tabela 5.16, mostrando acerto abaixo de 20 ms de 97,87 % e 96,33 %, respectivamente para surda e sonora.

Tabela 5.15: Comparação entre métricas utilizadas no refinamento (erro ms).

Classe Fonética: Plosiva		
Métricas	Surda	Sonora
MHMM-4	9,50	**7,50**
Regras	**7,50**	7,60

- Para a classe das vogais têm-se as seguintes conclusões, sumarizado na tabela 5.17:
  - Média: Esta classe obteve o seu melhor resultado com o MHMM-4 um

Tabela 5.16: Distribuição dos erros para a classe das consoantes plosivas.

Classe Fonética: Plosiva							
	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
Surda	46,39	77,16	96,57	**98,46**	98,93	99,17	99,53
Sonora	48,68	78,21	**96,13**	97,96	98,37	99,39	99,80

MAE de 23,2 ms, não apresentado neste tipo de transição valores abaixo de 20 ms.

- Anterior/Posterior: A taxa de erro para esta classe foi de 22,8 ms quando usando a métrica sobre os formantes $F1$ e $F2$.
- Anterior: Para este grupo, obteve-se a taxa de erro de 14,6 ms com a aplicação das regras fonéticas.
- Posterior: Neste caso, a taxa de erro foi de 37,0 ms, quando aplicado formantes, neste caso, para a fusão das medidas de distâncias BIC e Bhattacharyya o resultado é melhor com 26,9 ms, porém, abaixo de 20 ms a taxa de acertos foi de 42,86 %, enquanto para formantes foi de 75% abaixo de 20 ms. Assim, optou-se pela definição da métrica formantes.
- Transições entre vogais: para este tipo de transição a menor taxa de erro consegue-se com o uso das regras fonéticas, somente para as transições das vogais posteriores ou anteriores para as médias (ou vice-versa) que o resultado com o MHMM-4 obteve menor valor de MAE;
- Outras classes para vogais (e vice-versa): A tabela 5.18 mostram os resultados considerando todas as transições de/para as vogais.

Tabela 5.17: Comparação entre métricas utilizadas no refinamento (erro ms).

Classe Fonética: Vogais				
Métricas	Média	Anterior-Posterior	Anterior	Posterior
MHMM-4	23,2	23,8	29,4	29,5
Regras	24,9	22,8	14,6	30,1
Fusão F1+ F2	25	24,1	27,8	-
Fusão BIC + Bhatt	-	27,2	23,9	26,9

5.5 Desempenho do Sistema Misto

O desempenho do sistema usando a técnica MHMM-4 é sumarizado para cada classe fonética na tabela 5.19, o que pode ser diretamente comparado com os resulta-

Tabela 5.18: Distribuição dos erros para a classe das vogais.

Classe Fonética: Vogais								
Classes	MAE (ms)	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
Anterior	13,4	41,47	66,06	**83,61**	90,31	93,38	94,70	98,51
Média	11,1	47,08	72,29	**87,15**	92,52	95,40	96,26	98,85
Posterior	12,1	42,28	67,13	**85,27**	91,18	93,89	95,79	99,20
Nasal	8,1	51,33	80,96	**92,77**	94,46	96,63	98,07	99,76

dos de [1] antes da etapa de refinamento pelas regras fonéticas. Os resultados de ambos os sistemas são comparáveis, com significativas vantagens do sistema proposto nas classes das vogais e consoantes plosivas surdas, fricativas surdas, laterais e v2ibrantes.

Tabela 5.19: MAE do sistema MHMM.

	Limiar (ms)						
Classe Fonéticas	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
Silêncio	22,22	39,06	77,44	92,26	95,62	96,97	98,65
Vogal Anterior	40,65	66,23	83,11	89,65	93,05	94,29	98,34
Vogal Média	45,35	70,18	87,06	92,33	95,49	96,45	99,04
Vogal Posterior	42,08	65,63	84,17	90,68	93,69	95,69	99,00
Vogal Nasalizada	51,33	79,76	92,53	94,22	96,63	98,07	99,76
Plosiva Sonora	48,68	78,21	96,13	97,96	98,37	99,39	99,80
Plosiva Surda	39,29	55,50	97,40	98,93	99,17	99,29	99,65
Fricativa Surda	46,87	75,84	93,89	97,53	97,82	98,54	99,56
Fricativa Sonora	37,05	62,72	89,51	96,21	99,11	99,55	100,00
Africada	26,51	50,20	86,75	89,96	93,57	95,58	99,20
Consoante Nasal	51,04	75,07	88,12	91,38	93,73	94,65	98,96
Laterais	32,83	58,08	80,81	87,88	91,92	92,93	97,47
Vibrantes	42,25	68,84	87,44	94,45	96,57	97,88	98,86

O desempenho do sistema proposto com o uso das regras fonéticas é sumarizado na tabela 5.21, enquanto que o desempenho do sistema misto, que ativa ou não as regras fonéticas, é descrito na tabela 5.22. Destes conjuntos de dados, é simples perceber as vantagens de se usar o sistema misto, que explora as vantagens de cada uma das técnicas de refinamento.

5.6 Conclusão

Este capítulo apresentou os principais resultados experimentais do sistema proposto. Inicialmente, foi apresentado o desempenho da etapa de refinamento por mútiplos

Tabela 5.20: MAE antes do refinamento na tese de Selmini [1].

	Limiar (ms)						
Classe Fonéticas	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
Silêncio	24,25	53,25	82,25	95,75	98,00	99,25	100
Vogal Anterior	23,59	30,74	61,17	80,91	88,35	92,02	99,03
Vogal Média	3,88	10,72	46,52	83,12	93,16	96,01	99,32
Vogal Popsterior	8,25	17,54	51,57	74,63	84,97	91,23	98,33
Vogal Nasalizada	10,81	27,84	50,54	70,27	80,81	88,11	99,73
Plosiva Surda	31,45	55,91	74,88	80,87	86,36	90,68	99,00
Plosiva Sonora	40,00	69,30	90,70	94,93	98,11	99,72	100
Fricativa Surda	29,09	51,64	77,64	87,45	91,82	94,18	98,91
Fricativa Sonora	41,94	89,11	96,37	98,39	100	100	100
Africada	56,98	73,84	88,95	96,51	97,67	98,84	100
Consoante Nasal	54,42	74,36	87,75	90,88	94,02	96,30	99,72
Laterais	30,57	44,59	54,14	62,42	68,79	77,71	98,73
Vibrantes	22,75	37,75	70,50	84,00	87,75	91,00	98,75

Tabela 5.21: MAE do sistema MHMM após o uso das regras fonéticas.

	Limiar (ms)						
Classe Fonéticas	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
Silêncio	18,18	60,94	75,76	86,53	92,59	95,62	98,99
Vogal Anterior	31,04	56,29	77,65	88,66	93,05	94,78	98,51
Vogal Média	33,65	55,13	78,14	88,11	94,25	96,26	98,66
Vogal Posterior	36,67	57,41	78,06	87,47	92,69	95,89	99,10
Vogal Nasalizada	40,96	62,17	78,31	91,81	96,63	98,55	99,76
Plosiva Sonora	52,75	78,82	94,30	97,56	98,17	99,39	99,80
Plosiva Surda	46,39	77,16	96,57	98,46	98,93	99,17	99,53
Fricativa Surda	27,37	49,05	75,98	91,12	97,09	98,98	99,56
Fricativa Sonora	25,89	44,87	75,67	90,18	95,31	98,44	99,55
Africada	34,14	72,29	86,35	91,97	93,17	95,18	99,20
Consoante Nasal	31,33	56,01	80,03	90,73	93,99	95,17	98,83
Laterais	32,32	59,60	74,24	86,36	89,90	93,94	97,98
Vibrantes	21,21	46,82	79,28	89,72	95,92	97,39	98,86

HMMs, considerando quatro configurações distintas. Neste estudo, o melhor desempenho, com um MAE de aproximadamente 10,9 ms para a base de teste, foi obtido para a configuração que gera a fronteira desejada pela mediana de 7 HMMs para cada classe fonética. Em seguida foi considerado o refinamento por regras fonéticas, onde se discutiu o desempenho para cada classe fonética. Por fim, consideramos o sistema misto, combinando as técnicas de refinamento por MHMMs e por regras fonéticas e procurando otimizar o desempenho global explorando os bons desempenhos individuais de cada técnica. O sistema misto consolidado atingiu uma MAE

Tabela 5.22: MAE do sistema misto com seleção dos módulos MHMM e/ou regras fonéticas.

	Limiar (ms)						
Classe Fonéticas	≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100
Silêncio	18,18	60,94	75,76	86,53	92,59	95,62	98,99
Vogal Anterior	41,47	66,06	83,61	90,31	93,38	94,70	98,51
Vogal Media	47,08	72,29	87,15	92,52	95,40	96,26	98,85
Vogal Posterior	42,28	67,13	85,27	91,18	93,89	95,79	99,20
Vogal Nasalizada	51,33	80,96	92,77	94,46	96,63	98,07	99,76
Plosiva Sonora	49,69	80,04	96,33	97,96	98,37	99,39	99,80
Plosiva Surda	40,83	59,88	97,87	98,93	99,17	99,29	99,65
Fricativa Surda	47,02	75,69	93,89	97,53	97,82	98,54	99,56
Fricativa Sonora	43,75	69,20	91,74	96,43	97,54	98,66	99,78
Africada	36,55	70,28	86,35	91,57	94,78	96,79	99,20
Consoante Nasal	51,17	75,07	87,99	91,38	93,73	94,65	98,96
Laterais	32,83	58,08	80,81	87,88	91,92	92,93	97,47
Vibrantes	42,09	68,84	87,60	94,45	96,57	97,88	98,86

menor ou igual a 20 ms em 88,60% dos segmentos da base de teste.

Capítulo 6

Conclusão

6.1 Introdução

Nesta tese abordamos todas as etapas de um sistema de segmentação do sinal da fala, com ênfase na busca por parâmetros acústicos e de modelagem do HMM. A seguir são apresentadas uma descrição das propostas e contribuições contidas na tese e sugestões para trabalhos futuros.

6.2 Considerações Finais

Ao longo deste trabalho foi apresentado e desenvolvido um sistema que segmenta o sinal da fala em unidades acústicas e que proporcionou a produção de três artigos [25], [51] e [52]. Foi abordado, no segundo capítulo, uma aplicação para uma base de segmentos da fala, a síntese concatenativa da fala. Neste contexto, as unidades acústicas segmentadas são concatenadas conforme sua transcrição fonética, gerando o sinal de voz sintetizado, com prosódia determinada pelo método TD-PSOLA, que perfaz a manipulação dos contornos de pitch, duração e intensidade das devidas unidades fonéticas.

O capítulo 3, detalhou o problema da segmentação da fala, com ênfase na combinação de diferentes técnicas com o objetivo de aumentar a robustez do sistema, tais como Selmini [1], Park [6], Ting [49] e Adell [53]. Uma das técnicas básicas de segmentação é baseada nos modelos ocultos de Markov (HMM, do inglês *continuous hidden Markov models*) que gera as fronteiras à esquerda e à direita das unidades acústicas.

São propostos dois critérios de busca dos melhores parâmetros, dentro da faixa estipulada, que modelem o HMM na representação do sinal de fala. Os critérios de busca: *Critério de Treinamento Exaustivo* e *Critério de Treinamento Fones por Estados*. Para o critério treinamento exaustivo, variam-se os parâmetros número de

estados *(N)*, número de gaussianas *(M)*, coeficientes deltas (N_d) e a dimensão do vetor de características *(D)*, enquanto que no critério fones por estados, o número de estados e a dimensão do vetor de características são mantidos fixos, os demais, *(M)* e (N_d) são variáveis. Realiza-se o treinamento e testes com a base Yoho (língua inglesa), o *critério de treinamento fones por estados*, apresentou-se melhor, com erro médio de 7,7 ms enquanto o outro critério o foi de 9,7 ms de erro médio total. Dessa forma, pode-se inferir que o critério de fones por estados define melhor os parâmetros que modelam o HMM, e, por isso, foi utilizado na implementação do algoritmo MHMM proposto.

Com o propósito de avaliar técnicas de segmentação, ainda no capítulo 3, foi considerado o algoritmo utilizando múltiplos HMMs que geram diferentes estimativas iniciais dos limiares de segmentação. Numa etapa posterior, estas estimativas são avaliadas por método de cálculo de tendência (média ou mediana) para gerar uma nova estimativa do limiar. O processo de segmentação pode considerar ainda a aplicação das regras fonéticas no refinamento dos limiares das fronteiras para obter uma estimativa mais precisa ainda dos limiares de segmentação.

Assim, no capítulo 4, consideramos a combinação das técnicas de MHMM e das regras fonéticas, procurando explorar as vantagens destas modelagens estatística e determinística, respectivamente. Para este modelo misto foram testados diferentes formas de inicialização do HMM, resultando no uso de modelos-"semente"treinados por um subconjunto de fones segmentados manualmente. Uma instigante questão é a analise para obter uma métrica fonética que encontre a fronteira, por fusão de parâmetros, e que agregue informações da espontaneidade da fala, e não apenas um limiar determinístico e igual para todas as n repetições de um mesmo fone ou classe fonética. Neste sentido, algumas métricas fonéticas foram analisadas [1]. Foi proposto o uso de medida de distância para a separação de classes fonéticas, utilizando as características taxa de cruzamento de zeros e centróide. Assim, as medidas de distâncias por Bhattacharyya e BIC foram implementadas nas classes fonéticas, obtendo resultados satisfatórios. Outra métrica fonética avaliada foi o uso da duração média dos fones. Neste caso, aplicou-se em todas as classes, entretanto, somente as transições a partir das classes fricativa e plosiva apresentaram com descartes de fones devido ao tamanho muito reduzido. Nas demais classes não houve ganho para o sistema. Adotou-se a métrica de descarte somente para o limiar inferior da duração mínima do fone.

Finalmente, no capítulo 5, apresentamos os resultados experimentais alcançados pelo sistema proposto. Para a implementação final, foi necessário identificar a faixa de variação para extração dos parâmetros acústicos e para a modelagem dos HMMs, permitindo iniciar do treinamento. Após a seleção das variações paramétricas, foram identificadas 34 configurações de HMM que serviram de base para o sistema MHMM

proposto. Neste sentido, foram consideradas 4 diferentes configurações para o sistema MHMM e o uso das regras fonéticas. O sistema misto consolidado atingiu uma MAE menor ou igual a 20 ms em 88,67% dos segmentos da base de teste, conforme indicado na tabela 6.1.

Tabela 6.1: Distribuição dos erros para o sistema após o refinamento.

Resultado							
≤ 5	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 100	Erro Médio
43,53	69,69	88,67	93,42	95,59	96,77	99,13	10,90

6.3 Propostas Futuras

Neste trabalho a discussão é estimar limiares do sinal da fala usando regras determinísticas em conjunto com técnica estatísticas e, durante a pesquisa e o desenvolvimento do sistema, foram surgindo alternativas ou indagações que não foram implementadas. Estes pontos são sugeridos como propostas futuras.

- O sistema, na fase de estimação do limiar, tem se mostrado robusto às variações de determinadas classes fonéticas, como as classes africadas e fricativas. Neste sentido, devemos investigar o porquê deste comportamento comparando com o comportamento obtido para as outras classes a fim de otimizar o desempenho do sistema como um todo.

- Observando que os HMMs têm gerado melhor desempenho quando operando com poucos estados (2 ou 3) por fone e um número maior de gaussianas, é sugerido avaliar o uso de modelos GMM (do inglês *Gaussian mixture model*) na estimação de limiares iniciais ou na fase de refinamento.

- Avaliar a técnica proposta em outras bases do português brasileiro, possivelmente com outros sotaques ou mesmo com locutores do sexo feminino, a fim de analisar a robustez dos métodos aqui analisados em outros contextos linguísticos.

- O sistema utilizou ferramentas para separação dos vetores acústicos em grupos identificando modelos distintos de comportamento. Assim, outras metodologias poderiam ser aplicadas nesta etapa, como por exemplo, máquinas de suporte vetorial (SVM, do inglês *support vector machines*) ou outras técnicas de classificação.

Referências Bibliográficas

[1] SELMINI, A., M. *Sistema Baseado em Regras para o Refinamento da Segmentação Automática da Fala*. Tese de doutorado, UNICAMP, Campinas, SP, Brasil, 2008.

[2] RABINER, L. R., JUANG, B. H. *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA, Prentice-Hall, Inc, 1993.

[3] WANG, D., LU, L., ZHANG, H.-J. "Speech Segmentation Without Speech Recognition". In: *IEEE Trans. ASSP*, v. I, pp. 405–408, Apr. 2003.

[4] CALLOU, D., LEITE, Y. *Iniciação à Fonética e à Fonologia*. Rio de Janeiro,RJ, Jorge Zahar, 1995.

[5] O'SHAUQHNESSY, D. "Modern Methods of Speech Synthesis", *Circuits and Systems Magazine, IEEE*, v. 7, n. 3, pp. 6–23, 1ªquarter 2007.

[6] PARK, S. S., KIM, N. S. "On Using Multiple Models for Automatic Speech Segmentation", *Audio, Speech, and Language Processing, IEEE Transactions on*, v. 15, n. 8, pp. 2202 –2212, Nov. 2007.

[7] VIOLARO, F., BARBOSA, P. A., ALBANO, E., et al. "Um Conversor Texto-Fala para o Português Brasileiro com Processamento Linguístico de Alta Qualidade". In: *XIV Simpósio Brasileiro de Telecomunicações (SBrT 1996)*, pp. 361–366, Jul. 1996.

[8] FERREIRA, A. *Minidicionário Aurélio*. Nova Fronteira, 1977.

[9] LATSCH, V. *Construção de Banco de Unidades para a Síntese da fala por Concatenação no Domínio Temporal*. Dissertação de mestrado, UFRJ, Rio de Janeiro, Brasil, Abr. 2005.

[10] BARBOSA, P. A. "Estudos de Prosódia". cap. Revelar a Estrutura Rítmica de uma Língua Construindo Máquinas Falantes: pela Integração entre Ciência e Tecnologia de Fala, pp. 21–52, Editora da UNICAMP, 1999.

[11] FIGUEIREDO, F. L. *Segmentação Automática e Treinamento Discriminativo Aplicados a um Sistema de Reconhecimento de Dígitos Conectados*. Tese de doutorado, UNICAMP, Campinas, SP, Brasil, 2000.

[12] SIMÕES, F. S. *Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil*. Tese de doutorado, UNICAMP, Campinas, SP, Brasil, 1999.

[13] MOULINES, E. *Algoritmes de cadage et de Modification das Paramètres Prosodiques pour La Synthèse de La Parole à partir Du Texte*. Ph.d. thesis, École National Supérieure des Télécommunications, Paris, 1990.

[14] WANG, W. S.-Y., PETERSON, G. "Segment inventory for speech synthesis", *Journal of the Acoustical Society of America*, v. 30, pp. 743–746, Aug. 1958.

[15] SOLEWICZ, J., MORAES, J., ALCAIM, A. "Text-to-speech System for Brazilian Portuguese using a Reduced Set of Synthesis Units". In: *International Symposium on Speech, Image Processing and Neural Networks*, pp. 579 – 582, Hong Kong, Apr. 1994.

[16] JARIFI, S., PASTOR, D., ROSEC, O. "A Fusion Approach for Automatic Speech Segmentation of Large Corpora with Application to Speech Synthesis", *Speech Communication*, , n. 50, pp. 67–80, 2008.

[17] YOUNG, S., OTHERS. *The HTK Book*, 3.4 ed. Cambridge University Engineering Department, 2006.

[18] PARK, S., KIM, N. "Automatic Segmentation Based on Boundary Type Candidate Selection", *IEEE Signal Processing Letters*, v. 13, n. 10, pp. 640–643, Oct. 2006.

[19] KAWAI, H., TODA, T. "An Evaluation of Automatic Phone Segmentation for Concatenative Speech Synthesis". In: *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, v. 1, pp. 677–680, May 2004.

[20] TOLEDANO, D. T., GOMEZ, L. A. H., GRANDE, L. V. "Automatic Phonetic Segmentation", *IEEE Transactions on Speech and Audio Processing*, v. 11, n. 6, pp. 617–625, Nov. 2003.

[21] WANG, L., ZHAO, Y., CHU, M., et al. "Refining Segmental Boundaries for TTS Database using Fine Contextual-Dependent Boundary Models". In:

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), v. 1, pp. 641–644, May 2004.

[22] DEMUYNCK, K., LAUREYS, T., GILLIS, S. "Automatic Generation of Phonetic Transcriptions for Large Speech Corpora", *Proc. ICSLP, Denver*, v. 1, pp. 333– 336, 2002.

[23] HOSOM, J.-P. *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. Ph.d. thesis, Oregon Graduate Institute of Science and Technology, May 2000.

[24] LEE, K. "MLP-based Phone Boundary Refinement for a TTS Database", *Audio, Speech, and Language Processing, IEEE Transactions on*.

[25] PARANAGUÁ, E., NETTO, S. "Aplicação de Múltiplos HMMs com Refinamento para a Segmentação Automática da Fala". In: *Anais do XXVII Simpósio Brasileiro de Telecomunicações (SBrT 2009)*, Blumenau, SC, Set.-Out. 2009.

[26] DELLER, J. R., PROAKIS, J. G., HANSEN, J. H. L. *Discrete-time Processing of Speech Signals*. 1ª ed. Englewood Cliffs, Macmillan Publishing Company, 1993.

[27] PICONE, J. "Signal Modeling Techniques in Speech Recognition", *Proceedings of the IEEE*, v. 81, n. 09, pp. 1215–1246, Sep. 1993.

[28] JUNEJA, A. "Segmentation of Continuous Speech using Acoustic-phonetic Parameters and Statical Learning". In: *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'O2)*, v. 2, pp. 726–730, Nov. 2002.

[29] JUNEJA, A. *Speech Recognition Based on Phonetic Features and Acoustic Landmarks*. Ph.d. thesis, University of Maryland, College Park, USA, 2004.

[30] YNOGUTI, C. A. *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*. Tese de doutorado, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, 1999.

[31] UOL-EDUCAÇÃO, S. "Uol-Educação: Disciplina Português", *http://educacao.uol.com.br/portugues/ult1693u47.jhtm, (último acesso Agosto/2011)*.

[32] PCI-CONCURSOS, S. "Aulas de Português - Tipos de Fonemas", *http://www.pciconcursos.com.br/aulas/portugues/tipos-de-fonemas, (último acesso Agosto/2011)*.

[33] FRANCISCO, S. P. S. "Fonética - Estudos da Língua Portuguesa", *http://www.portalsaofrancisco.com.br/ alfa/fonetica/fonetica-3.php, (último acesso Agosto/2011)*.

[34] ARAÚJO, A. M. L. *Jogos Computacionais Fonoarticulatórios para Crianças com Deficiência Auditiva*. Tese de doutorado, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, 2000.

[35] CEGALLA, D. P. *Novíssima Gramática da Língua Portuguesa*. 17ª ed. São Paulo, Companhia Editora Nacional, 1993.

[36] OLIVEIRA, D. H. "Fonética e Fonologia", *http://portal.virtual.ufpb.br/biblioteca-virtual/files/pub _1291085011.pdf, (último acesso Abril/2012)*, Jul. 2009.

[37] SEARA, R. J., SEARA, I. C. S., KAFKA, S. G., et al. "Parâmetros Lingüísticos Utilizados para a Geração Automática de Prosódia em Sistemas de Síntese de Fala". In: *XXI Simpósio Brasileiro de Telecomunicações (SBrT 2004)*, pp. 1–6, Belém, PA, 2004.

[38] JUANG, B. H., RABINER, L. R. "The Segmental k-Means Alfgorithm for Estimating Parameters of Hidden Markov Models", *IEEE Trans. ASSP*, v. 38, n. 9, pp. 1639–1641, Nov. 1990.

[39] WILPON, J. G., RABINER, L. R. "Modified k-Means Clustering Algorithm for Use in Isolated Word Recognition", *IEEE Trans. ASSP*, v. 33, pp. 587–594, June 1990.

[40] ALBANO, E., BARBOSA, P., GAMA-ROSSI, A., et al. "A Interface Fonética-Fonologia e a Interação Prosódica-Segmentos". In: *Estudos Linguísticos XXVII, Anais do XLV Seminário do Grupo de Estudos Línguístico do Estado de São Paulo - GEL'97*, pp. 135–143. UNICAMP, 1997.

[41] FREUND, Y., SCHAPIRE, R. "A Decision-theoretic Generalization of On-line Learning and an Application to Boosting", *Computer and System Sciences*, v. 55, n. 1, pp. 119 – 139, 1995.

[42] CAMPBELL JR., J. P. "Speaker Recognition: A Tutorial", *Proceedings of the IEEE*, v. 85, n. 9, pp. 1437–1462, Sep. 1997.

[43] FUKUNAGA, K. *Introduction to Statistical Pattern Recognition*. 2nd ed. San Francisco, Morgan Kaufman, 1990.

[44] BONATT, M. "A Produção de Plosivas por Crianças de Três Anos Falantes do Português Brasileiro", *Revista CEFAC Speech, Language, Hearing Sciences and Education Journal*, v. 9, pp. 199–206, Apr-June 2007.

[45] GOLIPOUR, L., O'SHAUGHNESSY, D. "A New Approach for Phoneme Segmentation of Speech Signals". In: *8th Annual Conference of the International Speech Communication Association - INTERSPEECH/2007*, Aug. 2007.

[46] AMBIKAIRAJAH, E. . A. A. "Language Identification: A Tutorial", *IEEE Circuits and Systems*, v. 11, n. 02, pp. 82–108, 2ªquarter 2011.

[47] CHARONNAT, L., VIDAL, G., BOEFFARD, O. "Automatic Phone Segmentation of Expressive Speech". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008.

[48] MPORAS, I., GANCHEV, T., FAKOTAKIS, N. "Speech Segmentation using Regression Fusion of Boundary Predictions". In: *Computer Speech e Language*, v. 24, pp. 273–288, Apr. 2010.

[49] TING, C.-M., SALLEH, S.-H., TAN, T.-S., et al. "Automatic Phonetic Segmentation of Malay Speech Database". In: *Information, Communications Signal Processing, 2007 6th International Conference on*, pp. 1 –4, Dec. 2007.

[50] SANTOS, S. C. B. S. *Reconhecimento de Voz Contínua para o Português Utilizando Modelos de Markov Escondidos*. Tese de doutorado, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, Brasil, 1997.

[51] PARANAGUÁ, E., NETTO, S. "Parametrização de Sinais da Fala para Reconhecimento Automático de Locutor", *Revista Tecnologia&Cultura*, v. 08, n. 09, pp. 50–58, Jul-Dez. 2007.

[52] PARANAGUÁ, E., SILVA, D., NETTO, S., et al. "Segmentação Automática de Voz para Sistemas de Conversão Texto-fala". In: *Anais do X Congresso Internacional de Fonética e Fonologia & IV Congresso Internacional de Fonética e Fonologia*, Niterói, RJ, Brasil, Nov. 2008.

[53] ADELL, J., BONAFONTE, A., GÖMEZ, J. A., et al. "Comparative Study of Automatic Phone Segmentation Methods for TTS". In: *Proceedings*

of 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), v. 1, pp. 309–312, Philadelphia, Mar. 2005.