Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação
Departamento de Comunicações

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

Antônio Marcos Selmini
Orientador: Prof. Dr. Fábio Violaro

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para a obtenção do título de Doutor em Engenharia Elétrica. Área de Concentração: **Engenharia de Telecomunicações.**

Banca Examinadora

Fábio Violaro, Dr. ..... DECOM/FEEC/UNICAMP
Aldebaro Barreto da Rocha Klautau Júnior, Dr. ..... CT/DEEC/UFPA
Carlos Alberto Ynoguti, Dr. ..... INATEL
Jayme Garcia Arnal Barbedo, Dr. ..... DECOM/FEEC/UNICAMP
Plínio Almeida Barbosa, Dr. ..... IEL/UNICAMP

Campinas, SP
Agosto/2008

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

Selmini, Antônio Marcos
Se49s Sistema baseado em regras para o refinamento da segmentação automática de fala / Antônio Marcos Selmini. --Campinas, SP: [s.n.], 2008.

Orientador: Fábio Violaro.
Tese de Doutorado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Sistemas de processamento da fala. 2. Fonética acústica. 3. Reconhecimento automatico da voz. 4. Markov, Processo de. 5. Algoritmos. I. Violaro, Fábio. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Titulo em Inglês: Rule based system for refining the automatic speech segmentation
Palavras-chave em Inglês: Automatic speech segmentation, Refining the automatic speech segmentation, Acoustic-phonetic features, HMM modeling, Viterbi's algorithm
Área de concentração: Telecomunicações e Telemática
Titulação: Doutor em Engenharia Elétrica
Banca examinadora: Aldebaro Barreto da Rocha Klautau Júnior, Carlos Alberto Ynoguti, Jayme Garcia Arnal Barbedo, Plínio Almeida Barbosa
Data da defesa: 22/08/2008
Programa de Pós Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE DOUTORADO

Candidato: Antônio Marcos Selmini

Data da Defesa: 22 de agosto de 2008

Título da Tese: "Sistema Baseado em Regras para o Refinamento da Segmentaçã Automática de Fala"

Prof. Dr. Fábio Violaro (Presidente): _____

Prof. Dr. Carlos Alberto Ynoguti: _____

Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior: _____

Dr. Jayme Garcia Arnal Barbedo: _____

Prof. Dr. Plínio Almeida Barbosa: _____

Resumo

A demanda por uma segmentação automática de fala confiável vem crescendo e exigindo pesquisas para suportar o desenvolvimento de sistemas que usam fala para uma interação homem-máquina. Neste contexto, este trabalho relata o desenvolvimento e avaliação de um sistema para segmentação automática de fala usando o algoritmo de Viterbi e refinamento das fronteiras de segmentação baseado nas características fonético-acústicas das classes fonéticas. As subunidades fonéticas (dependentes de contexto) são representadas com Modelos Ocultos de Markov (HMM – *Hidden Markov Models*). Cada fronteira estimada pelo algoritmo de Viterbi é refinada usando características acústicas dependentes de classes de fones, uma vez que a identidade dos fones do lado direito e esquerdo da fronteira considerada é conhecida. O sistema proposto foi avaliado usando duas bases dependentes de locutor do Português do Brasil (uma masculina e outra feminina) e também uma base independente de locutor (TIMIT). A avaliação foi realizada comparando a segmentação automática com a segmentação manual. Depois do processo de refinamento, um ganho de 29% nas fronteiras com erro de segmentação abaixo de 20 ms foi obtido para a base de fala dependente de locutor masculino do Português Brasileiro.

Palavras-chave: Segmentação automática de fala, características fonético-acústicas, modelagem HMM, refinamento da segmentação automática de fala.

Abstract

The demand for reliable automatic speech segmentation is increasing and requiring additional research to support the development of systems that use speech for man-machine interface. In this context, this work reports the development and evaluation of a system for automatic speech segmentation using Viterbi's algorithm and a refinement of segmentation boundaries based on acoustic-phonetic features. Phonetic sub-units (context-dependent phones) are modeled with HMM (Hidden Markov Models). Each boundary estimated by Viterbi's algorithm is refined using class-dependent acoustic features, as the identity of the phones on the left and right side of the considered boundary is known. The proposed system was evaluated using two speaker dependent Brazilian Portuguese speech databases (one male and one female speaker), and a speaker independent English database (TIMIT). The evaluation was carried out comparing automatic against manual segmentation. After the refinement process, an improvement of 29% in the percentage of segmentation errors below 20 ms was achieved for the male speaker dependent Brazilian Portuguese speech database.

Keywords: Automatic speech segmentation, acoustic-phonetic features, HMM modeling, refining automatic speech segmentation.

Agradecimentos

Aos meus familiares e amigos pela paciência e compreensão nos diversos momentos em que tive que me ausentar do convívio social para me dedicar ao trabalho de doutorado.

Agradeço imensamente ao Prof. Dr. Fábio Violaro pelo rico conhecimento transmitido, por todos os ensinamentos, pelas inúmeras sugestões e discussões durante o desenvolvimento do trabalho, e também pela paciência.

Aos meus colegas de trabalho, pela força e coragem nos muitos e muitos momentos de desânimo durante o projeto.

Dedico esse trabalho aos meus pais e meus avós,
que infelizmente não tiveram todas as oportunidades
de estudo que puderam me proporcionar.

Sumário

Lista de Figuras

Lista de Tabelas

Glossário

BIC – Bayesian Information Criterion

CART – Classification and Regression Tree

DCF – Delta Cepstral Function

DFT – Discrete Fourier Transform

FFT – Fast Fourier Transform

GLR – Generalized Likelihood Ratio

HMM – Hidden Markov Models

HTK – Hidden Markov Models Toolkit

IA – Inteligência Artificial

IPA – International Phonetic Alphabet

LDA – Linear Discriminant Analysis

LPC – Linear Predictive Coding

MAP – Maximum a Posteriori

MFCC – Mel Frequency Cepstral Coefficient

MLLR – Maximum Likelihood Linear Regression

PB – Português do Brasil

PLP – Perceptual Linear Prediction

SVM – Support Vector Machine

TIMIT – Texas Instruments - Massachussets Institute of Technology

TTS – Text to Speech Synthesis

VOT – Voice Onset Time

Lista de Símbolos

a_{ij}	– Probabilidade de transição de estados
$b_j(k)$	– Probabilidade de observação dos símbolos
o_t	– vetor de características acústicas no instante de tempo *t*
dim	– dimensão do vetor de parâmetros de entrada
L	– número de Gaussianas na mistura de cada estado do modelo HMM
c_{jk}	– coeficiente de ponderação da Gaussiana
G	– função densidade de probabilidade Gaussiana multidimensional
μ_{jk}	– vetor média da Gaussiana *k* no estado *j* do modelo
U_{jk}	– matriz de covariância da Gaussiana no estado "j" do modelo
$\lvert U_{jk} \rvert$	– determinante da matriz de covariância
U_{jk}^{-1}	– matriz covariância inversa
$\eta(o_t)$	– conjunto de funções de probabilidade para o *codebook*
$f(o_t \mid v_k)$	– valor da *k*-ésima função densidade de probabilidade para o vetor de parâmetros de entrada o_t
v_k	– *k*-ésimo símbolo de saída
$F_t(j,m)$	– verossimilhança normalizada
$C_i(t)$	– *i*-ésimo coeficiente cepstral para a janela de análise *t*
$c(t)$	– função de custo para a variação cepstral
$BIC(i)$	– variação do valor do BIC entre dois modelos no instante de tempo *i*
P_0	– fator de penalização para o cálculo do BIC
F_β	– perfil de energia
z_n	– taxa de cruzamentos por zero
$\Delta Eb_i(n)$	– variação da energia espectral na banda de freqüência *i* para a janela de análise *n*.
$w(n)$	– janela de Hamming

Trabalhos Publicados Durante o Doutorado

SELMINI, A. M.; VIOLARO, F.; Segmentação Automática de Fala para o Português Brasileiro. Anais do XXVI Simpósio Brasileiro de Telecomunicações, Rio de Janeiro, Setembro, 2008. ISBN 978-85-89748-05-6.

SELMINI, A. M.; VIOLARO, F.; Acoustic-Phonetic Features for Refining Automatic Speech Segmentation. In: Proceedings of the INTERSPEECH 2007, Antwerp, Belgium, August, 2007.

SELMINI, A. M.; VIOLARO, F.; Improving the Explicit Speech Segmentation. In: Proceedings of the International Workshop on Telecommunications, IWT 2007, Santa Rita do Sapucaí (MG), February, 2007.

YARED, G. F. G.; VIOLARO, F.; SELMINI, A. M.; HMM Topology in Continuous Speech Recognition Systems. 6th International Telecommunications Symposium (ITS2006). CD-ROM, pp. 588-593, ISBN: 85-89748-04-9, IEEE Catalog Number: 06EX1453C.

Capítulo 1

Introdução

1.1. Introdução

A idéia de se construir sistemas dotados de inteligência com as características dos seres humanos tornou-se uma realidade com o surgimento da inteligência artificial (IA) aliada aos avanços tecnológicos em todas as áreas do conhecimento, notadamente na engenharia e na ciência da computação.

Dentre todas as características presentes nos seres humanos que um sistema inteligente deve apresentar, a mais importante é a capacidade de entender e se comunicar em uma língua natural. Como a fala representa o meio básico e primordial de comunicação entre as pessoas, os sistemas de reconhecimento automático e síntese de fala passaram a ser largamente estudados e aprimorados ao longo dos anos.

Um sistema de reconhecimento automático de fala visa transformar uma locução em uma "string" (texto em uma determinada língua natural) e, em um sistema de síntese a partir de um texto, é gerada a fala. As aplicações para os sistemas de reconhecimento de fala podem variar muito, desde sistemas com vocabulário limitado a poucas palavras até sistemas com vocabulário de milhares de palavras, sistemas dependentes ou independentes de locutor, etc. Independentemente da aplicação, o objetivo principal é aproximar ao máximo, tanto o reconhecimento automático quanto a síntese, do processamento natural realizado pelos seres humanos.

Dependendo da aplicação e da sofisticação desejada para os sistemas de reconhecimento automático e síntese de fala, cada vez mais se torna necessária a utilização de técnicas e modelagens que produzam um resultado com uma qualidade muito próxima à dos seres humanos.

Tanto em sistemas de reconhecimento automático quanto em sistemas para síntese de fala, uma tarefa indispensável é a segmentação. Segmentar uma determinada locução consiste em

determinar as fronteiras que separam os elementos essenciais da locução. Esses elementos essenciais dependem da aplicação em questão e podem ser simplesmente palavras, sílabas ou unidades acústicas menores, como os fones.

Em sistemas de reconhecimento de fala que utilizam Modelos Ocultos de Markov (HMM – *Hidden Markov Models*) para representar as subunidades fonéticas, uma base de fala segmentada é importante porque é utilizada como ponto de partida para o treinamento das subunidades (durante a fase de treinamento). Essas subunidades podem ser dependentes ou independentes de contexto. Sem uma base de fala segmentada, durante o treinamento as locuções são segmentadas de modo uniforme para gerar as primeiras estimativas que serão refinadas durante as próximas iterações do treinamento. A segmentação de fala também é importante em muitos algoritmos de treinamento discriminativo.

Nos sistemas de síntese ou conversão texto-fala (TTS – *Text To Speech Synthesis*), uma base de fala segmentada e rotulada (transcrição fonética presente) de um locutor também é necessária. Para a síntese de fala (no caso da síntese concatenativa), as subunidades fonéticas são extraídas da base de fala e concatenadas para produzir a fala sintetizada.

Uma outra aplicação em que uma base de fala segmentada é necessária são as animações sincronizadas com fala (*Talking Heads*). Nestes sistemas, os movimentos articulatórios de uma face virtual tridimensional imitam os movimentos produzidos por um locutor real. Estes movimentos devem estar sincronizados com os fones da locução que está sendo pronunciada. No caso de se usar um sistema de síntese para alimentar o *talking head*, a segmentação já é fornecida pelo próprio sistema, mas no caso de se usar uma base de fala real, a segmentação deve ser fornecida.

Preparar uma base de fala segmentada não é uma tarefa fácil, principalmente se a segmentação manual for utilizada. A segmentação manual que é realizada por foneticistas ou pessoas com conhecimento na área fonética é uma tarefa extremamente tediosa, cansativa, que consome muito tempo e é, portanto, economicamente inviável em muitas situações. Outra desvantagem da segmentação manual é a falta de consistência entre as segmentações produzidas por diferentes especialistas. Com a crescente necessidade de bases de fala cada vez maiores, uma segmentação automática de alta qualidade é extremamente desejada.

Normalmente, a segmentação manual é realizada utilizando-se diversas ferramentas tais como análise de espectrogramas, trajetória de formantes, análise de curva de energia e também testes subjetivos para avaliar a qualidade do segmento obtido.

1.2. Motivação e Objetivos

Como apresentado na seção anterior, com o crescente desenvolvimento de sistemas que usam fala para uma interface homem-máquina, a demanda por uma segmentação automática confiável também cresce, visto que a segmentação é um processo importante e necessário em diversas áreas do processamento de fala.

Uma vez que a segmentação manual torna-se economicamente inviável devido ao tamanho das bases de fala atualmente requeridas, é um processo tedioso, que consome muito tempo e que produz inconsistências por ser um processo subjetivo, este trabalho tem como objetivo principal desenvolver um sistema de segmentação automática de fala que possa produzir segmentos acústicos a partir de uma locução, evitando dessa forma as inconsistências e o tédio da segmentação manual. É importante destacar que o sistema para segmentação desenvolvido neste trabalho não é específico para nenhuma aplicação, ou seja, a segmentação produzida pode ser utilizada em diversas aplicações.

Vários procedimentos clássicos para a segmentação automática foram propostos na comunidade científica, apresentando vantagens e desvantagens. Neste trabalho, a segmentação será realizada empregando HMMs e efetuando o alinhamento inicial da locução com os estados do HMM através do algoritmo de Viterbi.

A transcrição fonética da locução que será segmentada é essencial para o sistema, pois a partir da transcrição fonética é montado o HMM da locução através da concatenação dos HMMs de seus fones constituintes. Em seguida um algoritmo de refinamento é aplicado em cada fronteira previamente estimada para melhorar a segmentação inicialmente obtida. O processo de refinamento será realizado levando em consideração os fones presentes no lado direito e esquerdo da fronteira que está sendo analisada.

Para cada fronteira previamente estimada, uma regra específica é aplicada em um intervalo de refinamento próximo à fronteira. Cada regra é composta por um ou mais parâmetros que indicarão a posição da nova fronteira. Alguns parâmetros apresentam um limiar previamente calculado enquanto que outros são baseados na detecção de picos (*peak picking*) resultantes da

derivada dos parâmetros. Esses limiares ou picos são responsáveis por indicar a janela de análise em que as características acústicas dos fones mudam.

Como as regras de refinamento são baseadas nos fones separados pela fronteira que está sendo analisada, a informação da transcrição fonética das locuções também é empregada. Uma vantagem da estratégia adotada para o refinamento é que não há a necessidade de material de treinamento para treinar "modelos" de fronteiras.

1.3. Contribuições da Tese

Como a segmentação automática de fala desempenha um papel primordial em diversos sistemas que usam fala para interação homem-máquina, este trabalho é de extrema importância para os próximos trabalhos que serão desenvolvidos no LPDF (Laboratório de Processamento Digital da Fala – DECOM/UNICAMP). As principais contribuições desse projeto são:

- Adquirir o conhecimento necessário e familiaridade sobre técnicas de segmentação automática de fala e refinamento para a construção de bases de fala segmentadas do Português do Brasil (PB). Não existe no Brasil uma base oficial e confiável que possa ser usada para testar os sistemas desenvolvidos, como ocorre nos Estados Unidos e na Europa. As bases são construídas nos diversos centros de pesquisas para sua própria utilização.
- Produzir segmentações confiáveis que possam ser aplicadas ao treinamento discriminativo dos HMMs para o reconhecimento automático de fala e nos trabalhos de conversão texto-fala.
- Fornecer a base de um sistema de segmentação e refinamento para que possa ser aprimorado e configurado para outras aplicações em que a segmentação de fala se faz necessária.

1.4. Organização da Tese

Esta tese está dividida em sete capítulos. No Capítulo 2 é discutida a modelagem estatística baseada nos HMMs, técnica esta que domina a área de reconhecimento automático de fala. O Capítulo também apresenta o algoritmo para o treinamento dos HMMs (algoritmo de Baum-Welch) e o algoritmo de Viterbi que será utilizado para gerar as primeiras fronteiras de segmentação.

No Capítulo 3 são apresentadas a teoria sobre segmentação automática de fala e as principais técnicas sobre refinamento ou pós-processamento da segmentação automática de fala. Uma revisão bibliográfica sobre os dois tópicos é apresentada. O capítulo tem como objetivo principal fornecer a base para justificar a escolha da técnica de refinamento apresentada neste trabalho.

O processo de produção da fala e sua parametrização são discutidos no Capítulo 4. Neste capítulo são apresentados os principais parâmetros representativos de cada classe fonética, parâmetros esses que serão de extrema importância para o processo de refinamento automático das marcas de segmentação.

Tendo como base os Capítulos 2, 3 e 4, no Capítulo 5 é apresentada a arquitetura do sistema desenvolvido, com ênfase nas regras e parâmetros utilizados para refinar cada marca de segmentação.

Uma vez apresentado o funcionamento do sistema, o Capítulo 6 descreve as bases de fala utilizadas nos testes, as configurações utilizadas no processo de treinamento dos HMMs, os resultados iniciais obtidos com a segmentação de Viterbi e os resultados finais obtidos com a aplicação do sistema de refinamento desenvolvido.

Finalmente, no Capítulo 7 são apresentadas as conclusões da tese e algumas propostas de trabalhos futuros.

Capítulo 2

Modelos Ocultos de Markov

Quando se fala em processamento digital de fala, é impossível não falar sobre modelagem estatística, principalmente sobre os HMMs e o algoritmo de Viterbi. Atualmente a maioria das aplicações que empregam fala usam HMMs para modelagem das subunidades fonéticas, mas esse cenário nem sempre foi assim.

Os primeiros sistemas de reconhecimento automático de fala empregavam métodos baseados em reconhecimento de padrões. A idéia básica consistia em gerar padrões acústicos de palavras ou de fones e, a partir desses padrões e usando medidas de distância espectral, seqüências de palavras ou de fones eram reconhecidas. Além das medidas de distância espectral, métodos baseados em programação dinâmica também foram empregados para alinhar os modelos de padrões acústicos com as informações acústicas que seriam reconhecidas. Esses métodos ainda são empregados, mas combinados com outras técnicas (Rabiner and Juang, 1993).

A introdução da modelagem estatística de sinais, notadamente o emprego dos Modelos Ocultos de Markov, trouxe grandes avanços ao processamento digital de fala. Através desses modelos foi possível caracterizar melhor as variações temporais e espectrais da fala e, conseqüentemente, obter melhores resultados em reconhecimento automático de fala, síntese texto-fala e também em segmentação de fala.

Neste Capítulo serão discutidos os Modelos Ocultos de Markov, o algoritmo básico de treinamento (algoritmo de Baum-Welch) e o alinhamento de Viterbi. Os Modelos Ocultos de Markov serão utilizados para representar as subunidades fonéticas e o alinhamento de Viterbi será utilizado para gerar a estimativa inicial das fronteiras de segmentação que serão posteriormente refinadas usando um algoritmo proposto neste trabalho.

2.1. Modelos Ocultos de Markov

A teoria de Modelos Ocultos de Markov foi introduzida na literatura na década de 60 por Baum (Baum, 1966). Sua utilização na área de reconhecimento automático de fala se deu na década de 70 e, desde então, passou a ser largamente utilizado em diversas aplicações tais como reconhecimento de fala, conversão texto-fala via HMM, etc.

Um HMM é um modelo estatístico baseado na teoria dos processos de Markov, utilizado para modelar processos estocásticos. Em reconhecimento de fala, HMMs são utilizados para modelar palavras e até mesmo unidades menores que são os fones, que é o caso de interesse em reconhecimento automático de fala para grandes vocabulários e também em segmentação.

Por definição, um HMM nada mais é do que um conjunto de *N* estados conectados por transições. Associada às transições entre os estados existe uma probabilidade ou densidade de probabilidade de emissão dos símbolos (dependendo se o HMM é discreto ou contínuo). Os símbolos representam as possíveis saídas físicas do sistema que está sendo modelado.

Dependendo do número de símbolos presentes no alfabeto de símbolos, o HMM pode ser discreto (alfabeto finito) ou contínuo (função densidade de probabilidade contínua). As transições de estados também são responsáveis pela modelagem das variabilidades temporais dos padrões de voz (Figueiredo, 1999).

A cada instante de tempo *t* existe uma mudança de estado (que pode ser para o mesmo estado) e um símbolo é emitido com uma determinada densidade de probabilidade de saída. Esta seqüência de símbolos emitidos é chamada de seqüência de observações, que por sua vez é a saída do HMM.

Em resumo, um HMM é constituído por (Rabiner and Juang, 1993):

1. Número *N* de estados q_j no modelo, sendo $1 \le j \le N$;
2. Número *M* de observações distintas e finitas (caso o seja HMM discreto);
3. Distribuição de probabilidade de transição entre os estados $A = \{a_{ij}\}$, onde

$$a_{ij} = P[q_{t+1} = j \mid q_t = i], \text{ onde } 1 \le i, j \le N; \tag{2.1}$$

Os coeficientes a_{ij} da matriz de transição *A* apresentam duas propriedades:

$$a_{ij} \ge 0 \text{ para } 1 \le i, j \le N \tag{2.2}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \text{ para } 1 \le i \le N \tag{2.3}$$

4. Distribuição de probabilidade de observação dos símbolos $B = \{b_j(k)\}$, onde $b_j(k) = P[\mathbf{o}_t = v_k \mid q_t = j]$, sendo $1 \le k \le M$, o é a seqüência de observações e v denota os símbolos de saída para um HMM discreto;
5. Distribuição do estado inicial $\Pi = \{\pi_i\}$, sendo $\pi_i = P[q_1 = i]$, onde $1 \le i \le N$.

A partir das observações acima pode-se notar que, para uma especificação completa de um HMM, é necessário definir o número de estados do modelo (*N*), o número de símbolos de observação (*M*) caso o HMM seja discreto, as probabilidades de transição entre os estados (*A*), de observação de símbolos (*B*) e do estado inicial (Π). Uma forma compacta é utilizada para indicar o conjunto completo de parâmetros do modelo:

$$\lambda = (A, B, \Pi) \tag{2.4}$$

2.1.1. Classificação dos HMMs

Os HMMs basicamente podem ser classificados segundo dois critérios: i) quanto à distribuição de probabilidade associada a cada estado e ii) quanto à topologia.

A cada estado do HMM é associada uma função densidade de probabilidade, que pode ser tanto uma função massa de probabilidade (caso discreto) quanto uma função densidade de probabilidade (caso contínuo). De acordo com essa densidade, os HMMs podem ser classificados em discreto, contínuo ou semicontínuos (uma combinação do HMM discreto com o HMM contínuo).

- **HMM discreto:**

No HMM discreto a principal operação envolvida é a quantização vetorial. Neste processo é definido um dicionário (*codebook*), que por sua vez é composto por palavras-código (*codewords*). Durante o processo de reconhecimento, a partir da janela de análise do sinal é obtido um conjunto de parâmetros acústicos que, após o processo de quantização vetorial é associado a um dos *M* possíveis *codewords*.

Como no HMM discreto a função massa de probabilidade de saída é modelada por uma função discreta, o número de possíveis símbolos de saída (*M*) é finito. A probabilidade de emitir o símbolo v_k no estado q_j é dada por $b_j(k)$, com as seguintes propriedades:

$$b_j(k) \geq 0 \text{ para } 1 \leq j \leq N \text{ e } 1 \leq k \leq M \tag{2.5}$$

$$\sum_{k=1}^{M} b_j(k) = 1 \tag{2.6}$$

onde:

N é o número de estados do HMM;

M é o número de símbolos discretos do modelo;

$b_j(k)$ é a probabilidade de emitir o símbolo v_k no estado q_j;

- **HMM contínuo:**

Para o caso do HMM contínuo, a função densidade de probabilidade de saída é contínua. Normalmente a função densidade de probabilidade é modelada por uma mistura de *L* Gaussianas multidimensionais (Rabiner and Juang, 1993). Esta função densidade de probabilidade é representada por:

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{L} c_{jk} G(\mathbf{o}_t, \mu_{jk}, U_{jk}) \tag{2.7}$$

sendo:

$\mathbf{o}_t$, o vetor de parâmetros de entrada de dimensão *dim* no instante de tempo *t*;

L, é o número de Gaussianas na mistura para cada estado;

c_{jk}, é o coeficiente de ponderação para a *k-ésima* mistura no estado j;

G, é a função densidade de probabilidade Gaussiana multidimensional com vetor média μ_{jk} e matriz de covariância U_{jk} para o componente da *k-ésima* mistura no estado j;

A função densidade de probabilidade Gaussiana multidimensional *G* é dada por (Huang et al., 1990):

$$G(\mathbf{o}_t, \mu_{jk}, U_{jk}) = \frac{1}{(2\pi)^{\dim/2} |U_{jk}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}_t - \mu_{jk})U_{jk}^{-1}(\mathbf{o}_t - \mu_{jk})'\right] \quad (2.8)$$

onde:

dim, a dimensão do vetor $\mathbf{o}_t$;

$|U_{jk}|$, o determinante da matriz de covariância;

U_{jk}^{-1}, a inversa da matriz de covariância;

O coeficiente de ponderação c_{jk} deve satisfazer a seguinte restrição:

$$\sum_{k=1}^{L} c_{jk} = 1,\ 1 \le j \le N,\ c_{jk} \ge 0,\ 1 \le k \le L \quad (2.9)$$

A função densidade de probabilidade por sua vez satisfaz a restrição representada pela Equação (2.10):

$$\int_{-\infty}^{\infty} b_j(\mathbf{o}) d\mathbf{o} = 1,\ 1 \le j \le N \quad (2.10)$$

Uma questão importante com o uso de HMMs contínuos é o número de parâmetros livres que devem ser estimados durante o processo de treinamento dos modelos. Uma forma de reduzir esse número em HMMs contínuos baseados em funções densidade de probabilidade Gaussianas é o uso de matriz de covariância diagonal (que corresponde a considerar os componentes do vetor de parâmetros independentes entre si). Essa estratégia diminui o esforço computacional para a estimação dos parâmetros e também exige menos dados de treinamento.

- **HMM semicontínuo:**

A idéia geral do HMM semicontínuo é combinar as vantagens do HMM discreto e do HMM contínuo, construindo dessa forma um novo modelo intermediário, o semicontínuo.

Uma das vantagens do HMM discreto é sua capacidade de modelar qualquer evento aleatório com um número razoável de parâmetros. Por outro lado, a operação de quantização vetorial divide o espaço acústico em diversas regiões usando tipicamente medidas de distorção espectral. Essa divisão é responsável por diminuir a precisão do modelo (Huang et al., 1990). Uma sugestão para contornar o problema descrito é representar o *codebook* por uma mistura de funções densidade de probabilidade, em que a distribuição seja sobreposta ao invés de ser dividida. Dessa forma, cada "palavra" do *codebook* é representada por uma das funções densidade de probabilidade.

As funções densidade de probabilidade das misturas no HMM contínuo podem ser compartilhadas com o *codebook*, reduzindo dessa forma o número de parâmetros livres que devem ser estimados e também reduzindo o problema da partição do espaço no HMM discreto.

No HMM semicontínuo a densidade de probabilidade de emissão dos símbolos de saída é dada pela seguinte expressão (Huang et al., 1988a, 1988b):

$$b_j(\mathbf{o}_t) = \sum_{v_k \in \eta(o_t)} c_j(k) f(\mathbf{o}_t \mid v_k) \qquad \text{para } 1 \le j \le N \tag{2.11}$$

onde:

N, o número de estados do modelo;

$\mathbf{o}_t$, é o vetor de parâmetros de entrada;

$\eta(\mathbf{o}_t)$, é o conjunto das funções densidade de probabilidade das "palavras" do *codebook* que apresentam os M maiores valores de $f(\mathbf{o}_t \mid v_k)$, $1 \le M \le K$. O valor adequado de M sugerido na literatura está na faixa de 2 a 8 (Huang et al., 1990).

K, é o número de funções densidade de probabilidade;

v_k, é o k-ésimo símbolo de saída;

$c_j(k)$, é o coeficiente de ponderação das Gaussianas;

$f(\mathbf{o}_t \mid v_k)$, é o valor da k-ésima função densidade de probabilidade para o vetor de parâmetros de entrada $\mathbf{o}_t$;

Como o HMM semicontínuo é uma combinação dos modelos discretos e contínuos, o HMM semicontínuo pode tornar-se discreto ou contínuo. Quando o valor de M é igual a 1, o

HMM semicontínuo torna-se um HMM discreto com um *codebook* formado por funções densidade de probabilidade. Neste caso, usa-se apenas a função $f(\mathbf{o}_t \mid v_k)$ que apresentar maior valor para calcular a densidade de probabilidade de emissão de símbolos de saída.

Já no caso em que o valor de *M* é igual a *K*, pode-se considerar o HMM semicontínuo como um HMM contínuo em que todas as misturas (vetor média μ_{jk} e matriz de covariância U_{jk}) são iguais para todos os estados e todos os modelos. A única variante de um estado para outro são os valores dos coeficientes de ponderação.

Em resumo, o modelo semicontínuo agrega as vantagens dos modelos discreto e contínuo. Desse modo, é possível melhorar a robustez do modelo discreto usando o modelo contínuo, mas reduzindo o número de parâmetros livres e a complexidade computacional. A complexidade computacional é reduzida em virtude do compartilhamento das funções densidade de probabilidade no *codebook*.

2.1.2. Topologia dos HMMs

Quanto à topologia, os HMMs normalmente são classificados em totalmente conectados ou ergódicos e *left-right*, também conhecido como modelo de Bakis (Rabiner and Juang, 1993).

Um HMM é ergódico quando a partir de um estado do modelo é possível atingir todos os outros, resultando em uma matriz de transição de estados totalmente preenchida. O segundo tipo, que é o mais utilizado para a modelagem da fala, foi proposto por Bakis em 1976 (Bakis, 1976). A Figura 2.1 mostra um exemplo de HMM com topologia do tipo *left-right* com salto duplo.

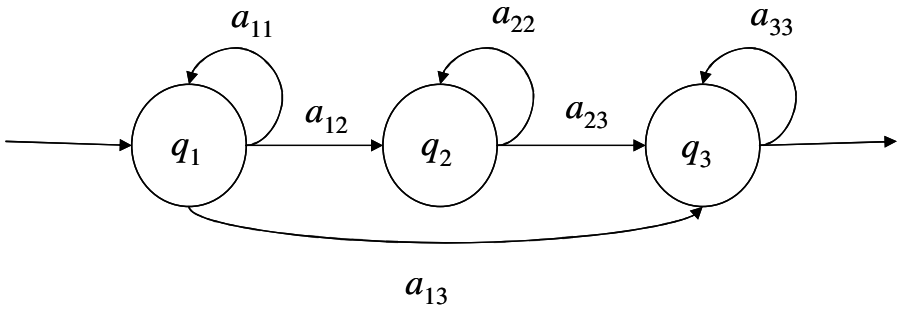


Figura 2.1: Topologia *left-right* com salto duplo.

No modelo *left-right*, a matriz de transição de estados apresenta a seguinte propriedade:

$$a_{ij} = 0 \text{ para } j < i; \tag{2.12}$$

Segundo a propriedade representada pela Equação (2.12), nenhuma transição é permitida para estados cujos índices são menores do que o estado corrente. Outra propriedade apresentada pelo modelo *left-right* é com relação à probabilidade do estado inicial. A seqüência de estados deve começar no primeiro estado e terminar no último (*N*). A probabilidade do estado inicial é representada matematicamente pela relação:

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases} \tag{2.13}$$

Uma vez estabelecidos os parâmetros que definem os HMMs, três problemas básicos surgem e devem ser resolvidos a fim de que os modelos possam ser utilizados em aplicações práticas (Rabiner and Juang, 1993):

Problema 1:

– Dada uma seqüência de eventos observados e dado um modelo, o primeiro problema está ligado ao cálculo da verossimilhança da seqüência observada ter sido gerada pelo modelo, ou seja, determinar uma medida que reflita o quão próximo a seqüência de observações se encontra do modelo. Este problema pode ser resolvido através do algoritmo *forward* ou *backward*.

Problema 2:

– Este problema está relacionado com a determinação da seqüência de estados ocultos associada aos eventos observados. Diversas seqüências de estados possíveis podem existir, porém deve-se estabelecer um critério para determinar a seqüência mais provável, como por exemplo aquela que fornece a maior verossimilhança $P(\mathbf{O} \mid \lambda)$, em que **O** é a seqüência de observações e λ é o conjunto de parâmetros que define o modelo. A determinação da seqüência ótima é realizada neste trabalho por meio do algoritmo de Viterbi, que será descrito nas próximas seções.

Problema 3:

– O último problema diz respeito à estimação dos parâmetros do modelo, que devem ser ajustados de acordo com um método de treinamento de tal forma que o sistema apresente um bom desempenho. A solução desse problema pode ser obtida através do algoritmo de treinamento Baum-Welch.

2.2. HMMs Dependentes de Contexto

Como já foi reportado, os HMMs desempenham papel fundamental na área de reconhecimento automático de fala devido à sua capacidade para modelar as variações acústicas e temporais das unidades fonéticas. Apesar desse excelente desempenho, o grande problema presente na fala contínua são os efeitos contextuais que provocam variações na maneira como os sons são produzidos, dificultando portanto uma correta modelagem.

Na prática, o ideal é modelar e treinar cada um dos diferentes contextos de um mesmo fone com um HMM diferente de forma a obter uma boa discriminação entre eles. Para contornar essa situação, ao invés de usar modelos de monofones, o ideal é usar no mínimo modelos de trifones.

Em um HMM que emprega trifones, cada fone possui um modelo distinto dependendo dos fones à sua direita e à sua esquerda. Os modelos trifones são representados por X-Y+Z, que significa a ocorrência do fone Y seguido por Z e precedido por X. O sinal de subtração (-) em X- representa o fone à esquerda e o sinal de adição (+) em +Z representa o fone à direita. Como exemplo a frase "boa tarde" pode ser representada por "# b o a t a R D y #" ou pelos trifones "# #-b+o b-o+a o-a+t a-t+a t-a+R a-R+D R-D+y D-y+# #". Uma outra vantagem do uso da modelagem com trifones é que as fronteiras entre as palavras podem ser determinadas com maior precisão em relação à modelagem com monofones.

Apesar de todas as vantagens reportadas com o uso de trifones, existem também desvantagens. A primeira grande desvantagem está relacionada com o material de treinamento, pois há a necessidade da ocorrência de todos os fones em todos os contextos para um treinamento adequado. A segunda grande desvantagem é um aumento da quantidade de parâmetros que devem ser treinados.

O problema de muitos parâmetros e normalmente poucos dados de treinamento é muito comum em reconhecimento automático de fala. A solução encontrada é a união de misturas (*tied-*

mixture), ou seja, compartilhar os componentes das misturas de Gaussianas entre os estados dos HMMs. Para esse compartilhamento é necessário primeiro que haja a união dos estados (*state tying*) que são acusticamente comuns.

A união entre os estados pode ser realizada através da construção de uma árvore binária de decisão para cada fone. Em cada nó da árvore existe uma pergunta cuja resposta é "sim/não". Com base nas respostas o conjunto de estados é dividido até que os estados atinjam os nós terminais da árvore. Todos os estados que atingem os mesmos nós terminais são unidos. A Figura 2.2 ilustra o processo de agrupamento de estados realizado pelo HTK.

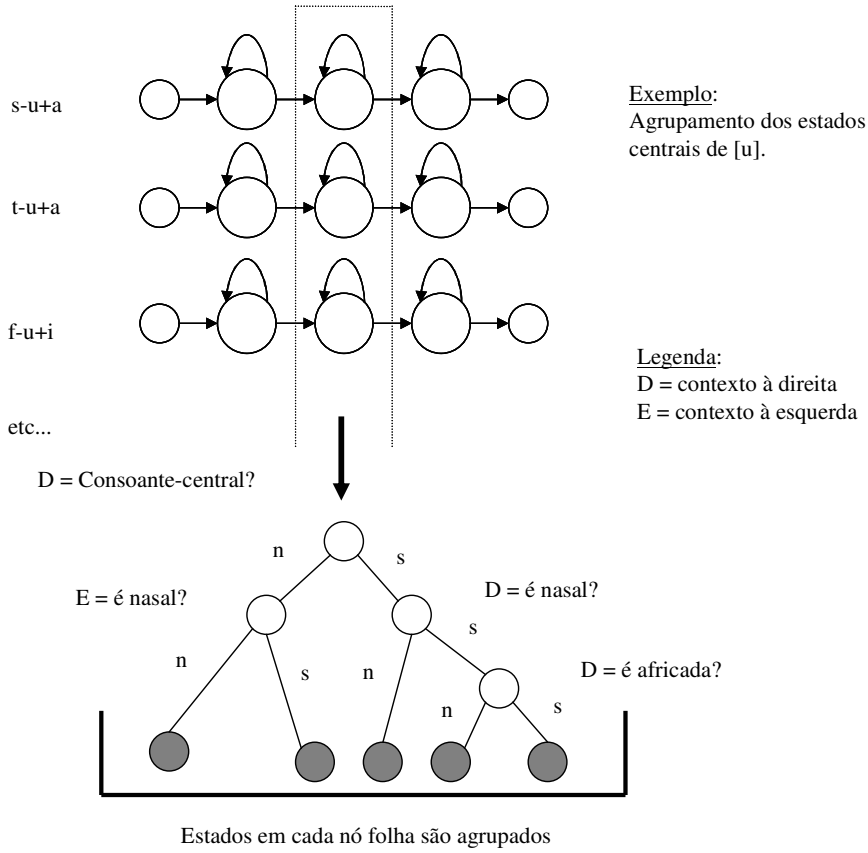


Figura 2.2: Agrupamento de estados baseado em árvore de decisão (Adaptado do HTKBook, 2006).

A grande vantagem da utilização de árvores de decisão é que o agrupamento de estados resultante pode estimar as probabilidades para qualquer contexto, quer eles apareçam ou não no

material de treinamento. Esse processo é realizado utilizando as distribuições de probabilidade associadas aos nós terminais.

2.3. Treinamento dos HMMs

Como citado na seção anterior, o treinamento dos HMMs consiste em ajustar os parâmetros do modelo de modo a satisfazer algum critério de otimização. Neste trabalho será empregado o critério da maximização da verossimilhança, ou seja, o processo de treinamento é repetido enquanto a verossimilhança na iteração atual é maior do que a verossimilhança da iteração anterior. O método mais conhecido e utilizado para o treinamento dos HMMs é o algoritmo de Baum-Welch, que consiste em um conjunto de equações recursivas.

Para definir o conjunto de equações de re-estimação dos parâmetros do modelo através do algoritmo de Baum-Welch é necessário definir dois outros algoritmos, *forward* e *backward*.

- **Algoritmo Forward**

Inicialmente para o desenvolvimento do algoritmo é definida a variável *forward* $\alpha_t(i)$ como:

$$\alpha_t(i) = P(\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_t, q_t = i \mid \lambda) \tag{2.14}$$

que representa a probabilidade da seqüência de observações parciais $\mathbf{o}_1$, $\mathbf{o}_2$, ..., $\mathbf{o}_t$ passando pelo estado i no instante de tempo t, dado o modelo λ. O algoritmo pode ser resumido em três passos:

1. Inicialização

$$\alpha_1(i) = \pi_i b_i(\mathbf{o}_1),\ 1 \le i \le N \tag{2.15}$$

2. Recursão

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(o_{t+1}),\ \begin{cases} 1 \le t \le T-1 \\ 1 \le j \le N \end{cases} \tag{2.16}$$

3. Término

$$P(\mathbf{O} \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{2.17}$$

O valor de $P(\mathbf{O} \mid \lambda)$ é uma medida da probabilidade de uma determinada locução formada pela seqüência de observações **O** ter sido produzida pela seqüência de estados $\mathbf{Q} = [q_1, q_2, q_3, ..., q_t, ..., q_T]$.

- **Algoritmo Backward**

De forma semelhante ao algoritmo *forward*, inicialmente é definida a variável *backward* $\beta_t(i)$ como:

$$\beta_t(i) = P(\mathbf{o}_{t+1}, \mathbf{o}_{t+2}, ..., \mathbf{o}_T \mid q_t = i, \lambda) \tag{2.18}$$

que corresponde à probabilidade da seqüência de observações parciais do instante t+1 até a última observação no instante T, dado que o caminho passa pelo estado i no instante t e dado o modelo λ. O algoritmo pode ser resumido em dois passos:

1. Inicialização

$$\beta_T(i) = 1, \; 1 \le i \le N \tag{2.19}$$

2. Recursão

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{o}_{t+1}) \, \beta_{t+1}(j), \quad \begin{cases} t = T-1, \; T-2, \; ..., \; 1 \\ 1 \le i \le N \end{cases} \tag{2.20}$$

Um dos grandes problemas enfrentado pelo algoritmo de treinamento do HMM é o *underflow*. Os valores fornecidos pelas variáveis *forward* e *backward* que são calculados de forma recursiva tendem a se tornar bem menores que 1 à medida que a seqüência de observações

é processada. Dessa forma esses valores excederão até a faixa de precisão dupla dos computadores, resultando em problemas numéricos.

Para contornar esse problema pode ser utilizado um fator de normalização (ou escalonamento) ou usar o logaritmo dos parâmetros. Tanto o fator de normalização quanto o logaritmo são aplicados para cada instante de tempo de forma a evitar o problema numérico de *underflow*. A idéia básica da normalização consiste em multiplicar os termos $\alpha_t(i)$ e $\beta_t(i)$ por um fator que é independente de *i*, mantendo estes termos dentro da faixa de precisão do computador para $1 \le t \le T$. O cálculo do fator de normalização ($\hat{c}_t$) e dos parâmetros normalizados ($\hat{\alpha}_t$) para o instante de tempo $t = 1$ é dado pela seguinte seqüência de equações (Huang et al., 1990):

1. Definir uma nova variável $\overline{\alpha}$, com o seguinte valor inicial:

$$\overline{\alpha}_1(i) = \alpha_1(i) \tag{2.21}$$

2. Definir o fator de normalização $\hat{c}_1$:

$$\hat{c}_1 = \frac{1}{\sum_{i=1}^{N} \overline{\alpha}_1(i)} \tag{2.22}$$

3. Definição de uma variável auxiliar $\hat{\alpha}$ com o seguinte valor inicial:

$$\hat{\alpha}_1(i) = \overline{\alpha}_1(i)\hat{c}_1 \tag{2.23}$$

4. Recursão:

$$\overline{\alpha}_{t+1}(j) = \left[\sum_{i=1}^{N} \hat{\alpha}_t(i)\ a_{ij} \right] b_j(\mathbf{o}_{t+1}), \quad \begin{cases} 1 \le t \le T-1 \\ 1 \le j \le N \end{cases} \tag{2.24}$$

$$\hat{c}_t = \frac{1}{\sum_{i=1}^{N} \overline{\alpha}_t(i)} \tag{2.25}$$

$$\hat{\alpha}_t(i) = \overline{\alpha}_t(i)\hat{c}_t \tag{2.26}$$

$$\hat{\beta}_t(i) = \hat{c}_t\overline{\beta}_t(i) \tag{2.27}$$

Na prática são empregadas seqüências de observações múltiplas **O** para o treinamento dos HMMs. Os parâmetros são calculados a partir das variáveis *forward* e *backward* normalizadas, do fator de normalização $\hat{c}$, dos vetores de parâmetros acústicos $\mathbf{o}_t$, dos parâmetros que compõem o modelo e dos valores de verossimilhança normalizada definida pela Equação (2.28):

$$F_t(j,m) = \frac{c_{jm}G(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm})}{\sum_{k=1}^{N_g} c_{jk}G(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk})} \tag{2.28}$$

onde N_g é o número de Gaussianas na mistura no estado j. A probabilidade de transição de estado, o peso, a média e a matriz de covariância são estimados pelas Equações (2.29)-(2.32).

- Probabilidade de transição de estados:

$$\overline{a}_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d-1} \hat{\alpha}_t^d(i)a_{ij}b_j(\mathbf{o}_{t+1}^d)\hat{\beta}_{t+1}^d(j)}{\sum_{d=1}^{D}\sum_{t=1}^{T_d-1} \hat{\alpha}_t^d(i)\hat{\beta}_t^d(i)/\hat{c}_t^d} \tag{2.29}$$

- Peso ou coeficiente de ponderação:

$$\overline{c}_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m)/\hat{c}_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)/\hat{c}_t^d} \tag{2.30}$$

- Média:

$$\overline{\boldsymbol{\mu}}_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m)\mathbf{o}_t^d / \hat{c}_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m) / \hat{c}_t^d} \tag{2.31}$$

- Matriz de covariância:

$$\overline{\mathbf{U}}_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m)(\mathbf{o}_t^d - \boldsymbol{\mu}_{jm})(\mathbf{o}_t^d - \boldsymbol{\mu}_{jm})' / \hat{c}_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m) / \hat{c}_t^d} \tag{2.32}$$

onde D é o número de sentenças de treinamento e T_d é o número de quadros extraídos da sentença d. O apóstrofo indicado na Equação (2.32) representa a operação de transposição da matriz.

Em reconhecimento automático de fala normalmente se emprega matriz de covariância diagonal (considerando que os componentes do vetor de parâmetros são independentes entre si). Neste caso, G apresentado na Equação (2.28) representa o produto de Gaussianas unidimensionais onde cada uma está associada a uma dimensão do vetor de parâmetros.

2.4. Algoritmo de Viterbi

O algoritmo de Viterbi é utilizado para encontrar uma seqüência ótima de estados relacionada com a seqüência de observação **O**. O algoritmo encontra a seqüência de estados ótima q_t^*, dentre todas as possíveis seqüências q, utilizando o seguinte critério:

$$q_t^* = \arg\max P(q_t = i, \mathbf{O} \mid \lambda) \tag{2.33}$$

Inicialmente, para a aplicação do algoritmo de Viterbi, é definida a probabilidade:

$$\delta_t(i) = \max_{q_1, q_2, \ldots, qt} P[q_1 q_2 \ldots q_{t-1}, q_t = i, \mathbf{o}_1 \mathbf{o}_2 \ldots \mathbf{o}_t \mid \lambda] \tag{2.34}$$

que representa o maior valor da probabilidade em um caminho, no instante de tempo *t*. A Equação (2.34) pode ser reescrita da seguinte forma:

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] . b_j(\mathbf{o}_{t+1}) \tag{2.35}$$

Os passos básicos para a implementação do algoritmo de Viterbi são:

1. Inicialização – para todos os estados *i*:

$$\delta_1(i) = \pi_i b_i(\mathbf{o}_1) \text{ , para } 1 \le i \le N \tag{2.36}$$

$$\psi_1(i) = 0 \tag{2.37}$$

2. Recursão

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] \; b_j(\mathbf{o}_t) \text{ , para } \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{2.38}$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) \; a_{ij}] \text{ , para } \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{2.39}$$

3. Término

$$p^* = \max_{1 \le i \le N} [\delta_T(i)] \tag{2.40}$$

$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)] \tag{2.41}$$

4. Seqüência de estados ótimos (*backtracking*)

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \text{ , para } t = T-1, T-2, ..., 1 \tag{2.42}$$

Na prática os valores assumidos pela variável δ podem ser significativamente menores do que 1, podendo exceder a faixa de precisão dupla dos computadores. De forma a contornar esse problema é utilizado o logaritmo da Equação (2.38):

$$\delta_t(j) = \log_{10}\left(\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]\ b_j(\mathbf{o}_t)\right) \tag{2.43}$$

Quando aplicado à segmentação automática de fala, para determinar as marcas de segmentação que estão armazenadas na variável q_t^*, é necessário "contar" o número de ocorrências de um determinado estado e fazer uma relação com as janelas da locução que foram processadas. A Figura 2.3 representa graficamente o funcionamento do algoritmo de Viterbi para um modelo HMM com três estados, do tipo *left-right*, sem salto duplo.

Cada coluna da Figura 2.3 armazena 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.

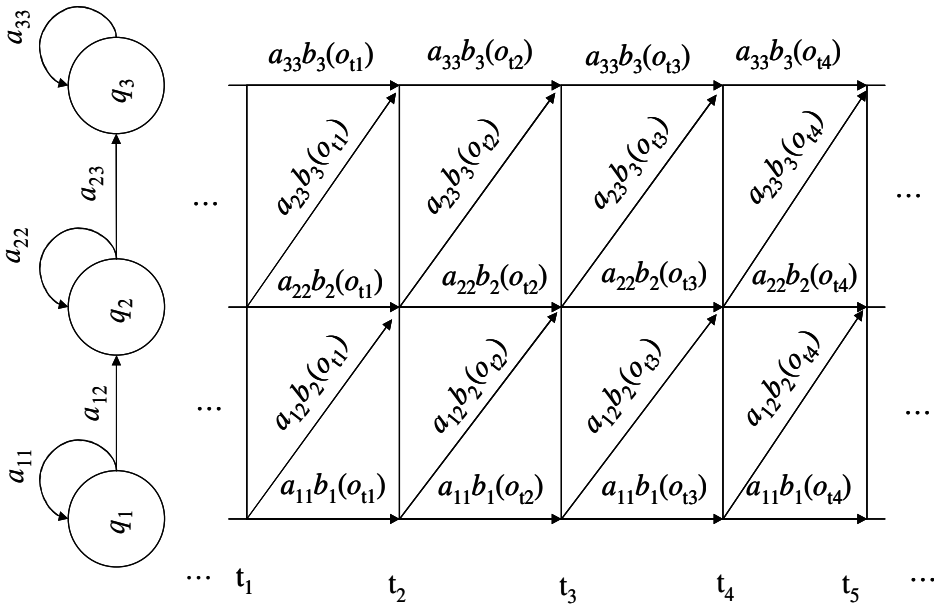


Figura 2.3: Exemplo de funcionamento do algoritmo de Viterbi.

2.5. HTK

O HTK (*Hidden Markov Models Toolkit*) é um kit de desenvolvimento para construir e manipular HMMs. A primeira versão do HTK foi disponibilizada em 1989 e foi desenvolvida pelos pesquisadores do Departamento de Engenharia da Universidade de Cambridge (HTKBook, 2006).

O kit é composto por um conjunto de módulos e ferramentas com código fonte escrito em linguagem C. Esse conjunto foi desenvolvido para aplicações em reconhecimento automático de fala, mas também pode ser utilizado para outras aplicações como síntese de fala, reconhecimento automático de caracteres dentre outras. Atualmente é uma das ferramentas mais divulgadas e utilizadas por permitir uma fácil integração com outras linguagens, como a linguagem de programação C.

As ferramentas contidas no HTK provêem funcionalidades para o treinamento de HMMs, reconhecimento de locuções e fones e análise dos resultados. O *software* também suporta modelos de misturas de Gaussianas para processamento de fala contínua e também distribuições discretas.

2.6. Considerações Finais

Neste Capítulo foram abordadas as principais tecnologias envolvidas em segmentação automática de fala: HMM e alinhamento de Viterbi. Como já destacado, os HMMs serão utilizados para a modelagem das subunidades acústicas e o algoritmo de Viterbi será responsável por gerar a estimativa das fronteiras de segmentação. Essas fronteiras serão corrigidas pelo algoritmo de refinamento.

No sistema proposto que será descrito no Capítulo 5, haverá um módulo responsável pelo treinamento dos HMMs e também um módulo para a segmentação automática. O módulo de treinamento implementa o algoritmo de Baum-Welch e o módulo de segmentação implementa o alinhamento de Viterbi. Tanto o treinamento dos HMMs quanto o alinhamento forçado de Viterbi foram realizados no HTK.

Capítulo 3

Segmentação Automática de Fala

Como já destacado, a segmentação automática de fala desempenha um papel de extrema importância nos processos de síntese e reconhecimento automático de fala.

Ao longo das várias décadas de pesquisa em segmentação de fala, diversos métodos e modelos foram propostos, todos apresentando vantagens, desvantagens e limitações. Dentre as várias pesquisas realizadas, diversos resultados importantes foram reportados.

Este Capítulo tem como objetivo principal apresentar uma revisão do estado da arte em segmentação e refinamento da segmentação automática de fala, destacando os principais métodos, modelos e resultados obtidos em diversas pesquisas. Uma revisão bibliográfica será apresentada para fornecer o panorama atual das pesquisas e também para justificar o método adotado neste trabalho.

3.1. Segmentação Automática de Fala

O problema geral da segmentação automática de fala pode ser definido dentro da estrutura geral de reconhecimento de padrões. A Figura 3.1 mostra de forma ilustrativa esse processo (Vidal and Marzal, 1990).

Na Figura 3.1, para que o sistema possa gerar automaticamente as fronteiras de segmentação, é necessário que o mesmo tenha como entrada o sinal de fala, a partir do qual podem ser extraídas informações no domínio do tempo ou da freqüência. Dependendo também do sistema, a transcrição fonética das locuções pode ou não estar presente.

O sistema pode também ter um processo de aprendizagem que é "alimentado" com as informações da representação da fala e também com as próprias informações sobre as fronteiras de segmentação obtidas. Todas essas informações, por sua vez, são armazenadas em forma de

"conhecimento", que será utilizado posteriormente pelo sistema para gerar novas fronteiras de segmentação mais refinadas.

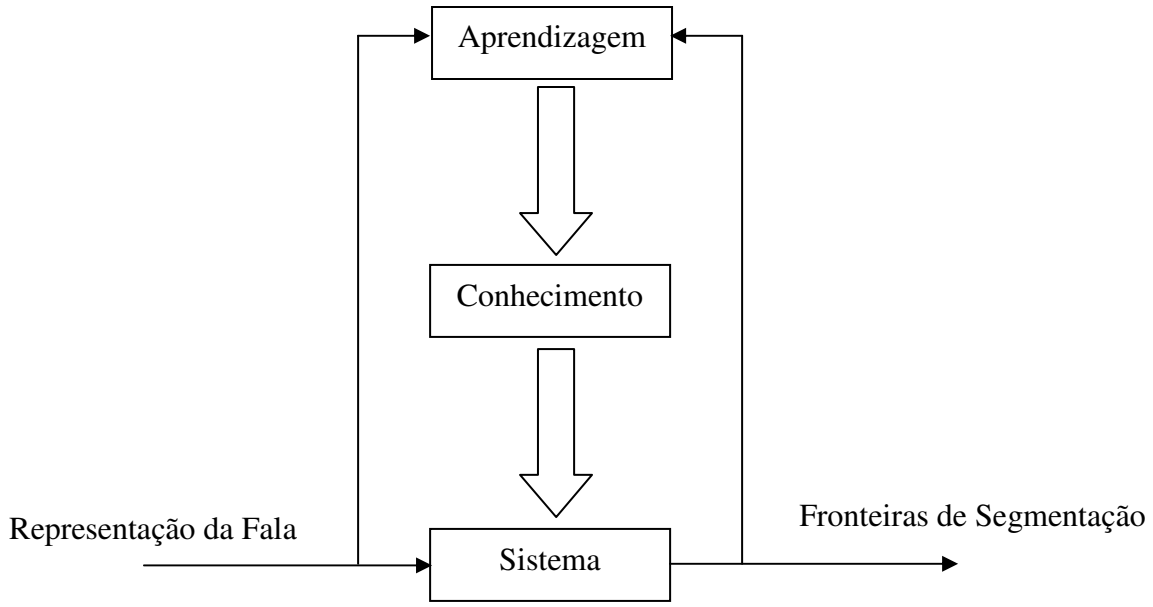


Figura 3.1: Segmentação vista como um problema de reconhecimento de padrões (Adaptado de Vidal e Marzal, 1990).

Por ser um sistema fundamentado nas idéias do reconhecimento de padrões, o seu comportamento pode ser descrito matematicamente por uma função que faça um mapeamento entrada-saída. Nos sistemas para segmentação, a entrada é representada por informações extraídas do sinal de fala que será segmentado, que neste trabalho serão referenciadas por **Observações Acústicas**. Uma outra informação que pode ou não ser apresentada são as informações lingüísticas que serão referenciadas por **Categoria Lingüística** (Vidal and Marzal, 1990). Por outro lado, a saída do sistema é composta por uma seqüência de valores que representam os instantes onde ocorre a transição entre uma unidade fonética e outra.

Os métodos de segmentação podem ser classificados de acordo com a presença ou ausência da categoria lingüística. Se a categoria lingüística não estiver presente, a segmentação é dita **Lingüisticamente Irrestrita** e, neste caso, apenas com as observações acústicas o sistema irá gerar as fronteiras de segmentação. Por outro lado, se a categoria lingüística estiver presente, a segmentação é dita **Lingüisticamente Restrita**.

Na prática é muito comum fazer referência às Observações Acústicas como simplesmente janelas de análise. Normalmente as janelas são representadas por um vetor de parâmetros de dimensão *dim*, que representa as informações do sinal de fala em um curto intervalo de tempo. A Categoria Lingüística pode ser representada pela transcrição fonética da locução que pode ou não ser apresentada como entrada para o sistema.

A classificação Lingüisticamente Irrestrita e Lingüisticamente Restrita é referenciada por (van Hemert, 1991) como **Segmentação Implícita** e **Segmentação Explícita** respectivamente.

Os métodos de segmentação implícita dividem uma dada locução em segmentos (fones, palavras ou sílabas) sem o conhecimento de nenhuma informação explícita, como por exemplo, a transcrição fonética da locução. Os segmentos são calculados com base nas informações espectrais contidas na própria locução, ou seja, com base em informações implícitas.

Por outro lado, os métodos baseados na segmentação explícita realizam a segmentação de uma locução com base na transcrição fonética (informação explícita). O número de segmentos calculados é determinado pela transcrição fonética da locução que deve ser fornecida ao sistema. A princípio, uma grande desvantagem desses métodos é que a transcrição fonética deve ser gerada antes do processo de segmentação. A Tabela 3.1 resume as principais características dos dois métodos de segmentação apresentados.

Tabela 3.1: Principais características dos métodos de segmentação implícita e segmentação explícita.

Segmentação Implícita	**Segmentação Explícita**
▪ Nem sempre o número apropriado de segmentos é determinado; ▪ Os segmentos não são rotulados; ▪ Pode haver inserções e/ou reduções de fronteiras entre os fones; ▪ É comum ocorrer sobre-segmentação, ou seja, surgimento de um número de fronteiras acima do apropriado;	▪ O número de segmentos é determinado em função da transcrição fonética; ▪ Os segmentos são rotulados baseados na transcrição fonética; ▪ Pode haver inconsistências devido a uma fraca modelagem das unidades acústicas;

Na segmentação implícita, Hemert (van Hemert, 1991) destaca que nem sempre os métodos determinam o número correto de segmentos, podendo haver inserções de segmentos adicionais e até mesmo a remoção. Isso pode ser explicado pela falta de informações explícitas sobre a locução que está sendo segmentada.

Em Vidal e Marzal (1990), duas outras classificações são apresentadas para a segmentação. A primeira delas se refere à presença ou não de algum tipo de modelo para as transições entre os segmentos acústicos. Caso existam modelos, a segmentação é dita **Dependente de Modelo**, caso contrário **Independente de Modelo**. A segunda classificação se refere à presença ou não do bloco de treinamento, conforme mostrado na Figura 3.1. Se o bloco de treinamento estiver presente, a segmentação é denominada **Segmentação Supervisionada** e, neste caso, há necessidade de material de treinamento para treinar o sistema de segmentação. Por outro lado, se o bloco de treinamento não estiver presente, tem-se uma **Segmentação Não-Supervisionada**. Por exemplo, no caso de modelar as transições entre os fones, é necessário treinar os modelos de transição para, em seguida, segmentar as locuções.

3.2. Segmentação Implícita ou Lingüisticamente Irrestrita

Este tipo de segmentação corresponde ao processo no qual uma dada locução é segmentada em unidades fonéticas (fones, sílabas ou palavras) sem nenhum conhecimento externo (transcrição fonética, por exemplo) a respeito da locução. Toda a informação necessária para o processo é obtida a partir da própria locução.

Uma série de técnicas para realizar a segmentação implícita surgiu na literatura durante as várias décadas de pesquisa. Dentre as principais destacam-se: segmentação baseada na variação espectral, segmentação multinível, segmentação por decomposição temporal, segmentação baseada em filtragem e segmentação baseada na máxima verossimilhança. A seguir, é apresentada uma breve discussão sobre os principais métodos baseados na segmentação implícita.

3.2.1. Função de Variação Espectral

A segmentação baseada na variação espectral é uma técnica implícita independente de modelo. Considere o vetor de parâmetros $a_t \in$ A, representando as características espectrais do sinal no instante de tempo *t*. A variação espectral do vetor a_t é dada por:

$$a_t^{'} = \frac{\partial a_t}{\partial t} \tag{3.1}$$

A magnitude da derivada primeira $\left\| a_t^{'} \right\|$ representa a taxa na qual o vetor de características espectrais muda com o tempo *t*. Dessa forma, a segmentação baseada na variação espectral procura o ponto (pico) onde a variação espectral é máxima e o define como uma marca de transição entre os fones da locução.

A derivada da Equação (3.1), em aplicações onde as observações são discretas, pode ser aproximada por uma equação de diferenças. Segundo Rabiner (Rabiner and Juang, 1993), tanto as equações de diferenças de primeira ordem como de segunda ordem são aproximações ruidosas da derivada, e o ideal é utilizar uma aproximação polinomial obtida através de uma estimativa dos mínimos quadrados. Normalmente, a magnitude da variação espectral é definida utilizando-se uma janela de ponderação w_θ, para $-\Theta \leq \theta \leq \Theta$. A expressão descrita pela Equação (3.1) pode ser reescrita da seguinte forma:

$$\left\| a_t^{'} \right\| = \frac{\left\| \sum_{-\Theta \leq \theta \leq \Theta} \theta \; w_\theta \; a_{t+\theta} \right\|}{\sum_{-\Theta \leq \theta \leq \Theta} w_\theta} \tag{3.2}$$

Em aplicações práticas, normalmente é utilizada uma janela retangular para w_θ, definida por (Figueiredo, 1999):

$$w_\theta = \begin{cases} 1, & \\ 0, & \text{caso contrário} \end{cases} \tag{3.3}$$

Com base na Equação (3.3), a Equação (3.2) pode ser reescrita da seguinte forma:

$$\left\| a_t^{'} \right\| = \frac{\left\| \sum_{-\Theta \leq \theta \leq \Theta} \theta \; a_{t+\theta} \right\|}{(2\Theta + 1)} \tag{3.4}$$

Uma questão importante a ser analisada é o parâmetro Θ que deve ser escolhido com muito cuidado, pois para valores pequenos é comum ocorrer sobre-segmentação, e valores altos podem dificultar a localização de segmentos curtos como, por exemplo, para as consoantes plosivas. Vidal (Vidal and Marzal, 1990) destaca que o cálculo da variação espectral deve levar em consideração várias observações acústicas adjacentes ao tempo *t*. A escolha dessas informações é feita através de uma janela de análise, e o seu comprimento e forma devem ser cuidadosamente escolhidos. Os autores não sugerem nenhuma estimativa de valor para o parâmetro Θ.

Algumas alterações da função de variação espectral foram propostas na literatura. Mitchell (Mitchell et al., 1995) propôs o uso da função de variação cepstral (DCF – *Delta Cepstral Function*) utilizada para estimar a variação espectral através da soma da derivada dos componentes cepstrais do sinal. Seu cálculo é dado pela expressão:

$$DCF_i(t) = C_i(t+1) - C_i(t-1) \text{ , para } i = 1,..., Q_c \tag{3.5}$$

onde $C_i(t)$ representa o *i*-ésimo coeficiente cepstral para a janela de análise *t* e Q_c é o número de coeficientes cepstrais. A Equação (3.5) é utilizada para calcular uma função de custo $c(t)$ responsável por detectar as mudanças espectrais que estão associadas às transições entre os fones. A função de custo $c(t)$ é computada utilizando a Equação (3.6).

$$c(t)_{DCF} = \frac{\sum_{i=1}^{Q_c} \frac{DCF_i(t)}{\max_i \left| DCF_i(t) \right|}}{\max_i \sum_{i=1}^{Q_c} \frac{DCF_i(t)}{\max_i \left| DCF_i(t) \right|}} \tag{3.6}$$

A função de variação espectral também já foi utilizada por alguns autores (Brugnara et al., 1992 e Mitchell et al., 1995) para estimar a variação espectral como o ângulo entre dois vetores cepstrais normalizados, separados no domínio do tempo por um número fixo de janelas de análise. Esta operação é realizada pela Equação (3.7):

$$F(t) = \frac{\hat{C}(t-1) \bullet \hat{C}(t+1)}{\left\|\hat{C}(t-1)\right\| \cdot \left\|\hat{C}(t+1)\right\|} \tag{3.7}$$

onde $\hat{C}(t)$ é a diferença entre o vetor cepstral e a média dos vetores cepstrais dentro de uma janela de análise centrada em *t,* e $\bullet$ indica a operação de produto escalar. A função de custo $c(t)$ para essa operação é calculada através da Equação (3.8).

$$c(t)_F = \frac{1}{2}\left(1 - \frac{F(t)}{\max\left|F(t)\right|}\right) \tag{3.8}$$

Como observado na prática, a segmentação através da técnica de variação espectral não apresenta bom desempenho devido ao problema da sobre-segmentação. Por outro lado, pode ser uma boa ferramenta a ser utilizada como ponto de partida para a segmentação manual. Uma vantagem dessa técnica é estimar as fronteiras entre os fones independente de modelos e também sem a necessidade de estimar limiares.

Algumas formas de atenuar os problemas encontrados nas técnicas baseadas em variação espectral foram propostas na literatura. Dentre as técnicas mais citadas destacam-se o espaço escalonado, proposta por Witkin (Witkin, 1989), que consiste em obter uma descrição estruturada do sinal de fala em diferentes "escalas de resolução" (variando o parâmetro Θ). A segunda, proposta por Glass (Glass and Zue, 1988), consiste em um procedimento de clusterização das observações acústicas.

3.2.2. Quantização Vetorial por Agrupamento

Uma estratégia alternativa para o problema de segmentação foi proposta por Svendsen (Svendsen, 1987) e consiste em agrupar os vetores acústicos de modo a maximizar o grau de homogeneidade de cada segmento. Para tanto, propõe-se uma medida de distorção que é utilizada para encontrar a segmentação ótima, através de um algoritmo de programação dinâmica.

A idéia consiste em obter uma seqüência de vetores acústicos de entrada (a_i , i = 1...m), organizados em segmentos delimitados pelo conjunto de fronteiras (b_j, j = 1...l). Para cada segmento tem-se o valor de seu centróide (c_k, k = 1...l) e, dessa forma, pode-se aplicar uma

medida que representa a distorção global da segmentação. Com base nessa medida de distorção global, a idéia da segmentação baseada na quantização vetorial por agrupamento é determinar a segmentação ótima da seqüência de entrada, minimizando o valor da distorção global.

3.2.3. Redes Neurais Artificiais

Com o avanço das redes neurais artificiais e seu sucesso em diferentes áreas, novos métodos de segmentação automática irrestrita, que utilizam diretamente técnicas de Reconhecimento de Padrões baseadas em Redes Neurais, têm sido propostos na literatura.

Em (Suh and Lee, 1996), utiliza-se um método supervisionado baseado em uma rede do tipo MLP (*MultiLayer Perceptron*), treinada com um algoritmo *Back-Propagation* modificado, a fim de obter uma estimativa das fronteiras dos segmentos acústicos em fala contínua. Em (Rubio and Reilly, 1995) e (Fukada et al., 1997), são propostos métodos que utilizam Redes Recorrentes para determinar as fronteiras dos segmentos em aplicações envolvendo fala contínua, uma vez que tais estruturas são mais apropriadas para modelar a variabilidade temporal dos padrões da voz humana.

Modelos de redes neurais artificiais baseados nos mapas auto-organizáveis de Kohonen também foram aplicados à segmentação automática de fala (Figueiredo, 1999). Os mapas auto-organizáveis de Kohonen apresentam uma técnica de rede neural artificial baseada em treinamento não supervisionado.

3.2.4. Razão de Verossimilhança Generalizada de Brandt

O principal objetivo do método da razão de verossimilhança generalizada (GLR – *Generalized Likelihood Ratio*) de Brandt consiste em detectar descontinuidades no sinal de fala sem a necessidade do conhecimento prévio da transcrição fonética da locução. Os pontos de descontinuidade detectados pelo método representam as fronteiras entre os fones da locução (Jafiri et al., 2006).

O método assume que cada fone, composto por um conjunto de amostras, obedece a um modelo auto-regressivo de ordem p que é constante para todos os fones presentes na locução. A partir de uma janela w_0 do sinal com n amostras e parâmetros Θ_0, o método decide se a janela será dividida em dois subsegmentos w_1 e w_2 através da detecção de descontinuidade entre seus

parâmetros Θ_1 e Θ_2. A descontinuidade em uma janela do sinal é calculada usando a Equação (3.9).

$$D(i) = n\log\hat{\sigma}_0 - r\log\hat{\sigma}_1 - (n-r)\log\hat{\sigma}_2 \tag{3.9}$$

onde r é o tamanho da janela w_1, e $\hat{\sigma}_0$, $\hat{\sigma}_1$ e $\hat{\sigma}_2$ são respectivamente as estimativas do desvio padrão do ruído dos modelos caracterizados pelos parâmetros Θ_0, Θ_1 e Θ_2. A divisão da janela (ponto de descontinuidade entre os fones) ocorre na amostra em que o valor de $D(i)$ é máximo. A Equação (3.9) define a razão de verossimilhança.

A grande desvantagem do método é o número de inserções e deleções em virtude de ser um método de segmentação lingüisticamente irrestrito.

3.2.5. Critério de Informação Bayesiana

Como já destacado, a teoria estatística está intimamente relacionada com o reconhecimento automático e a síntese de fala.

Métodos utilizados para testes de hipótese e para seleção de modelos também têm sido empregados para a segmentação automática de fala. Dentre esses métodos destaca-se o Critério de Informação Bayesiana (BIC – *Bayesian Information Criterion*) (Chen and Gopalakrishnan 1998), (Almpanidis and Kotropoulos, 2008). Este método é largamente utilizado para a identificação de modelos em modelagem estatística, séries temporais, regressão linear, determinação do número de Gaussianas em HMM para o reconhecimento de fala (Yared et al., 2006), dentre outras aplicações.

Para detectar pontos de mudança acústica no sinal (segmentação automática de fala) usando o BIC, segmentos adjacentes são modelados usando diferentes distribuições de Gaussianas multivariadas. A concatenação desses segmentos por sua vez obedece a uma terceira distribuição. O problema consiste em decidir se o modelo definido pelo segmento maior (concatenação entre os dois segmentos adjacentes) é melhor do que a representação através de dois segmentos menores. Neste caso o BIC é utilizado para selecionar o melhor modelo que se adapta à modelagem proposta.

Considere que H_0:(a_1, a_2, ..., a_n) ~ N(μ_0, Σ_0) seja a seqüência de vetores de características acústicas para o segmento maior, e H_1:(a_1, a_2, ..., a_m) ~ N(μ_1, Σ_1) e H_2:(a_{m+1}, a_{m+2}, ..., a_n) ~ N(μ_2, Σ_2) a seqüência de vetores de características acústicas do primeiro e do segundo segmento respectivamente. Os vetores a_i de dimensão p podem representar os coeficientes mel-cepstrais obtidos a partir de cada segmento, Σ_0, Σ_1 e Σ_2 são as matrizes de covariância completas para cada segmento. A Figura 3.2 representa o modelo para dois segmentos adjacentes hipotéticos.

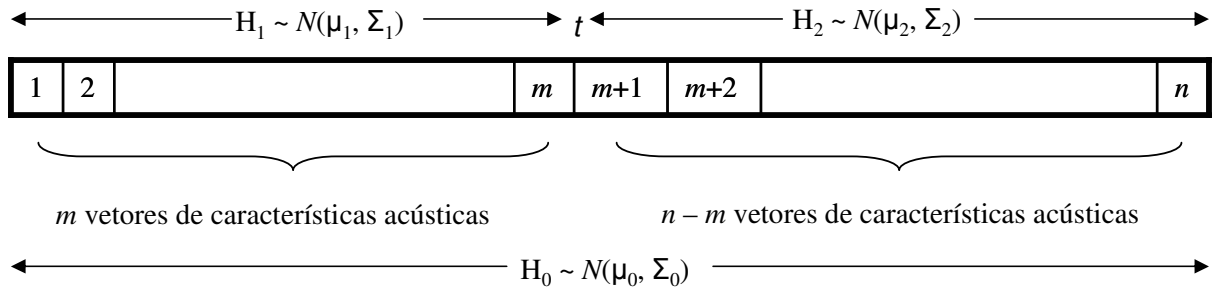


Figura 3.2: Modelo para dois segmentos adjacentes de fala.

A variação do valor do BIC entre os modelos é dada por:

$$BIC(i) = R(i) - \lambda P_0 \tag{3.10}$$

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

$$R(i) = n\log|\Sigma_0| - m\log|\Sigma_1| - (n-m)\log|\Sigma_2| \tag{3.11}$$

O parâmetro P_0 é o fator de penalização para a complexidade do modelo, e seu valor é calculado usando a Equação (3.12):

$$P_0 = \frac{1}{2}\left(p + \frac{1}{2}p(p+1)\right)\log n \tag{3.12}$$

Na Equação (3.10), o parâmetro λ representa o peso para o fator de penalização. Seu valor pré-definido é 1. O ponto de mudança acústica calculado através do BIC ocorre no centro da janela de análise em que o valor de $BIC(i)$ é máximo, para todos os valores de i.

3.3. Segmentação Explícita ou Lingüisticamente Restrita

Como apresentado na seção anterior, a grande desvantagem dos métodos baseados na segmentação implícita é a ocorrência da sobre-segmentação ou da omissão de fronteiras que pode ser prejudicial para as aplicações que utilizam a segmentação automática.

Uma possibilidade para reduzir esses problemas é a utilização de informações explícitas da locução que está sendo segmentada, evitando dessa forma a sobre-segmentação e a omissão de fronteiras. A segmentação explícita é também chamada de restrita porque o número de fronteiras que serão determinadas pelo processo de segmentação é restrito ao número de símbolos presentes na transcrição fonética da locução.

As informações de natureza lingüística a serem adicionadas ao sistema de segmentação são obtidas a partir de uma seqüência de elementos do conjunto das categorias lingüísticas como descrito na Seção 3.1. O conjunto de categorias lingüísticas é representado pelos fones que compõem a transcrição fonética de cada locução.

A grande vantagem da segmentação explícita em relação à segmentação implícita é que o número de fronteiras geradas corresponde ao número de fones presentes na transcrição fonética, não havendo, portanto, uma sobre-segmentação ou omissão de fronteiras. Por outro lado, a desvantagem é que a transcrição fonética de cada locução que será segmentada deve ser gerada antes do processo de segmentação.

O método mais popular usado para a segmentação explícita é o HMM baseado no alinhamento forçado de Viterbi que, a partir de uma locução e de sua transcrição fonética gera as estimativas das fronteiras entre os fones que compõem a locução. Os locais das fronteiras estimadas não são necessariamente os locais das fronteiras da segmentação manual, uma vez que o algoritmo de Viterbi procura pela seqüência mais provável de fones e não por pontos de descontinuidade acústica para gerar as fronteiras de segmentação. Outro ponto a ser destacado no alinhamento de Viterbi é com relação à precisão das fronteiras que está limitada pelo tamanho do quadro (intervalo entre as janelas de análise adjacentes), normalmente na ordem de 10 ms. Para

corrigir essa discrepância, métodos de refinamento são empregados para corrigir a limitação do HMM quando usado em segmentação automática de fala.

3.4. Avaliação da Segmentação Automática

Quando um sistema para segmentação automática de fala é implementado ou técnicas de refinamento são aplicadas às marcas de segmentação já existentes, deseja-se obter medidas que possam avaliar a qualidade da saída de tais sistemas.

Na prática não existe uma medida ou métrica que seja amplamente difundida na comunidade científica. Basicamente a avaliação da segmentação automática pode ser realizada de forma subjetiva ou objetiva. Na avaliação subjetiva, os segmentos de fala determinados pelas marcas de segmentação podem ser ouvidos para analisar a qualidade da segmentação obtida.

A forma mais comum de avaliação e também a mais direta consiste em fazer uma comparação entre as marcas de segmentação obtidas através de um segmentador e as marcas de segmentação manual obtidas por um foneticista. Na comparação entre as segmentações normalmente é reportada uma porcentagem das fronteiras cujo erro de segmentação está dentro de um certo limiar. O limiar mais utilizado é de 20 ms (Toledano et al., 1998, 2000, 2003), (Honsom, 2003).

Várias outras medidas podem ser feitas de forma a refletir o erro entre a segmentação automática e a segmentação manual. Dentre essas medidas destacam-se a raiz quadrada do erro médio quadrático e o erro médio absoluto que pode ser representado pela Equação (3.13)(Figueira and Oliveira, 2008).

$$EMA = \frac{1}{N_F} \sum_{i=0}^{N_F} |e_i| \tag{3.13}$$

onde $|e_i|$ representa o erro para a fronteira i e N_F é o número total de fronteiras.

3.5. Estado da Arte em Segmentação Automática de Fala

Nesta Seção, pretende-se fornecer um panorama abrangente sobre as principais técnicas utilizadas para a segmentação automática de fala. As técnicas apresentadas nas Seções anteriores são conhecidas na literatura como técnicas clássicas de segmentação. Basicamente os trabalhos

apresentados nesta Seção são fundamentados nas técnicas apresentadas nas Seções anteriores. Algumas técnicas apresentadas nesta revisão bibliográfica utilizam a combinação de duas ou mais técnicas anteriores e também novas técnicas foram desenvolvidas.

Como apresentado na Seção 3.1, as técnicas de segmentação dividem-se em dois grupos: segmentação implícita e segmentação explícita. Como ambas as técnicas apresentam vantagens e desvantagens, alguns trabalhos utilizaram uma combinação das duas com o objetivo de maximizar os resultados da segmentação.

Um dos primeiros relatos da combinação dos dois métodos data de 1991, publicado por Jan P. van Hemert (van Hemert, 1991). Em seu trabalho, inicialmente as locuções são submetidas à segmentação implícita e, em seguida, à segmentação explícita. Para a segmentação implícita, as fronteiras entre os fones são determinadas através do grau de similaridade entres duas janelas vizinhas com base no espectro de freqüências. Este grau de similaridade adotado por Hemert foi a correlação entre as janelas.

Na segmentação explícita, a seqüência de fones da transcrição fonética é convertida em "estados espectrais" que atuam como uma referência do espectro para cada fone. O algoritmo de segmentação determina as fronteiras entre os fones baseado na correlação entre o espectro da locução e o espectro de referência criado a partir da transcrição fonética.

Tendo os resultados da segmentação implícita e explícita, um algoritmo de combinação analisa sempre uma fronteira da segmentação implícita com uma fronteira da segmentação explícita, sempre preservando a precisão da segmentação implícita. Se a segmentação implícita gerar mais fronteiras do que a segmentação explícita, as fronteiras excedentes são ignoradas; caso gere menos, as fronteiras são completadas com as da segmentação explícita. O autor aplicou os métodos de segmentação em um conjunto de 90 locuções. Dos resultados obtidos para a segmentação híbrida, 95% das fronteiras apresentam um erro de segmentação abaixo de 20 ms, enquanto que para a segmentação explícita a taxa foi de 82%.

Como já destacado, os HMMs desempenham papel essencial na área de reconhecimento automático de fala e também são largamente utilizados para segmentação automática de fala como um primeiro estágio. Um dos primeiros relatos de trabalho utilizando HMM para a segmentação de fala data de 1988 (Nakagawa and Hashimoto, 1988). O trabalho descreve o processo de segmentação implícita onde as locuções são segmentadas em sílabas. Os testes foram realizados em uma base de fala da língua Japonesa e a porcentagem de acertos foi de 97,5% (para

erros abaixo de 20 ms), tendo apenas 1,2% de erros de inserção e 1,2% para erros de omissão de fronteiras.

Uma técnica baseada na segmentação implícita realizada em dois estágios foi proposta por Aversano em 2001 (Aversano et al., 2001). O primeiro estágio consiste em realizar um pré-processamento no sinal de fala seguido por um processo de detecção das fronteiras entre os fones. No pré-processamento, cada janela do sinal é submetida a uma análise baseada em psicoacústica, que é capaz de extrair características acústicas que capturam informações sobre a transição entre os fones. O método desenvolvido utiliza o espectro modificado para a detecção das fronteiras entre os fones sem haver a necessidade de qualquer tipo de filtragem no sinal ou modelagem preditiva.

O algoritmo proposto por Aversano para determinar as fronteiras consiste em detectar picos na seqüência de fala. Esses picos correspondem às janelas em que as características acústicas obtidas na fase de pré-processamento mudam significativamente. Os picos são calculados utilizando a diferença absoluta entre o valor médio da seqüência de fala no instante de tempo *i*, calculado usando *n* frames antes e após esse instante.

Os autores destacam que a grande vantagem de usar conceitos da psicofísica da audição com o objetivo de estimar o espectro auditivo é gerar um número relativamente baixo de parâmetros para cada janela. Outro ponto destacado é que cada parâmetro quantifica a energia espectral em um intervalo de freqüência determinado. Os autores obtiveram uma taxa de 73,58% de fronteiras de segmentação detectadas com erro menor que 20 ms, e nenhum problema de inserção de novas fronteiras.

Uma combinação do uso de regras fuzzy com redes neurais artificiais foi proposta em 1999 por Hsieh (Hsieh et al., 1999). A idéia foi desenvolver um sistema para segmentação automática de fala contínua em sílabas, dividido em duas fases. Na primeira fase, uma rede neuro-fuzzy é utilizada para classificar o sinal de fala em três tipos diferentes (silêncio, consoante e vogal). Para realizar a classificação dos segmentos são utilizados dois parâmetros temporais: a taxa de cruzamentos por zero e a energia total da janela de análise.

A segunda fase (após a classificação em três classes fonéticas) consiste na segmentação propriamente dita, utilizando uma rede neural com o algoritmo de retropropagação (*backprogagation*). As locuções utilizadas nos testes e no treinamento da rede neural foram extraídas de jornal e pronunciadas por dois locutores do sexo masculino e dois do sexo feminino,

em Mandarin. Uma análise dos resultados mostra que as taxas de erro quando comparadas com a segmentação manual são baixas. A grande desvantagem do método é a fraca generalização das redes neurais aplicadas ao processamento de fala.

Os vários trabalhos apresentados até este ponto fazem normalmente uso de um modelo para representar as unidades fonéticas (HMM, por exemplo). Nagarajan (Nagarajan et al., 2003) (Nagarajan and Murthy, 2004) apresenta uma técnica de segmentação implícita baseada na técnica de atraso de grupo de fase mínima (*minimum phase group delay*) do sinal para segmentar a fala contínua em sílabas.

Toda a técnica desenvolvida por Nagarajan é baseada no cálculo da energia do sinal em curtos intervalos de tempo mas, como destacado pelos autores, alguns problemas podem ocorrer quando a energia é utilizada para a segmentação automática. Um dos problemas consiste na necessidade de um limiar para decidir a posição das fronteiras de segmentação, e o outro são as flutuações da energia que podem ocorrer devido à presença de consoantes na locução.

Para contornar os problemas apresentados, ao invés de usar a energia obtida diretamente do sinal, um sinal de fase mínima é obtido a partir da função de energia em curtos intervalos de tempo, como se fosse a magnitude do espectro do sinal. Segundo os autores, a função de atraso de grupo do sinal de fase mínima é a melhor representação da energia em curtos intervalos de tempo para realizar a segmentação. O atraso de grupo é definido como a derivada negativa da fase da transformada de Fourier do sinal (Nagarajan et al., 2001).

A fundamentação teórica do trabalho desenvolvido por Nagarajan já havia sido desenvolvida por Hema (Murthy and Yegnanarayana, 1991) e (Murthy, 1997). Os autores afirmam que, uma vez que o sinal é derivado a partir de uma função positiva (similar ao espectro da magnitude), o sinal resultante sempre apresentará fase mínima. Hema destaca que se o sinal apresenta fase mínima, as funções de atraso de grupo resolvem bem os picos e os vales do espectro do sinal. Os locais dos picos da função de atraso de grupo do sinal de fase mínima representam as fronteiras entre as sílabas. Os testes foram realizados em duas bases de fala da língua Hindi. Em uma das bases, 66,93% das fronteiras apresentaram um erro abaixo de 25 ms e, em outra base, 76,58%, com uma taxa de aproximadamente 5% de inserção e 4,38% de omissão.

Técnicas de processamento digital de sinais, tais como a filtragem, também já foram aplicadas em segmentação automática de fala. No trabalho desenvolvido por Li e Gibson (Li and Gibson, 1996), uma técnica chamada de filtragem paramétrica pelos autores é empregada com o

objetivo de detectar mudanças no sinal de fala e, dessa forma, detectar as fronteiras de segmentação. O sinal a ser segmentado é filtrado por um banco de filtros com atraso unitário. Em seguida, é realizada uma análise da autocorrelação do sinal e medidas de distorções são empregadas para determinar as similaridades entre as janelas.

A motivação do trabalho está no fato de que a estrutura de autocorrelação de um sinal estacionário pode ser caracterizada por certas estatísticas da saída de um banco de filtros, como destacado por Li em trabalhos anteriores (Li, 1996), (Li and Gibson, 1994).

Os testes foram realizados com diversas palavras corrompidas por ruído gaussiano branco e os resultados mostraram um posicionamento correto das marcas de segmentação quando comparadas com a segmentação manual.

Uma técnica relativamente nova e bastante difundida na área de classificação de padrões é baseada nas Máquinas de Vetor de Suporte (*Support Vector Machines* – SVMs). Em seu trabalho inicial, Vapnik (Vapinik, 1995) mostrou que essa técnica é muito eficiente na detecção de características e na classificação de fones em fala contínua (Niyogi, 1998). Como reportado em diversos trabalhos, as máquinas de vetor de suporte têm capacidade de aprendizagem a partir de pequenas quantidades de dados quando comparadas a outras técnicas como redes neurais artificiais, mas ainda são técnicas estatísticas limitadas para modelar as variações temporais e as coarticulações da fala.

Amit Juneja em 2003 (Juneja and Espy-Wilson, 2003) propôs um método que combina máquinas de vetor de suporte com características fonéticas para segmentar fala contínua em cinco classes: vogais, consoantes sonoras, fricativas, plosivas e silêncio. No sistema desenvolvido pelos autores, denominado de Sistema Baseado em Eventos (*event-based system*), a fala é primeiro segmentada nas classes já mencionadas através de uma hierarquia de características fonéticas e máquinas de vetor de suporte.

A idéia do sistema é estender os conceitos apresentados para o reconhecimento de fones, usando características de sonoridade e também do local de articulação. Para a classificação e o reconhecimento são utilizados treze parâmetros baseados no conhecimento das características acústicas dos fones.

O método implementado foi testado usando a base TIMIT e uma comparação foi realizada com o modelo HMM. Para um HMM com 39 parâmetros acústicos a taxa de acerto para erros menores do que 20 ms foi de 69,6% contra 79,8% do sistema baseado em eventos.

Uma técnica muito interessante para a segmentação automática de fala foi proposta em 2005 por Keshet (Keshet et al., 2005). O método desenvolvido utiliza um algoritmo de aprendizagem supervisionada discriminativa para realizar o alinhamento entre a transcrição fonética e a locução.

O alinhamento dos fones é realizado através de funções que são utilizadas para mapear uma representação acústico-fonética de uma locução em um vetor de espaços representando as fronteiras de segmentação. Os autores definem sete funções de alinhamento.

As primeiras quatro funções têm por objetivo capturar a transição entre os fones e são baseadas na distância entre os vetores acústicos dos dois lados da suposta marca de segmentação. A medida de distância empregada foi a Euclidiana. A quinta função empregada é uma medida de confiança garantindo que um determinado fone foi pronunciado em um determinado frame. A sexta função é baseada na duração do fone, e a sétima e última é baseada na taxa de elocução de um locutor.

O algoritmo de treinamento recebe como entrada para cada locução um conjunto formado pelo vetor de parâmetros, um símbolo representando o fone e um valor representando a suposta marca de segmentação. O treinamento consiste em determinar a partir do conjunto de treinamento um vetor de pesos que, quando multiplicado pelo conjunto de funções, determina a marca correta de segmentação. O método proposto foi testado na TIMIT e 92,3% das fronteiras apresentam um erro menor que 20 ms.

Técnicas de segmentação que levam em conta a variação de energia do sinal para determinar as fronteiras entre os fones estão sempre sendo propostas. Golipour e O'Shaughnesy (2007) propõem um método de segmentação automática de fala em fones que localiza os principais picos de variação de energia no domínio da freqüência e os define como as fronteiras de segmentação.

Em um primeiro momento a transformada de Fourier do sinal é calculada para determinar o espectro de potência. Em seguida, o espectrograma é suavizado através de um filtro de médias para eliminar flutuações de energia. Após a suavização, o gradiente é calculado para obter uma medida da variação de energia. Os autores utilizam quatro sub-bandas de freqüências (0-500 Hz, 500-1420 Hz, 1420-2386 Hz e 2386-8000 Hz). Os valores da derivada da energia em cada sub-banda são somados a cada instante de tempo de forma a realçar os picos que caracterizam as fronteiras entre os fones da locução.

Apesar da suavização aplicada no espectrograma, os autores ainda aplicam a função de atraso de grupo para reduzir algumas flutuações indesejadas nas altas freqüências. Por último, o sinal ainda é normalizado e em seguida um algoritmo determina as fronteiras através da análise dos picos produzidos. O algoritmo foi testado em 1300 sentenças da TIMIT e 86,96% das fronteiras apresentam um erro de segmentação abaixo de 20 ms.

3.6. Refinamento da Segmentação Automática de Fala

Quando as técnicas de segmentação automática de fala são aplicadas a uma locução, uma estimativa de fronteira é obtida. Esta estimativa 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. Descobrir esses instantes de tempo é a tarefa da segmentação automática.

Dependendo da técnica utilizada para segmentar uma locução, normalmente as estimativas de fronteiras, quando comparadas com a segmentação manual, apresentam uma discrepância. Essa discrepância pode ser prejudicial em determinadas aplicações, como por exemplo, em síntese de fala. Dessa forma, um pós-processamento ou refinamento da segmentação automática faz-se necessário.

O refinamento da segmentação automática consiste em realizar um processamento em cada fronteira previamente estimada, com o objetivo de diminuir a diferença entre a segmentação automática e a manual. Este processamento é realizado dentro de um intervalo específico, de forma a deslocar as fronteiras inicialmente estimadas para uma nova posição.

Na próxima Seção serão apresentados e analisados alguns parâmetros e também algumas técnicas que normalmente são empregados para o refinamento da segmentação automática de fala.

3.7. Estado da Arte em Refinamento da Segmentação Automática de Fala

Em 2002, Sethy e Narayanan (Sethy and Narayanan, 2002) propuseram uma técnica de refinamento baseado em modelos de fronteiras dependentes de contexto. Segundo os autores, uma fronteira próxima a um fone deve ser determinada pelo seu contexto, uma vez que a mesma pode apresentar características diferentes dependendo do contexto.

A primeira etapa do processo de refinamento consiste em modelar e treinar as fronteiras. Para modelar as fronteiras, os autores usam uma base segmentada e rotulada manualmente. Primeiro um número *n* de frames é estabelecido de ambos os lados da fronteira (totalizando 2*n* frames) e, em seguida, algumas características fonéticas são calculadas para cada frame. Cada fronteira é modelada como um HMM e os parâmetros calculados foram: energia média, parâmetros mel-cepstrais com suas derivadas primeira e segunda e uma medida baseada na autocorrelação para indicar se o frame é surdo ou sonoro.

Durante o processo de refinamento, os mesmos passos utilizados para o treinamento são repetidos nos vizinhos de cada fronteira previamente estimada. A cada passo é calculada uma medida de probabilidade e a posição da nova fronteira é estabelecida no centro do frame que apresenta a maior probabilidade.

O treinamento das fronteiras foi realizado com uma base dependente de locutor contendo 400 locuções. A técnica proposta foi testada em uma base contendo apenas 100 locuções pronunciadas pelo mesmo locutor do treinamento. Pelos testes realizados, 87% das fronteiras apresentam um erro menor que 16 ms.

Doroteo Toledano (Toledano et al., 2003) emprega uma técnica de refinamento baseado em características acústicas e modelos de mistura de Gaussianas para encontrar a fronteira ótima dada a fronteira inicial (estimada pelo alinhamento de Viterbi) e a transcrição fonética da locução.

A partir da fronteira inicial é definida uma região de refinamento que compreende as fronteiras imediatamente anterior e posterior à fronteira que está sendo refinada. De cada lado da fronteira são calculados alguns parâmetros acústicos que são agrupados em um vetor. Este vetor por sua vez é utilizado para estimar uma medida de probabilidade a partir de um modelo de mistura de Gaussianas. Os parâmetros calculados foram: energia, correlação entre o logaritmo da energia do lado esquerdo e direito da fronteira que está sendo refinada, taxa de cruzamentos por zero e derivada da energia. Para o cálculo dos parâmetros foi utilizada uma janela de 20 ms.

Os HMMs foram treinados usando uma base independente de locutor no Espanhol Castelhano. Em seguida foi realizada uma adaptação de locutor e, após o refinamento, 96,01% das fronteiras apresentaram um erro de segmentação abaixo de 20 ms.

Trabalho semelhante ao proposto por Sethy e Narayanan foi proposto em 2004 por Wang (Wang et al., 2004). Neste trabalho, as fronteiras inicialmente estimadas pelo alinhamento forçado de Viterbi são também refinadas usando modelos de fronteiras dependentes de contexto.

Os modelos de fronteiras definidos por Wang seguem os mesmos padrões dos modelos definidos por Sethy, salvo que cada fronteira, ao invés de ser modelada por um HMM, é modelada por uma mistura de Gaussianas.

Para o treinamento, um número n de frames é determinado de ambos os lados de cada fronteira, centrados na fronteira que está sendo analisada (total de $2n+1$ frames). Em seguida, para cada frame são extraídas m características acústicas que por sua vez são agrupadas em um super vetor de dimensão $(2n+1)m$. Este super vetor de características é utilizado para treinar um modelo de mistura de Gaussianas.

O ideal nesta técnica de refinamento é treinar um modelo para cada par de fones que define uma fronteira. O grande problema é a falta de material de treinamento manualmente segmentado e rotulado para cada modelo. Para contornar esse problema, os autores utilizam árvore de regressão e classificação para agrupar modelos de fronteiras similares em uma mesma categoria. Os modelos de fronteiras que apresentam pouca ocorrência dentro do material de treinamento, ou mesmo não aparecem, são mapeados para um nó específico da árvore. Para cada nó folha da árvore, um modelo de mistura de Gaussianas é treinado para refinar fronteiras que pertencem àquele tipo.

Para o processo de refinamento, as novas fronteiras são estimadas nas proximidades das fronteiras previamente estabelecidas pelo alinhamento forçado de Viterbi. Nesta região, as características fonéticas são calculadas e um nó folha na árvore que corresponde à fronteira que está sendo refinada é encontrado. Em seguida, a verossimilhança para cada frame dentro do espaço de busca é calculada usando os modelos de misturas de Gaussianas para o nó folha encontrado. O frame que apresenta a maior verossimilhança é definido como a fronteira ótima.

Para testar o sistema, foi utilizada uma base de fala do Mandarim manualmente segmentada. Durante o treinamento dos modelos de fronteiras, foram utilizados 5 frames com duração de 20 ms e deslocamento a cada 30 ms. Para cada frame, um vetor de parâmetros de ordem 39 foi calculado. Após o refinamento, 91,9% das fronteiras apresentaram um erro de segmentação menor que 20 ms. Em relação ao trabalho de Sethy houve uma melhora de 4,3%.

Como observado nos resultados dos dois trabalhos anteriores, modelos de fronteiras dependentes de contexto apresentam bons resultados quando aplicados no refinamento da segmentação automática de fala. Uma desvantagem deste método é a exigência de uma grande base de fala segmentada contendo todas as ocorrências dos fones em todos os contextos para o treinamento dos modelos. Nos dois trabalhos descritos o modelo das fronteiras é estimado usando uma base de fala dependente de locutor.

Pensando na produção de novas falas e também em uma forma de minimizar esforços manuais, Zhao (Zhao et al., 2005) propôs adaptar os modelos de fronteiras dependentes de contexto e de locutor para novos locutores. O procedimento para adaptação utilizado pelos autores é semelhante ao utilizado no processo de reconhecimento de fala.

A adaptação ocorre em duas etapas. Na primeira etapa os parâmetros utilizados pelo HMM para gerar as estimativas iniciais das fronteiras são adaptados e, na segunda etapa, os parâmetros dos modelos de fronteiras dependentes de contexto são alterados para cobrir as características do novo locutor. Os autores empregam e avaliam três métodos de adaptação: MAP (*Maximum a Posteriori*), MLLR (*Maximum Likelihood Linear Regression*) e uma combinação dos dois métodos MAP+MLLR.

Os HMMs foram treinados com uma grande base de fala (aproximadamente 12.000 locuções) dependente de locutor e adaptado para outro locutor. Os testes de refinamento foram realizados usando quatro bases menores (aproximadamente 200 locuções). Foi observado que 90% das fronteiras apresentam um erro de segmentação menor que 20 ms. Esses resultados mostram-se bastante expressivos uma vez que os modelos de fronteiras foram adaptados.

Máquinas de vetor de suporte (SVM – *Support Vector Machine*) têm sido aplicadas largamente em reconhecimento automático de fala pelo seu grande poder de classificação. Algumas aplicações começam a surgir em segmentação e refinamento automático de fala. No trabalho desenvolvido por Lo e Wang (2007), máquinas de vetor de suporte são empregadas para refinar marcas de segmentação previamente determinadas pelo alinhamento forçado de Viterbi.

Durante o processo de refinamento, para cada fronteira detectada pelo alinhamento forçado, 16 hipóteses de fronteiras são extraídas a cada 5 ms dentro de uma faixa de ± 40 ms, centrada na fronteira que está sendo analisada. Cada uma dessas fronteiras determinada será examinada pelo classificador SVM dependente de transição e, finalmente, a fronteira mais provável é selecionada para substituir a fronteira inicial.

Para o treinamento dos classificadores, 46 SVM dependentes de transição entre os fones são gerados e usados para o refinamento das marcas de segmentação. Os classificadores são treinados utilizando vetores de parâmetros de ordem 45 formados por: 39 parâmetros mel-cepstrais, taxa de cruzamento por zero, freqüência *bisector*, *burst degree*, entropia espectral, entropia geral ponderada e energia em sub-bandas.

Os testes foram realizados usando a TIMIT e, para uma tolerância de erro de 20 ms, 92,47% das fronteiras resultaram abaixo desse erro.

Um outro sistema de refinamento dependente de contexto também foi proposto por Boonsuk (Boonsuk et al., 2007). Nesta técnica os autores propõem o uso de características acústicas que são selecionadas automaticamente para refinar tipos específicos de fronteiras de segmentação estimadas pelo alinhamento forçado de Viterbi.

Para o refinamento, um conjunto de características acústicas é construído contendo uma lista dos tipos de fronteiras e dos parâmetros que melhor caracterizam essa transição. A janela de análise que representa a transição de um fone para outro será aquela em que as características fonéticas mudam significativamente de uma janela para outra, dependendo do tipo de fronteira e das características acústicas empregadas.

Durante o processo de refinamento, para cada fronteira uma região com duração específica é definida, de forma que a fronteira que está sendo analisada seja o centro da região. Os autores testaram duas regiões de refinamento com durações diferentes: uma de ±20 ms e a outra de ±30 ms. Após a definição da região de refinamento, um conjunto de parâmetros acústicos é selecionado com base no tipo de fronteira. Uma análise de variância é empregada para determinar os parâmetros que têm maior habilidade para separar as janelas de análise que correspondem à fronteira que está sendo analisada e automaticamente descartar os parâmetros que não demonstram diferença significativa.

A janela de análise que carrega as mudanças acústicas mais significativas é escolhida através de um classificador LDA (*Linear Discriminant Analysis*). Esse classificador é usado para julgar o quão próxima uma janela candidata está da verdadeira fronteira. Para cada janela candidata, o classificador LDA faz a decisão baseada nas características acústicas que foram selecionadas de acordo com o tipo da fronteira. Essa proximidade é calculada através de probabilidades.

Para testar o sistema de refinamento proposto, os autores utilizaram uma base de fala contínua da língua Tailandesa e definiram apenas 21 tipos de fronteiras que serão refinadas. As fronteiras que não se encaixam nesses grupos não são refinadas. Os autores observaram que 90,24% das fronteiras refinadas apresentaram um erro menor que 20 ms quando comparadas com a segmentação manual.

3.8. Considerações Finais

Neste Capítulo foi apresentada uma introdução ao processo de segmentação de fala e os problemas envolvidos. Como destacado, a segmentação automática pode ser dividida em dois grandes grupos: segmentação irrestrita (implícita) e segmentação restrita (explícita). Pode haver também a utilização de modelos com a finalidade de representar as unidades acústicas a serem segmentadas. Neste caso a segmentação é dita dependente de modelo e, caso contrário, independente de modelo.

Uma revisão bibliográfica foi levantada com o intuito de destacar as principais técnicas e modelos que estão sendo utilizados na comunidade científica. Pelos trabalhos apresentados, percebe-se uma grande variedade de técnicas utilizadas, sendo as principais: medida de variação espectral, filtragem paramétrica, redes neurais artificiais, utilização de características fonético-acústicas, aprendizagem discriminativa, dentre outras.

A técnica baseada em variação espectral tem a vantagem de não precisar de treinamento e também não precisar que a transcrição fonética esteja disponível, mas por ser uma técnica de segmentação implícita apresenta algumas desvantagens (inserção e remoção de fronteiras. As redes neurais artificiais, algoritmos evolutivos, máquinas de vetor de suporte e aprendizagem discriminativa precisam de um treinamento prévio.

Dentre todos os métodos apresentados, os HMMs se destacam devido à sua capacidade de modelar a dinâmica das variações temporais presentes no sinal de fala e, em conjunto com o algoritmo de Viterbi, produzem um alinhamento da transcrição fonética com a locução. Esse alinhamento produz uma estimativa das primeiras marcas de segmentação que, por sua vez, podem ser refinadas para se aproximarem da segmentação manual.

Neste Capítulo também foram apresentadas algumas técnicas de refinamento da segmentação automática de fala. Dentre as técnicas apresentadas destacam-se aquelas que utilizam modelos de fronteiras para o refinamento.

Capítulo 4

Produção e Parametrização da Fala

A fala representa uma das principais características dos seres humanos que os diferencia de outros seres vivos. O processo de produção e decodificação da fala sempre foi um objeto de interesse, visando o seu reconhecimento e a sua reprodução através de computador, e conseqüentemente a criação de "seres artificiais" que possam se comunicar usando a linguagem natural.

Entender o processo de produção e as principais características dos sons é extremamente importante para qualquer aplicação computacional que utiliza a fala. Neste Capítulo serão abordados o modelo fisiológico de produção da fala, sua representação, as principais classes de sons presentes no português do Brasil (PB) e também os parâmetros que serão importantes para o processo de refinamento das marcas de segmentação.

Todas as figuras apresentadas neste Capítulo foram geradas a partir de palavras isoladas pronunciadas por um locutor masculino paulista. As locuções foram amostradas a 22,05 kHz e quantizadas com 16 bits/amostra. As figuras foram geradas no Matlab®.

4.1. Modelo Fisiológico de Produção de Fala

Os sinais de fala são formados por uma seqüência de sons que, por sua vez, representam a informação transmitida (Rabiner and Schafer, 1978). Cada língua apresenta regras diferentes para o arranjo desses sons, dando significado à comunicação. O estudo da linguagem é denominado Lingüística e o estudo da produção dos sons é chamado de Fonética.

A voz é uma onda de pressão acústica que se origina a partir de movimentos voluntários dos órgãos vocais humanos (Deller et al., 1993). O aparelho oral humano pode ser considerado um tubo acústico, tendo um comprimento aproximado de 17 cm. O aparelho oral inicia-se nas pregas vocais e termina nos lábios, onde o ar é irradiado.

Outro componente muito importante durante o processo de fonação é o véu palatino ou palato. Este componente desempenha papel importante durante a geração de sons nasais. Para a geração destes sons, o véu palatino é baixado e o trato nasal é acoplado acusticamente ao trato vocal. Por outro lado, durante a produção de sons não nasais, a cavidade nasal é bloqueada pela ação do véu palatino de modo que nenhum som seja irradiado pelo nariz.

Durante o processo de fonação, o ar armazenado nos pulmões é expelido passando pelas pregas vocais, que por sua vez podem vibrar ou não, constituindo dessa forma a fonte de excitação para a produção de alguns sons. As pregas vocais não são a única fonte de excitação para a produção de sons. Para as consoantes fricativas, plosivas e africadas ocorre uma constrição em alguma região do trato vocal, cujo ar passando por essa constrição forma uma turbulência que por sua vez atua como a fonte de excitação. No trato vocal, esse fluxo de ar sofre influência das estruturas ressonantes, produzindo os diversos sons da fala.

Em resumo, a produção da fala é composta por três partes fundamentais, conforme mostra o diagrama em blocos na Figura 4.1. As três partes que compõem o processo são: fonte de excitação, trato vocal e radiação.

Figura 4.1: Etapas do processo de produção da fala.

A fonte de excitação é responsável por excitar (alimentar) o trato vocal, que por sua vez atua como um filtro acústico, variante no tempo, que modela o sinal de fala através da ação de seus articuladores. Por último, este sinal é irradiado pela boca ou nariz, dependendo do tipo de som em consideração.

4.2. Os Sons da Fala

Os sons gerados pelo sistema vocal são basicamente divididos em dois grandes grupos: sons sonoros (gerados em resposta à vibração das pregas vocais) e não sonoros (não ocorre vibração das pregas vocais).

Os sons sonoros são resultantes da passagem do ar vindo dos pulmões e forçado na glote (abertura entre as pregas vocais). Durante a passagem do ar, as pregas vocais vibram (abrem e fecham) de forma oscilatória produzindo pulsos quase regulares responsáveis pela excitação do trato vocal, que por sua vez produz sons com característica periódica.

A taxa de vibração das pregas vocais depende das características do locutor, tais como: sexo, idade, comprimento do trato vocal, pressão do ar, dentre outros fatores. Essa taxa de vibração das pregas vocais é definida no sinal acústico como freqüência fundamental da fala. Exemplos de sons sonoros são as vogais e também algumas consoantes sonoras como os fones [b], [d], [g], [j], [v], [m], [n]. Por outro lado, para a geração dos sons não sonoros, também chamados de surdos, não há vibração das pregas vocais e o ar vindo dos pulmões tem livre acesso na glote. Exemplo: [t], [k], etc.

Em uma língua, um fonema é responsável por estabelecer o contraste de significado para diferenciar as palavras. Um fone é a realização de um fonema. No PB, basicamente os fonemas são divididos em duas grandes classes: vogais e consoantes.

A Tabela 4.1 mostra todos os 39 símbolos utilizados para representar os fonemas do português do Brasil que são adotados neste trabalho, e que são utilizados na transcrição fonética das locuções. A notação utilizada neste trabalho é particular e diferente do alfabeto fonético internacional (IPA – *International Phonetic Alphabet*). Os símbolos utilizados na TIMIT serão mostrados no Capítulo 5.

A Tabela 4.2 mostra os símbolos da Tabela 4.1 representados em suas classes fonéticas.

4.2.1. Vogais

As vogais são fonemas silábicos, ou seja, representam o núcleo da sílaba. Durante sua produção o ar vindo dos pulmões causa uma vibração quase periódica das pregas vocais. A variação da área do trato vocal para a produção das vogais determina as freqüências ressonantes, também chamadas de freqüências formantes.

Durante a produção das vogais, o ar expelido dos pulmões encontra apenas a barreira imposta pelas pregas vocais, diferente de algumas consoantes onde alguns articuladores do trato vocal também impõem barreiras à passagem do ar.

Tabela 4.1: Subunidades acústicas utilizadas na transcrição fonética das locuções (Adaptado de Ynoguti, 1999).

Símbolo Utilizado	**Exemplo**
a	açafrão = ***a*** s a f r an u
e	elevador = ***e*** l e v a vcl d o R
E	pele = cl p ***E*** l y
i	sino = s ***i*** n u
y	fui = f u ***y***
o	bolo = vcl b ***o*** l u
O	bola = vcl b ***O*** l a
u	lua = l ***u*** a
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***
b	bela = vcl ***b*** E l a
d	dádiva = vcl ***d*** a vcl d i v a
D	dia = vcl ***D*** i a (djia)
f	feira = ***f*** e i r a
g	gorila = ***g*** o r i l a
j	jiló = ***j*** i l O
k	cachoeira = cl ***k*** a x o e i r a
l	leão = ***l*** e an u
L	lhama = ***lh*** an m a
m	montanha = ***m*** on cl t an N a
n	névoa = ***n*** E v o a
N	inhame = i ***N*** an m y
p	poente = cl ***p*** o en cl T y
r	cera = s e ***r*** a
rr	cerrado = s e ***rr*** a vcl d u
R	carta = cl k a ***R*** cl t a
s	sapo = ***s*** a cl p u
t	tempestade = cl ***t*** en cl p e s cl ***t*** a vcl D y
T	tia = cl ***T*** i a (tchia)
v	verão = ***v*** e r an u
x	chave = ***x*** a v y
z	zabumba = ***z*** a vcl b un vcl b a
cl	Período de constrição para as plosivas surdas
vcl	Período de constrição para as plosivas sonoras
sp	Pausa entre palavras
#	Silêncio presente no início e no fim das locuções

Tabela 4.2: Classificação dos fones do Português do Brasil.

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

O principal articulador envolvido na produção das vogais é a língua, mas a posição dos outros articuladores também influencia. A posição da língua durante a produção das vogais pode ser utilizada para classificá-las em: vogais anteriores, central e posteriores. A altura da língua também pode ser utilizada para a classificação em vogais altas, médias e baixas.

A vogal é dita central quando durante a articulação a posição da língua permanece quase em repouso. A vogal central do PB é o [a]. Para as vogais anteriores, durante sua produção a língua se eleva, avançando em relação ao palato duro, diminuindo a abertura bucal.

As vogais anteriores são: [e], [E], [i] e [y]. Por último, para as vogais posteriores o dorso da língua se eleva e recua em direção ao palato mole. As vogais posteriores são: [o], [O] e [u].

4.2.2. Consoantes

Diferente das vogais, na produção das consoantes o ar expelido dos pulmões encontra barreiras impostas pelos articuladores, dando origem dessa forma aos vários sons. As consoantes do PB podem ser classificadas com base em quatro critérios: modo de articulação (plosivas, fricativas e líquidas(róticas e laterais)), quanto ao ponto de articulação (bilabiais, labiodentais, alveolares, palatais e velares), quanto ao papel das pregas vocais (sonoras e surdas) e quanto ao

papel das cavidades bucal e nasal (consoantes orais e nasais). Em seguida é apresentada uma descrição das consoantes quanto ao modo de articulação.

4.2.2.1. Fricativas

Na produção das consoantes fricativas, em algum ponto do trato vocal forma-se uma constrição. Em seguida, uma quantidade de ar forçada através desta constrição produz uma turbulência. Esta turbulência por sua vez constitui a fonte de excitação para o trato vocal que, em resposta, gera os sons fricativos. As fricativas do PB são: [f], [s], [z], [v], [x] e [j]. A Figura 4.2 mostra a forma de onda e o espectrograma para a locução "chuva", destacando a ocorrência da fricativa surda [x] e da fricativa sonora [v].

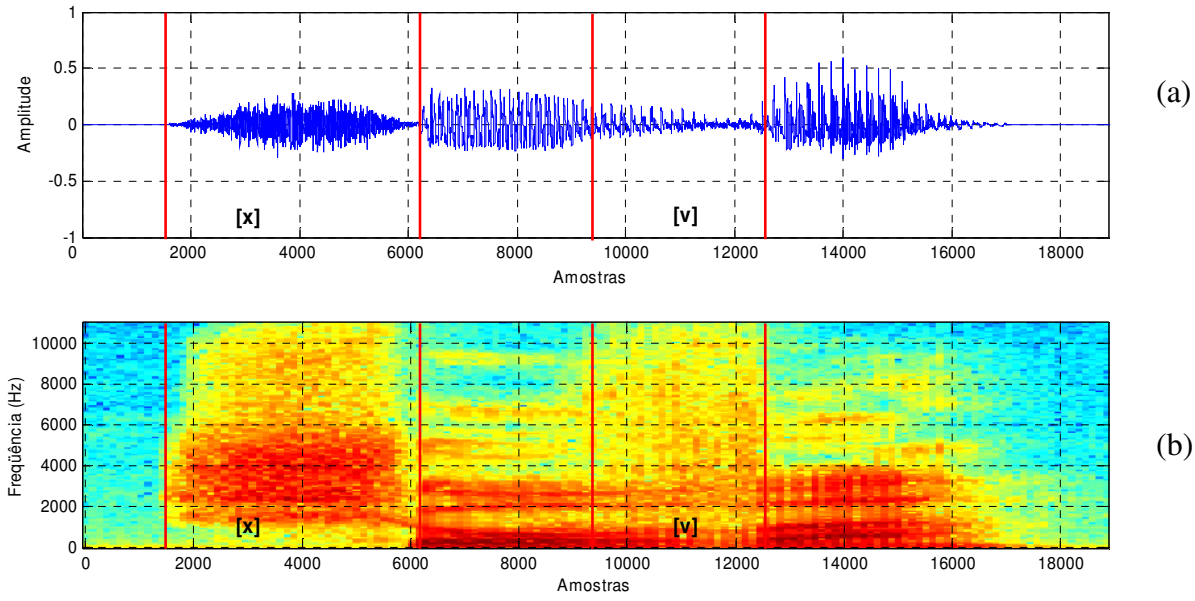


Figura 4.2: Locução "**ch**u**v**a": (a) Forma de onda. (b) Espectrograma.

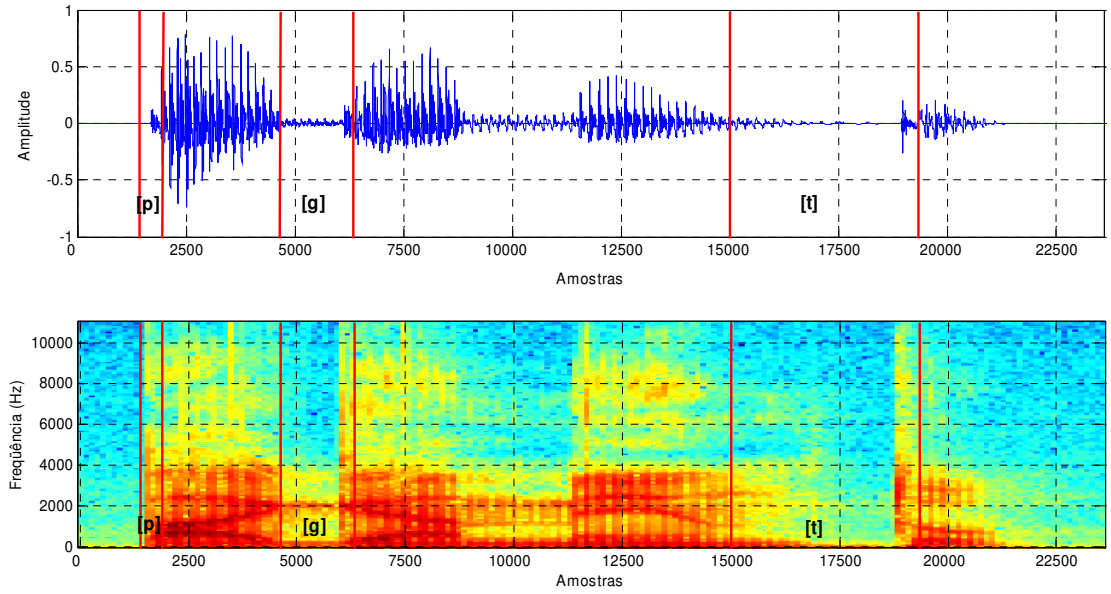
Analisando a figura, percebe-se claramente que o trecho que corresponde às consoantes fricativas assemelha-se a um ruído, uma vez que é resultado de uma fonte de excitação ruidosa (turbulência) que estimula o trato vocal.

Dentre os vários critérios de classificação das fricativas, o mais importante neste trabalho é com relação ao papel das pregas vocais. Através deste critério pode-se classificar as fricativas em surdas e sonoras.

Na produção das fricativas surdas não há vibração das pregas vocais. As fricativas surdas são: [f], [s] e [x]. Ao contrário das fricativas surdas, para a produção das fricativas sonoras há a excitação de duas fontes, a glote e o ponto de constrição em algum local do trato vocal. As fricativas sonoras são: [v], [z] e [j].

4.2.2.2. Plosivas

Os sons plosivos são resultantes de um fechamento completo do trato vocal. A pressão do ar é aumentada no ponto de fechamento e subitamente liberada. Durante o período de tempo em que ocorre a constrição total, nenhum som é irradiado. Dessa forma, estes sons são caracterizados por um período de oclusão seguido de uma "explosão". As consoantes plosivas do PB são [p], [t], [k], [b], [d] e [g]. A Figura 4.3 mostra a forma de onda e o espectrograma para a locução "pagamento", destacando a ocorrência das plosivas surdas [p] e [t], e a plosiva sonora [g].



(a)

(b)

Figura 4.3: Locução "**p**a**g**amen**t**o": (a) Forma de onda. (b) Espectrograma.

Como pode ser observado na Figura 4.3, as plosivas apresentam duas características na forma de onda: i) um período caracterizado por baixa amplitude (período de constrição) e ii) período caracterizado pela "explosão" (liberação do ar).

As plosivas também podem ser classificadas em surdas e sonoras. Para as plosivas sonoras ([b], [d] e [g]), durante o período de constrição total do trato vocal há uma pequena

quantidade de energia nas baixas freqüências irradiada através das paredes da garganta. Isso é explicado pelo fato de que as pregas vocais estão hábeis para vibrar mesmo com a constrição em algum ponto do trato vocal (Rabiner and Schafer, 1978).

Para as plosivas surdas ([p], [t] e [k]), durante o período de total constrição no trato vocal, não há vibração das pregas vocais e esses sons são caracterizados por um período de oclusão seguido de uma explosão.

4.2.2.3. Africadas

As consoantes africadas, também chamadas de plosivas africadas, são sons complexos formados pela combinação das consoantes plosivas e das fricativas. Como as consoantes plosivas, durante a produção das consoantes africadas ocorre uma constrição total em algum ponto do trato vocal e, após a liberação do ar, tem-se um som caracterizado por um ruído de fricação (como nas consoantes fricativas).

Como na língua inglesa, o PB tem duas consoantes africadas ([T] e [D]). Esses sons ocorrem na combinação das consoantes plosivas [t] ou [d] seguidas pela vogal posterior [i]. A consoante [T] é surda enquanto que a consoante [D] é sonora. A Figura 4.4 mostra a forma de onda e o espectrograma para a locução “titia”, destacando a africada surda [T].

Como pode ser observado na Figura 4.4, a forma de onda para a africada [T] é muito semelhante à forma de onda de algumas consoantes fricativas. Para a africada surda no início da locução, o período de oclusão é fundido com o período de oclusão que caracteriza o início da locução.

Para a segunda ocorrência da consoante é possível distinguir claramente o período de oclusão seguido por uma região muito semelhante a uma fricativa. Em contrapartida, a Figura 4.5 mostra a forma de onda e o espectrograma para a locução “didi”. A figura destaca as ocorrências para a consoante africada sonora [D], no início e no meio da locução. Nas duas ocorrências percebe-se claramente o período que caracteriza a oclusão sonora, região típica das plosivas sonoras como já observado.

A distinção entre as consoantes africadas e as fricativas está no período de oclusão, que não existe para as fricativas. Por outro lado, a diferença entre as africadas e as plosivas está na duração da região após a liberação do ar, que para as africadas é longo e para as plosivas é curto.

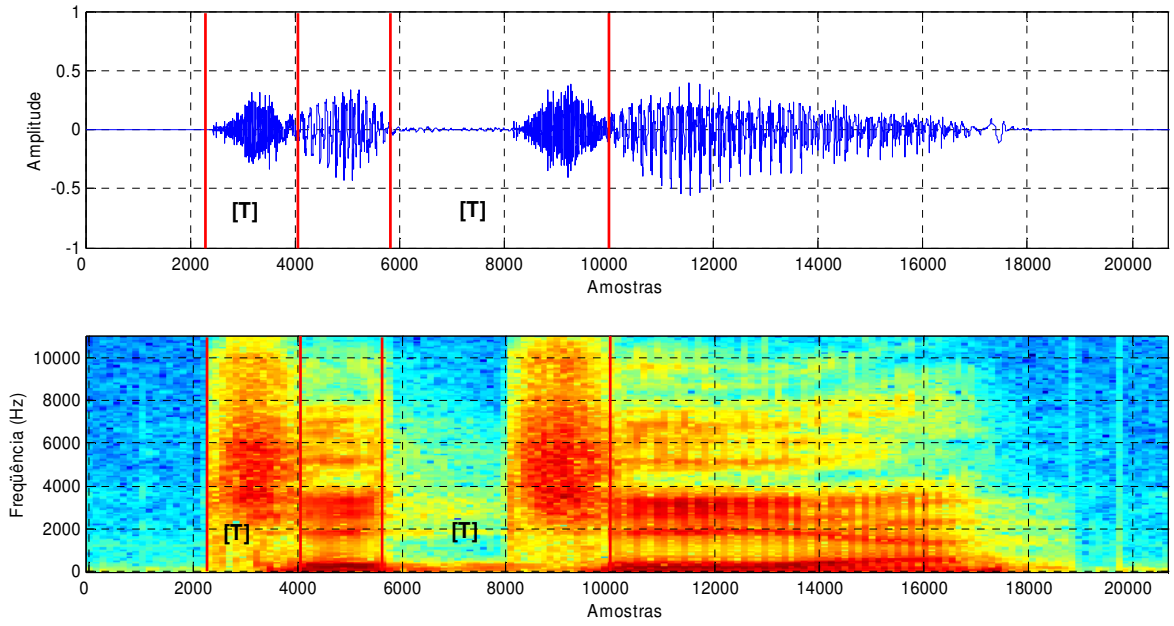


Figura 4.4: Locução "**ti**ti**a**": (a) Forma de onda. (b) Espectrograma.

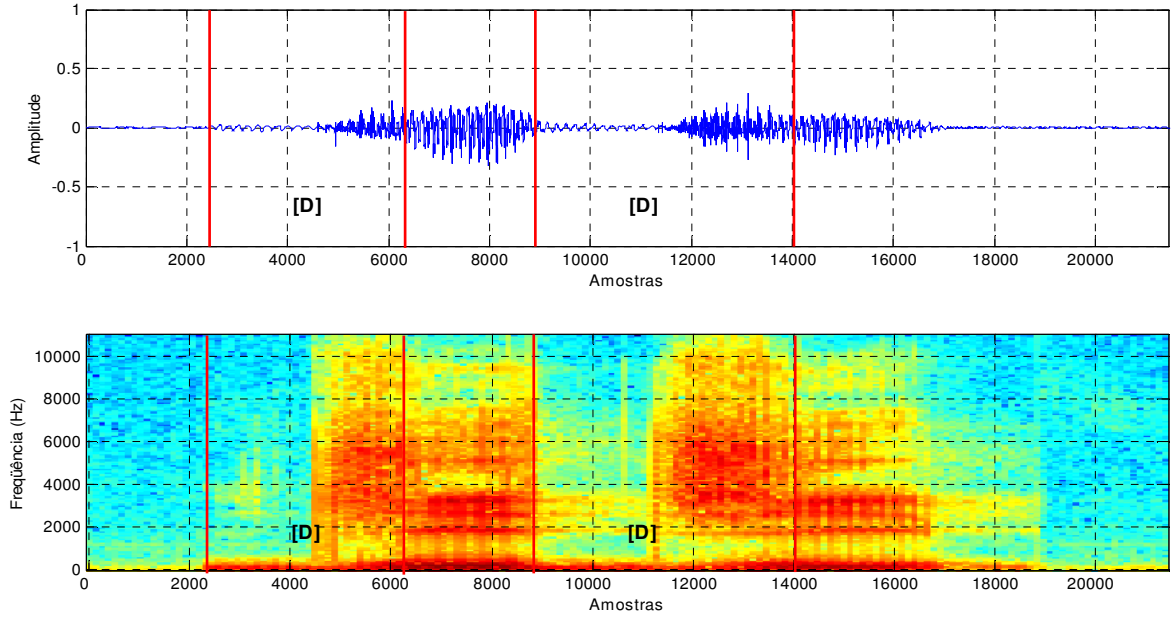


Figura 4.5: Locução "**di**di". (a) Forma de onda. (b) Espectrograma.

4.2.2.4. Nasais

As consoantes nasais são produzidas com excitação glotal e também com uma constrição em algum ponto do trato vocal. O véu palatino que bloqueia a passagem do ar para o trato nasal

durante a produção das consoantes e vogais não-nasalizadas agora é baixado, fazendo com que o fluxo de ar seja irradiado não pela boca e sim pelo nariz.

Apesar do bloqueio realizado pelo véu palatino, o trato vocal ainda permanece acusticamente acoplado à faringe e a boca serve como uma cavidade ressonante que emite energia acústica com uma certa freqüência. Essa energia nas baixas freqüências é chamada de murmúrio. A Figura 4.6 mostra a forma de onda e o espectrograma para a locução "amazonas". No PB, as três consoantes nasais são: [m], [n] e [N].

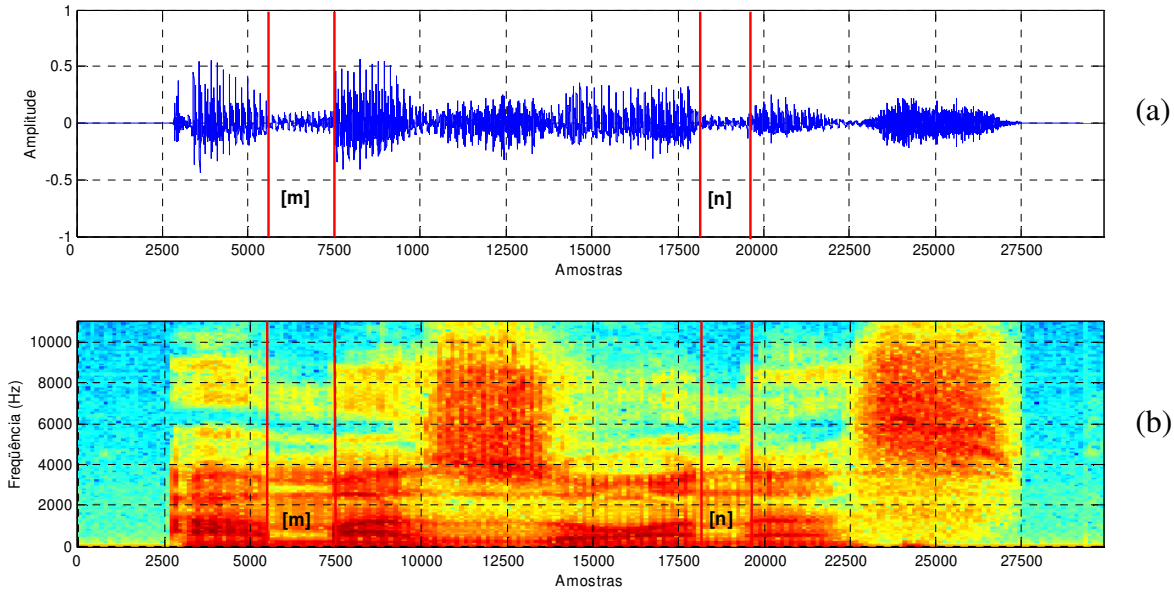


Figura 4.6: Locução "a**m**azo**n**as": (a) Forma de onda. (b) Espectrograma.

4.2.2.5. Laterais

As consoantes laterais apresentam essa denominação porque, durante a sua produção, a ponta da língua é posicionada no palato duro e o ar expelido dos pulmões, que passa pela glote vibrando as pregas vocais, "escapa" pelas laterais da constrição formada pela língua. As consoantes laterais do PB são: [l] e [L]. A Figura 4.7 mostra a forma de onda e o espectrograma para a locução "leite".

Nota-se pela análise da Figura 4.7 que a consoante lateral [l] apresenta uma forma de onda bem definida, muito parecida com as vogais, sendo diferenciada apenas pela amplitude. Tanto as consoantes laterais quanto as consoantes róticas na língua inglesa são classificadas como

aproximantes e representam uma classe fonética difícil de ser caracterizada justamente pela sua natureza acústica semelhante às vogais (Rabiner and Juang, 1993).

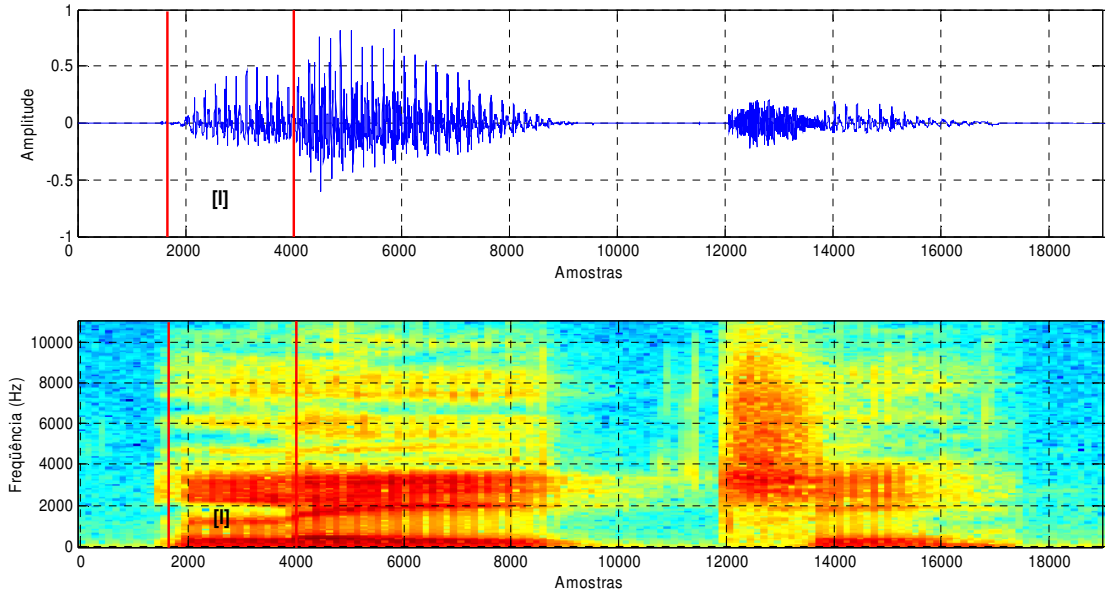


Figura 4.7: Locução "**l**eite": (a) Forma de onda. (b) Espectrograma.

4.2.2.6. Róticas

As consoantes róticas (*rhotics* em inglês) representam uma classe fonética bastante difícil de ser caracterizada acusticamente por apresentar diversas possibilidades de pronúncias como destacado por Silva (Silva and Albano, 1999).

Diferente da produção das outras consoantes, para as róticas não há uma constrição no trato vocal e sim apenas um estreitamento na região palatal. A corrente de ar que passa pela glote vibrando as pregas vocais, também provoca vibrações na região do estreitamento. As consoantes róticas do PB são: [r] e [R]. Neste trabalho, a consoante fricativa posterior sonora [rr] será incluída no grupo das consoantes róticas em virtude do seu desempenho durante o processo de refinamento.

A Figura 4.8 mostra a forma de onda e o espectrograma para a locução "realidade". Como destacado para as consoantes laterais, as consoantes róticas também apresentam uma forma de onda muito semelhante às vogais.

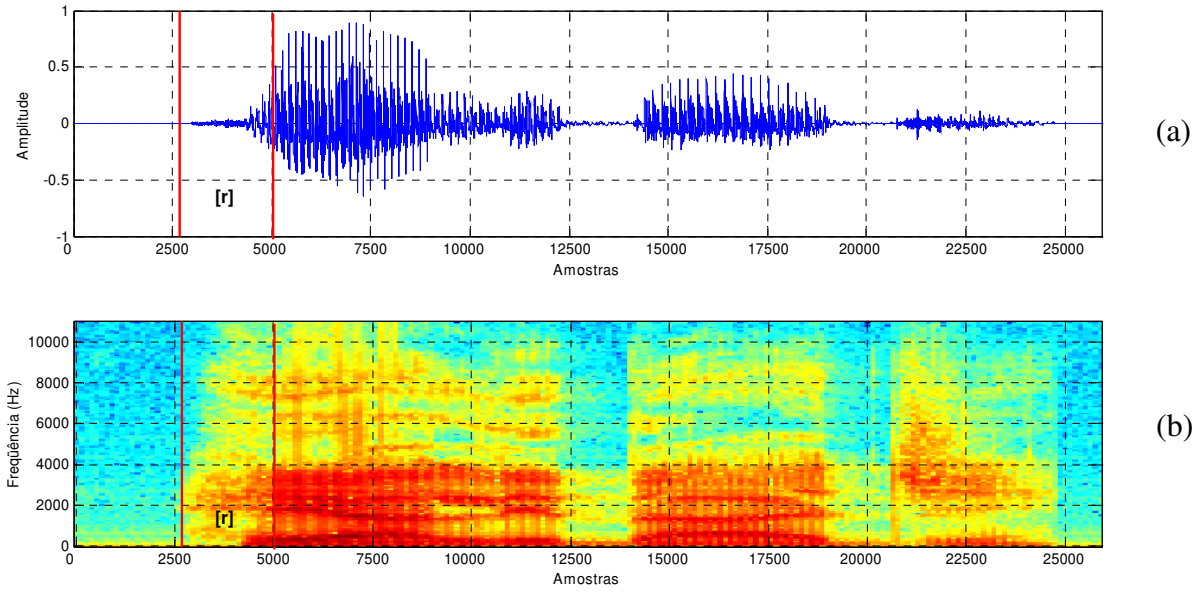


Figura 4.8: Locução "**r**ealidade": (a) Forma de onda. (b) Espectrograma.

4.3. Análise e Parametrização dos Fones

Na Seção 4.2, os fonemas do PB foram agrupados em duas grandes classes (vogais e consoantes) e o processo de produção de cada classe foi apresentado. Para aplicações em segmentação automática de fala, que é o propósito deste trabalho, as características acústicas devem ser levadas em consideração.

O objetivo desta Seção é apresentar as principais características acústicas de cada classe fonética. Apenas as características necessárias para diferenciar as classes entre si serão abordadas, uma vez que apenas essas características serão empregadas para o refinamento das marcas de segmentação.

4.3.1. Vogais

As vogais representam a classe fonética que desempenha papel primordial em qualquer língua, representando para o português o núcleo de uma sílaba. A descrição acústica é simples e poucos parâmetros são necessários para diferenciá-la de outras classes e mesmo diferenciar uma vogal de outra. Os principais parâmetros a serem descritos são os formantes.

Os formantes ou simplesmente freqüências de ressonância são as ressonâncias do trato vocal. Durante a excitação o trato vocal apresenta um determinado número de ressonâncias, que

são localizadas em freqüências específicas, freqüências essas que dependem do tamanho do trato vocal. Kent e Read (Kent and Read, 1992) definem a Equação (4.1) para determinar a freqüência em que ocorre uma determinada ressonância. A Equação é dada por:

$$F_n = \frac{(2n-1)c}{4l} \tag{4.1}$$

onde:

n, um número inteiro que representa a freqüência de ressonância a ser calculada;

c, é a constante que representa a velocidade do som (340 m/s), nas condições normais de pressão e temperatura;

l, é o comprimento do tubo ressonador, neste caso o trato vocal (em m).

Durante o processo de produção das vogais, a língua, que é o principal articulador, impõe constrições no trato vocal. Por exemplo, para a produção da vogal [i] há uma constrição próxima à região dos lábios e uma região de grande abertura perto da laringe e da faringe onde o fluxo de ar é maior. Um outro exemplo é para a vogal [a], onde a constrição é na região anterior próxima à região da laringe.

Essas constrições formadas pela língua durante o processo de fonação das vogais, alterando a configuração do trato vocal, fazem com que os valores das freqüências ressonantes também sejam alterados.

A freqüência do primeiro formante (F1) varia inversamente com a altura da língua durante a produção da vogal, ou seja, as vogais altas apresentam um valor de F1 baixo enquanto que as vogais baixas apresentam um valor alto para F1. A variação da freqüência do segundo formante (F2) está relacionada com o avanço da posição da língua (anterior – posterior), ou seja, F2 aumenta quando a posição move-se em direção a boca. Para as vogais posteriores o valor de F2 é baixo e para as anteriores o valor é alto.

A Figura 4.9 mostra todas as vogais tônicas do PB descritas através dos valores de F1 e F2. Esse diagrama é conhecido como triângulo das vogais e representa os valores médios de cada formante obtidos a partir da pronúncia de nove palavras por nove diferentes locutores do sexo masculino.

O valor das freqüências formantes pode ser afetado pela participação dos lábios no processo de produção da voz. O arredondamento dos lábios que normalmente ocorre nas vogais posteriores e nas vogais médias pode diminuir o valor dos formantes. Isso é explicado pela teoria da perturbação (Kent and Read, 1992). Em resumo, quanto maior o comprimento do trato vocal, menor tende a ser o valor dos formantes e, como o arredondamento dos lábios tende a aumentar o trato vocal, isso provoca uma diminuição da freqüência dos formantes.

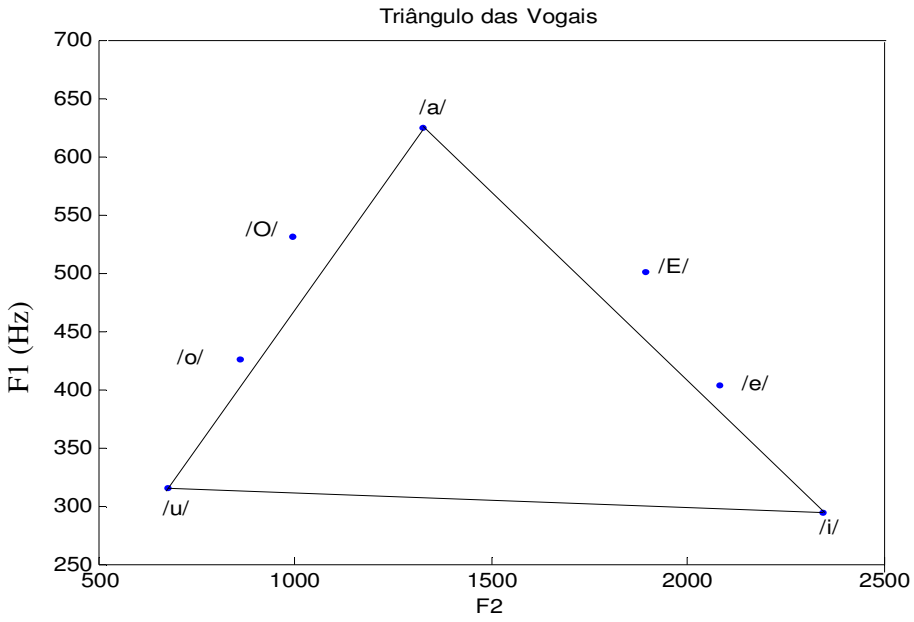


Figura 4.9: Triângulo das vogais.

Para a identificação das vogais, ou mesmo identificar vogais presentes em um ditongo (vogal+semivogal), uma análise detalhada dos dois primeiros formantes faz-se necessária e, neste caso, algoritmos para realizar um acompanhamento dos formantes ao longo do tempo é de extrema importância. Um problema que pode ser observado na prática é a proximidade de alguns valores de formantes como, por exemplo, o valor de F1 para as vogais [e] e [i], como pode ser observado na Figura 4.9.

Como sugerido por Araújo (Araújo, 2000), uma medida de distância absoluta e relativa entre os formantes é uma característica importante que deve ser considerada na classificação e parametrização das vogais. Como destacado pelo autor, a medida de distância absoluta,

representada por ΔF_α, entre as vogais, pode ser determinada tomando-se o valor absoluto da diferença entre o valor dos formantes para as vogais em questão. A variável α em ΔF_α pode assumir dois possíveis valores: $\alpha = 1$ corresponde a F1 e $\alpha = 2$ corresponde a F2.

A distância relativa representada por ΔF_α, pode ser obtida por:

$$\Delta F_\alpha = \frac{\left| F_{\alpha/v1/} - F_{\alpha/v2/} \right|}{\min\left[F_{\alpha/v1/}, \quad F_{\alpha/v2/} \right]} 100\%, \quad \alpha = 1,2 \tag{4.2}$$

onde $v1$ e $v2$ correspondem às vogais que estão sendo analisadas.

Araújo também sugere uma medida de distribuição de energia no espectro de freqüências para separar as vogais em anteriores e posteriores. A fundamentação reside no fato de que a energia das vogais posteriores está mais concentrada nas baixas freqüências, e que a energia das vogais anteriores está distribuída em uma larga faixa de freqüências. A medida sugerida é o perfil de energia.

O perfil de energia (F_β) representa a freqüência em Hz abaixo da qual está contida uma determinada porcentagem (β) da energia total calculada sobre o espectro de freqüência, que por sua vez é calculado a partir da DFT de uma janela do sinal de N amostras. A expressão para o cálculo do perfil de energia é dada por:

$$F_\beta = k_\beta \frac{f_a}{N} \tag{4.3}$$

onde f_a é a freqüência de amostragem do sinal e N é número de pontos utilizados no cálculo da FFT. O valor de k_β é um número inteiro, satisfazendo a seguinte relação:

$$k_\beta \leq \begin{cases} \dfrac{N}{2} & \text{para } N \text{ par;} \\ \dfrac{N-1}{2} & \text{para } N \text{ ímpar;} \end{cases} \tag{4.4}$$

A expressão para o cálculo de k_β também sofre uma leve variação em função do valor de *N*, que pode ser par ou ímpar. As Equações (4.5) e (4.6) são utilizadas para calcular a energia até a freqüência F_β. A Equação (4.5) é utilizada quando *N* é par e a Equação (4.6) para *N* ímpar.

$$E_\beta = \frac{1}{N}\left[|X(0)|^2 + 2\sum_{i=1}^{k_\beta} |X(i)|^2 \right] = \left(\frac{\beta}{100}\right)\frac{1}{N}\left[|X(0)|^2 + \left|X\left(\frac{N}{2}\right)\right|^2 + 2\sum_{i=1}^{\frac{N}{2}-1} |X(i)|^2 \right] \quad (4.5)$$

$$E_\beta = \frac{1}{N}\left[|X(0)|^2 + 2\sum_{i=1}^{k_\beta} |X(i)|^2 \right] = \left(\frac{\beta}{100}\right)\frac{1}{N}\left[|X(0)|^2 + 2\sum_{i=1}^{\frac{N}{2}} |X(i)|^2 \right] \quad (4.6)$$

onde $X(i)$, $i = 0...N-1$ é a DFT de uma janela do sinal de *N* amostras.

Alguns testes iniciais foram realizados com o objetivo de descobrir o melhor valor de β para contribuir no processo de refinamento das marcas de segmentação entre as vogais. Como sugerido por Araújo, o valor da porcentagem β foi fixado entre 5 e 95%, e foi variado em diversos testes para descobrir o melhor valor, ou seja, aquele que apresenta uma boa diferenciação entre as vogais. O melhor valor encontrado para a base dependente de locutor do PB, e que será utilizado durante o processo de refinamento, foi 75%. Para a TIMIT, cuja freqüência de amostragem é 16 kHz, será utilizada a taxa de 70%.

4.3.2. Fricativas

As fricativas representam uma classe de consoantes que são produzidas através de constrições impostas no trato vocal, por onde o ar vindo dos pulmões é forçado produzindo um ruído de fricção (Fant, 1960). Segundo Kent e Read (Kent and Read, 1992), as fricativas podem ser divididas em duas grandes classes, segundo a intensidade da energia do ruído: sibilantes e não-sibilantes. As sibilantes também são chamadas de fricativas fortes e as não-sibilantes de fricativas fracas.

Tanto as fricativas fortes quanto as fracas também podem ser subclassificadas em surdas e sonoras, de acordo com a presença ou não de atividade das pregas vocais. A fricativas fortes do

PB são [s], [x], [z] e [j], e as fricativas fracas são o [f] e o [v]. Neste trabalho será adotada apenas a classificação para fricativas surdas e sonoras.

As faixas de freqüências para as fricativas já foram estudadas por diversos pesquisadores. Russo e Behlau (Russo and Behlau, 1993) obtiveram para as consoantes fricativas uma faixa de freqüências entre 1200 e 7000 Hz para as anteriores e freqüências entre 2500 e 6000 Hz para as posteriores. Os fones [f] e [v] apresentam a menor intensidade entre as fricativas.

As Figuras 4.10 a 4.15 mostram os espectrogramas para as fricativas do PB. As Figuras 4.10, 4.12 e 4.14 representam as fricativas surdas enquanto que as Figuras 4.11, 4.13 e 4.15 as fricativas sonoras.

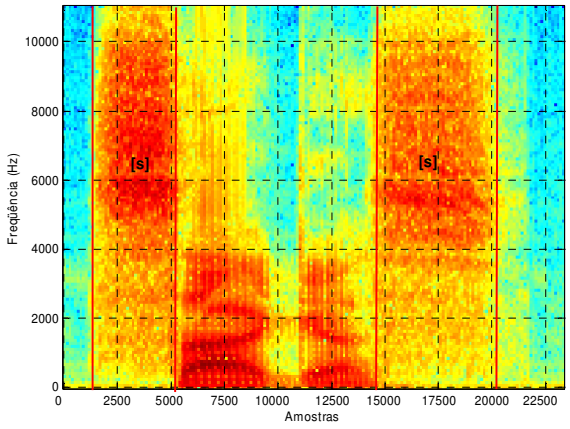


Figura 4.10: Espectrograma para a locução "**s**aga**s**".

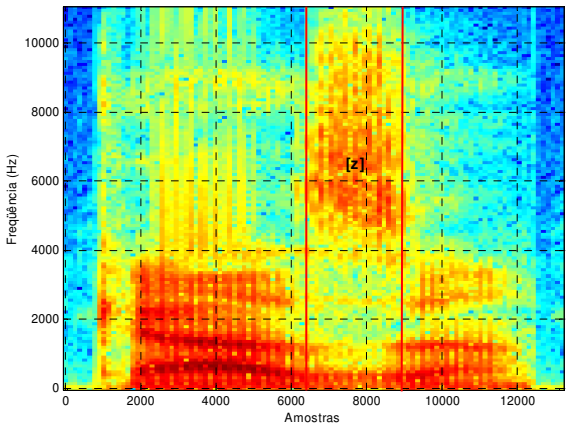


Figura 4.11: Espectrograma para a locução "ca**s**a".

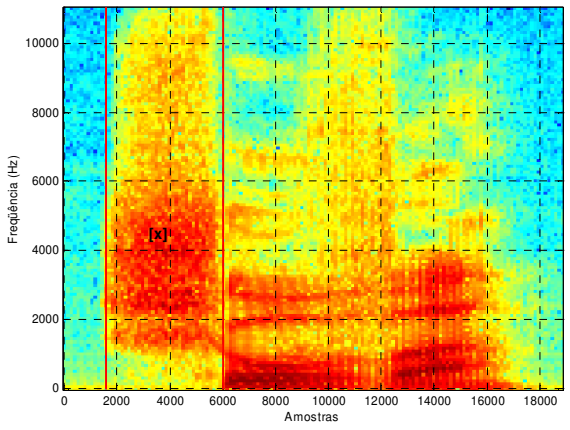


Figura 4.12: Espectrograma para a locução "**ch**uva".

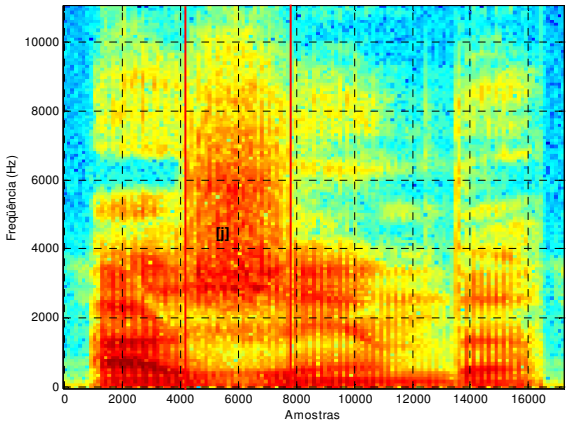


Figura 4.13: Espectrograma para a locução "a**g**enda".

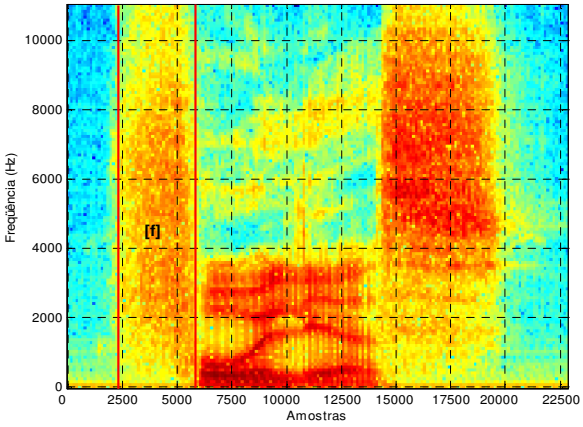


Figura 4.14: Espectrograma para a locução "**f**olhas".

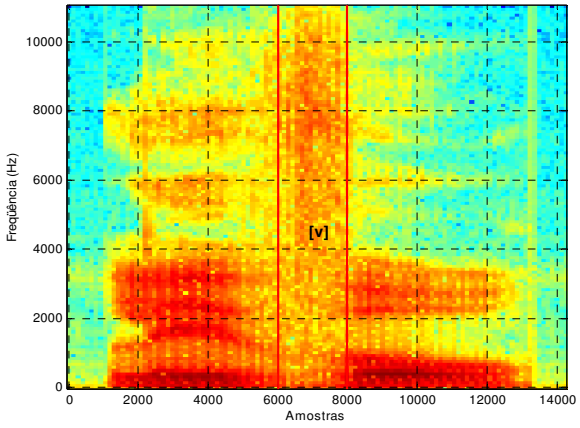


Figura 4.15: Espectrograma para a locução "le**v**ou".

Como observado nos espectrogramas, as fricativas anteriores apresentam componentes em alta freqüência e um dos parâmetros que pode ser utilizado para classificar janelas contendo fricativas é a taxa de cruzamentos por zero (Rabiner and Schafer, 1978), (Vieira, 1989) e (Maciel, 2003). Um cruzamento por zero ocorre quando amostras sucessivas do sinal apresentam diferentes sinais algébricos. O cálculo da taxa de cruzamentos por zero em uma janela de comprimento $2M+1$ centrada na amostra n é dado por:

$$z_n = \frac{1}{2M+1} \sum_{m=n-M+1}^{n+M} \left| \text{sgn}[x(m)] - \text{sgn}[x(m-1)] \right| \tag{4.7}$$

onde:

$$\text{sgn}[x(m)] = \begin{cases} 1, & para \quad x(m) \geq 0 \\ -1, & para \quad x(m) < 0 \end{cases} \tag{4.8}$$

As Figuras de 4.16 a 4.21 mostram as variações da taxa de cruzamentos por zero das locuções cujos espectrogramas foram apresentados nas Figuras de 4.10 a 4.15. A taxa de

cruzamentos por zero foi determinada a partir de janelas de análise de duração de 20 ms deslocadas a cada 10 ms.

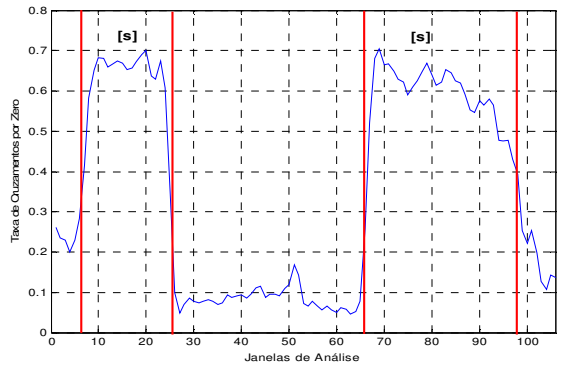


Figura 4.16: Taxa de cruzamentos por zero para a locução "**s**aga**s**".

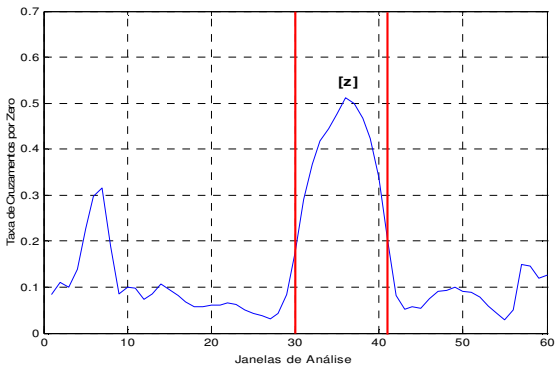


Figura 4.17: Taxa de cruzamentos por zero para a locução "ca**s**a".

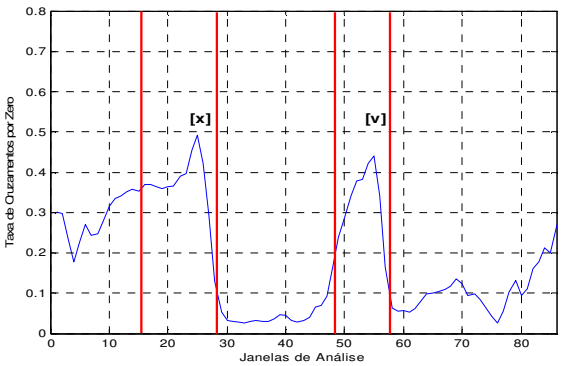


Figura 4.18: Taxa de cruzamentos por zero para a locução "**ch**uva".

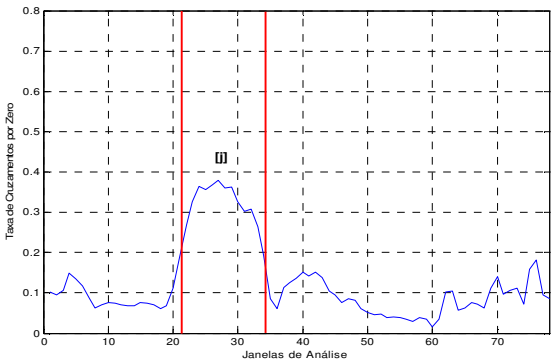


Figura 4.19: Taxa de cruzamentos por zero para a locução "a**g**enda".

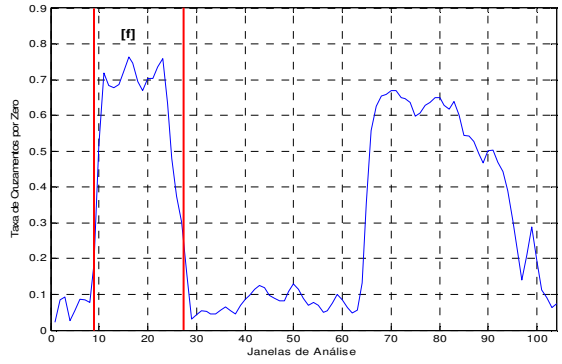


Figura 4.20: Taxa de cruzamentos por zero para a locução "**f**olhas".

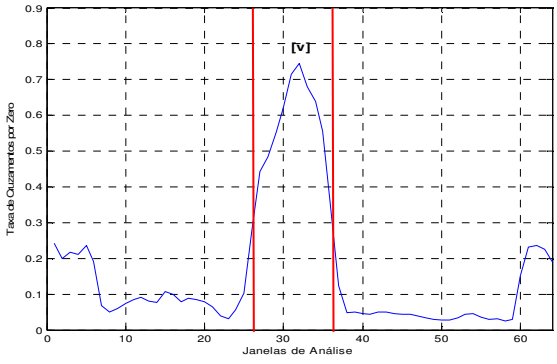


Figura 4.21: Taxa de cruzamentos por zero para locução "le**v**ou".

Segundo Araújo (Araújo, 2000) e (Araújo and Violaro, 2000), um outro parâmetro que pode ser utilizado na identificação das consoantes fricativas é o centro de gravidade espectral. O centro de gravidade espectral é a freqüência abaixo (ou acima) da qual estão concentrados 50% da energia da janela de análise. O cálculo do centro de gravidade espectral pode ser realizado utilizando-se a expressão para o cálculo do perfil de energia dado pela Equação (4.3) em que β = 50%.

As Figuras de 4.22 a 4.27 mostram a variação do centro de gravidade espectral das locuções cujos espectrogramas foram apresentados nas figuras de 4.10 a 4.15. O centro de gravidade espectral foi calculado usando janelas de análise com duração de 20 ms deslocadas a cada 10 ms. Como os testes foram realizados com locuções amostradas a 22,05 kHz, o valor do centro de gravidade espectral está abaixo de 11,025 kHz.

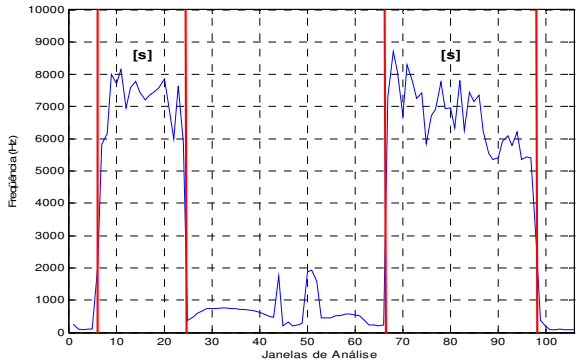


Figura 4.22: Centro de gravidade espectral para a locução "**s**aga**s**".

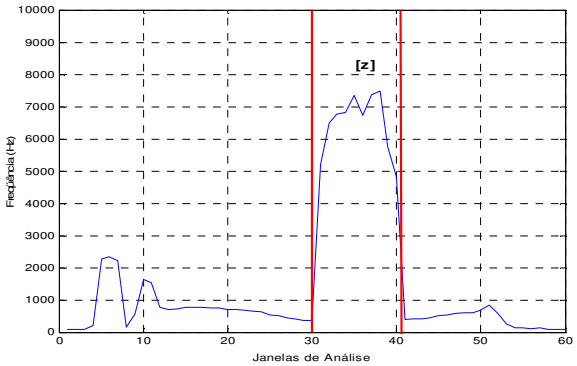


Figura 4.23: Centro de gravidade espectral para a locução "ca**s**a".

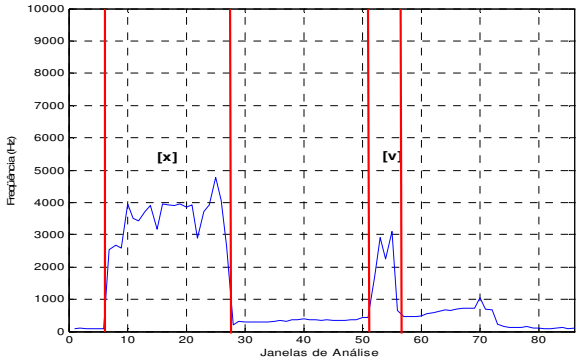


Figura 4.24: Centro de gravidade espectral para a locução "**ch**uva".

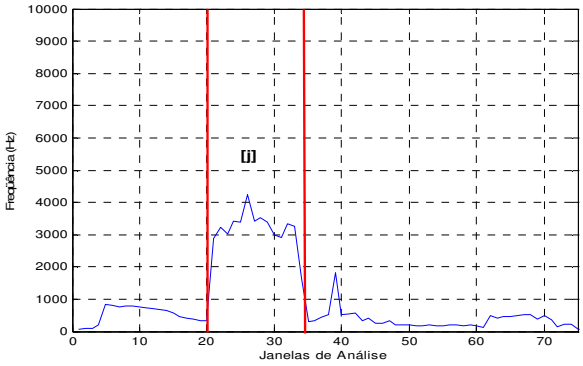


Figura 4.25: Centro de gravidade espectral para a locução "a**g**enda".

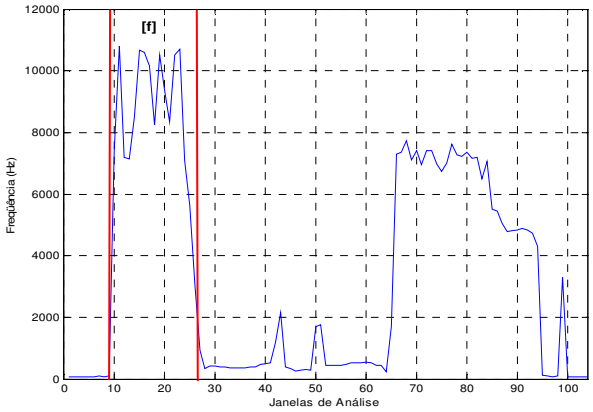


Figura 4.26: Centro de gravidade espectral para a locução "**f**olhas".

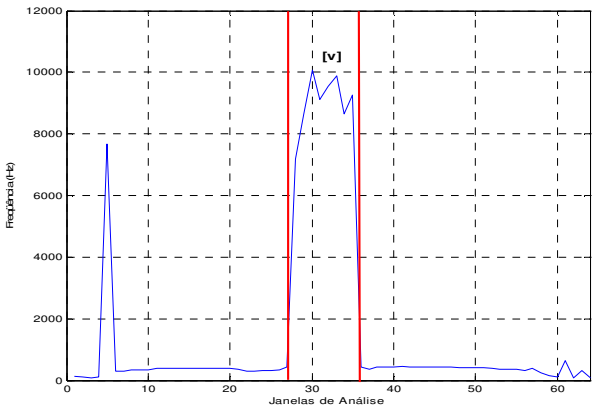


Figura 4.27: Centro de gravidade espectral para a locução "le**v**ou".

Um outro parâmetro sugerido por Kent e Read (Kent and Read, 1992) é a freqüência de pico para a qual a amplitude do espectro é máxima. Testes realizados por Araújo indicam, entretanto, que a taxa de cruzamentos por zero e o centro de gravidade espectral apresentam as menores variações (entre fones) quando comparadas com a freqüência de pico.

Para o propósito deste trabalho, que é realizar um refinamento das marcas de segmentação, e não um reconhecimento automático de fones, tanto a taxa de cruzamento por zeros quanto o centro de gravidade espectral parecem parâmetros promissores para identificar janelas contendo as consoantes fricativas. A freqüência de pico não será utilizada durante o processo de refinamento.

4.3.3. Laterais e Róticas

As consoantes laterais e as róticas são duas classes fonéticas difíceis de serem caracterizadas acusticamente devido à presença de propriedades similares às plosivas e também similares às vogais (Kent and Read, 1992).

A similaridade com as plosivas está na natureza dinâmica dos fonemas que apresentam movimentos articulatórios rápidos e dependentes do contexto fonético. A similaridade com as vogais é a natureza sonora e uma estrutura bem definida dos formantes.

Vários estudos realizados com as consoantes laterais e róticas, principalmente com fonemas do inglês americano, mostram que essa classe fonética apresenta mudanças acústicas relativamente rápidas, sendo que o fone [l] varia mais rápido que o [r], principalmente em relação

ao formante F1 (Espy-Wilson et al., 2000). Para os fonemas americanos, a principal característica espectral utilizada para distinguir a rótica [r] é o seu baixo valor para o formante F3. Essas observações não parecem acontecer com os fonemas do PB, conforme testes preliminares ‘realizados neste trabalho.

As Figuras 4.28 e 4.29 mostram os espectrogramas de duas locuções contendo a consoante lateral [l] e a rótica [r].

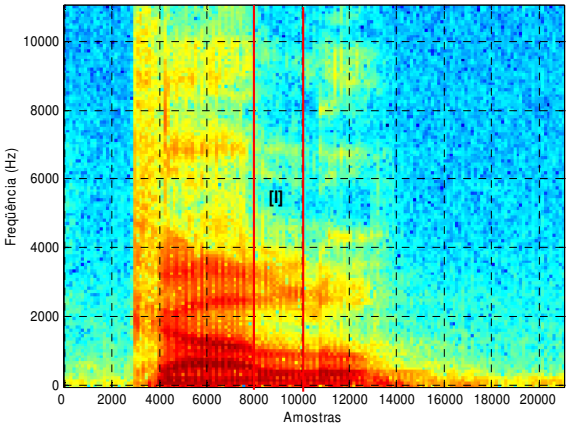


Figura 4.28: Espectrograma para a locução “ca**l**o”.

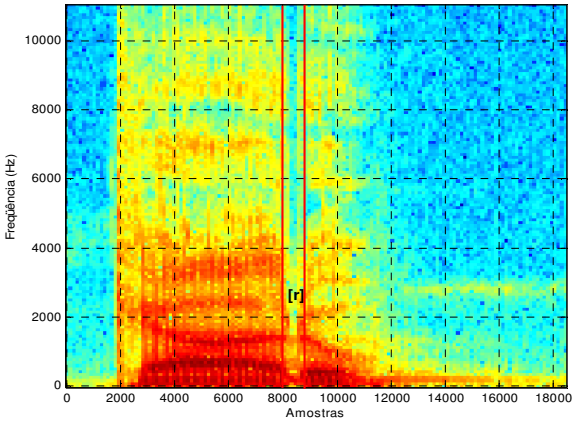


Figura 4.29: Espectrograma para a locução “ca**r**o”.

Como já discutido anteriormente, pela análise dos espectrogramas é fácil perceber que a duração da lateral [l] é maior do que a duração da rótica [r]. Outro ponto que merece destaque é a distribuição da energia nas diferentes bandas de freqüência. Para as consoantes laterais a distribuição de energia é muito próxima da distribuição de energia das vogais, ocorrendo abaixo de 4 kHz. Por outro lado, a distribuição de energia para as consoantes róticas ocorre nas baixas freqüências, normalmente abaixo de 1 kHz.

As Figuras de 4.30 a 4.33 mostram a variação da energia total por janela para algumas locuções contendo consoantes laterais e róticas. A energia total foi calculada a partir de janelas de análise com duração de 20 ms e deslocadas a cada 10 ms.

A análise dos gráficos da variação total da energia mostra que, para as consoantes róticas, há um vale mais acentuado e também com duração menor em relação às consoantes laterais, como já discutido. Para as consoantes laterais, como pode ser observado nas figuras, existe uma variação de energia, mas essa variação é mais suave em relação às consoantes róticas. Esse

parâmetro não apresenta grandes variações e, portanto, apenas a sua utilização não é totalmente confiável para detectar a transição entre as consoantes laterais e róticas com os outros fones.

A idéia sugerida neste trabalho para caracterizar a região de transição entre as consoantes laterais e róticas e as demais classes fonéticas é determinar o principal ponto de variação de energia no domínio da freqüência. Com esse objetivo, cinco sub-bandas de freqüências são determinadas, e a variação da energia é calculada nessas bandas.

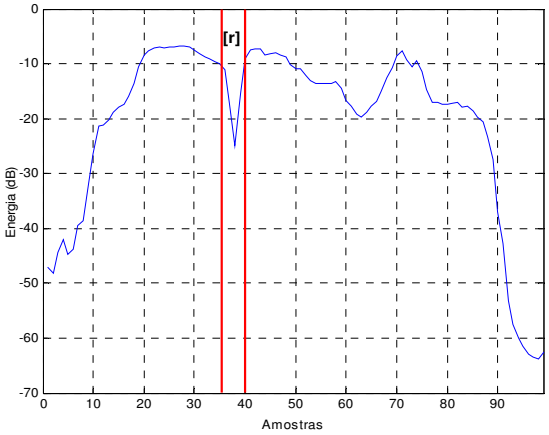


Figura 4.30: Energia total por janela para a locução "la**r**anja".

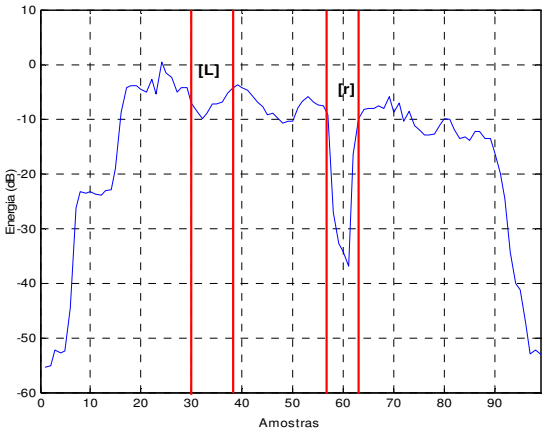


Figura 4.31: Energia total por janela para a locução "me**lh**o**r**ia".

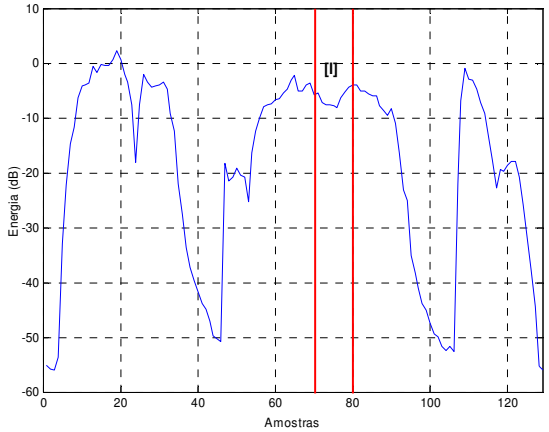


Figura 4.32: Energia total por janela para a locução "choco**l**ate".

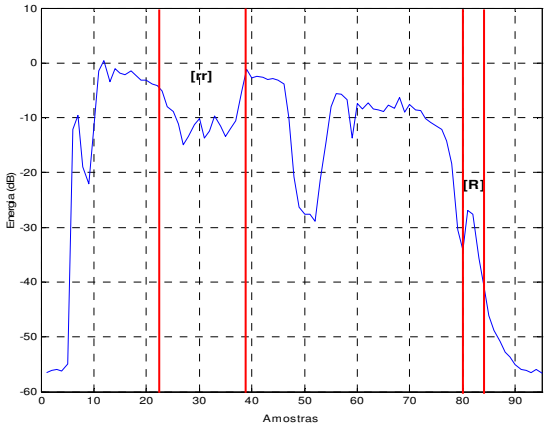


Figura 4.33: Energia total por janela para a locução "ca**rr**ega**r**".

A variação da energia espectral em cada banda de freqüência foi calculada conforme sugerido por Landan (Golipour and O'Shaughnessy, 2007). Para o cálculo da variação da energia, inicialmente é calculada a DFT de 1024 pontos, $X(k)$, do sinal de fala janelado $v(t)$ (janelas de

análise de 20 ms com deslocamento a cada 1 ms, ponderadas pela janela de Hamming). Em seguida, é calculado o módulo ao quadrado da DFT para se obter a energia $E(n)$ do sinal em cada janela de análise n. A variação foi calculada empregando a Equação (4.10), com 7 janelas adjacentes (1 janela central, 3 janelas à direita e 3 à esquerda).

$$\Delta Eb_i(n) = \frac{\left\| \sum_{\theta=-3}^{3} \theta . Eb_i(n+\theta) \right\|}{7} \tag{4.10}$$

onde $\Delta Eb_i(n)$ é a variação da energia espectral na banda de freqüência i para a janela de análise n.

Para o propósito de segmentação, a variação da energia espectral foi calculada em cinco sub-bandas: banda 1: 0 – f_a/2 Hz, banda 2: 0 – 500 Hz, banda 3: 500 – 1500 Hz, banda 4: 1500 – 2400 Hz e banda 5: 2400 – f_a /2 Hz, sendo f_a a freqüência de amostragem do sinal. A variação da energia em cada janela do sinal é somada para todas as bandas de freqüência para determinar a variação total da energia (ΔE), conforme Equação (4.11)

$$\Delta E(n) = \sum_{i=1}^{5} \Delta Eb_i(n) \tag{4.11}$$

As Figuras 4.34 a 4.37 mostram a variação da energia obtida a partir da soma das variações de energia nas cinco sub-bandas especificadas. As linhas vermelhas traçadas nos gráficos de variação da energia total representam as fronteiras para as consoantes laterais e consoantes róticas.

Como pode ser observado nas figuras, as transições são caracterizadas por picos de variação de energia. A amplitude dos picos para as consoantes [l] e [rr] é menor em relação às outras consoantes, uma vez que a distribuição da energia dessas consoantes é muito próxima da distribuição de energia das vogais e, portanto, a transição é bastante suave.

Outro ponto a ser destacado é que o cálculo da variação de energia apenas a partir da energia total de cada janela de análise não produz os mesmo resultados obtidos a partir da soma

da variação de energia nas sub-bandas especificadas. A soma das variações da energia nas bandas de freqüência determinadas realça os picos na transição entre os fones.

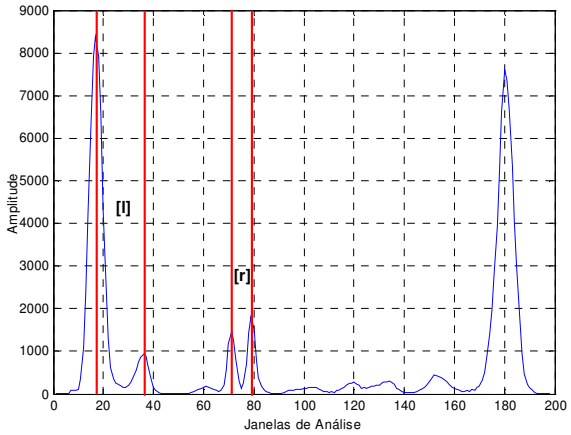


Figura 4.34: Variação da energia total para a locução "**l**a**r**anja".

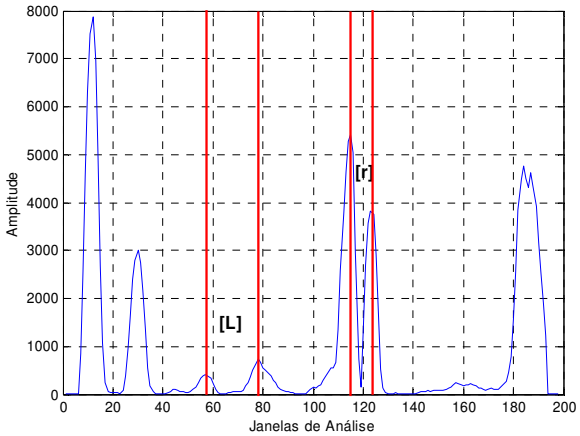


Figura 4.35: Variação da energia total para a locução "me**lh**o**r**ia".

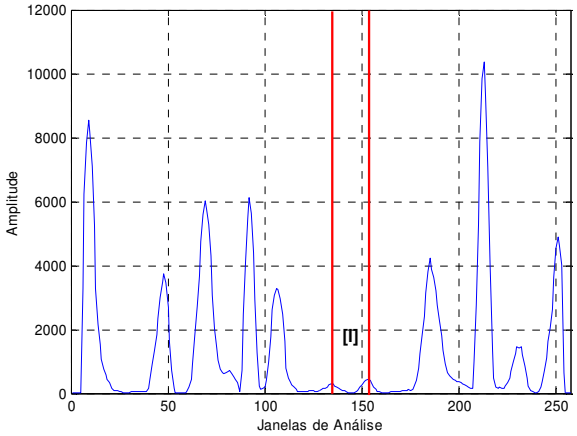


Figura 4.36: Variação da energia total para a locução "choco**l**ate".

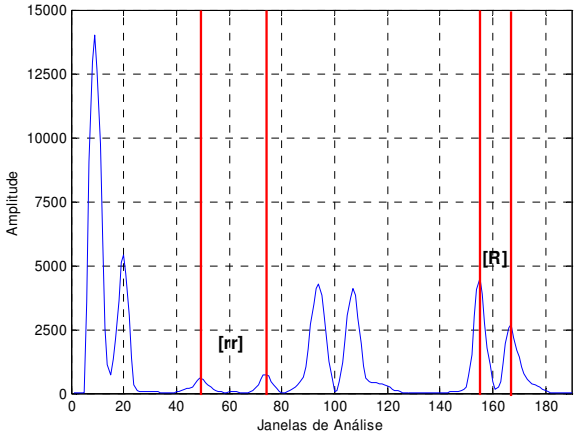


Figura 4.37: Variação da energia total para a locução "ca**rr**ega**r**".

4.3.4. Plosivas

As consoantes plosivas, como já descrito na seção anterior, representam uma classe de fonemas que é produzida a partir de uma constrição total do trato vocal por onde o ar expelido dos pulmões é forçado. Essa classe fonética é caracterizada por um período de oclusão, podendo haver ou não atividade sonora, e em seguida uma explosão que caracteriza esses sons.

Apesar do longo período de pesquisa reportado pela comunidade científica com a finalidade de descobrir características acústicas das consoantes plosivas que possam ser utilizadas de forma segura em sistemas de segmentação automática de fala, ainda não foram encontrados parâmetros suficientes ou totalmente confiáveis que possam descrever e distinguir esses fonemas. Essa dificuldade pode ser explicada pelo fato de que as consoantes plosivas são de natureza dinâmica, apresentam informações acústicas de curta duração, são dependentes de contexto e de locutor (Ali and Spiegel, 2001) e (Ali et al., 2001).

Como as consoantes fricativas, as plosivas também podem ser divididas em subclasses. Uma das principais classificações utilizadas é com relação ao ponto de oclusão na cavidade oral, podendo ser bilabiais ([p] e [b]), alveolares ([t] e [d]) e palatais ([k] e [g]). Durante o período no qual ocorre a oclusão, as pregas vocais podem ou não vibrar, dando origem às plosivas surdas e sonoras como já dito anteriormente.

Para Ahmed Ali (Ali et al., 2001), a classificação das consoantes plosivas envolve a detecção da sonoridade e o modo de articulação. Para a detecção da sonoridade, o tempo de início de vozeamento (*voice onset time* - VOT) pode ser utilizado. Com relação ao modo de articulação, um conjunto de informações deve ser utilizado tais como: o valor do segundo formante da vogal seguinte à explosão e a decisão surdo ou sonoro.

O tempo de início de vozeamento (VOT) é especificado por um número que corresponde ao intervalo entre a soltura da obstrução na articulação das consoantes plosivas e o início do vozeamento, ou seja, a vibração das pregas vocais (Lisker and Abramsom, 1964). Para as plosivas sonoras (inglês americano), o valor assumido por VOT é zero ou próximo de zero, ou seja, a vibração das pregas vocais tem início no momento da liberação da pressão do ar. Isso significa que a liberação do ar ocorre simultaneamente com o início da vibração das pregas vocais que dará início à produção do próximo fone. Um valor de VOT negativo significa que o início da sonoridade ocorre antes da liberação do ar e um valor positivo maior que zero significa que o início da sonoridade ocorre após a liberação do ar. De acordo com Kent e Read, o valor para o VOT para as plosivas sonoras pode variar aproximadamente entre -20 ms a 20 ms e para as plosivas não sonoras de 25 ms a 100 ms. Esses valores são específicos para o inglês americano.

Para Ficker (Ficker, 2003), o silêncio é uma pista necessária para a percepção de uma plosiva. Além do silêncio também existe o *burst*. Em um espectrograma, o *burst* é facilmente

localizado, pois assemelha-se ao ruído das fricativas e é representado por uma breve "faixa" vertical de energia. Segundo Kent e Read, as plosivas bilabiais apresentam concentração de energia nas baixas freqüências (500 a 1500 Hz) e pouca concentração de energia nas altas freqüências, com um espectro de freqüência decrescente. Para as plosivas alveolares, o espectro de energia é crescente, tendo uma concentração de energia nas altas freqüências (acima de 4000 Hz). A concentração de energia para as plosivas palatais está em uma faixa intermediária (1500 a 4000 Hz), entre as bilabiais e as alveolares.

A Figura 4.38 mostra a forma de onda e o espectrograma para a locução "bumbum" que contém a plosiva sonora bilabial [b] no início e no meio da locução.

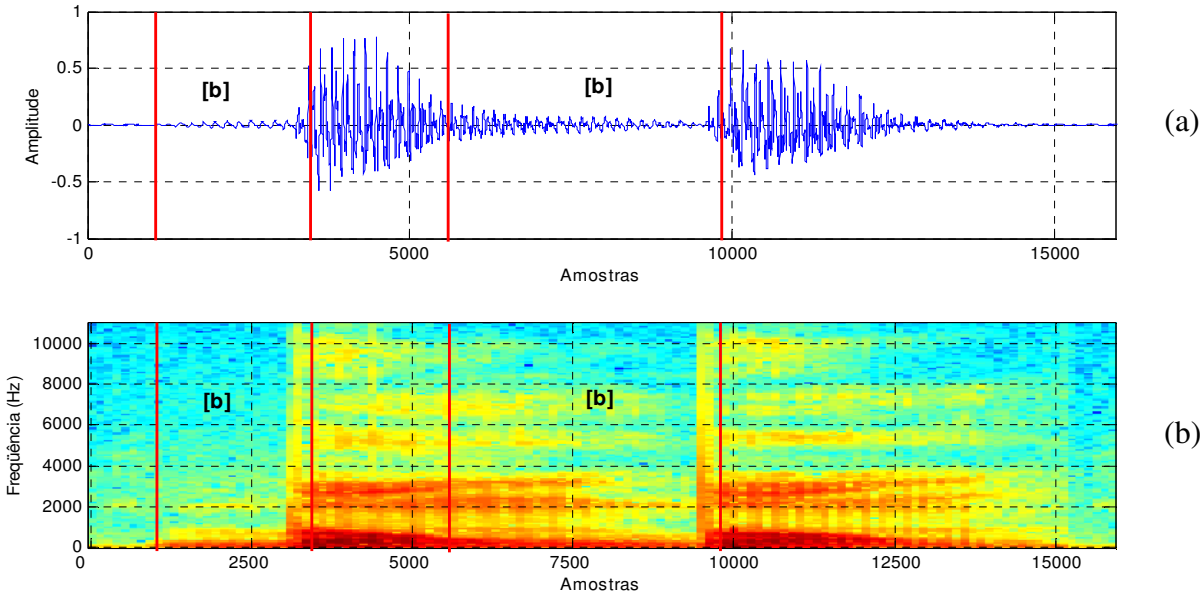


Figura 4.38: Locução "**b**um**b**um": (a) Forma de onda. (b) Espectrograma.

A análise da figura mostra que para a plosiva sonora [b] existe uma distribuição de energia nas baixas freqüências durante o período de oclusão. Essa energia é resultante da vibração das pregas vocais durante esse período. As mesmas observações são válidas para as outras plosivas sonoras [d] e [g]. O instante em que o ar é liberado não é tão nítido nas figuras, uma vez que ocorre num intervalo de tempo muito curto.

Para as plosivas surdas [p], [t] e [k] o período de liberação do ar é mais visível, uma vez que não é mascarado pela energia nas baixas freqüências. Essas observações podem ser

conferidas nas Figuras 4.39 a 4.40, que mostram as formas de onda e os espectrogramas para as locuções “tese” e “casa” que contêm as plosivas surdas [t] e [k] respectivamente.

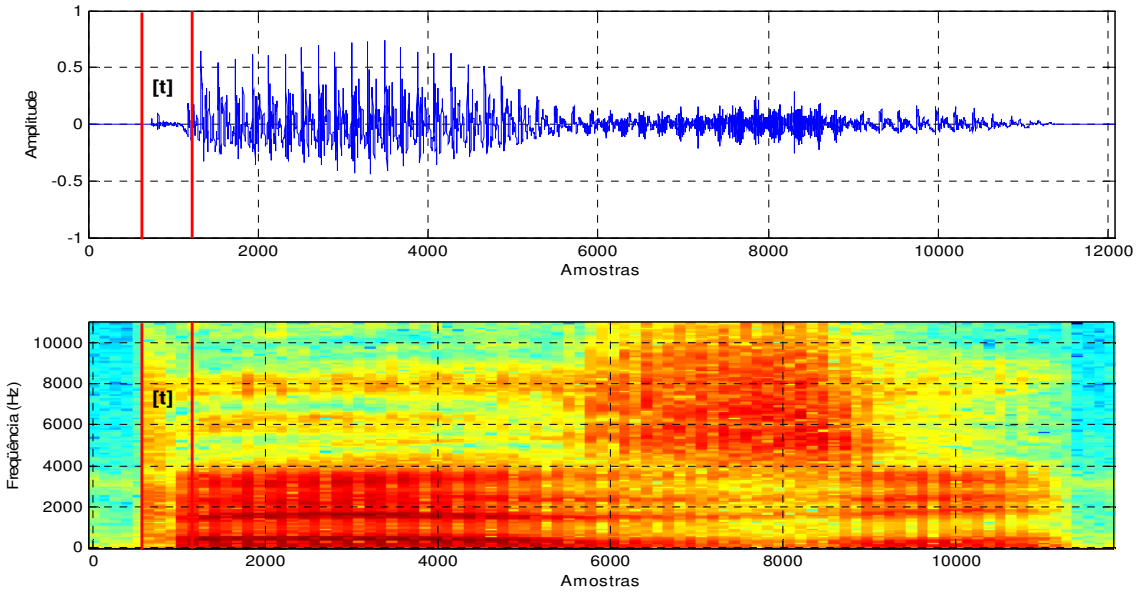


Figura 4.39: Locução “**t**ese”: (a) Forma de onda. (b) Espectrograma.

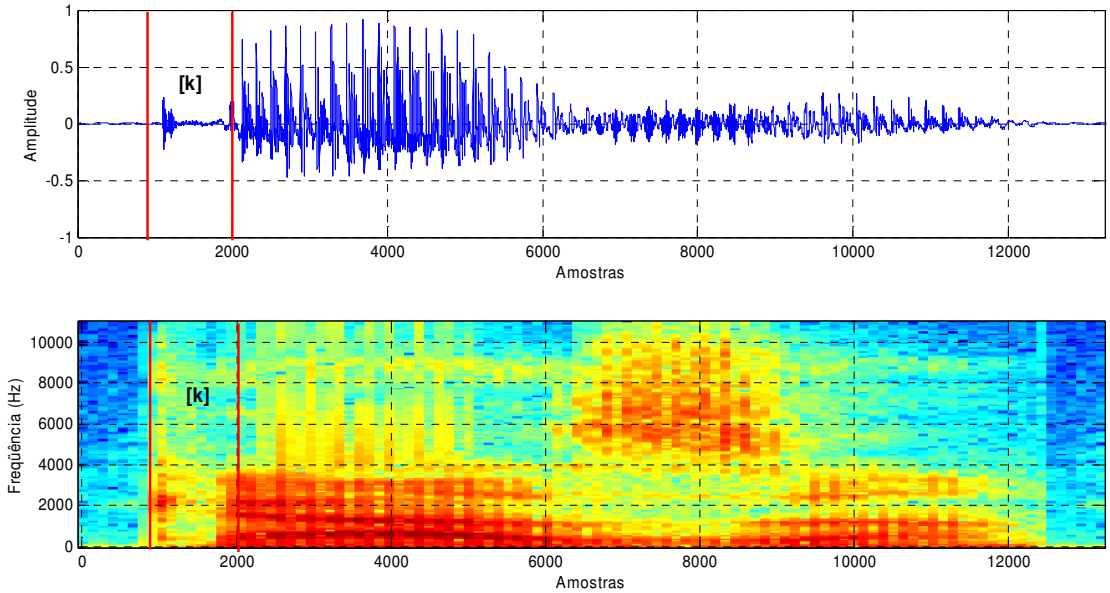


Figura 4.40: Locução “**c**asa”: (a) Forma de onda. (b) Espectrograma.

Amit Juneja (Juneja, 2004), em sua tese de doutorado, sugere dois parâmetros para classificar janelas de análise como plosivas: a energia na banda 0-F3 e F3- f_a /2, onde F3 é o valor

da terceira freqüência formante e f_a é a freqüência de amostragem do sinal. Neste trabalho, esses parâmetros serão alterados e combinados para detectar a transição entre as consoantes plosivas e outras classes de fone (normalmente as vogais).

Para a detecção das transições, primeiro é calculada a variação de energia nas duas bandas de freqüência. Em seguida, a variação nas duas bandas é somada a cada instante de tempo obtendo a variação total de energia na banda 0- f_a /2, conforme Equação (4.11). Os picos gerados representam os principais pontos de mudança de energia, que por sua vez representam as transições entre os fones.

As Figuras 4.41 e 4.42 mostram a forma de onda para os fones [# p a] extraídos da locução "pagamento" e a soma da variação de energia nas bandas especificadas. Como as plosivas têm duração curta, foram utilizadas janelas de análise com duração de 10 ms, deslocadas a cada 1 ms.

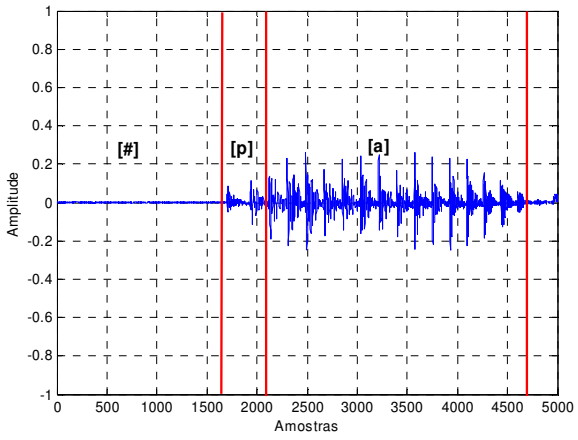


Figura 4.41: Forma de onda para o segmento [# p a] da locução "**p**agamento".

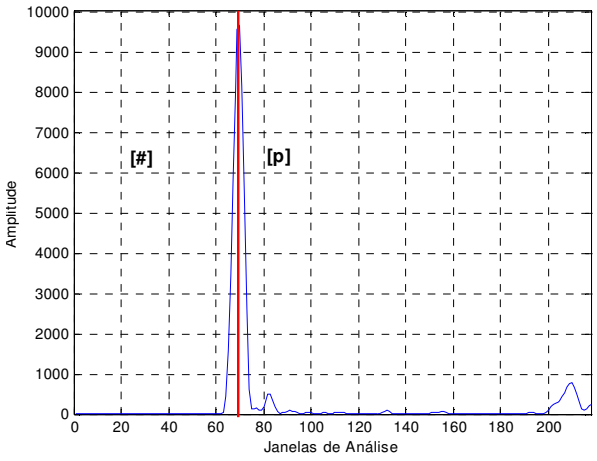


Figura 4.42: Variação da energia espectral para o segmento [# p a] da locução "**p**agamento".

O pico gerado pela variação da energia espectral ocorre na janela de análise de número 70, cujo centro corresponde à amostra 1631. Como pode ser observado na Figura 4.41, a amostra de valor 1631 não corresponde à transição entre a plosiva [p] e a vogal central [a], e sim marca o início do período em que o ar começa a ser liberado.

A princípio parecia que a alteração nos parâmetros não seria suficiente para detectar o instante de transição. Para resolver esse problema, adotou-se a seguinte estratégia: a partir do

início da liberação do ar (início da explosão), a variação da energia é novamente calculada, gerando novos picos como indicado na Figura 4.43. O pico gerado marca a transição entre a plosiva [p] e a vogal [a].

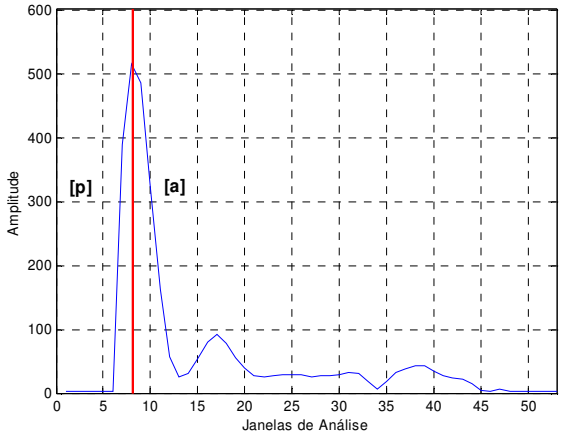


Figura 4.43: Variação da energia espectral para o segmento [p a] da locução "**p**agamento".

O primeiro pico gerado ocorre aproximadamente na janela de análise de número 8, cujo centro corresponde à amostra 264. Como a amostra inicial para o processamento foi em 1628, esse pico na verdade corresponde à amostra 1892 (1628 + 264) na Figura 4.41.

Em resumo, o processo de detecção das fronteiras entre as consoantes plosivas e outros fones é realizado em duas etapas. Na primeira etapa determina-se, através da variação da energia espectral, o início da explosão da consoante plosiva e, na segunda etapa, a partir do início da explosão, calcula-se novamente a variação da energia espectral e determina-se a transição entre a consoante plosiva e o fone seguinte (freqüentemente uma vogal). Outro ponto importante a ser mencionado é que o tamanho da janela de análise a ser empregado deve ser menor em relação à empregada para as outras classes de fones em virtude da curta duração após a soltura do ar das plosivas.

4.3.5. Africadas

Como as consoantes africadas são uma combinação das plosivas com as fricativas, a caracterização acústica também pode ser realizada através da combinação dos parâmetros acústicos das duas classes.

Das plosivas, as africadas “herdaram” o período de oclusão (que pode ser surdo ou sonoro), que é caracterizado por uma menor concentração de energia espectral em relação aos outros fones.

A segunda parte das consoantes africadas corresponde à região que segue a liberação do ar. Esta região, muito semelhante às fricativas, é caracterizada por alta concentração de energia espectral nas altas freqüências. Esta observação pode ser comprovada pela análise dos espectrogramas das Figuras 4.44 e 4.45, que mostra energia concentrada entre 2 e 6 kHz.

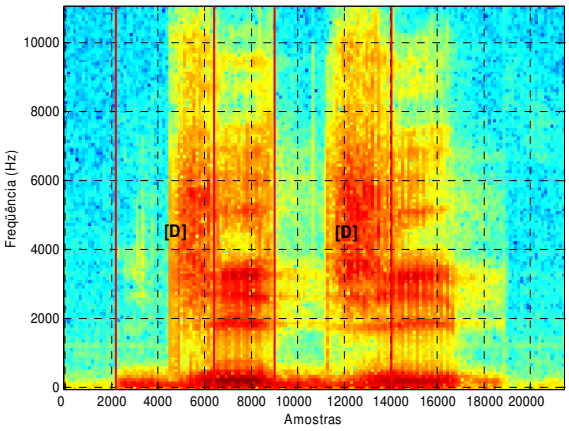


Figura 4.44: Espectrograma para a locução “**did**i”.

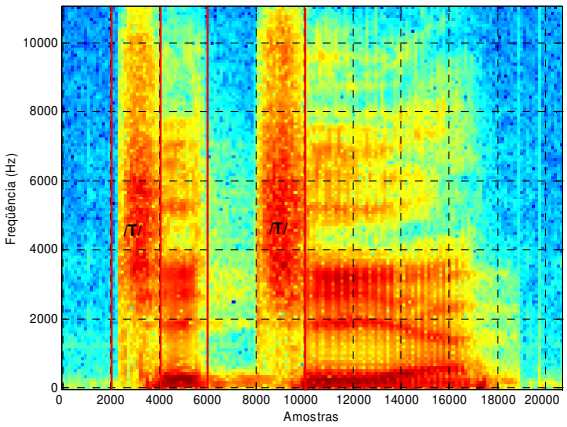


Figura 4.45: Espectrograma para a locução “**tit**ia”.

Em virtude da semelhança acústica com as fricativas, os mesmos parâmetros podem ser utilizados para caracterizar as consoantes africadas. Tendo em vista o PB, as transições entre as consoantes africadas sempre ocorrerão com a vogal anterior [i]. Tanto a taxa de cruzamentos por zero quanto o centro de gravidade espectral podem caracterizar a região da locução em que ocorrem as consoantes africadas.

Nas Figuras 4.46 e 4.47 são representadas as taxas de cruzamentos por zero para as locuções mostradas nas Figuras 4.44 e 4.45. Como era de se esperar, o comportamento da taxa de cruzamentos por zero das consoantes africadas segue o mesmo comportamento das consoantes fricativas. Através da análise das figuras percebe-se nitidamente uma separação entre as africadas e a vogal seguinte.

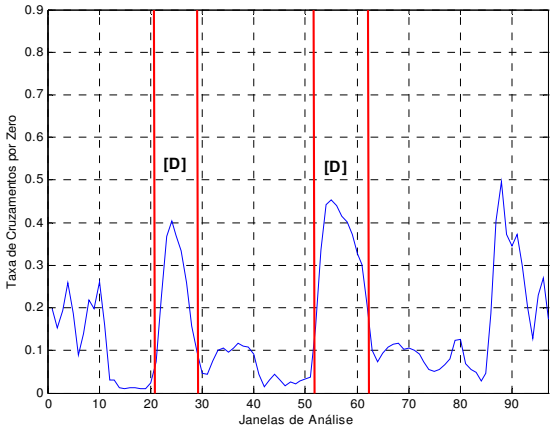


Figura 4.46: Taxa de cruzamentos por zero para a locução "**d**i**d**i".

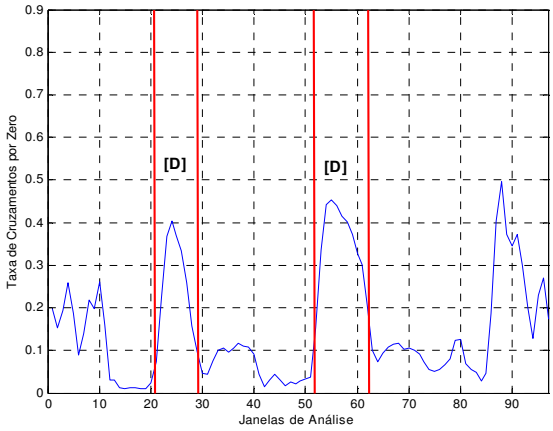


Figura 4.47: Taxa de cruzamentos por zero para a locução "**t**i**t**ia".

Outro parâmetro acústico muito característico das fricativas é o centro de gravidade espectral, que também pode ser estendido para as africadas, conforme mostrado nas Figuras 4.48 e 4.49.

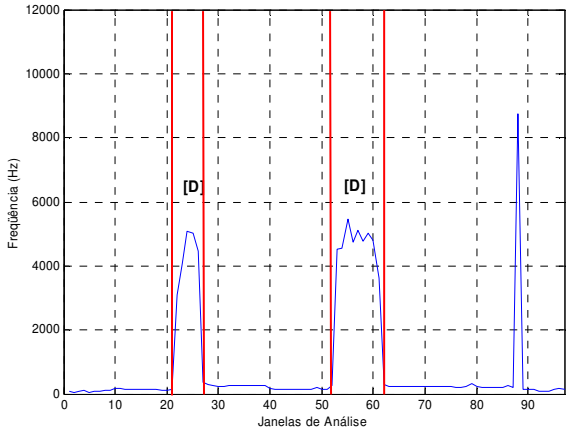


Figura 4.48: Centro de gravidade espectral para a locução "**d**i**d**i".

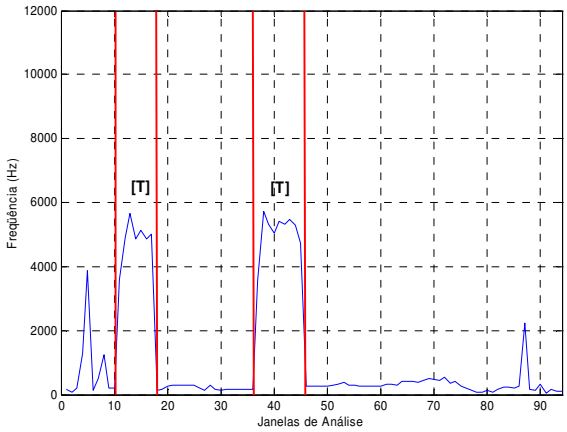


Figura 4.49: Centro de gravidade espectral para a locução "**t**i**t**ia".

As duas consoantes africadas do PB apresentam centro de gravidade espectral acima de 4 kHz, devido à concentração de energia espectral nas altas freqüências. A combinação da taxa de cruzamentos por zero com o centro de gravidade espectral mostra-se bastante promissora na detecção das fronteiras entre as africadas e as vogais. Estes dois parâmetros serão utilizados no sistema de refinamento para essa classe.

4.3.6. Consoantes Nasais

Tanto as consoantes nasais quanto as vogais nasalizadas têm sido objeto de estudo por várias décadas, estudo este motivado por aplicações em reconhecimento automático de fala e de locutor, melhoria da qualidade dos sinais de fala, avaliação clínica da qualidade da fala nasalizada, dentre outros (Pruthi, 2006). No PB existem três consoantes nasais ([m], [n] e [N]) e cinco vogais nasalizadas ([an], [en], [in], [on] e [un]).

Na literatura clássica, diversos parâmetros foram apontados com o intuito de detectar essa importante classe fonética. Um conjunto de parâmetros que retrate corretamente o comportamento desses fones de forma independente de locutor ainda precisa ser determinado. Um motivo para essa dificuldade está na variação acústica que ocorre nos diferentes locutores.

As consoantes nasais são uma classe fonética que apresenta, além de pólos, zeros na sua função de transferência, o que dificulta o seu reconhecimento automático. A presença de uma consoante nasal em um trecho de um segmento acústico pode ser detectada através de uma mudança espectral abrupta que ocorre entre a consoante nasal e um outro fone sonoro adjacente. Outra grande pista para a detecção de uma consoante nasal é a presença do murmúrio que caracteriza esses sons (Pruthi and Espy-Wilson, 2003).

Uma característica acústica importante das consoantes nasais foi detectada na década de 50 por House e Stevens (House and Stevens, 1956). Neste trabalho é mostrado que com a utilização da cavidade nasal, a amplitude do primeiro formante é reduzida e, há um aumento da largura de banda e da freqüência do formante. Foi observada também uma leve redução na amplitude do segundo e do terceiro formantes (Hattori et al., 1958), (Fant, 1960) e (Dickson, 1962).

A principal característica das consoantes nasalizadas é a presença do murmúrio nasal. O murmúrio é o segmento acústico associado exclusivamente com a irradiação nasal da energia do sinal. Segundo Fujimura (Fujimura, 1962(a, b)), o murmúrio nasal pode ser caracterizado de um modo geral através de quatro propriedades:

1. Existência do primeiro formante tendo uma baixa freqüência (normalmente abaixo dos 300 Hz) e bem separado dos outros formantes;
2. Os formantes apresentam uma taxa de decaimento rápido;
3. Alta densidade de formantes nas altas freqüências;
4. Existência de zeros no espectro do sinal;

Kent e Read também afirmam que o murmúrio nasal pode ser facilmente distinguido de outros fones não nasais através de uma comparação da energia total entre os fones.

Pruthi e Espy-Wilson (Pruthi and Espy-Wilson, 2003) desenvolveram um sistema para classificação automática das consoantes nasais e das semivogais presentes no inglês americano. Os autores utilizaram um conjunto de quatro parâmetros acústicos:

1. Medida de energia de início (*onset*) e fim (*offset*) para capturar a natureza consonantal das nasais (mudança espectral abrupta que freqüentemente ocorre no início e no momento de irradiação das nasais);
2. Medida de razão de energia nas faixas de 0 a 358 Hz (concentração de energia das consoantes nasais) e 358 a 5373 Hz (concentração de energia das vogais);
3. Medida da densidade (número) de formantes na faixa de 0 a 2500 Hz.

Na análise dos resultados obtidos, a combinação dos quatros parâmetros acústicos utilizados em um classificador baseado em máquina de vetor de suporte foi responsável por 92,4% de classificações corretas para as consoantes nasais e 88,1% para as semivogais. Segundo os autores, a baixa taxa para as semivogais deve-se, principalmente, à baixa ocorrência de alguns fones nas locuções utilizadas.

As Figuras 4.50 e 4.51 mostram os espectrogramas para as três consoantes nasais do PB ([m], [n] e [N]). Uma análise detalhada das regiões onde se encontram as consoantes nasais confirma a maior concentração de energia nas baixas freqüências e também a característica de sonoridade presente nessas consoantes.

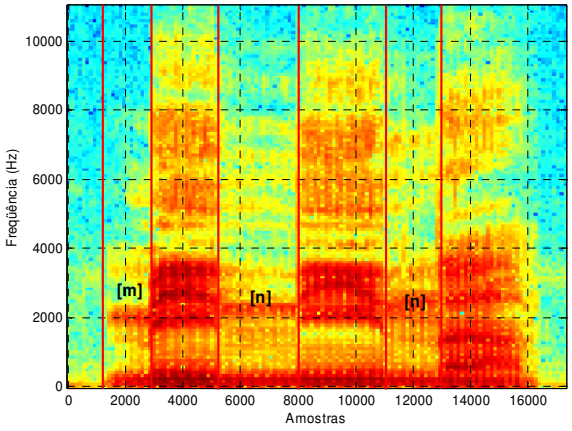


Figura 4.50: Espectrograma para a locução "**me**n**i**n**a**".

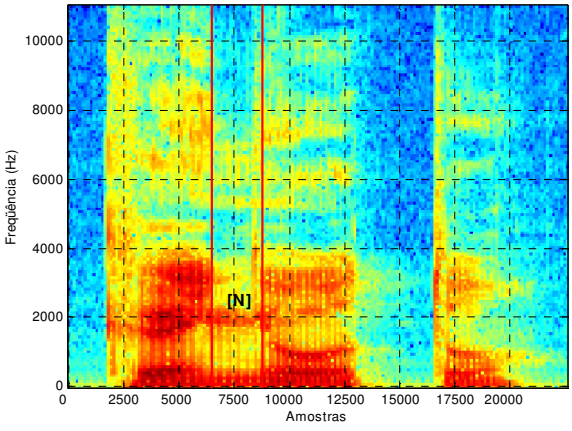


Figura 4.51: Espectrograma para a locução "ca**nh**oto".

Quanto à distribuição de energia, as Figuras 4.52 e 4.53 mostram a energia total por janela para as locuções mostradas nos espectrogramas.

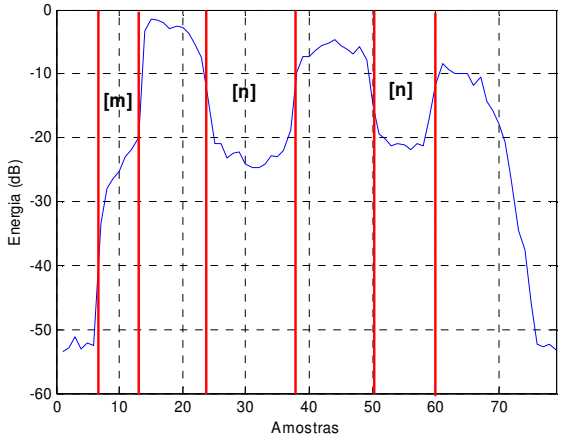


Figura 4.52: Energia para a locução "**m**e**n**i**n**a".

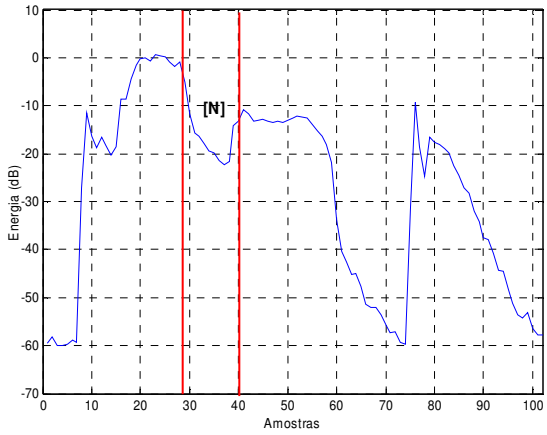


Figura 4.53: Energia para a locução "ca**nh**oto".

Percebe-se que nas janelas onde há a ocorrência das consoantes nasais, existe um vale indicando uma queda da energia. A energia total foi calculada utilizando janelas de análise com duração de 20 ms e deslocadas a cada 10 ms.

Para a detecção da fronteira entre as consoantes nasais e as outras classes de fone (as vogais) neste trabalho foi utilizada a variação da energia espectral em duas bandas de freqüência 0-358 Hz e 358-5378 Hz conforme sugeridos por Pruthi e Espy-Wilson. A variação da energia foi calculada usando a Equação (4.10) e somada a cada instante de tempo usando a Equação (4.11) de forma a destacar os picos em que ocorre a mudança de energia.

Na Figura 4.54 é apresentada a variação da energia espectral para a locução menina em que ocorrem as consoantes nasais [m] e [n] e, na Figura 4.55, para o segmento [a N o] da locução canhoto. A variação da energia foi calculada a partir de janelas de análise com duração de 20 ms e deslocadas a cada 1 ms.

Na Figura 4.54 a soma da variação da energia espectral nas duas bandas de freqüência sugeridas realça os picos da variação de energia, destacando as fronteiras entre as classes fonéticas. As mesmas observações podem ser feitas para a consoante nasal [N], mostrada na Figura 4.55. A grande diferença nas transições para as consoantes nasais mostradas na Figura

4.54 e 4.55 está na amplitude da derivada da energia espectral. A consoante nasal [N] apresenta variação de energia mais suave do que as consoantes [m] e [n].

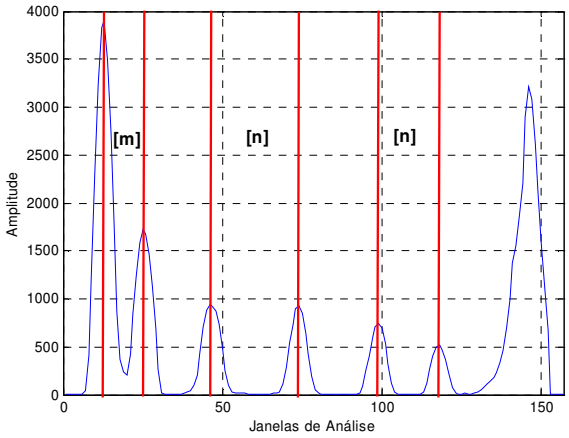


Figura 4.54: Variação da energia espectral nas bandas [0-358 Hz] e [358-5378 Hz] para a locução "**me**n**in**a".

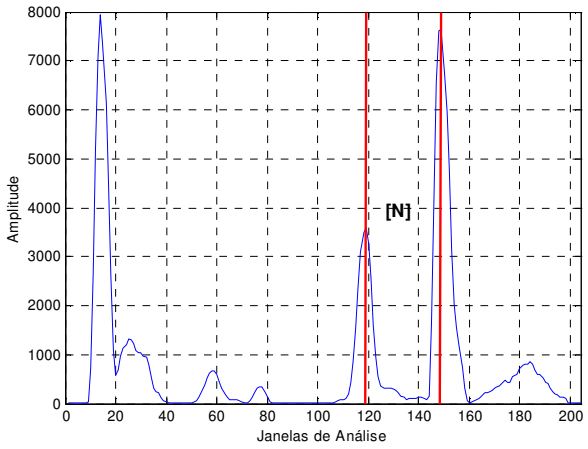


Figura 4.55: Variação da energia espectral nas bandas [0-358 Hz] e [358-5378 Hz] para o segmento [a N o] da locução "ca**nh**oto".

4.3.7. Silêncio

Neste trabalho, o silêncio, presente no início e no final de cada locução, é tratado como um fone e a sua representação na transcrição fonética adotada é [#]. Como uma classe fonética, sua caracterização acústica é a mais simples, levando em consideração que tanto as locuções de treinamento quanto às de teste utilizadas para testar o sistema são livres de ruído.

Durante o período de oclusão, nenhuma atividade sonora ou quase nenhuma é realizada e, portanto, um bom parâmetro para caracterizar o silêncio é a energia total por janela, que é muito baixa em relação à de outros fones.

A detecção do silêncio presente no início e no final da locução já foi objeto de estudo na comunidade científica. Rabiner (Rabiner and Hunag, 1993) sugere a utilização da energia combinada com a taxa de cruzamentos por zero. Pelos testes realizados pôde-se verificar que apenas a energia seria suficiente para caracterizar período de silêncio no início e no final de cada locução.

A Figura 4.56 mostra a forma de onda e o espectrograma da locução "faixa", destacando os instantes de silêncio que ocorrem no início e no final da locução.

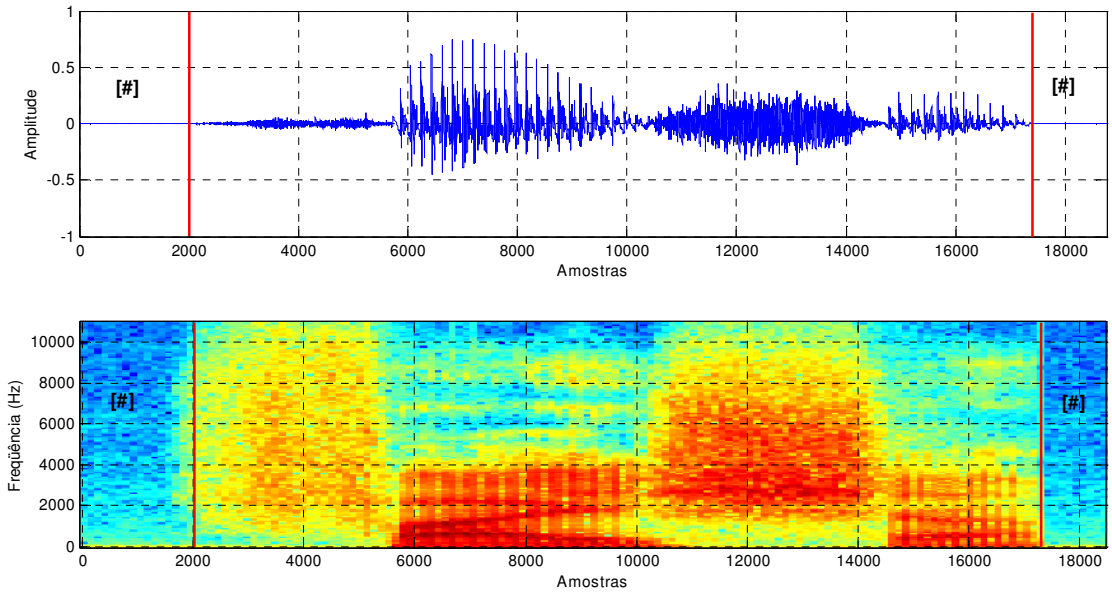


Figura 4.56: Locução “faixa”: (a) Forma de onda. (b) Espectrograma.

4.4. Considerações Finais

Neste Capítulo foi descrito o mecanismo básico responsável pela produção da voz humana, bem como a classificação dos diferentes fones e suas características acústicas. Entender esse processo é vital em sistemas de reconhecimento automático e segmentação automática de fala, bem como nos processo de síntese.

Como destacado, a fala é uma onda de pressão acústica originada a partir dos órgãos do aparelho fonador humano, que por sua vez é composto por vários articuladores. A classificação para os sons da fala pode ser realizada de diversas maneiras, levando em consideração desde a fonte de excitação (sons sonoros ou surdos), até a posição dos articuladores durante o processo de produção.

As vogais, que podem ser nasalizadas ou não, representam o núcleo central das sílabas produzidas no PB. A produção desses sons está diretamente relacionada com a vibração das pregas vocais. Apresentam relativamente alta energia nas freqüências abaixo de 4 kHz e podem ser diferenciadas entre si pelas freqüências de ressonância do trato vocal. Outra característica das vogais é uma baixa taxa de cruzamentos por zero e um perfil de energia concentrado em 2 kHz.

Na produção das consoantes, os articuladores do trato vocal impõem obstáculos à passagem do ar, produzindo dessa forma sons com diferentes características acústicas que podem ser subdivididos em seis classes: fricativas, plosivas, africadas, laterais, róticas e nasais.

As consoantes fricativas (surdas ou sonoras) são caracterizadas por uma alta taxa de cruzamentos por zero e pelo centro de gravidade espectral acima de 3500 Hz. As consoantes laterais e róticas apresentam características acústicas muito próximas das vogais, com uma estrutura bem definida de formantes. Para a sua caracterização foi proposta a variação de energia espectral em algumas bandas de freqüências.

Dentre as consoantes, a classe mais difícil de ser acusticamente modelada é a das plosivas, pela sua reduzida duração e também por ser muito dependente do contexto. Sua caracterização também foi realizada pela variação da energia espectral. As africadas representam uma combinação entre as consoantes plosivas e as consoantes fricativas, e são caracterizadas acusticamente pela taxa de cruzamentos por zero e pelo centro de gravidade espectral.

As consoantes nasais têm como característica principal a concentração de energia nas baixas freqüências, o que pode ser muitas vezes confundido com as vogais posteriores. Sua caracterização também é realizada através da variação da energia espectral.

Os parâmetros acústicos descritos neste Capítulo são os principais representantes de cada classe fonética e são largamente utilizados em classificação de fones. Todos os parâmetros foram estudados e determinados a partir da base de fala dependente de locutor masculino do PB, com taxa de amostragem de 22,05 kHz. Os limiares de cada parâmetro permanecerão inalterados durante o processo de refinamento para a base de fala independente de locutor (TIMIT), com exceção do perfil energia que será utilizada a taxa de 70%.

Capítulo 5

Refinamento da Segmentação Automática de Fala

Os Capítulos 2, 3 e 4 tiveram como objetivo central apresentar os fundamentos teóricos sobre a modelagem estatística a ser empregada neste trabalho: o algoritmo de Viterbi responsável por gerar as estimativas das fronteiras de segmentação, o estado da arte em segmentação automática de fala e o refinamento da segmentação automática. Um estudo sobre o processo de produção de fala e as características das principais classes de fones presentes no português do Brasil também foram apresentados, uma vez que serão necessários para o refinamento da segmentação automática de fala.

Neste capítulo será apresentada a arquitetura do sistema desenvolvido, que por sua vez é composta basicamente por três módulos: treinamento, segmentação e refinamento. Cada módulo será descrito detalhadamente, enfatizando-se o módulo de refinamento e as regras para refinar cada classe de fone.

5.1. Arquitetura do Sistema Baseado em Regras

Como mostrado no Capítulo 3, existe uma diversidade muito grande de técnicas de segmentação automática de fala, todas com suas vantagens e desvantagens. As técnicas básicas variam entre a segmentação implícita e a segmentação explícita.

Com base na revisão bibliográfica e nas fundamentações teóricas sobre a produção e parametrização da fala exposta no Capítulo 3, definiu-se como objetivo da tese o desenvolvimento de um sistema para segmentação automática de fala baseado no algoritmo de Viterbi, seguido de um processo de refinamento automático das marcas de segmentação baseado nas características acústicas de cada classe de fones. Apesar de o algoritmo de Viterbi apresentar bons resultados para a segmentação, o refinamento das marcas de segmentação é necessário para diminuir erros que normalmente ocorrem. O motivo de se ter escolhido o algoritmo de Viterbi

para realizar a segmentação de fala justifica-se por ser um algoritmo simples, e apresentar bons resultados, como anteriormente observado.

Durante a fase de revisão bibliográfica, nenhum trabalho explorando as características acústicas dos fones para o refinamento das marcas de segmentação foi encontrado. A maioria dos trabalhos utiliza técnicas recentes que normalmente apresentam uma complexidade computacional alta ou necessitam de grande quantidade de material de treinamento.

O sistema proposto tem inspiração no trabalho de Amit Juneja (Juneja, 2004), em que características acústicas dos fones foram utilizadas para realizar uma classificação fonética em cinco grandes classes (silêncio, vogais, fricativas, plosivas e consoantes sonoras). Após a classificação, máquinas de vetor de suporte foram utilizadas para o reconhecimento dos fones.

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

A Figura 5.1 ilustra a arquitetura dos módulos de treinamento e segmentação e a Figura 5.2 o módulo de refinamento do sistema desenvolvido.

Para o módulo de treinamento dos HMMs duas informações são essenciais: as locuções de treinamento e as respectivas transcrições fonéticas. Para o módulo de segmentação é necessário apresentar ao sistema as locuções de teste e também suas respectivas transcrições fonéticas.

Como pode ser observado nas Figuras 5.1 e 5.2, a transcrição fonética das locuções é necessária para todos os módulos do sistema desenvolvido. Essa informação utilizada pelo algoritmo de Viterbi caracteriza uma segmentação explícita.

No início do desenvolvimento do sistema proposto, todos os módulos foram desenvolvidos em C++ Builder para ambiente Windows. Com a idéia de se trabalhar com fones dependentes de contexto, foi adotado para treinamento e segmentação o *software* HTK. Este software apresenta facilidades na geração e também no treinamento dos HMMs dependentes de contexto.

O HTK foi adotado apenas para o treinamento e segmentação, uma vez que o objetivo deste trabalho não é desenvolver uma técnica específica de treinamento voltada para a segmentação. O módulo de refinamento que será descrito nas próximas seções não faz parte do HTK.

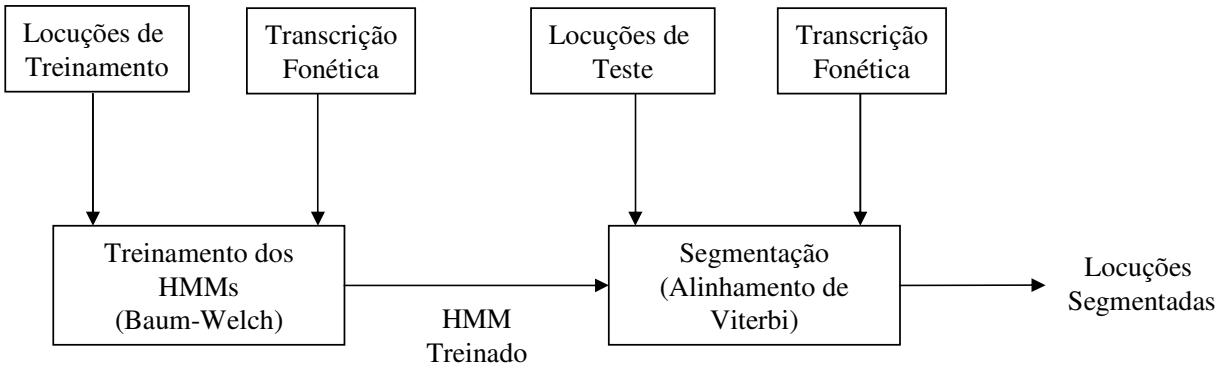


Figura 5.1: Arquitetura dos módulos de treinamento e segmentação do sistema proposto.

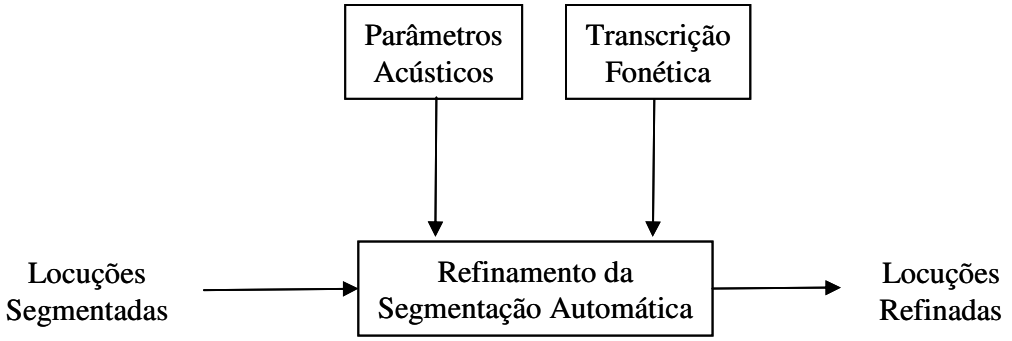


Figura 5.2: Arquitetura do módulo de refinamento do sistema proposto.

5.2. Módulo de Treinamento

Como mostrado na Figura 5.1, para o treinamento dos HMMs associados às unidades acústicas são necessárias as locuções de treinamento e as respectivas transcrições fonéticas. A partir da transcrição fonética são gerados os modelos acústicos de cada unidade que compõe a locução. O modelo da locução é gerado concatenando-se os HMMs de cada unidade fonética. O treinamento foi realizado para modelos acústicos dependentes e independentes de contexto.

Para a geração dos modelos acústicos, cada unidade fonética foi representada por um HMM contínuo com três estados com topologia do tipo *left-right***,** sem salto duplo, conforme mostra a Figura 5.3.

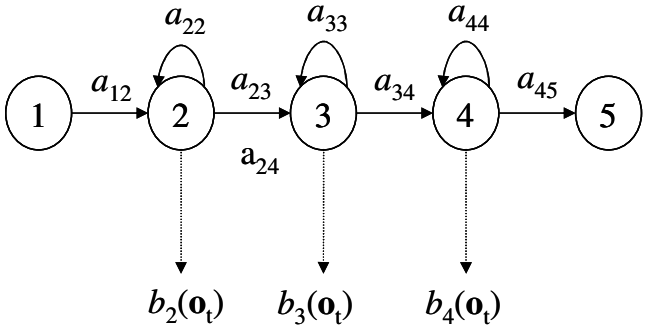


Figura 5.3: Modelo acústico de um fone baseado em HMM.

Dois estados adicionais (estado de entrada e saída), não emissores, foram adicionados ao modelo como sugestão do *software* HTK em que os modelos acústicos foram treinados. A idéia dos estados de entrada e saída é facilitar a junção entre os modelos, ou seja, o estado de saída do modelo de um fone é unido ao estado de entrada de outro, criando dessa forma um HMM composto.

O HTK também sugere a introdução de um modelo de pausa entre as palavras. A pausa é um modelo simplificado, com um estado de entrada, um estado emissor e um estado de saída.

A modelagem da função densidade de probabilidade para emissão dos símbolos foi feita através de uma mistura de Gaussianas, em que o número de componentes na mistura foi determinado através de testes e será descrito no próximo Capítulo 6. No modelo será considerada uma matriz de covariância diagonal dos vetores de parâmetros, considerando componentes independentes entre si.

Tendo o modelo das unidades fonéticas, cada locução passa por uma etapa de pré-processamento, conforme mostrado na Figura 5.4.

Durante a fase de pré-processamento, inicialmente o nível DC de cada sentença é subtraído. Em seguida, as sentenças são submetidas a uma filtragem de pré-ênfase através do filtro passa-altas ($1-cz^{-1}$), em que c é o coeficiente do filtro. Nas simulações usando a base de fala dependente de locutor, foi utilizado o coeficiente $c = 0{,}95$ e, nas simulações usando a TIMIT, foi

utilizado c = 0,97 (HTKBook, 2006). A pré-ênfase é empregada para compensar a queda espectral que ocorre no sinal de fala devido à combinação dos efeitos do espectro dos pulsos glotais (-12 dB/oitava) e da irradiação dos lábios (6 dB/oitava).

Figura 5.4: Diagrama em blocos da etapa de pré-processamento do sinal de fala.

Após a filtragem, cada sentença é submetida ao janelamento de Hamming, $v(n) = x(n)w(n)$, em que $v(n)$ é o sinal janelado, $x(n)$ é o sinal antes do janelamento e $w(n)$ é a janela de Hamming, definida conforme Equação (5.1):

$$w(n) = \begin{cases} 0{,}54 - 0{,}46\cos\left(\dfrac{2\pi n}{N-1}\right) & 0 \le n \le N\text{-}1 \\ 0 & \text{para } \forall\, n \text{ fora do intervalo} \end{cases} \tag{5.1}$$

Para o treinamento foram utilizadas janelas com duração de 20 ms e deslocamento a cada 10 ms. Dessa forma, para a base de fala independente de locutor, amostrada a 22,05 kHz, N = 441 e para a TIMIT, amostrada a 16 kHz, N = 320.

O sinal janelado $v(n)$ é submetido a um banco de filtros cujas saídas serão utilizadas para calcular os parâmetros mel-cepstrais. Inicialmente calcula-se a DFT do sinal com 1024 pontos para as duas bases de fala. Como o treinamento foi realizado no HTK, este utiliza um banco ligeiramente diferente do proposto por Picone (Picone, 1993). São utilizados 26 filtros. Na saída de cada filtro é calculado o logaritmo da energia e então determinam-se os coeficientes mel-cepstrais de acordo com a Equação (5.2):

$$MFCC_i = \sum_{m=1}^{M} \left(E_m \cos\left[i\left(m - \frac{1}{2} \right) \frac{\pi}{M} \right] \right) \quad i = 1, 2, 3, ..., 12 \tag{5.2}$$

onde M é o número total de filtros do banco, i é o número do coeficiente mel-cepstral (neste trabalho foram utilizados 12 coeficientes) e E_m é o logaritmo da energia calculada na saída do *m-ésimo* filtro.

Além dos coeficientes mel-cepstrais, o logaritmo da energia total também é calculado para compor o vetor de características acústicas. O logaritmo da energia é calculado usando a Equação (5.3) para o sinal janelado conforme descrito na Equação (5.1).

$$LogEnergia = 10\log_{10}\left(\sum_{i=0}^{N-1} v(i)^2 \right) \tag{5.3}$$

Para a normalização da energia determina-se o maior valor do logaritmo da energia calculado para cada locução completa. Em seguida subtrai-se o maior valor do logaritmo da energia dos outros quadros da locução.

A última etapa durante a determinação do vetor de características consiste em calcular a derivada de primeira e segunda ordem dos coeficientes mel-cepstrais e do logaritmo da energia normalizada. A aproximação para a derivada é calculada usando a Equação (5.4), como definida no HTKBook (HTKBook, 2006).

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{5.4}$$

onde d_t são os coeficientes delta no instante de tempo t em função dos coeficientes estáticos $c_{t+\theta}$ e $c_{t-\theta}$. Para a base dependente de locutor foi utilizada apenas uma janela adjacente de cada lado para o cálculo dos parâmetros delta, e para a TIMIT foram utilizadas duas janelas adjacentes de cada lado da janela em consideração.

Após o cálculo de todos os parâmetros, estes são agrupados em um único vetor de dimensão 39. O vetor é constituído por: 1 LogEnergia, 12 MFCC, 1 ΔLogEnergia, 12 ΔMFCC, 1 ΔΔLogEnergia, 12 ΔΔMFCC.

Uma vez modeladas as unidades acústicas que constituem cada uma das locuções, estas são submetidas ao processo de treinamento. O módulo de treinamento utiliza o algoritmo de Baum-Welch, conforme descrito no Capítulo 2. Para o algoritmo de treinamento deverão ser fornecidas as seguintes informações: o conjunto das subunidades fonéticas utilizadas na transcrição fonética, a transcrição fonética de cada locução de treinamento, e as locuções parametrizadas.

5.3. Módulo de Segmentação

A segmentação das locuções é realizada pelo algoritmo de Viterbi. Para executar a segmentação é necessário que os HMMs associados às unidades fonéticas estejam treinados, pois o algoritmo de Viterbi utiliza no alinhamento, a cada instante de tempo *t*, a matriz de transição de estados e as probabilidades de emissão dos símbolos calculadas na fase de treinamento.

Para segmentar uma determinada locução, inicialmente é gerado o modelo da locução com base na transcrição fonética da mesma. Em seguida, é realizada a fase de pré-processamento como descrito na seção anterior e, a cada instante de tempo *t*, os parâmetros acústicos de cada janela da locução são apresentados ao algoritmo de Viterbi, que calcula a probabilidade do modelo emitir os símbolos acústicos. Durante a fase de treinamento e de segmentação, os parâmetros acústicos são calculados a partir de janelas de análise com duração de 20 ms e deslocadas a cada 10 ms.

Após apresentar todos os parâmetros acústicos de todas as janelas da locução, é possível recuperar o caminho que contenha as maiores probabilidades determinadas, ou seja, o caminho ótimo de Viterbi. Através desse caminho ótimo estima-se o número de janelas em cada estado do HMM e, através desse número, determinam-se os instantes de tempo que correspondem às transições entre os fones adjacentes.

As fronteiras calculadas pelo algoritmo de Viterbi apresentam erro de quantização de 10 ms (valor de deslocamento entre as janelas de análise). Alguns trabalhos empregam deslocamentos menores durante a fase de segmentação, como por exemplo, 3 ms (Toledano et al., 2003). Neste trabalho, deslocamentos menores não apresentaram bons resultados.

5.4. Módulo de Refinamento

O último módulo do sistema proposto, e que é a proposta deste trabalho, é o módulo de refinamento.

Como o refinamento é baseado nas características acústicas dos fones, primeiro os 38 fones utilizados para a base dependente de locutor foram agrupados em 15 classes fonéticas, e os 48 fones usados para a base independente de locutor foram agrupados em 13 classes. A Tabela 5.1 mostra a descrição das classes e dos respectivos fones determinados para as bases em Português dependente de locutor (masculino e feminino) e a Tabela 5.2 mostra os fones utilizados para a base independente de locutor (TIMIT).

Tabela 5.1: Classes fonéticas para as bases em Português dependente de locutor.

Classes	**Fones**
Silêncio	[#]
Fricativas Surdas	[x], [s], [f]
Fricativas Sonoras	[v], [z], [j]
Vogais Anteriores	[i], [e], [E], [y]
Vogal Central	[a]
Vogais Posteriores	[o], [O], [u]
Vogais Nasais	[an], [en], [in], [on], [un]
Plosivas Surdas	[p], [t], [k]
Plosivas Sonoras	[b], [d], [g]
Africadas	[T], [D]
Laterais	[l], [L]
Róticas	[r], [rr], [R]
Consoantes Nasais	[m], [n], [N]
Voiced closure	[vcl]
Unvoiced closure	[cl]

Para a base independente de locutor, as consoantes laterais e róticas foram agrupadas em uma única classe que foi chamada de semi-vogais. Outra classe que não está presente na base independente de locutor é a das vogais nasais. A classificação utilizada neste trabalho é a mesma sugerida por Pruthi, ou seja, uma vogal é considerada nasal quando é imediatamente seguida por uma consoante nasal (Pruthi and Espy-Wilson, 2007). Essa observação vale para o Inglês e nem sempre acontece no PB.

O processo de refinamento proposto é baseado em um conjunto de regras. Cada regra, por sua vez, é formada por alguns parâmetros. Os parâmetros determinados são os mais representativos de cada classe e também são largamente utilizados para a classificação de fones. Alguns parâmetros utilizam um limiar para determinar a nova posição da marca de segmentação e outros são baseados na detecção de picos (*peak-picking*). O número de parâmetros para cada classe é variável.

Tabela 5.2 – Classes fonéticas para a base independente de locutor (TIMIT).

Classes	**Fones**
Silêncio	[sil]
Fricativas Surdas	[sh], [s], [th], [f]
Fricativas Sonoras	/zh/, [z], [v], [dh]
Vogais Anteriores	[eh], [ey], [ih], [ix], [iy], [er]
Vogal Central	[aa], [ae], [ah], [ao], [aw], [ax], [ay]
Vogais Posteriores	[oy], [ow], [uh], [uw], [ux]
Plosivas Surdas	[p], [t], [k], [dx]
Plosivas Sonoras	[b], [d], [g]
Africadas	[jh], [ch]
Semi-vogais	[l], [r], [w], [y], [hh], [el]
Consoantes Nasais	[m], [n], [ng], [en]
Voiced closure	[vcl]
Unvoiced closure	[cl]

Para a determinação dos parâmetros, um módulo adicional foi desenvolvido com a finalidade de calcular os parâmetros de uma determinada janela de análise. Inicialmente esse módulo foi utilizado para estimar os limiares de cada parâmetro acústico que compõem as regras de refinamento. Os limiares determinados representam os pontos mais prováveis onde ocorre a transição entre dois fones.

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

Os limiares determinados usando a base dependente de locutor masculino são os mesmos limiares empregados para o refinamento da base de fala com locutor feminino e também com a TIMIT. Embora os limiares tenham sido calculados para um locutor masculino, eles apresentaram bom desempenho tanto na base de fala feminina quanto na TIMIT, como será discutido no próximo Capítulo 6.

O processo de refinamento, para cada uma das fronteiras previamente determinadas, leva em consideração a classe fonética do lado direito e do lado esquerdo da fronteira em análise. Um levantamento de todas as possíveis transições foi realizado com o objetivo de determinar qual ou quais os parâmetros que melhor se aplicam a cada tipo de transição e dessa forma construir as regras.

Essa estratégia foi adotada com o objetivo de tentar reduzir o número de possíveis combinações de parâmetros nas regras e também poder trabalhar com os melhores parâmetros para cada tipo de transição. Por exemplo, em uma transição onde estão presentes as consoantes fricativas e as vogais, uma série de parâmetros poderiam ser empregados tais como: taxa de cruzamentos por zero, centro de gravidade espectral, energia total por janela de análise, trajetória de formantes, etc. Através desse levantamento pode-se concluir que apenas a taxa de cruzamentos por zero e o centro de gravidade espectral são suficientes para determinar a fronteira com precisão.

Na Tabela 5.3 é apresentado um resumo das principais transições mapeadas e os parâmetros acústicos empregados.

Tabela 5.3: Parâmetros acústicos utilizados em cada tipo de transição fonética.

Tipo de Transição	Parâmetro Acústico
Fricativas/africadas + demais classes	- Taxa de cruzamentos por zero - Centro de gravidade espectral
Demais classes + fricativas/africadas	
Silêncio + demais classes	- Energia total da janela de análise
Demais classes + silêncio	
Laterais/Róticas + demais classes	- Pico resultante da soma da variação da energia espectral em 5 bandas: Et+E[0-500]+E[500-1500]+E[1500- 2400]+E[2400-fs/2]
Demais classes + laterais/róticas	
Consoantes nasais + demais classes	- Pico resultante da soma da variação da energia espectral em 2 bandas de freqüência: E[0-358]+E[358-5378]
Demais classes + consoantes nasais	
Plosivas + demais classes	- Pico resultante da soma da variação da energia espectral em duas bandas de freqüência: E[0-F3]+E[F3-fs/2]
Demais classes + plosivas	
Vogal central + vogal posterior/anterior	- Variação de F1 e F2
Vogal posterior/anterior + vogal central	
Vogal posterior + vogal anterior	- Variação de F2 - Perfil energia (75%)
Vogal anterior + vogal posterior	
Entre fones da mesma classe fonética	- Critério de informação Bayesiana (BIC)

Durante o processo de refinamento, cada marca de segmentação é analisada separadamente. Ao analisar cada marca de segmentação sabe-se, através da transcrição fonética, quais são os fones que cada marca separa. Com base nos fones, determina-se a qual classe eles pertencem e, conseqüentemente, qual o tipo de transição envolvido. Com base no tipo de transição um conjunto de parâmetros acústicos é empregado para refinar as fronteiras. Por exemplo, se uma marca de segmentação separa o silêncio inicial de uma locução com uma fricativa, sabe-se que na transição para a consoante fricativa o valor da energia total vai aumentar

até ultrapassar um determinado limiar. O instante de tempo em que o limiar é atingido é definido como a fronteira entre os dois fones.

Um ponto que deve ser levado em consideração ao analisar as marcas de segmentação para o refinamento é definir o intervalo na qual os parâmetros acústicos serão calculados e, conseqüentemente, em que intervalo as marcas de segmentação poderão ser deslocadas. O intervalo de refinamento foi inicialmente proposto como sendo limitado ao intervalo formado pelas marcas imediatamente anterior e posterior à marca de segmentação que está sendo refinada, e que foram determinadas pelo alinhamento forçado de Viterbi. Conseqüentemente, a marca de segmentação em análise só poderá ser deslocada neste intervalo. A Figura 5.5 ilustra o intervalo de refinamento inicialmente proposto.

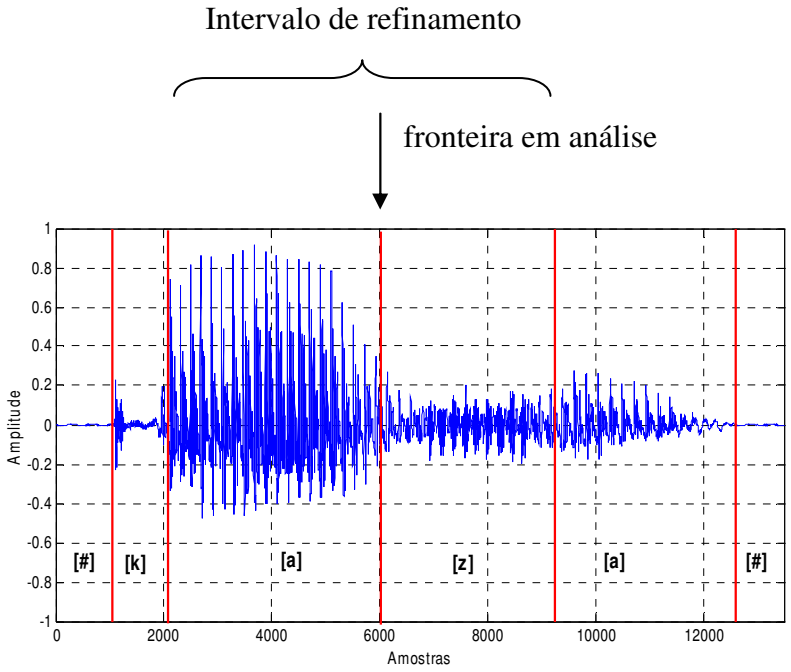


Figura 5.5: Intervalo de refinamento inicialmente proposto.

Com o intervalo de refinamento inicialmente proposto pôde-se verificar que, em algumas situações, o processo de refinamento foi prejudicado. Essa situação ocorre quando o início do intervalo de refinamento não ocorre na fronteira anterior a fronteira que está sendo refinada. Ao calcular os parâmetros acústicos, parte do fone imediatamente anterior é levada em consideração, o que por sua vez prejudica a detecção da nova posição da fronteira em análise. Esse problema foi constatado apenas no início do intervalo de refinamento. A solução adotada foi definir o início

do intervalo de refinamento como o ponto médio entre a fronteira em análise e a fronteira imediatamente anterior, que seria o início do intervalo. Essa alteração é mostrada na Figura 5.6, onde o início do intervalo é mostrado pela linha pontilhada. Esse processo é adotado para todas as fronteiras que serão refinadas.

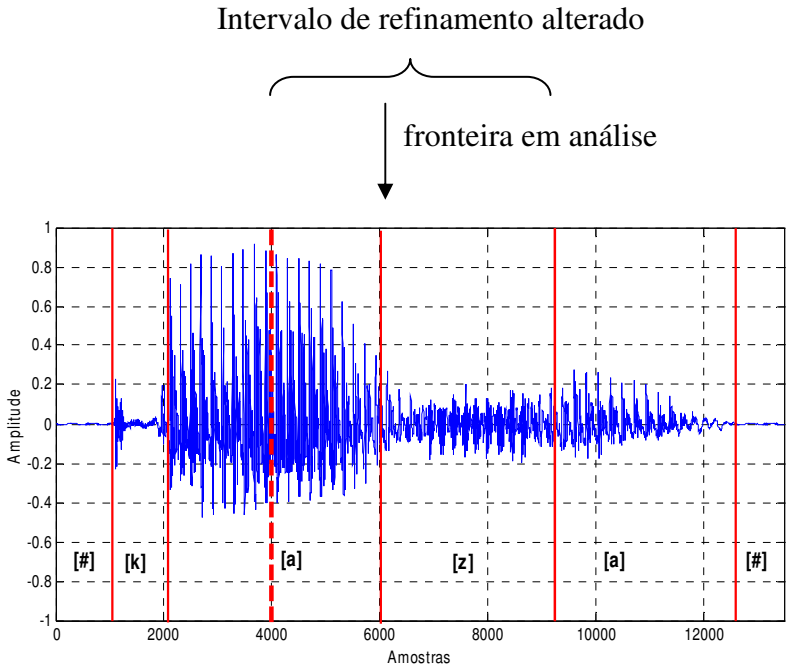


Figura 5.6: Intervalo de refinamento alterado.

Durante a fase de refinamento são calculados os parâmetros especificados nas regras apenas no intervalo proposto. O tamanho da janela de análise é de 20 ms, com exceção para as plosivas que é de 10 ms. Todas as janelas são deslocadas a cada 1 ms de forma a obter maior precisão na segmentação durante a fase de refinamento. A nova fronteira de segmentação será o centro da janela de análise cujos parâmetros cruzam os limiares previamente estabelecidos ou apresentam o maior pico para o parâmetro empregado, conforme será descrito nas próximas subseções para cada classe fonética.

Todas as figuras apresentadas nas subseções a seguir foram geradas no Matlab®, usando palavras isoladas pronunciadas por um locutor masculino. Todas as locuções foram amostradas a 22,05 kHz e quantizadas com 16 bits por amostra. O Matlab® normaliza a amplitude do sinal de fala entre -1 e 1.

5.4.1. Refinamento do Silêncio

Esta classe representa o silêncio presente no início e no final de cada locução, e também as possíveis pausas entre as palavras. Dentre todas as classes presentes neste trabalho, o silêncio é a mais simples de ser refinada e também a mais simples de ser parametrizada.

O silêncio do início e do final de cada locução e também as pausas entre as palavras correspondem a um período em que não há nenhuma atividade relacionada à fala e, dessa forma, a energia nessa região é a menor em relação a outras regiões em que existe fala. Portanto, o melhor parâmetro para caracterizar o intervalo em que ocorre o silêncio é a energia por janela de análise. Nenhum outro parâmetro foi combinado com a energia para refinar as marcas de segmentação e, portanto, a regra para o refinamento é constituída apenas pela energia média por janela.

O limiar definido para a energia foi de -60 dB, ou seja, no intervalo de refinamento o centro da janela de análise que cruzar o valor -60 dB representa a fronteira de separação entre o silêncio e outro fone. Durante o refinamento duas situações são possíveis: transição do silêncio para qualquer classe fonética ou, de qualquer classe fonética para o silêncio. Na primeira situação é procurada a janela em que a energia ultrapassa o limiar de -60 dB e, na segunda situação, a janela em que a energia fica abaixo de -60 dB.

A Figura 5.7 mostra um exemplo do refinamento para o silêncio no início e no final da locução “fundamental”. As linhas vermelhas mostradas na Figura 5.7 representam o intervalo de refinamento definido para o cálculo dos parâmetros, a linha verde indica a fronteira em análise inicialmente definida pelo alinhamento forçado de Viterbi e a linha preta indica a posição da nova fronteira determinada de acordo com os parâmetros específicos de cada classe fonética.

Para o início da locução, o intervalo de refinamento compreende as amostras de 330 a 5720, e a fronteira determinada pelo alinhamento de Viterbi foi definida na amostra 660, conforme indica a linha verde na Figura 5.7 (a). A primeira linha preta na Figura 5.7 (c) indica o ponto correspondente ao limiar de -60 dB na curva de energia e também o número da janela de análise em que esse limiar é encontrado (janela 61). Como o processamento foi realizado usando janelas de análise com duração de 20 ms (441 amostras) e deslocamento de 1 ms (aproximadamente 22 amostras), a fronteira determinada corresponde à amostra 1345, ou seja, houve um deslocamento de 685 amostras para a direita em relação à fronteira inicialmente calculada, conforme indica a primeira linha preta na Figura 5.7 (a).

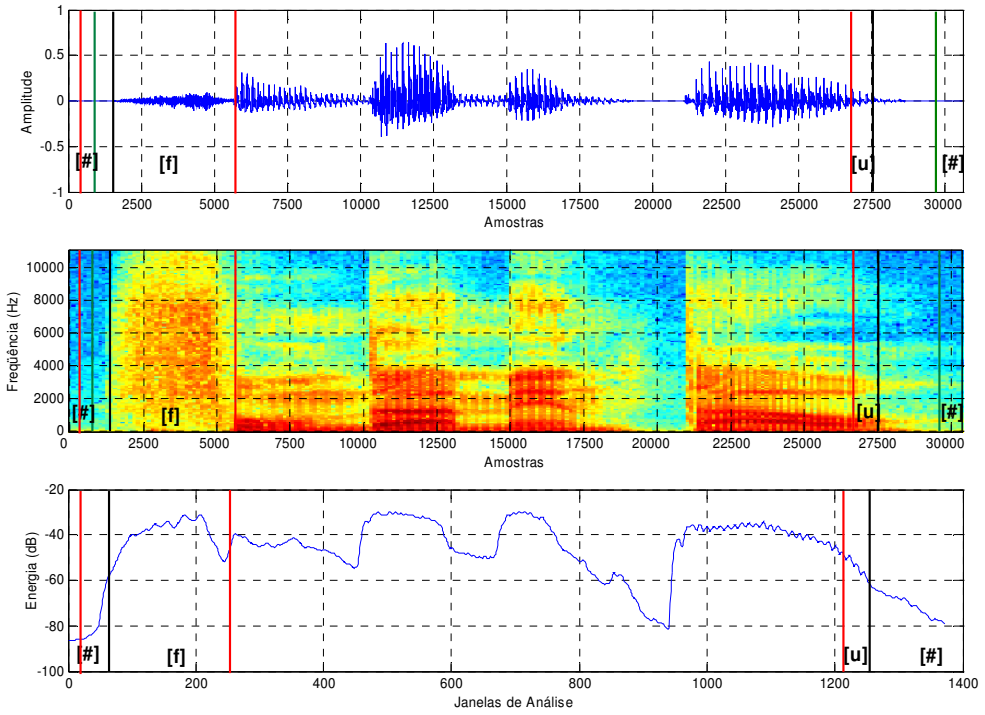


Figura 5.7: Locução “fundamental”: (a) Forma de onda. (b) Espectrograma. (c) Energia.

No final da locução, o intervalo de refinamento foi definido entra as amostras 27050 e 30630 (final da locução). A última fronteira determinada pelo alinhamento de Viterbi ocorre na amostra 29480 (última linha verde indicada na Figura 5.7 (a)). Após o refinamento, o limiar de energia é atingido no centro da janela de análise de número 1251, que corresponde à amostra 27782. Neste caso, houve um deslocamento de 1698 amostras para a esquerda.

5.4.2. Refinamento das Fricativas

As fronteiras entre as fricativas e as demais classes fonéticas, assim como o silêncio, são relativamente simples de serem caracterizadas porque necessitam de poucos parâmetros. Para as fricativas apenas dois parâmetros foram utilizados: a taxa de cruzamentos por zero e o centro de gravidade espectral.

As fricativas surdas apresentam um maior valor para a taxa de cruzamentos por zero em relação às fricativas sonoras. Para as fricativas surdas o limiar estabelecido foi de 0,52 e para as fricativas sonoras 0,28. As janelas de análise que apresentam valor para a taxa de cruzamentos por zero menor que os limiares estabelecidos não são classificadas como fricativas.

Além do valor da taxa de cruzamentos por zero, o centro de gravidade espectral também é combinado para obter uma fronteira com mais precisão. O limiar estabelecido para o centro de gravidade espectral foi de 2500 Hz. A fronteira entre uma fricativa e outra classe fonética é estabelecida na janela em que tanto o valor da taxa de cruzamentos por zero quanto o valor do centro de gravidade espectral tornam-se menores do que os limiares estabelecidos (considerando a transição entre uma fricativa e uma outra classe fonética) ou maiores caso a transição seja de qualquer classe fonética para uma fricativa.

A Figura 5.8 mostra o processo de refinamento entre a fricativa surda [s] e a vogal central [a] presentes na locução “sagas”.

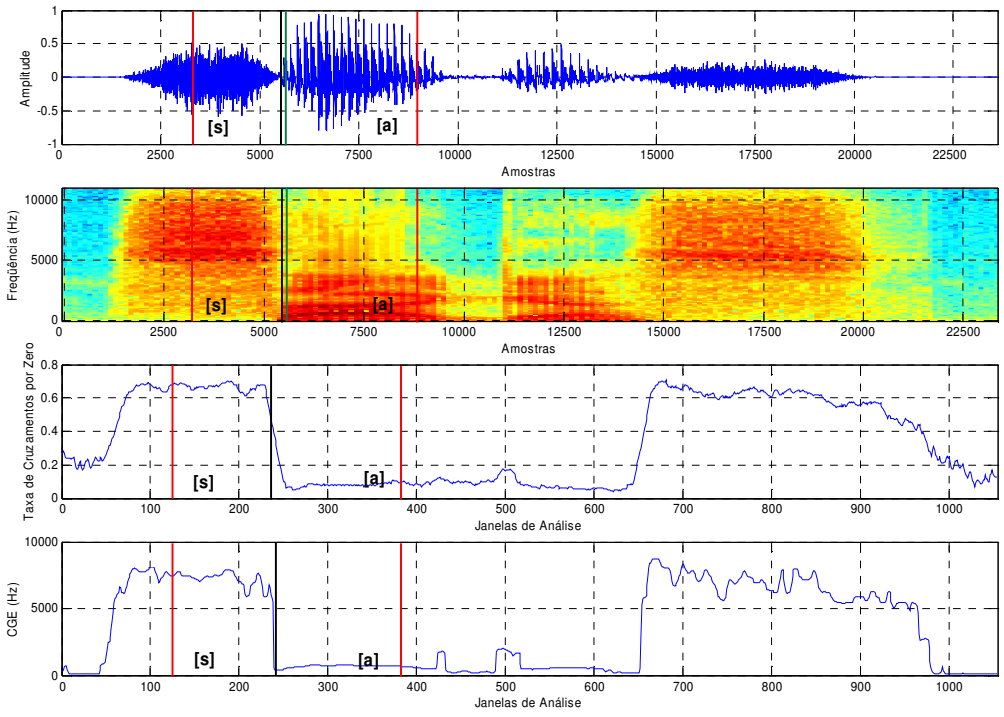


Figura 5.8: Locução “sagas”: (a) Forma de onda. (b) Espectrograma. (c) Taxa de cruzamentos por zero. (d) Centro de gravidade espectral (CGE).

Na Figura 5.8 (a), a região entre as linhas vermelhas compreende o intervalo de refinamento (2976 a 8800). A linha verde representa a fronteira inicial em análise, que por sua vez ocorre na amostra 5700. Na Figura 5.8 (c) é indicado o ponto em que a taxa de cruzamentos por zero cruza o limiar previamente estabelecido para as fricativas surdas (janela de análise de

número 235, que corresponde à amostra 5379). O limiar de 2500 Hz para o centro de gravidade espectral ocorre na janela de análise de número 239 (amostra 5467).

A transição entre os fones é definida na janela em que tanto a taxa de cruzamentos por zero quanto o centro de gravidade espectral tornam-se menores que os limiares estabelecidos. Isso ocorre na janela de número 239, que corresponde à amostra 5467, que é indicado pela linha preta nas Figuras 5.9 (a) e (b). Em relação à fronteira previamente determinada pelo alinhamento forçado de Viterbi, a nova fronteira é deslocada 233 amostras para a esquerda.

Na Figura 5.9 é repetida a mesma análise para a transição entre a fricativa sonora [z] e a vogal central [a], que ocorre na locução "casa".

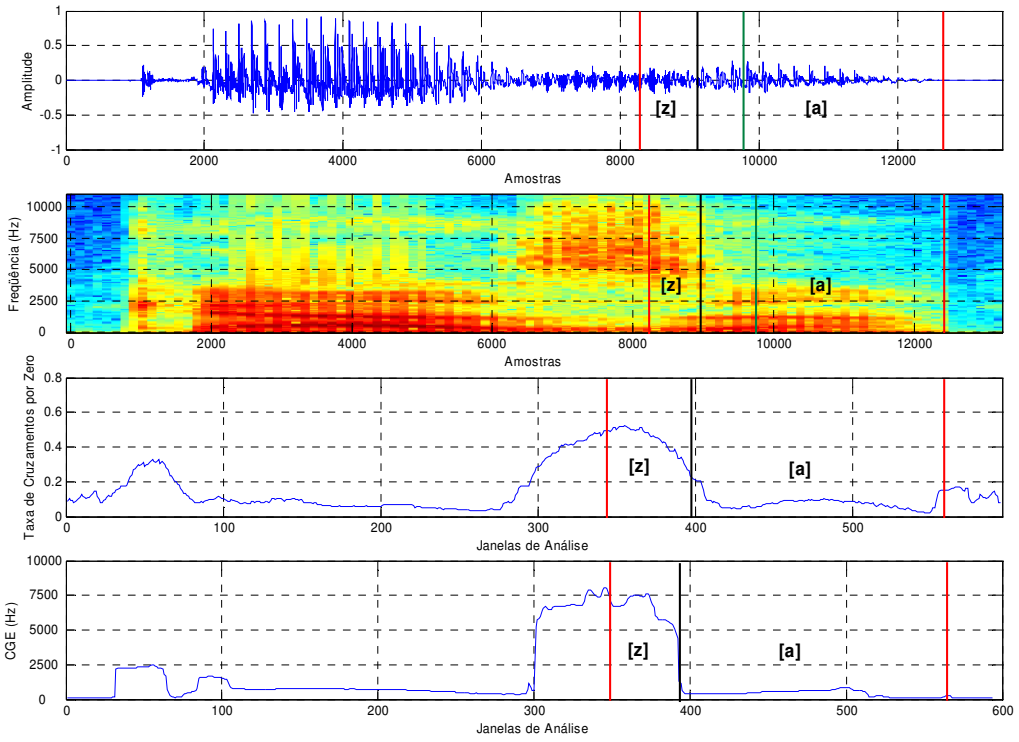


Figura 5.9: Locução "casa": (a) Forma de onda. (b) Espectrograma. (c) Taxa de cruzamentos por zero. (d) Centro de gravidade espectral.

O intervalo de refinamento foi estabelecido entre as amostras 8150 e 12540, e a fronteira inicialmente proposta ocorre na amostra 9900. Para o centro de gravidade espectral o limiar é atingido na janela de análise de número 393 (amostra 8863) e para a taxa de cruzamentos por

zero na janela de número 395 (amostra 8907). Como a fronteira é definida no centro da janela em que os dois limares são menores, o centro da janela de número 395 será definido como a nova fronteira (amostra 8907). A fronteira inicial foi determinada na amostra 9900. Após o refinamento a nova fronteira é deslocada para a amostra 8907, ou seja, 993 amostras para a esquerda.

5.4.3. Refinamento das Consoantes Laterais e Róticas

Como foi exposto no Capítulo 4, a transição entre as consoantes laterais e as demais classes fonéticas é marcada por uma leve variação de energia, uma vez que esses sons são muito parecidos com as vogais. Por outro lado, as consoantes róticas apresentam maior variação de energia. Como já sugerido e exemplificado, a variação da energia espectral em algumas bandas de freqüência representa um bom parâmetro para determinar a transição entre essas consoantes e outras classes de fones (normalmente as vogais no PB). Esse parâmetro tem como objetivo principal detectar os principais pontos de variação abrupta da energia, o que corresponde às possíveis transições entre os fones.

O limite de cada banda de freqüência foi determinado de forma a coincidir com a região de ocorrência das freqüências formantes (regiões caracterizadas por alta concentração de energia), conforme definido no Capítulo 4. Esses valores foram determinados através de testes, e a derivada da energia espectral foi calculada utilizando-se 3 janelas adjacentes de cada lado. A combinação da variação da energia espectral de todas as bandas é feita através da soma de seus valores a cada instante de tempo. Esse procedimento permite realçar os picos da variação de energia e dessa forma localizar as fronteiras corretamente, conforme exposto no Capítulo 4.

Na Figura 5.10 é mostrada a transição entre a consoante lateral [l] e a vogal central [a] da locução “chocolate” determinada durante a fase de refinamento.

O intervalo de refinamento foi definido entre as amostras 16140 e 20082, o que corresponde às janelas de análise de números 721 e 900. Como pode ser observado na Figura 5.10 (c), a transição entre a consoante [l] e a vogal [a] é marcada por um pico de variação de energia. Esse pico por sua vez apresenta amplitude relativamente baixa, o que pode ser explicado pela distribuição de energia da consoante [l] muito próxima das vogais. Conseqüentemente, a derivada produz uma transição mais suave. O pico ocorre na janela de análise 753, que corresponde à amostra 16889 na Figura 5.10 (a), indicado pela linha verde. A fronteira

inicialmente proposta ocorre na amostra 17380 (491 amostras deslocadas à direita em relação à fronteira correta).

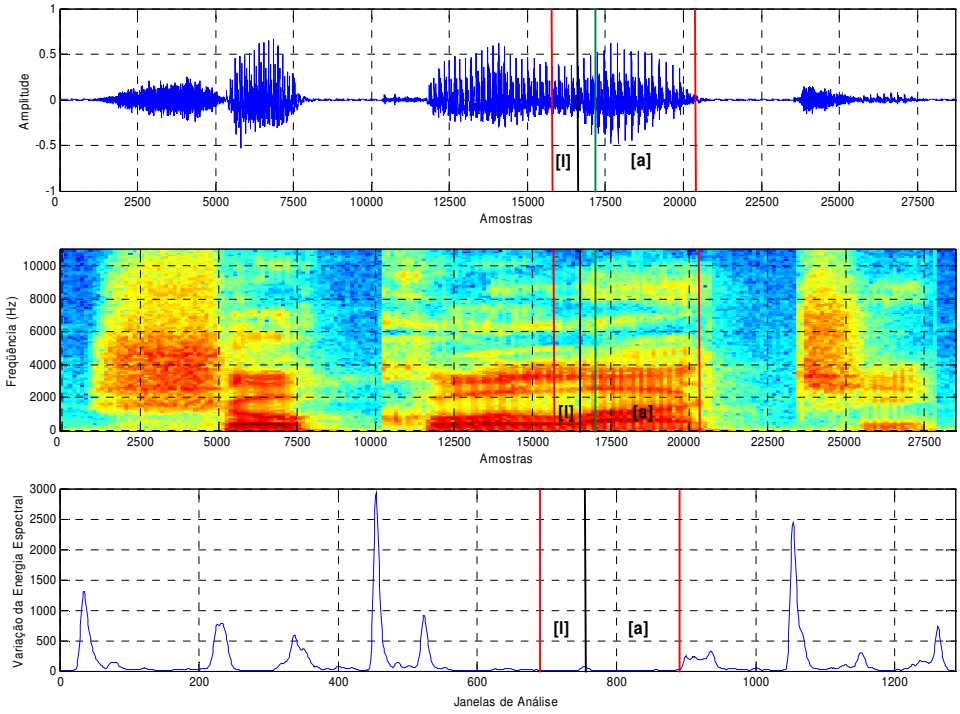


Figura 5.10: Locução “chocolate”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

Na Figura 5.11 é mostrada a transição entre a consoante lateral [L] e a vogal central [a] na locução “calha”. A transição da consoante [L] para a vogal é marcada por uma variação da energia espectral mais acentuada em relação à transição da consoante [l]. Conforme mostra a Figura 5.11 (c) o pico de transição ocorre na janela de análise 386, que corresponde à amostra 8709 na Figura 5.11 (a) (indicada pela linha preta).

A fronteira inicialmente determinada pelo alinhamento forçado de Viterbi ocorre na amostra 10000. Com o refinamento empregado, a fronteira é deslocada para uma nova posição (8709) que se encontra 1291 amostras à esquerda da fronteira inicialmente estimada.

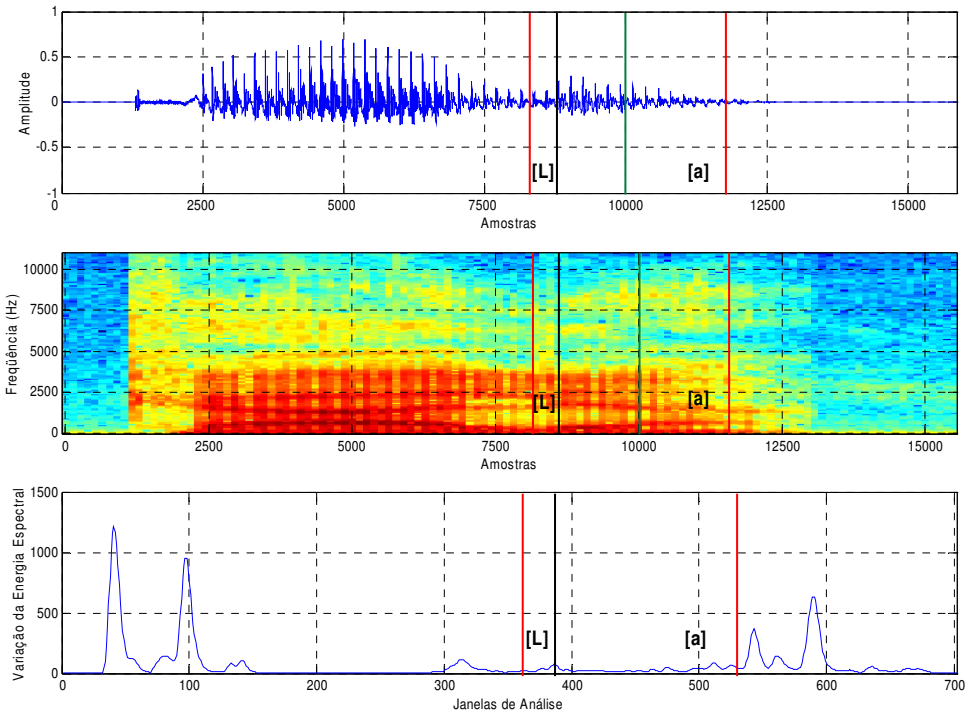


Figura 5.11: Locução “calha”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

A Figura 5.12 descreve a transição entre a consoante rótica [r] e a vogal anterior [i] na locução “gostaria”. Como pode ser comprovado em relação às consoantes laterais, a variação de energia para as consoantes róticas é mais acentuada. Isso pode ser observado pela amplitude do pico de variação da energia espectral descrito na Figura 5.12 (c).

O intervalo de refinamento compreende as amostras 16060 a 22220. A fronteira estimada pelo algoritmo de Viterbi ocorre na amostra 16940 (linha verde na Figura 5.12 (a)). A transição determinada pelo algoritmo de refinamento ocorre na janela de análise de número 745, que corresponde à amostra 16625. A fronteira em análise é deslocada para a nova posição 315 amostras à esquerda em relação à posição inicial.

A transição das classes fonéticas tanto para a consoante rótica [R], como para a lateral [l] também é marcada por uma variação suave da energia espectral (amplitude do pico de variação espectral é baixa), como mostrado na Figura 5.13.

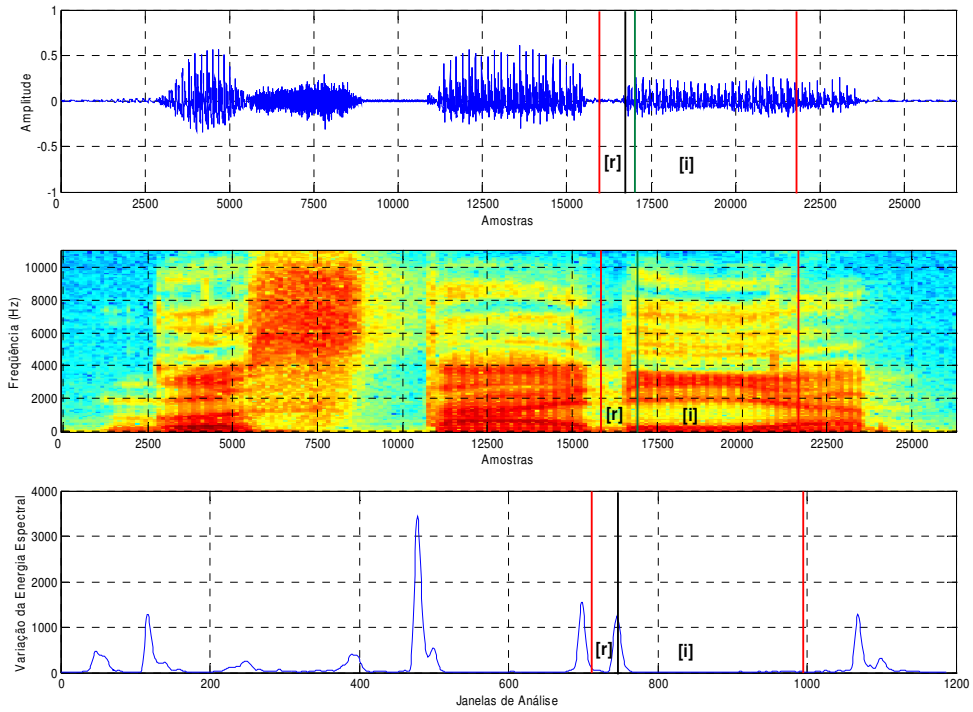


Figura 5.12: Locução “gostaria”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

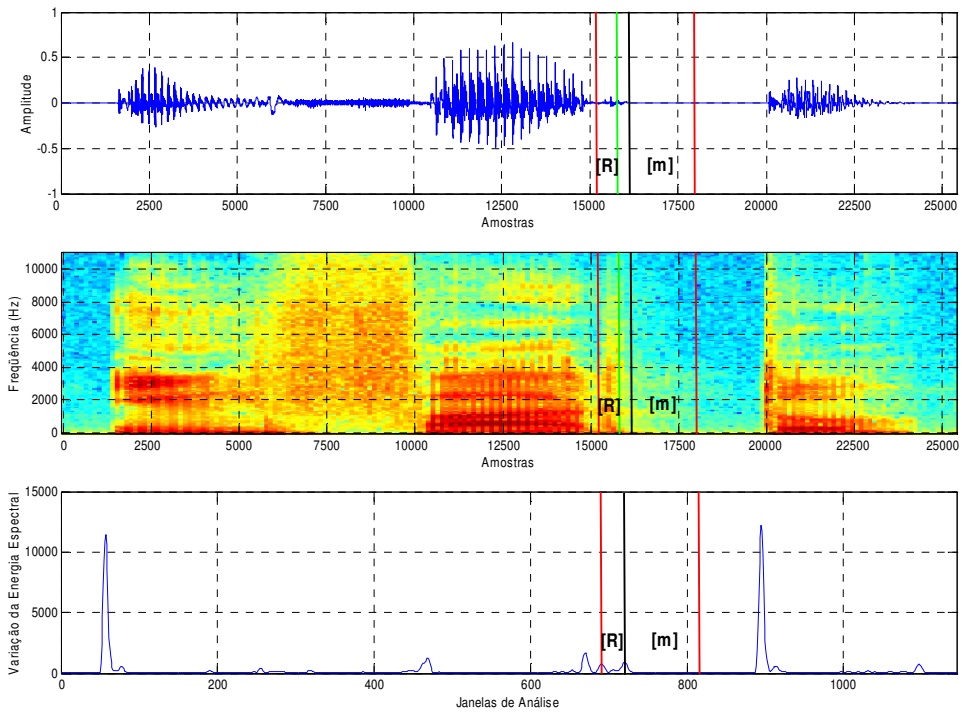


Figura 5.13: Locução “infarto”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

O intervalo de refinamento foi considerado entre as amostras 15322 e 18186, que por sua vez corresponde às janelas de análise de número 690 e 820. A transição ocorre na janela de análise de número 721, que corresponde à amostra 15986. Como inicialmente a fronteira foi determinada na amostra 15670, a nova posição se encontra 316 amostras à direita em relação à fronteira inicial.

Analisando as Figuras 5.11 a 5.14, pode-se concluir de forma geral que a variação de energia é realmente um bom parâmetro para a detecção das fronteiras entre as consoantes laterais e róticas e as demais classes fonéticas (normalmente as vogais no PB). Como a detecção é baseada em picos resultantes da derivada da energia, esse parâmetro é sensível ao intervalo de refinamento, ou seja, dependendo do intervalo determinado pode ocorrer a detecção de um pico que não corresponde à fronteira real em análise.

5.4.4. Refinamento das Consoantes Nasais

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

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

Como sugestão de Amit Juneja (Juneja, 2004), para a detecção das fronteiras foi utilizada a energia em duas bandas de freqüência: 0-358 Hz e 358-5378 Hz. A primeira banda está relacionada com a concentração da energia nas baixas freqüências, característica marcante das consoantes nasais. A segunda banda por sua vez está relacionada com a concentração de energia das vogais. A variação da energia em cada banda de freqüência é determinada e, em seguida, os valores de cada banda são somados de forma a destacar os pontos de variação.

A Figura 5.14 mostra o ponto de transição entre a consoante nasal [m] e a vogal central [a] da locução “amazonas”. A variação da energia espectral nas duas bandas de freqüência especificadas gera picos nas fronteiras entre os fones. Como pode ser observado na Figura 5.14 (c), a transição entre a consoante nasal [m] e a vogal central [a] ocorre na janela de análise de

número 329. O centro da janela 329 corresponde à amostra 7452 nas Figuras 5.15 (a) e (b). A fronteira inicial ocorre na amostra 7920. A nova posição para a fronteira foi deslocada 468 amostras para a esquerda.

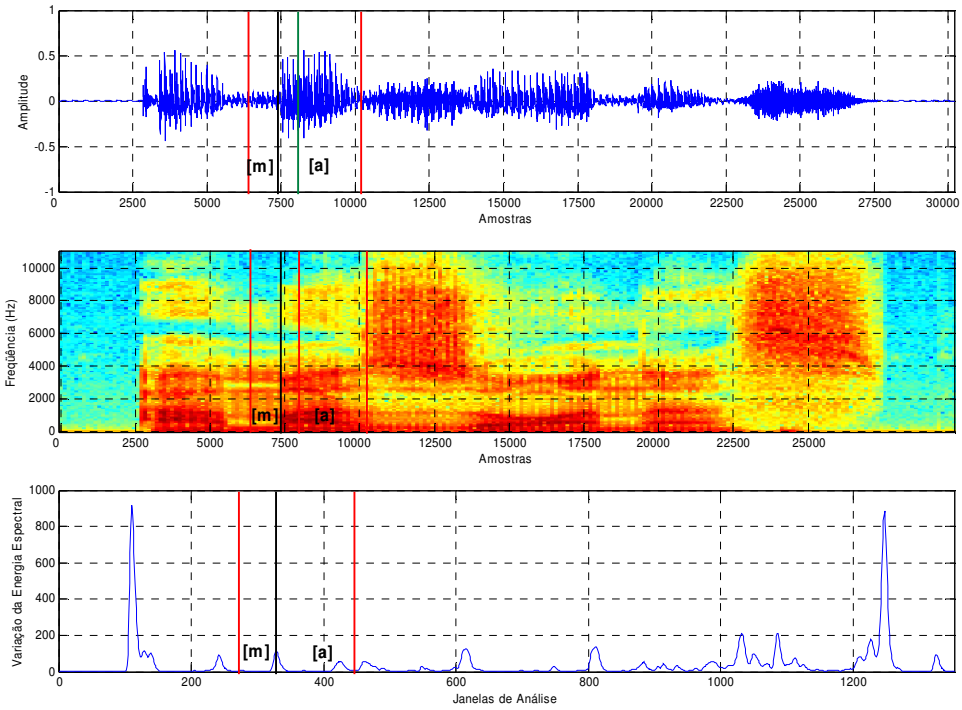


Figura 5.14: Locução “amazonas”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

As mesmas considerações podem ser feitas para a consoante nasal [n], como mostrado na Figura 5.15. Na Figura 5.15 (a) é mostrada a forma de onda para a locução “autonomia”, e na Figura 5.15 (b) é mostrado o espectrograma da locução.

O intervalo de refinamento tem início na amostra 13310 (janela de análise 593) e fim na amostra 16720 (janela de análise 641). Nesse intervalo, que compreende a transição entre a consoante nasal [n] e a vogal anterior [o], é detectado um pico de variação da energia espectral na janela de análise de número 641 (amostra 14332). A fronteira determinada pelo alinhamento forçado de Viterbi foi posicionada na amostra 14740, portanto a nova fronteira foi deslocada 408 amostras para a esquerda.

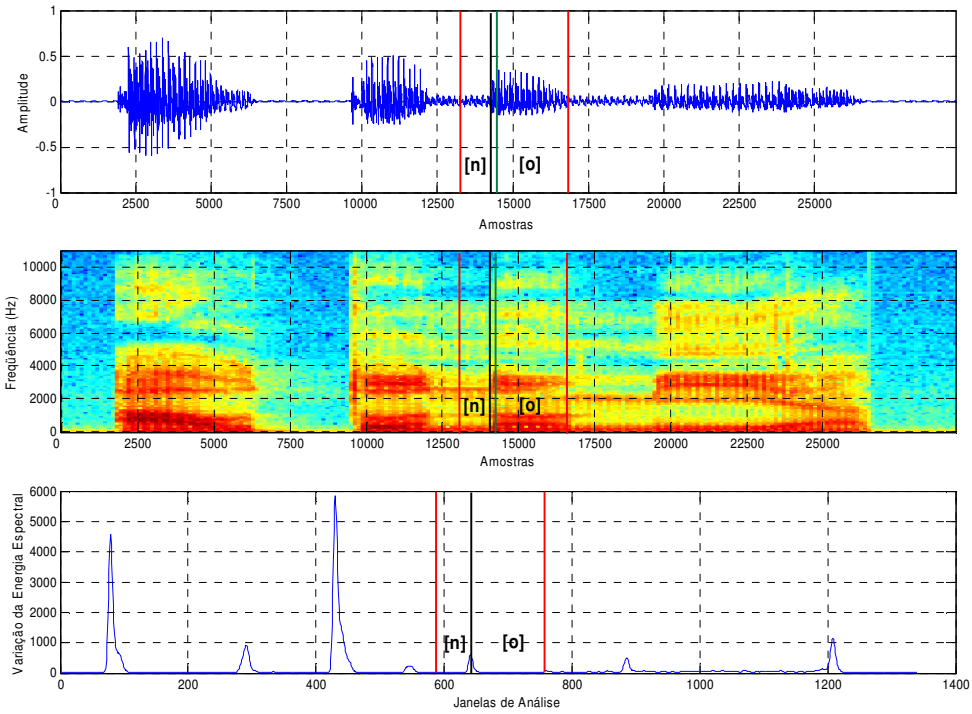


Figura 5.15: Locução “autonomia”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

Como última análise do refinamento das consoantes nasais, a Figura 5.16 mostra a transição entre a consoante nasal [N] e a vogal central [a].

Para a análise da locução “contenha”, o intervalo de refinamento compreende as amostras 14900 e 20010 (janelas de análise 665 a 897). A fronteira determinada pelo algoritmo de Viterbi encontra-se na amostra 18920, avançando sobre a vogal central [a]. O pico resultante da variação de energia espectral foi encontrado na janela de análise de número 756 (amostra 16867). Neste caso há um deslocamento de 2052 amostras para a esquerda.

Como pode ser comprovado na análise das Figuras 5.15, 5.16 e 5.17, a variação da energia espectral nas bandas [0-358 Hz] e [358-5378] Hz permite definir com precisão a transição entre as consoantes nasais e as vogais (únicos fones que seguem as consoantes nasais). Em todos os casos analisados, a variação da energia espectral foi suavizada para eliminar possíveis picos espúrios. Para a suavização foi considerado que cada ponto da variação de energia

é dado pela média dos três pontos imediatamente anteriores e dos três pontos imediatamente posteriores (incluindo o ponto central).

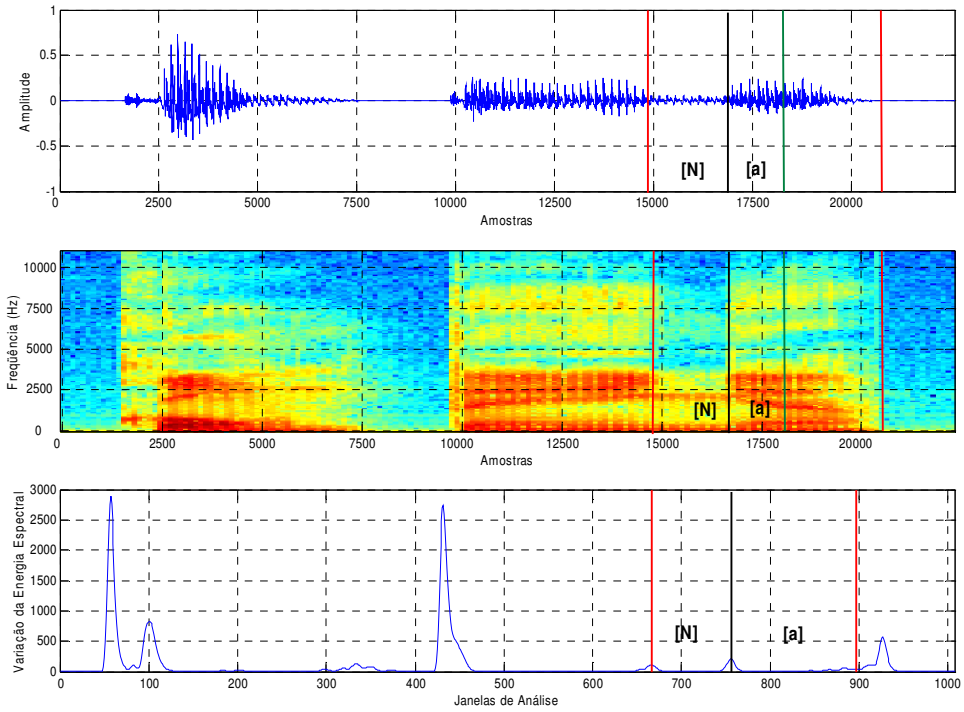


Figura 5.16: Locução “contenha”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

5.4.5. Refinamento das Plosivas

As consoantes plosivas são os fones mais difíceis para serem refinados. Como já foi observado no Capítulo 4, as plosivas são caracterizadas por um longo período de oclusão e por uma “explosão” que corresponde à liberação do ar. O período que constitui a explosão é muito rápido, o que dificulta o seu processamento e, conseqüentemente, a localização do instante em que ocorre a transição para o fone seguinte.

Para poder refinar corretamente as plosivas, neste trabalho esses sons foram caracterizados como dois fones diferentes, ou seja, existe uma representação para o período de oclusão, que antecede a explosão, e uma representação para a explosão propriamente dita. As características do período de oclusão dependem da consoante plosiva que está sendo gerada. Por exemplo, se a plosiva é surda, o período de oclusão é caracterizado por não existir nenhuma

vibração das pregas vocais e, conseqüentemente, baixa energia espectral. Caso a plosiva seja sonora, existe uma pequena concentração de energia nas baixas freqüências antes da explosão.

Durante alguns testes de refinamento com as plosivas foi observado que o intervalo de refinamento como estabelecido para as outras classes fonéticas não era apropriado, pois o intervalo abrangia o período de oclusão da plosiva, dificultando o refinamento entre a plosiva e as outras classes de fones. Dessa forma, o primeiro passo para o refinamento consiste em determinar o início da explosão que será definido como o início do intervalo de refinamento. O final do intervalo de refinamento será representado pela marca de segmentação posterior à marca que está sendo analisada.

Tendo em vista que as plosivas são tratadas como dois fones independentes durante a fase de refinamento, a transição entre qualquer classe fonética (com exceção do silêncio inicial) e uma plosiva ocorre em uma região do espectro caracterizada por baixa energia espectral (silêncio que antecede a liberação do ar das plosivas). O ponto de transição é marcado então por uma queda abrupta da energia espectral.

Para uma plosiva surda que ocorre no início da locução, o seu período de oclusão é mascarado pelo silêncio inicial da locução e, portanto, a transição entre o silêncio inicial e a plosiva ocorre no início da explosão. Se a plosiva for sonora, o período de oclusão é caracterizado por uma determinada quantidade de energia espectral nas baixas freqüências e, portanto, a transição entre o silêncio inicial e a plosiva ocorre quando a energia espectral ultrapassa um determinado limiar previamente determinado. A Figura 5.17 mostra a transição entre o silêncio inicial e uma consoante plosiva surda.

A transição entre o silêncio inicial da locução e a plosiva [p] é definida na janela de análise de número 48, em que a energia total se torna maior que -60 dB (mesma regra já definida para a transição entre o silêncio inicial das locuções e as demais classes fonéticas). A janela de número 48 corresponde à amostra 1256. Nesta regra de transição também pode ser utilizada a derivada primeira da energia, que gera um pico na transição entre os fones devido à mudança abrupta na energia.

A fronteira inicial foi definida na amostra 1100 havendo, portanto, um deslocamento de 156 amostras para a direita. Nota-se pela análise da Figura 5.17 que a nova fronteira é definida na região próxima à explosão.

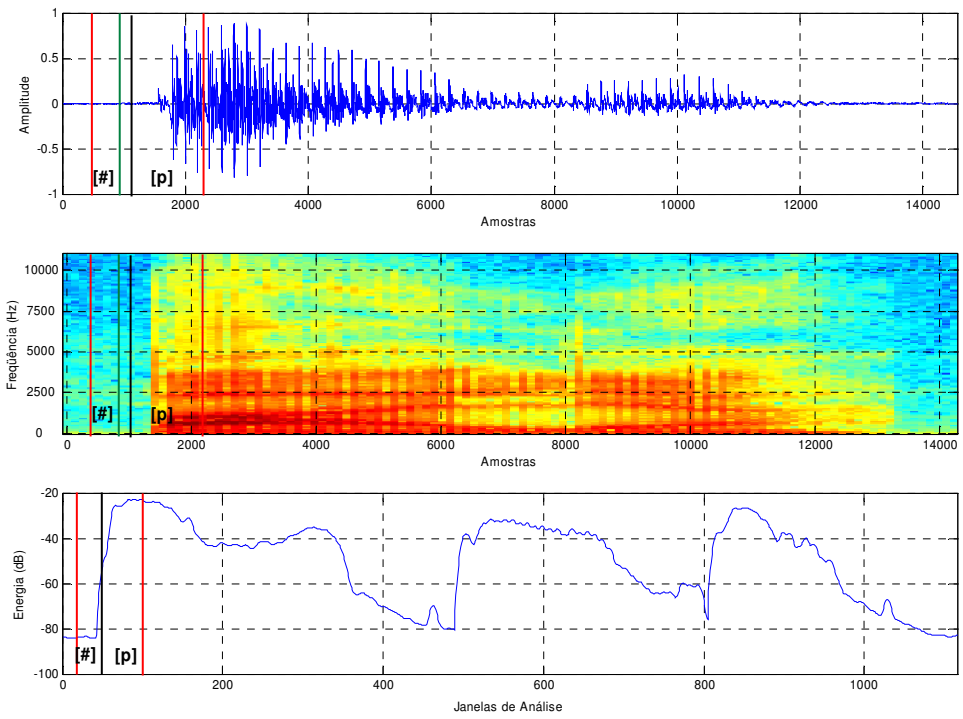


Figura 5.17: Locução “palha”: (a) Forma de onda. (b) Espectrograma. (c) Energia.

Para as plosivas sonoras no início das locuções, o mesmo limiar de energia usado para as plosivas surdas não é suficiente para detectar a transição do silêncio inicial para a plosiva sonora. É necessário definir um limiar de energia que possa detectar a transição do silêncio inicial para o período de vozeamento que antecede a explosão da plosiva sonora. O limiar de energia definido foi -70 dB, ou seja, o centro da janela de análise em que a energia se tornar maior que -70 dB é definida como a amostra de transição entre os fones.

Uma importante observação para as plosivas sonoras que ocorrem no início da locução é que a derivada da energia não representa um bom parâmetro para detectar a transição, tendo em vista que o período de vozeamento apresenta baixa energia e, portanto, a transição apresenta picos com pequena amplitude.

Na Figura 5.18 é mostrada a transição entre o silêncio inicial da locução e a plosiva sonora [d]. Na Figura 5.18 (c) percebe-se claramente a leve variação de energia entre o silêncio inicial e o período de vozeamento que antecede a plosiva. Na análise detalhada da Figura 5.18 (a)

nota-se que a linha preta define a transição na posição em que há alguma alteração na forma de onda do sinal (região definida por baixa energia espectral).

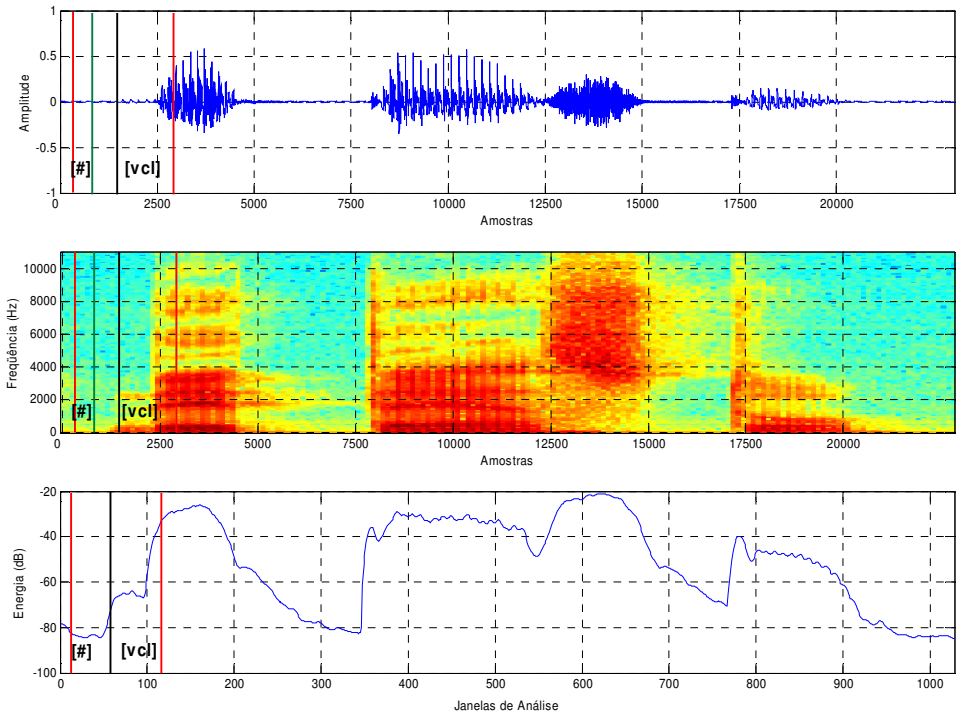


Figura 5.18: Locução “detesto”: (a) Forma de onda. (b) Espectrograma. (c) Energia.

As observações feitas nas Figuras 5.18 e 5.19 ocorrem quando a consoante plosiva está no início da locução. Quando estas ocorrem em outras posições na locução, a transição para o período de oclusão (seja ele surdo ou sonoro) é marcada por uma variação abrupta da energia espectral, que é detectada através de um pico na derivada da energia. Um exemplo da transição entre uma vogal e uma plosiva surda é mostrado na Figura 5.19.

O intervalo de refinamento utilizado para determinar a fronteira entre a vogal anterior [i] e o intervalo de oclusão [cl] que antecede a plosiva [t] teve início na amostra 7250 e fim na amostra 11880 da locução “irritante”. Este intervalo composto por 4630 amostras corresponde a parte da vogal [i] e parte do intervalo de oclusão [cl]. O pico de variação da energia foi determinado na janela de análise 354, que por sua vez representa a amostra 8003, como pode ser comprovado pela análise da Figura 5.19. A fronteira é definida em uma região do espectro de freqüências em

que ocorre uma queda brusca da energia espectral, representada pelo pico da variação da energia espectral.

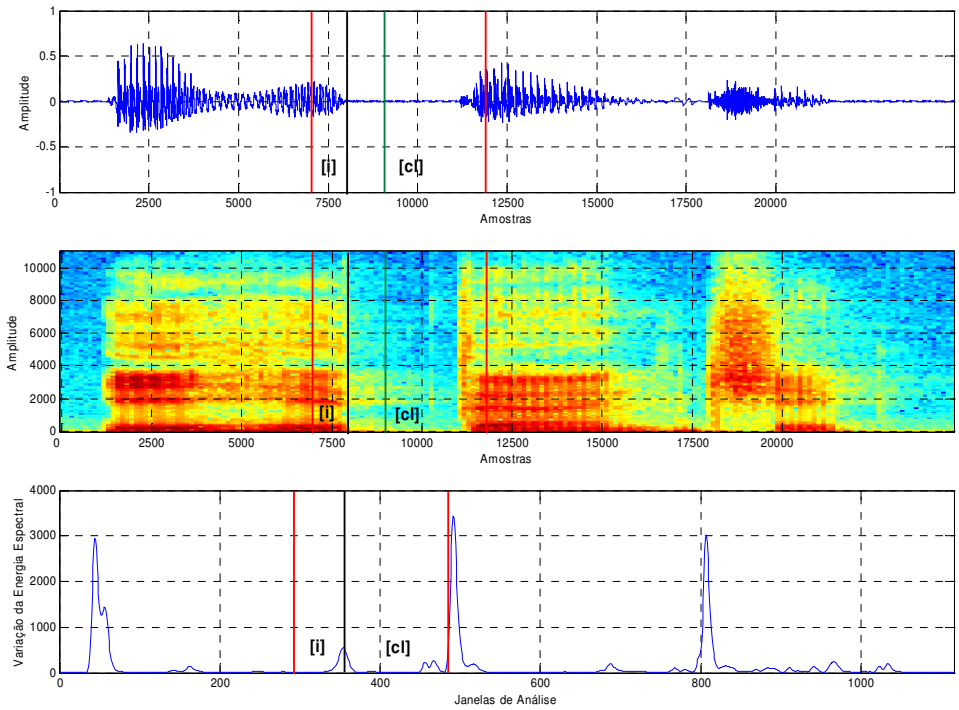


Figura 5.19: Locução “irritante”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

O mesmo comportamento verificado na transição para uma plosiva surda ocorre para a plosiva sonora, conforme mostra a Figura 5.20. Para exemplificar esse tipo de transição foi analisada a transição entre a vogal anterior [o] e o período de constrição sonoro [vcl] da locução “poderosa”. O intervalo de refinamento teve início na amostra 3720 (centro da vogal [i]) e fim na amostra 7040 (início da explosão para a plosiva sonora [d]). Como mostrado na Figura 5.20 (c), o pico de variação da energia ocorre na janela de análise 228, cujo centro corresponde à amostra 5225.

Para a determinação da transição entre as consoantes plosivas e as outras classes fonéticas, duas etapas são empregadas. A primeira etapa consiste em determinar o início da explosão, e a segunda consiste na localização da transição propriamente dita. Como a regra de refinamento das consoantes plosivas leva em consideração a variação da energia espectral e também, devido à presença do silêncio que antecede a liberação do ar, a variação da energia gera

um pico com grande amplitude que corresponde ao instante da liberação do ar (início da explosão) e não a transição para o fone adjacente. Após a detecção do início da explosão, um segundo pico (com a segunda maior amplitude) é procurado. Esse segundo pico corresponde à transição entre a consoante plosiva em análise e o fone seguinte.

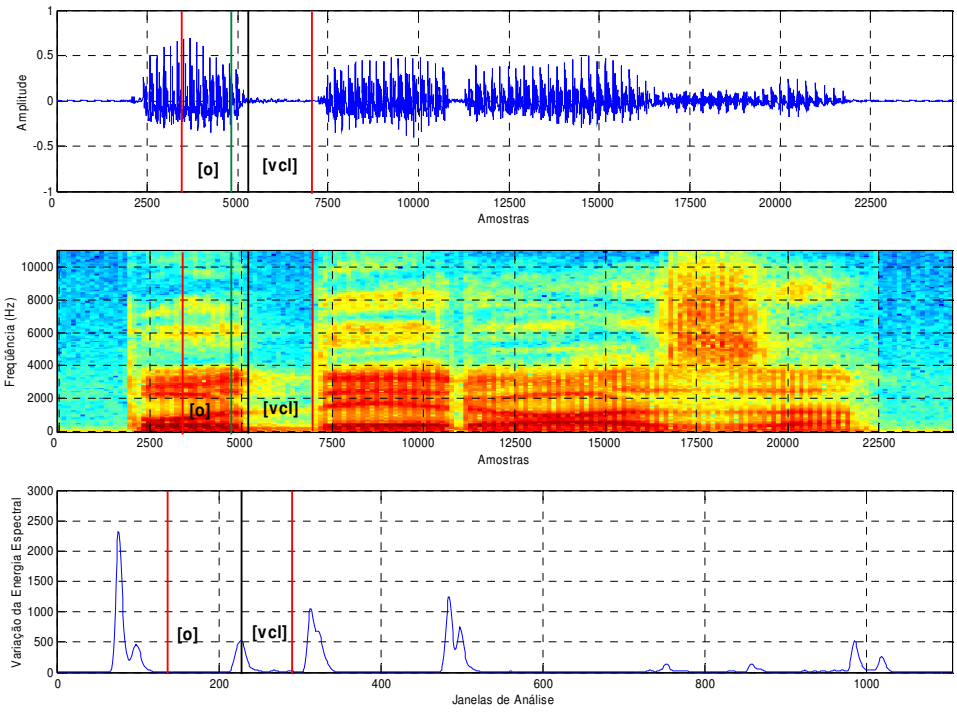


Figura 5.20: Locução “poderosa”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

Outra diferença adotada para o refinamento das consoantes plosivas é a alteração no tamanho da janela de análise. Ao invés de utilizar janelas de análise com duração de 20 ms, foram utilizadas janelas com duração de 10 ms para detectar com maior precisão a descontinuidade no sinal de fala e, dessa forma, estabelecer a fronteira entre a consoante plosiva e o fone seguinte. O intervalo de deslocamento para o cálculo dos parâmetros acústicos não foi alterado e continua em 1 ms. Essa alteração fez-se necessária uma vez que a duração entre a liberação do ar e a transição para o fone seguinte é muito curta.

A estratégia utilizada para a detecção das consoantes plosivas leva em consideração a variação da energia espectral em duas bandas de freqüência. A variação da energia nas duas bandas é calculada e, em seguida, seus valores são somados destacando os instantes de transição entre os fones. A Figura 5.21 exemplifica o processo de detecção da fronteira entre a plosiva surda [k] e a vogal central [a].

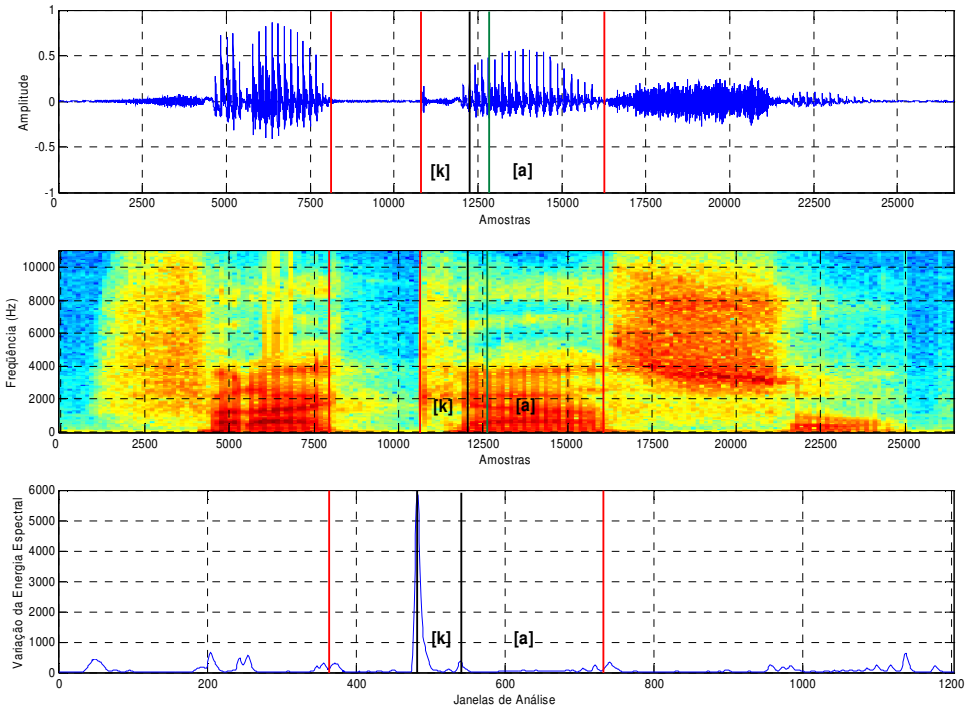


Figura 5.21: Locução “fracasso”: (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

O intervalo de refinamento utilizado na Figura 5.21 teve início na amostra 7250 e fim na amostra 16100. O primeiro pico gerado na Figura 5.21 (c) corresponde à transição entre o silêncio e o início da plosiva surda [k]. Essa transição ocorre na janela de análise 482, que corresponde à amostra 10826. O segundo maior pico, mas com amplitude menor que o primeiro, corresponde ao instante de transição entre a plosiva surda [k] e a vogal central [a]. A transição foi detectada na janela de análise 540 (amostra 12104).

As mesmas observações feitas para as plosivas surdas podem ser feitas para as plosivas sonoras, como mostrado na Figura 5.22 para a locução “corredores”.

Na Figura 5.22 é mostrada a transição entre a plosiva sonora [d] e a vogal posterior [o]. A transição das plosivas sonoras é mais suave do que as transições das plosivas surdas, como pode ser comprovado analisando as Figuras 5.22 e 5.23. A nova fronteira foi determinada no centro da janela de análise de número 548, que corresponde à amostra 12281. A posição da nova fronteira está 479 amostras à esquerda da fronteira inicialmente determinada pelo alinhamento forçado de Viterbi.

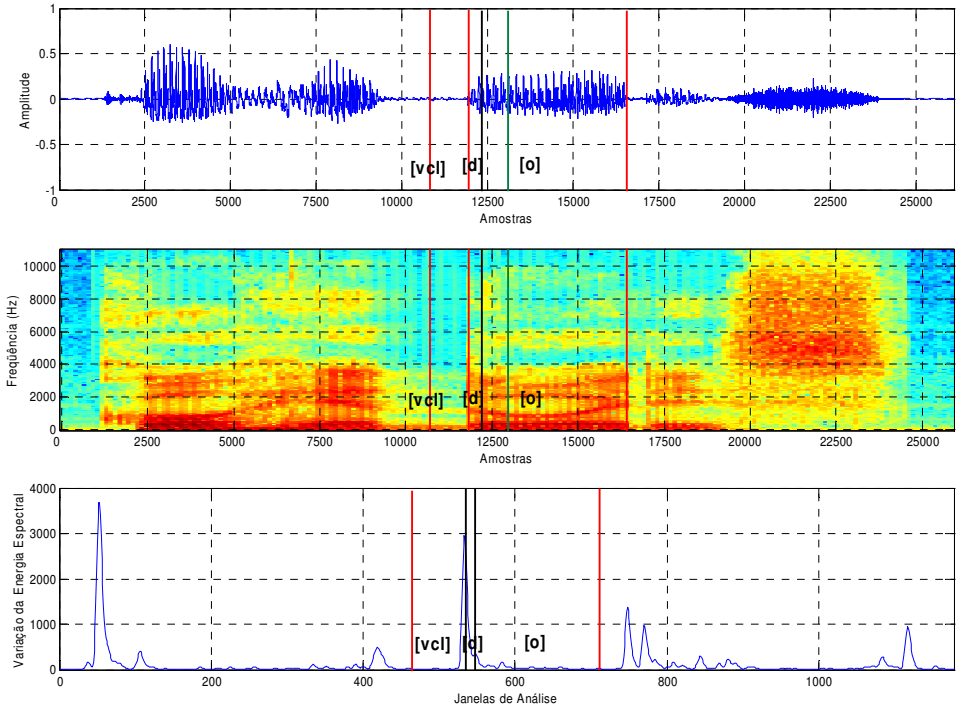


Figura 5.22: Locução "corredores": (a) Forma de onda. (b) Espectrograma. (c) Variação da energia espectral.

As plosivas surdas apresentam picos mais acentuados na transição com as vogais em relação às plosivas sonoras, que apresentam características acústicas próximas das vogais.

5.4.6. Refinamento das Africadas

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

As africadas podem ocorrer em qualquer posição dentro de uma locução, mas como já observado, no PB são sempre seguidas pela vogal anterior [i]. Acusticamente, as africadas são

caracterizadas por um período de baixa energia espectral nas baixas freqüências (período de constrição), seguida por uma região que apresenta uma alta taxa de cruzamentos por zero e também centro de gravidade espectral abaixo de 2 kHz. A combinação dessas características acústicas é suficiente para determinar com precisão a fronteira entre uma consoante africada e a vogal [i].

A transição entre qualquer classe fonética e uma consoante africada é definida na região em que ocorre uma queda abrupta de energia. Essa queda abrupta é determinada através da derivada primeira da energia. Essa grande variação de energia espectral ocorre devido ao período de constrição da africada (caracterizado por baixa energia espectral).

O refinamento (transição entre a africada e a vogal [i]) ocorre em duas etapas. Na primeira é determinado o instante em que ocorre a liberação do ar e, na segunda, a transição propriamente dita. O intervalo de refinamento para essa classe fonética segue o mesmo padrão das outras classes. Na Figura 5.23 é mostrado o resultado da aplicação das regras de refinamento para a consoante africada sonora [D] na locução "ódio".

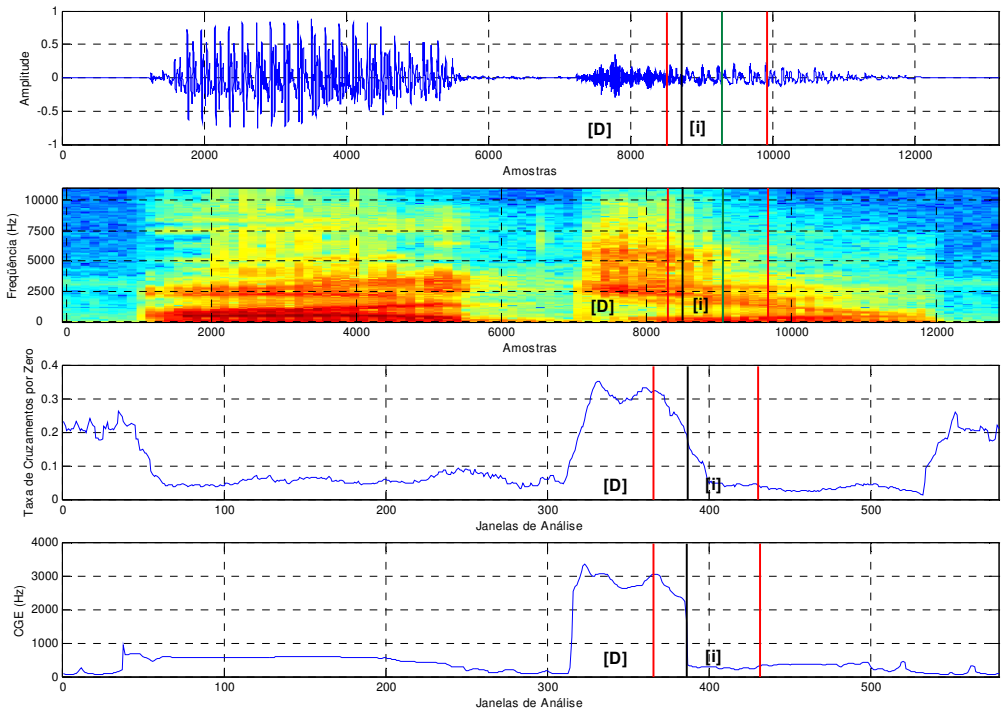


Figura 5.23: Locução "ódio": (a) Forma de onda. (b) Espectrograma. (c) Taxa de cruzamentos por zero. (d) Centro de gravidade espectral.

Para determinar a fronteira entre a consoante [D] e a vogal [i] na locução "ódio", primeiro foi estabelecido o intervalo de refinamento tendo início na amostra 8373 e fim na amostra 9900. Seguindo o padrão definido para as plosivas, foi determinado o instante em que ocorre a liberação do ar. Esse instante foi calculado usando a variação da energia espectral e foi definido na amostra 7287.

Tendo o início da liberação do ar da consoante [D], novamente é calculado o início do intervalo de refinamento para o cálculo dos parâmetros (ponto médio entre o início da liberação do ar e a fronteira estabelecida pelo algoritmo de Viterbi). Nesse intervalo foram calculadas a taxa de cruzamentos por zero e o centro de gravidade espectral, cujas variações são indicadas nas Figuras 5.24 (b) e 5.24 (c) respectivamente. A fronteira entre a consoante africada e a vogal foi definida na janela de análise em que a taxa de cruzamentos por zero é menor que 0,2 e que o centro de gravidade é menor que 2000 Hz. Esses limiares são cruzados na janela de análise de número 386, que corresponde à amostra 8705 (indicado nas Figuras 5.24 (a) e (b) pela linha preta).

É importante destacar que, apesar de toda semelhança com as fricativas, os limiares para as africadas (tanto sonora quanto surda) tiveram seus valores diminuídos. As fronteiras para a consoante [T] são mostradas na Figura 5.24.

Assim como as fricativas, as regras de refinamento definidas para as consoantes africadas produzem bons resultados quando comparados com a segmentação manual.

5.4.7. Refinamento das Vogais e Vogais Nasais

Como já discutido anteriormente, as vogais representam o coração das sílabas no PB e, portanto, podem estar "ligadas" a todas as outras classes de fones. É muito comum no PB a ocorrência de ditongos que, juntamente com as plosivas, representam as classes mais difíceis de serem refinadas. Nesta subseção será apenas considerado o refinamento entre as vogais quando ocorrem em ditongos, uma vez que a transição entre as vogais e as demais classes fonéticas já foi considerada anteriormente.

O primeiro tipo de transição a ser considerado é a ocorrência da vogal central [a] com as vogais anteriores ([e], [E], [i] e [y]) ou com as vogais posteriores ([o], [O], [u]). Analisando o triângulo das vogais na Figura 4.9, nota-se que o valor do primeiro e do segundo formante estão em faixas diferentes para as três classes de fone, o que é útil para o refinamento.

Para a determinação dos formantes em cada janela de análise foi realizada uma análise LPC com ordem 14. Os coeficientes do preditor linear foram determinados aplicando o algoritmo de Levinson-Durbin (Rabiner and Schafer, 1978). Tendo os coeficientes do preditor, a DFT com 1024 pontos é empregada para calcular a resposta em freqüência do filtro de síntese. Os picos resultantes da análise LPC representam as freqüências formantes. A análise LPC foi escolhida neste trabalho pela sua simplicidade em relação a outros métodos e também pelo baixo custo computacional. Nenhum outro algoritmo foi testado.

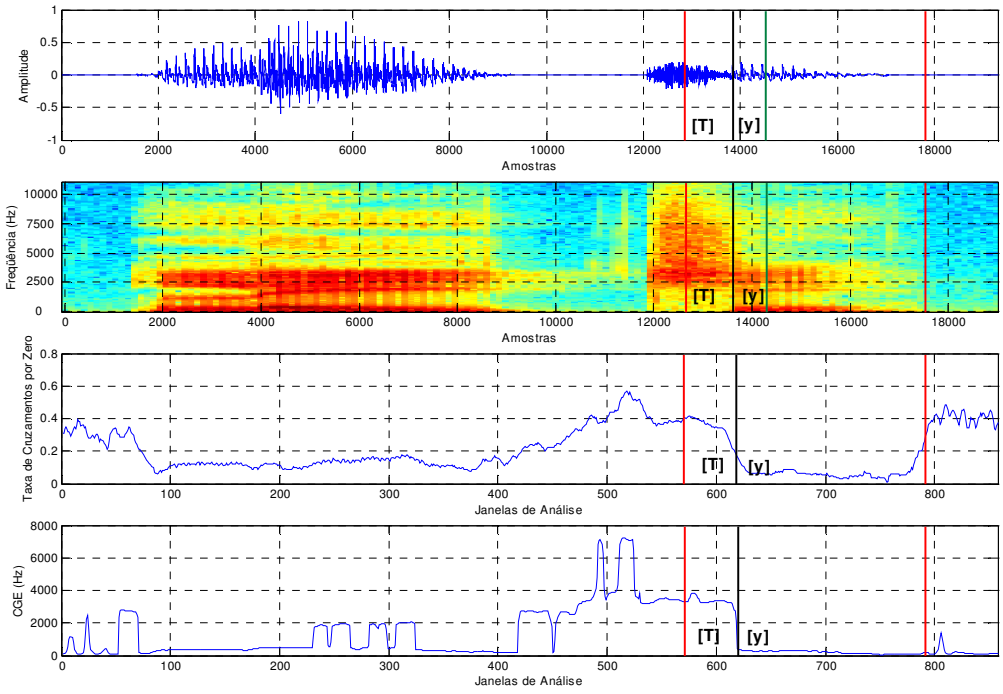


Figura 5.24: Locução “leite”: (a) Forma de onda. (b) Espectrograma. c) Taxa de cruzamentos por zero. (d) Centro de gravidade espectral.

Através de testes foram determinados os limiares para o primeiro e o segundo formantes para as vogais. Para as vogais anteriores e posteriores, o valor do primeiro formante é menor que 450 Hz e, para a vogal central, acima de 450 Hz. Para o segundo formante cada classe apresenta valores diferentes. O valor do segundo formante para as vogais anteriores está acima de 1845 Hz e, para as vogais posteriores, abaixo de 1135 Hz. Já a vogal central apresenta valores intermediários (maior que 1135 Hz e menor que 1845 Hz).

Na Figura 5.25 é mostrado o resultado do refinamento para o hiato fonético /i a/ da locução “autonomia”, onde tem-se a transição entre da vogal anterior [i] para a vogal central [a].

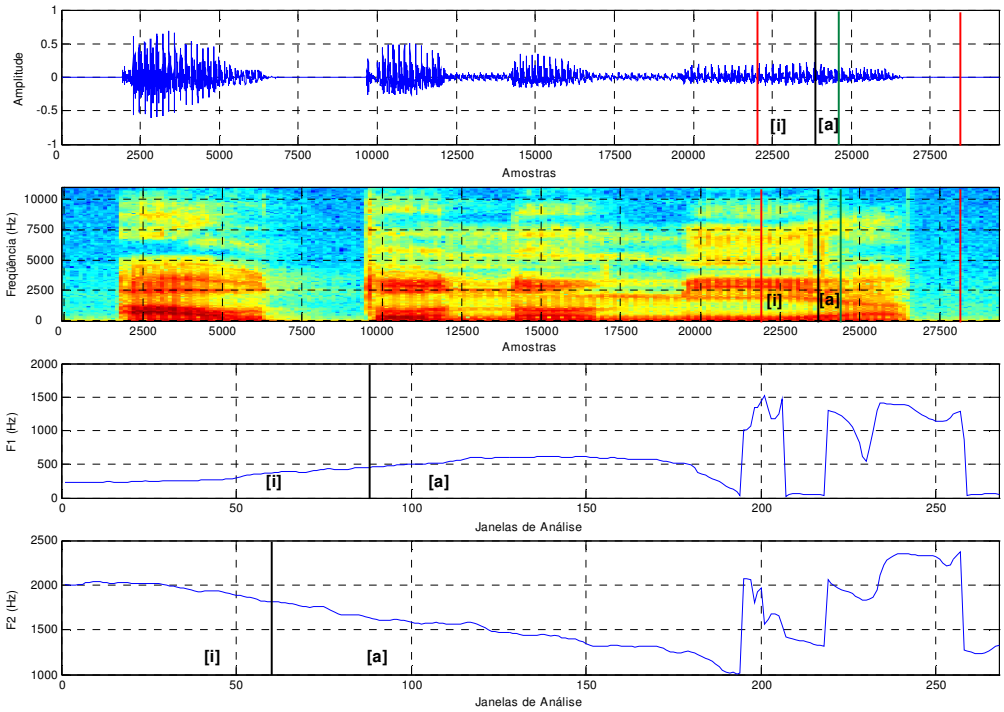


Figura 5.25: Locução “autonomia”: (a) Forma de onda. (b) Espectrograma. (c) Variação de F1 no intervalo de refinamento [22280-28600]. (d) Variação de F2 no intervalo de refinamento [22280-28600].

O intervalo de refinamento para o exemplo foi estabelecido entre as amostras 22280 e 26860. Como a vogal anterior [i] apresenta valor de F1 menor que a vogal central [a], mas valor de F2 maior, ao percorrer o sinal em direção à vogal [a] o limiar de F1 é detectado na janela de análise de número 86 e o limiar de F2 na janela de número 56. Como o intervalo de refinamento teve início na amostra 22280, o centro da janela de análise de número 86 corresponde à amostra 24374 e o centro da janela 56 à amostra 23712. A nova posição da fronteira é definida na amostra em que os dois parâmetros ultrapassam os limiares definidos, portanto na amostra 24374 (486 amostras à esquerda em relação à fronteira inicial).

Comportamento semelhante ao apresentado na Figura 5.25 também pode ser conferido na Figura 5.26, onde é apresentado um exemplo do refinamento entre a vogal central [a] e a

semivogal posterior [u]. O intervalo de refinamento utilizado está compreendido entre as amostras 2600 e 9703 da locução “Áustria”.

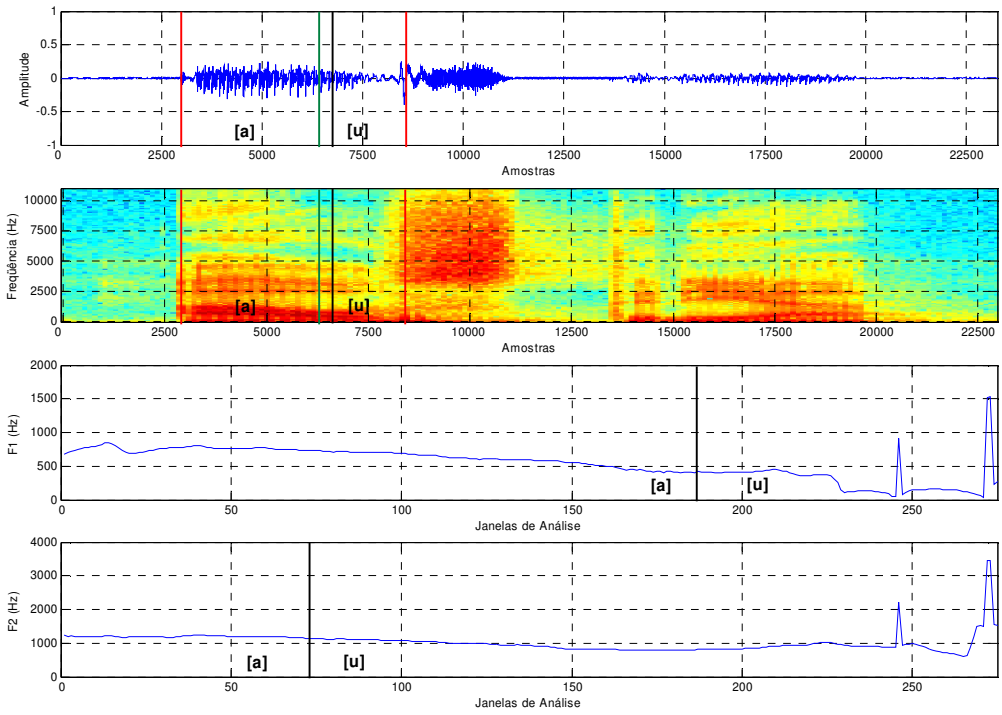


Figura 5.26: Locução “Austria”: (a) Forma de onda. (b) Espectrograma. (b) Variação de F1 no intervalo de refinamento [2600-9703]. (d) Variação de F2 no intervalo de refinamento [2600-9703].

Como o sinal é percorrido partindo-se da vogal [a] em direção à semivogal [u], os formantes tendem a diminuir à medida que a vogal [a] termina e começa a vogal [u]. O limiar para o primeiro formante é detectado no centro da janela de análise de número 196 (7031) e, para o segundo formante, no centro da janela 104 (amostra 4451). A fronteira foi definida na amostra 7031 (431 amostras à direita em relação à fronteira inicial).

O segundo tipo de transição em um ditongo ocorre quando se tem uma vogal anterior e uma vogal posterior ou vice-versa. Neste caso dois parâmetros foram utilizados: a variação do segundo formante e o perfil energia, como discutido no Capítulo 4. Para o perfil energia foi utilizada a taxa de 75% para o PB e 70% para a TIMIT, que é suficiente para discriminar entre as duas classes de vogais. O limiar foi estabelecido em 1550 Hz.

O valor do segundo formante tende a ser menor para as vogais posteriores em relação às vogais anteriores, como já foi discutido. O limiar para o segundo formante foi estabelecido em 1490 Hz neste tipo de transição, tendo como objetivo detectar a fronteira no instante da transição entre os fones.

Um exemplo da transição entre vogal anterior e posterior é mostrado na Figura 5.27 entre a vogal posterior [o] e a vogal anterior [e] da locução "coelho". O intervalo de refinamento teve início na amostra 2900 e fim na amostra 10780.

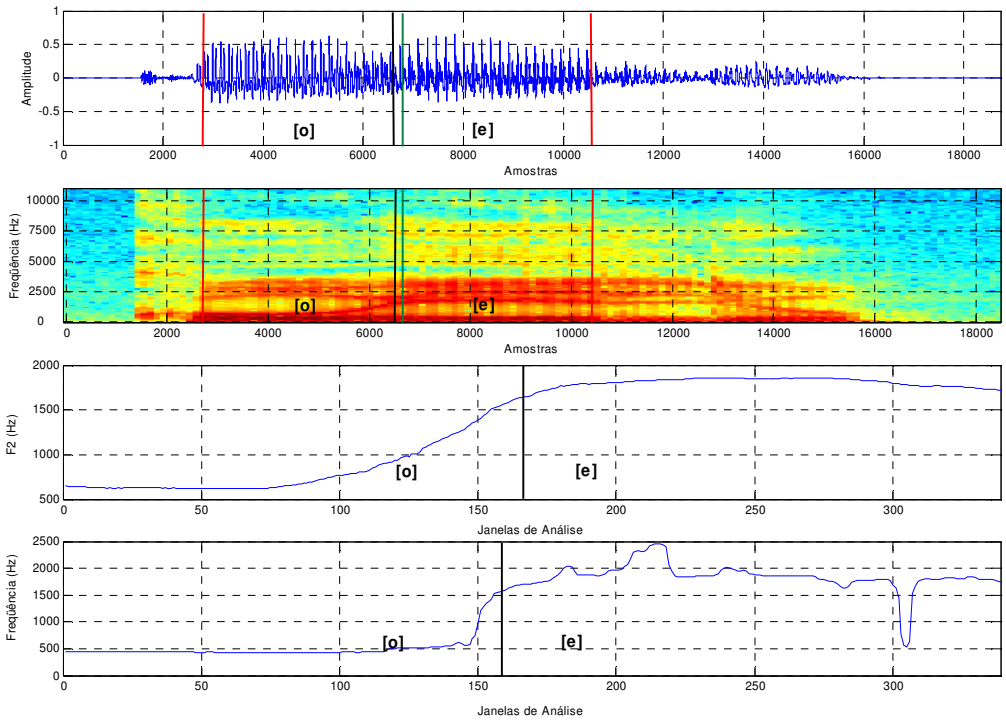


Figura 5.27: Locução "coelho": (a) Forma de onda. (b) Espectrograma. (c) Variação de F2 no intervalo de refinamento [2900-10780]. (c) Variação do perfil energia no intervalo de refinamento [2900-10780].

O limiar para o segundo formante foi detectado na janela de análise 155 (amostra 6515) e o perfil energia na janela 156 (amostra 6537). A fronteira é estabelecida na amostra 6537, amostra essa em que os dois parâmetros ultrapassam os limiares estabelecidos.

A grande dificuldade no refinamento das vogais quando ocorrem em ditongos é a presença das vogais pertencentes à mesma classe fonética, por exemplo a ocorrência de duas vogais anteriores ou duas vogais posteriores.

Neste terceiro e último caso abordado, como as vogais apresentam as mesmas características acústicas, a utilização das freqüências formantes não auxilia no processo de refinamento. A estratégia então adotada é o uso do critério de informação Bayesiana, como descrito no Capítulo 3.

Na Figura 5.28 é mostrado um exemplo do refinamento entre duas vogais anteriores (vogal [e] e vogal [i]) da locução "dinheiro". O intervalo de refinamento foi definido entre as amostras 7600 e 14300.

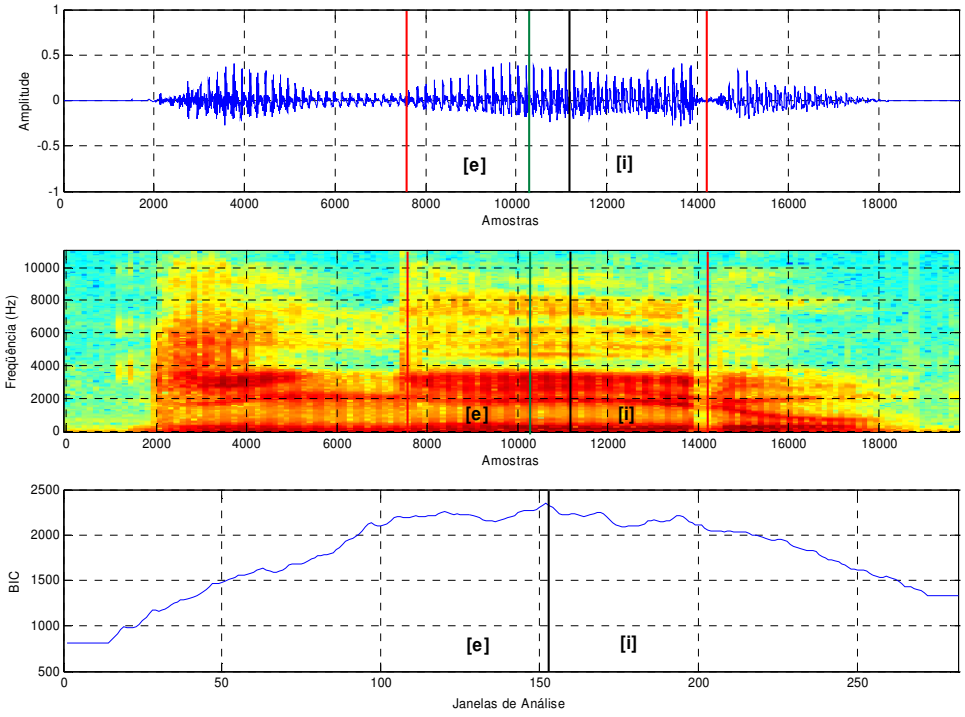


Figura 5.28: Locução "dinheiro": (a) Forma de onda. (b) Espectrograma. (b) Critério de informação Bayesiana no intervalo de refinamento [7600-14300].

Para o cálculo do BIC, inicialmente no intervalo de refinamento são calculados apenas os parâmetros mel-cepstrais (ordem 12) sem as derivadas de primeira e segunda ordem. Esses parâmetros por sua vez são modelados por uma Gaussiana multivariada.

A determinação da fronteira entre os dois fones envolvidos leva em consideração a existência de dois segmentos inicialmente de tamanhos diferentes, onde cada um também é modelado por uma Gaussiana multivariada. O objetivo do BIC consiste em variar, a cada instante de tempo, o tamanho dos segmentos e calcular a variação entre eles usando a Equação (3.10). Por último uma análise é realizada para localizar o ponto máximo de variação entre os dois segmentos e defini-lo como a fronteira de segmentação.

Como pode ser acompanhado na Figura 5.28 (c), o BIC tende a exibir um pico na transição entre as vogais. O pico mostrado na figura é detectado na janela de análise de número 152, que por sua vez corresponde à amostra 11149 (definida como a transição entre os fones).

5.5. Considerações Finais

Neste capítulo foram apresentados de modo detalhado os módulos que compõem o sistema para refinamento da segmentação automática de fala. O sistema proposto é composto por três módulos principais: treinamento, segmentação e refinamento. Inicialmente, no desenvolvimento do trabalho, os módulos de treinamento e segmentação foram implementados usando a linguagem de programação C++ para ambiente Windows. Com o uso de fones dependentes de contexto, optou-se por trabalhar somente com o HTK para o treinamento e segmentação via Viterbi.

O módulo de refinamento que foi a proposta do trabalho é realizado com base nos fones que compõem a locução, ou seja, é necessário que a transcrição fonética da locução esteja presente. Os fones foram agrupados nas suas classes mais representativas e, para cada classe, foi determinado um conjunto de parâmetros acústicos para serem empregados no refinamento.

Algumas classes fonéticas são fáceis de serem refinadas, como por exemplo, o silêncio, as fricativas e as africadas. Para essas classes pôde-se perceber que poucos parâmetros são suficientes. Outras apresentam uma complexidade maior, como por exemplo, as plosivas e as vogais (principalmente em ditongos).

Pelos exemplos apresentados para cada classe neste capítulo pôde-se também verificar que é possível determinar com grande precisão as fronteiras.

No próximo capítulo serão apresentados os resultados do sistema de refinamento em três bases de fala diferentes.

Capítulo 6

Resultados e Discussão

No capítulo anterior, o sistema de refinamento da segmentação automática de fala baseado em regras foi detalhadamente apresentado. Cada regra específica de refinamento, para cada uma das diferentes classes fonéticas, foi explicada e exemplificada.

Neste capítulo são apresentados os resultados dos testes realizados com três bases de fala diferentes, duas bases do PB e uma base do inglês Americano. Uma avaliação da qualidade da segmentação automática produzida pelo alinhamento de Viterbi também é apresentada, destacando a influência de diversos parâmetros no resultado final.

6.1. Bases de Fala

Os testes realizados neste trabalho utilizaram três bases de fala com o objetivo de avaliar tanto a segmentação realizada pelo algoritmo de Viterbi quanto o sistema baseado em regras para o refinamento da segmentação automática. Através dos testes realizados com as bases foi possível avaliar o sistema quanto à dependência ou não de locutor, avaliar quanto ao sexo do locutor e os parâmetros que afetam a segmentação de fala.

Das três bases utilizadas uma é independente de locutor para o inglês americano (TIMIT) e, as outras duas são dependentes de locutor (uma gravada por um locutor masculino e a outra por um locutor feminino).

6.1.1. Base de Fala Dependente de Locutor Masculino

A base de fala contínua do Português do Brasil, dependente de locutor masculino, é composta por 1226 locuções. Desse total, 1026 locuções são utilizadas para o treinamento dos HMMs e as 200 locuções restantes, diferentes das locuções de treinamento, são utilizadas para teste de segmentação e refinamento.

As locuções foram gravadas por um locutor paulista, do interior do Estado de São Paulo, na ocasião com 29 anos de idade. Por ser considerada uma base pequena, os regionalismos do país e as diferentes pronúncias para algumas locuções não são abordados. Durante a elaboração das frases houve uma preocupação em manter o número médio de ocorrências de cada fone.

Todas as locuções foram capturadas usando uma placa Sound Blaster, amostradas a uma taxa de 22,05 kHz e quantizadas com 16 bits por amostra. Durante a gravação de cada locução houve um cuidado para manter o mínimo de ruído possível. A transcrição de cada locução foi gerada de forma manual, utilizando 39 unidades fonéticas, conforme mostrado na Tabela 4.1. As sentenças gravadas têm uma duração média de 3 segundos e a lista completa encontra-se no Apêndice A.

As locuções do conjunto de teste também foram manualmente segmentadas para a gerar a segmentação de referência utilizada para avaliar a segmentação automática de fala. A segmentação manual foi realizada usando o software livre *Praat* (www.praat.org), mas não foi feita por um foneticista, e sim pelo próprio autor do trabalho. Cada fronteira foi determinada analisando-se a forma de onda do sinal, o espectrograma, curvas de energia e trajetória de formantes.

6.1.2. Base de Fala Dependente de Locutor Feminino

A base de fala contínua do Português do Brasil, dependente de locutor feminino, foi cedida gentilmente pela empresa VOCALIZE - Soluções em Tecnologias da Fala e da Linguagem Ltda, empresa de base tecnológica incubada na UNICAMP, com o propósito de avaliar a qualidade do sistema de refinamento desenvolvido quando aplicado à fala feminina.

Em comparação com a base dependente de locutor masculino, a base de fala feminina é menor. A base é composta apenas por 100 locuções gravadas por uma locutora feminina, paulista, falante nativa do português brasileiro. A locutora selecionada é musicista, especializada em Canto Popular e Lírico, contando na época da gravação com 29 anos de idade.

A gravação das locuções foi acompanhada por profissionais habilitados, realizada em sala acusticamente tratada. Houve preocupação com a postura da locutora em relação ao microfone e às locuções lidas. Para a gravação foi utilizado também equipamento profissional (M-Box Digidesign, microfone Beringher B2-Pro). As locuções foram amostradas a uma taxa de 22,05 kHz, com resolução de 16 bits por amostra.

A base é foneticamente balanceada e as locuções foram selecionadas do CETENFolha (Corpus de Extractos de Textos Eletrônicos NILC/Folha de São Paulo), disponível no endereço eletrônico http://acdc.linguateca.pt/cetenfolha.

Junto com as locuções também foi cedida pela empresa a transcrição fonética de cada locução e a segmentação manual. Tanto a transcrição fonética quanto a segmentação manual foram realizadas por profissional habilitado, utilizando a representação SAMPA-PB conforme proposta de Morais (Morais et al., 2005).

Como os símbolos da representação SAMPA-PB diferem dos símbolos utilizados neste trabalho, uma alteração foi feita para adaptar as transcrições para o padrão utilizado.

6.1.3. Base de Fala Independente de Locutor (TIMIT)

A TIMIT (*Texas Instruments/Massachussets Institute of Technology*) é uma base de fala independente de locutor desenvolvida inicialmente com a finalidade de desenvolver e testar sistemas de reconhecimento de fala. Devido ao aumento crescente de pesquisas em diversas áreas de processamento digital da fala, a TIMIT também passou a ser utilizada para testes em segmentação.

Para a confecção da TIMIT foram utilizados 630 locutores, abrangendo os 8 principais dialetos do Inglês Americano. Cada locutor pronunciou 10 locuções, totalizando 6300 locuções foneticamente balanceadas. Desse total apenas 5040 são utilizadas para os testes. Esse conjunto é dividido em locuções de treinamento (3696 locuções) e locuções de teste (1344). É muito importante destacar que do total de locuções de teste, 624 são distintas.

Diferente das locuções das bases de fala dependente de locutor, as locuções da TIMIT foram amostradas a 16 kHz, com uma resolução de 16 bits por amostra. Cada locução do conjunto de treinamento e teste apresenta sua transcrição fonética, a segmentação manual em termos de palavras e também em termos de fones. Essa segmentação manual fornecida será utilizada como referência para avaliar a segmentação automática produzida pelo sistema desenvolvido.

Na transcrição fonética das locuções é utilizado um conjunto de 64 fones distintos. A documentação da TIMIT sugere que apenas 48 fones sejam utilizados no treinamento dos modelos acústicos e sugere ainda que, na fase de teste esse conjunto seja simplificado para 39.

Na Tabela 6.1 são mostrados os fones representativos das plosivas, fricativas, africadas, consoantes nasais e o silêncio. Como pode ser observado na tabela há dois símbolos diferentes para representar o silêncio: silêncio epentético ("epi"), silêncio que é freqüentemente encontrado entre uma fricativa e uma semi-vogal ou nasal (como exemplo a palavra "slow"), e o silêncio prolongado ("sil"), normalmente presente no início, no final e nas pausas das locuções.

Tabela 6.1: Símbolos utilizados na transcrição fonética das plosivas, fricativas, consoantes nasais e silêncio.

Classe Fonética	**Símbolo**	**Exemplo**
Plosiva	b	bee = **vcl b** iy
Plosiva	d	day = **vcl d** ey
Plosiva	g	gay = **vcl g** ey
Plosiva	p	pea = **cl p** iy
Plosiva	t	tea = **cl t** iy
Plosiva	k	key = **cl k** iy
Plosiva	dx	dirty = vcl d er **dx** iy
Fricativa	s	sea = **s** iy
Fricativa	sh	she = **sh** iy
Fricativa	z	zone = **z** ow n
Fricativa	zh	azure = ae **zh** er
Fricativa	f	fin = **f** ih n
Fricativa	th	thin = **th** ih n
Fricativa	v	van = **v** ae n
Fricativa	dh	then = **dh** e n
Nasais	m	mom = **m** aa **m**
Nasais	n	noon = **n** uw n
Nasais	ng	sing = s ih **ng**
Nasais	en	button = vcl b ah cl t **en**
Africadas	jh	joke = **vcl jh** ow cl k
Africadas	ch	choke = **cl ch** ow cl k
Voiced closure	vcl	
Unvoiced closure	cl	
Silêncio epentético	epi	
Silêncio	sil	

Uma das simplificações sugerida pela documentação da TIMIT é quanto aos símbolos que representam o período de constrição das plosivas (*voiced and unvoiced closure*). A transcrição original fornecida pela TIMIT utiliza símbolos diferentes para representar o período de constrição de cada uma das plosivas. Por exemplo, para as plosivas surdas [p], [t] e [k] são utilizados os

símbolos “pcl”, “tcl” e “kcl”. Da mesma forma, para as plosivas sonoras [b], [d] e [g], são utilizados os símbolos “bcl”, “dcl” e “gcl”. Segundo a documentação, o período de contrição de todas as plosivas surdas deverá ser representado por “cl” e, das plosivas sonoras, deverá representado por “vcl”.

As outras alterações sugeridas são: os fones “ax-h”, “ux”, “axr”, “em”, “nx”, “eng”, “j”, “hv” e “pau” são substituídos por “ax”, “uw”, “er”, “m”, “n”, “ng”, “jh”, “hh” e “sil”, respectivamente. O fone “q” é removido da transcrição, por sugestão da documentação. Na tabela 6.2 são listados os símbolos utilizados na transcrição fonética das vogais e semi-vogais da TIMIT.

Tabela 6.2: Símbolos utilizados na transcrição fonética das vogais e semi-vogais.

Classe Fonética	**Símbolo**	**Exemplo**
Vogal	iy	beet = vcl b **iy** cl t
Vogal	ih	bit = vcl b **ih** cl t
Vogal	eh	bet = vcl b **eh** cl t
Vogal	ey	bait = vcl b **ey** cl t
Vogal	ae	bat = vcl b **ae** cl t
Vogal	aa	bott = vcl b **aa** cl t
Vogal	aw	bout = vcl b **aw** cl t
Vogal	ay	bite = vcl b **ay** cl t
Vogal	ah	but = vcl b **ah** t
Vogal	ao	bougth = vcl b **ao** cl t
Vogal	oy	boy = vcl b **oy**
Vogal	ow	boat = vcl b **ow** cl t
Vogal	uh	book = vcl b **uh** cl k
Vogal	uw	boot = vcl b **uw** cl t
Vogal	er	bird = vcl b **er** vcl d
Vogal	ax	about = **ax** vcl b aw cl t
Vogal	ix	debit = vcl d eh vcl b **ix** cl t
Semi-Vogais	l	lay = **l** ey
Semi-Vogais	r	ray = **r** ey
Semi-Vogais	w	way = **w** ey
Semi-Vogais	y	yacht = **y** aa cl t
Semi-Vogais	hh	hay = **hh** ey
Semi-Vogais	el	bottle = vcl b aa cl t **el**

Durante a fase de segmentação e teste foram utilizados apenas 48 símbolos diferentes, como sugerido pela documentação da TIMIT. A TIMIT também não faz nenhuma referência às vogais nasalizadas.

6.2. Avaliação do Alinhamento de Viterbi

O ponto de partida para a avaliação do sistema foi verificar a qualidade da segmentação realizada pelo algoritmo de Viterbi, amplamente utilizado em segmentação automática de fala.

Através dessa avaliação foi possível verificar a influência de alguns parâmetros na segmentação automática de fala, tais como o número de Gaussianas na mistura, quantidade de janelas adjacentes utilizadas para o cálculo dos parâmetros delta, e fones dependentes de contexto. O principal objetivo dessa avaliação foi determinar a combinação de parâmetros que produza o melhor segmentador e, em seguida, aplicar as regras de refinamento para aproximar a segmentação automática da segmentação manual.

Os resultados se referem à base de fala independente de locutor TIMIT em que o treinamento dos modelos acústicos e o alinhamento forçado de Viterbi foi realizado no HTK. A TIMIT foi escolhida para analisar a qualidade da segmentação automática justamente por ser uma base independente de locutor.

Os resultados serão apresentados através de porcentagens, indicando o número de marcas de segmentação que apresentam erro (diferença entre a segmentação manual e a automática) abaixo de um limiar previamente estabelecido. Esta forma de avaliação é a mais utilizada entre os pesquisadores da área de segmentação automática de fala, como descrito no Capítulo 3. A idéia é maximizar a porcentagem com erros de segmentação inferiores a 20 ms. Outra medida também empregada foi o valor médio do módulo do erro, conforme mostrado na Equação (3.13).

Teste 1 – Fones dependentes e independentes de contexto:

Neste primeiro teste é verificada a influência de se usar fones independentes de contexto ou fones dependentes de contexto para a segmentação automática. Para o primeiro teste foi utilizado um vetor acústico de dimensão 39 (Mel + LogEnergia + ΔMel + ΔLogEnergia + ΔΔMel + ΔΔLogEnergia), HMM com 3 estados, 5 Gaussianas na mistura, janela de análise de 20 ms com deslocamento a cada 10 ms e apenas uma janela para o cálculo dos parâmetros delta.

Os resultados para a segmentação com fones independentes e dependentes de contexto são mostrados na Tabela 6.3.

Tabela 6.3: Resultado da segmentação automática para fones independentes e dependentes de contexto.

Limiar (ms)	**Porcentagem de Erro**	
	Fones Independentes de Contexto	**Fones Dependentes de Contexto**
<= 5	24,23	26,98
<= 10	46,45	51,15
<= 20	**75,61**	**81,01**
<= 30	86,63	90,87
<= 40	90,62	95,09
<= 50	92,92	97,06
<= 100	97,45	99,63

Do reconhecimento automático de fala sabe-se que sistemas com fones dependentes de contexto tendem a apresentar taxas de reconhecimentos melhores em relação aos que não empregam fones dependentes de contexto. Analisando a Tabela 6.3 pode-se concluir que para a segmentação automática de fala os melhores resultados também são obtidos com a modelagem dependente de contexto. Para o limiar de 20 ms há uma melhora de 5,4% sobre a modelagem independente de contexto, mas em todos os outros limiares também existe um ganho da modelagem dependente de contexto. Outra observação é com relação ao valor médio do módulo do erro que foi de 14,19 ms para a modelagem independente de contexto e 14,03 para a modelagem dependente de contexto.

Os resultados apresentados na Tabela 6.3 divergem dos resultados apresentados por Toledano (Toledano et al., 2003). Em seu artigo é mencionado que HMMs dependentes de contexto tendem a produzir resultados de segmentação menos precisos do que os resultados produzidos por HMMs independentes de contexto. A explicação dada pelos autores é que durante o treinamento dos HMMs dependentes de contexto ocorre perda de alinhamento entre o fone e o seu contexto. Como os HMMs dependentes de contexto são sempre treinados com realizações de fones dentro de um mesmo contexto, eles não têm informações para diferenciar entre o fone e o contexto. Dos resultados apresentados pelos autores, HMMs independentes de contexto apresentaram melhora de 4,49% sobre os HMMs dependentes de contexto para erros de segmentação menor que 20 ms, divergindo totalmente dos resultados aqui obtidos (5,4% de melhora sobre HMMs independentes de contexto).

Como os fones dependentes de contexto sempre apresentam os melhores resultados, em todos os testes seguintes só será adotada essa modelagem.

Teste 2 – Número de Gaussianas na mistura:

Neste segundo teste foi verificada a influência do número de Gaussianas nas mistura dos HMMs que modelam a densidade de emissão de símbolos. Para o treinamento foi utilizado um vetor acústico de dimensão 39 (Mel + LogEnergia + ΔMel + ΔLogEnergia + ΔΔMel + ΔΔLogEnergia), HMM com 3 estados e com número de Gaussianas por mistura variável, janela de análise de 20 ms com deslocamento a cada 10 ms e apenas uma janela para o cálculo dos parâmetros delta.

Na Tabela 6.4 são mostrados os resultados da segmentação automática variando o número de Gaussianas nas misturas entre 1 e 7, na tabela 6.5 variando entre 8 e 13 e, por fim na tabela 6.6 variando entre 14 e 20.

Tabela 6.4: Resultado da segmentação automática variando o número de Gaussianas na mistura entre 1 e 7.

Limiar (ms)	**Número de Gaussianas na Mistura**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
<= 5	25,91	25,91	26,39	26,59	26,41	26,37	26,51
<= 10	51,37	54,43	51,93	51,90	51,43	51,22	51,20
<= 20	**81,75**	**81,95**	**81,84**	**81,65**	**81,39**	**81,18**	**80,92**
<= 30	91,38	91,35	91,19	91,11	91,18	91,20	91,02
<= 40	95,34	95,36	95,29	95,31	95,31	95,33	95,33
<= 50	97,12	97,16	97,18	97,24	97,27	97,31	97,29
<= 100	99,60	99,62	99,64	99,67	99,68	99,69	99,68

Tabela 6.5: Resultado da segmentação automática variando o número de Gaussianas na mistura entre 8 e 14.

Limiar (ms)	**Número de Gaussianas na Mistura**						
	8	**9**	**10**	**11**	**12**	**13**	**14**
<= 5	23,21	26,14	26,14	26,12	26,12	25,94	25,70
<= 10	51,01	50,78	50,78	50,51	50,51	50,08	49,79
<= 20	**80,63**	**80,07**	**80,04**	**79,81**	**79,81**	**79,38**	**79,17**
<= 30	90,90	90,55	90,56	90,43	90,43	90,26	90,33
<= 40	95,29	95,16	95,16	95,11	95,11	95,09	95,23
<= 50	97,24	97,19	97,18	97,16	97,16	97,22	97,37
<= 100	99,68	99,69	99,68	99,69	99,69	99,67	99,68

Tabela 6.6: Resultado da segmentação automática variando o número de Gaussianas na mistura entre 15 e 20.

Limiar (ms)	Número de Gaussianas na Mistura					
	15	**16**	**17**	**18**	**19**	**20**
<= 5	25,61	25,49	25,37	25,34	25,29	25,29
<= 10	49,61	49,40	49,22	49,14	49,08	49,08
<= 20	**78,97**	**78,83**	**78,70**	**78,57**	**78,48**	**78,48**
<= 30	90,21	90,06	89,97	89,89	89,82	89,82
<= 40	95,12	95,04	94,97	94,94	94,88	94,88
<= 50	97,30	97,31	97,27	97,27	97,25	97,25
<= 100	99,66	99,67	99,68	99,67	99,66	99,66

A análise dessas três tabelas revela um resultado muito interessante. O melhor segmentador é aquele cujo HMM apresenta apenas 2 Gaussianas na mistura. O aumento gradual do número de Gaussianas em cada modelo provoca uma leve degradação do alinhamento de Viterbi, mas por outro lado, pode aumentar a taxa de reconhecimento.

Já era de conhecimento que o melhor reconhecedor não é necessariamente o melhor segmentador (Yared, 2006), mas é muito comum empregar em segmentação o HMM com as mesmas características do HMM empregado para o reconhecimento.

Na Figura 6.1 é apresentado um gráfico com a evolução do valor médio do módulo do erro para os resultados apresentados nas tabelas 6.4, 6.5 e 6.6.

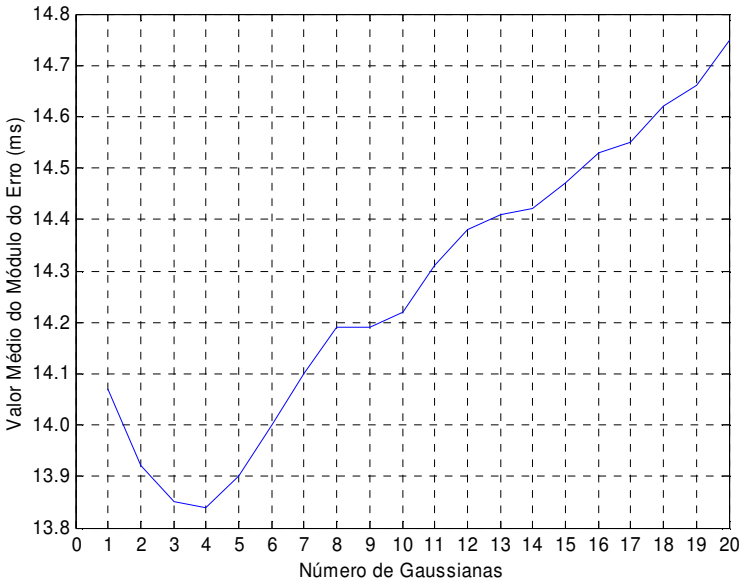


Figura 6.1: Evolução do valor médio do módulo do erro.

Do gráfico apresentado na Figura 6.1, o menor valor médio do erro é obtido para a modelagem com apenas 4 Gaussianas por mistura, e à medida que o número de Gaussianas é incrementado, o valor médio do erro também aumenta.

O objetivo do HMM não é modelar as fronteiras entre os fones das locuções, e sim modelar os fones propriamente dito. Um aumento do número de Gaussianas promove uma melhor modelagem dos fones e, conseqüentemente, um aumento na taxa de reconhecimento, o que não é necessariamente seguido pela taxa de segmentação, conforme comprovado pelas Tabelas 6.4, 6.5 e 6.6 e também pela Figura 6.1.

Teste 3 – Parâmetros Acústicos

No terceiro teste alguns parâmetros acústicos são avaliados para verificar o desempenho com relação à segmentação automática de fala. Para o treinamento dos HMMs foi utilizado um vetor acústico de dimensão variada (de acordo com os parâmetros empregados). O HMM foi modelado com 3 estados e apenas 2 Gaussianas na mistura, janela de análise de 20 ms com deslocamento a cada 10 ms e apenas uma janela para o cálculo dos parâmetros delta (quando utilizado esse parâmetro).

Na tabela 6.7 são mostrados os resultados da segmentação automática de fala para os diversos parâmetros acústicos.

Tabela 6.7: Resultado da Segmentação automática variando os parâmetros acústicos.

Limiar (ms)	**Parâmetros Acústicos**				
	Mel+Energia+DMel+ DEnergia+DDMel+DDEnergia	**Mel**	**Mel+DMel +DDMel**	**Mel+Energia**	**PLP**
<= 5	25,91	27,87	28,68	**29,73**	27,34
<= 10	54,43	51,86	53,32	**54,97**	51,37
<= 20	**81,95**	78,09	80,83	81,78	77,98
<= 30	**91,35**	87,11	90,47	89,77	87,23
<= 40	**95,36**	91,65	94,62	93,46	91,77
<= 50	**97,16**	94,21	96,54	95,58	94,45
<= 100	**99,62**	98,83	99,46	99,24	98,93
Erro Médio	**13,92**	15,05	14,13	14,00	15,10

Neste teste, cinco combinações de parâmetros foram empregadas: parâmetros clássicos (Mel + LogEnergia + ΔMel + ΔLogEnergia + ΔΔMel + ΔΔLogEnergia), apenas os parâmetros

mel-cepstrais, parâmetros mel-cepstrais com as respectivas derivadas primeira e segunda, parâmetros mel-cepstrais com o parâmetro energia e também a análise de predição linear perceptual (PLP – *Perceptual Linear Prediction*).

A análise da tabela 6.7 revela alguns pontos interessantes sobre a influência dos parâmetros acústicos para a segmentação automática de fala. A combinação dos parâmetros clássicos tem o melhor resultado para erros menores que 20 ms e inclusive para o valor médio do módulo do erro.

Para os erros menores que 20 ms, o uso dos parâmetros mel-cepstrais com as derivadas primeira e segunda e mel-cepstrais com energia produzem os melhores resultados, com destaque para a combinação entre os parâmetros mel-cepstrais e a energia que apresentam a melhor taxa.

Vale a pena destacar também que a combinação dos parâmetros mel-cepstrais com o parâmetro energia produziu um resultado muito próximo dos parâmetros clássicos para o limiar de 20 ms, tendo apenas uma diferença de 0,17%.

Para os limiares de erro acima de 20 ms, os melhores resultados obtidos são com o uso dos parâmetros clássicos, tendo grande vantagem sobre os outros parâmetros. Apesar de todas essas divergências iniciais, independente do conjunto de parâmetros utilizados, as regras de refinamento são suficientes para aproximar a segmentação automática da segmentação manual.

É muito importante ressaltar que a segmentação automática de fala não está totalmente relacionada com o reconhecimento automático, ou seja, parâmetros que são conhecidos por produzir boas taxas de reconhecimento não produzem necessariamente bons resultados na segmentação automática de fala.

6.3. Avaliação do Refinamento da Segmentação Automática de Fala

As análises produzidas na seção anterior tiveram como objetivo principal determinar o melhor conjunto de parâmetros que produza os melhores resultados de segmentação. Os parâmetros determinados são então empregados para o treinamento dos HMMs e também são utilizados no alinhamento forçado de Viterbi.

A Tabela 6.8 mostra os resultados da segmentação automática de fala fornecida pelo alinhamento forçado de Viterbi. Os resultados foram obtidos a partir de fones dependentes de contexto, vetor de parâmetros acústicos com dimensão 39, HMM com 3 estados, 2 Gaussianas por mistura, janelas de análise com duração de 20 ms com deslocamento a cada 10 ms e 1 janela

para o cálculo dos parâmetros delta. Para a base no PB foram utilizados 38 fones diferentes e, para a TIMIT, 48 fones. Na base de fala masculina o HTK gerou 2918 unidades dependentes de contexto enquanto que, na TIMIT, foram geradas 10180 unidades dependentes de contexto.

Tabela 6.8: Resultados da segmentação automática de fala fornecida pelo alinhamento forçado de Viterbi.

Limiar (ms)	**Porcentagem de Erro**		
	Base Masculina	**Base Feminina**	**TIMIT**
<= 5	21,98	17,86	25,91
<= 10	40,00	32,09	54,43
<= 20	66,49	55,13	81,95
<= 30	83,24	70,59	91,35
<= 40	89,77	81,78	95,36
<= 50	93,48	89,23	97,16
<= 100	99,18	95,12	99,62
Erro Médio	16,47	17,20	13,92

Os resultados da segmentação obtidos para a base independente de locutor são melhores do que os resultados obtidos para a base dependente de locutor masculino e dependente de locutor feminino. Os melhores resultados obtidos para a TIMIT justificam-se pelo número de locuções disponíveis para treinamento e, conseqüentemente, melhor modelagem dos fones dependentes de contexto. Por outro lado, a base dependente de locutor feminino apresentou os piores resultados, uma vez que a segmentação foi realizada a partir dos HMMs treinados com a fala masculina.

Os resultados da TIMIT, para o limiar de 20 ms, são 14,52% melhores em relação à base de fala masculina, e 25,88% melhores em relação à base feminina. Os resultados para a base masculina em relação à base feminina são 11,36% melhores.

Uma análise de erros por classe fonética também foi realizada antes do processo de refinamento. Essa análise é importante porque fornece uma visão sobre o comportamento dos parâmetros acústicos sobre os resultados da segmentação para cada classe fonética. Além disso, a análise por classes também é importante para verificar quais são as que não apresentam bom desempenho no alinhamento de Viterbi. Na Tabela 6.9 são apresentados os erros de segmentação por classes fonéticas para a base de fala TIMIT.

Tabela 6.9: Erros por classes fonéticas para a TIMIT.

Classes Fonéticas	Limiar (ms)						
	<= 5	<= 10	<= 20	<= 30	<= 40	<= 50	<= 100
Silêncio	21,69	42,37	65,74	77,08	87,43	94,79	99,11
Vogal Anterior	29,91	54,54	81,36	91,52	95,26	97,23	99,63
Vogal Central	28,25	50,11	78,03	90,33	95,80	97,74	99,75
Vogal Posterior	26,29	47,56	75,83	88,31	93,46	96,70	99,34
Plosiva Surda	27,70	50,43	80,06	91,66	96,40	98,29	99,83
Plosiva Sonora	29,33	53,69	81,55	93,85	97,66	98,58	99,59
Fricativa Surda	26,56	48,11	77,54	90,85	97,19	98,74	99,79
Fricativa Sonora	30,19	54,00	83,02	94,00	97,08	98,25	99,73
Africada	33,75	57,22	83,94	94,58	97,83	98,92	99,46
Consoante Nasal	28,76	52,37	78,44	89,29	94,93	97,38	99,68

Como pode ser comprovado pela análise da Tabela 6.9, as classes não apresentam uma discrepância muito grande. Para erros abaixo de 20 ms, a classe representada pelo silêncio apresenta o pior desempenho, e o melhor desempenho é apresentado pelas fricativas e africadas.

Na Tabela 6.10 são apresentados os erros de segmentação por classes fonéticas para a base dependente de locutor masculino. Diferente da TIMIT, na base dependente de locutor masculino há uma discrepância maior entre os erros. Para erros abaixo de 20 ms ela tem comportamento pior em relação à TIMIT, mas para erros acima de 20 ms a base dependente de locutor apresenta melhores resultados.

Tabela 6.10: Erros por classes fonéticas para a base dependente de locutor masculino.

Classes Fonéticas	Limiar (ms)						
	<= 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 Central	3,88	10,72	46,52	83,12	93,16	96,01	99,32
Vogal Posterior	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
Róticas	22,75	37,75	70,50	84,00	87,75	91,00	98,75

Levando em consideração o limiar de 20 ms, as classes representadas pelas vogais central, posterior e nasal apresentam os piores resultados. Por outro lado, as plosivas sonoras, fricativas sonoras e africadas apresentam excelente resultados. De uma forma geral, os resultados por classes fonéticas são melhores do que para a TIMIT.

Na Tabela 6.11 são apresentados os resultados após a aplicação das regras de refinamento às fronteiras de segmentação inicialmente determinadas pelo alinhamento forçado de Viterbi.

Tabela 6.11: Resultados da segmentação automática de fala após o refinamento.

Limiar (ms)	**Porcentagem de Erro**		
	Base Masculina	**Base Feminina**	**TIMIT**
<= 5	61,00	35,20	57,10
<= 10	73,00	50,50	68,69
<= 20	95,55	78,02	92,97
<= 30	96,98	82,67	94,50
<= 40	98,23	92,14	96,32
<= 50	98,50	95,92	97,70
<= 100	100	98,10	100
Erro Médio	10,69	14,40	11,52

Após o refinamento, a base de fala masculina apresentou os melhores resultados. Os limiares utilizados para as três bases foram determinados a partir da base de fala masculina, o que justifica o melhor desempenho. A base de fala feminina não apresentou bons resultados em virtude de apresentar características acústicas e fonéticas diferentes da fala masculina ao qual o HMM foi treinado. A solução adotada para tentar melhorar os resultados com a base de fala feminina é aplicar adaptação de locutor.

Para a adaptação de locutor foi empregada uma técnica baseada em transformações lineares chamada *Maximum Likelihood Linear Regression* (MLLR), realizada no próprio *software* HTK. Nenhuma outra técnica de adaptação foi empregada ou testada durante este trabalho.

A técnica consiste em estimar um conjunto de transformações lineares para a média e a variância dos parâmetros dos HMMs do sistema que está sendo adaptado, provocando um "deslocamento" desses componentes de forma que cada estado do HMM do sistema seja capaz de gerar os dados adaptados.

Como a base de fala feminina não dispõe de muitos dados, foi realizada uma adaptação global onde as transformações são aplicadas a cada Gaussiana do modelo. A tabela 6.12 apresenta uma comparação entre a segmentação produzida pelo alinhamento de Viterbi, sem aplicar adaptação, e após a aplicação de adaptação de locutor. Os resultados apresentados não foram submetidos ao refinamento (apenas com o alinhamento de Viterbi).

Tabela 6.12: Resultados da adaptação de locutor para a base de fala feminina, sem o refinamento.

Limiar (ms)	**Porcentagem de Erro**	
	Sem Adaptação de Locutor	**Com Adaptação de Locutor (MLLR)**
<= 5	17,86	19,73
<= 10	32,09	34,54
<= 20	55,13	56,22
<= 30	70,59	71,58
<= 40	81,78	82,35
<= 50	89,23	90,53
<= 100	95,12	96,67
Erro Médio	17,20	16,90

Com a adaptação de locutor, para o limiar de 20 ms, houve uma melhora de apenas 1,09%, mas uma melhora em todos os limiares também pôde ser observada. Na tabela 6.13 são reportados os resultados do refinamento para a base de fala após a adaptação de locutor.

Tabela 6.13: Comparação entre os resultados de refinamento para a base de fala feminina antes e após a adaptação de locutor.

Limiar (ms)	**Porcentagem de Erro**	
	Sem Adaptação de Locutor	**Com Adaptação de Locutor (MLLR)**
<= 5	35,20	36,29
<= 10	50,50	52,30
<= 20	78,02	79,52
<= 30	82,67	83,57
<= 40	92,14	92,89
<= 50	95,92	96,50
<= 100	98,10	98,82
Erro Médio	14,40	14,06

Apesar da adaptação de locutor empregada, mesmo com o refinamento, a precisão da segmentação não sofreu grandes alterações. Para o limite de tolerância de 20 ms houve uma melhora de apenas 1,5%.

Da análise das tabelas 6.12 e 6.13 pode-se concluir que, neste caso, a adaptação de locutor não é suficiente para melhorar a precisão da segmentação automática de fala. Uma justificativa para o baixo desempenho da adaptação de locutor é que o material disponível para a base de fala feminina é muito pequeno. As fronteiras devem estar fora do intervalo de refinamento e, dessa forma, os limiares dos parâmetros definidos não são suficientes para aproximar as fronteira em relação à segmentação manual.

6.4. Correção de Erros Sistemáticos

A partir das figuras do capítulo anterior, mostrando o funcionamento de cada regra de refinamento, é possível perceber que, dependendo da classe fonética, a marca de segmentação pode estar um pouco antes ou um pouco depois da marca de segmentação de referência (segmentação manual). Essa diferença é conhecida na literatura como erro sistemático ou *bias*.

Esse erro sistemático foi observado tanto antes quanto após o refinamento e é totalmente dependente das classes fonéticas presentes do lado esquerdo e do lado direito de cada fronteira. Para a sua remoção foi implementado um módulo adicional no sistema, que só foi aplicado antes do processo de refinamento.

Alguns autores (Wang et al., 2004 e Demuynck and Laureys, 2002) sugerem o uso de técnicas de pós-processamento para remover possíveis erros sistemáticos das fronteiras inicialmente estimadas.

No trabalho proposto por Wang é utilizada uma árvore de regressão e classificação (CART), semelhante ao procedimento utilizado pelo HTK. O método tem por objetivo fazer a correção de erros sistemáticos baseado em modelos de misturas de Gaussianas. Inicialmente as fronteiras de uma base de fala segmentada manualmente são clusterizadas de acordo com as características fonéticas dos fones e modelos de misturas são empregados a cada cluster.

A estratégia utilizada nesta tese é similar à proposta por Demuynck e Laureys. Um estudo do comportamento da segmentação produzida pelo alinhamento de Viterbi é realizado e um desvio médio é estimado. Para determinar o desvio médio, as classes fonéticas utilizadas pelo

módulo de refinamento também são utilizadas e, em seguida, todas as possíveis transições entre as classes fonéticas são mapeadas.

Para determinar o valor do *bias* para cada classe fonética, a diferença entre a segmentação manual e a segmentação automática é calculada para cada par de classes fonética. O valor médio dessa diferença representa o *bias*. É interessante notar que, para algumas classes, a posição da fronteira estimada está sempre antes da fronteira correta e, para outras, sempre depois. O valor do *bias* foi determinado para cada uma das bases de fala utilizando apenas as locuções de treinamento, com exceção para a base de fala feminina onde só existem as locuções de teste.

A tabela 6.14 mostra os erros de segmentação após a aplicação da remoção do *bias* na segmentação inicialmente gerada pelo alinhamento forçado de Viterbi, sem a aplicação das regras de refinamento.

Tabela 6.14: Resultados da segmentação automática de fala após a remoção do *bias*.

Limiar (ms)	Porcentagem de Erro		
	Base Masculina	Base Feminina	TIMIT
<= 5	35,77	19,32	39.20
<= 10	58,82	32,09	52.15
<= 20	81,26	57,05	85.70
<= 30	86,09	73,90	89.82
<= 40	90,69	82,55	92.54
<= 50	92,38	92,10	95.73
<= 100	97,80	96,95	99.68
Erro Médio	14,19	16,31	12,98

Comparando a tabela 6.14 com a tabela 6.8, após a remoção do *bias* há uma melhora em todos os resultados da segmentação para todas as bases de fala e em todos os limiares. Os melhores resultados são obtidos para a base dependente de locutor masculino, justamente por ser uma base dependente de locutor. Para o limiar de 20 ms, a base masculina teve uma melhora de 14,77%, a base feminina 1,92% e a TIMIT 4,69%.

Após a remoção do *bias*, as locuções foram submetidas ao processo de refinamento, e os resultados obtidos foram os mesmos mostrados na tabela 6.11. A conclusão a partir desses resultados é que a remoção do *bia*s antes do processo de refinamento não traz nenhuma vantagem, uma vez que, após a aplicação do refinamento, os resultados são os mesmos (Selmini e Violaro, 2007).

6.5. Considerações Finais

Neste capítulo o sistema baseado em regras para o refinamento da segmentação automática foi avaliado. O procedimento adotado para a avaliação foi através de comparação com a segmentação manual (segmentação de referência).

O ponto de partida para a avaliação foi gerar a melhor segmentação e, em seguida, aplicar o sistema de regras para o refinamento. Para gerar a melhor segmentação, uma série de testes foi realizada de forma a determinar o melhor conjunto de parâmetros a ser utilizado pelo algoritmo de Viterbi.

Apesar de todos os parâmetros testados, optou-se pelo conjunto clássico utilizado em reconhecimento automático de fala (parâmetros mel-cepstrais, logaritmo da energia normalizada e derivadas de primeira e segunda ordem). Outros parâmetros também produziram resultados próximos, representando possíveis alternativas.

Outro teste com resultado interessante foi o número de Gaussianas por estado. Ao contrário do que se imaginava, a redução do número de Gaussianas por estado contribui para os melhores resultados. Testes com a remoção do *bias*, estimados a partir das locuções de treinamento de cada base, também foram realizados. A partir dos resultados pôde-se concluir que a remoção de erros sistemáticos antes do refinamento não traz melhoras na segmentação uma vez que o processo de refinamento é capaz de aproximar as fronteiras obtidas da segmentação automática e da segmentação manual.

A base de fala feminina empregada apenas para testes de segmentação não apresentou bons resultados quando comparadas às outras duas bases. Uma forma de melhorar os resultados foi empregar técnicas de adaptação de locutor, que por usa vez apresentou uma pequena melhora. Uma possível justificativa é a pequena quantidade de material disponível.

Capítulo 7

Conclusões

7.1. Discussão Geral

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

A segmentação de fala desempenha papel primordial em diversas aplicações em que a fala é utilizada (reconhecimento automático de fala, síntese texto-fala, animações tridimensionais sincronizadas com a fala, etc). Atualmente essas aplicações utilizam bases de fala cada vez maiores, o que exige um processo de segmentação automática ao invés da segmentação manual. Apesar da segmentação automática produzir fronteiras com boa qualidade e precisão, dependendo da aplicação o refinamento é desejado para aproximar as fronteiras obtidas a partir da segmentação automática das que seriam produzidas pela segmentação manual.

No Capítulo 2 diversas técnicas relacionadas à segmentação automática foram apresentadas. Como o foco principal do trabalho não está no desenvolvimento de uma técnica específica para segmentação automática, optou-se por trabalhar com HMMs e o alinhamento forçado de Viterbi, que são largamente empregados em reconhecimento automático de fala, e produzem bom resultados. Apesar dos bons resultados, o refinamento faz-se necessário.

Quanto ao processo de refinamento da segmentação, que é o objetivo principal da tese, a estratégia adotada leva em consideração o tipo de transição, ou seja, os fones presentes no lado direito e esquerdo da fronteira que está sendo refinada. De acordo com o tipo de transição, características dependentes dos fones são empregadas para mover a fronteira em análise para uma

nova posição, de forma a aproximar das fronteiras que seriam obtidas a partir da segmentação manual.

O sistema proposto foi avaliado em três bases de fala, duas bases dependentes de locutor do PB (uma para locutor masculino e outra para locutor feminino), e uma base independente de locutor do inglês Americano (TIMIT).

7.2. Avaliação da Segmentação Automática de Fala

Apesar da segmentação automática não ser o objetivo principal da tese, o ponto de partida foi determinar o melhor conjunto de parâmetros para o treinamento dos HMMs de forma a obter o melhor resultado de segmentação, e dessa forma, produzir melhores resultados após o refinamento.

Inicialmente acreditava-se que o mesmo treinamento utilizado para o reconhecimento automático também seria suficiente para a segmentação. A partir dos testes realizados com os parâmetros conclui-se que o conjunto de parâmetros que produz o melhor reconhecedor não necessariamente produz o melhor segmentador.

Quando se empregam HMMs para segmentação é muito comum treinar os modelos das subunidades acústicas com o conjunto de parâmetros clássicos (12 parâmetros mel-cepstrais, 1 parâmetro log-energia normalizado e as derivadas de primeira e segunda ordem) empregados em reconhecimento automático. Com a avaliação de diversas combinações de parâmetros pôde-se perceber que, além do conjunto de parâmetros clássicos, outras alternativas também podem ser utilizadas.

Um ponto interessante que também pôde ser comprovado através dos testes foi o número de Gaussianas utilizadas na mistura dos HMMs, que foi variado entre 1 e 20. Novamente, diferente do reconhecimento, onde a redução do número de Gaussianas não produz bons resultados, os melhores resultados para a segmentação foram produzidos com apenas duas Gaussianas na mistura.

Outro teste realizado foi quanto à influência da dependência ou não de contexto dos fones na segmentação automática. De acordo com o teste realizado usando a TIMIT, também pôde-se comprovar que, para a segmentação, a modelagem através de unidades dependentes de contexto leva a melhores resultados em relação à modelagem de unidades independentes de contexto. Para

a TIMIT, considerando o limiar de 20 ms, há uma melhora de 5,4% das unidades dependentes de contexto sobre as unidades independentes de contexto.

Os resultados obtidos na fase de segmentação estão em conformidade com os resultados apresentados na literatura. Para a TIMIT, 81,01% das fronteiras apresentam erro de segmentação abaixo de 20 ms. Para as bases dependentes de locutor fica mais difícil fazer uma análise comparativa de desempenho porque não foi encontrado nenhum trabalho reportando resultados de segmentação para o Português do Brasil.

7.3. Avaliação do Refinamento da Segmentação Automática de Fala

Dentre as diversas possibilidades de refinamento, a estratégia proposta é baseada no tipo de transição ao qual a fronteira em análise pertence, ou seja, é totalmente dependente dos fones presentes do lado direito e esquerdo da fronteira. A técnica desenvolvida permite que cada fronteira da locução seja refinada de acordo com as características acústicas presentes na região de transição entre os fones envolvidos. Essa região de análise, para decidir a nova posição da fronteira, foi chamada de intervalo de refinamento.

Para cada classe fonética definida existem diversos parâmetros que podem ser empregados durante o refinamento. A combinação de todos esses parâmetros pode levar a regras bastante complexas, o que pode aumentar também o tempo de processamento para cada fronteira. Para evitar esse problema, optou-se por empregar parâmetros que pudessem caracterizar o tipo de transição entre as classes envolvidas, e não empregar todos os parâmetros de cada uma das classes no refinamento. Desse modo, houve uma redução no número de parâmetros para cada tipo de transição.

Uma das vantagens da técnica proposta é ser independente de modelo e treinamento, ou seja, não é necessário definir modelos para as fronteiras e, conseqüentemente, treinar cada um desses modelos com uma base manualmente segmentada.

A comparação de desempenho entre os resultados obtidos nesta tese com outros sistemas de segmentação e refinamento para o PB não é possível, pois não foram encontrados trabalhos reportando tais resultados.

Do estado da arte em refinamento da segmentação automática de fala descrito na Seção 3.7, Doroteo Toledano (Toledano et al., 2003) descreve os melhores resultados para o modo dependente de locutor, 96,01% de fronteiras com erro de segmentação abaixo de 20 ms. Neste

trabalho, após o refinamento, é obtido o valor, 95,55%. O sistema proposto por Toledano tem uma vantagem de 0,46% sobre o sistema proposto neste trabalho. É importante destacar que os resultados apresentados por Toledano foram obtidos após a adaptação de locutor e também após um processo estatístico de correção das fronteiras.

Para a TIMIT, os melhores resultados reportados após o processo de refinamento para o limiar de 20 ms foram: 92,47% com o uso de SVM (Lo and Wang, 2007) e 90,24% com o uso de um classificador LDA (Boonsuk et al., 2007). Os resultados obtidos no sistema de refinamento proposto são da ordem de 92.97%. O refinamento da TIMIT usando características acústicas apresentado neste trabalho leva vantagem de 0,5% sobre o refinamento usando Máquinas de Vetor de Suporte apresentado na revisão bibliográfica

Os resultados obtidos com a base de fala dependente de locutor feminino não foram bons se comparados com os outros resultados, uma vez que a segmentação foi realizada com HMMs treinados com a base de fala masculina. O refinamento aplicado não foi muito eficiente em corrigir as fronteiras o que pode mostrar que, se as fronteiras determinadas pelo alinhamento de Viterbi estiverem fora da região em que se situa o fone na locução, o refinamento poderá ser prejudicado. Mesmo adaptando a fala feminina à base dependente de locutor masculino os resultados não foram bons, apresentando apenas uma melhora de 1,5%.

7.4. Trabalhos Futuros

Durante o desenvolvimento deste projeto algumas idéias surgiram, mas não foram estudadas e implementadas. Algumas delas são possíveis de serem implementadas no curto prazo, porém outras exigiriam um estudo mais detalhado.

A primeira idéia proposta consiste em estudar o desempenho das características fonético-acústicas determinadas para cada classe fonética e aplicadas durante o processo de refinamento para segmentar as locuções. As locuções seriam segmentadas usando o mesmo processo utilizado para o refinamento, ou seja, com base na transcrição fonética da locução o algoritmo emprega um parâmetro ou conjunto de parâmetros de forma a localizar a fronteira entre fones adjacentes.

A segunda proposta também está relacionada com o processo de segmentação e não diretamente com o refinamento das locuções. A idéia consiste em aplicar treinamento discriminativo nos HMMs de forma a produzir uma segmentação com maior precisão.

O algoritmo clássico de treinamento dos HMMs baseado na Máxima Verossimilhança define o problema da segmentação de fala como um processo de estimação de parâmetros. Durante o alinhamento forçado de Viterbi, o algoritmo procura por uma região da locução que corresponde ao fone que está sendo alinhado, e não pela fronteira entre os fones da locução.

Com o treinamento discriminativo é definida uma função que traduz o erro de segmentação entre a segmentação automática e a manual durante o treinamento dos HMMs. Para que isso possa ser realizado para o treinamento é necessário fornecer, além das locuções e das respectivas transcrições fonéticas, a segmentação de cada locução.

Uma terceira sugestão para futuros trabalhos é acrescentar modelos de duração dos fones para auxiliar no processo de segmentação e também no refinamento.

A quarta sugestão consiste em realizar um estudo mais profundo sobre as transições entre fones da mesma classe fonética, o que é comum no PB. O uso do Critério de Informação Bayesiana produz bons resultados, mas um estudo mais detalhado faz-se necessário para obter fronteiras com mais precisão.

Apêndice A

Lista de Locuções da Base de Fala Masculina

A1. Locuções de Treinamento

1. O grêmio ganhou a quadra de esportes.
2. Hoje irei à vila sem meu filho.
3. Essa magia não acontece todo dia.
4. Será bom que você estude esse assunto.
5. O menu incluía pratos bem saborosos.
6. Podia dizer as horas, por favor?
7. A casa é ornamentada com flores do campo.
8. A terra é farta, mas infinita.
9. O sinal emitido é captado por receptores.
10. A mensalidade aumentou mais que a inflação.
11. O tele-jornal termina às sete da noite.
12. A cabine telefônica fica na próxima rua.
13. Defender a ecologia é manter a vida.
14. Nesse verão o calor está insuportável.
15. Um jardim exige muito trabalho.
16. O mamão que eu comprei estava ótimo.
17. Meu primo falará com a gerência amanhã.
18. De dia apague a luz sempre.
19. A sociedade uruguaia tem que se mobilizar.
20. Suas atitudes são bem calmas.
21. Dezenas de cabos eleitorais buscavam apoio.
22. A vitória foi paga com muito sangue.
23. Nossa filha tem amor por animais.
24. Esse peixe é mais fatal que certas cobras.
25. O time continua lutando pelo sucesso.
26. Essa medida foi devidamente alterada.
27. O estilete é uma arma perigosa.
28. Quinta eu venho jantar em sua casa.
29. A mudança é lenta, porém duradoura.
30. O clima não é mais seco no interior.
31. A sensibilidade indicará a escolha.
32. A Amazônia é a reserva ecológica do globo.
33. O ministério mudou demais com a eleição.
34. Novos rumos se abrem para a informática.
35. O capital de uma empresa depende da produção.

36. Se não fosse ela, tudo seria contido.
37. A principal personagem do filme é uma gueixa.
38. Receba seu jornal em sua casa.
39. A juventude tinha que revolucionar a escola.
40. A atriz terá quatro meses para ensaiar seu canto.
41. A velha leoa ainda aceita combater.
42. É hora do homem se humanizar mais.
43. Ela ficou na fazenda por uma hora.
44. Seu crime foi totalmente encoberto.
45. A escuridão da garagem assustou a criança.
46. Ontem não pude fazer minha ginástica.
47. Comer quindim é sempre uma boa pedida.
48. Hoje eu irei precisar de você.
49. Sem ele o tempo flui num ritmo suave.
50. A sujeira lançada no rio contamina os peixes.
51. O jogo será transmitido bem tarde.
52. É possível que ele já esteja fora de perigo.
53. A explicação pode ser encontrada na tese.
54. Meu vôo tinha sido marcado para às cinco.
55. Daqui a pouco a gente irá pousar.
56. Estou certo que mereço a atenção dela.
57. Era um belo enfeite todo de palha.
58. O comércio daqui tem funcionado bem.
59. É a minha chance de esclarecer a notícia.
60. A visita transformou-se numa reunião íntima.
61. O cenário da história é um subúrbio do rio.
62. Eu tenho ótima razão para festejar.
63. A pequena nave medirá o campo magnético.
64. O prêmio será entregue em sessão solene.
65. A ação se passa numa cidade calma.
66. Ela e o namorado vão a Portugal de navio.
67. O adiamento surpreendeu a mim e a todos.
68. A gente sempre colhe o que plantou.
69. Aqui é onde existem as flores mais interessantes.
70. A locomotiva vem sem muita carga.
71. Esse empreendimento será de enorme sucesso.
72. As feiras livres não funcionam amanhã.
73. Fumar é muito prejudicial à saúde.
74. Entre com o seu código e o número da conta.
75. Reflita antes e discuta depois.
76. As aulas dele são bastante agradáveis.
77. Usar aditivos pode ser desastroso.
78. O clima não é mau em Calcutá.
79. A locomotiva vem sem muita carga.
80. Ainda é uma boa temporada para o cinema.
81. A questão foi retomada no congresso.

82. Leila tem um lindo jardim.
83. O analfabetismo é a vergonha do país.
84. A casa foi vendida sem pressa.
85. Trabalhando com união rende muito mais.
86. Recebi nosso amigo para almoçar.
87. A justiça é a única vencedora.
88. Isso se resolverá de forma tranqüila.
89. Os pesquisadores acreditam nessa teoria.
90. Sei que atingiremos o objetivo.
91. Nosso telefone quebrou.
92. Desculpe se magoei o velho.
93. Queremos discutir o orçamento.
94. Ela tem muita fome.
95. Uma índia andava na mata.
96. Zé vá mais rápido!
97. Hoje dormirei bem.
98. João deu pouco dinheiro.
99. Ainda são seis horas.
100. Ela saía discretamente.
101. Este é o mês dos judeus.
102. Alguns professores dão muitas chances.
103. Os fóruns de justiça não funcionam amanhã.
104. Escolha o melhor curso de engenharia.
105. O padre lhe lançou um olhar intrigado.
106. O corredor do andar superior era amplo.
107. Abra os dedos e relaxe os músculos.
108. O carro subia cada vez mais a montanha.
109. O bispo tinha uma ametista roxa em seu anel.
110. O trem resfolegava e chiava.
111. Eu vi logo a Ioiô e o Léo.
112. Eles abaixaram-se enquanto as sirenes iram ficando mais baixas.
113. Seu avô lhe entregou uma chave de ouro.
114. Todo o lugar foi examinado.
115. Ao franzir o cenho examinou a ponta do eixo.
116. Aproximou deles a chave e examinou a beirada do metal.
117. O facho de luz foi dirigido para o carro.
118. Ela fechou a cara.
119. Naquela noite não podia ser exigente.
120. Assim podemos chocar um pouco o turista americano.
121. Sua cabeça tornou-se a encher de imagens.
122. Um homem não caminha sem um fim.
123. Tenho um milhão de coisas para lhe contar.
124. Conseguiram chegar a embaixada de trem.
125. O tenente foi levado para um aposento próximo.
126. Feche bem o cerco e rápido.
127. A polícia não lhes deixaria opções.

128. Era uma crônica espantosa composta de segredos e chantagens.
129. O rei detinha um poderoso segredo.
130. Vi Zé fazer essas viagens seis vezes.
131. O atabaque do Tito é coberto com pele de gato.
132. Ele lê no leito de palha.
133. Paira um ar de arara rara no Rio Real.
134. Foi muito difícil entender a canção.
135. Depois do almoço te encontro.
136. Esses são nossos times.
137. Procurei Maria na copa.
138. A pesca é proibida nesse lago.
139. Espero te achar bem quando voltar.
140. Temos muito orgulho da nossa gente.
141. O inspetor fez a vistoria completa.
142. Ainda não se sabe o dia da maratona.
143. Será muito difícil conseguir que eu venha.
144. A paixão dele é a natureza.
145. Você quer me dizer a data?
146. Desculpe, mas me atrasei no casamento.
147. Faz um desvio em direção ao mar!
148. Os maiores picos da terra ficam debaixo d'água.
149. A inauguração da vila é quarta-feira.
150. Só vota quem tiver o título de eleitor.
151. É fundamental buscar a razão da existência.
152. A temperatura só é boa mais cedo.
153. Em muitas regiões a população está diminuindo.
154. Nunca se pode ficar em cima do muro.
155. Pra quem vê de fora o panorama é desolador.
156. É bom te ver colhendo flores.
157. Eu me banho no lago ao amanhecer.
158. É fundamental chegar a uma solução comum.
159. Há previsão de muito nevoeiro no rio.
160. Muitos móveis virão às cinco da tarde.
161. A casa pode desabar em algumas horas.
162. O candidato falou como se estivesse eleito.
163. A idéia é falha, mas interessa.
164. O dia está bom para passear no quintal.
165. Minhas correspondências não estão em casa.
166. A saída para a crise dele é o diálogo.
167. Finalmente o mau tempo deixou o continente.
168. Um casal de gatos come no telhado.
169. A cantora foi apresentar seu último sucesso.
170. Lá é um lugar ótimo para tomar uns chopinhos.
171. O musical consumiu sete meses de ensaio.
172. Nosso baile inicia após às nove.
173. Apesar desses resultados, tomarei uma decisão.

174. A verdade não poupa nem as celebridades.
175. As queimadas devem diminuir este ano.
176. O vão entre o trem e a plataforma é muito grande.
177. Infelizmente não comparecei ao encontro.
178. As crianças conheceram o filhote de ema.
179. A bolsa de valores ficou em baixa.
180. O congresso volta atrás em sua palavra.
181. A médica receitou que eles mudassem de clima.
182. Não é permitido fumar no interior do ônibus.
183. A apresentação foi cancelada por causa do som.
184. Uma garota foi presa ontem à noite.
185. O prato do dia é couve com atum.
186. Eu viajarei ao Canadá amanhã.
187. A balsa é o meio de transporte aqui.
188. Muito prazer em conhecê-lo.
189. Eles estavam sem um bom equipamento.
190. O sol ilumina a fachada da tarde.
191. A correção do exame está coerente.
192. As portas são antigas.
193. Sobrevoamos Natal acima das nuvens.
194. Trabalhei mais do que podia.
195. Hoje eu acordei muito calmo.
196. Esse canal é pouco informativo.
197. Parece que nascemos ontem.
198. Receba meus parabéns pela apresentação.
199. Eu planejo uma viajem no feriado.
200. No lado de cá do rio há uma boa sombra.
201. A maioria dos visitantes gosta deste monumento.
202. Minha filha é especialista em música sacra.
203. A casa só tem um quarto.
204. A duração do simpósio é de cinco dias.
205. Ao contrário da nossa expectativa, correu tranqüilo.
206. A intenção é obter apoio do governante.
207. A fila aumentou ao longo do dia.
208. À noite a temperatura deve ir a zero.
209. A proposta foi inspecionada pela gerência.
210. O quadro mostra uma face do quotidiano.
211. Já era bem tarde quando ele me abordou.
212. O canário canta ao amanhecer.
213. A lojinha fica bem na esquina de casa.
214. Meu time se consagrou como o melhor.
215. Um instituto deve servir a sua meta.
216. Ele entende quando se fala pausadamente.
217. Seu saldo bancário está baixo.
218. O termômetro marcava um grau.
219. O discurso de abertura é bem longo.

220. Eu precisei do microfone na conferência.
221. Joyce esticou sua temporada até quinta.
222. Nada como um almoço ao ar livre.
223. Nossa filha é a primeira aluna da classe.
224. Gostaria de deitar um pouco.
225. Não fizemos uma viagem muito cansativa.
226. Ainda tenho cinco telefonemas para dar.
227. Os hotéis do sudoeste são fantásticos.
228. O presidente decidiu manter o ministro.
229. Herança e polimorfismo são conceitos de orientação a objetos.
230. As denúncias ainda podem sufocá-lo.
231. Não tenho remédio para me livrar desse tédio.
232. Ele se firma como o fiador da estabilidade.
233. Esse banco tem muitos investimentos.
234. A carroça estava carregada.
235. O rato roeu a roupa do rei de Roma.
236. Aqui é o lugar certo para investir o seu dinheiro.
237. Prepare-se para correr.
238. A maternidade aguça a inteligência da mulher.
239. Os moradores de rua no foco.
240. A mais importante joalheria do país.
241. São seis décadas de brilho.
242. Não temos autocrítica e estamos pagando por isso.
243. A sua estrela brilha hoje.
244. Desafie a mais poderosa força da natureza.
245. Palavras se desgastam como pedras roladas em fundo de rio.
246. No silêncio escutamos nossos próprios desejos.
247. A ética tem sido expulsa de muitos cenários atuais.
248. Confira alguns dos lançamentos do mês de dezembro.
249. Descobre-se agora que ele mentiu novamente.
250. Os computadores estão mais rápidos.
251. Inteligência artificial será o futuro.
252. O papa tirou a autonomia dos frades.
253. Mais uma idéia para salvar Veneza.
254. O presidente tem o seu pior desempenho.
255. A farsa cruel de um ponto de exclamação.
256. A história das bebidas que mudaram o mundo.
257. Vinho é a bebida preferida dos gregos e romanos.
258. Novela boa é, realmente, coisa de mulher.
259. Tenho grandes planos para o futuro.
260. Este trabalho é muito cansativo.
261. Essa tese dará muito trabalho.
262. A Internet revolucionou o mundo.
263. A gente é o que pensa.
264. Quem foi que disse que a gente não pode voar?
265. Cada um tem o mar que merece.

266. A comida baiana é muito forte.
267. Gosto muito de comida italiana.
268. Meus amigos foram para a Europa.
269. As melhores dicas do mundo não têm preço.
270. O apartamento está em reformas.
271. Seu coração precisava mais que cuidados.
272. A Argentina tem alguns símbolos.
273. Não importa o tamanho ou o segmento da sua empresa.
274. É tanta luz no natal que as pessoas ficam iluminadas.
275. Chuvas inundam bairros da zona norte.
276. Sempre pago as contas em dia.
277. É uma pequena cidade a uma hora da capital.
278. São décadas de trabalho para dar o melhor para você.
279. A turma estava muito a vontade no governo.
280. Come-se muito bem naquele restaurante.
281. Pato com maçã é uma boa combinação.
282. O futuro será decidido no tribunal.
283. Todos estavam a seu favor.
284. Tapetes grandes e espessos são caros.
285. O exame de matemática estava muito difícil.
286. A professora ficou muito zangada com a turma.
287. A árvore caiu com a tempestade.
288. Xadrez é um jogo para pessoas muito inteligentes.
289. Maria sempre come pudim como sobremesa.
290. Todos esperavam pela grande final.
291. O rio estava muito cheio por causa das chuvas.
292. As correntezas são muito fortes nesta época.
293. Todo dia é um grande dia para um homem.
294. Ela tem uma criação de gado.
295. O leite e o queijo são produzidos na fazenda.
296. A casa foi pintada de amarelo.
297. Compramos muitos livros de programação.
298. A lógica é a fundamental na programação de computadores.
299. A engenharia é responsável pelas grandes inovações.
300. Pedro cortou o dedo na foice afiada.
301. O namoro deles terminou de forma trágica.
302. Vamos velejar neste final de semana?
303. Minha mãe ficou muito brava com todo o barulho.
304. Será o sim mais esperado do ano.
305. O Brasil está blindado contra o populismo.
306. Vítimas de derrame recuperam parte dos movimentos.
307. O homem que explica o mundo.
308. Ele tornou-se um fenômeno desvendando o lado oculto do cotidiano.
309. Viver e não ter a vergonha de ser feliz.
310. Tudo vale a pena se a alma não é pequena.
311. O professor que tinha onze personalidades.

312. O filme conta a historia de um bruxo.
313. A crença no progresso nos impede de progredir.
314. Também na ciência o progresso é uma ilusão nociva?
315. O clima é imprevisível.
316. O homem é um sucesso evolutivo.
317. Ainda não conquistamos a terra
318. As habilidades humanas sempre trazem conseqüências negativas.
319. O genocídio é uma prática dos tempos modernos.
320. Somos a espécie dominante.
321. Se chover não vamos ao cinema.
322. Existem religiões que conflitam menos que a ciência.
323. Seu livro pode deixar desesperança.
324. Qual o papel da religião hoje?
325. Este ano houve muitos tornados.
326. A tempestade foi muito furiosa.
327. O livro traz contos eróticos muito picantes.
328. Presenteie quem você gosta de coração.
329. Ele sugere um conserto para a universidade pública.
330. Ajude o instituto e ganhe descontos na compra de produtos.
331. O carro estava descontrolado.
332. O controle passará para suas filhas.
333. Sou um cara comum.
334. Os americanos não cobram pouco de seus alunos.
335. A viagem está marcada para maio.
336. O Detran não liberou a carteira.
337. A bela mulher estava sentada no rio.
338. O novo papa chama-se Bento.
339. Não é o valor do presente que faz a emoção.
340. O jogador quebrou a perna durante a partida.
341. Penso diferente dela sobre equilíbrio fiscal.
342. Foi uma imolação pela pátria.
343. O navio do cruzeiro é imenso.
344. O macaco pedia por bananas.
345. Que matéria melhor uma escola poderia ensinar?
346. Seu livro mostra que o país tem condições privilegiadas.
347. Na verdade o que a gente entrega são bons negócios.
348. Dezembro é o mês mais esperado do ano.
349. Você aciona um botão e a tela abaixa.
350. Adoro as férias de julho.
351. Cinco cinco sete um.
352. Este filme é de terror.
353. Comprei um conjunto de facas para churrasco.
354. Faz um bom tempo para tomar sorvete.
355. Os supermercados estão lotados aos sábados.
356. Era uma dádiva divina.
357. A página estava toda grifada.

358. Meu jornal está chegando rasgado.
359. Vou assinar quatro revistas.
360. O torneio dos campeões começa no domingo.
361. Sua tv ficou menor que um controle remoto.
362. A divina comédia é muito boa.
363. Tudo correu muito bem.
364. É o mundo em tempo real no seu celular.
365. Aquele carro é mais rápido que um avião.
366. Somos um anjo para você.
367. Baba de moça é muito doce.
368. Renato come muito brigadeiro.
369. Ela é a mais badalada artista do momento.
370. As bebidas estão cada vez mais caras.
371. Durmo cheio de problemas.
372. Quadro tem personalidade própria.
373. A obra faz uma ponte entre o antigo e o moderno.
374. Zuleica adora cozinhar milho.
375. As fofocas são comuns entre as mulheres.
376. Focas adoram peixes.
377. Converse sobre seus jogos favoritos.
378. Palpite sobre as novelas que estão no ar.
379. Saiba como usar os recursos do bate papo.
380. Grafos é uma teoria muito interessante.
381. Fale com o crítico de cinema hoje a tarde.
382. Cantor apresenta novo repertório em show.
383. O ministro foi agredido ontem de manhã.
384. Pornografia infantil é crime.
385. Regras e dicas de segurança.
386. Ar condicionado a partir de duzentos reais.
387. Se tiver dificuldades para ler, troque a imagem.
388. Não termino relação, acumulo.
389. Duda tenta justificar o dinheiro recebido.
390. Invasão deve ser crime hediondo.
391. Um dia eu tinha que realizar o sonho das minhas filhas.
392. Quem vê a foto quer levar para casa.
393. Tudo mudou na rotina de nossa casa.
394. Parece uma serenata em dó maior.
395. Eles latem como se estivessem uivando.
396. Conheça os novos candidatos.
397. Ele é muito sociável, brinca com outras raças.
398. Um elefante incomoda muito a gente.
399. O mesmo conhecimento, um novo nome.
400. Todo telhado é perigoso.
401. A telha quebrou com a chuva.
402. De dia caminho bastante.
403. É uma diva da música clássica.

404. A minilipo elimina gorduras.
405. O dado do jogo estava viciado.
406. Doda vai se casar na sexta-feira.
407. Titia fez um bom chá.
408. Totó saiu ao encontro de Suzana.
409. Aqui se ensina, aqui se aprende.
410. Seu filho aprende certo.
411. Certeza é tudo que não posso dar.
412. Fogo não pega em lenha molhada.
413. Nossas escolas oferecem ensino de qualidade.
414. Grupos de internautas pregam a intolerância.
415. Racismo já foi detectado como crime.
416. Ver a copa do mundo não tem preço.
417. Sai da fila e venha conosco.
418. Uma cartilha para a prevenção de doenças.
419. Deve-se praticar uma hora de exercícios.
420. Faça uma dieta baseada em frutas.
421. Alimentos ricos em gorduras só uma vez por semana.
422. Não sobrecarregue uma criança.
423. Eles são o sangue novo.
424. Deram idéias para melhor a vida.
425. Conheça os vencedores dos prêmios.
426. São jovens cientistas.
427. É o carro mais caro da categoria.
428. Para ter opinião você precisa de informação.
429. A vida em tempo real.
430. A maneira mais barata de viajar.
431. Dicas para não perder a mala.
432. Ao achar uma pechincha, feche negócio.
433. Faça o melhor negócio em imóveis.
434. Enfrente um leão bem mansinho no ano que vem.
435. Aprenda a usar o crédito de maneira correta.
436. Remova adesivos de companhias aéreas.
437. Amarre fitas e grude colantes coloridos na bagagem.
438. Seja obsessivo antes de deixar sua bagagem na esteira.
439. Os pacotes foram cotados em janeiro, mês de férias.
440. Qual editora leva você para a Alemanha?
441. Agora sua foto pode virar notícia.
442. Sonhe com o mundo.
443. Você tem o seu estilo.
444. A revolta na França.
445. As cobras estavam nervosas.
446. Mais alimentos e mais músculos.
447. Um programa de alimentação e de exercícios.
448. Reserve seis semanas para consumir os alimentos certos.
449. Saiba o que beber e o que não beber.

450. Cálcio, o futuro do combate a gordura.
451. Afinal, o que é diabetes?
452. Ele já demonstrava um talento singular como contador de histórias.
453. O livro mergulha no intrigante universo dos sistemas de informação.
454. Ela precisa encontrar a chave do engenhoso código.
455. Presa numa teia de segredos e mentiras.
456. É o melhor e mais realístico suspense tecnológico.
457. Impossível não ficar arrepiado a cada página.
458. A agência de inteligência mais poderosa do mundo.
459. Ficou pensando se devia perturbar.
460. Uma história comovente e bem forjada.
461. Alcançou a lista dos livros mais vendidos.
462. Aranhas são insetos perigosos.
463. Essa impressora não funciona bem.
464. Comer tatu é bom.
465. Traficantes queimam ônibus no Rio.
466. Zâmbia expulsa igreja sob acusação de satanismo.
467. Agrediram-no com uma bengala.
468. Retornemos a esse ponto mais tarde.
469. Falarei sobre a hidrofobia dela.
470. O homem precisa desesperadamente de ajuda.
471. Lamento a complexidade da segunda vida.
472. Notou alterações em seus sentimentos.
473. Mais do que uma crônica dos acontecimentos recentes.
474. Homens armados roubam filhote de leão.
475. Todas as informações foram divulgadas pela Internet.
476. O filhote de panda apareceu para o público.
477. É um dos mamíferos com o maior risco de extinção.
478. Um lago para descansar.
479. Ele não se apresentará na China.
480. Animais também têm direitos.
481. A tartaruguinha é muito resistente.
482. A legislação proíbe a importação de espécies estrangeiras.
483. Zebras são encontradas nas savanas africanas.
484. Laranjas contêm vitamina C.
485. Chocolates são calóricos e tem muita gordura.
486. Existem duas subespécies de jibóias.
487. Formigas são metódicas e disciplinadas.
488. Alguns possuem cenários em três dimensões.
489. Temos um imenso prazer em recebê-lo.
490. Digite sua senha para receber o boletim.
491. É proibida a reprodução do documento.
492. Nessa vitamina não tem papaia.
493. Nos menus abaixo você tem acesso a tudo.
494. Lontras vivem na água.
495. Boa parte do país deve ter chuva.

496. Nuvem carregada sobre a cidade.
497. Patos cruzam a rodovia.
498. Festa para a copa do mundo.
499. A serpente exalou um veneno muito forte.
500. Velas pela vida de um australiano.
501. Zélia comprou telhas na roça.
502. Zuzu ama correr na floresta.
503. Ele nos oferece uma jornada fascinante e cândida.
504. Bem-vindos ao mundo dos importantes.
505. Conforto e qualidade de vida para a família.
506. Experimente a força do conhecimento.
507. Pulseirinhas coloridas definem o poder.
508. O consulado americano promete coibir a ação.
509. Prepare sua casa para o natal.
510. Aqui jazia um muro branco.
511. Entrega com condições especiais.
512. Seja um bicho de respeito.
513. Veja o que ele pode fazer por sua empresa.
514. O que é melhor chileno ou francês?
515. A um passo de virar a bambambã.
516. Da sala de aula para o consultório.
517. Três navios e muitos destinos para você.
518. Curta as atrações da avenida Vilares.
519. Ensaio fotográfico revela surpresas.
520. Longa jornada dentro da noite.
521. Garanta seu lugar nessa festa.
522. Ofertas especiais de inauguração.
523. Aqui tudo faz a diferença.
524. O cantinho virou depósito.
525. Economize e compre o que você quiser.
526. O enterro foi ontem.
527. Não gostei dos atores.
528. Tico canta toda tarde.
529. Faça sua barba e seu cabelo.
530. Tonico e Tinoco são cantores sertanejos.
531. Estrutura é apenas uma forma de organização.
532. Tudo o que eu gosto eu repito.
533. Admite-se chapeiro e cozinheira.
534. Minha namorada me acha uma pessoa intensa.
535. Quando gosto de uma música, ouço o dia todo.
536. Ela não só respeita como adora.
537. Ninguém faz igual.
538. Dirceu fica sem direitos políticos.
539. Cresce o número de corredores acima de 70.
540. A qualidade de vida aumentou o mês passado.
541. O partido é um grande risco.

542. Ao sinal vermelho pare.
543. Teste e comprove a melhoria.
544. A polícia prendeu uma quadrilha.
545. As notas foram divulgadas.
546. Carlos vigia seu gato durante o dia.
547. Tina bebe chá apenas de manhã.
548. Do sonho a grande realidade.
549. Tirson comprou um celular da Tim.
550. Frases podem ter várias palavras.
551. Esse conjunto de jantar é de prata.
552. A mesa custou muito caro.
553. As ruas estavam cheias de gente.
554. O parque espera por ajuda.
555. É um símbolo de perdição.
556. Pasta de trabalho em branco.
557. Isso que é doce vida.
558. Livro inspirado na obra de Rubem Fonseca.
559. Tereza foi nomeada como ministra.
560. Por muito tempo resisti a medicação.
561. Achava que iria comprometer a minha vida.
562. O distúrbio afetou sua carreira.
563. Não existe democracia.
564. O dado de Daiane está viciado.
565. Crianças gostam de dadinho de leite.
566. Eu acompanho todo o noticiário político.
567. Minha diversão é tentar adivinhar o que se passa.
568. Não preciso de mais uma frase perdida.
569. Artigos são extensos e pouco informativos.
570. O professor estava sem giz.
571. Zorro é um personagem lendário.
572. O zepelim cruzou a cidade.
573. A amizade que gerou sagas.
574. Muita aventura e um pouco de religião.
575. A terra do lado de lá.
576. Tenha um presente no armário.
577. Nunca pergunte o que a pessoa quer.
578. É melhor fazer sondagens.
579. Eles fazem o orçamento estourar.
580. Fuja dos presentes de última hora.
581. Organize uma lista para as pessoas.
582. Anote na agenda o nome das lojas.
583. Sua primeira máquina de escrever.
584. Uma amiga foi sorteada.
585. Devolveram a moto em dois dias.
586. Entregaram o livro ontem.
587. Meus pais eram contra.

588. Foi a grande sorte dela.
589. Sinto a cinta apertada.
590. Não faça dieta drástica.
591. Prepare uma lista com os pratos.
592. Brinde com espumantes nacionais.
593. Sandra é refém do produto importado.
594. Ande depois do jantar.
595. Coma porções iguais de cada comida.
596. Há prejuízo para o sabor.
597. Incrementa a festa e sai barato.
598. As camas têm cabeceiras a parte.
599. Hoje tem mais carros na rua.
600. Cachorros correm pela praia.
601. Como chegar as melhores ofertas.
602. Considere a idéia de alugar.
603. Dava dó de ver.
604. Um tiro na dor.
605. Remédio promete alivio imediato.
606. A cada dia um novo time.
607. O jornalista foi para a guerra.
608. Há tanto para aproveitar.
609. Maior conforto com flocos de gel.
610. Proteção garantida e maior economia.
611. Não chega não.
612. Toda probidade é suspeita.
613. Azar de quem não comprar.
614. Vi o último dos imperadores.
615. Pelo mesmo motivo chegarei cedo.
616. Ele pensa o contrário.
617. Nunca mais em minha vida.
618. Minha mãe é pequena.
619. Decida ca comigo.
620. Pago para investigar crimes.
621. Na sola da bota.
622. Num arroubo totalitário.
623. A agência vai atuar na América.
624. A sorrir eu pretendo levar a vida.
625. Novo recorde no mercado editorial.
626. Tenista brasileiro terminou a temporada.
627. Vamos nos casar.
628. Foi um gigante épico.
629. Lara lamenta a situação.
630. Quero queijo fresco.
631. Modelos gerais de manuscrito.
632. Foi uma reação desmedida e descabida.

633. Consegui saber o que ele gerou.
634. Laudas de teste para corrigir.
635. Tarifa zero para falar.
636. Lados do mesmo lago.
637. O foguete foi para a lua.
638. Maresia e calor em Maceió.
639. O rio Amazonas passa na Amazônia.
640. Tito e Teco andam na avenida.
641. Viagem ao mundo dos sonhos.
642. Pacto entre gerações é saída para educação.
643. Visão sistêmica pode ajudar na tomada de decisões.
644. Fórum reafirma autonomia universitária.
645. Novos projetos prometem aquecer o setor.
646. O maestro dissemina musica erudita.
647. Aqui tem conteúdo sob medida para os alunos.
648. Tire suas dúvidas pela Internet em tempo real.
649. Universidades abrem espaço para a criatividade.
650. Uma nova forma de arte congrega diferentes áreas.
651. O ensino é o grande desafio nacional.
652. Batermos perna em centenas de lojas.
653. Tão útil e pequeno que nem enche o saco.
654. Uma seleção de presentes em todas as faixas de preço.
655. Acho bacana a reportagem sobre o sucesso.
656. Talento e persistência são virtudes.
657. Nossos governantes deveriam ser honestos.
658. Gostaria de agradecer a homenagem ao meu avô.
659. Ir ao teatro ficou mais fácil.
660. Conheça a nova boutique da cidade.
661. Maneco me fez uma surpresa.
662. Vamos comemorar juntos o evento.
663. Venha visitar as peças.
664. Estão fazendo mais que o possível.
665. Medimos a temperatura em vinte lojas.
666. Promoção volta ao mundo.
667. Lalo comeu doce de laranja.
668. Segunda-feira, lá em casa.
669. O cara de pau e o enrosco das camisetas.
670. Celulares com preços inacreditáveis.
671. Cintura fina é bom para a saúde.
672. A morte lenta da democracia.
673. A gente nem percebe o que é fácil.
674. É o maior felino das Américas.
675. Lages é uma pequena cidade.
676. Lutar e vencer são coisas diferentes.
677. Luto pelo que quero e acredito.
678. A medicina descobriu a cura.

679. Gordura demais é prejudicial.
680. Lula é um prato perfeito.
681. Larissa infringiu as leis.
682. Salmão e tilápia são peixes
683. Detesto filme preto e branco.
684. Esse som é irritante.
685. Compilar é diferente de interpretar.
686. Tenha paciência que ele virá.
687. Ganhamos o suficiente para poder comprar.
688. Banho de sol só depois das três.
689. Seu cachorro deve tomar banho.
690. O coelho escondeu-se na toca.
691. Sente-se e sirva-se.
692. Pagamento apenas em dinheiro.
693. Semente de linhaça tem proteínas.
694. Sinto uma sensação zonza.
695. Odeio o zorra total.
696. Tenho pavor de cobras.
697. Compre castanhas e amêndoas para a ceia.
698. Diga adeus ao pneuzinho.
699. A gordura localizada é a mais nociva à saúde
700. Meu filho será um campeão.
701. O filhote brincava no jardim.
702. Como esse lombo a califórnia.
703. O presidente está destruindo a democracia
704. Isso é uma questão de tempo.
705. Coma muitas frutas e legumes.
706. Receba os meus cumprimentos.
707. Dinda lançou o dardo.
708. A flecha atingiu a manga.
709. Os valores não podem ultrapassar dois.
710. Carregue as caixas de dinheiro para o banco
711. Imprimas os ingressos com antecedência.
712. Compare as notas e decida pelo melhor.
713. Divergência e convergência são opostos.
714. Busque novas oportunidades.
715. A reunião terminou com fracasso.
716. Dia sim, dia não.
717. Nesse jardim tem dálias.
718. Confiança envolve transparência.
719. A bactéria vive no intestino.
720. O manual ensina a se comportar em viagens.
721. É um paraíso negro em Brasília.
722. A lhama habita o Peru.
723. Um cafezinho por favor.
724. Disque o número e informe a senha.

725. O senhor viaja no próximo trem.
726. A cachaça tem muito álcool.
727. É possível monitorar os sinais vitais.
728. Que todos seus sonhos se realizem.
729. Você vai pagar caro se não ler o anuncio.
730. Compro a veja para ler a vejinha.
731. Não cometa mais erros.
732. Erre, mas admita.
733. Componha suas histórias de aventura.
734. Quero lançar um grande desafio.
735. Contenha a doença com vacinas.
736. O mundo é plano para todos.
737. Conquiste o sucesso merecido.
738. A paca comeu o coco.
739. O capim cresceu no quintal.
740. A ida ao banheiro será vigiada.
741. Todo mundo de olho no Brasil.
742. O grupo foi considerado fácil.
743. Poderá haver problemas na segunda fase.
744. No outono as folhas caem.
745. É um grupo balanceado.
746. Beba seu leitinho todas as manhãs.
747. Cafeína é prejudicial ao organismo.
748. A festa será brega na Alemanha.
749. Seu prestígio foi o privilégio no sorteio.
750. Mande-o comprar seu remédio.
751. O quarto mais barato custa quatrocentos.
752. São duas seleções fortíssimas.
753. Logo ligarei para Londres.
754. Torcedores com as cores da bandeira.
755. Nada será como antigamente.
756. Nade o mais veloz possível.
757. A francesa que renasceu.
758. Foi o primeiro transplante de face.
759. A menina levada mordeu a língua.
760. Levante a lavadeira com cuidado.
761. Coma esse lambari.
762. Meça todas as paredes.
763. Este relatório é o comprovante.
764. Esta é sua solicitação de matrícula.
765. Consulte amanhã o resultado da validação.
766. Valide o bilhete antes de sair do estacionamento.
767. Estacione sempre em lugares autorizados.
768. Mantenha sempre a distância no trânsito.
769. A solicitação será enviada ao diretor.
770. Pague as contas sempre em dia.

771. O imposto será reajustado.
772. Beba todo o suco de maracujá.
773. Arraste este móvel.
774. Se precisar, conte com a gente.
775. Muito prazer, eu sou o Marcos.
776. Espero que tenha muitas noites felizes.
777. Cuidado com o fogo.
778. Hoje tem festa no gueto.
779. O boi saiu do pasto.
780. Não gosto de azeite de dendê.
781. Meu dente está amarelo.
782. Vá correndo avisá-lo.
783. O show foi perfeito.
784. Pegue o meu xampu e a toalha.
785. Este sapato machuca a unha.
786. Gina tem um imenso coração.
787. Estude na faculdade de engenharia.
788. Seja sempre muito paciente.
789. Tenha todos os documentos em mãos.
790. Adoro frango com polenta.
791. A duração do ensino fundamental será ampliada.
792. Aqui o ensino começa tarde.
793. As crianças não têm estudo.
794. O analfabetismo ainda é alto.
795. Deposite em um fundo de investimentos.
796. Haverá churrasco e caipirinha.
797. Feijoada é o prato típico.
798. Descubra as belezas naturais.
799. Durma oito horas por dia.
800. Vi Lele levando o gato.
801. Subi no pé de lima.
802. Sonhar é viver.
803. O aluno colou no exame.
804. Ser consciente é respeitar.
805. Sempre respeite os mais velhos.
806. Você é o que você deseja.
807. Não tinha ninguém no laboratório.
808. A grade foi alterada.
809. O público não se sustenta por si só.
810. Pesquisa revela degradação do mar.
811. O navio afundou.
812. Caqui e jaca são frutas.
813. Pendure-a na porta para dar boas-vindas.
814. Há diversas opções disponíveis no mercado.
815. Uma tendência interessante neste ano são as borboletas.
816. Soluções práticas para enfeitar a casa.

817. Uma boa opção é escolher um galho natural.
818. Os vasos devem ser da mesma altura.
819. As menores são ideais para acompanhar o guardanapo.
820. Posicione-o em um lugar de destaque na sala.
821. Nenhuma festa é completa sem música.
822. É o único planetário fixo de São Paulo.
823. Tão perto que você vai querer morar aqui.
824. O dia em que as loiras deram uma forcinha.
825. Duas entidades beneficentes dividirão a porcentagem.
826. As mensagens devem trazer assinatura.
827. A empresa que nasceu para voar.
828. Tiveram um chilique com a chuva.
829. A chuva tirou as telhas do telhado.
830. Como jiló após o jejum.
831. A intenção ontem foi boa.
832. Banhe seu cachorro com sabão.
833. Não pigarre na colher de prata.
834. Venha compartilhar do churrasco comigo.
835. Minha cunhada surrou as meninas.
836. Um menino empurrou a campainha.
837. Fingi que gostei do pudim de berinjela.
838. O jegue carregou o bumbo.
839. O chefe tirou o chassi do carro.
840. O cheque veio da China.
841. O serviço contra cupins foi cumprido.
842. Encontrei um cupom jogado no banheiro.
843. Traga-lhe o jogo de xadrez.
844. Unte o fundo da frigideira.
845. Chame todos para jantar.
846. Meu xará usava um chalé.
847. A galinha era o xodó do xeique.
848. O burro beijou a lhama.
849. A jarra cheirava xarope.
850. O carrinho carregava capim.
851. A abelha pousou na coalhada do conselheiro.
852. Eu tomo banho após malhar.
853. Um é pouco, mas dois é bom.
854. Fiquei zonzo após a punhalada.
855. Julia comi bombom no telhado.
856. Assisti xuxinha e o guto no cinema.
857. Cumpra com seus deveres.
858. José soltava bombinhas no cantinho da casa.
859. Diga-lhe bom-dia ao entrar.
860. A sua bondade foi clara no bonde.
861. A cintura da modelo tinha cinqüenta centímetros.
862. Lasanha e nhoqui são pratos italianos.

863. O cacho de gengibre caiu.
864. Junte todo esse calhamaço.
865. O ódio é o empecilho para a bondade.
866. No calhambeque estão as quinquilharias.
867. Seu empenho foi satisfatório.
868. A porta emperrou com a tempestade.
869. A calha entope quando chove.
870. Dumbo era um elefante charmoso.
871. Juca juntou-se a Janete na ilha.
872. Julho é o mês das boas intenções.
873. o marinheiro saiu da marinha apressado.
874. A intenção dele é a globalização mundial.
875. O tempo passa rápido nesse mundinho.
876. Ele tinha uma vida mundana ao lado de Janete.
877. A carroça corria em alta velocidade.
878. Chame todos para procurar o ouro escondido.
879. Juliana tem um charme exótico.
880. Chico veste-se de palhaço todo ano.
881. O mundo de Juliano é tonto.
882. A palha queimava junto com o palheiro.
883. junte-se a nós no mundo perdido.
884. Nunca esqueça do orgulho dela.
885. Junior telefonou no orelhão.
886. O burro era orelhudo.
887. O sangue jorrava da orelha.
888. Onde estão as chaves?
889. Ontologia é uma ciência.
890. A pomba voou sobre o ponto.
891. A reunião é pontual.
892. Jaqueline levou um tombo.
893. Os remorsos pungem o criminoso.
894. Proteja as anilhas dos fungos.
895. A informação estava quente.
896. O médico foi cúmplice do checape.
897. A lâmina do serrote foi aumentada.
898. Eles eram uns pulhas.
899. Todos dançaram rumba e beberam rum a noite toda.
900. O canhoto pegou o bumerangue.
901. O pimpolho brincava de bingo.
902. A serragem provocou arrotos na criança
903. Os húngaros se divertem no aterro da praia.
904. O jesuíta come muita jabuticaba.
905. Caminhe sempre pelo atalho.
906. Levou um arranhão no bumbum no passear pelo engenho.
907. O discurso foi uma verdadeira palhaçada.
908. Coloque um punhado de sal.

909. Chega de apanhar verduras.
910. Todos tem uma parreira de uva.
911. O talher foi talhado pelo artesão.
912. Ele cumprimentou Jonas pelo trabalho.
913. Usufrua do poder do urucum.
914. A inveja lhe pungia o coração.
915. A roça tinha caminhos de alho.
916. Foi fundamental a fundação da sede.
917. Vinha um zumbido de dentro.
918. Junte todas as peças na carroça.
919. Faça a criança arrotar.
920. Arrear o cavalo é fundamental para a segurança.
921. Chega de xingar os animais.
922. O marrom predominou no desfile.
923. Gincém é um energético poderoso.
924. Arrombe a porta para roubar o dinheiro.
925. É importante varrer toda a casa.
926. Umbanda é um culto religioso.
927. Nhá Benta morreu de infarto.
928. O chocolate foi talhado no mármore.
929. A camiseta foi reduzida a frangalhos.
930. Espinha é uma designação comum das saliências.
931. Estranho é entrar sem bater na porta.
932. Comprei alguns espelhos para meu tio.
933. O zunzum correu toda a cidade.
934. O espelheiro fez um preço bom.
935. O leão de uma boa espichada.
936. Foi uma humilhação mundial.
937. Era a linha mais elevada do telhado.
938. Chame algumas pessoas para ajudar.
939. Farroupilha foi a revolução dos farrapos.
940. A pinguela quebrou com o pingüim.
941. Narciso se espelhava todos os dias.
942. Os farrapos viviam no sul do país.
943. O fechamento da bolsa provocou tumulto.
944. Aquele homem ferrenho está doente.
945. A rolha manchou a toalha de mesa.
946. O apanhador de alho ganhou o prêmio.
947. A tundra invadiu o atalho.
948. A falha no sistema provocou a demissão.
949. Ganhei muitos mexilhões.
950. Leio o folhetim diariamente.
951. A torneira pingava sem parar.
952. O espertalhão tentou pegar a chuveiro alheio.
953. Aproveite a chance que a vida oferece.
954. Este é um folhado de creme com baunilha.

955. A tocha ficou acesa com querosene.
956. Chumbo e mercúrio são tóxicos.
957. O fundiário morreu na fundição.
958. O homem galhardo tem um galgo.
959. A tumba do faraó era azul/
960. O jumbo pousou na água.
961. O chimpanzé foi empalhado.
962. A criança empurrou a chupeta.
963. A pilha de ferro caiu sobre a menina.
964. O presidente tinha uma grande camarilha.
965. Ela preferiu uma grande camarinha.
966. Junte todos os documentos.
967. O terremoto chacoalhou os telhados.
968. O palhaço deu muitas cambalhotas.
969. O trem descarilhou dos trilhos.
970. A campanha pela caminhada deu certo.
971. A pinha quebrou a unha do velho.
972. O caminhoneiro decidiu dirigir a caminhonete.
973. Minha mãe chama-se Ivonete.
974. Cânhamo é rico em fibras.
975. Ser acanhado é comum.
976. A estrela usa brinco com pingente.
977. Perdi o molho de chaves.
978. O carrapicho grudou na roupa do Nho Chico.
979. Os sinos do carrilhão tocaram.
980. Suas fronhas estão no fundo armário.
981. O arraial foi comemorado com bumba meu boi.
982. Espinhaço é a coluna vertebral.
983. Tinha cascalho no casaquinho.
984. Um pássaro não faz verão.
985. Ele tirou uma casquinha.
986. Tive uma caxumba perigosa.
987. A chalupa pegou fogo.
988. Bombeiros foram chamados para conter as chamas.
989. A chacina ocorreu em frente a igreja.
990. A gerencia informou-lhe o prejuízo.
991. A chácara estava rodeada de mato.
992. Foram motivos de chacota.
993. O fungo causou um arranhão no banhista.
994. O arrocho salarial foi cumprido pelo calhorda.
995. Estanho é um elemento químico branco.
996. Umbrela é apenas um guarda chuva grande
997. Esta chacrinha é rodeada por flores.
998. Tem chaleira no chalé.
999. Mantenha a tundra for a do arremesso do canhão.
1000. O cacto é espinhoso.

1001. Coma champignon e beba champanhe.
1002. A Austrália vive um chauvinismo.
1003. Deu um chega-pra-la no bandido.
1004. Toda criança tem um chocalho.
1005. O chope da choperia esta chocho.
1006. Jussara gerencia toda a empresa.
1007. Fagulhas pulavam dos brasas.
1008. A jarra enferrujou com o ar.
1009. O clima chuvoso dificulta a colheita.
1010. A circunferência tinha um bom diâmetro.
1011. Um católico reza joelhado.
1012. Judeus e hebreus são circuncisados.
1013. Um mendigo esfarrapado pedia comida.
1014. A circunvizinhança tinha mato e bichos.
1015. Xena varreu toda a sala.
1016. O leite coalhou na mamadeira.
1017. Cochicho é falta de respeito.
1018. A ferradura caiu na palha.
1019. Sua arrogância causou um arrombo nos cofres.
1020. A batalha mantinha o curso normal.
1021. Cochinilhas são insetos.
1022. Ouvia-se o zurro no xilindró.
1023. A xícara continha café quente.
1024. Use páprica e cominho no tempero.
1025. Os erros chegaram até mim.
1026. A choradeira foi grande por causa do choque.

A2. Locuções de Teste

1. Muitas pessoas participam da construção de um texto.
2. Tenho vergonha do meu país.
3. Cada aluno fez a sua avaliação.
4. Escute o seu coração.
5. Um dos encontros mais emocionantes foi o último do ano.
6. Para Karla, educar é uma atitude de esperança.
7. Ela afirmou que tinha sido um fato.
8. O jabuti come muita jabuticaba.
9. O texto da professora trazia bons pensamentos.
10. Minha mãe estava nervosa.
11. O cavalo pulou a cerca.
12. A chave do chaveiro enferrujo.
13. Conhecera grandes pescadores no mar.
14. Ricardo escreve de modo fácil.
15. Fica aqui o convite para sua leitura.
16. A essa altura todos estavam emocionados.
17. Ele gostava mais das músicas animadas.
18. A atriz recebeu o prêmio com entusiasmo.

19. Todos correram para pegar o cachorro.
20. O aparelho de jantar caiu no chão.
21. O sol faz propaganda do verão.
22. O povo não estava feliz com o governo.
23. Levante a vela para limpar a mesa.
24. A chuva derrubou várias casas.
25. O feijão da feijoada estava azedo.
26. O seu primo comeu todas as laranjas.
27. A guerra acabou com fracasso.
28. Tenha uma boa noite.
29. Era uma ótima opção.
30. O calor me deixou de garganta seca.
31. O anjo caiu do céu.
32. Cobras são animais peçonhentos.
33. A gente sempre faz o que pode.
34. Meu computador não conversa com a impressora.
35. Os padres se reuniram na sacristia.
36. Meu gravador de DVD quebrou.
37. Diga sempre a verdade para seus pais.
38. Diga pata bem baixinho.
39. Beba muita água durante o dia.
40. Ganhei um novo aparelho de DVD.
41. Temos um belo presente para você.
42. Bebidas são prejudiciais a saúde.
43. Por favor, fume longe de mim.
44. Corra Lola corra.
45. É a coisa mais linda do mundo.
46. Bonito é a cidade de Bonito.
47. A Rússia tem uma cozinha com história.
48. A faculdade abriu concurso.
49. Para ter boas notas é preciso estudar.
50. A criança acredita em fadas.
51. O filme foi filmado na Argentina.
52. Paraguai e Chile são os novos representantes.
53. Empréstimos não são concedidos aos pobres.
54. O governador socorreu as vítimas.
55. Acidentes são comuns nesta época do ano.
56. O ganso correu atrás do frango.
57. O quarto tinha uma cama imensa.
58. Todas as palavras são plosivas.
59. As fricativas ocorrem com freqüência.
60. A família viaja com muito prazer.
61. Foram confirmadas as escolhas.
62. O sangue corria firme nas veias.
63. A velha caiu e machucou o braço.
64. Avós e avôs costumam ser bonzinhos.

65. Tita e titio vieram jantar conosco.
66. Papai e mamãe compram frutas.
67. Alface e tomate têm vitaminas.
68. Casa e roupa lavada é o que elas querem.
69. Nova chance será dada ao aluno.
70. Eles tinham muita fome e sede.
71. Galinhas caipiras botam ovos todos os dias.
72. A reprodução do documento foi perfeita.
73. A torcida estava furiosa com o time.
74. Todo mundo quer uma teve de plasma.
75. As plantas foram fracasso total.
76. Aqui tem autorama para os mais velhos.
77. A política acelera a economia.
78. O programa causou uma alegria.
79. Sempre compro cadernos e livros nesta loja.
80. É uma experiência de fluxo contínuo.
81. Tem muitas canetas nesse estojo.
82. A revolução ocorreu em todo estado.
83. O programa terminou de maneira inesperada.
84. Preferência para idosos e gestantes.
85. A festa aconteceu com muita animação.
86. Meninas manhosas gostam de brincar.
87. Todos foram para a festa no iate.
88. Adoro o gosto gostoso da pêra.
89. Compramos muitas batatas.
90. Tudo saiu conforme o esperado.
91. A informação veio de repente.
92. O vento é o reflexo do mau tempo.
93. As chuvas de inverno afetaram as colheitas.
94. Ele ficou mudo ao ver a namorada.
95. A manhã será fria e com nevoeiro.
96. A Internet revolucionou a vida.
97. Pessoas comuns pagam muitos impostos.
98. Velocidade limite deve ser imposta.
99. Clique aqui para obter mais informações.
100. Saindo do trabalho passo na sua casa.
101. Chuvas provocam soterramentos e mortes.
102. Advogada é a segunda vítima.
103. Ônibus é incendiado durante o protesto.
104. Umberto casou-se com Solineuza anteontem.
105. A jaca caiu no telhado e depois de escorregar caiu na lenha.
106. Tenho muitas caixas de papelão para guardar os utensílios.
107. Ganhei muito dinheiro ajudando o xerife no cassino.
108. O furacão chegou com muita força no sul e sudeste do país.
109. Zélia usa o mármore para talhar o chocolate.
110. Zoraide vai toda segunda ao terrero de umbanda de cheveti.

111. A zona de convergência provocou uma série de temporais.
112. Chame sempre o xerife quando sentir medo.
113. Depois de fazer muita zoeira os alunos penduraram as chuteiras.
114. O filhote correu o dia todo atrás do caixote.
115. A lhama tinha arrastado a zebra para dentro da jaula.
116. A aula foi atrapalhada pela zoeira dos alunos.
117. A garrafa de marmelada foi arrastada pela correnteza.
118. Os bichos do zoológico têm acompanhamento diário.
119. Pinte as paredes do quarto de vermelho e a sala de azul.
120. O bebê ficou com o bumbum assado.
121. Tenho muita certeza que ela chegará num carro vermelho.
122. A noiva deixou o noivo aguardando por duas horas.
123. Os filhotes de cachorro atravessaram a ponte sozinhos.
124. Venha e visualize as novas tendências de mercado.
125. O cigarro queimou o xale cinza da Zezinha.
126. O carrinho enferrujado quebrou as rodas durante a corrida.
127. Nós deixamos de pintar as paredes para correr no parque.
128. O elefantinho Dumbo chora todas as noites.
129. A tumba da múmia foi derrubada pelo caminhão tanque.
130. Atrás da xoupana tinha lenha molhada para colher.
131. As colheres de prata foram trazidas da Noruega.
132. Quem nasce nas Antilhas é antilhano.
133. Tive a grande satisfação de me juntar a eles no jantar.
134. A cozinha estava cheirando ao cozido de jacaré.
135. As pontes estavam molhadas por causa do temporal.
136. O molho de tomate ficou bastante ácido.
137. Finja que está feliz e tudo voltará ao normal.
138. A raposa chegou de mansinho pegou o filhote e correu para a floresta.
139. Balhufa significa nada coisa nenhuma.
140. Bumba meu boi é um bailado popular do nordeste do país.
141. Bundão é uma pessoa sem coragem sem iniciativa.
142. Bunda mole é a mesma coisa que bundão.
143. Kazuza sempre criticou a burguesia em suas músicas.
144. Burrinho também é a bomba de freio dos automóveis.
145. Um burro é resultante do cruzamento da égua com o jumento.
146. A calha da casa do meu cunhado entupiu.
147. O caixeiro é um operário que faz caixas.
148. Pode ser também um empregado que tem a seu cargo as vendas a retalho.
149. Os cupins invadiram a fazenda do cupineiro.
150. Cumpra sempre com suas obrigações.
151. As pessoas ficaram presas nas ferragens durante o terremoto.
152. A palha queimou durante horas no armazém.
153. A lasanha a bolonhesa do cozinheiro ficou maravilhosa.
154. Bengalas são produzidas a partir de lenha de boa qualidade.
155. A lenha é uma madeira geralmente utilizada como combustível em fogões.
156. O leonino comeu muita lentilha na festa.

157. O cachorro malvado mordeu a perna da menina malvada.
158. O mundo não gira ao seu redor.
159. Nunca desista dos seus ideais por nada.
160. Dançaram rumba a noite toda.
161. A terra desta região está molhada por causa da chuva.
162. O chuveiro da Zenaide é muito caro.
163. A leoa garrida saiu lentamente da toca.
164. Metidabundu é uma pessoa muito pensativa.
165. Tenho certeza que a lhama morreu de fome.
166. Lixeiro lixou a porta com uma lixa de borracha.
167. Juditi vai a academia com freqüência para malhar.
168. O atum atravessou o rio nadando muito rápido.
169. Cancun é uma cidade maravilhosa que fica no México.
170. Fui assistir ao show da Xuxa e fiquei com cheiro de fumaça.
171. O casamento ocorreu no meio da floresta de carvalho.
172. No carvalho tem uma caixa de abelhas muito bravas.
173. O orvalho caiu no ramalhete que estava no telhado.
174. O palhaço imitou um espantalho cansado.
175. Juvenal adora correr toda manhã na floresta de carvalhos.
176. Junte todos os ingredientes e misture o alho por último.
177. O rato roeu a roupa do rei de Roma.
178. Ontem vi o jumento arrastando a carroça de milho.
179. A roupa do palhaço encolheu depois da chuva.
180. Eu também corri do terremoto de ontem.
181. Colhi framboesas e coloquei em caixas com palhas.
182. O carro roxo decolou feito um avião.
183. Jardim Zaíra é um bairro chique no Chile.
184. Joguei palha no fundo do caixote.
185. O rouxinol brinca com o jasmim.
186. A girafa Zuzu gira em torno do ramalhete.
187. Comprei um jipe malhado de verde e vermelho.
188. O chumbo é um elemento metálico azulado.
189. Unte a frigideira antes de colocar o cozido.
190. Isso não é assunto para jumento.
191. Use tomilho para cozinhar os mantimentos.
192. A agulha furou o colchão da dona Zuleika.
193. Ouvir-lhe me faz sentir uma moleza absurda.
194. Nunca junte as pontas da mesma cor.
195. Finja interesse pelo dinheiro alheio.
196. A buzinha disparou com o bombardeio.
197. A colcha de retalhos será suntuosa.
198. A sunga azul chegou na loja.
199. O eu trunfo foi ganhar um carro roxo.
200. Sorvete de baunilha é bom com semente de linhaça.

Apêndice B

Lista de Locuções da Base de Fala Feminina

1. Londres quase cobriu a primeira Guerra da Chechênia.
2. Ele não é do gênero que namora filha de ninguém.
3. Guimarães anunciou o recuo, mas criticou a Executiva do partido.
4. Acusado de assassinar a mãe se entrega à polícia.
5. A fé é a raiz de todas as orações.
6. As duas tomam conta da mãe abertamente.
7. Globais preocupados com a silueta lotam agenda de mago das vitaminas.
8. Após reagir, os ladrões teriam lhe dado vários golpes de facão.
9. A família está preocupada e a mãe adoeceu.
10. Queria saber também onde posso obter informações sobre morar fora do Brasil.
11. Mãe e filhos ficaram juntos no fundo do plenário.
12. Família de mãe assassina pede perdão a negros.
13. Mas antes, Cristo lhe perguntou se O amava.
14. A história de quatro crianças que enterram a mãe no próprio porão.
15. Minha mãe esboçou resistência, mas eu aceitei prontamente.
16. E a rainha mãe passa a visitar uma casa de apostas.
17. Mostram cães, gatos, parentes, amigos e prostitutas.
18. Durante o processo, ele voltou a te ameaçar?
19. Secretaria vacina cães de Camboriú contra a raiva.
20. Seu pai é advogado e sua mãe, química.
21. Técnico do Milan é o melhor da Itália.
22. É que não tens mãe, nem futuro, nem rosto.
23. Uso de micro pesa na escolha da escola.
24. Entre os demitidos estão quatro majores, quatro capitães e dois capelães.
25. Agora eu queria apresentar minha mãe aos senhores.
26. Os presos terão contato humano freqüente com guardas, psicólogos e capelães.
27. Até que ponto sua mãe influenciou sua carreira?
28. No trecho abaixo, Guimarães Rosa descreve as veredas do sertão.
29. Mistérios da vida, dos cães e dos carros.
30. Sérvios atacam encrave na Bósnia e aviões da Otan fazem advertência.
31. Araci tinha vergonha de ser mãe solteira.
32. Mas minha mãe era cristã e não queria música em casa.
33. No momento do crime, a mãe não estava em casa.
34. Só Caetano pra rimar mãe com champanhe!
35. Minha mãe cheira e meu pai também.
36. Mas é verdade que minha mãe não toma banho de chuveiro.
37. Guimarães ainda não havia sido preso ontem à tarde.

38. O aparelho emite ondas ultrassônicas que afastam os animais.
39. Com a mãe o furo é mais em cima.
40. Segundo ele, a mãe chegou de helicóptero durante a tarde.
41. Mas se for uma mãe amorosa e assassina, as coisas se complicam.
42. Pelo que estou informado, mãe só tem uma embora todas a tenham.
43. Sempre fizemos tudo juntas, sem os limites de mãe ou tia.
44. Imitava a mãe, a irmã e os tios.
45. Mãe gosta mesmo é de uma foto de sua cria.
46. Sua mãe, Lúcia Ferreira, foi atingida por quatro tiros.
47. Se a mãe não estranha aqueles sadomasoquismos todos?
48. E Machado de Assis, Guimarães Rosa, Euclides da Cunha?
49. As duas mães negam que seus filhos tivessem envolvimento com o tráfico.
50. Justiça condena mãe por tirar filha da escola.
51. Só a mãe aceitava a nossa união.
52. As mães vêem pontos em comum nos desaparecimentos.
53. Eu cutucava minha mãe: por que não posso desfilar?
54. Erguida e ornada por uma bela tiara de brilhantes.
55. Perderam ou perderão a mãe ou o pai ou os dois.
56. Na capa, Lula aparece com o olho esquerdo roxo e avermelhado.
57. As ações foram ótimas opções neste início de ano.
58. Mãe enterra filho vivo em casa e é presa.
59. O presidente da Febraban observa que as convenções partidárias acontecerão somente em abril.
60. Mamãe embaixatriz foi acompanhar a formatura do filho Luiz em seu endereço favorito.
61. Com Guimarães Rosa, ou se faz assim, ou não se faz.
62. Desanimado de regenerar a mãe, o poeta a adota, como filha.
63. A Alemanha já baniu quatro organizações neonazistas e xenófobas nos últimos dois anos.
64. Sua mãe sonhava em ter uma filha.
65. Juntos, compartilhamos a agonia e êxtase da busca.
66. Mãe sempre sabe se o filho está com fome.
67. Como nos velhos e bons tempos, Eunícia Guimarães vai de frasqueira.
68. Para Túlio, só instruções ilustradas evitariam dúvidas sobre a colocação.
69. Esse homem que fez tudo isso é o Capitão Guimarães.
70. Helicópteros e cães policiais participam da caçada humana.
71. Foram presos também sua mãe e dois irmãos.
72. Ele vai governar com a mãe do Tancredo e as netas do Juscelino.
73. Alheio à crise do país, o turismo em Cuba vai bem.
74. O meia Nélio foi atingido em joelho pelo zagueiro Cléber.
75. As duas viajam em companhia do pai ou da mãe.
76. Segundo Arraes, isso é invenção dos adversários.
77. Com problemas sérios de saúde, vive com a mãe.
78. Até ontem, patrões e empregados não tinham chegado a nenhum acordo.
79. As pessoas que não olham nos olhos me afligem.
80. Faria tudo, menos matar o pai ou roubar a mãe.
81. As duas seleções femininas de vôlei voltam a jogar amanhã em Osaka.
82. Para elas, Émerson matou o pai para defender a mãe.

83. Afora a paixão por aviões, sempre gostou de carros.
84. Se não for do seu tempo, pergunte a sua mãe.
85. Extrair óvulos maduros de uma mulher é desconfortável e arriscado.
86. Também puxou a mãe nos defeitos: nenhum mais notório que a misoginia.
87. São paraguaios e bolivianos que vêm comprar os botijões do lado de cá.
88. Seu pai era sérvio, sua mãe muçulmana.
89. Mas meus pai, minha mãe e irmãos são muçulmanos.
90. Segundo Reze, a educação é inócua para conter o ágio.
91. Os dois teriam se xingado e trocado empurrões.
92. Alguns dos registros da hospedaria mostram as regiões para onde eles foram encaminhados.
93. A volta à casa de mamãe e papai inclui também namorada nova.
94. Aílton Guimarães Júnior visitou o pai anteontem e ontem.
95. Aqui o filho chora e a mãe não ouve.
96. A mãe, em estado de choque, ainda não foi ouvida.
97. Mas os olhos de Parreira só enxergam o futebol de Branco.
98. O relacionamento era melhor com sua mãe ou com seu pai?
99. O senhor se relaciona melhor com sua mãe ou com seu pai?
100. Romário abraçou a mãe, Manuela, e acordou o pai.

Referências Bibliográficas

Ali, A. M. A.: van Spiegel, J. V. Acoustic-Phonetic Features for the Automatic Classification of Stop Consonants. IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 8, November, 2001.

Ali, A. M. A.; van Spiegel, J. V.; Mueller, P. Robust Classification of Stop Consonants Using Auditory-Based Speech Processing. Proceeding of the International Conference on Acoustic, Speech and Signal Processing. Vol. 1, pp. 81-84, Salt Lake City, USA, 2001.

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

Araújo, A. M. L. Jogos Computacionais Fonoarticulatórios para Crianças com Deficiência Auditiva. Tese de Doutorado, Universidade Estadual de Campinas (Unicamp), Campinas, SP, 2000.

Araújo, A. M.; Violaro, F. Análise e parametrização de fricativas. *A*nais do XVIII Simpósio Brasileiro de Telecomunicações, Gramado–RS, 2000.

Aversano, G.; Esposito, A.; Marinaro, M. A new Text-Independent for Phoneme Segmentation. Proceedings of the IEEE International Workshop on Circuits and Systems, Volume 2, ISBN 0-7803-7150X, August, 2001, 516-519.

Bakis, R. Continuous speech word recognition via centisecond acoustic states. In Proceedings ASA Meeting, Washington , DC, USA, April, 1976.

Baum, L. E.; Petrie, T. Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat., Vol. 37, pp. 1554-1563, 1966.

Boonsuk, S.; Punyabukkana, P.; Suchato, A. Phone Boundary Detection using Selective Refinements and Context-dependent Acoustic Features. INTERSPEECH, August, 2007, Antwerp, Belgium.

Chen, S.; Gopalakrishnam, P.; Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In: DARPA Speech Recognition Workshop, 1998.

Deller, J. R.; Proakis, J. G.; Hansen, J. H. L. Discrete-Time Processing of Speech Signals. Englewood Cliffs, N. J.: Prentice-Hall, 1993.

Demuynck, K. and Laureys, T. "A comparison of different approaches to automatic speech segmentation". Proceedings of the 5th International Conference on Text, Speech and Dialogue, Brno, Czech Republic, September, 2002.

Dickson, D. Acoustic study of nasality. Journal of Speech and Hearing Research. Vol. 5, No. 2, pp. 103-111, 1962.

Espy-Wilson, C. Y.; Boyce, S. E.; Jackson, M.; Narayanan, S.; Alwan, A. Acoustic Modeling of American English [r]. Journal of the Acoustical Society of America, vol. 108, pp. 343-356, 2000.

Fant, G. Acoustic theory of speech production. The Hague: Mouton, 1960.

Ficker, L. B. Produção e percepção das plosivas do português brasileiro: estudo fonético-acústico da fala de um sujeito com deficiência auditiva. Tese de Doutorado, Pontifícia Universidade Católica (PUC/SP), São Paulo, SP, 2003.

Figueira, L. and Oliveira, L. C. Comparison of Phonetic Segmentation Tools for European Portuguese. PROPOR 2008, International Conference on Computational Processing of Portuguese Language, 8-10th September, Aveiro, Portugal, 2008.

Figueiredo, F. L. Segmentação Automática e Treinamento Discriminativo Aplicados a um Sistema de Reconhecimento de Dígitos Conectados. Tese de Mestrado, Universidade Estadual de Campinas (Unicamp), Campinas, SP, 1999.

Fujimura, O. Analysis of Nasal Consonants, Journal of Acoustical Society of America, Vol. 34, No. 12, pp 1865-1875, 1962(a).

Fujimura, O. Formant-Antiformant structure of Nasal Murmurs, Proceedings of the Speech Communication Seminar, Vol 1, Stockholm: Royal Institute of Technology, Speech Transmission Laboratory, pp 1-9, 1962(b).

Fukada, T.; Aveline, S.; Schuster, M.; Sagiska, Y. Segment boundary estimation using recurrent neural networks", Proceedings of EUROSPEECH'97, pp. 2839-2842, 1997

Glass, J. R.; Zue, V. W. Multi-level segmentation of continuous speech. Proceedings on the International on Acoustic, Speech and Signal Processing, pp. 429-432, 1988.

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

Hattori, S.; Yamamoto, K. F. Nasalization of vowels in relation to nasals. *Journal of Acoustical Society of America*. Vol. 30, No. 4, pp.267-274, 1958.

Honsom, J. P. Automatic phoneme alignment based on acoustic-phonetic modeling. International Conference on Spoken Language Processing. Boulder, CO, USA, Vol. I, pp. 357-300, 2003.

House, A.; Stevens, K. Analog studies of the nasalization of vowels. Journal of the Speech and Hearing Disorders. Vol. 21, No. 2, pp. 218-232, 1956.

Hsieh, C. T.; Su, M. C.; Lai, E. A segmentation method for continuous speech utilizing hybrid neuro-fuzzy network. Journal of Information Science and Engineering, Vol. 15, No. 4, pp. 615-628, 1999.

Huang, X. D.; Ariki, Y.; Jack, M. A. Hidden Markov Models for Speech Recognition. Edinburgh University Press, 1990.

Huang, X. D.; Jack, M. A. Hidden Markov Modelling of Speech based on a Semicontinuous Model. Eletronics Letters, No. 24, Vol. 1, January, 1988(a).

Huang, X. D.; Jack, M. A. Performance Comparison between Semicontinuous and discrete Hidden Markov Models of Speech. Eletronics Letters, No. 24, Vol. 3, pp. 149-151, February, 1988(b).

Jafiri, S.; Pastor, D.; Rosec, O.; Cooperation between global and local methods for automatic segmentation of speech synthesis corpora. INTERSPEECH, September, 2006, Pittsburgh, USA.

Juneja, A. Speech Recognition Based on Phonetic Features and Acoustic Landmarks. Ph.D. Thesis, University of Maryland, College Park, USA, 2004.

Juneja, A.; Espy-Wilson, C. Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. International Joint Conference on Neural Networks, 2003.

Kent, R. D.; Read, C. The acoustic analysis of speech. San Diego: Singular Publishing, 1992.

Keshet, J.; Shalev-Shwartz, S.; Singer, Y.; Chazan, D. Phoneme Alignment Based on Discriminative Learning. INTERSPEECH, September, 2005, Lisbon, Portugal.

Li, T. H. Discrimination of time series by parametric filtering. Journal of the American Statistical Association, Vol. 91, pp. 284-293, 1996.

Li, T. H.; Gibson, J. D. Discrimination analysis of speech by parametric filtering . Proceedings on IEEE conference on Information Science Systems. Princeton, NJ, pp. 575-580, march, 1994.

Li, T. H.; Gibson, J. D. Speech Analysis and Segmentation by Parametric Filtering. IEEE Transactions on Speech and Audio Processing, Vol. 4, No. 3, may, 1996.

Lisker, L.; Abramson, A. A cross language study of voicing in initial stops: acoustical measurements. Word, No. 20, Vol. 3, pp. 384-422, 1964.

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

Maciel, R. C. V. Melhoria da Qualidade de Sinais de Fala Degradados por Ruído Através da Utilização de Sinais Sintetizados. Tese de Mestrado, Universidade de São Paulo, Escola Politécnica, São Paulo, SP, 2003.

Mitchell, C.; Harper, M.; Janieson, L.; Using explicit segmentation to improve HMM phone recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1995, vol. I, pp. 229-232.

Morais, E. S.; Vieira, J. M.; Arantes, P.; MATTE, A. C. Metodologias para Projeto e Aquisição de uma Base de Dados Lingüísticos para Treinamentos e Avaliações de Sistemas de Reconhecimento de Fala. In: III TIL - Workshop em Tecnologia da Informação e da Linguagem, 2005, São Leopoldo. III TIL - Workshop em Tecnologia da Informação e da Linguagem, 2005.

Murthy, H. A. The real root cepstrum and its application to speech processing. National Conference on Communication, IIT Madras, Chennai, India, pp. 180-183, January, 1997.

Murthy, H. A.; Yegnanarayana, B. Formant extraction from minimum phase group delay function. Speech Communication, Vol. 10, pp. 209-221, 1991.

Nagarajan, T.; Kamakshi, V.; Murthy, H. M. The minimum phase signal derived from the magnitude spectrum and its applications to speech segmentation. 6th Biennial Conference of Signal Processing and Communications, july, 2001.

Nagarajan, T.; Murthy, H. M. Group Delay based Segmentation of Spontaneous Speech into Syllable-like units. EURASIP Journal of Applied Signal Processing, Vol.17, pp.2614-2625, 2004.

Nagarajan, T.; Murthy, H. M.; Hedge, R. M. Segmentation of speech into syllable-like units. 8th Conference on Speech Communication and Tecnology (EUROSPEECH), Geneva, Switzerland, September, 2003.

Nakagawa, S.; Hashimoto, Y. A Method for Continuous Speech Segmentation using HMM. 9th International Conference on Pattern Recognition , Rome, Italy, 1988, pp. 960-962.

Niyogi, P. Distinctive feature detection using support vector machines, International Conference on Acoustics, Speech and Signal Processing, pp. 425–428, 1998.

Picone, J. W., Signal Modeling Techniques in Speech Recognition. Proceedings of the IEEE, 81(9):1215-1223, 1993.

Pruthi, T. Analysis, Vocal-tract Modeling and Automatic Detection of Vowel Nasalization. Ph.D. Thesis Proposal. University of Maryland, College Park, USA, 2006.

Pruthi, T. and Espy-Wilson, C. Y. “Acoustic Parameters for the Automatic Detection of Vowel Nasalization. INTERSPEECH, August, 2007, Antwerp, Belgium.

Pruthi, T.; Espy-Wilson, C. Y. Automatic Classification of Nasals and Semivowels, 15th International Congress of Phonetic Sciences (ICPhS) 2003, Barcelona, Spain, August 2003.

Rabiner, L. R.; Juang, B. H. Fundamentals of Speech Recognition, Prentice Hall, 1993.

Rabiner, L. R.; Schafer, R. W. Digital Processing of Speech Signals. Prentice Hall, New Jersey, 1978.

Rubio, A. J; Reilly, R. G. Preliminary results on speech signal segmentation with recurrent neural networks. Proceedings of EUROSPEECH'95, pp. 2197-2200, 1995.

Russo, I. C. P.; Behlau, M. Percepção da fala: análise acústica do português brasileiro. São Paulo: Lovise, 1993.

Selmini, A. M.; Violaro, F.; Acoustic-Phonetic Features for Refining Automatic Speech Segmentation. In: Proceedings of the INTERSPEECH 2007, Antwerp, Belgium, August, 2007.

Sethy, A.; Narayanan, S.; Refined Speech Segmentation for Concatenative Speech Synthesis. Proceedings of the International Conference on Spoken Language Processing. September, 2002, pp. 16-20, Dever, Colorado, USA.

Silva, A. H. P.; Albano, E.; Brazilian Portuguese Rhotics and the Phonetics/Phonology Boundary. In: Proceedings ICPh'99, 1999, São Francisco, University of California, v. 3, p. 2211-2214.

Suh, Y.; Lee, Y. Phoneme segmentation of continuous speech using multilayer perceptron. Proceedings of ICSLP'96, pp. 273-275, 1996.

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

The HTK Book. Cambridge University Engineering Department, 2006.

Toledano, D. L. Neural network boundary refining for automatic speech segmentation. Proceedings of the International Conference on Acoustic Speech and Signal Processing. Istanbul, Turkey, june, 2000.

Toledano, D. L.; Gómez, L. A. H.; Grande, L. V. Automatic Phonetic Segmentation. IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 6, November 2003.

Toledano, D. L.; Gómez, L. A. H.; Grande, L. V. Trying to mimic human segmentation os speech using HMM and fuzzy logic post-correction rules. Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, Jenolan Caves, Australia, 1998, pp. 207-212.

van Hemert, J. P. Automatic Segmentation of Speech. IEEE Transactions on Signal Processing. Volume 39, Issue 4, April 1991, 1008-1012.

Vapnik, V. The Nature of Statistical Learning Theory, Springer Verlag, 1995.

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

Vieira, M. N. Módulo frontal para um sistema de reconhecimento automático de voz. Tese de Mestrado, Universidade Estadual de Campinas (Unicamp), Campinas, 1989.

Wang, L.; Zhao, Y.; Chu, M.; Zhou, J.; Cao, Z. Refining Segmental Boundaries for TTS Database Using Fine Contextual-Dependent Boundary Models. Proceedings of the International Conference on Acoustics, Speech and Signal Processing. May, 2004, Vol. 1, pp. 641-644. Montreal, Canada.

Witkin, A. P. Scale-space filtering: a new approach to multi-scale description. Proceedings on International Conference on Acoustic, Speech and Signal Processing, 1989.

Yared, G. F. G.; Método para a Determinação do Número de Gaussianas em Modelos Ocultos de Markov para Sistemas de Reconhecimento de Fala Contínua. Tese de Doutorado, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, 2006.

Yared, G. F. G.; Violaro, V.; Selmini, A. M.; HMM Topology in Continuous Speech Recognition Systems. 6th International Telecomunications Symposium (ITS2006), September, 2006, Fortaleza, Brazil, ISBN: 85-89748-04-9.

Ynoguti, C. A. Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov. Tese de Doutorado, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, 1999.

Zhao, Y.; Wang, L.; Chu, M.; Zhou, J.; Cao, Z. Refining Phoneme Segmentations Using Speaker-Adaptative Context Dependent Boundary Models. INTERSPEECH, September, 2005, Lisbon, Portugal.