CONSTRUÇÃO DE BANCO DE UNIDADES PARA SÍNTESE DA FALA POR CONCATENAÇÃO NO DOMÍNIO TEMPORAL

Vagner Luis Latsch

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

_____

Prof. Sergio Lima Netto, Ph.D.

_____

Prof. Marcio Nogueira de Souza, D.Sc.

_____

Prof. Luiz Pereira Calôba, Dr.Ing.

RIO DE JANEIRO, RJ – BRASIL

ABRIL DE 2005

LATSCH, VAGNER LUIS

Construção de Banco de Unidades para Síntese da Fala por Concatenação no Domínio Temporal [Rio de Janeiro] 2005

VIII, 125 p.29,7 cm (COPPE/UFRJ, M.Sc. Engenharia Elétrica, 2002)

Tese – Universidade Federal do Rio de Janeiro, COPPE

1. Síntese da fala

I. COPPE/UFRJ II. Título (série)

Dedico esta tese à minha filha Luisa.

AGRADECIMENTOS

Agradeço primeiramente à minha mãe e ao meu pai, por terem sido o início de tudo, e pelo apoio ao longo deste trabalho.

Agradeço também ao grupo do laboratório de processamento da fala da Universidade do Porto, por terem sido fundamentais no início desta caminhada insistente em trabalhar com processamento da fala.

Ao prof. Sergio Lima Netto, pela paciência e orientação nos momentos em que este trabalho parecia estar perdido.

Aos amigos Paulo Brandão e Patrícia, pela atenção dedicada aos preparatórios do nascimento da Luisa, que nasceu quase conjuntamente com este trabalho.

Ao amigo José Augusto, pela força nas horas de vacilo.

E agradeço à minha esposa Adriana, por último, porém primeira, pela compreensão de meu isolamento nestes últimos dias e pelo apoio ao longo de todo este trabalho.

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

CONSTRUÇÃO DE BANCO DE UNIDADES PARA SÍNTESE DA FALA POR CONCATENAÇÃO NO DOMÍNIO TEMPORAL

Vagner Luis Latsch

Abril/2005

Orientador: Sergio Lima Netto
Programa: Engenharia Elétrica

Este trabalho aborda uma das etapas de desenvolvimento de um sistema conversor texto-fala, que ocupa grande parte do tempo de desenvolvimento e que está fortemente ligada à qualidade da fala sintetizada. Trata-se da determinação do inventário de unidades, da gravação, da segmentação e do recorte de segmentos do sinal de fala a serem concatenados no domínio temporal. Neste processo, as etapas que demandam mais trabalho manual são a inserção de marcas de pitch para o algoritmo TD-PSOLA e a segmentação do sinal ao nível fonético. Estas tarefas foram abordadas particularmente, na busca por um método semi-automático, na qual se propõe um método alternativo, de baixo custo e preciso para a detecção das marcas de pitch, utilizando um sinal auxiliar de um microfone de contato e para a segmentação do sinal ao nível fonético, é revisitado um método de segmentação acústica, que apresenta resultados razoáveis quando são impostas algumas condições.

Foi desenvolvido um aplicativo, chamado de *Editor*, que permite agregar toda a etapa de desenvolvimento do banco em um único ambiente e onde os métodos semi-automáticos foram implementados. O resultado desta tese é uma metodologia geral para a construção do banco de unidades utilizando o *Editor*.

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

BUILDING UNITS FOR SPEECH CONCATENATION SYNTHESIS IN TEMPORAL DOMAIN

Vagner Luis Latsch

April/2005

Advisor: Sergio Lima Netto
Department: Electrical Engineering

This work deals with one stage in the development of a text-to-speech (TTS) system which demands a great amount of time and effort and which is strongly related to the final speech quality. Such stage is the determination of the database of speech units for the TTS system. This whole process includes recording, segmentation and labeling of speech units to be concatenated in the time domain.

In general, the most demanding steps are the determination of pitch marks for the TD-PSOLA algorithm and the speech segmentation at the phonetic level. For that matter, we propose a low-cost and precise method for determining the pitch marks utilizing an auxiliary signal obtained from a contact (throat) microphone. For the phonetic speech segmentation, we revise an algorithm for accustic segmentation which yields interesting results when proper operation conditions are imposed. A new software tool, the so-called Editor, was developed integrating all previously addressed techniques in a single environment. The result is a general methodology for developing a speech database for TTS systems using the Editor framework.

ÍNDICE

Capítulo 1

Introdução

1.1. Apresentação

Os conversores texto-fala, também conhecidos como TTS (*text-to-speech*), são sistemas que produzem fala sintética correspondente à leitura de um texto. Estes sistemas têm sido aplicados em diferentes áreas. Nas telecomunicações, por exemplo, vê-se a aplicação na consulta de *e-mails* por telefone e também em centros de atendimento eletrônico, na reprodução da leitura de menus e orientações. Outro exemplo é a aplicação ao princípio de acessibilidade para todos, onde estes sistemas podem permitir que deficientes visuais tenham acesso, a caixas eletrônicos, à Internet, etc. quando usados na leitura de telas do computador ou na leitura de *sites*. A título informativo, "*tramita no Congresso o Projeto de Lei nº 7.432/02, que propõe a alteração nos artigos 2º e 17º da Lei 10.098 (a Lei da Acessibilidade), que obriga os sites do governo a*

obedecerem à lei de acessibilidade facilitando a navegação de todos os internautas, inclusive dos portadores de necessidades especiais" (Jornal do Brasil 30 de março de 2004). Isto aponta para uma popularização dos sistemas de conversão texto-fala.

Os exemplos citados acima tratam da conversão do texto irrestrito em fala, ou seja, o texto não está restrito a um conjunto de palavras ou frases. Diferente, por exemplo, nos serviços de auxilio as listas telefônicas, que na pronúncia automática dos números telefônicos (soletrados), só há a ocorrência dos algarismos de 0 a 9. Neste caso, é intuitivo sugerir a concatenação dos sons correspondentes a cada algarismo para formar a pronúncia do número. No entanto, retornando ao caso do texto irrestrito, devido à enorme possibilidade de combinação de sons, ocorre a proposta do uso de segmentos menores, como sílabas, fones, ou outros recortes de som ainda menores, buscando minimizar a quantidade de sons necessários. Estes segmentos de som são chamados de unidade de concatenação.

É preciso levar em conta que as unidades de concatenação, sejam elas quais forem, estão sujeitas à variação, de acordo com a posição ocupada dentro de uma frase ou com a entoação aplicada. Por exemplo, no caso dos números de telefone a pronúncia do **2** no número **2**555-5555 será diferente no número 555**2**-5555. Assim, para manter uma entoação correta e natural, seria necessário considerar todas as variantes do número **2** como uma unidade de concatenação, ou usar um método que fosse capaz de modificá-las, principalmente em intensidade, duração e freqüência fundamental, que são os principais fatores para caracterizar a entoação, ou prosódia.

Atualmente existem sistemas de seleção automática de unidades que utilizam bancos com várias unidades diferenciadas. Em tempo de síntese, o sistema busca no banco uma unidade que melhor se adequar ao contexto fonético e prosódico. Este método requer pouca manipulação prosódica, porém o processo de busca e seleção automática de unidades não é trivial e além disso é necessário um extenso *corpus* previamente etiquetado.

Considerando um banco reduzido de unidades e um método que as manipule de acordo com a prosódia desejada, um método que se tornou popular devido à sua simplicidade

de implementação e baixo custo computacional, sendo usado em vários sistemas de conversão texto-fala atuais, é o TD-PSOLA (*time domain - pitch synchronous overlap and add*) (CHARPENTIER, 1989). Apesar de sua simplicidade, este método requer um grande esforço na preparação do banco das unidades de concatenação e possui uma capacidade de modificação limitada, o que exige em alguns casos algumas variantes de uma mesma unidade (DUTOIT, 1997).

De forma geral, a qualidade da fala sintética produzida está diretamente relacionada à qualidade do banco de unidades, ou seja, ao correto inventário de unidades, à qualidade do sinal gravado, ao correto recorte das unidades e à precisão na incorporação de outras informações necessárias ao método de concatenação e modificação. Isto principalmente para o método TD-PSOLA. Usualmente as unidades são identificadas e recortadas manualmente, por um profissional experiente, exigindo grande consumo de tempo. Em (KRAFT, 1992) os autores relatam que foram necessários 18 meses para a construção de uma banco de unidades para síntese do Alemão.

Além disso, considerando que o método TD-PSOLA opera de maneira síncrona com o pitch, este requer que as unidades possuam marcas de pitch precisas. Idealmente, as marcas de pitch devem coincidir com o instante de fechamento glotal (*glottal closure instant* - GCI), porém a detecção dos GCIs de maneira automática e precisa ainda é um problema com solução não trivial (DUTOIT, 1997). Por isso geralmente estas marcas são inseridas também de maneira semi-automática. Atualmente, muitos autores têm utilizado um equipamento chamado eletroglotógrafo, que monitora o movimento das cordas vocais, como uma fonte de sinal auxiliar para a marcação dos GCIs.

1.2. Objetivos

Geralmente o desenvolvimento de um sistema TTS por concatenação de unidades pode ser dividido em três partes: a extração de características lingüísticas do texto, como a seqüência fonética, os limites de palavras, frases, sílabas e informações de tonicidade; a determinação do padrão prosódico a partir destas informações do texto; e a síntese da

fala que a partir da seqüência fonética e do padrão prosódico determinado irá concatenar as unidades correspondentes e impor a prosódia à seqüência concatenada. Dada a dependência entre as partes, observa-se que inicialmente é preciso um sistema de síntese, até mesmo para validação e teste das outras partes.

Assim, assumindo o método TD-PSOLA como um método de manipulação prosódica eficiente e de baixa complexidade de implementação, para a concepção do sintetizador de fala o próximo passo é a construção do banco de unidades, que será abordada neste trabalho desde seu planejamento à sua implementação propriamente dita.

Conforme dito anteriormente, a construção do banco de unidades geralmente é feita de forma manual consumindo grande parte do tempo de desenvolvimento. Este processo precisa ser repedido ainda se for de interesse criar novas vozes, diferentes em sexo, idade, ou ainda típicas de uma região, que no caso do Brasil são muitas. Isto implica além de um consumo de tempo considerável um trabalho tedioso que justifica a busca por uma metodologia semi-automática. Cabe salientar que é considerado aqui semi-automático qualquer método que seja necessário a validação ou correção do resultado por parte do usuário.

Deste modo, inicialmente este trabalho se destina a coletar experiências anteriores auxiliando a etapa de planejamento. Em seguida, após a gravação dos trechos onde estão contidas as unidades, as duas fases que demandam mais tempo por parte do usuário são: a identificação da unidade dentro do trecho gravado, incluindo a marcação das fronteiras entre fones, e a marcação dos GCIs a serem usadas pelo TD-PSOLA.

Na busca por automatizar a primeira fase de segmentação e com enfoque em métodos não supervisionados, ou seja, que não precisem de corpus previamente etiquetados, neste trabalho será apresentado um método de segmentação baseado na clusterização de segmentos com características acústicas semelhantes. Por ser baseada somente em características acústica, a segmentação oferecida pelo método nem sempre corresponde às fronteiras fonéticas, porém será visto que a redução do intervalo de busca e o acréscimo da informação sobre sonoridade, resultam em um método semi-automático

satisfatório para a marcação de fronteiras fonéticas e localização do segmento a ser recortado.

Quanto à marcação dos GCIs nos segmentos, vários autores consideram o eletroglotógrafo como uma solução atrativa, porém é uma solução de custo alto. Na busca por uma alternativa de custo baixo, neste trabalho será apresentado um método de obtenção precisa dos GCIs que associa um sinal auxiliar obtido por um microfone de contato com um método clássico de detecção automática.

Por fim, este trabalho tem por o objetivo principal descrever uma metodologia, incluindo algumas soluções semi-automáticas, descritas anteriormente, para a construção futura de outras vozes.

1.3. Estrutura da tese

Esta tese está organizada em 5 capítulos. No Capítulo 1 é feita uma apresentação do assunto contendo as motivações e os objetivos deste trabalho. Considerando uma etapa de planejamento, na qual observa-se os sons presentes na língua e define-se o inventário de unidades.

No Capítulo 2 serão apresentados inicialmente alguns conceitos da fonética e fonologia considerados fundamentais, seguido de uma compilação de diferentes experiências obtidas em outros trabalhos a respeito da escolha do tipo de unidade, do inventário das unidades, da gravação do corpus de fala, da etiquetagem de segmentos fonéticos e do recorte.

Em busca de um método de etiquetagem não supervisionado ou semi-automático, no Capítulo 3 será apresentado e implementado um método de segmentação do sinal de fala a partir somente de características acústicas do sinal, baseado em um algoritmo de clusterização e quantização vetorial.

No Capítulo 4 será discutida a problemática de inserção de marcas de pitch nas unidades, coincidentes com os GCIs, a serem utilizadas pelo algoritmo TD-PSOLA. Serão também apresentados alguns métodos clássicos de detecção automática dos GCIs e proposto um método de detecção precisa e semi-automática utilizando um sinal auxiliar de um microfone de contato.

No Capítulo 5 será descrita uma metodologia para a obtenção semi-automática de um banco de unidades, utilizando um aplicativo desenvolvido para esta proposta, chamado de *Editor*, destinado ao usuário inexperiente. Procuraremos ainda validar os resultados apresentados ao longo da tese construindo um pequeno banco de unidades, como exemplo para uma metodologia geral de construção de um banco de unidades utilizando o *Editor*.

Finalmente no Capitulo 6 serão descritas as conclusões obtidas neste trabalho, onde serão apresentadas possíveis continuações do mesmo.

Capítulo 2

Banco de unidades para síntese por concatenação temporal

2.1. Introdução

Um banco de unidades para a síntese por concatenação consiste em um conjunto de segmentos obtidos de sinais de fala que serão concatenados em uma dada seqüência. Estes segmentos possuem particularidades associadas ao método de síntese que será utilizado. Por exemplo, no caso de unidades paramétricas, um conjunto de parâmetros é extraído de cada segmento de fala e em seguida armazenado no banco. No caso da síntese por concatenação de unidades temporais, abordada neste trabalho, os segmentos de fala usualmente são adicionados ao banco tal como são recortados do sinal de fala.

Como citado anteriormente, as unidades são geralmente identificadas e marcadas manualmente, por um profissional experiente e requer um certo nível de experimentação. Assim, as etapas de construção nem sempre seguem um roteiro cronológico. Por exemplo, ao final da construção do banco pode-se concluir que novas unidades são necessárias para melhorar a inteligibilidade da fala sintética ou que algumas unidades podem ser suprimidas de modo a reduzir o espaço em disco necessário, sem prejudicar a inteligibilidade.

De maneira a minimizar o caráter experimental da construção do banco, uma etapa de planejamento que antecede a gravação de frases e recorte de segmentos torna-se necessária. Neste caso, para este planejamento algumas perguntas precisam ser respondidas:

- Quais os sons produzidos na língua em questão e quais as suas características dinâmicas?
- Quais os tipos de segmentos que apresentam boa possibilidade de concatenação? sílabas? fones?
- Quais segmentos então serão gravados e recortados?
- De que maneira gravar os segmentos? Em frases faladas naturalmente?
- E em que região recortá-los?

Uma proposta interessante a ser citada é o método chamado de COC (*context oriented clustering*), que dispensa a necessidade de planejamento do banco de unidades, sendo este construído automaticamente (DUTOIT, 1997). Considerado como precursor dos métodos recentes de seleção automática de unidades, neste método as unidades de concatenação são determinadas e geradas automaticamente usando um corpus de fala natural previamente etiquetado.

Portanto, este capítulo se propõe a recolher experiências anteriores, tanto para o português quanto para outras línguas, de modo a oferecer um ponto de partida para que as perguntas acima possam ser respondidas. Assim, este capítulo foi estruturado da seguinte forma: na seção 2.2 será feita uma apresentação dos conceitos da fonética e da fonologia, considerados fundamentais, onde serão descritos os sons produzidos no português, incluindo exemplos de ocorrência, nomenclatura e características acústicas.

Neste capítulo e no restante da tese, os caracteres utilizados para transcrição fonética seguem o padrão IPA (*international phonetic association*). Na representação das unidades de concatenação, por não ser conhecida uma norma bem definida, será adotada posteriormente um padrão para este trabalho. Porém, em alguns casos específicos, na citação de outros trabalhos, será utilizada a mesma notação empregada pelos autores.

Na seção 2.3 será discutido o processo de transcrição fonética, ou seja, o processo de representação do som da fala, através de uma seqüência de símbolos específicos, que carregam informação acústica e articulatória do som a ser transcrito. Em seguida, na seção 2.4, serão apresentadas as unidades mais comumente usadas e algumas questões a serem consideradas na concatenação de segmentos que são relevantes para a determinação do tipo de unidade a ser escolhido. Na seção 2.5, serão apresentados inicialmente alguns métodos de especificação do inventário de unidades baseados em conceitos fonéticos e em seguida será apresentado brevemente o método COC de geração automática de unidades de síntese. A seção 2.6 contém algumas considerações sobre a gravação do *corpus* de onde serão recortadas as unidades e finalmente, na seção 2.7 serão apresentadas algumas considerações a respeito do local de recorte das unidades.

2.2. Conceitos de fonética e fonologia

A fonologia é o ramo da lingüística que estuda os sons constituintes da fala a partir seu aspecto funcional, ou seja, trata do papel exercido por eles no sistema de organização da fala, sem se preocupar com as propriedades de produção ou acústicas os sons (SIMÕES, 1999) .

A unidade básica de estudo da fonologia, ou fonêmica (SILVA, 2003), é o *fonema,* que é uma unidade abstrata restrita ao domínio psicológico e não físico e portanto não apresenta características acústicas. O que diferencia um fonema de outro é o seu papel distintivo dentro da língua. Por exemplo, para o português ***/p/*** e ***/b/*** representam fonemas diferentes, pois diferenciam palavras como ***pasta*** e ***basta*** (SIMÕES, 1999).

Os fonemas representam *classes de sons*, ou seja, dois sons diferentes entre si, mas que não representam papeis distintivos dentro da língua, pertencem a uma mesma classe. Por exemplo, o ***"r"*** da palavra ***porta,*** dependendo da região do Brasil, aparecerá pronunciado de diferentes maneiras diferentes, de modo que cada variante do fonema ***/r/*** constitui um *alofone* de ***/r/.*** Normalmente os fonemas são representados entre barras (/ /), enquanto que os alofones são representados entre colchetes ([ ]).

A realização física de um fonema, ou seja, um trecho de sinal acústico, é denominado de *fone*. Para cada fonema existe um número infinito de fones, sendo assim realizações acústicas pertencentes à mesma classe (SIMÕES, 1999).

A fonética é o ramo da lingüística que estuda os sons constituintes da fala de acordo com suas características articulatórias e acústicas. A fonética articulatória descreve os diversos sons da língua de acordo com a dinâmica (posição e movimentação) dos articuladores que constituem o trato vocal humano (lábios, língua, mandíbula, etc.).

Os diferentes modos de modificação do fluxo de ar permitem o estabelecimento de duas classes de sons: **consoantes** e **vogais**. As consoantes são produzidas com algum tipo de obstrução total ou parcial da passagem do fluxo de ar, podendo ou não haver fricção. Por outro lado, na produção das vogais o fluxo de ar não é interrompido, não havendo obstrução ou fricção. Alguns segmentos não têm características fonéticas tão precisas são denominados **glides** ou semivogais (SILVA, 2003).

Segmentos Consonantais: As consoantes podem ser descritas considerando tanto o *modo de articulação* quanto o *ponto de articulação*. O modo de articulação se refere aos diferentes graus de obstrução do fluxo de ar, podendo ser divididos nas seguintes categorias (SILVA, 2003):

- **Oclusiva ou Plosiva**: ocorre a obstrução total à passagem do fluxo de ar pela boca, seguida de uma liberação abrupta do fluxo retido. Ex.: **p**á, **t**á, **c**á, **b**ar, **d**á, **g**ol.

- **Nasal**: ocorre a obstrução total à passagem do fluxo de ar pela boca e o ar que vem dos pulmões dirige-se às cavidades nasal e oral. Ex.: **m**á, **n**ua, ba**nh**o.
- **Fricativa**: ocorre a obstrução parcial à passagem do fluxo de ar o que resulta na geração de um ruído de turbulência ou fricção. Ex.: **f**é, **v**á, **j**á, **r**ata, **Z**apata, **c**há, **s**apa.
- **Africada**: ocorre em duas fases, uma oclusiva seguida de uma fricativa. Para o português carioca ocorre por exemplo em **t**ia e **d**ia.
- **Vibrante (múltipla)**: ocorre quando um articulador móvel (a ponta da língua ou a úvula) bate repetidas vezes causando vibração. Ocorre por exemplo no dialeto (sotaque) paulista em **r**ua e **r**ato.
- **Lateral**: ocorre a obstrução do ar no centro do trato vocal, e conseqüentemente o fluxo é deslocado pelos lados desta obstrução. Ex.: **l**á, pa**lh**a, sa**l** (pronunciada no Sul do Brasil).
- **Tepe (vibrante simples)**: ocorre com uma batida do articulador móvel (ponta da língua) ocorrendo uma rápida obstrução do fluxo de ar pela boca. Ex.: ca**r**o, b**r**ava.

As vibrantes e laterais podem aparecer em outra nomenclatura como **líquidas** (CALLOU, 1990).

O lugar de articulação ou ponto de articulação, refere-se ao lugar do trato vocal onde ocorre a máxima obstrução à passagem do fluxo de ar. As principais ocorrências podem ser classificadas em:

- **Bilabiais**: são formadas a partir da constrição dos lábios superior e inferior. Ex.: **p**á, **b**oa,**m**á.
- **Labiodentais**: ocorre a constrição entre o lábio inferior e os dentes incisivos superiores. Ex.: **f**aca, **v**á.
- **Alveolares**: a constrição ocorre entre a ponta da língua e os alvéolos (porção do trato situada entre o palato e os incisivos superiores). Ex.: **s**apo, **l**ata, **n**ada, ca**r**a.
- **Velar**: a constrição ocorre entre o dorso da língua e o palato mole. Ex.: **c**asa, **g**ata.

- **Palatal**: são produzidas com o dorso da língua próximo à parte final do palato duro. Ex.: ba**nh**a, pa**lh**a.
- **AlveoPalatal**: o dorso da língua toca a área próxima aos alvéolos. Ex.: **t**ia e **d**ia no dialeto carioca.

Assim, os segmentos consonantais são classificados seguindo a seguinte notação:

(*modo de articulação* + *lugar de articulação* + *grau de vozeamento*).

Ex.: [p] Oclusiva bilabial desvozeada, [b] Oclusiva bilabial vozeada .

Segmentos vocálicos: No caso das vogais, não há obstrução do fluxo de ar no trato vocal. Estes segmentos são descritos considerando-se: a altura do corpo da língua (sentido vertical), a posição da língua em termos de avanço ou recuo (sentido horizontal) e arredondamento ou não dos lábios. Pela altura da língua as vogais podem ser classificadas em **altas**, **média-alta**, **média-baixa** ou **baixas**. Alguns autores referem-se à altura em termos de fechamento/abertura da boca, utilizando baixo=**aberta** e alta=**fechada**. Quanto a posição da língua, classifica-se em **anteriores, centrais e posteriores** e se houver ou não arredondamento dos lábios classificam-se em **arredondadas** ou **não-arredondadas** (SILVA, 2003).

Assim, na classificação das vogais, é comum a seguinte notação:

(*altura* + *anterioridade* + *arredondamento*).

E quando ocorre a nasalização da vogal, esta é indicada como último parâmetro.

Ex.: [i] vogal alta anterior não arredondada , [ũ] vogal alta posterior arredondada nasal.

Dada a classificação dos segmentos, cabe ressaltar que a descrição dos segmentos fonéticos que ocorrem na fala depende não só da língua, mas também do dialeto e do próprio *idioleto* (características próprias do falante quanto ao seu modo de falar). Na seção a seguir, serão identificados alguns segmentos que ocorrem no português, mais particularmente para o dialeto carioca. Serão também mostrados alguns exemplos de trancrição fonética ampla, ou seja, da transcrição que explicita somente os aspectos que não são condicionados pelo contexto ou por características específicas da língua ou dialeto (SILVA, 2003).

2.3. Transcrição fonética

Tendo em vista a necessidade de determinar quais segmentos considerar para a construção do inventário de unidades de concatenação, a seguir serão dadas algumas definições necessárias, sem a intenção de se aprofundar no assunto.

Como referência, no apêndice A são listados os principais segmentos consonantais e vocálicos que ocorrem com maior freqüência no português brasileiro, tipicamente no dialeto carioca, utilizando os símbolos IPA, seguidos de um exemplo ortográfico e da transcrição fonética ampla. Uma descrição completa e minuciosa dos segmentos que ocorrem no português brasileiro pode ser encontrada em (SILVA, 2003).

Em relação ao sistema vocálico do português, as vogais podem ser tônicas, pretônicas (antecedem a tônica) ou postônicas (precedem a tônica). Na transcrição fonética, as sílabas tônicas são marcadas por uma apóstrofe. Para a maioria dos falantes, as vogais postônicas finais são distintas das vogais tônicas e pretônicas e são pronunciadas como [ɪ,ə,ʊ]. Por exemplo, para a vogal postônica final na palavra "mato" ocorre a semivogal [ʊ], sendo a palavra transcrita por ['matʊ]. No caso da fala considerada informal, o mesmo pode acontecer para as vogais postônicas mediais. Por exemplo, a palavra "pérola" ocorre no estilo informal como ['pɛrʊlə] e no formal por ['pɛrolə].

Glides: Os glides ou semivogais, são segmentos que podem apresentar características fonéticas de segmentos consonantais ou vocálicos. Em português classificamos os glides como segmentos vocálicos. As vogais [ɪ] e [ʊ] (sons sem proeminência, equivalentes aos segmentos [i] e [u]) ocorrem no português como glides dentro dos ditongos, podendo também ocorrer em posição final átona, como por exemplo nas palavras "safar**i**" [sa'farɪ] e "pat**u**" ['patʊ]. Alguns autores utilizam respectivamente os símbolos [j] e [w] para indicar as semivogais.

Ditongos: os ditongos são geralmente tratados como uma seqüência de segmentos. Um deles é interpretado como uma vogal e o outro por uma semivogal (vogal assilábica ou glide). Para SILVA, (2003) um ditongo é uma vogal que apresenta mudanças continuas

dentro de um percurso vocálico, podendo ser identificado pelos segmentos inicial e final do movimento. Por exemplo, na palavra “pais” ocorre o movimento contínuo e gradual entre as posições articulatórias vocálicas de [a] até [ɪ]. Neste caso os dois segmentos ocupam uma única sílaba, onde [a] é o núcleo da sílaba e o outro segmento [ɪ] é assilábico e corresponde ao glide. O movimento articulatório de um ditongo se difere de duas vogais em seqüência, chamadas de **hiato**, como por exemplo na palavra “país” onde as vogais ocorrem em sílabas diferentes. A transcrição fonética ampla destas palavras seria respectivamente: ['paɪs] e [pa'is]. No ditongo, as seqüências vogais-glides são chamadas de ditongos decrescentes, por outro lado, uma seqüência glide-vogal é chamada ditongo crescente, como por exemplo “saudade” [saʊ'dadʒɪ] e “quieto” [kɪɛ'tʊ] respectivamente. No apêndice A também são descritas as principais ocorrências dos ditongos.

Ocorrem casos em que uma transcrição fonética rigorosa, faz diferenciação entre alguns segmentos fonéticos que, sob o ponto de vista da determinação de um inventário de unidades, talvez não seja preciso diferenciá-los. Um teste de inteligibilidade poderá confirmar essa suposição. Alguns casos deverão ser considerados, como:

- Os **tritongos**, que podem ocorrer na presença das consoantes [k] e [g] quando precedidas por um ditongo, podem ser transcritos, por exemplo, como nas palavras “quais” ['kʊaɪs] e “iguais” [i'gʊaɪs].
- A seqüência consoante lateral-glide, por exemplo, as palavras “julho” e “Julio” podem ser transcritas da mesma forma por ['ʒʊlɪʊ], consideradas portanto como homófonas.
- O dígrafo “nh”, que ocorre sempre em posição intervocálica, pode ser transcrito como um segmento vocálico [i] nasalizado [ĩ]. Isto ocorre na maioria dos dialetos do português brasileiro. A palavra “banha” por exemplo pode ser transcrita por ['bãĩə].

- Para alguns autores, quando ocorre a vocalização da lateral, como no caso das palavras “calda” e “mural”, a lateral é transcrita como [ʊ], e portanto as palavras seriam transcritas por [mʊ'ɾaʊ] e ['kaʊdə].

- No caso da palavra “cueca” a transcrição poderá ser [kuˈɛkə] e no caso da palavra “seqüela” [seˈkʊɛlə].

Um caso particular a ser considerado é um fenômeno chamado de **êpentese**. Neste fenômeno, geralmente ocorre um pequeno trecho vocálico entre uma seqüência oclusiva-vibrante (e provavelmente laterais também) quando estão na mesma sílaba. Isto significa que na pronúncia de uma sílaba como “bra”, por exemplo, na verdade se pronuncia um pequeno “a” entre o “b” e o “r”. A vogal inserida entre as consoantes, conhecida por vogal epentética , é breve e espectralmente possui a qualidade do núcleo da sílaba (SOLEWICZ. 1994). Outro exemplo é a palavra “adstringente”, que pode ser transcrita por [adistɾĩˈʒẽˈtʃɪ], onde é explicitada a vogal epentética [i]. O mesmo surgimento da vogal epentética pode ser considerada no encontro consonantal entre [p] e [n] ou [s], por exemplo as palavras “pneu” e “psicologia”, que seriam transcritas por [piˈneʊ] e [pisikoloˈʒiə].

Até o momento, os exemplos citados foram de palavras isoladas, porém quando as palavras são faladas naturalmente dentro de um contexto, ocorre um fenômeno de **coarticulação** entre os segmentos vizinhos, que tendem a se modificar. Por exemplo, a frase “...fazemos uso...” pode ser transcrita como [...faˈzẽmʊz ˈuzʊ...]. Em alguns casos ocorre a transformação das vogais átonas no final de palavras, em vogais átonas não finais (KAFKA, 2002), como por exemplo na frase “que me leve” [ˈkɪ ˈmɪ ˈlɛvɪ] que se transforma em uma única palavra fonológica [kimiˈlɛvɪ]. Neste caso as vogais átonas finais [ɪ] se transformaram em [i]. Este fenômeno de ressilabação é objeto de estudo da fonologia, ou fonêmica, na qual são estudados os processos dinâmicos da língua. KAFKA, (2002) aborda este assunto com aplicação em um sistema de síntese de fala.

2.4. Definição das unidades de concatenação

O processo de definição das unidades representa um compromisso com a naturalidade e inteligibilidade requerida. Por exemplo, a palavra como uma unidade de concatenação, agrupa uma grande quantidade de articulações, apresenta boa conexão, no entanto existem em grande quantidade e apresenta ampla variação (prosódica) com o contexto. Assim, algumas propriedades básicas precisam ser consideradas na definição das unidades (DUTOIT, 1993):

- devem agrupar tantos efeitos articulatórios quanto possíveis;
- sua quantidade e largura devem ser tão pequenas quanto possíveis;
- dada a capacidade restrita de suavização na etapa de concatenação, elas devem ser facilmente conectáveis (conectividade), ou seja, as descontinuidades acústicas na vizinhança precisam ser minimizadas e
- a diferença acústica entre a unidade disponível e a unidade necessária no contexto da concatenação (representatividade) também precisa ser minimizada.

É fato que quanto maior a quantidade de unidades, retiradas de diferentes contextos, maior é a naturalidade e inteligibilidade do sistema. Por exemplo, HUNT, (1996) utiliza várias instâncias de uma mesma unidade como segmentos em potencial, cada um com um diferente contexto fonético e prosódico. Com isso, o processamento de sinal necessário para corrigir as características prosódicas das unidades foi reduzido drasticamente. A conseqüência de se utilizar um banco contendo várias ocorrências da mesma unidade (no caso aproximadamente 115.000 unidades), implica na necessidade de um processo de busca e seleção automática de unidades que não é trivial. Além disso, a alta qualidade destes sistemas é conseguida com um extenso *corpus* de fala, o que implica na necessidade de etiquetagem automática (DUTOIT, 1993).

No entanto, não quer dizer que com um número reduzido de unidades não se consiga naturalidade e inteligibilidade razoáveis e até muito boas. Neste caso é necessário um esforço maior na definição do inventário e recorte de unidades, considerando tanto fatores fonológicos da língua como aspectos acústicos e ainda um eficiente processo de concatenação (DUTOIT, 1993). É preciso levar em conta também que quanto menor a

unidade, mais pontos de concatenação existirão no momento da síntese exigindo maior desempenho do método de concatenação para uma boa qualidade de síntese.

Algumas unidades tradicionalmente utilizadas, são descritas a seguir, a começar pelos **fones,** que são unidades diretamente ligadas á fonética. Foram unidades de concatenação desconsideradas no passado, devido ao efeito de modificação que a unidade sofre de acordo com o contexto fonético onde ela se insere (coarticulação), o que requer várias amostras da mesma unidade em diferentes contextos (alofones). Porém, recentemente, com a introdução dos algoritmos de seleção automática de unidades, como o citado acima em (HUNT, 1996), os fones têm sido considerados novamente como possíveis unidades de síntese.

Os **difones** são unidades ou segmentos que englobam somente uma transição entre um fone e outro e tem em seus limites os pontos de máxima estabilidade espectral, ou de mínima dinâmica articulatória (SOLEWICZ. 1994). Um conjunto de difones (da ordem de 1600 para o francês) consegue capturar muitas transições fonéticas e por isso são amplamente usados. Por outro lado, eles englobam somente em parte os vários efeitos coarticulatórios da língua falada, que geralmente afetam um fone inteiro (DUTOIT, 1993). Na verdade, a estabilidade espectral é uma noção abstrata que neste caso não corresponde à realidade (SOLEWICZ, 1994). Tais efeitos foram especialmente registrados quando algum fone transiente, por exemplo uma consoante líquida e semivogais (pior de todas) estão interconectadas (DUTOIT, 1993).

Esta desvantagem dos difones, levou a proposição de uma unidade. chamada **trifone.** que engloba um fone inteiro e suas transições à direita e à esquerda e que constituem um complemento aos difones podendo solucionar os efeitos dinâmicos citados (DUTOIT, 1993). Associar difones, trifones e tetrafones (engloba 4 transições) para cobrir um número limitado de efeitos contextuais, chama-se de técnica por polifones. BIGORNE et al., (1993) na construção de um banco de polifones, para o francês, (com 1290 difones mais 1047 unidades mais longas), relata uma redução no erro de inteligibilidade de 20% quando foram adicionadas unidades mais longas do que os difones. BARBOSA, (1999) utiliza um banco de polifones de aproximadamente 2500 unidades, entre difones a tetrafones, para o português brasileiro.

A **sílaba** já foi uma unidade apontada como lingüisticamente atraente como uma unidade de concatenação (SOLEWICZ, 1994). Em seu projeto de fim de curso, SOLIMAR, (2001) concluiu que a qualidade segmental do sintetizador desenvolvido utilizando a sílaba como unidade básica para o português brasileiro foi menor que a qualidade de um sintetizador baseado em difones, onde a perda de qualidade foi atribuída à perda de coarticulação entre sílabas.

Baseado no conceito fonológico de formação da sílaba, foi proposto a divisão da sílaba em duas partes parcialmente superpostas, as **demissílabas**, na qual o pico silábico pertence a ambas as partes. Por exemplo, a sílaba *tar* na palavra *tar.de* possui uma demissílaba inicial *ta* e uma demissílaba final *ar*. Uma conseqüência desta divisão é que não há relação de dependência entre as duas partes da sílaba em respeito à sonoridade (SOLEWICZ, 1994). Porém, apesar desta técnica funcionar bem para a maioria dos casos, existem certas consoantes intervocálicas as quais não são bem sintetizadas da mesma forma, por exemplo, o caso do tepe alveolar [ɾ] intervocálico. Neste caso, alguns trifones contendo unidades CVC (consoante vogal consoante) podem ser usados para complementar o banco de demissílabas (BHASKARARAO, 1991).

Todo o tipo de unidade proposto acima está baseado em algum conhecimento fonético prévio, mas nada impede que estas unidades sejam automaticamente encontradas baseando-se nos requerimentos de conectividade e representatividade. Isto levou a proposta de um algoritmo automático de seleção de unidades chamado COC (NAKAGIMA, 1988) que será brevemente apresentado adiante.

2.5. Especificação do inventário de unidades

A especificação do inventário de unidades, isto é, quais serão os difones, trifones ou até tetrafones para se obter uma boa qualidade segmental e um número de unidades aceitáveis para um dado nível de naturalidade, é um processo que está necessariamente fundamentado no conhecimento da língua em questão e na experiência adquirida. Deste

modo, a seguir serão apresentadas algumas experiências na definição deste inventário, descritas em outros trabalhos.

Usualmente muitos pesquisadores têm suplementado um banco de difones com unidades mais longas. Estas unidades tem sido selecionadas manualmente buscando proteger fones com alta coarticulação (DONOVAN, 1996). Em outros casos, vê-se a proposição de um grande número de unidades, ideais à concatenação e que garantiria boa qualidade segmental, seguida de uma análise de casos particulares de modo a reduzir o inventário.

Como exemplo do primeiro caso, BIGORNE et al. (1993) utilizam um banco de difones para o francês e descrevem que testes de inteligibilidade mostraram alguns problemas na concatenação de alguns fonemas específicos, principalmente as semivogais e líquidas, que são altamente articuladas. Deste modo, um *corpus* suplementar de trifones e quadrifones foi definido, onde estes sons são "protegidos" dentro de unidades largas. Um total de 1047 novas unidades foi adicionado ao *corpus* de 1290 difones, que foram 153 trifones incluindo [j], 54 trifones incluindo [w], 544 difones incluindo [l] ou [ɾ] e 288 quadrifones incluindo [j] e [l] ou [r].

Como um segundo caso, ou seja, da proposição inicial de um inventário contendo um grande número de unidades, em (KRAFT, 1992), os autores utilizam o conceito de demissílabas e propõem que todas as sílabas possíveis para o alemão podem ser descritas por C_iVC_j, onde C_i representa um grupo de consoantes iniciais (contendo de 0 a 3 consoantes), V o núcleo contendo uma vogal ou ditongo e C_j um grupo de consoantes finais (contendo de 0 a 5 consoantes). Deste modo, foi proposto um método para determinar o inventário de demissílabas combinando todos os grupos de consoantes iniciais e finais, com todas as vogais. Em seguida, um dicionário léxico é utilizado para validar a ocorrência das demissílabas, eliminado as combinações não existentes. O resultado obtido foi um *corpus* de 1080 demissílabas iniciais (IDS) e 804 finais (FDS). Em seguida, outros 160 difones, contendo transições *VV*, foram acrescentados ao inventário.

SOLEWICZ, (1994) propõe um inventário de unidades baseado na formação da sílaba no Português. Apesar do sistema de síntese utilizado ser paramétrico, permitindo a interpolação de parâmetros na região de concatenação, alguns conceitos utilizados pelos autores poderão ser aproveitados na concatenação temporal. Inicialmente os autores observam que o Português é uma língua com forte estrutura silábica do tipo CV, onde C representa um grupo de consoantes, e V um agrupamento de vogais, como nas sílabas destacadas em "**Br**a-sil" e "m**au".** e baseado neste fato propõem um conjunto de regras para a formação da sílaba a partir desta unidade CV. O conjunto completo de regras pode ser encontrado em (SOLEWICZ, 1994), porém duas regras de formação merecem destaque. A primeira delas é baseada no fenômeno chamado de êpentese (vide seção 2.3), onde ocorre o surgimento de uma pequena vogal em um encontro vocálico. Deste modo, os autores propõem que as sílabas C_1C_2V (bra) sejam formadas concatenando-se C_1V (ba) e C_2V (ra). O outro artifício utilizado é baseado na hipótese de que há uma breve fase oral (20 ms) na transição vogal oral – vogal nasal. Assim as sílabas CV~ (sã) são formadas concatenando-se CV (sa) e V~ (ã).

Em ALBANO et al. (1997), BARBOSA et al. (1999) e SIMÕES et al. (1999) é proposto um inventário de polifones baseados em uma aprimorada análise lingüística. Na visão dos autores, muita da variabilidade segmental do Português brasileiro é devida à estrutura da sílaba, ao acento e às fronteiras de palavras ou frases. Um segmento (um fonema com diferentes alofones) aparece em duas versões: pleno, ocorrendo em forte ambiente prosódico; e reduzido, ocorrendo em ambiente prosódico fraco, e foneticamente mais variável. Os fatores de redução são as fronteiras de constituintes da sílaba e a posição relativa ao acento.

Os segmentos plenos correspondem às vogais tônicas e pretônicas, assim como as consoantes de ataque[1] silábico (inicio de sílaba). Os segmentos reduzidos correspondem às vogais postônicas, às semivogais, às consoantes de coda silábica (final de sílaba) e as líquidas de ataque silábico (inicio de sílaba). A Tabela 2.5-1 mostram agrupados os 34 segmentos geralmente considerados para o português (ALBANO et al., 1997), classificados em plenos e reduzidos. A notação usada pelos autores é intermediária entre

[1] Uma sílaba consiste em um ataque (*onset*) e uma rima; a rima por sua vez consiste em um núcleo e em uma coda. Qualquer categoria, exceto o núcleo, pode ser vazia (BISOL, 1999).

a fonética e a fonologia, no qual os segmentos plenos aparecem em letras minúsculas e os segmentos reduzidos em maiúsculas.

Em conseqüência desta análise ALBANO et al. (1997) concluem que a segmentação do sinal de fala em ambientes reduzidos deve levar uma concatenação ruim. Assim, um inventário ideal para síntese por concatenação para o Português brasileiro seria de unidades recortadas somente em pontos estacionários de segmentos plenos. Isto garante que boa parte dos fenômenos coarticulatórios que se manifestam predominantemente nos segmentos reduzidos estariam contidos nas próprias unidades. No entanto, a adoção deste principio implicaria em um inventário de aproximadamente 20.000 unidades (SIMÔES, 1999).

Tabela 2.5-1: Classificação dos segmentos em reduzidos (em minúsculo) e plenos (em maiúsculo) conforme ALBANO et al. (1997).

Consoantes plenas

	labiais	coronais	pal./vel.
Oclusivas	p	t	k
	b	d	g
Fricativas	f	s	sh
	v	z	zh
Nasais	m	n	nh
Laterais		l	lh
Vibrante		r	

Consoantes reduzidas

S
N
L
R

Vogais plenas

i u
e o
eh oh
a

Vogais reduzidas

I U
E O
A

Para reduzir o inventário, a partir de uma análise fonética de alguns casos, os autores utilizam alguns critérios, como:

- As vogais pretônicas não são diferenciadas das tônicas.
- As vogais postônicas em posição de núcleo silábico são segmentadas. Neste caso equivalem a unidades demissilábicas. O corte é efetuado no final da transição com a consoante precedente. Por exemplo: ótimo - /ohtimo/ - /oh+oht+**tI+Im**+mO+O/ (somente a transição entre t e I está contida em tI)

- As vogais nasais e ditongos nasais são concatenados com ataques (inícios) orais. SOLEWICZ, (1994) utiliza o mesmo princípio. O corte na vogal de ataque oral é feito logo após a transição desta com a consoante precedente. Portanto, tais unidades podem ser concatenadas tanto com unidades contendo vogais ou ditongos orais como com aqueles contendo vogais e ditongos nasais. Por exemplo: bomba - /boNbA/ - /b+**bo+oNb**+bA+A/.
- Segmenta-se S, R e L quando ocorrerem no início de encontros consonantais. Por exemplo: pasta - /paStA/ - /p+pa+**aS+St**+tA/

Utilizando estes critérios os autores chegam a um banco de 2193 unidades e declaram uma ótima qualidade de concatenação. O inventário completo dos polifones utilizados pode ser encontrado em (SIMÔES, 1999).

Seguindo o caso de redução do banco, baseado em características fonéticas peculiares, KAFKA et al. (2002) propõem um método de reduzir o inventário de demissílabas. Este método é baseado no conceito de que um único segmento de transição VC (vogal consoante), que apresente o mesmo ponto de articulação (uma mesma classe homorgânica), poderia ser usado para sintetizar diferentes segmentos. Por exemplo, sendo os segmentos [t] e [ʃ] (representados como segmentos de concatenação por {t} e {X}) homorgânicos, a palavra "machado" seria concatenada na forma {_m ma at XA At do o_}[2], onde se observa que a demissílaba {at} se une a {XA} sem nenhum prejuízo perceptual (KAFKA et al., 2002). Na Tabela 2.5-2 é mostrado o esquema de substituição utilizado.

Tabela 2.5-2: Correspondência de segmentos consonantais homorgânicos (KAFKA et al., 2002).

Segmentos sintetizados	Correspondente
Vogal + [p]	- oclusicas bilabiais [p b] - nasal [m]
Vogal + [t]	- oclusivas alveolares [t d] - vibrante [r] - tepe [ɾ]

[2] A notação utilizada neste trecho da tese segue a mesma de (KAFKA et al., 2002) onde _ corresponde ao silêncio; as letras maiúsculas indicam o contexto tônico e as minúsculas o átono.

	- lateral [l] - nasal [n] - fricativas dental-alveolares [s z] - fricativas palato-alveolares [ʃ ʒ]
Vogal + [f]	- fricativas lábio-dentais [f v]
Vogal + [k]	- oclusivas velares [k g]

Com este método de equivalência das demissílabas VC, os autores relatam a redução de 204 unidades VC para 48. Porém, nos testes de naturalidade houve o estranhamento por parte dos ouvintes de palavras com as consoantes "s" e "r" em final de sílaba, por exemplo, a palavra "vesgo", onde neste caso seriam necessárias demissílabas específicas.

De maneira similar à apresentada acima, de substituição de um grupo de segmentos por um segmento equivalente, YOSHIDA et al. (1991) propõem o uso de duas técnicas para efetuar este agrupamento, reduzindo um inventário de trifones construído para o Japonês. A primeira é baseada em 3 regras fonológicas e a segunda é baseada no agrupamento de trifones através de quantização vetorial, baseado em características acústicas. As 3 regras são as seguintes:

- As vogais longas são tratadas como normais se estiverem na primeira ou terceira posição. Por exemplo[3], o trifone $^{A}M^{I}$ é uma representação para $^{a}M^{I}$, $^{a}M^{i}$,...
- São agrupados os trifones onde os dois últimos fonemas são os mesmos e o fonema do meio é uma plosiva não vozeada. Por exemplo: $^{U}K^{O}$ é uma representação para $^{A}K^{O}$, $^{I}K^{O}$,...
- São agrupados os trifones onde os dois primeiros fonemas são os mesmos e o último fonema é uma plosiva não vozeada. Por exemplo: $^{b}E^{p}$ é uma representação para $^{b}E^{t}$, $^{b}E^{k}$,...

[3] A notação aqui utilizada, segue a notação de YOSHIDA, (1991), onde $^{A}M^{I}$ denota o fonema /M/ precedido por /A/ e seguido por /I/.

Desta maneira um inventário de 15.000 trifones foi reduzido para 6.000. A segunda técnica é baseada no algoritmo COC onde cada trifone é quantizado e agrupado de acordo com o seu contexto fonético. O algoritmo inicia agrupando todos os trifones que possuem o mesmo fonema precedente, por exemplo, ${}^{p}E^{g}$, ${}^{p}E^{f}$,... Em seguida, os grupo são divididos recursivamente, conforme sua variância interna, até que sejam atingidas as condições de número mínimo de unidades dentro do grupo e do número máximo de grupos. Por fim, cada grupo possuirá um protótipo, ou elemento centróide, que será a unidade de síntese representativa do grupo. Por exemplo, se um protótipo tem o contexto fonético ${}^{p}a^{d}$ e o grupo contém mais dois elementos de contexto ${}^{p}a^{z}$ e ${}^{p}a^{s}$, então este protótipo será usado como unidade de síntese para o contexto fonético precedido por [p] e seguido por [d],[z] ou [s].

2.5.1. Agrupamento orientado pelo contexto (COC)

No método chamado de COC, as unidades podem ser geradas automaticamente sem nenhum conhecimento fonológico a priori, e a geração das unidades alofônicas de síntese não é dependente da decisão humana mas das características estatísticas dos parâmetros espectrais na fala natural (DUTOIT, 1993).

No algoritmo, o espaço de segmentos fonéticos (fones) é particionado iterativamente, mantendo a correspondência entre os subespaços divididos e seus contextos fonéticos. A partir de um *corpus* etiquetado foneticamente, primeiro é feita a extração de parâmetros obtendo um vetor de características com marcas fonéticas. Cada *cluster* inicial é um conjunto de segmentos com a mesma identificação fonética. Em seguida, o processo de divisão do *cluster* é aplicado interativamente de modo a formar grupos de acordo com o contexto fonético. Após um número de interações, quando as condições de término são satisfeitas, o segmento centróide de cada *cluster* será a unidade de síntese e o contexto fonético indicará onde a unidade será usada (NAKAGIMA, 1988).

Na Figura 2.5-1, temos um exemplo do processo de partição. O grupo inicial w_1, que representa um conjunto de segmentos com a mesma etiqueta fonética [a], é dividido em

2 grupos: w_{11} ([a] precedido por [b]) e w_{12} ([a] precedido por qualquer fonema com exceção de [b]). Estes grupos podem ser divididos em outro dois grupos, por exemplo, w_{11} se divide em: w_{111} ([a] precedido por [b] e sucedido por [s]) e w_{112} ([a] precedido por [b] e sucedido por qualquer fonema com exceção de [s]). Este processo é continuado até que alguma condição de término seja satisfeita. Note na Figura 2.5-1 que cada grupo possui uma representação simbólica que identifica o fonema do grupo e o seu contexto.

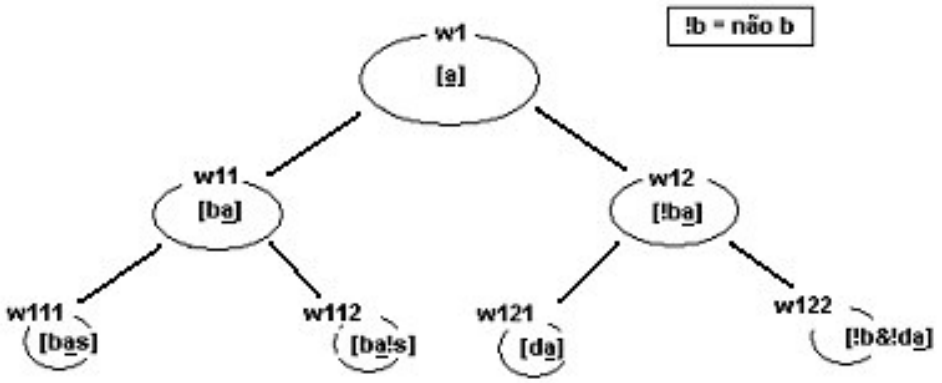


Figura 2.5-1: Exemplo do processo de partição automática.

O algoritmo consiste no seguinte: considere v_i^2 um valor de avaliação interno ao grupo w_i, dado pela variância interna do grupo, normalizada pelo tamanho do grupo. Para uma medida de avaliação total J, considere a média dos valores de avaliação dos grupos. Considere ainda um valor de avaliação de partição G dado por: $G = v_1^2 - (v_{11}^2 + v_{12}^2)/2$, para o caso do grupo w_1 ser particionado em w_{11} e w_{12}. Assim, seja N_{min} o número de segmentos por grupo e J_{max} o limiar para o valor de J

i) Tome cada grupo inicial w_i formado pelos segmentos com a mesma etiqueta e calcule v_i^2;

ii) Encontre w_k que tenha o máximo v_k^2 e tenha mais segmentos do que N_{min}.

iii) Entre todos os contextos vizinhos de w_k, encontre o contexto que produz o máximo valor de G. O contexto vizinho é o contexto anterior ou posterior além da etiqueta do contexto corrente.

iv) Divida o *cluster* em dois *clusters* disjuntos.

v) Se o valor de avaliação total J é maior do que J_{max} volte ao passo ii).

Os autores concluem que o método produz resultados que coincidem com fatos fonológicos. Para verificação das unidades geradas, consulte (NAKAGIMA, 1988).

2.6. Geração do *corpus*

Com exceção dos métodos de seleção automática de unidades que utilizam um extenso *corpus* de fala natural, as unidades são geralmente extraídas de um *corpus* especializado porque a freqüência relativa de ocorrência em fala natural é altamente variável. No caso de difones, por exemplo, uma lista de 100 sentenças foneticamente balanceadas cobre somente 43% das 1200 unidades necessárias para o francês, com uma redundância de 80% (BIGORNE et al., 1993). Por isso, o projeto e gravação do *corpus* devem ter atenção especial.

Em alguns casos os bancos usados para preparar o inventário de unidades são compostos por palavras isoladas sem significado (logotomas) construídas especialmente para conter as unidades especificadas, ao invés de palavras reais (DONOVAN, 1996). O uso de logotomas no lugar de palavras reais é ainda indicado quando se deseja usar métodos de segmentação automática. BIGORNE et al., (1993) sugerem que o uso de palavras reais na segmentação automática de difones para o alemão resultou em uma performance relativamente ruim.

Porém, dependendo da especificação das unidades, algumas não são possíveis de serem obtidas com palavras sem significado. Em (SIMÕES, 1999), para as unidades que não acontecem no interior de palavras, por exemplo, no encontro vocálico do tipo (vogal

postônica + vogal tônica), foram utilizadas palavras reais. Estas palavras foram escolhidas de modo a obter uma seqüência (Nome + Adjetivo), por exemplo, a frase "bol**a o**ca".

SIMÕES, (1999) e ALBANO, (1997) argumentam que a pronúncia de palavras ou logotomas isoladamente não seria uma boa estratégia pois as unidades deveriam ocorrer dentro de um enunciado maior, garantindo a neutralidade prosódica. Porém, esta neutralidade prosódica pode ser obtida também com outros mecanismos, impondo ao leitor condições de ritmo e de *pitch* (LENZO, 2002).

Neste sentido, a escolha do contexto fonético parece ser mais importante, de modo a minimizar a coarticulação da unidade com os segmentos adjacentes. É consensual que a unidade precisa estar em um contexto fonético neutro, porém a forma de criar esta neutralidade é variável. Por exemplo, em (SIMÕES, 1999) foram utilizadas as [p] e [b] e a vogal [a] como adjacentes. Para unidades que ocorrem no meio de uma sentença, utilizou-se a forma ***/pa**+unidade/* inserida na frase veículo: *Digo _**b**aixinho.* No caso de unidades que podem ocorrer no inicio ou no fim das sentenças, foram gravadas respectivamente as formas */unidade+**pa**/* e ***/pa**+unidade/*, inseridas nas frases: *_ **d**igo baixinho* e *Baixinho digo _* .

A posição da unidade em relação à tonicidade também é uma questão a ser considerada. As sílabas acentuadas são longas e portanto menos submetidas a coarticulação, no qual resulta em unidades facilmente concatenáveis, no entanto, as sílabas não acentuadas são mais numerosas na fala natural, portanto a produção delas pode aumentar a qualidade segmental (DUTOIT, 1997). Em (KRAFT, 1992) todas as unidades foram inseridas em palavras com três sílabas, de modo que ocupassem a posição de acento secundário, ou seja, a posição entre uma sílaba não acentuada e o acento primário. Isto foi determinado para que a unidade obtida apresentasse boa manipulação prosódica em ambos os sentidos. Em alguns casos é aconselhada a gravação de duas versões da mesma unidade, uma situada na posição tônica e outra não tônica, diminuindo ainda mais a modificação prosódica .

Em (KRAFT, 1992) os logotomas foram inseridos em sentenças sem significado e deste modo foi fácil ao falante manter a freqüência fundamental e os padrões de duração constantes durante a gravação. Do mesmo modo, em LENZO (2002), na construção de um banco de difones para o sistema festival, foram usadas sentenças sem sentido, onde mais de um difone é extraído de cada sentença.

Para buscar um ambiente prosódico neutro, pode ser imposto ao falante que faça uma pronúncia monótona (robotizada) e de forma lenta. Isto resulta em unidades mais inteligíveis, porém altamente articuladas. Estes acontecimentos podem ser vistos como parte de um lamentável, mas necessário, compromisso entre inteligibilidade e naturalidade (DUTOIT, 1997).

Um recurso interessante usado na construção de novas vozes para o sistema Festival (LENZO, 2002) é a sugestão de leitura. Na etapa de gravação do corpus, o leitor ouve primeiramente uma sugestão, sintetizada com *pitch* e velocidade constantes, demonstrando como a sentença deve ser lida e induzindo-o a uma leitura monótona. Observa-se que a produção de sons monótonos é mais fácil no caso de palavras sem sentido (LENZO, 2002).

2.7. Marcação fonética e recorte de unidades

Após a gravação do *corpus*, a etapa que prossegue é a identificação e recorte das unidades. Considere o sinal de uma sentença gravada, que contêm uma certa unidade em seu interior. Geralmente a identificação desta unidade dentro da frase será feita visualmente observando o sinal no domínio do tempo e também no domínio da freqüência, com o auxilio de um espectrograma. Se esta unidade é limitada pela transição entre segmentos fonéticos, o fone, por exemplo, então o recorte da unidade será feito na própria fronteira entre os segmentos. Como regra geral os difones e trifones são recortados em sua região mais estável, correspondendo geralmente à região central do segmento. Uma exceção a esta regra são os segmentos oclusivos, que devem ser recortados preferencialmente na região de oclusão.

Na redução do inventário de unidades, geralmente são utilizados artifícios fonéticos que determinam o instante de recorte. Como visto na seção 2.5, em (SIMÕES, 1999) os autores fazem a concatenação de vogais nasais e ditongos, com segmentos de inicio oral. Desta forma, o recorte deverá ocorrer logo após a transição da consoante precedente com a vogal oral em aproximadamente 20 ms (SOLEWICZ, 1994).

Os sistemas de síntese por concatenação utilizam a fronteira entre as unidades fonéticas como referência para a aplicação de regras de manipulação prosódica. Por exemplo, suponha que uma sílaba terá a freqüência fundamental e a duração aumentadas, se tornando uma sílaba com características de tônica. Assim, após a concatenação das unidades, o algoritmo de síntese precisará encontrar no sinal concatenado a informação de onde começa e termina a sílaba a ser modificada. Portanto, as fronteiras entre segmentos fonéticos precisam estar marcadas nas unidades.

Tradicionalmente a identificação, marcação e o recorte das unidades são feitos manualmente por um foneticista experiente, o que requer um consumo considerável de tempo além de ser um trabalho tedioso. Tendo em vista estes fatores, um sistema de etiquetagem automática poderá ser uma possibilidade atrativa.

A marcação fonética e o recorte totalmente automáticos são dificilmente concebíveis no contexto de sintetizadores baseados em concatenação, exceto para técnicas como a usada em COC, na qual explicitamente dispensa a intervenção no processo de geração dos segmentos. Porém, isto não implica necessariamente que métodos de segmentação automática são inúteis, pelo contrário, são interessantes em uma perspectiva multilingües ou para a criação de novas vozes (DUTOIT, 1993).

Quando as unidades escolhidas são difones ou polifones, a etiquetagem automática, ou segmentação dos fones, atua somente como um guia para um recorte mais acurado e portanto requer uma acurácia limitada (DONOVAN, 1996).

2.8. Conclusão

Neste capítulo foram revistos alguns conceitos da fonética e da fonologia, necessários tanto para o entendimento da literatura específica, assim como o próprio conhecimento da língua e foram também apresentados alguns fenômenos fonéticos peculiares, úteis na consideração do inventário de unidades.

Foram apresentadas também as unidades mais comumente utilizadas, observando algumas considerações como a capacidade de envolverem a maior quantidade possível de fenômenos articulatórios, a redução de pontos de concatenação, largura e tamanho. Foi observada também a preferência em utilizar bancos com unidades de tamanhos diferenciados, buscando proteger segmentos altamente coarticulados ou reduzir um inventário super dimensionado.

Na determinação do inventário de unidades observou-se que o processo de definição é geralmente recorrente, ou seja, parte-se de um inventário inicial, extenso, que garantiria teoricamente boa inteligibilidade e naturalidade e em seguida, a partir do exame fonético e fonológico de casos particulares, o inventário é reduzido. Sobre este procedimento, foram recolhidas diferentes experiências, de onde se poderá agrupar o melhor de cada.

Foi visto também que após a gravação do *corpus*, na etapa de identificação, marcação e recorte das unidades, seria interessante um método de marcação ou segmentação semi-automático que auxiliasse o usuário nesta tediosa tarefa. No capítulo 3 a seguir será abordado um método de segmentação acústica que será usado nesta proposta.

Na introdução deste capítulo algumas perguntas foram colocadas de maneira que precisariam ser respondidas antes da construção do banco de unidades, como uma forma de planejamento. Deste modo, ao longo do capítulo se procurou fornecer diferentes possibilidades de resposta a estas questões, na busca por acrescentar diferentes abordagens do problema. No capítulo 5 algumas destas questões serão respondidas, no âmbito desta tese, onde será descrita a construção de um banco de unidades utilizando um aplicativo desenvolvido para este fim.

Capítulo 3

Segmentação automática

3.1. Introdução

Foi visto no Capítulo 2 que o processo manual de recorte e marcação das unidades é tedioso, que exige um consumo de tempo considerável e que um processo automático ou semi-automático é uma possibilidade atraente, que será abordada neste capítulo.

VAN HEMERT, (1991) classificou as formas de segmentação em implícita ou explícita. Na segmentação implícita, o sinal é fracionado somente a partir de características intrínsecas, geralmente acústicas. Na segmentação explícita é utilizado algum conhecimento prévio a respeito dos segmentos, tal como a transcrição fonética dos segmentos.

Na segmentação implícita, a característica principal é a exploração da continuidade (ou descontinuidade) de características acústicas do sinal. Por exemplo, SVENDSEN, (1987) considera o sinal de voz como uma seqüência de segmentos quasi-estacionários, considerando cada segmento como uma configuração quasi-estacionária do trato vocal, caracterizados por um centróide com flutuações estatísticas. Assim, o problema de segmentação consistiria em minimizar a distorção total em torno do centróide.

Um caso de segmentação explícita é dado também por SVENDSEN (1987), onde os autores referenciam a técnica como *template matching*. Neste método foram coletados 41 segmentos fonéticos e através de um método de clusterização uma única janela de análise foi utilizada como padrão (*template*) para cada fonema, com exceção dos ditongos e plosivas, onde foram utilizadas duas janelas como padrão. Em seguida, a partir da transcrição fonética dada, a distorção acumulada entre os padrões e o sinal a ser segmentado é minimizada por programação dinâmica. Uma variação deste método é proposta em (VAN HEMERT, 1991), onde a janela padrão é selecionada manualmente no próprio sinal a ser segmentado.

Utilizando esta distinção entre segmentação implícita e explicita, VAN HEMERT, (1991) compara estes dois métodos onde concluí que na segmentação implícita, nem sempre o número correto de segmentos é fornecido; os segmentos não são nomeados e as vizinhanças entre segmentos são determinadas de maneira acurada sob o ponto de vista acústico. Por outro lado, na segmentação explícita o número de segmentos é dado pela transcrição fonética; os segmentos são nomeados de acordo com a transcrição fonética e as vizinhanças podem ser não acuradas.

As propostas de segmentação explícita deram origem a métodos mais recentes que podem ser separados em duas principais vertentes. A primeira, a partir da transcrição fonética equivalente do sinal a ser segmentado, um sistema de síntese é utilizado para gerar a fala sintética correspondente e em seguida o sinal é segmentado e etiquetado através de alinhamento temporal, ou seja, DTW (*dynamic time warping*), conforme proposto em (MALFRÈRE, 1997). Em uma segunda vertente, conforme enunciado por RAPP (1993), o problema de alinhamento automático pode ser considerado como um problema simplificado de reconhecimento contínuo e portanto é natural adotar um

paradigma bem sucedido nos sistemas de reconhecimento, nomeadamente os HMMs (*hidden Markov models*).

O método desenvolvido por LJOLJE, (1991), utilizando HMMs, é considerado precursor desta proposta, tendo sido posteriormente utilizado por ele mesmo em (LJOLJE, 1993) e (LJOLJE et al., 1996) e por outros (VAN SANTEN, 1999) na segmentação automática de unidade para síntese por concatenação. Do mesmo modo, LENZO, (1999) e MALFRÈRE, (1997) utilizam a proposta por DTW com o mesmo fim.

Conclui-se portanto que para a concepção de um método de segmentação automática, conforme as propostas mais recentes, comparadas por ADELL (2004), é preciso ter em mãos um sistema de reconhecimento contínuo ou um sistema de síntese, que implica na posse de um banco de unidades inicial.

Não tendo a disposição nem um nem outro, neste trabalho será dado enfoque à segmentação do sinal a partir das informações acústicas (segmentação implícita), na qual o conhecimento prévio da transcrição fonética será usado para a etiquetagem dos segmentos e ainda para reduzir a complexidade do problema de segmentação.

O problema de segmentação do sinal de fala por uma metodologia de segmentação implícita, ou seja, pela obtenção das fronteiras entre segmentos somente a partir da estrutura acústica do sinal, sem nenhum conhecimento fonético prévio, foi anteriormente abordado em (SHARMA, 1996) e (SVENDSEN, 1987), onde os autores utilizam a programação dinâmica como solução.

Observa-se que o uso da programação dinâmica nestes métodos, apresenta uma forte similaridade com as soluções para o problema de reconhecimento de dígitos conectados, apresentados em (MAYER, 1981) e (RABINER, 1993) como *Two-level dynamic programming* (TLDP) e *level building dynamic programming* (LBDP), porém esta similaridade não é tão óbvia. Deste modo, na seção 3.2 e 3.3 serão descritos e implementados os métodos citados acima para um melhor entendimento dos algoritmos.

Em seguida, na busca por uma abordagem teórica para o problema de segmentação usando a programação dinâmica como solução, na seção 3.4, será apresentado um método de segmentação unidimensional de um sinal, conforme proposto por KAY, (2001). Observando a similaridade com os algoritmos citados para reconhecimento de dígitos, os algoritmos serão modificados para a segmentação unidimensional e os resultados apresentados no tópico seguinte.

Como solução para o problema de segmentação, dita implícita, do sinal de voz, verifica-se um caso multidimensional da abordagem acima. Assim, na seção 3.5 os algoritmos foram adaptados para variáveis vetoriais, onde os resultados são apresentados. Ao longo deste tópico, outras interpretações são dadas ao problema, assim como algumas características do método são discutidas e por fim, são apresentadas as conclusões.

3.2. Two-level dynamic programming (TLDP)

Originalmente, o método TLDP foi proposto como solução para o problema de reconhecimento de dígitos conectados em (RABINER, 1993), que de uma forma geral, significa encontrar dentro de um dado sinal a ocorrência de um ou mais padrões de referência, através de uma medida de similaridade. Por exemplo, dada uma seqüência de dígitos falados continuamente, e na posse de um padrão para cada dígito, deseja-se encontrar a melhor coincidência entre os padrões e a seqüência de dígitos.

Considere porém um caso simplificado do reconhecimento de dígitos, onde o vetor de entrada contém uma seqüência de ocorrências de um mesmo padrão.

A idéia deste algoritmo é dividir o problema em dois estágios ou níveis. Seja um vetor padrão de referência R e um vetor de entrada X de tamanho M. No primeiro nível o algoritmo compara o padrão R com uma porção do vetor X.

Seja *b* e *e* amostras do vetor de entrada, sendo que *e* é sempre maior que *b*. No primeiro nível do algoritmo, partindo de cada *b*, $1 \le b \le M$, são encontrados pelo método DTW as menores distâncias que levam a cada *e,* para $b < e \le M$. Na Figura 3.2-1 é ilustrado o processo de busca pelos caminhos que levam de *b* até *e*.

A distância mínima entre o padrão e cada intervalo (*b, e*) é dada por:

$$D(b,e) = \min_{w(m)} \sum_{m=b}^{e} d(x(m), r\,(w(m))) \tag{3.1}$$

onde *w(m)* é uma função de *warping,* ou seja, a função que fornece o alinhamento temporal entre o vetor padrão e o intervalo (*b,e*) no vetor de entrada. Como dito anteriormente esta função é obtida por DTW, vide ANEXO D.

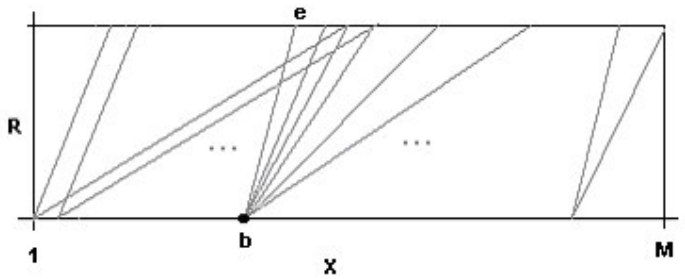


Figura 3.2-1: Ilustração do processo de busca de similaridade entre o padrão *R* e um segmento do vetor *X* no intervalo (*b,e*).

Para o segundo nível do algoritmo, considere que em um certo ponto *e*, do vetor de entrada, o padrão de referência ocorreu exatamente *l* vezes no sinal de entrada (o que corresponderia no caso dos dígitos conectados a *l* referências). Dado que o melhor ponto para se chegar a *e* é partindo de *b,* o melhor caminho para se chegar a *e*, é a soma da distância $\bar{D}_{l-1}(b-1)$ para se chegar até *b*-1, mais a distância $D(b,e)$ entre *b* e *e*. A distância para se chegar a *b*-1 equivale a passar por $l-1$ referências e terminar no ponto *b*-1. Deste modo, a distância do melhor caminho terminando no índice *e,* usando uma seqüência de *l* ocorrências da referência, é dada por (MYERS, 1981):

$$\bar{D}_l(e) = \min_{1 \le b \le e} [D(b,e) + \bar{D}_{l-1}(b-1)] \tag{3.2}$$

Aplicando o método acima recursivamente, para L_{max} ocorrências, o segundo nível do algoritmo é formulado da seguinte forma:

Início: $\bar{D}_0(0) = 0, \quad \bar{D}_l(0) = \infty, \quad \text{para} \quad 1 \le l \le L_{\max}$

para l =1

$$\bar{D}_1(e) = D(1, e), \quad \text{para} \quad 2 \le e \le M$$

para l=2,3,...,L_{max}

$$\bar{D}_2(e) = \min_{1 \le b \le e}[D(b, e) + \bar{D}_1(b-1)], \quad \text{para} \quad 3 \le e \le M$$

$$\bar{D}_3(e) = \min_{1 \le b \le e}[D(b, e) + \bar{D}_2(b-1)], \quad \text{para} \quad 4 \le e \le M$$

$$\vdots$$

$$\bar{D}_l(e) = \min_{1 \le b \le e}[D(b, e) + \bar{D}_{l-1}(b-1)], \quad \text{para} \quad l+1 \le e \le M$$

$$F_l(e) = \arg\min_{1 \le b \le e}[D(b, e) + \bar{D}_l(b-1)] - 1$$

Final: $L = \arg\min_{L_{\min} \le l \le L_{\max}}[\bar{D}_l(M)]$

$$D = \min_{1 \le l \le L_{\max}}[\bar{D}_l(m)]$$

A função $F_l(e)$ é o ponto final, ao longo do sinal de entrada, da ocorrência $l-1$ do padrão na seqüência na qual gera a mínima distância $\bar{D}_l(e)$.

Os instantes de fronteira de cada ocorrência são obtidos através retroativamente por:

$e_L = M$

$e_l = F_l(e_{l+1}), \quad \text{para} \quad l = L_{\max} - 1, L_{\max} - 2, ..., 1$

Experimento 3.1: Implementação do método TLDP para o caso simplificado de um único padrão.

Na Figura 3.2-2, observa-se no sinal de entrada, à esquerda, a ocorrência de três padrões. O pequeno sinal à direita é tomado como a referência. Deseja-se encontrar no sinal de teste os instantes de fronteira das três ocorrências desta referência.

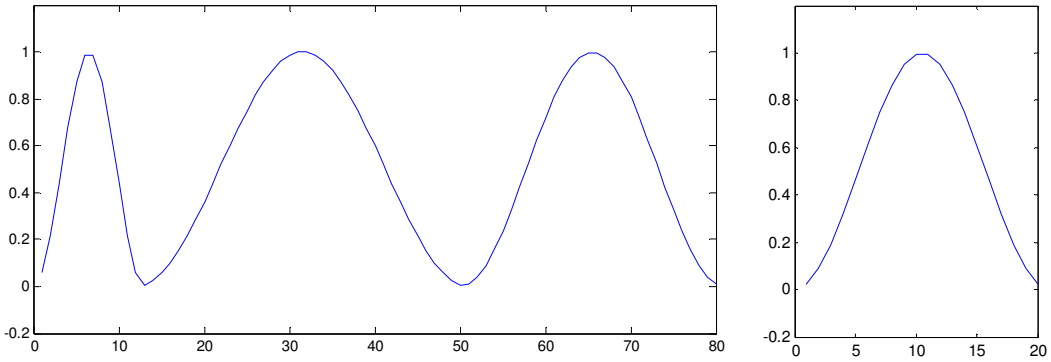


Figura 3.2-2: Exemplo de um sinal de entrada e sinal de referência.

No primeiro nível do algoritmo, aplica-se recursivamente o DTW para encontrar o melhor caminho entre todos os pares (*b,e*). Na Figura 3.2-3, é mostrada a matriz $D(b,e)$, onde a parte abaixo da diagonal principal, representa os pontos onde (*e*<*b*) que não foram calculados. As regiões acima da diagonal, que aparecem mais claras, exibem os intervalos (*b,e*) onde acontecem as maiores similaridades com a referência.

Observando o sinal de teste, é esperado que um dos padrões seja identificado no intervalo (13 , 50), por exemplo, dado *b* = 13, é esperado que em torno de *e* = 50 exista alta similaridade, como pode ser confirmado na figura.

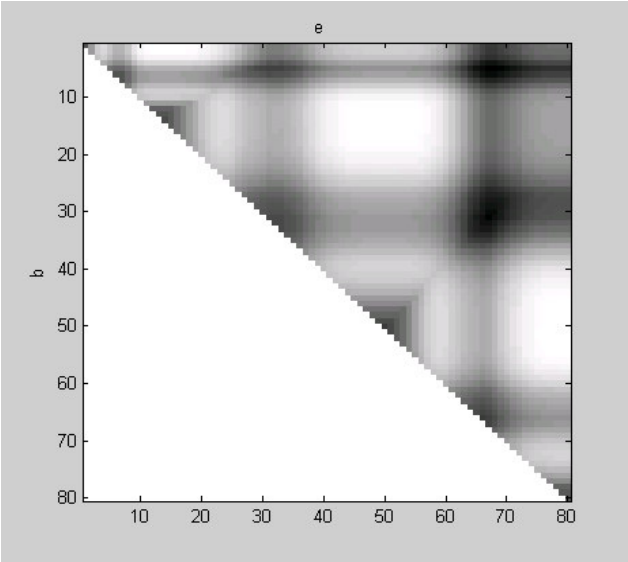


Figura 3.2-3: Imagem da matriz de similaridade *D(b,e)* entre o sinal de teste e a referência.

Após o cálculo da matriz, foram escolhidos 3 níveis e a recursividade foi aplicada, resultando na marcação da Figura 3.2-4.

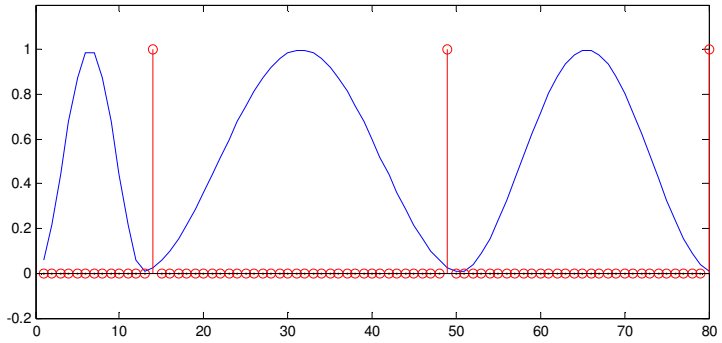


Figura 3.2-4: Resultado do algoritmo quando selecionados 3 níveis. Os traços em vermelho indicam os pontos de fronteira entre as ocorrências.

Em um exemplo seguinte, as amplitudes dos padrões foram modificadas e foi acrescentado um intervalo entre dois padrões. Executando o algoritmo com 4 níveis, concluímos pela Figura 3.2-5 que o algoritmo é capaz de identificar os padrões.

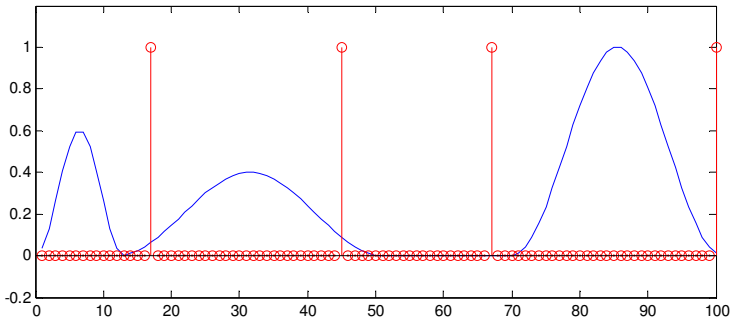


Figura 3.2-5: Resultado do algoritmo de identificação de padrões para um segundo caso.

3.3. Level building dynamic programming (LBDP)

Na solução do mesmo problema, da identificação de padrões conectados, encontra-se o LBDP (RABINER, 1993) e (MYERS, 1981). Por simplicidade, considere o mesmo caso citado na seção anterior de um único padrão e um sinal de entrada que consiste em ocorrências seqüenciais deste padrão. O algoritmo consiste em dividir o alinhamento entre o padrão e as ocorrências da entrada em níveis horizontais intermediários. No primeiro nível, comparando o sinal de entrada com o sinal padrão, todos os caminhos que tocam o primeiro nível horizontal, ou seja, o fim do primeiro padrão de referência, têm a sua distância acumulada guardada e então passa-se ao próximo nível.

Nota-se que o conceito de níveis neste caso difere do algoritmo em dois níveis descrito anteriormente. A diferença fundamental destes dois métodos é que em LBDP o resultado do alinhamento que ocorre em um certo nível é usado para reduzir o intervalo de busca nos níveis posteriores, enquanto que em TLDP o resultado do alinhamento é acumulado até o fim de toda a entrada. Em LBDP a computação necessária é significantemente reduzida. A descrição do algoritmo pode fornecer um melhor entendimento do método.

Considere um sinal de entrada de tamanho *M*, contendo *L* ocorrências do padrão, e um sinal padrão de tamanho *N*. Dado um ponto *m* da entrada e um ponto *n* do padrão de referência num dado nível *l*, ou seja, na busca da ocorrência *l* do padrão na entrada. Define-se: $d(m,n)$ como a distância entre a amostra no ponto *m* da entrada e a amostra no ponto *n* do padrão de referência; $D_l(m,n)$ como a distância acumulada até ponto (*m*, *n*) no nível *l*, e $\tilde{D}_l(m)$ como a distância acumulada do ponto *m* ao último ponto da referência, no nível *l*.

Observa-se que no caso do LBDP, o alinhamento temporal entre o padrão de referência e o sinal de entrada está implícito no algoritmo no cálculo de $D_l(m,n)$. Isto implica que a imposição imposta ao caminho de busca precisa estar contida no algoritmo e não mais visto a parte como no caso do TLDP. Considere que as funções $L(m)$, $U(m)$, $\tilde{L}(n)$ e

$\tilde{U}(n)$ definem a imposição ao caminho. No ANEXO D são fornecidos mais detalhes sobre a imposição aos caminhos de transição, no DTW.

Assim, o algoritmo é descrito pelos seguintes passos (MYERS, 1981):

Início: $\tilde{D}_0(m) = \begin{cases} 0 & m = 0 \\ \infty & m \neq 0 \end{cases}$ $\tilde{D}_l(0) = \infty, \text{para} \quad l = 1, 2, ..., L_{\max}$

para $l = 1, 2, ..., L_{\max}$

para $m = 0, 1, 2, ..., M$

$\tilde{D}_l(m, 0) = \tilde{D}_{l-1}(m)$, $F_l(m, 0) = m$

para $m = 1, 2, ..., M$ e $n = L_l(m), ..., U_l(m)$

$\hat{n} = \underset{\tilde{L}(n) \le n' \le \tilde{U}(n)}{\arg\min} [D_l(m-1, n')]$

$D_l(m, n) = d_l(m, n) + D_l(m-1, \hat{n})$

$F_l(m, n) = F_l(m-1, \hat{n})$

para $m = 1, 2, ..., M$

$\tilde{D}_l(m) = \tilde{D}_l(m, N)$

$\tilde{F}_l(m) = \tilde{F}_l(m, N)$

volta

volta

Fim: $L = \underset{L_{\min} \le l \le L_{\max}}{\arg\min} [D_l(M)]$

$D = D_L(m)$

Os instantes de fronteira de cada ocorrência são obtidos retroativamente por:

$e_L = M$

$e_l = \tilde{F}_{l+1}(e_{l+1}), \quad \text{para} \quad l = L-1, L-2, ..., 1$

O algoritmo implementado apresenta resultados similares ao método de programação em dois níveis. No entanto, a complexidade computacional é reduzida significativamente.

3.4. Segmentação unidimensional por programação dinâmica

Nas seções 3.2 e 3.3 foram vistos os algoritmos que identificam dentro de um sinal as ocorrências de um dado padrão. Porém, deseja-se estudar o caso de segmentação de um sinal quando não existe um padrão definido anteriormente. Neste caso, a segmentação do sinal utiliza características intrínsecas do próprio sinal para determinar fronteiras entre possíveis segmentos.

KAY, (2001) propõe um método de segmentação ótima, aplicado a detecção de bordas em imagens. Considere um sinal $\{x(0), x(1), ...x(N-1)\}$, conforme a Figura 3.4-1, composto de N_s segmentos de diferentes estatísticas. Deste modo, existem $N_s - 1$ tempos de transição compondo o conjunto $\{n_0, n_1, ...n_{N_s}\}$. O problema mais geral de segmentação é estimar o número de segmentos N_s, os tempos de transição, e os parâmetros estatísticos de cada segmento.

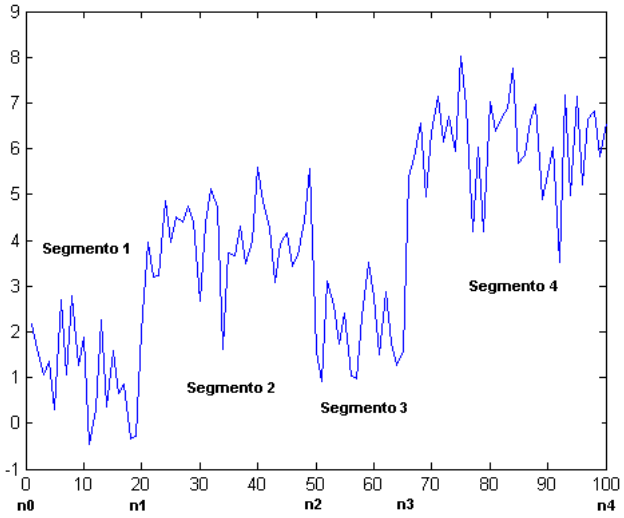


Figura 3.4-1: Exemplo de um sinal composto por 4 segmentos.

A proposta estatisticamente ótima para resolver o problema é encontrar o estimador de máxima verossimilhança (*maximum likelihood estimator*, MLE) dos parâmetros não conhecidos (KAY, 2001).

Conforme a Figura 3.4-1 , assume-se que o *iésimo* segmento é caracterizado pela função de densidade de probabilidade conjunta ou função de verossimilhança $p_i(x(n_{i-1}),...,x(n_i-1)\,|\,\theta_i)$, onde θ_i é um vetor de parâmetros não conhecidos. Além disso, assumindo que cada segmento é estatisticamente independente dos outros segmentos, podemos definir que para todos os segmentos a função será (KAY, 2001):

$$\prod_{i=1}^{N_s} p_i(x(n_{i-1}),...,x(n_i-1)\,|\,\theta_i) \tag{3.3}$$

O produtório acima pode ser reescrito como o somatório de logaritmos:

$$\sum_{i=1}^{N_s} \ln\, p_i(x(n_{i-1}),...,x(n_i-1)\,|\,\theta_i) \tag{3.4}$$

Deste modo, deseja-se maximizar a equação acima de modo a encontrar $\{n_1, n_2, ...n_{N_s-1}, \theta_1, \theta_2, ...\theta_{N_s-1}\}$ e assim, seria necessário:

- escolher um conjunto de tempos de transição;
- estimar $\hat{\theta}_i$ para cada segmento;
- efetuar o somatório dos logaritmos;
- repetir o cálculo para cada conjunto de tempos de transição possíveis.

A dificuldade neste procedimento é a quantidade de combinações possíveis para os tempos de transição. Este problema pode ser resolvido por programação dinâmica (KAY, 2001).

Reescrevendo $\mathbf{x}[n_{i-1}, n_i-1] = x(n_{i-1}),...,x(n_i-1)$, temos que maximizar a função de máxima verossimilhança pode ser substituída por minimizar $d_i[n_{i-1}, n_i-1] = -\ln\, p_i(\mathbf{x}[n_{i-1}, n_i-1]\,|\,\hat{\theta}_i)$. Deste modo, seja *k* o número de segmentos que se deseja segmentar, tem-se que a função de mínimo negativo do log da verossimilhança para *k* partições é dado por (KAY, 2001):

$$D_k(m) = \min_{\substack{n_1, n_2, ..., n_{k-1} \\ n_0=0, n_k=N+1}} \sum_{1}^{k} d_i[n_{i-1}, n_i-1] \tag{3.5}$$

Assim, sendo um sinal com *m* amostras, e considerando que $1 < n_1 < n_2 ... < n_{k-1} < m$, o somatório acima é reescrito por:

$$\begin{aligned} D_k(m) &= \min_{\substack{n_{k-1} \\ n_k = L+1}} \min_{\substack{n_1, n_2, \ldots, n_{k-2} \\ n_0 = 0}} \sum_{1}^{k} d_i[n_{i-1}, n_i - 1] \\ &= \min_{\substack{n_{k-1} \\ n_k = L+1}} \min_{\substack{n_1, n_2, \ldots, n_{k-2} \\ n_0 = 0}} \sum_{1}^{k-1} d_i[n_{i-1}, n_i - 1] + d_k[n_{k-1}, n_k - 1] \\ &= \min_{\substack{n_{k-1} \\ n_k = L+1}} \left[D_{k-1}(n_{k-1} - 1) + d_k[n_{k-1}, n_k - 1] \right] \\ &= \min_{n_{k-1}} \left[D_{k-1}(n_{k-1} - 1) + d_k[n_{k-1}, m] \right] \end{aligned} \tag{3.6}$$

Impondo a condição de monotonicidade ao caminho, de modo que $k - 1 \le n_{k-1} \le m$, tem-se que $D_k(m) = \min_{k-1 \le n_{k-1} \le L} \left[D_{k-1}(n_{k-1} - 1) + d_k[n_{k-1}, m] \right]$, para *m = k-1,k,...,N-1*, na qual pode ser resolvida por programação dinâmica.

A solução para o problema original, mostrado na Figura 3.4-1, acontece quando $k = N_s$ e *m=N-1*. Na figura, é fácil verificar que o sinal é composto por trechos com níveis médios bem definidos. Deste modo, modelando o sinal como níveis DC de valor A, contaminados por um ruído branco gaussiano de média zero e variância σ^2, tem-se que:

$$p_i(\mathbf{x}[n_{i-1}, n_i - 1] \mid A_i, n_i) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=n_{i-1}}^{n_i - 1} [\mathbf{x}[n] - A_i]^2 \right] \tag{3.7}$$

e portanto

$$\begin{aligned} d_i[n_{i-1}, n_i - 1] &= -\ln p_i(\mathbf{x}[n_{i-1}, n_i - 1] \mid \mathbf{A}, \mathbf{n}) \\ &= \frac{1}{2\sigma^2} \sum_{n=n_{i-1}}^{n_i - 1} [\mathbf{x}[n] - A_i]^2 + \frac{n_i - n_{i-1}}{2} \ln(2\pi\sigma^2) \end{aligned} \tag{3.8}$$

Assim, para minimizar $d_i[n_{i-1}, n_i - 1]$, tem-se que

$$d_i[n_{i-1}, n_i - 1] = \sum_{n=n_{i-1}}^{n_i - 1} [\mathbf{x}[n] - \hat{A}_i]^2 \tag{3.9}$$

onde $\hat{A}_i = \frac{1}{n_i - n_{i-1}} \sum_{n=n_{i-1}}^{n_i - 1} \mathbf{x}[n]$ representa o nível DC do segmento.

Nota-se que minimizar a expressão $d_i[n_{i-1}, n_i - 1]$ é similar a minimizar o erro quadrático interno ao segmento.

Comparando a recursividade encontrada acima com o algoritmo LBDP, no algoritmo, $\tilde{D}_l(m)$ era a distância acumulada entre o ponto *m* do sinal de teste e o último ponto da referência de ocorrência *l*. No caso do modelo acima $\tilde{D}_l(m)$ pode ser substituído pelo acúmulo do erro quadrático até o ponto *m* do sinal, particionando o sinal de teste em *l* segmentos.

Dado $d_l(i,m) = \sum_{n=i}^{m} [\mathbf{x}[n] - \hat{A}_{i,m}]^2$ e $\hat{A}_{i,m} = \frac{1}{m-i+1} \sum_{n=i}^{m} \mathbf{x}[n]$ o algoritmo LBDP pode ser modificado para:

Início: $l = 1$ e $m = 1, 2, ..., N - L + 1$

$$D_l(m) = d_l(1, m)$$

$$F_l(m) = 1$$

para $l = 2, ..., L\max - 1$ e $m = l, 2, ..., N - L\max + 1$

$$D_l(m) = \min_{i<n} [d_l(i+1, m) + D_{l-1}(i)]$$

$$F_l(m) = \arg\min_{i<n} [d_l(i+1, m) + D_{l-1}(i)]$$

Fim:

Por recursividade, a fronteira entre os segmentos é obtida por:

$$e_L = M$$

$$e_l = \tilde{F}_{l+1}(e_{l+1}) \quad l = L-1, L-2, ..., 1$$

O algoritmo de programação linear em dois níveis, apresentado como TLDP, também pode ser modificado para o propósito de segmentação. Neste caso, cada elemento da matriz $\tilde{D}(b,e)$ é calculado como o erro quadrático no segmento que começa em *b* e termina em *e*.

Experimento 3.2: Implementação da segmentação de um sinal unidimensional por LBDP e TLDP

Foram implementados em Matlab os algoritmos LBDP e TLDP para o fim de segmentação do sinal unidimensional, onde os nomes foram substituídos por LBDPseg e TLDPseg. O sinal utilizado como exemplo, mostrado na Figura 3.4-2, foi gerado com níveis contínuos acrescidos de um ruído gaussiano de média zero e variância unitária.

Utilizando o algoritmo LBDPseg , na Figura 3.4-2, o traço em vermelho é o resultado da segmentação, dado o número de segmentos igual a 4, onde se verifica visualmente a detecção correta dos segmentos. A desvantagem deste algoritmo está no tempo de processamento. Algumas características podem diminuir a quantidade de cálculo, por exemplo, a imposição de tamanho mínimo e máximo ao segmento.

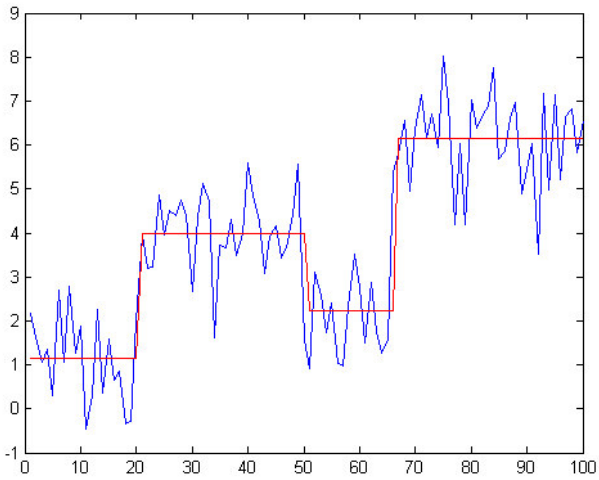


Figura 3.4-2: Resultado da segmentação pelo algoritmo LBDPseg.

O algoritmo TLDPseg apresenta o mesmo resultado que LBDPseg. A principal distinção entre os dois métodos é o cálculo prévio da matriz $\tilde{D}(b,e)$, o que anteriormente era uma desvantagem ao método. No entanto, para o fim de segmentação este fato se torna uma vantagem, pois o número de segmentos escolhidos pode ser variado sem que seja necessário recalcular a matriz $\tilde{D}(b,e)$, o que no caso do LBDP não seria possível. O algoritmo ainda pode ser otimizado quando é imposta uma

condição de tamanho mínimo e máximo para os segmentos limitando o intervalo de busca entre *b* e *e*.

No exemplo acima, o sinal representa fielmente o modelo considerado, onde a média estatística de cada segmento é bem definida e a variância é constante para todo o sinal. Na Figura 3.4-3, é mostrado o sinal utilizado no experimento 3.1, onde observa-se que o algoritmo não é capaz de segmentar o sinal da mesma maneira quando era fornecido o padrão.

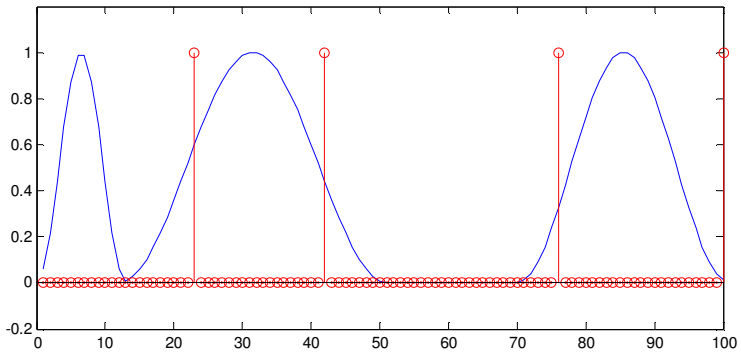


Figura 3.4-3: Segmentação de um sinal formado pela ocorrência de três padrões.

Observa-se que neste caso, o número de segmentos é dado a priori. Uma interpretação a ser explorada na proposta, é quando o número de segmentos a serem estimados é super dimensionado. Na Figura 3.4-4, é mostrado um exemplo deste caso e observa-se que o processo de segmentação se comporta como método de quantização, porém com referência temporal. Este conceito será explorado com mais detalhes posteriormente.

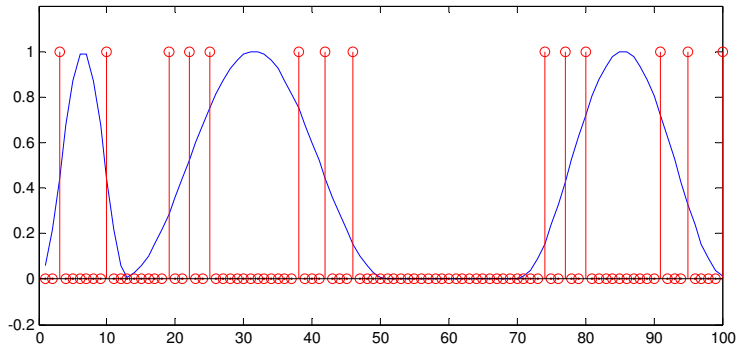


Figura 3.4-4: Segmentação com número de segmentos super-dimensionado.

3.5. Segmentação do sinal de voz

Um modelo mais aprimorado do que o modelo assumido acima seria considerar que cada segmento pode ser separado por outras características estatísticas além da média. SVENDSEN, (1987) apresenta um método de segmentação, aplicado ao sinal de voz, onde o sinal é considerado como uma seqüência de elementos quase-estacionários caracterizados por uma configuração quase-estacionária do trato vocal. Em outras palavras, dados *m* segmentos a serem segmentados, o método proposto consiste em *clusterizar* o sinal em *m* consecutivos e não sobrepostos *clusters*, ou equivalentemente, encontrar um *codebook* de tamanho *m* por quantização vetorial, sujeito à restrição de que todos os vetores contidos em um *cluster* são contínuos no tempo (SVENDSEN, 1987).

SVENDSEN, (1987) e posteriormente SHARMA, (1996) utilizaram um processo de quantização vetorial baseado no algoritmo LBDP e utilizam os coeficientes cepstrais como parâmetros para a representação espectral. Este método de quantização pode ser interpretado como o caso multidimensional do problema de segmentação apresentado no tópico anterior. Porém agora, cada suposto segmento é caracterizado por um centróide no domínio espectral ao invés da média temporal e a função a minimizar torna-se a distância acumulada entre o centróide e todos os elementos do intervalo considerado.

A seguir serão descritos os métodos para obtenção dos coeficientes cepstrais e mel cepstrais. Em seguida os métodos de segmentação LBDPseg e TLDPseg serão modificados para o caso multidimensional e alguns casos de segmentação do sinal de voz serão apresentados.

3.5.1. Cálculo dos coeficientes cepstrais

O cepstrum complexo de um sinal é definido como a transformada inversa de Fourier do logaritmo do espectro do sinal (DELLER, 1993). Para um espectro de potência $S(\omega)$, na qual é simétrico em relação a $\omega = 0$ e periódico, então:

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega} \tag{3.10}$$

onde $c_n = c_{-n}$ são reais e citados como coeficientes cepstrais.

A distância espectral entre um par de espectros $S(\omega)$ e $S'(\omega)$, pode ser obtida por (RABINER, 1993):

$$d_2^2 = \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)| \frac{d\omega}{2\pi} = \sum_{n=-\infty}^{\infty} (c_n - c_n^{'})^2 \tag{3.11}$$

Sendo o cepstrum uma seqüência com decaimento, o somatório acima não requer um número infinito de termos. Para o sinal de voz são necessários somente de 10 a 30 valores (RABINER, 1993). Os primeiros coeficientes (excluindo c_0) determina o filtro recursivo de fase mínima. Deste modo, a distância espectral fica definida como:

$$d_c^2(L) = \sum_{n=1}^{L} (c_n - c_n^{'})^2 \tag{3.12}$$

Pode ser visto que sobre certas condições regulares, os coeficientes cepstrais, exceto c_0, possuem média zero e variância inversamente proporcional ao quadrado do índice dos coeficientes, ou seja: $E\{c_n^2\} \sim \frac{1}{n^2}$. Interpretando a medida de distância como uma métrica euclidiana, podemos considerar a pesagem dos coeficientes como fator de normalização, utilizando:

$$d_{cw}^2(L) = \sum_{n=1}^{L} (nc_n - nc_n^{'})^2 . \tag{3.13}$$

Um tipo de pesagem utilizada principalmente para diminuir as características próprias do falante, presente nos coeficientes mais baixos, e a interferência da excitação presente nos coeficientes mais altos, é o processo de *liftragem*, onde é aplicada uma janela aos coeficientes. Um exemplo de janela é mostrado abaixo:

$$w(n) = \begin{cases} 1 + \frac{L}{2}\sin(\frac{n\pi}{L}) & n = 1, 2, ..., L \\ 0 & n \le 0, n > L \end{cases} \tag{3.14}$$

Em sistemas de reconhecimento independente do falante, geralmente é usada esta técnica, no entanto no caso de segmentação seu uso é questionável.
Os coeficientes cepstrais foram obtidos da seguinte maneira:

- FFT de 512 pontos,
- Log10 do módulo da FFT,
- IFFT de 512 pontos,
- Pesagem ou liftragem conforme enunciado acima com L=16.

3.5.2. Cálculo dos coeficientes mel-cepstrais

A percepção não-linear da freqüência tem levado a um modelo computacional objetivo que fornece um mecanismo para converter a medida física do espectro de um dado som, para um "espectro subjetivo" (RABINER, 1993).

Um dos métodos de simulação do espectro subjetivo é o uso de banco de filtros triangulares espaçados uniformemente em um escala não-linear da freqüência modificada, tal como a escala mel ou Bark (RABINER, 1993).

O pitch percebido (ou subjetivo) em função da freqüência, se mostra aproximadamente linear abaixo de 1000 Hz. Acima deste valor a relação entre o pitch percebido e a freqüência é aproximadamente linear com o logaritmo da freqüência (DELLER, 1993).

Originalmente DAVIS, (1980) usou filtros triangulares, onde os coeficientes mel cepstrais (*mel-frequency cepstrum coefficients,* MFCC) são obtido através da relação:

$$c_n = \sum_{k=1}^{K} (\log S_k) \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right] \qquad n = 1, 2, ..., L \tag{3.15}$$

onde S_k é a energia da saída do filtro *k*, dentre os *K* filtros e *L* o número de coeficientes desejados.

O cálculo da energia de cada filtro suaviza o espectro reduzindo os efeitos da excitação periódica, enquanto a escala de freqüência modificada oferece uma sensibilidade variável ao espectro, inspirado no sistema auditivo humano (DAVIS, 1980).

De acordo com SLANEY, (1998), o banco de filtros é construído usando 13 filtros linearmente espaçados (com diferença de 133,33 Hz entre as freqüências centrais) seguidos de 27 filtros espaçados em escala logarítmica (separados por um fator de 1.071103 na freqüência), totalizando 40 filtros. Na Figura 3.5-1 abaixo, é mostrada a responda em freqüência dos 40 filtros.

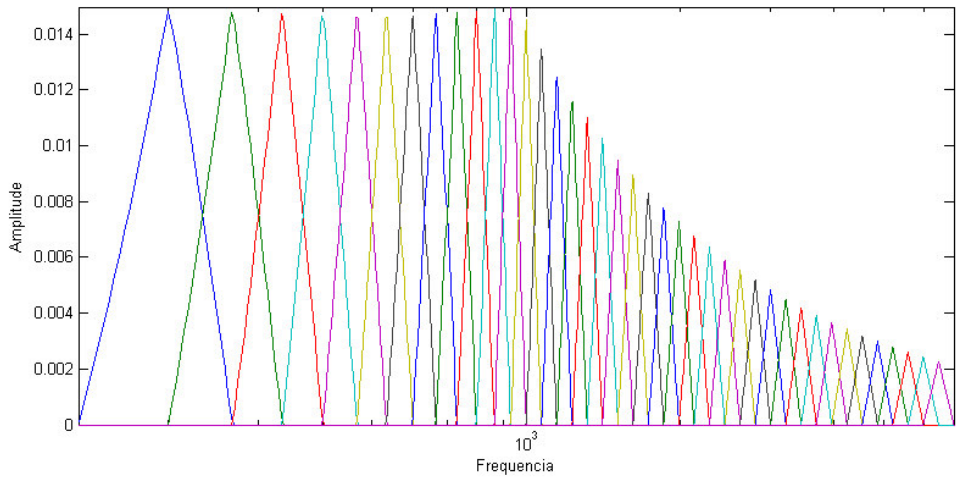


Figura 3.5-1: Resposta em freqüência dos 40 filtros triangulares utilizados para obter os coeficientes mel cepstrais.

Em resumo, a obtenção dos coeficientes é feita nos seguintes passos:

- FFT de 512 pontos;
- Obtenção da energia de saída de cada filtro pela combinação do módulo da FFT com a resposta em freqüência de cada filtro triangular;
- Aplica-se o log10 aos valores obtidos acima;
- Aplica-se a DCT para obter os *L* coeficientes mel-cepstrais.

De maneira análoga, a distância na freqüência mel-cepstral fica definida como (RABINER, 1993):

$$d_{\tilde{c}}^2(L) \quad = \sum_{n=1}^{L} (c_{\tilde{n}} - c_{\tilde{n}}^{'})^2 \tag{3.16}$$

As mesmas considerações de pesagem e *liftragem* podem ser consideradas também para os coeficientes mel cepstrais.

Experimento 5.3: Implementação do método de segmentação do sinal de voz por DP.

Foram implementados em Matlab, os algoritmos chamados de LBDPsegQ e TLDPsegQ, onde inicialmente foram utilizados os coeficientes cepstrais como representação do espectro e em seguida foram utilizados os coeficientes mel-cepstrais. É importante ressaltar que neste caso, nem todos os artifícios usados nos sistemas de reconhecimentos podem ser úteis, pois neste caso não importa a variabilidade do locutor ou outros fatores independentes do locutor.

Nos algoritmos, o erro quadrático acumulado no segmento, dado por $d_i[n_{i-1}, n_i - 1]$, torna-se o acúmulo da distância euclidiana entre o vetor centróide e os vetores do segmento, dado por:

$$d_i[n_{i-1}, n_i - 1] = \sum_{n=n_{i-1}}^{n_{i-1}} \left(\mathbf{c_n} - \mathbf{c_i}\right)^2 \tag{3.17}$$

onde $\mathbf{c_i}$ é o vetor centróide, ao longo do segmento $[n_{i-1}, n_i - 1]$ e $\mathbf{c_n}$ é o vetor correspondente à *n-ésima* janela de análise interna ao segmento.

Inicialmente o sinal foi janelado aplicando uma janela de Hanning e em seguida foram extraídos 16 coeficientes cepstrais de cada janela de análise. O tamanho da janela e o intervalo entre janelas são parâmetros a serem escolhidos de acordo com algumas questões: uma janela muito extensa apresenta boa resolução espectral, no entanto torna o encontro de fronteiras impreciso; por outro lado, uma janela estreita apresentaria uma fronteira precisa se não fosse a baixa resolução espectral. O passo curto entre janelas, ou seja, longa sobreposição, melhora o problema de precisão nas fronteiras, no entanto aumenta consideravelmente o custo computacional. Um tamanho usual para a janela, encontrado na literatura (DELLER, 1993) varia entre 10 e 20 ms e um passo entre

janelas de 5 ms. Nos exemplo a seguir foram utilizadas 10 ms por janela, com intervalo entre janelas de 5 ms, os quais apresentaram ser uma escolha razoável.

Na Figura 3.5-2 é mostrado o sinal da seqüência de sons [_ae_] (o símbolo "_" indica silêncio). Os sons foram produzidos propositadamente sustentados de modo a caracterizar períodos estacionários. O algoritmo LBDPsegQ foi executado para encontrar 4 segmentos. Na figura inferior, estão marcadas no espectrograma as janelas finais (inclusive) de cada segmento e na figura superior, estão marcados, ao longo das amostras, os intervalos entre o inicio do último frame do segmento até a sobreposição com o frame seguinte.

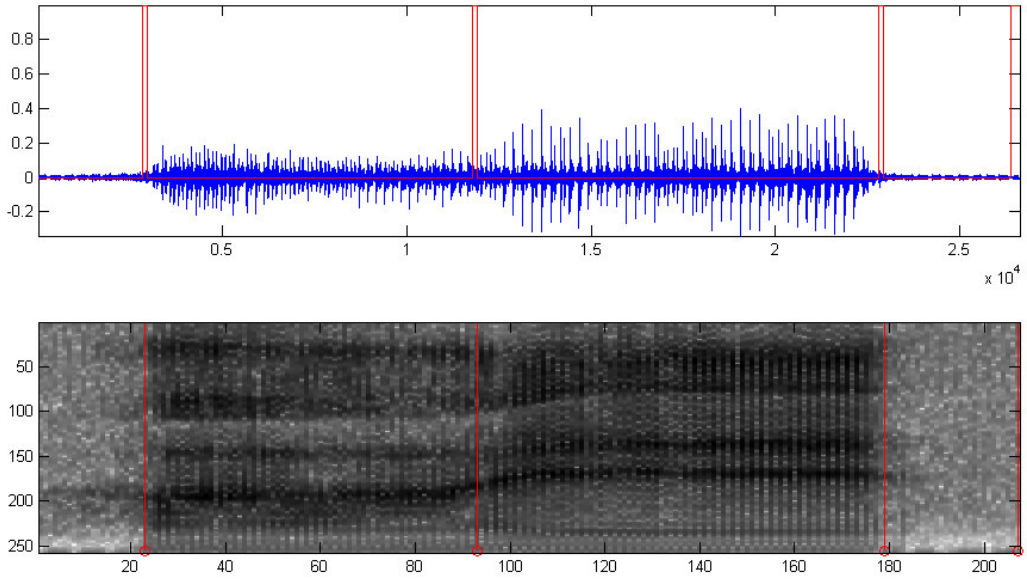


Figura 3.5-2: Segmentação do sinal [_ae_]. Os traços em vermelho indicam as fronteiras dos segmentos.

Para este caso verifica-se que os segmentos selecionados pelo algoritmo parecem coerentes com as fronteiras entre os segmentos fonéticos, no entanto, este exemplo não representa um sinal de fala natural. Na Figura 3.5-3, é mostrada a segmentação da palavra "sopa" [_ʃopa_], pronunciada naturalmente. Observa-se que as marcas inseridas parecem novamente coerentes, no entanto este é um caso onde existe pouca influência

contextual entre segmentos e boa estacionaridade em cada segmento, conforme pode ser observado no espectrograma.

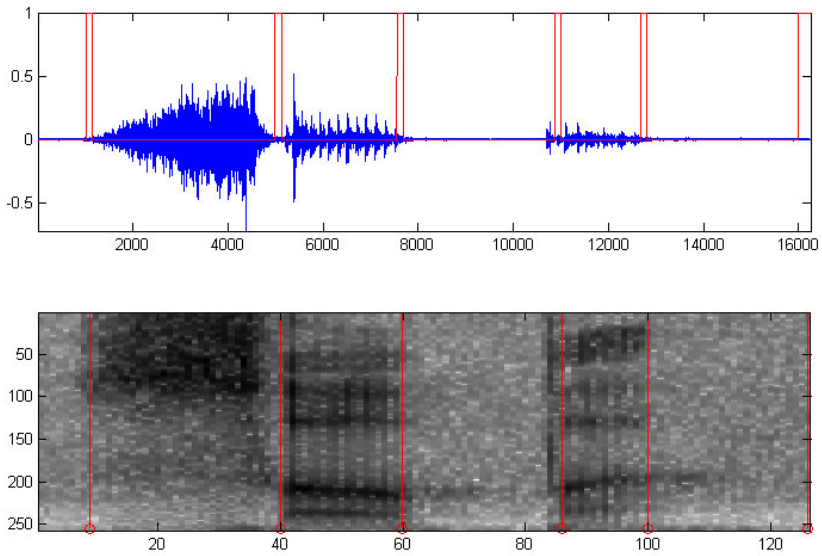


Figura 3.5-3: Sinal e o espectrograma da palavra sopa, segmentada em 6 segmentos.

No exemplo da Figura 3.5-4, é mostrada a palavra "teimoso" [_teɪmozu_], pronunciada mais rapidamente, no qual representa um caso crítico devido a alta coarticulação entre segmentos, tornando-os difíceis de serem marcados.

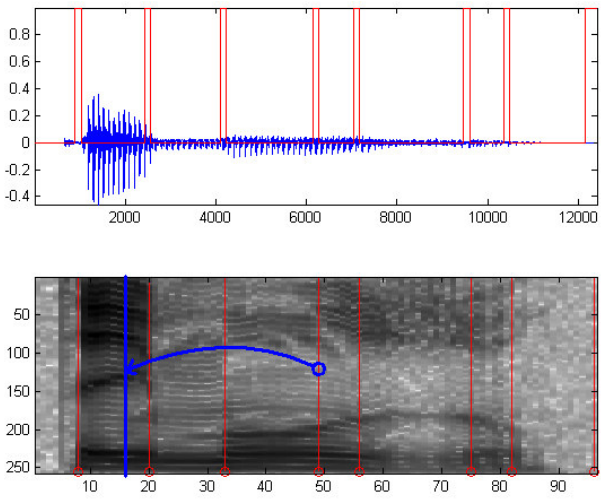


Figura 3.5-4: Sinal da palavra "teimoso" [_teɪmozu_] segmentado em 8 segmentos.

Sendo o número de segmentos determinados a priori, 8 no caso, o algoritmo selecionou os segmentos de maneira inadequada, sob o ponto de vista fonético. Neste exemplo, observa-se no espectrograma que o algoritmo "prefere" marcar a região de transição [oz] como um segmento, ao invés de separar os segmentos [e] e [ɪ], considerando-os como um único segmento.

Nota-se que em um segmento onde é difícil definir uma região estável, o algoritmo tem a tendência de particionar esta região em pequenos segmentos. Este fato pode ser observado claramente na Figura 3.4-4, o que é um comportamento esperado para um processo de quantização, porém inconveniente para a proposta de segmentação. Observa-se ainda que se existir um único segmento com esta característica, inserido no sinal a ser segmentado, resulta que algumas marcas serão deslocadas para este segmento buscando minimizar o erro da quantização. É o que se observa na Figura 3.5-4.

3.5.3. Super-segmentação

Um caso a ser observado é a super-segmentação, ou seja, o aumento do número de segmentos a serem encontrados. No exemplo da Figura 3.5-5, tem o espectrograma do sinal [_ae_] onde o número de segmentos a serem encontrados é aumentado de 4 (vide Figura 3.5-2) para 5. Observa-se que a região de transição é marcada como um segmento.

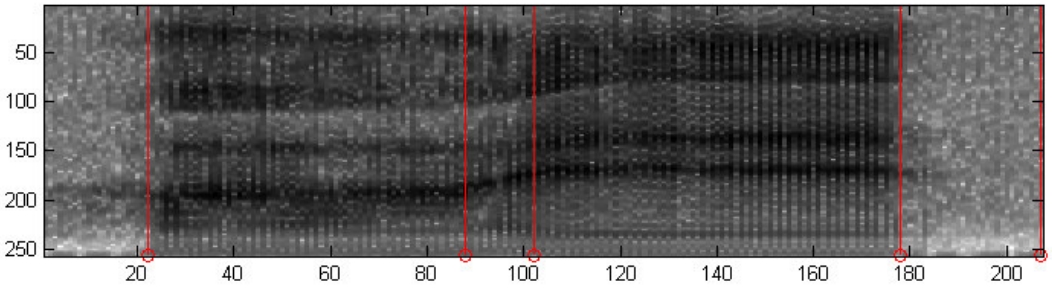


Figura 3.5-5: Super segmentação do sinal [_ae_].

A super-segmentação pode ser interpretada como um aumento na resolução do processo de quantização. Vários graus de super-segmentação podem ser definidos, dado o

número de segmentos desejados a priori. Por exemplo, na Figura 3.5-6, é mostrado o espectrograma da palavra "teimoso", cujo sinal foi segmentado em 12 segmentos ao invés de 8 (vide Figura 3.5-4), o que equivale a uma supersegmentação de 50%. Neste caso, observa-se a separação dos segmentos [e] e [ɪ], no entanto são selecionados outros segmentos que não possuem correspondência com segmentos fonéticos.

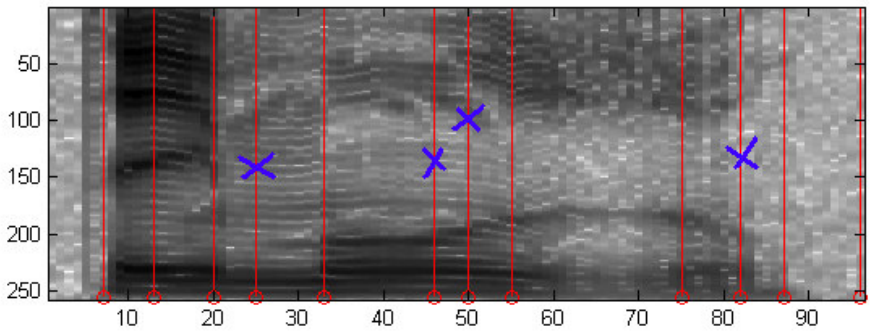


Figura 3.5-6: Espectrograma do sinal da palavra teimoso segmentado em 12 segmentos.

Em relação à coincidência entre os segmentos fonéticos e os segmentos determinados pelo algoritmo, SVENDSEN, (1987) comparou as marcas obtidas automaticamente com marcas fonéticas inseridas manualmente, chegando a conclusão que era necessária uma supersegmentação de 75% para que 99% das marcas manuais aparecessem com um erro de 30 ms. Observando que o algoritmo seleciona segmentos de transição como um segmento, é esperado que em alguns casos uma fronteira fonética posicionada manualmente esteja contida dentro de um segmento selecionado. A supersegmentação faz com que nestes casos, um segmento de transição venha a ser dividido, onde a divisão faz com que a coincidência com a marca fonética venha a aparecer. No entanto pode ser que isto aconteça somente para altos níveis de super-segmentação.

Na Figura 3.5-7 é mostrado o acúmulo do erro quadrático em função do número de segmentos escolhidos, onde observa-se que para uma supersegmentação até aproximadamente 75%, a curva apresenta um decaimento acentuado, e inicia em seguida um decaimento suave.

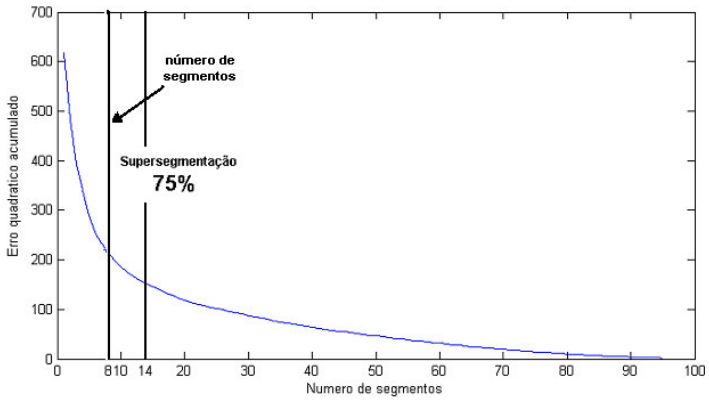


Figura 3.5-7: Erro quadrático acumulado em função do número de segmentos.

Foram utilizados os coeficientes mel-cepstrais como parâmetros, de modo a verificar o comportamento do algoritmo na colocação das fronteiras. Sendo os coeficientes mel-cepstrais mais próximos da percepção humana, espera-se que a coincidência entre as marcas fonéticas e as fronteiras entre segmentos seja maior. Porém na segmentação do sinal da palavra "teimoso" com o número esperado de segmentos, a inserção das marcas foi similar aos coeficientes cepstrais, incidindo no mesmo comportamento. Na Figura 3.5-8 é mostrado o resultado de uma supersegmentação de 100% para a mesma palavra utilizando os dois tipos de parâmetros. Comparando a colocação das marcas, verifica-se que não há uma vantagem clara em usar um ou outro tipo de parâmetro.

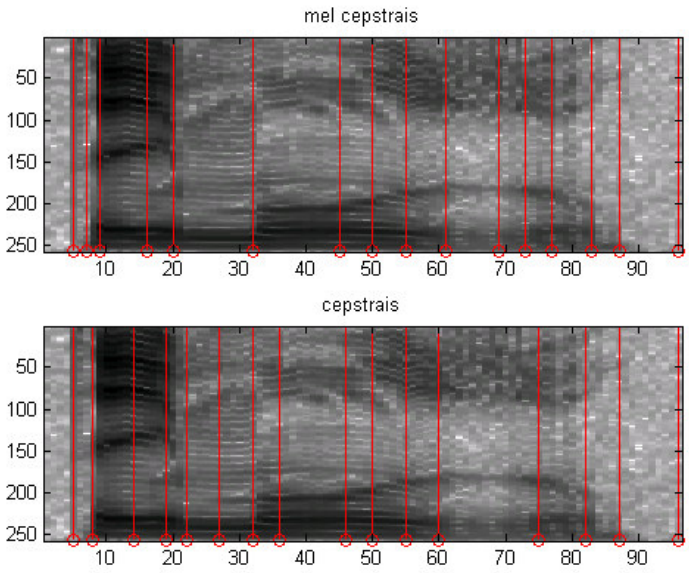


Figura 3.5-8: Supersegmentação da palavra teimoso utilizando coeficientes mel cepstrais e cepstrais.

3.6. Conclusão

Neste capítulo foram revistos e implementados os métodos utilizados no reconhecimento de dígitos conectados, nomeados de LBDP e TLDP, para um caso simplificado de reconhecer as ocorrências de um único sinal padrão dentro de um sinal de entrada.

Em seguida, foi apresentada uma abordagem teórica sobre o problema de segmentação de um sinal baseado no estimador de máxima semelhança dos parâmetros estatísticos dos segmentos, onde foi utilizada a programação dinâmica como uma solução eficiente para o problema de maximização (ou minimização como foi visto na seção 3.4).

Observada a semelhança entre os métodos LBDP, TLDP e a proposta teórica desenvolvida acima, os métodos foram modificados para efetuar a segmentação de um sinal. Neste ponto, se concluiu que o método TLDP é mais adequado a esta proposta, na qual não se encontrou referência sobre o assunto.

Em seguida, os métodos de segmentação foram adaptados com o fim de segmentar o sinal de fala baseado no conteúdo espectral do sinal. Foi concluído que o problema de segmentação sob este ponto de vista, se resume a um método de quantização ou *clusterização,* onde a seqüência temporal dos parâmetros é considerada. HOSOM, (2000) referencia esta metodologia por *sequence constrained vetorial quantization* (SCVQ). A função a critério a minimizar foi a soma dos erros quadráticos e de acordo com HART, (1973) o método pode ser classificado como um método de partição por *mínima variância*.

Foram também descritos e implementados os algoritmos para obtenção dos coeficientes cepstrais e mel-cepstrais. Sendo os coeficientes mel-cepstrais mais próximos da percepção humana, esperava-se que a coincidência entre as marcas fonéticas e as fronteiras entre segmentos fosse maior. Os resultados, porém, indicaram uma certa equivalência entre os coeficientes cepstrais e mel-cepstrais neste tipo de problema.

Os exemplos apresentados, com exceção do primeiro ([_ae_]), tratam-se de segmentos de fala natural, onde não foi imposta nenhuma condição ao falante. Porém o método de segmentação descrito tem por objetivo auxiliar na construção de um banco de unidades para síntese por concatenação. Neste caso, se forem utilizados logotomas para o recorte das unidades, acredita-se que os resultados da segmentação podem oferecer uma boa estimativa para as fronteiras entre segmentos, pois estes são falados de maneira bem mais articulada do que na fala natural. Isto será verificado adiante nesta tese, no Capítulo 5.

Será visto também no Capítulo 5 que a redução do intervalo de segmentação, pode resultar na inserção de marcas de forma mais satisfatória. Isto pode ser feito inicialmente através de uma detecção de sonoridade, de modo que a segmentação ocorrerá limitada aos intervalos de sonoro ou surdo. No Capítulo 4 a seguir será apresentado um método para detecção precisa de marcas de pitch, e conseqüentemente de sonoridade, no qual os resultados serão utilizados no Capítulo 5.

Foi observado que o método é consistente como quantizador vetorial, porém as fronteiras inseridas entre segmentos nem sempre correspondem às fronteiras fonéticas. Verificou-se que aumentando o número de segmentos a serem encontrados, processo na qual denomina-se de super-segmentação, ocorre maior coincidência com as marcas fonéticas. Na prática o fato da segmentação acústica não coincidir com as marcas fonéticas, não se deve a um erro de estimação, pois nem sempre esta correspondência existe (HOSOM, 2000).

Capítulo 4

Obtenção das marcas de pitch

4.1. Introdução

O algoritmo TD-PSOLA é baseado na técnica de *overlap and add* (OLA), na qual através do janelamento do sinal de maneira síncrona com o pitch, reconstrói-se um sinal periódico com diferente escala temporal e/ou pitch. O janelamento síncrono é feito aplicando ao sinal janelas centradas em marcas de pitch que são introduzidas das seguintes formas: nos trechos sonoros as janelas são inseridas na posição de um evento específico no ciclo de pitch, e nos trechos surdos são regularmente espaçadas (DUTOIT, 1997).

A qualidade oferecida pelo TD-PSOLA no contexto de síntese por cópia é perto da perfeição. Porém quando o algoritmo é utilizado na modificação de segmentos concatenados, provenientes de outros contextos, se as marcas não são posicionadas de

forma consistente, o resultado são erros de fase na superposição das janelas, principalmente na vizinhança de concatenação, conforme é mostrado na Figura 4.1-1 (DUTOIT, 1997)

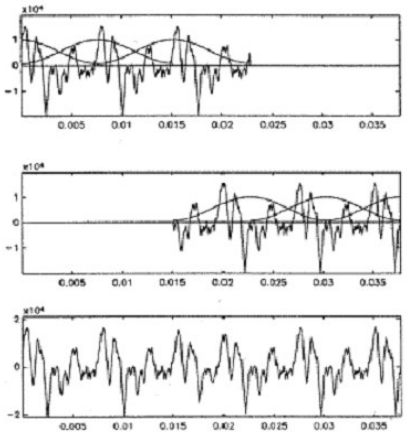


Figura 4.1-1: Erro de concatenação devido à irregularidade no posicionamento das marcas de pitch (DUTOIT, 1997).

Para a determinação das marcas, podem ser usados algoritmos de detecção de pitch chamados de PDAs (*pitch detection algorithms*). Estes algoritmos podem ser divididos em duas categorias: que operam no domínio do tempo, tal como o método SIFT (*simple inverse filter tracking*) (DELLER, 1993) ou por análise em termo curto, tal como os métodos de autocorrelação (DELLER, 1993), AMDF (*average magnitude difference function*) (DELLER, 1993) e cepstrum (DELLER, 1993). Os PDAs no domínio do tempo oferecem a estimativa do pitch período a período, mas são sensíveis às degradações do sinal no segmento de análise. PDAs de termo curto, por outro lado, são mais robustos mas oferecem uma estimativa apenas do pitch médio, isto porque o método usa a similaridade do sinal entre segmentos adjacentes. Assim, se um largo número de períodos de pitch estão contidos no segmento de análise, o valor do pitch estimado é um valor médio para o segmento (KRISHNAMURTHY, 1986).

De modo geral as marcas de pitch inseridas por PDAs não são referenciadas por um evento específico no período de pitch e portanto, na concatenação de segmentos é comum acontecer erros de fase na região de concatenação. Para contornar este problema, um evento disponível no ciclo de pitch é o instante de fechamento glotal

(*glotal closure instant,* GCI). Por simplicidade, os GCIs do sinal de voz às vezes são chamados de épocas. No entanto, a detecção automática e precisa deste evento apresenta grande dificuldade.

Na construção do banco de unidades para síntese por concatenação utilizando o TD-PSOLA, usualmente as marcas de pitch são inseridas nos GCIs de forma semi-automática, isto é, inicialmente é feita uma indicação automática prévia seguida de inspeção visual para a eliminação de erros de inserção ou para a inclusão de marcas adicionais necessárias. Porém esta inspeção visual requer um grande consumo de tempo (DUTOIT, 1997).

Uma solução possível para a detecção precisa do instante de fechamento glotal é o uso de um equipamento chamado *eletroglotógrafo.* Este equipamento mede a atividade das cordas vocais, na qual é possível caracterizar, com maior precisão, os GCIs. Este equipamento tem sido usado intensivamente na construção de banco de unidades (LENZO, 2002), porém este é um equipamento de custo alto.

Este capítulo consiste na implementação de um método de detecção de forma robusta e simples, como uma alternativa de baixo custo para o *eletroglotógrafo*. Isto será feito utilizando um captador de vibração em contato com a garganta, chamado de microfone de contato, supondo que este possa oferecer maior informação sobre o ciclo glotal do que o próprio sinal de voz, aumentando assim a possibilidade de uma detecção automática. Esta proposta já foi considerada como um meio de obtenção da informação de pitch com finalidade médica em (ASKENFELT, 1980). Também se observa o uso de microfones em contato com o pescoço, chamados de *throat-microphone*, para telefones celulares, na captação da voz em ambientes ruidosos e ainda como informação adicional para alguns sistemas de reconhecimento (GRACIARENA, 2003).

Na organização do capítulo, na seção 4.2. serão apresentados e implementados dois métodos clássicos de detecção automática dos GCIs, ou de detecção de épocas, a partir do sinal de voz, considerados precursores de várias outras propostas. Na seção 4.3. será feita uma breve apresentação do instrumento *eletroglotógrafo* e da sua aplicabilidade. Será também descrito nesta seção um método de obtenção dos GCIs a partir do sinal

proveniente deste instrumento. Na tentativa de estimar os GCIs diretamente a partir do sinal de voz, usando o sinal do EGG somente em uma etapa de treino, serão realizados também alguns experimentos de estimação utilizando redes neurais.

Na seção 4.4. será proposto o método para obtenção dos GCIs a partir do sinal de um microfone de contato. Será apresentado o dispositivo utilizado como microfone, incluindo o circuito de pré-amplificação para captação simultânea do sinal de contato e do sinal de voz, a serem adquiridos por uma placa de som convencional para computador. Serão realizados experimentos de detecção automática dos GCIs, a partir do sinal obtido pelo contato, utilizando os métodos descritos na seção 4.2., onde serão apresentados os resultados para alguns casos críticos de detecção. Os resultados indicam que a detecção usando o equipamento proposto é uma alternativa eficaz para a obtenção dos GCIs, em unidades de síntese por concatenação temporal, tendo sido apresentada em (LATSCH, 2005).

4.2. Métodos de detecção automática dos GCIs

Muitos algoritmos precursores da detecção automática de épocas no sinal de voz, entre eles (STRUBE, 1974) e (WONG, 1979), (ANANTHAPADMANABHA, 1979) e (CHENG, 1989), estão baseados essencialmente na idéia de que em segmentos de análise curtos (menores do que um período de pitch), quando um instante de excitação ou um GCI está incluído no segmento, o erro de predição linear é maior. Deste modo, o ponto onde ocorre um grande erro de predição indica um GCI (CHANGXUE, 1994).

A partir da suposição de que o resíduo de predição linear exibe picos correspondentes aos GCIs, ANANTHAPADMANABHA, (1979) observa que existe ambigüidade na detecção direta dos picos e propõem um método para reduzir estas ambigüidades. Semelhante à teoria de filtros casados, CHENG, (1989) propõe a detecção dos GCIs a partir de máximos na função de verossimilhança entre um sinal modelo e o sinal de voz, apresentando boa performance ao ruído, porém com uma imprecisão sistemática. Entre os trabalhos mais atuais destaca-se (SMITS, 1995), baseado no fato de que o valor

médio da função de atraso de grupo dentro de um segmento de análise indica a ocorrência de um GCI.

Nas seções 4.3.1 e 4.3.2 os métodos propostos por ANANTHAPADMANABHA, (1979) e CHENG, (1989), são descritos e implementados. Estes métodos foram escolhidos devido à reduzida complexidade, baixo custo computacional, simplicidade e por serem considerados precursores dos métodos de detecção automática dos GCIs.

4.2.1. Estimação a partir do resíduo de predição linear

Conforme citado anteriormente, um valor alto no resíduo LP (*linear prediction*) supostamente indica a localização de um GCI. No entanto devido a alguns fatores como por exemplo, a estimação não acurada das formantes e das larguras de banda na etapa de análise, múltiplos picos podem ocorrer tornando difícil a estimação precisa dos GCIs (ANANTHAPADMANABHA, 1979).

Deste modo, o método proposto em (ANANTHAPADMANABHA, 1979) se propõe a reduzir as ambigüidades envolvidas no uso direto do resíduo LP. Inicialmente o sinal é filtrado (filtragem passa-faixa) no domínio da freqüência, aplicando-se uma janela de hanning à FFT do resíduo, para reduzir as componentes de baixa freqüência e o ruído de mais alta freqüência. Em seguida é obtido o contorno da transformada de Hilbert (*Hilbert envelope*) de modo a atenuar os efeitos de fase introduzidos na obtenção do resíduo (ANANTHAPADMANABHA, 1979).

O contorno da transformada de Hilbert de $t(n)$ é dado por: $t_0(n) = [t^2(n) + t_H^2(n)]^{1/2}$

onde $t_H(n)$ é a transformada de $t(n)$ (ANANTHAPADMANABHA, 1979).

Geralmente os algoritmos que implementam a transformada de Hilbert retornam um sinal complexo chamado de sinal analítico contendo na parte real o sinal original e na parte imaginária a transformada de Hilbert. Assim o contorno da transformada de Hilbert é equivalente ao módulo do sinal analítico.

Na Figura 4.2-1 é mostrado o diagrama em blocos do método, extraído de (ANANTHAPADMANABHA, 1979).

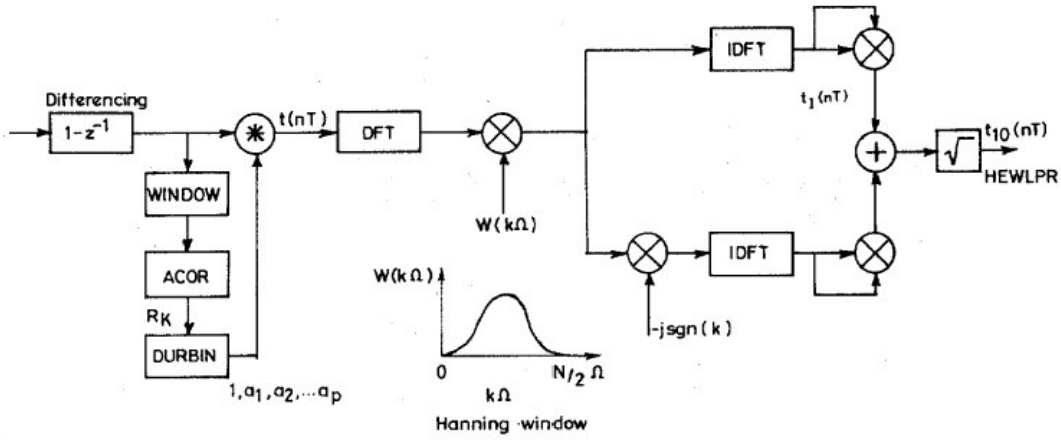


Figura 4.2-1: Diagrama em blocos do método, extraído de (ANANTHAPADMANABHA, 1979).

Experimento 4.3: **Implementação do método de tratamento do resíduo LP.**

O algoritmo foi implementado em Matlab nos seguintes passos:

i Pré-ênfase do sinal de voz, amostrado a 11025 kHz;

ii Segmentação do sinal com janelas do tipo Hanning, de 20 ms, com 50% de sobreposição.

iii Em seguida, para cada segmento é feito:

- a extração de 12 coeficientes de predição linear;
- b o resíduo ou erro de predição é obtido passando cada segmento através do filtro inverso (coeficientes predição de predição no numerador);
- c FFT do resíduo;
- d aplicação de uma janela de Hanning à FFT;
- e FFT inversa;
- f obtenção da transformada de Hilbert do segmento obtido no passo anterior;
- g obtenção do módulo do sinal analítico conforme descrito anteriormente;

iv O sinal resultante, referenciado como HEWLPR (*Hilbert envelope of windowed LP residual*) é montado através da sobreposição e soma dos segmentos.

Na Figura 4.2-2 são mostrados trechos de um sinal de voz, seguidos do resíduo LP correspondente e por fim a aplicação do algoritmo aos resíduos. Na Figura 4.2-2 (a) Observa-se que a ambigüidade na localização do GCI, devido a efeitos de fase, é resolvida e que as componentes de baixa freqüência são atenuadas, sendo possível detectar os GCIs com facilidade. Porém no trecho mostrado na Figura 4.2-2 (b), observa-se que nem sempre o isolamento dos GCI são triviais.

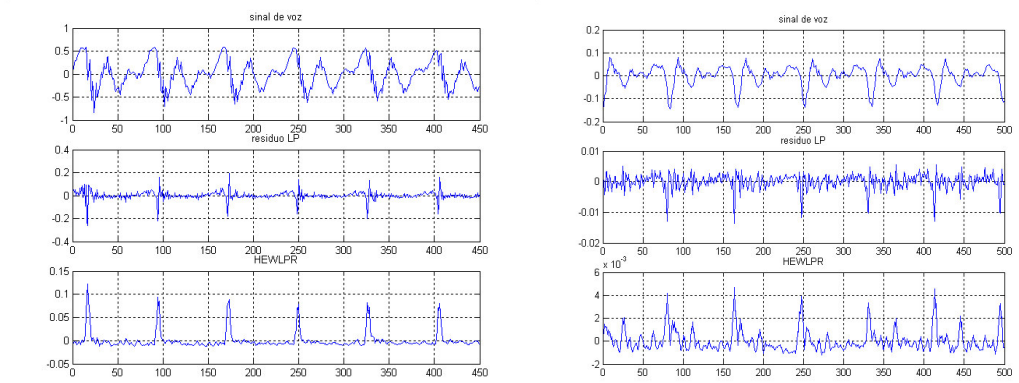


Figura 4.2-2: Sinais de voz, resíduo de predição linear e o sinal HEWLPR obtidos na aplicação do algoritmo. (a) caso ideal; (b) caso crítico de detecção dos GCIs.

ANANTHAPADMANABHA, (1979) mostra diferentes casos da aplicação do método, concluindo que o método é eficiente na diminuição das ambigüidades contidas no resíduo de predição, porém a eficiência do método depende da eficiência da etapa de análise em obter o resíduo LP. Nota-se ainda a dificuldade de isolar os GCIs de sons mistos e consoantes sonoras, devido à alta intensidade de ruído e à baixa sonoridade no sinal de voz destes segmentos. Em sinais compostos por trechos sonoros e surdos, é ainda necessária a identificação a priori dos trechos sonoros, onde somente nestes trechos será obtido o resíduo de predição. Isto porque a filtragem inversa enfatiza as altas freqüências e conseqüentemente aumentando o nível do ruído, o que torna difícil diferenciar os trechos sonoros dos surdos.

4.2.2. Estimação por máxima verossimilhança

Esta metodologia foi proposta por CHENG, (1989), para estimar os GCIs, adaptada da teoria de detecção de épocas por máxima verossimilhança em aplicações para radar. Assim como na seção anterior, este método assume que o sinal de voz dentro de um período de pitch é induzido por um pulso em uma época (ou evento) geralmente definida como a representação de um GCI.

Assumindo que a produção da voz pode ser modelada por um sistema linear AR (autoregressivo), o sinal modelo $\hat{s}(n)$ devido a uma época pode ser expresso como:

$$\hat{s}(n) = \begin{cases} \sum_{i=1}^{p} a_i \hat{s}(n-i) & 0 < n \leq \infty \\ G & n = 0 \\ 0 & n < 0 \end{cases} \tag{4.1}$$

onde *G* é uma constante positiva arbitrária, a os coeficientes do polinômio de ordem *p*.

Em seguida, é suposto que a diferença entre o sinal observado, $s(n+n_0) \quad n \in [0, N-1]$ (onde n_0 é uma seqüência de atrasos de alinhamento) e o sinal modelo é um processo gaussiano e que as N observações constroem um processo gaussiano com N dimensões independentes e variância uniforme σ.

Assim, dado $x(n) = s(n+n_0) - \hat{s}(n)$, a densidade de probabilidade condicional, ou função de verossimilhança, será descrita por:

$$p(X \mid \theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} . \exp\left\{ -\sum_{n=0}^{N-1} \left[s(n+n_0) - \hat{s}(n) \right]^2 \Big/ 2\sigma^2 \right\} \tag{4.2}$$

onde θ é o espaço de parâmetros $\theta = \{\sigma, a_1, a_2, a_3, ..., a_p, n_0\}$.

Deste modo, quando o valor dos parâmetros maximizar a função de verossimilhança, significa que uma época ocorreu. Maximizar a função de verossimilhança pode ser

substituído por maximizar o seu logaritmo e portanto a função a ser maximizada torna-se:

$$\ln\left[p(X \mid \theta)\right] = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{\sum_{n=0}^{N-1}\left[s(n+n_0) - \hat{s}(n)\right]^2}{2\sigma^2} \tag{4.3}$$

Assim conclui-se que (CHENG, 1989):

- Não é possível encontrar uma expressão explícita para um valor ótimo de n_0. Porém, resolvendo algebricamente a potência interna ao somatório, e observando as possibilidades de máximos na função de $\ln[p(X \mid \theta)]$ em função de n_0, tem-se que o termo $\sum_{n=0}^{N-1}\left[s(n+n_0)\hat{s}(n)\right]$ domina. Este termo é chamado de sinal MLED (*maximum-likelihood epoch determination*) e trata-se da correlação cruzada entre o sinal de voz e o sinal modelo. Portanto, em função de n_0, os máximos no sinal MLED equivalem aos máximos na função de verossimilhança,

- Em segundo lugar, os coeficientes do sinal modelo que produzem um máximo na função de verossimilhança são deduzidos como os coeficientes de predição linear obtidos pelo método da autocorrelação. Neste sentido, o sinal modelo é considerado como os coeficientes de um filtro casado.

Aplicando o método ao sinal de voz, o pulso positivo de maior amplitude indica o GCI dentro de um período. Porém, os autores observaram experimentalmente que o indicativo para o GCI é melhor definido a 50% da amplitude do máximo (de zero até o ponto máximo do pulso, à esquerda) e este critério é empírico (CHENG, 1989).

A seguir o método de obtenção do MLED foi implementado em Matlab para verificação do método.

Experimento 4.4: obtenção do sinal de máxima verossimilhança (MLED).

O algoritmo foi implementado nos seguintes passos:

i. Segmentação do sinal com janelas do tipo Hamming de 20 ms de largura e com 50% de sobreposição.

ii. Em seguida, para cada segmento é feito:

   a extração de 12 coeficientes de predição linear;

   b o sinal modelo de 5 ms é gerado obtendo a resposta ao impulso do filtro AR com os coeficientes obtidos acima;

   c é obtida a correlação cruzada entre o segmento e o sinal modelo.

iii. O sinal MLED de todo o sinal de voz é montado através da sobreposição e soma dos segmentos obtidos na correlação acima.

Foram utilizados sinais sintéticos para verificar a precisão do método, pois os GCIs são conhecidos a priori. Os sinais sintéticos foram gerados segundo a teoria fonte-filtro, a qual o sinal de voz é produzido pela excitação de um filtro digital que modela o trato vocal. A excitação foi obtida conforme o modelo LF (*Liljencrants and Fant*) para a derivada do fluxo do volume de ar glotal. Cada período no sinal de excitação possui uma fase de abertura da glote seguida do instante de fechamento glotal (representado por um pico negativo) e uma fase onde a glote permanece fechada. O filtro AR que modela o trato vocal é determinado a partir das freqüências e larguras de banda de 5 formantes escolhidas (LATSCH, 2002).

Na Figura 4.2-4 são dados dois exemplos de sinais de voz sintéticos, gerados pelos sinais de excitação mostrados logo abaixo e por último os sinais MLED. A imprecisão no posicionamento do GCI relatada pelos autores pode ser verificada nas figuras, porém o critério de correção sugerido pelos autores nem sempre é eficiente.

Em um período do sinal MLED aparecem não só os máximos locais correspondentes aos GCIs, mas também outros máximos que correspondem a falsos candidatos. A razão em amplitude entre o pulso principal e os outros pulsos varia substancialmente e depende das propriedades do sinal, criando ambigüidade na decisão (CHENG, 1989). Para contornar este problema, os autores propõem o uso de um "sinal de seleção",

similar à aplicação de uma janela, para enfatizar o contraste entre o pulso principal e os pulsos secundários.

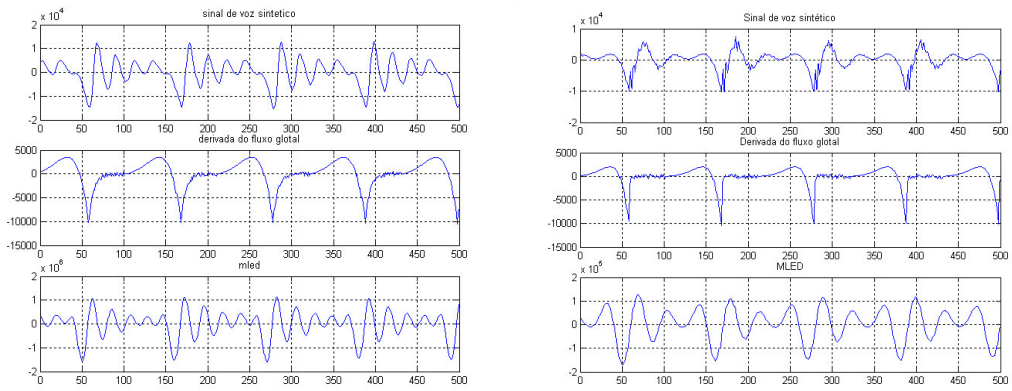


Figura 4.2-3: Sinal de voz sintético, derivada do fluxo de ar glotal e sinal MLED. (a) caso ideal; (b) caso crítico.

Seja $\hat{f}(n_0)$ o sinal MLED e $g_{\triangle}(n_0)$ o sinal de seleção desejado, os autores demonstram que o contorno da transformada de Hilbert (ou módulo do sinal analítico) do sinal MLED pode ser utilizado como sinal de seleção. Observa-se que neste caso, diferente da seção 4.3.1., o contorno de Hilbert é utilizado para enfatizar os pulsos no sinal MLED.

A média de $g_{\triangle}(n_0)$ pode ainda ser subtraída para tornar o sinal de seleção mais parecido com um pulso, sendo possível anular o sinal entre pulsos. Assim,

$$\hat{g}_{\triangle}(n_0) = \begin{cases} g_{\triangle}(n_0) - \overline{g_{\triangle}(n_0)} & \text{se} \quad g_{\triangle}(n_0) \geq \overline{g_{\triangle}(n_0)} \\ 0 & \text{se} \quad g_{\triangle}(n_0) < \overline{g_{\triangle}(n_0)} \end{cases} \tag{4.4}$$

Onde $\overline{g_{\triangle}(n_0)}$ é a média aritmética das amostras.

Assim, o sinal de determinação dos GCIs, chamado aqui de HMLED (*Hilbert envelope of MLED*) é obtido por $\hat{f}(n_0) \cdot \hat{g}_{\triangle}(n_0)$. O método de seleção do máximo principal foi acrescentado ao algoritmo descrito no experimento 4.4 e demonstrado a seguir.

Experimento 4.5.: método de seleção dos GCIs.

Para cada segmento do sinal MLED obtido no item ii.c, foi encontrado o sinal de seleção descrito anteriormente. Na Figura 4.2-4, são mostrados respectivamente os sinais MLED, referentes aos sinais sintéticos da Figura 4.2-3, os sinais de seleção obtidos pelo módulo do sinal analítico gerado pela transformada de Hilbert e por fim os sinais de determinação dos GCIs.

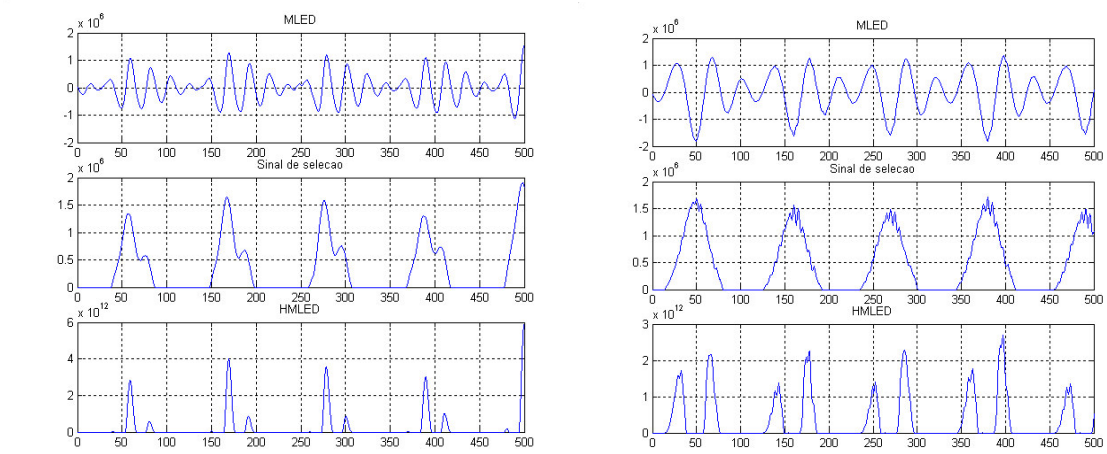


Figura 4.2-4: Sinal de voz sintético, sinal de seleção e sinal de determinação do GCIs. (a) caso ideal; (b) caso crítico.

Observa-se que na primeira figura os máximos são isolados com relativo sucesso, porém na segunda figura os máximos mais próximos também são enfatizados, permanecendo a ambigüidade, porém como possuem amplitudes menores, são possíveis de serem eliminados.

Na Figura 4.2-5 é mostrada a aplicação do método a um sinal de voz real, gravado em condições de ruído ambiente, contendo trechos sonoros e surdos. Observa-se que é possível estabelecer um limiar em amplitude de modo a eliminar a maioria dos falsos GCIs nos trechos surdos, porém eliminando também alguns verdadeiros GCIs de baixa amplitude nos trechos sonoros. Uma forma de amplificar os pulsos em trechos de baixa amplitude é normalizar a correlação cruzada na obtenção do sinal MLED. Isto torna as amplitudes dos pulsos mais uniformes, como pode ser visto no terceiro sinal da Figura 4.2-5, porém aumenta a amplitude nos trechos surdos, o que torna a escolha de um

limiar de amplitude mais difícil de ser estabelecido. Uma solução para este fato é determinação prévia das regiões de sonoridade.

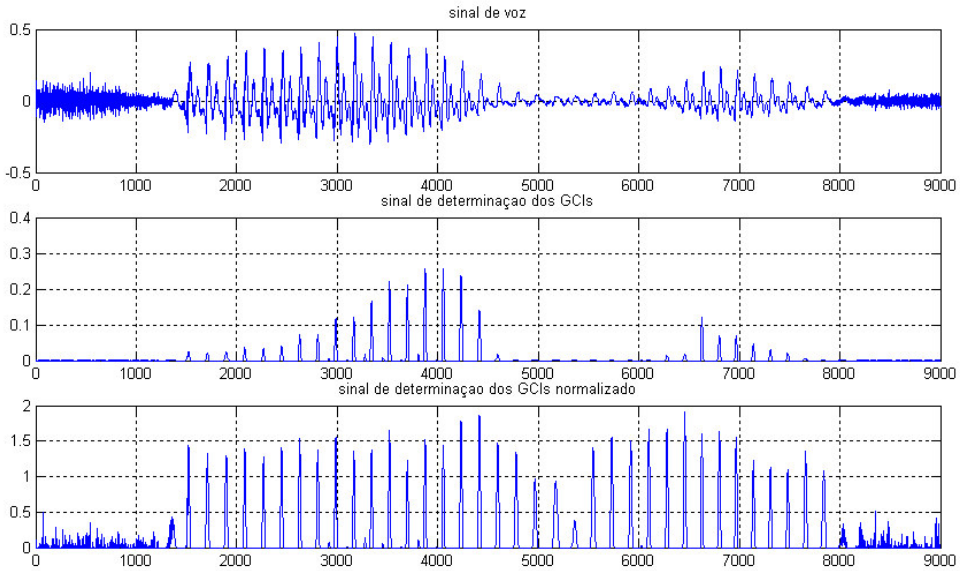


Figura 4.2-5: Sinal de voz real, sinal de determinação dos GCIs e com a normalização da correlação cruzada.

A estimação dos GCIs no trecho do sinal mostrado na Figura 4.2-5 pode ser considerada bem sucedida apesar da imprecisão no posicionamento dos GCIs. Foi observado ainda que o método é eficiente na presença de ruído, capaz de sugerir os GCIs para sons mistos, porém com pulsos de pequena amplitude. No entanto, os resultados do método mostrado na figura não se mantêm para todo o sinal, onde foram observados erros grosseiros na indicação dos GCIs.

Um fato interessante que pode ser verificado na Figura 4.2-4 é que o máximo no sinal de seleção, produzido pelo módulo da transformada de Hilbert do sinal MLED, é um indicativo mais próximo do GCIs verdadeiro do que o máximo no sinal MLED.

4.3. Eletroglotógrafo

A eletroglotografia é uma medida de impedância elétrica (KRISHNAMURTHY, 1986). O instrumento Eletroglotógrafo (do inglês *eletroglottograph,* também conhecido como *laryngograph*), mede a variação da impedância elétrica da laringe usando um par de eletrodos mantidos em contato com a pele em ambos os lados da cartilagem da tiróide, vide Figura 4.3-1 (a). A operação básica do circuito consiste na geração de um sinal de rádio freqüência, da ordem de 1 a 3 MHz, com corrente constante, o qual será modulado em amplitude pela variação da impedância do tecido devido à vibração das cordas vocais. Os eletrodos, montados com anéis de guarda para a redução do ruído, mostrados na Figura 4.3-1 (b), detectam o sinal modulado que será então demodulado por um circuito detector. A percentagem de modulação no sinal reflete a mudança percentual na impedância do tecido, tipicamente 0,1% em torno de 150 ohms. A forma de onda demodulada, chamada de EGG, é por hipótese uma função da área lateral de contato entre as cordas vocais (KRISHNAMURTHY, 1986).

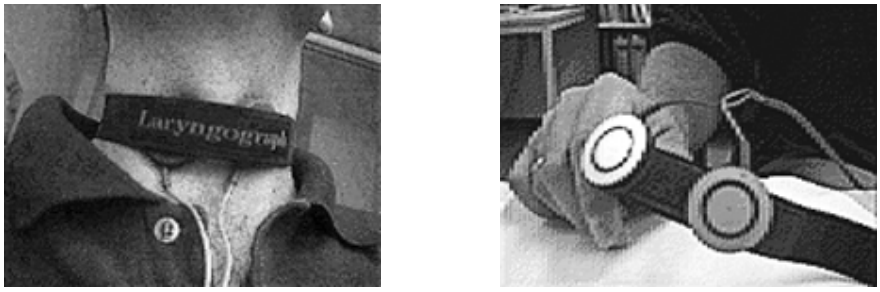


Figura 4.3-1: (a) Colocação dos eletrodos; (b) Eletrodos.

Observam-se na Figura 4.3-2, o sinal de voz e o sinal EGG gravados em simultâneo, e em seguida a derivada do sinal EGG (DEGG). O GCI é marcado, em um período, por um rápido decréscimo no EGG coincidindo com um pico em amplitude na derivada. Assim o mínimo no DEGG ocorre perto do instante de abertura da glote (BARNER, 1994) .

Os sinais mostrados na Figura 4.3-2 e utilizados nos experimentos a seguir consistem na gravação simultânea do sinal de voz e do EGG e foram cedidos pelo Departamento de

Eletrônica e Telecomunicações do Instituto de Engenharia, Electrônica e Telemática de Aveiro (IEETA) – Portugal. Os sinais foram gravados originalmente por um único falante e consiste na repetição da frase: “o pássaro toca na tuneca”.

Para classificação das regiões sonoras e inserção dos GCIs nestas regiões, o sinal EGG é uma alternativa interessante pois é extremamente simples de ser tratado, conforme descrito no experimento a seguir.

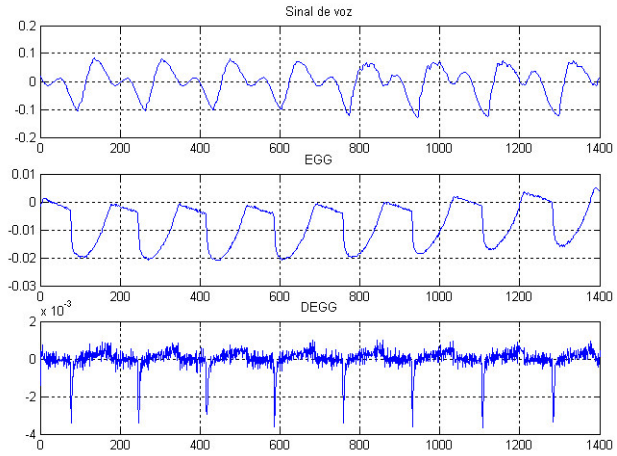


Figura 4.3-2: Trecho de sinal de voz sonoro, do sinal EGG e da derivada do EGG indicando a posição dos GCIs.

Experimento 4.1: Detecção dos GCIs a partir do sinal EGG

Os movimentos resultantes da laringe alteram a impedância vista pelos eletrodos, que acarreta em um sinal de baixa freqüência sobreposto ao sinal EGG normal. Estas baixas freqüências podem ser eliminadas aplicando-se um filtro passa altas em torno de 80 Hz. Em seguida as regiões equivalentes à sonoridade podem ser classificadas estabelecendo-se limiares para a energia e passagem por zero de segmentos curtos e assim somente nas regiões classificadas como sonoras os GCIs são detectados (KRISHNAMURTHY, 1986). Porém é possível a obtenção dos GCIs diretamente a partir do DEGG sem a detecção prévia dos trechos sonoros (BARNER, 1994) opção adotada e descrita a seguir:

i. Primeiramente, o EGG é derivado por um filtro passa-altas FIR. Este filtro é suficiente para obter o sinal DEGG e para eliminar as baixa freqüências provocadas pelo movimento da laringe.
ii. Em seguida, o sinal é invertido e limitado pela amplitude do trecho inicial de ruído.
iii. Os máximos no sinal resultante representam os GCIs. Através da busca local de pontos máximos, com janelas de 8 ms com superposição de 50%, as marcas do GCIs são pontualmente definidas.
iv. Devido ao atraso do sinal de voz em relação ao EGG, proporcional à distância entre o microfone e os eletrodos, as marcas dos GCIs encontradas no DEGG podem ser deslocadas uniformemente. No entanto neste trabalho, assim como em (BARNER, 1994), cada marca sugerida pelo DEGG foi indicada na derivada do sinal de voz e em seguida ajustada para o primeiro inteiro positivo após a passagem por zero.

Na Figura 4.3-3, na primeira figura é mostrado o sinal de voz da palavra "pássaro" com a aplicação do algoritmo ao sinal EGG gravado em simultâneo. Pode-se verificar a ausência de marcação nos trechos onde não há sonoridade. Na figura logo abaixo, tem-se um trecho mostrado em detalhe onde pode-se observar a precisão da inserção dos GCIs.

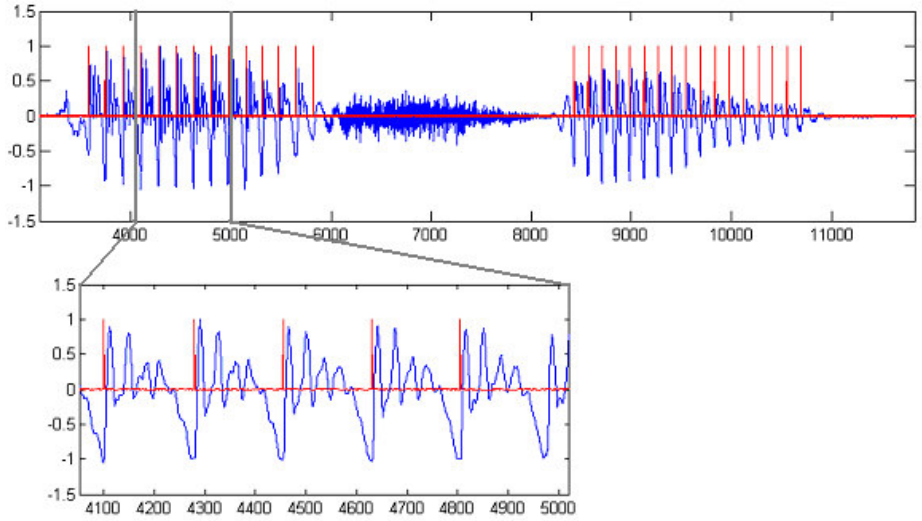


Figura 4.3-3: Sinais de voz e GCIs determinados a partir do DEGG. (a) sinal completo; (b) trecho do sinal em detalhe.

O uso do EGG para marcação dos GCIs, e conseqüentemente a estimação do pitch, comparado a outros métodos, apresenta as principais características (KRISHNAMURTHY, 1986):

- Há casos onde o EGG indica a ocorrência de regiões de excitação sonora, enquanto que no sinal de voz não é possível identificá-las. Por exemplo, na transição de regiões sonoras para surdas, onde ocorre uma redução na amplitude, chegando aos níveis de ruído da gravação, ou ainda no caso de excitações mistas.
- O EGG suporta um método extremamente preciso e fidedigno para a obtenção do contorno de pitch.

Apesar dos resultados favoráveis ao uso do sinal EGG, dependendo da aplicação, a obtenção do sinal em simultâneo com a voz nem sempre é possível (BARNER, 1994). Geralmente o eletroglotógrafo é utilizado em laboratório, nos casos em que se deseja obter um padrão para avaliar o desempenho de métodos de estimação automática dos GCIs ou de detecção de pitch somente a partir do sinal de voz. Além disso é usado ainda na construção de bancos de unidades para síntese por concatenação temporal, na inserção das marcas de pitch para o algoritmo TD-PSOLA (CHARPENTIER, 1989). A utilização do equipamento por um usuário comum é inviável devido principalmente ao custo relativamente alto (U$ 1.500,00 aproximadamente)[4]. Por estas razões, no experimento a seguir, o sinal EGG foi utilizado somente como sinal de treino em um sistema de estimação dos GCIs a partir da voz, cujo uso não seria mais necessário na fase de operação.

Experimento 4.2: estimação do GCIs a partir do sinal de voz utilizando Redes Neurais.

Para estimar os GCIs a partir do sinal de voz foi utilizada uma Rede Neural *feed-forward backpropagation* treinada com as marcas de GCIs obtidas do sinal EGG. É

[4] Orçamento fornecido pela empresa Laryngograph® através do endereço eletrônico http://www.laryngograph.com em 10/03/2004 . Os custos de importação não estão incluídos.

esperado que a rede seja capaz de indicar a posição dos GCIs e conseqüentemente indicar as regiões de sonoridade nos sinais usados para treino. Em (BARNER, 1994) os autores usaram um método para estimar o sinal DEGG a partir da voz através da otimização de um filtro não-linear adaptativo, porém os resultados apresentados são insuficientes para avaliar o método.

A melhor configuração da rede, encontrada ao longo do experimento, é mostrada na Figura 4.3-4. A rede é composta por 2 camadas, com 12 neurônios na primeira camada, com função de ativação tangente hiperbólica, e 1 neurônio na última camada, com função de ativação do tipo sigmóide. A rede possui ainda 21 entradas e 1 saída.

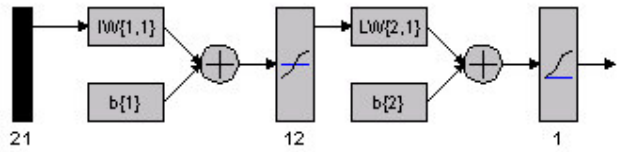


Figura 4.3-4: Arquitetura da Rede Neural

Para a formação dos conjuntos de treino da rede, os GCIs foram encontrados a partir do sinal EGG, de acordo com o experimento 4.1. e normalizados em amplitude indo de 0 e 1. Ao sinal de voz foi aplicado um filtro de pré-ênfase e em seguida, ambos os sinais foram decimados por 4, o que diminui o tempo de treino e operação. Este valor de decimação foi determinado experimentalmente de forma a obter o melhor compromisso entre tempo de treino e eficiência da estimação. Em seguida os sinais foram formatados de acordo com as entradas e saídas da rede, de forma que para as 21 amostras de entrada exista uma única saída desejada. Estas 21 amostras têm origem na suposição de que um GCI no sinal de voz possa ser detectado pelas amostras passadas e futuras a ele, que no caso foram escolhidas como 10 amostras passadas e 10 futuras, o que resulta em 21 amostras. Assim, o suposto GCI ocupa o centro da janela. O conjunto de entrada foi então obtido através do janelamento do sinal de voz, com janela retangular de 21 amostras, com deslocamento de 1 amostra. O conjunto de saída foi obtido do sinal indicativo dos GCIs com a defasagem de 10 amostras.

Foi observado que a suavização dos impulsos de marcação dos GCIs resultava em melhor performance da rede. Para isso, ao trem de impulsos dos GCIs foi convoluído um sinal simétrico com decaimento exponencial.

Dos sinais disponíveis foram selecionados 3 conjuntos: treino, validação e teste contendo 25000, 7000 e 5000 amostras respectivamente. O conjunto de validação foi utilizado como parâmetro de parada de treinamento, para assegurar a capacidade de generalização da rede. O conjunto de teste foi utilizado para avaliar a capacidade de aprendizado.

A rede foi implementada em Matlab e treinada pelo algoritmo *Levenberg-Marquardt*, que aumenta a velocidade de treino. Após o treino, para avaliação do desempenho de estimação da rede, foram utilizados critérios qualitativos, considerando a estimação bem sucedida se:

(1) for possível distinguir no sinal estimado as regiões sonoras das surdas;

(2) ocorrem nas regiões sonoras pontos de máximo na localização dos GCIs com amplitude suficiente para serem detectados e isolados com facilidade.

Operando a rede com o conjunto de teste, observa-se que é possível estabelecer um limiar de energia ou amplitude que identifique as regiões de sonoridade cumprindo o requisito (1) descrito acima. Quanto ao cumprimento requisito (2), observando trechos do sinal estimado na Figura 4.3-5 (a) nota-se que a estimação dos GCIs é bem sucedida enquanto que na Figura 4.3-5 (b) surgem máximos que não correspondem aos GCIs. Apesar da amplitude destes máximos ser pequena quando comparada aos CGIs verdadeiros, ocorrem casos, observados nos conjuntos de treino e validação, em que estes falsos máximos aparecem com amplitudes elevadas se confundindo aos máximos dos GCIs.

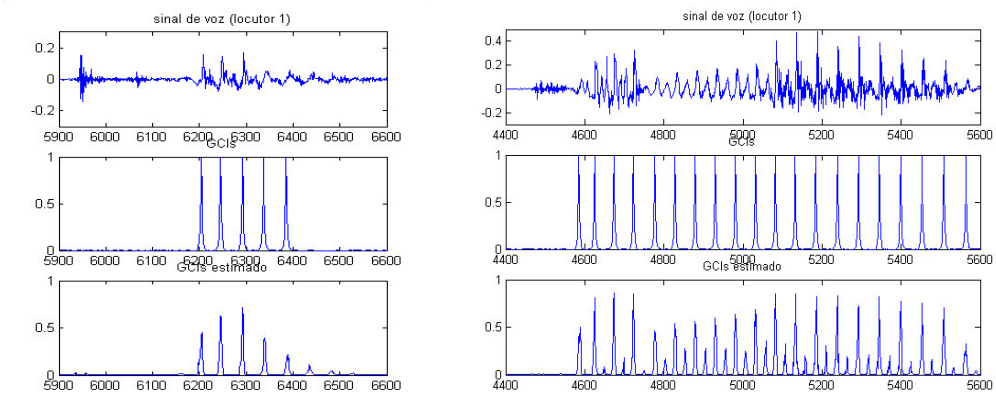


Figura 4.3-5: Sinais de voz, dos GCIs originais e estimados pela rede. (a) caso bem sucedido; (b) surgimento de falsos indicativos dos GCIs.

O conjunto de sinais submetidos à rede para treino e operação, até o instante, foram obtidos numa mesma sessão de gravação, com o mesmo falante (locutor 1), com o mesmo microfone e mesmas condições de ruído. Para observar a capacidade da rede em generalizar a estimação para qualquer falante, a mesma frase gravada pelo locutor 1 foi gravada por um segundo falante (locutor 2) e submetida à rede, na qual se observou que em poucos casos é possível a detecção precisa dos máximos, correspondentes aos CGIs.

De maneira a tornar os sinais de ambos os falantes mais semelhantes, o sinal de voz dos conjuntos de treino foi substituído pelo sinal obtido por filtragem inversa. A filtragem inversa foi obtida passando o sinal de voz pelo filtro inverso ao filtro que modela o trato vocal[5], onde os coeficientes do filtro foram obtidos pelo método da autocorrelação.

A rede foi treinada com o novo conjunto de dados e quando em operação se observou que no conjunto de teste a utilização da filtragem inversa implicou no aparecimento de máximos espúrios nos trechos surdos, não cumprindo o requisito (1). No entanto, dentro dos trechos sonoros, conforme a Figura 4.3-6, os falsos GCIs ocorreram com amplitude menor e em menor número do que na utilização do sinal de voz diretamente, cumprindo o requisito (2). A filtragem inversa reduz a indicação de falsos GCIs, porém a relação de

[5] Na filtragem inversa aplica-se um integrador após a passagem do sinal pelo inverso do filtro AR, para compensar o efeito de radiação dos lábios, obtendo um sinal característico da fonte de excitação glotal.

amplitude entre os trechos sonoros e surdos torna-se menor, fazendo com que a rede se torne mais sensível aos trechos surdos.

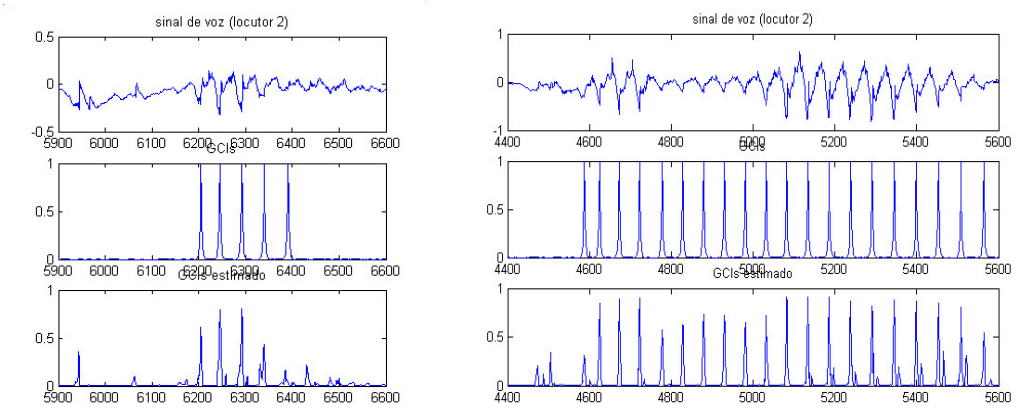


Figura 4.3-6: Sinais de filtragem inversa do sinal de voz, dos GCIs originais e estimados pela rede. (a) e (b) máximos espúrios nos trechos surdos.

O sinal de voz gravado pelo locutor 2, passou pela filtragem inversa e foi submetido à rede, onde se observou que a separação das regiões surdas é igualmente mal sucedida, porém nos trechos sonoros os GCIs foram estimados com sucesso, sendo possível detectá-los com facilidade.

Com este experimento, conclui-se que a rede é capaz de estimar os GCIs a partir de um número reduzido de amostras passadas e futuras e utilizando a filtragem inversa consegue-se boa independência do locutor, aumentando a capacidade de generalização da rede. No entanto, torna-se necessário uma etapa prévia de classificação de sonoridade para que a rede opere somente nestas regiões.

Os experimentos acima apontam para um método promissor de estimação dos GCIs, no entanto para que o método seja efetivamente validado, é preciso um conjunto de dados maior, com maior variabilidade acústica e obtida de diferentes falantes, inclusive do sexo feminino, o que não foi considerado nestes experimentos. O conjunto de dados não requer necessariamente que os GCIs sejam obtidos do EGG, poderiam ser obtidos por inspeção visual, porém o consumo de tempo seria considerável.

Também se sugere que outras arquiteturas para a rede sejam experimentadas, por exemplo, as redes recursivas que utilizam amostras passadas de saída como entrada.

4.4. Método proposto utilizando o microfone de contato

A obtenção de um sinal adicional que monitore o ciclo glotal de forma não invasiva, representa uma solução alternativa ao problema de identificação correta e precisa dos GCIs. O sinal EGG é um exemplo, porém a obtenção deste sinal é feita por um instrumento de custo elevado e usado tipicamente em laboratório.

Alguns sistemas de aquisição de voz em ambientes extremamente ruidosos têm usado microfones em contato com o pescoço, chamado de *throat microphone*, por apresentar reduzida captação de ruído ambiente. Alguns autores (GRACIARENA, 2003) têm proposto a utilização deste tipo de microfone para melhorar o desempenho de sistemas de reconhecimento em ambientes ruidosos. Em (ASKENFELT, 1980) os autores utilizaram um acelerômetro em contato com a pele, na altura das cordas vocais, e observaram que o sinal captado, quando comparado com o sinal de um EGG, representa o som gerado pela vibração das cordas vocais.

Neste trabalho propõe-se o uso de um microfone de contato para a captação da vibração das cordas vocais com o objetivo de auxiliar na detecção dos GCIs. O "microfone" utilizado trata-se de um disco piezoelétrico cerâmico, mostrado na Figura 4.4-1 , geralmente utilizado como captador em instrumentos musicais acústicos, como violão, violino, etc.

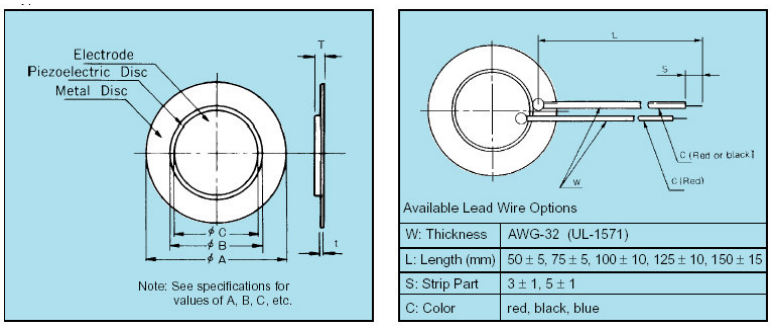


Figura 4.4-1: Formato de um disco piezoelétrico cerâmico, extraído do catálogo da Kyocera.

O disco piezoelétrico é colado a uma base plástica somente pelas bordas, de maneira que o centro fique livre. A base plástica é então fixada a uma fita de velcro, conforme mostrado na Figura 4.4-2 (a). A fita de velcro é colocada ao redor do pescoço de maneira que o disco piezoelétrico fique localizado na região frontal do pescoço, o mais baixo possível, conforme mostrado na Figura 4.4-2 (b). O ajuste de pressão do colar (apertado ou frouxo) está diretamente associado à qualidade do sinal, portanto a pressão ideal é aquela em que o colar fique o mais justo possível, sem causar grande desconforto.

Figura 4.4-2: (a) Disco piezoelétrico fixado ao colar de velcro; (b) colocação do colar na base do pescoço.

Além do sinal do microfone de contato, o sinal de voz precisa ser captado em simultâneo. No entanto, a maioria das placas de som não possui entrada para dois microfones em simultâneo (estéreo). Uma solução é utilizar a entrada *line-in* da placa de som, que pode ser utilizada em modo estéreo, porém é necessário um pré-amplificador para os microfones. Deste modo, foram montados dois pré-amplificadores, conforme a nota referenciada em (CITTADINNI, 2002).

Na Figura 4.4-3, é mostrado o esquemático dos pré-amplificadores[6]. No caso do microfone de contato, a polarização necessária para o microfone de eletreto, realizada por R1, R2 e C1, é dispensável. O valor de R7, que determina o ganho do amplificador juntamente com R8, foi modificado para 2,2 KΩ, de maneira que o ganho dos pré-amplificadores não saturasse a entrada *line-in* da placa de som.

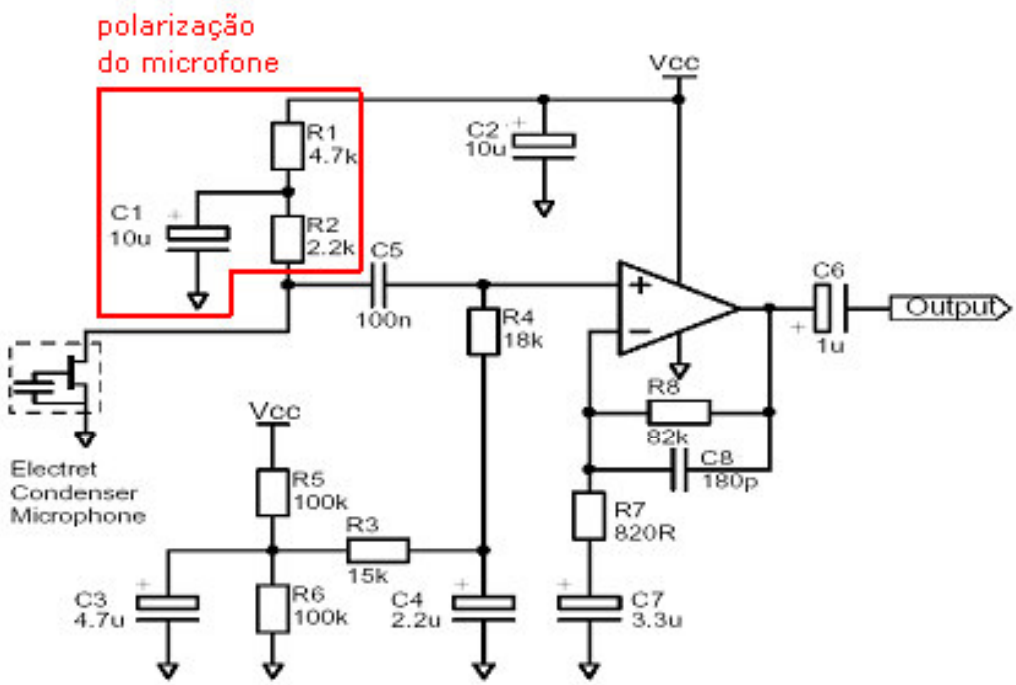


Figura 4.4-3: Circuito do pré-amplificador, adaptado de (CITTADINNI, 2002).

O circuito foi implementado de acordo com as características do microfone de eletreto. No entanto, no caso do captador piezoelétrico, outros circuitos de instrumentação podem ser mais adequados devido à característica particular deste tipo de sensor. A alimentação do circuito é feita por uma fonte de 9 V externa, não regulada. Então no circuito realizado incluiu-se um circuito integrado regulador de voltagem 7806. Na Figura 4.4-4 é mostrado a montagem do circuito em uma placa padrão de circuito impresso.

[6] Por questões práticas, neste circuito foi usado um amplificador operacional TL072 ao invés do TS971 indicado na *Application Notes*.

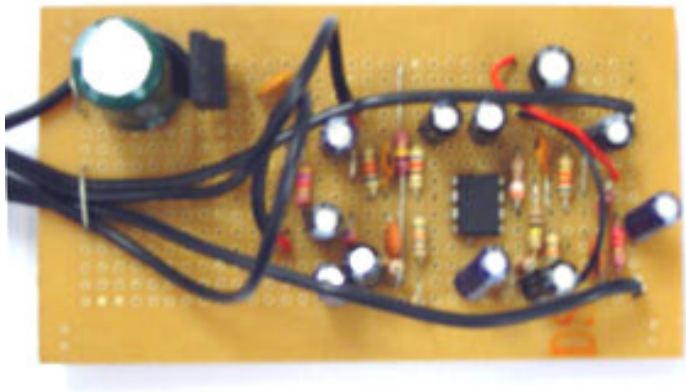


Figura 4.4-4: Circuito pré-amplificador e regulador de voltagem.

O nível de ruído do pré-amplificador, observado no sinal do microfone de eletreto, é similar ao ruído gerado pela entrada da placa de som específica para o microfone.

Convém citar que os sinais obtidos em simultâneo pelos dois microfones estarão defasados de acordo com a distância entre os microfones. Deste modo, para manter esta distância fixa ao longo da gravação, foi usado um microfone acoplado aos fones de ouvido e próximo à boca. Assim, o atraso dependerá principalmente das características físicas do usuário. Em média este atraso equivale ao tempo de propagação para o som percorrer uma distância típica de (20 ± 5) cm. Deste modo, temos um atraso no sinal do microfone convencional da ordem de (0.6 ± 0.15) ms. Os sinais obtidos foram amostrados a freqüência de 22050 Hz o que resulta em um atraso em torno de 12 amostras.

Na Figura 4.4-5 são mostrados nas figuras superiores os sinais obtidos pelo microfone convencional, nos quais foi compensado um atraso de 12 amostras, e abaixo os sinais do microfone de contato para sons com diferentes características.

Verificando os sinais obtidos pelo microfone de contato observa-se que durante os trechos sonoros, ou seja onde ocorre atividade glotal, o sinal demonstra características mais estacionárias do que o sinal de voz, que somente pode ser considerado estacionário

em curtos intervalos de tempo (20 ms aprox.). Esta característica se justifica pelo fato de que as vibrações captadas pelo contato provêm principalmente da laringe (considerada um tubo com dimensões constantes) que produz harmônicos quase invariantes ao longo do tempo. Para o sinal de voz, as diferentes configurações do trato vocal para a produção de diferentes sons, produzem diferentes harmônicos ao longo do tempo.

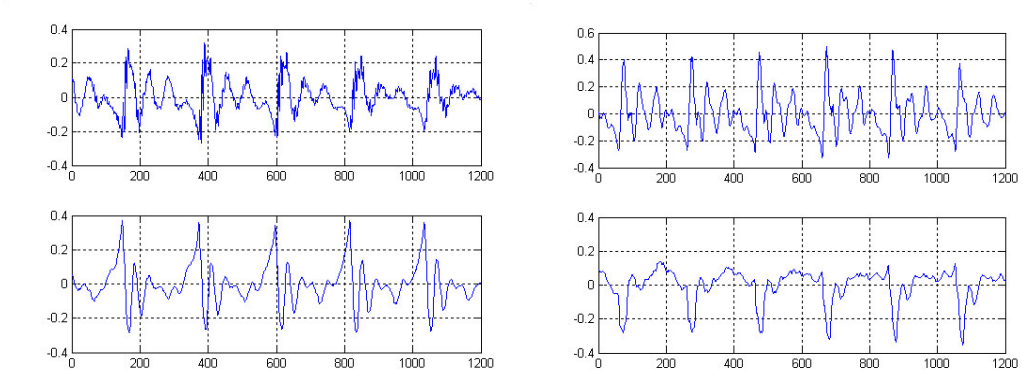


Figura 4.4-5: Sinal do microfone convencional e do microfone de contato para diferentes sons. (a) GCIs detectáveis como máximos locais; (b) caso crítico de detecção.

No primeiro sinal mostrado na Figura 4.4-5, observa-se que a ocorrência de picos no sinal do microfone de contato são bons indicativos para os GCIs e são facilmente detectáveis como máximos locais. Porém, no segundo sinal, os picos têm sua amplitude reduzida não sendo mais possível detectá-los como máximos locais. Isto implica no fato de que a obtenção dos GCIs diretamente do sinal do microfone de contato não é possível, sendo necessário um procedimento de detecção de épocas.

Os métodos apresentados nas seções 4.2.1. e 4.2.2. foram aplicados ao sinal obtido pelo microfone de contado sendo descritos a seguir nos experimentos 4.5 e 4.6, respectivamente.

Experimento 4.5.: obtenção dos GCIs pelo resíduo de predição linear .

Inicialmente é aplicado ao sinal do microfone de contato um filtro de pré-ênfase enfatizando as altas freqüências para tornar seu decaimento espectral similar ao sinal de

voz. Em seguida a aplicação do método de detecção pode ser observada na Figura 4.4-6.

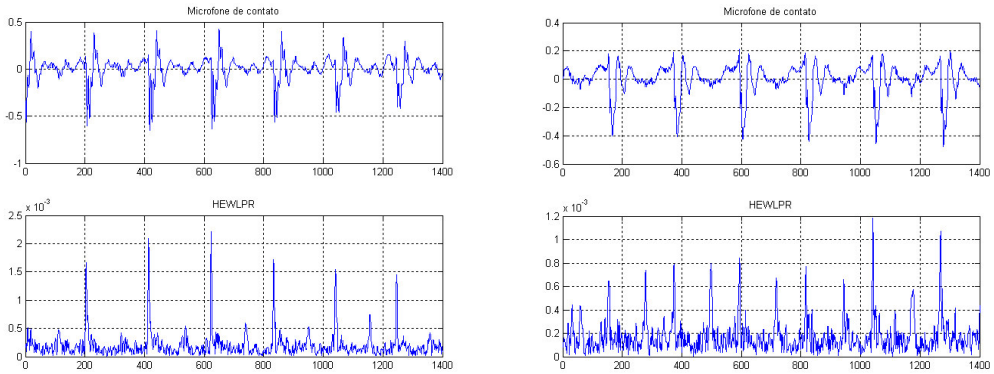


Figura 4.4-6: Detecção dos GCIs a partir do resíduo de predição linear.

Observa-se que os máximos, correspondentes aos pontos de excitação do trato vocal, ocorrem com maior amplitude do que no sinal de voz, sendo mais fácil detectá-los em meio ao ruído, como pode ser visto na primeira figura. Porém, os picos intermediários aos GCIs também ocorrem com maior amplitude, como pode ser observado na segunda figura. Conclui-se portanto que o sinal do microfone de contato fornece informações mais intensas não somente do instante de fechamento da glote, mas também de outros instantes de movimento das cordas vogais, por exemplo, o instante de abertura da glote.

Por um lado, o método confirma a suposição de que um microfone de contato forneceria informações mais nítidas sobre o movimento da glote, por outro lado, a dificuldade em separar os GCIs torna o método ineficiente para esta proposta.

Experimento 4.6.: obtenção dos GCIs por máxima verossimilhança.

Aplicando o método ao mesmo sinal do experimento 4.5, obtido do microfone de contado, o sinal é passado igualmente por um filtro de pré-ênfase e em seguida o método é aplicado, onde o resultado da detecção é mostrado na Figura 4.4-7.

O método quando aplicado ao sinal de voz apresentava uma imprecisão na detecção dos GCIs, no qual era necessário um método empírico de correção. Surpreendentemente, o

método aplicado ao sinal do microfone de contato, apresenta esta imprecisão em alguns poucos casos e da ordem de apenas 0.3 ms. Além disso, foi constatado que o módulo da transformada de Hilbert do sinal MLED é mais preciso do que próprio sinal MLED ou da multiplicação dos dois sinais conforme proposto em (CHENG, 1989). Além disso, a subtração da média global do módulo da transformada de Hilbert permite a diferenciação entre os trechos sonoros e surdos.

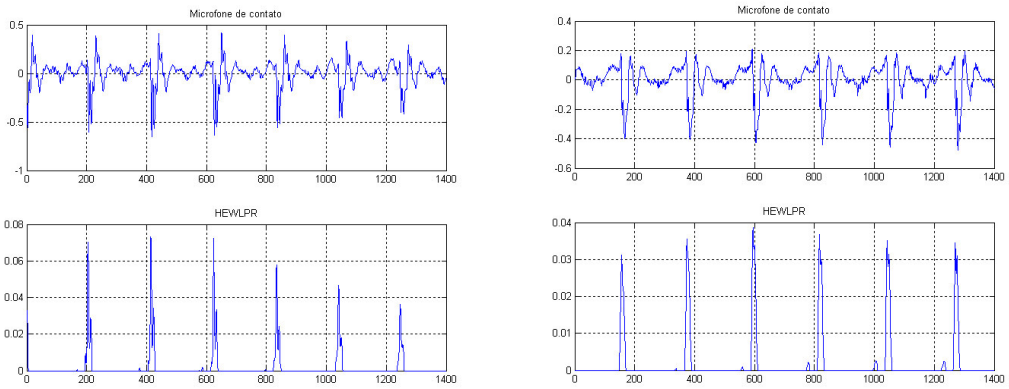


Figura 4.4-7: Detecção dos GCIs por máxima verossimilhança.

Finalmente, para a determinação pontual dos GCIs a partir do sinal do módulo da transformada de Hilbert do sinal MLED, foi implementado um algoritmo de busca por máximos locais, que consiste na aplicação de janelas sobrepostas de curta duração. A sobreposição é definida maior do que 50%, de maneira que um indicativo de um GCI ocorra como um máximo pelo menos duas vezes em janelas adjacentes. O tamanho da janela e a sobreposição são escolhidos de acordo com o pitch médio do sinal, podendo ser determinado a priori ou reajustado de acordo com o pitch característico do locutor.

Para verificação do método, foram observados casos onde a detecção dos GCIs a partir do sinal de voz é extremamente difícil, principalmente para consoantes vozeadas. Nas Figura 4.4-8, 4.4-9, 4.4-10 e 4.4-11 são mostrados trechos de sinais de voz e os GCIs obtidos a partir do sinal de voz e a partir do sinal do microfone de contato. Para notação dos segmentos e transcrição fonética foram utilizados os símbolos IPA e a nomenclatura utilizada para as consoantes segue a definida em (SILVA, 2003).

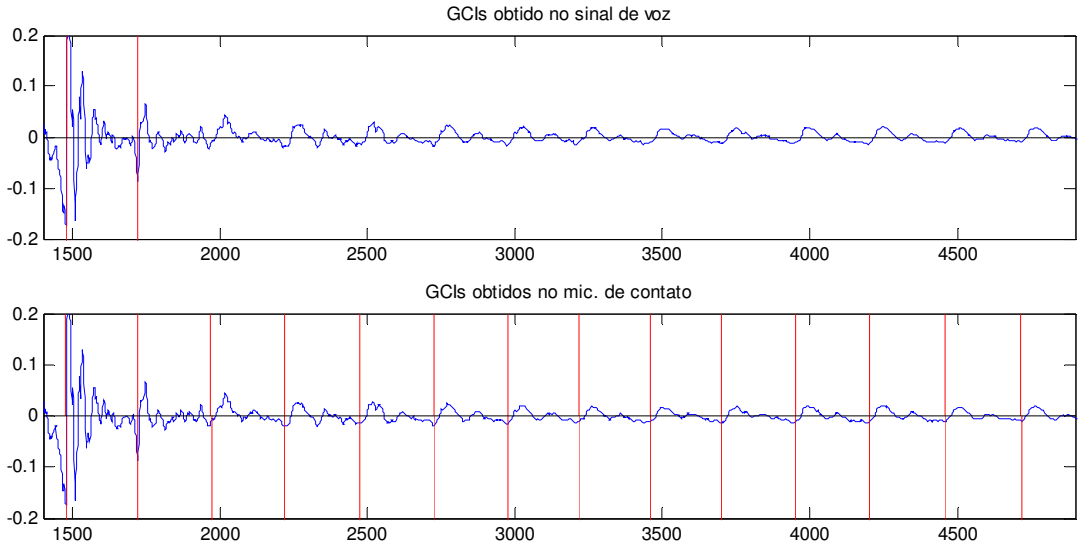


Figura 4.4-8: Segmento do sinal de voz contendo a consoante Oclusiva Bilabial Vozeada /b/, recortada da palavra "abril" - /aˈbriu/, e as marcas dos GCIs obtidos do sinal de voz e do microfone de contato.

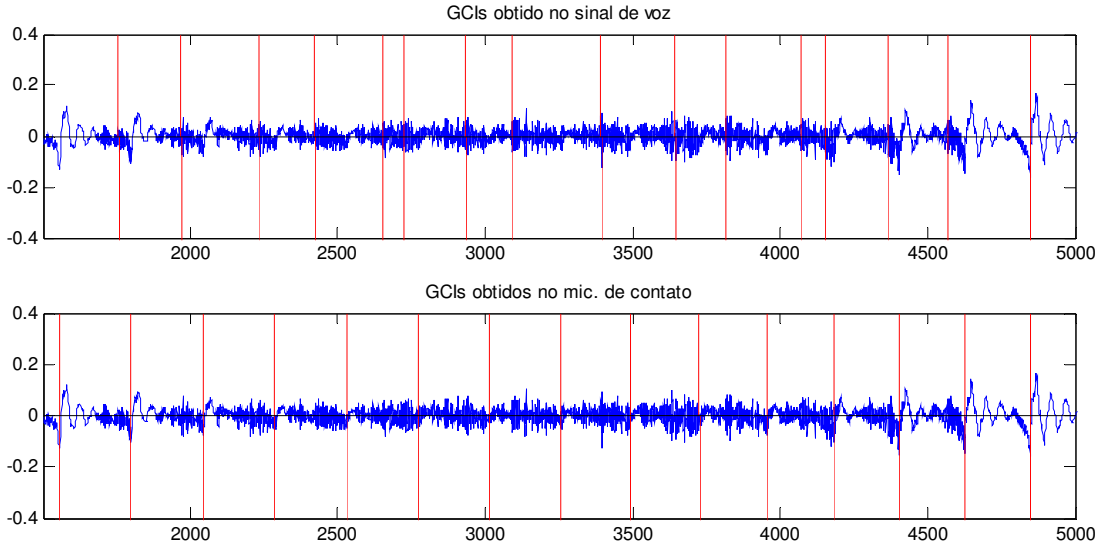


Figura 4.4-9: Segmento do sinal de voz contendo a consoante Fricativa Alveolar Vozeada /z/, recortada da palavra "casa" - /ˈkaza/, e as marcas dos GCIs obtidos do sinal de voz e do microfone de contato.

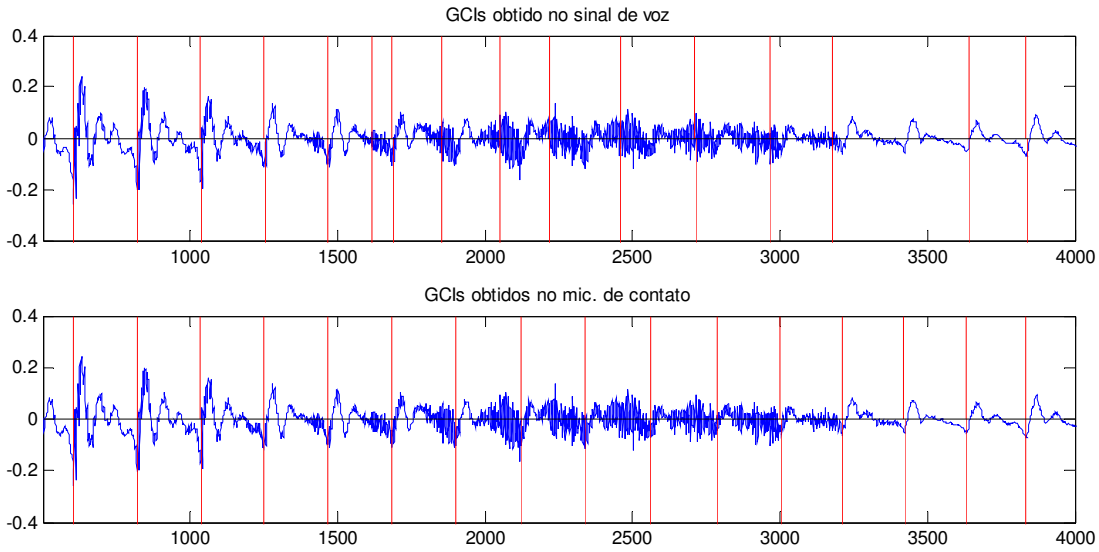


Figura 4.4-10: Segmento do sinal de voz contendo a consoante Fricativa Alveopalatal Vozeada /ʒ/, recortada da palavra "mesmo" - /ˈmeʒmʊ/ e as marcas dos GCIs obtidos do sinal de voz e do microfone de contato.

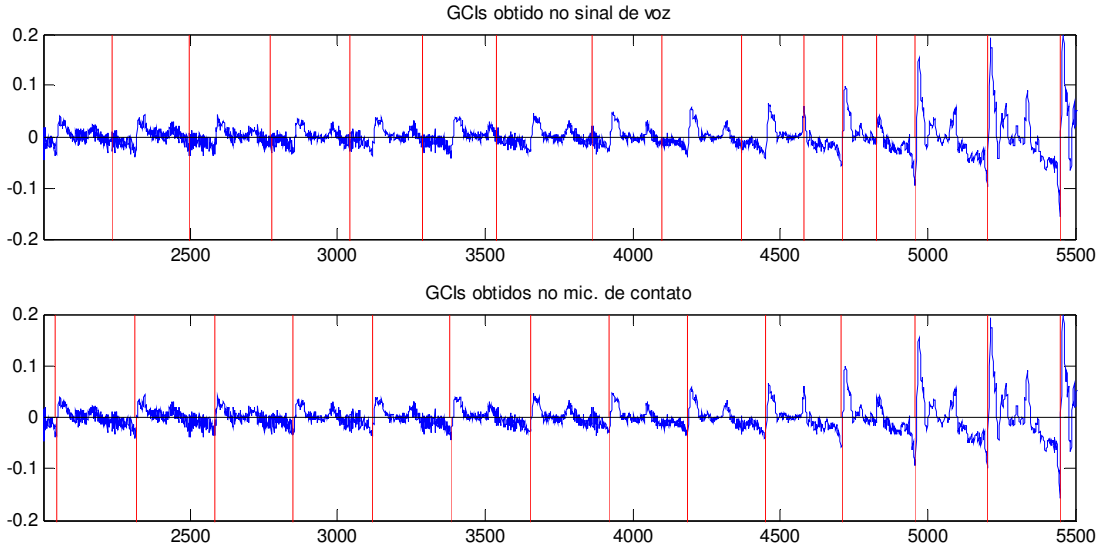


Figura 4.4-11: Segmento do sinal de voz contendo a consoante Fricativa Alveolar Vozeada /v/, recortada da palavra "avante" - /aˈvãtʃi/, e as marcas dos GCIs obtidos do sinal de voz e do microfone de contato.

Apesar dos resultados promissores, verificados nos sinais observados, o algoritmo não está livre de erros de imprecisão, de não inserção ou de inserção indevida de uma marca, sendo prudente a verificação manual das marcas pelo usuário. Neste sentido, a

utilização do microfone de contato se justifica também pela facilidade de se encontrar visualmente os GCIs no sinal.

4.5. Conclusão

Foi estabelecido um método semi-automático para a obtenção das marcas de pitch em segmentos de sinais de voz, a serem utilizadas pelo algoritmo TD-PSOLA na concatenação destes segmentos. As marcas coincidem com os instantes de fechamento glotal, reduzindo a possibilidade de erros de fase na concatenação.

O método utiliza a gravação simultânea do sinal de voz e o sinal obtido por um disco piezoelétrico em contato com a pele, localizado na base do pescoço, no qual se mostrou conter mais informações sobre a atividade glotal do que o próprio sinal de voz. Observa-se ainda que o sinal obtido pelo microfone de contato é menos sujeito a ruído ambiente; que em trechos sonoros o sinal apresenta características mais estacionárias do que o sinal de voz; e que as consoantes sonoras são enfatizadas.

Foi proposto um método para detecção automática do GCIs. Os resultados se mostraram promissores quando aplicado a casos críticos onde a obtenção precisa dos GCIs a partir do sinal de voz apresenta grande dificuldade. Apesar dos resultados, o algoritmo não está livre de erros de imprecisão, de não inserção ou de inserção indevida de uma marca, sendo prudente a verificação manual das marcas pelo usuário. Isto faz com que a metodologia proposta seja denominada de semi-automática. Neste sentido, a utilização do microfone de contato se justifica não só pelo aumento de eficiência no processo de detecção automática, reduzindo a ocorrência de erros, mas também pela facilidade de se encontrar visualmente os GCIs.

O disco piezelétrico usado como microfone de contato assim como a montagem do circuito pré-amplificador, tiveram caráter experimental. Sugere-se que futuramente seja dado mais enfoque a fatores de instrumentação propondo diferentes dispositivos e/ou circuitos, considerando sempre a proposta inicial de baixo custo e enfocada ao usuário comum.

Capítulo 5

Construção do banco de unidades

5.1. Introdução

Inicialmente, os algoritmos de detecção do pitch e de segmentação do sinal, foram implementados em Matlab. Porém, foi observado que muito tempo era gasto na manipulação de diferentes aplicativos, na escrita de arquivos, etc. Deste modo, buscando diminuir esse tempo, foi desenvolvido um aplicativo, na linguagem C++, para o sistema operacional Windows, chamado de *Editor*. Este aplicativo permite que todo o processo de construção do banco seja feito em um único ambiente de trabalho.

Neste capítulo será descrita a criação de um pequeno banco de unidades, onde serão utilizados os conceitos e técnicas abordadas nos capítulos anteriores. O uso do

aplicativo *Editor* será apresentado como uma metodologia geral para a construção futura de uma banco de unidades. Para exemplificar este processo, serão feitos o recorte e seleção de um conjunto simples de unidades.

No Capítulo 3, foi obtido um método de segmentação acústica do sinal de fala, no qual as fronteira detectadas entre segmentos nem sempre coincidiam com as fronteiras fonéticas. Neste capítulo será apresentado um método de segmentação, que utiliza a informação de sonoridade e a transcrição fonética dos segmentos para melhorar a eficiência da segmentação e fornecer as etiquetas dos segmentos. Este método foi implementado no *Editor*.

No Capítulo 4 foi apresentado um método de inserção de marcas de pitch, baseado na detecção dos GCIs, e conseqüentemente de determinação de sonoridade. Para essa detecção foi usado um sinal auxiliar, obtido de um microfone de contato, que é adquirido simultaneamente com o sinal de voz. No *Editor* foram implementadas as rotinas para aquisição e reprodução dos sinais, através da placa de som, utilizando a entrada *line-in* da placa para aquisição em dois canais. Também foram implementadas as rotinas para edição destes sons, como recorte e gravação em disco, assim como um método de determinação automática do atraso entre os sinais e a técnica proposta para detecção dos GCIs.

Considerando que a construção do banco é um processo recursivo, foi implementado no *Editor* um método de busca e concatenação de unidades, e o método TD-PSOLA para manipulação das durações e do contorno de pitch. Para validar o recorte das unidades, o sinal concatenado pode ser observado tanto no domínio temporal quanto no domínio espectral, através de um espectrograma.

Para avaliar a marcação das fronteiras entre segmentos, inseridas nas unidades, que serão utilizadas futuramente para a aplicação de regras prosódicas, foi implementado no *Editor* um método de *transplante de prosódia*, isto é, a partir de uma palavra ou frase previamente gravada e etiquetada foneticamente, as durações dos segmentos e o contorno de pitch são extraídos e aplicados a um sinal concatenado com a mesma seqüência fonética. Dado que os segmentos do sinal gravado e do sinal concatenado

possuem durações e contorno de pitch distintos, é possível determinar quais serão as durações e o contorno de pitch a serem impostas ao sinal concatenado, para que ele possua as mesmas características prosódicas do sinal natural, através de um alinhamento temporal entre as marcas fonéticas dos dois sinais.

Ao longo deste Capítulo será descrita uma metodologia geral para a construção do banco de unidades, auxiliada pelo aplicativo *Editor*. Na seção 5.2. serão descritos o processo de gravação dos logotomas e obtenção das marcas de pitch coincidentes com os GCIs, implementado pelo *Editor*. O método proposto de segmentação semi-automática, implementado no *Editor*, utilizando as fronteiras de vozeamento como segmentação inicial e a transcrição fonética para etiquetagem, será apresentado na seção 5.3, onde serão mostrados alguns casos de segmentação. Na seção 5.4 discutimos o processo de seleção e recorte das unidades que formarão o banco. Na seção 5.5., procuramos validar toda a metodologia aqui desenvolvida através da montagem de um pequeno inventário de unidades.

5.2. Gravação dos logotomas e detecção das marcas de pitch

Foi visto no Capítulo 2 que o uso de logotomas para a extração das unidades é uma proposta adequada, principalmente quando se pretende uma segmentação automática ou semi-automática da unidade. Além disso, foi visto que a unidade precisa estar inserida em um ambiente fonético e prosódico neutro. Deste modo, foi escolhido inserir os logotomas em frases sem sentido.

Os caracteres utilizados no *Editor* para identificar as unidades de concatenação estão no formato SAMPA (S*peech Assessment Methods Phonetic Alphabet*), considerados mais adequados para o ambiente computacional do que os caracteres IPA, pois utilizam os caracteres do código ASCII. No ANEXO C são descritos os caracteres utilizados, a correspondência com os caracteres IPA e um código identificador usado internamente no *Editor*. Ao longo do texto, na representação de uma unidade de concatenação, ou de

uma seqüência, os caracteres aparecem entre chaves "{}" seguindo a mesma notação que KAFKA, (2002).

Para a formação dos logotomas, foram utilizadas até 3 sílabas, formadas por [p], [t], [b] e [a]. Por exemplo, a unidade {zE} foi inserida no logotoma "pazé" {pazE}. Quando possível as vogais foram colocadas em posição tônica a não ser em poucos casos em que será preciso obtê-las em contexto átono. Para facilitar a segmentação, buscou-se sempre formar logotomas associando um segmento surdo com um sonoro. Por isso em alguns casos foi usado [p] ou [t] em outros [b].

O logotoma foi inserido em frases sem sentido seguindo o seguinte critério:

- unidades no interior do logotoma: ***Falo*** *logotoma* ***pausado.***.
- unidades precedidas por silêncio: *_logotoma* ***pausado***.
- unidades sucedidas por silêncio: ***Falo*** *logotoma_*.

Observa-se que quando a frase sem sentido é falada naturalmente, o final do logotoma se une à palavra "pausado", está é a razão de ter sido escolhida uma palavra paroxítona que começasse em p. A escolha pela palavra "falo" se deve ao fato de ocorrem poucos erros na detecção de sonoridade.

As frases a serem gravadas são escritas em um arquivo com as transcrições fonéticas em código SAMPA, precedidas pelo nome da unidade que será recortada. Em seguida, este arquivo é aberto pelo *Editor* que irá automaticamente criar um diretório, com o mesmo nome do arquivo, onde os logotomas serão gravados.

Na Figura 5.2-1 é mostrada a tela principal do editor e uma lista de logotomas contendo 3 exemplos. Nota-se que inicialmente a leitura a partir da transcrição fonética é difícil, por isso é conveniente que as frases sejam escritas em outro arquivo no formato normal de leitura.

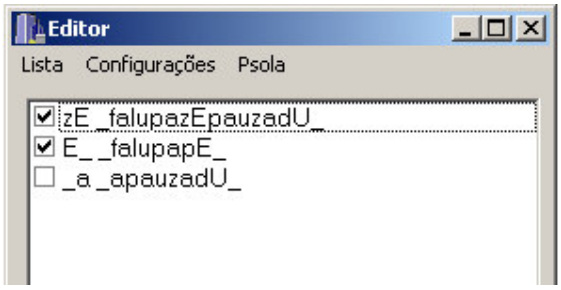


Figura 5.2-1: Parte da tela principal do aplicativo *Editor* com uma lista de 3 logotomas.

Antes de iniciar a gravação é preciso configurar o nível de ruído e determinar o alinhamento entre os sinais do microfone de contato e o sinal de voz. O nível de ruído é determinado segmentando 0,5 s de silêncio com janelas de 20 ms, de onde é obtida a energia em cada janela e em seguida calculada a energia média. Este nível de ruído será utilizado na fase de gravação, como limiar de energia, para eliminar grandes trechos de silêncio inicial e final, que podem estar contidos nas gravações.

O atraso contido no sinal de fala em relação ao microfone de contato é indicado pela máxima correlação entre os resíduos de predição linear dos dois sinais. No entanto, para eliminar problemas de fase e diminuir o ruído contido nos resíduos, foi aplicado o método descrito no Capítulo 4, na seção 4.2.1 para reduzir as ambigüidades de ambos os resíduos. Para a correlação, basta uma única janela de análise num ponto estável de uma vogal. Na tela de configuração do *Editor* o usuário grava o som de um [a] e em seguida ele seleciona somente um trecho estável da vogal onde o método de ajuste será aplicado. O atraso geralmente está em torno de 12 amostras. A partir daí, o sinal de fala e o sinal do microfone de contato serão defasados automaticamente na gravação dos logotomas.

Retornando à gravação dos logotomas, ao clicar sobre um logotoma da lista, uma segunda janela estará disponível para a gravação. O usuário inicia a gravação, fala o logotoma, termina a gravação e os sinais serão traçados em duas figuras. Se o usuário desejar gravar todos os logotomas antes de prosseguir com o processo de recorte da unidade, existe a opção na tela principal do *Editor*, chamada de *gravação seqüencial.* Quando selecionada, logo que o usuário seleciona outro logotoma da lista o logotoma atual é salvo em disco. Quando um logotoma é salvo, ele aparece marcado na lista.

Na Figura 5.2-2 é mostrada a tela de gravação e inserção das marcas de pitch do *Editor*, onde na figura superior é mostrado o sinal de voz e na figura inferior o sinal do microfone de contato. Todos os sinais adquiridos são amostrados à freqüência de 22050 Hz. É possível selecionar um trecho do sinal com o *mouse* e visualizar a seleção aumentada na escala temporal, que será útil para verificar a precisão das marcas de pitch. O *Editor* também oferece a possibilidade de reproduzir o sinal ou um trecho do sinal quando selecionado.

Depois de gravado o logotoma, a seguir as marcas de pitch serão detectadas, conforme o método descrito no Capítulo 4, na seção 4.4. O usuário poderá escolher o tamanho da janela e a sobreposição, usadas para a busca dos máximos locais, para melhor se adequar ao seu pitch médio. Na Figura 5.2-2 são mostradas as marcas de pitch obtidas após a busca pelos máximos locais.

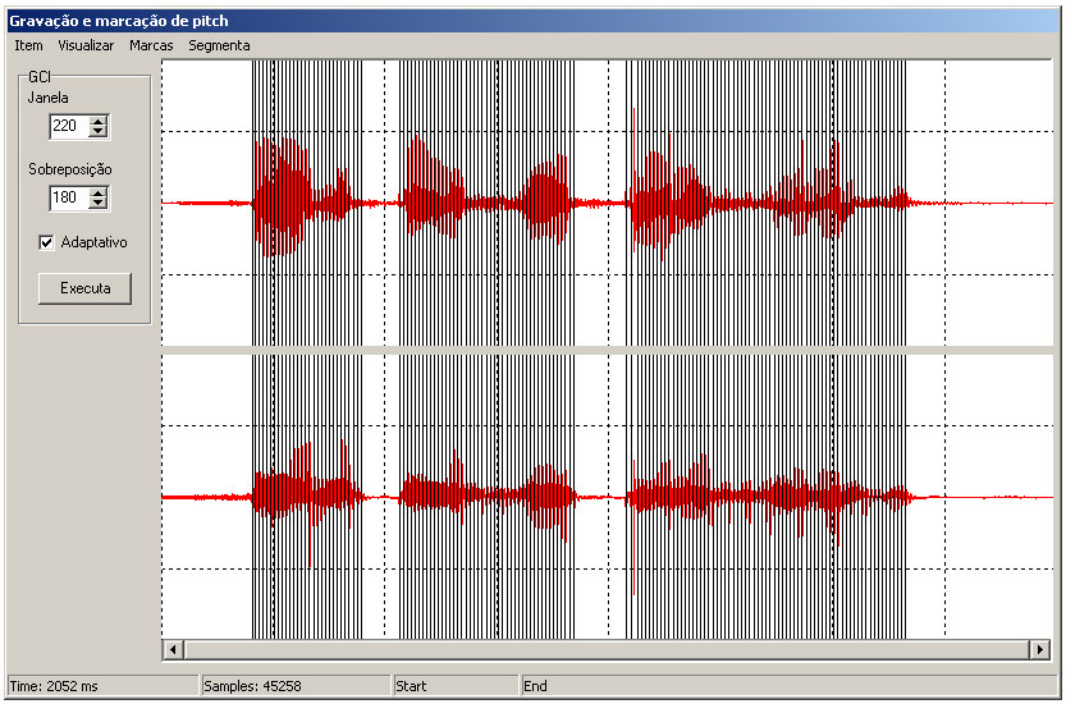


Figura 5.2-2: Tela de gravação e inserção das marcas de pitch do *Editor*. Na Imagem superior é mostrado o sinal de voz e na inferior o sinal do microfone de contato. Os traços verticais contínuos indicam as marcas obtidas automaticamente.

Foi verificado que pequenos erros de inserção de pitch geralmente ocorrem no inicio das oclusivas não vozeadas, nas vogais finais, e em alguns poucos casos, nos sons

mistos. O usuário pode editar estas marcas com o uso do *mouse*. Porém, para que o tempo não seja desperdiçado, atenta-se para o fato de que a precisão nas marcas só será necessária dentro da unidade. A razão por determinar as marcas de pitch antes da seleção da unidade deve-se ao fato de que, sendo o método robusto na detecção dos trechos sonoros e surdos, a segmentação semi-automática fará uso também desta informação para encontrar a unidade dentro do logotoma. Foi utilizado o critério de que trechos maiores do que 20 ms entre uma marca e outra são considerados surdos. Assim, nesta etapa é preciso estar atento somente aos erros grosseiros de detecção que afetariam a classificação da sonoridade.

5.3. Segmentação e etiquetagem semi-automática

No Capítulo 5, foi visto que o método proposto de segmentação acústica insere marcas de fronteira entre segmentos que nem sempre coincidem com as fronteiras fonéticas. Devido à natureza do método, acredita-se que se o sinal for segmentado primeiramente em trechos sonoros e surdos, e só em seguida aplicar o processo de segmentação acústica para cada trecho, espera-se por melhores resultados. O fato é que a determinação de sonoridade já estabelece uma fronteira entre os segmentos sonoros e surdos.

Outra questão a considerar é a etiquetagem dos segmentos. Se a segmentação fornecesse uma coincidência perfeita com as fronteiras fonéticas, tendo disponível a transcrição fonética do logotoma a etiquetagem seria imediata. No entanto, tendo em mãos um método de marcação robusta das fronteiras de sonoridade, a transcrição fonética do logotoma, e o conhecimento a priori de quais os segmentos da transcrição são sonoros ou surdos, é possível etiquetar os segmentos nestas fronteiras.

Na Figura 5.3-1 é mostrada a tela de segmentação, etiquetagem e recorte do *Editor*. Na figura superior é mostrado o sinal de fala correspondente ao logotoma “falo pazé pausado”. { _falupazEpauzadU_}. Na figura inferior é mostrado o espectrograma do sinal. Para a obtenção do espectrograma, o sinal foi segmentação em janelas de 10 ms,

com sobreposição de 5 ms, porém estes valores podem ser modificados no canto esquerdo da tela. O espectro foi obtido pelo módulo da FFT de 512 pontos. O limite superior é da freqüência máxima de 11025 Hz.

As marcas verticais em azul, indicam as fronteiras de sonoridade obtidas da marcação de pitch nas quais foram etiquetas de acordo com a informação de sonoridade contida na transcrição fonética. Por exemplo, no caso do logotoma mostrado na figura, é sabido a priori que os segmentos grifados são sonoros { _f**alu**p**azE**p**auzadU**_}.

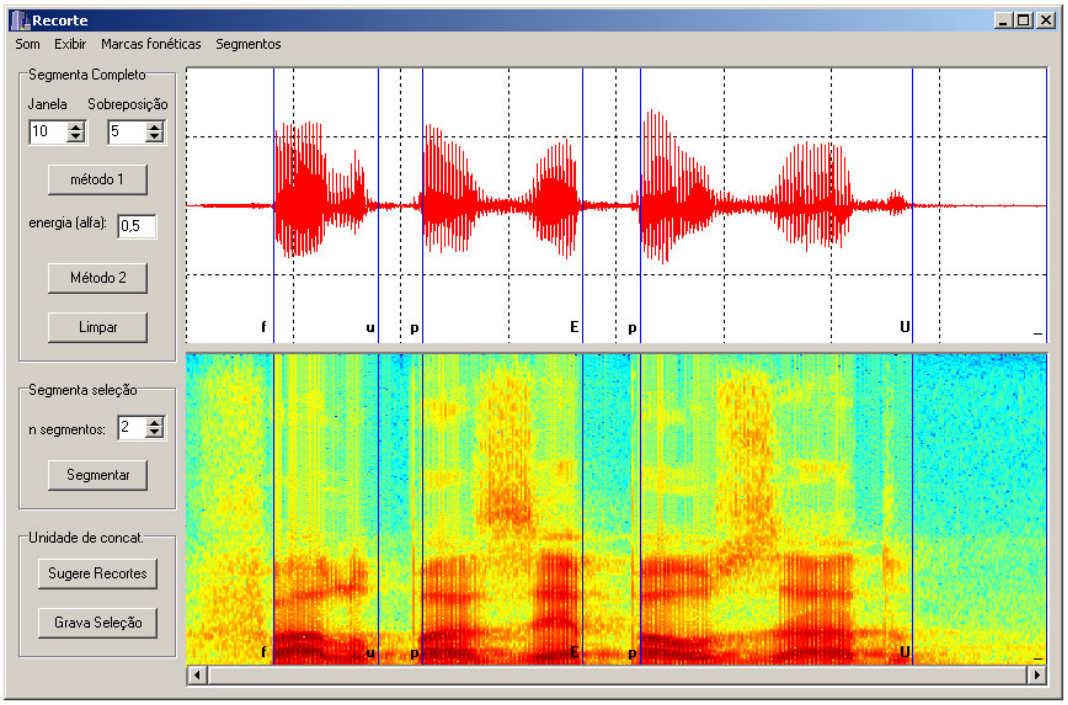


Figura 5.3-1: Tela de segmentação e recorte do aplicativo *Editor*. Na figura superior o sinal de fala e na inferior o espectrograma correspondente. As marcas indicam as fronteiras de sonoridade, etiquetadas de acordo com a informação prévia de sonoridade obtida da transcrição fonética.

As marcas podem ser editadas igualmente como as marcas de pitch. Com a exceção de que para manter a etiquetagem coerente é preciso inserir as marcas da esquerda para a direita.

A partir desta segmentação inicial, a segmentação acústica foi aplicada separadamente em cada trecho. O método implementado no *Editor* para a segmentação acústica utiliza 16 coeficientes cepstrais como parâmetros, extraídos do sinal janelado da mesma forma

que na montagem do espectrograma. A medida de dispersão usada entre os vetores de parâmetros é a distância Euclidiana, conforme definida no Capítulo 3. Foi adicionada à medida de dispersão espectral uma medida de disperção de energia, conforme sugerida por RABINER, (1993) e SVENDSEN, (1987) que consiste no erro quadrático entre os coeficientes cepstrais de ordem "0" (não considerados na dispersão espectral). Esta dispersão de energia foi pesada por um parâmetro *alfa*, ($0 \leq alfa \leq 1$) que pode ser determinada pelo usuário. Apesar de SVENDSEN, (1987) relatar que 0,5 foi um valor adequado para *alfa*, verificou-se que ocorrem casos em que a não consideração da energia ($alfa = 0$) resulta em melhores resultados. No *Editor,* este *alfa* é um parâmetro a ser escolhido pelo usuário.

Conforme observado no Capítulo 3, os segmentos que não possuem uma região estável bem definida, "atraem" as marcas buscando minimizar o acúmulo do erro quadrático. Assim, tendo sido a segmentação dividida em trechos menores, ocorre que nos trechos onde os segmentos são mais estáveis, existe uma maior possibilidade das fronteiras entre os segmentos coincidirem com as fronteiras fonéticas.

Na Figura 5.3-2, é mostrado a segmentação do logotoma {pazE} em detalhe, inserido na frase mostrada na Figura 5.3-1. A segmentação nos outros trechos apresentou resultado nem tão regulares, porém, na segmentação da unidade {zE} em questão, observa-se um posicionamento coerente das fronteiras dos segmentos e a etiquetagem correta.

Foi observado que geralmente nos trechos tônicos as fronteiras entre segmentos coincidem razoavelmente bem com as fronteiras fonéticas. Isso se justifica pelo fato de que a coarticulação na região de tonicidade é reduzida. Um caso crítico são os ditongos, pois quase não apresentam regiões de estabilidade. Neste caso, é preciso corrigir manualmente as marcas inseridas. Nota-se também que os logotomas inseridos nas frases são montados propositadamente para serem pouco articulados e portanto a segmentação acústica para estes casos oferece resultados bastante úteis.

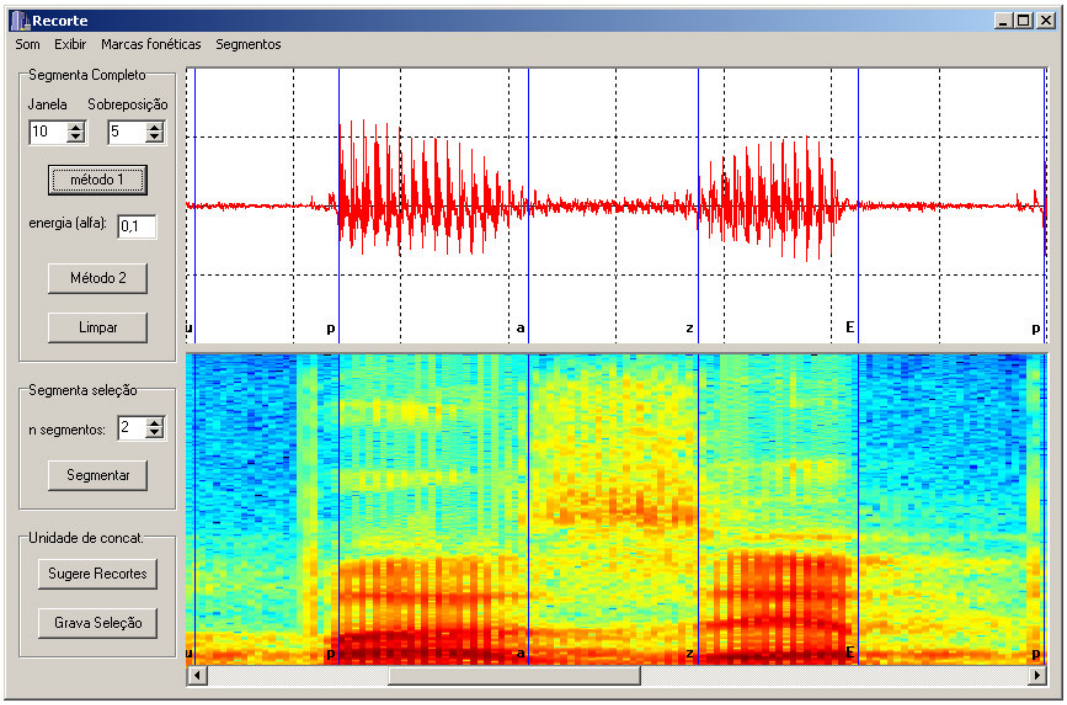


Figura 5.3-2: Segmentação do logotoma {pazE} em detalhe. Observa-se uma coincidência satisfatória das marcas de segmentação com as fronteiras dos segmentos fonéticos.

5.4. Seleção e recorte das unidades

Ao longo desta seção, será descrita a forma de recorte de algumas unidades simples, porém buscando estabelecer uma metologia geral para a futura construção completa do banco utilizando o aplicativo *Editor*. Ao longo desta descrição serão examinados alguns casos particulares que podem reduzir o inventário de unidades.

O recorte da unidade no *Editor* é feito selecionando o segmento na qual se deseja recortar e acionando o comando *gravar seleção*. Cada unidade será salva em um arquivo binário com o nome dado pela concatenação dos códigos dos segmentos fonéticos a que ela corresponde. Estes segmentos são indicados no início do logotoma. Cada unidade conterá o sinal de fala, as marcas de pitch e as marcas de fronteira fonéticas que estiverem dentro do segmento selecionado. Os limites de recorte que ocorrerem em trechos sonoros serão aproximados para a marca de pitch mais próxima, para que não ocorram erros de fase na concatenação.

A seguir serão apresentados os critérios de recorte e o inventário de algumas unidades utilizadas neste trabalho, na qual o método de obtenção destas unidades servirá como uma metologia geral para construção futura de outros bancos, ou a continuação deste. Esta apresentação foi dividida em grupos, com a seguinte notação: **C** indica uma consoante, **V** indica uma vogal e **v** indica uma semivogal e o til "**~**" representa a nasalidade nas vogais.

Unidades _C, _V, _V~, V_, V~_, v_ e C_ : Nestes casos foram usados difones. As unidades **V_** foram inseridas em logotomas com duas sílabas, sendo pronunciadas em sílaba tônica, com exceção do segmento {@} que ocorre predominantemente em final não tônico, assim como as semivogais **v_**. As vogais nas unidades **_V e _V~** não precisaram ser inseridas em logotomas e são pronunciadas como tônicas, por exemplo, na unidade {_o}, tem-se a frase: "ô pausado" {_opauzadU}. Para as unidades **_C** foi utilizada uma única sílaba ligada a vogal [a].

As unidades **C_** são somente duas: {X_} e {S_}. Estas são as únicas consoantes que aparecem na posição final (coda silábica) (SILVA, 2003).

Para este inventário, a primeira observação de redução do inventário é que nas unidades **_C**, as oclusivas surdas, inclusive o segmento {tS}, podem dividir o mesmo difone que no caso será usado o {_t}. Este critério não afeta a inteligibilidade do segmento, dado que este difone nada mais é do que um pequeno silêncio com uma marca de fronteira no fim pois o segmento de oclusão está contido na unidade seguinte. Para o segmento {_dZ} também foi utilizado o mesmo segmento que {_d} pela mesma razão. Posteriormente poderá ser avaliada a possibilidade de utilizar o mesmo princípio para as oclusivas sonoras.

Outro método de redução que será considerado posteriormente, mas que afeta as unidades V_ e V~_ é o de concatenar as V~_ com inícios orais. Isto implica que a maior parte da vogal estará contida nestes segmentos.

Unidades CV, CV~ e Cv: As unidades **Cv** foram inseridas na posição final de um logotoma de duas sílabas, em posição átona, assim como a unidade contendo a vogal {@}. As outras unidades foram inseridas em logotomas de três sílabas, ocupando a posição central, em posição tônica, por exemplo, a unidade {tE} foi inserida na frase: "Falo pateta pauzado" {_falUpatEtapauzadU_}

Conforme dito anteriormente, a redução do inventário concatenando as nasais com inícios orais elimina a necessidade de recorte das unidades **CV~**. Neste caso as unidades **CV** foram recortadas 1 período de pitch após o fim da consoante. Na Figura 5.4-1 são mostrados o sinal e o espectrograma da concatenação das unidades {_f f@ a_ _f f@ a~_ _f f@ @_}, utilizando o método de concatenação descrito a seguir. Observa-se que este critério parece oferecer bons resultados, porém é preciso validá-lo para os restantes das unidades, e principalmente para os segmentos mais problemáticos como os ditongos e as líquidas.

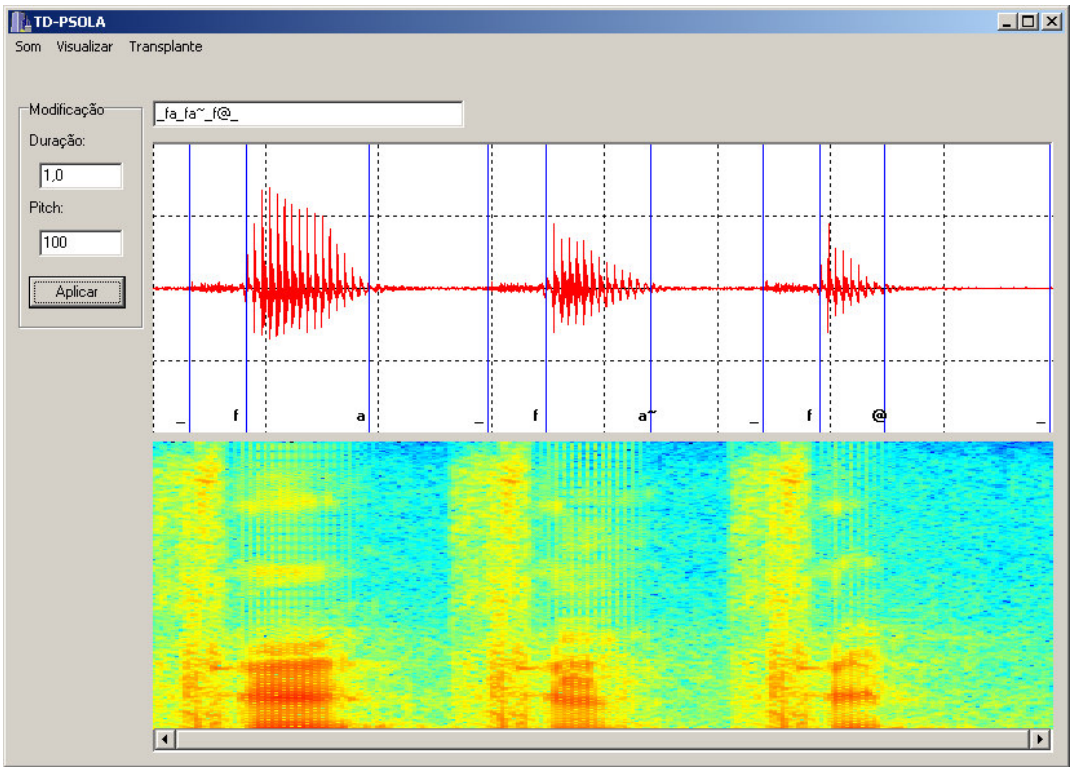


Figura 5.4-1: Sinal e espectrograma da concatenação de vogais nasais e semivogais com inícios orais. Os segmentos concatenados foram {_f f@ a_ _f f@ a~_ _f f@ @_}.

5.5. Concatenação e síntese das unidades

Após o recorte de uma unidade, é prudente observar como que esta se comporta na concatenação e se for preciso recortá-la novamente, ou gravar outro logotoma. Geralmente este é um processo interativo e por isso foi implementado no *Editor* um método de síntese por concatenação, utilizando o método TD-PSOLA para manipulação prosódica.

A partir de uma seqüência de caracteres fonéticos no formato SAMPA, o sistema irá buscar no banco as unidades correspondentes e concatená-las. O banco de unidades neste caso específico, trata-se de é um diretório contento uma unidade por arquivo, na qual o nome do arquivo se refere aos códigos das unidades correspondentes. O esquema de busca pela unidade foi implementado de maneira que fosse possível utilizar unidades de tamanhos diferenciados, no máximo com 5 segmentos. Isto foi feito buscando da seguinte forma: dada a seqüência fonética o algoritmo procura primeiramente pela maior unidade possível contendo 5 segmentos, se não encontra ele procura pela unidade contendo 4 segmentos e assim por diante. Isso permite utilizar desde palavras até fones como unidades de concatenação, mas que não será o caso.

Optou-se por primeiramente concatenar as unidades e somente em seguida manipulá-las com o uso do TD-PSOLA. Na concatenação de segmentos sonoros foi feita a aplicação da metade uma janela de hanning (do centro para a esquerda), síncrona com a penúltima marca de pitch. Esta janela foi sobreposta ao seu segmento adjacente, janelado da mesma forma, porém com a parte da janela com decaimento inverso. Inicialmente foi feita uma equalização em amplitude nos segmentos adjacentes, baseado na energia dos segmentos adjacentes, mas não ofereceu bons resultados.

As marcas de pitch das unidades também são concatenadas e serão usadas como parâmetro para o método TD-PSOLA. É permitido ao usuário impor duração e pitch constantes ao sinal concatenado. Nesta etapa, é possível validar o desempenho do método proposto para a obtenção das marcas de pitch, na qual se observou, na concatenação de algumas unidades, a ausência de erros de fase no sinal sintético. Esta

conclusão foi obtida tanto visualmente, observando o sinal no domínio do tempo e no domínio da freqüência, através do espectrograma, quanto de maneira auditiva.

Na Figura 5.5-1 é mostrada a tela do *Editor* utilizada para a síntese. O exemplo mostrado é da palavra "tinta" {_tSi~ta_}. Observa-se, principalmente no espectrograma, a ausência de descontinuidades graves que seriam percebidas na reprodução do sinal.

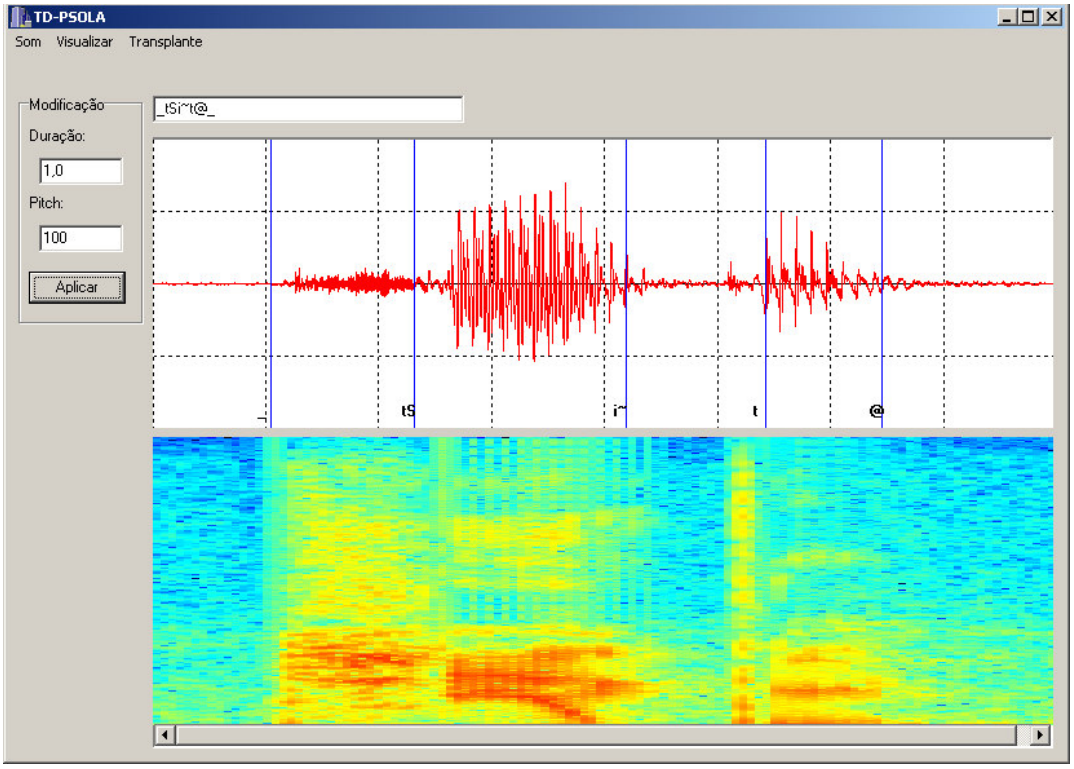


Figura 5.5-1: Tela do sistema de síntese por concatenação temporal, implementado no *Editor*. Observa-se no espectrograma a ausência de descontinuidades graves, demonstrando a eficiência do método de obtenção das marcas de pitch.

Na Figura 5.5-1 também são mostradas as fronteiras etiquetadas dos segmentos fonéticos. Conforme dito anteriormente, estas fronteiras serão utilizadas futuramente para a aplicação de regras prosódicas. Uma validação do posicionamento das marcas, poderia ser obtida pela imposição de regras de variação do contorno de pitch e das durações, porém este método estaria mais sujeito a erros do modelo de regras aplicado do que do posicionamento das marcas. Deste modo, optou-se por um processo interessante que é o transplante de prosódia. Isto significa impor ao sinal concatenado o mesmo contorno de pitch e durações obtidas de um sinal de fala natural.

O transplante de prosódia é feito no *Editor* na mesma tela mostrada na Figura 5.5-1. Inicialmente é preciso um arquivo de fala natural, etiquetado, cujo nome do arquivo é a própria transcrição fonética do sinal falado. Em seguida este arquivo é aberto através do *Editor*, que irá concatenar a seqüência fonética contida no nome do arquivo. Depois disso, fará a extração do contorno de pitch e comparando as marcas de fronteira fonética do sinal natural com o sinal concatenado, irá determinar o padrão de modificação das durações. A seguir, o sinal concatenado, as marcas de pitch, o contorno de pitch e a curva de durações, é submetido ao TD-PSOLA. O resultado deste processo é bem interessante, pois é possível obter um sinal sintetizado com características bem naturais. A vantagem deste método, na avaliação das unidades, é que ele permite que uma unidade seja avaliada dentro de um contexto natural, o que pode levar o usuário a concluir, por exemplo, que sua unidade está curta ou longa demais, ou que a gravação da unidade poderia ter sido feita em um tom mais baixo ou mais alto. Isto permite um processo interativo de refinamento das unidades que pode chegar a uma qualidade de síntese muito boa.

5.6. Conclusão

Neste capítulo foi descrita a implementação de um aplicativo chamado de *Editor*. Este aplicativo permite que todo o processo de construção do banco seja feito em um único ambiente de trabalho.

Para a detecção automática das marcas de pitch no *Editor*, foi implementado um método de determinação automática do atraso do sinal de voz em relação ao microfone de contato, assim como o próprio método de detecção dos GCIs apresentado no Capítulo 4.

A técnica de segmentação apresentada no Capítulo 3 foi implementada no *Editor*, porém para melhorar a eficiência na segmentação, ou seja, aumentar a coincidência das fronteiras acústicas com as fonéticas, foi explorado a eficiência do algoritmo de detecção de pitch em determinar os trechos sonoros e surdos. Foi observado que o

método quando usado no recorte de logotomas apresenta resultados bastante úteis para auxiliar na identificação e recorte das unidades.

Sendo a construção do banco um processo recursivo, foi implementado no *Editor* um método de busca e concatenação de unidades e o método TD-PSOLA para manipulação das durações e do contorno de pitch. Nesta etapa, é possível validar o desempenho do método proposto para a obtenção das marcas de pitch, na qual se observou, na concatenação de algumas unidades, a ausência de erros de fase no sinal sintético. E foi também apresentado o conceito de transplante de prosódia, que pode ser utilizado para avaliar o método de inserção de marcas entre segmentos.

Capítulo 6

Conclusão

Nesta tese foi abordada uma das etapas de desenvolvimento de um sistema de conversão texto-fala, que é a construção do banco de unidades para a síntese por concatenação, na qual foi assumido o método TD-PSOLA para a manipulação prosódica dos segmentos.

Inicialmente, foi feita uma revisão de alguns conceitos da fonética e da fonologia, considerados necessários tanto para o entendimento da literatura específica, quanto para o próprio conhecimento da língua. Este assunto foi apresentado no Capítulo 2, onde também foi feita uma compilação de trabalhos que relatam diferentes experiências na construção de bancos de unidades. Esta abordagem foi capaz de oferecer um melhor entendimento dos problemas envolvidos no processo de construção, sendo referenciada algumas vezes como uma etapa de planejamento.

Ao fim desta etapa, é constatado que o processo de construção demanda de um laborioso processo manual que consome grande parte do tempo de desenvolvimento. Deste modo para as tarefas que exigem maior quantidade de trabalho manual, que são a

segmentação do sinal ao nível fonético e a detecção das marcas de pitch nos segmentos, foram propostos métodos semi-automáticos, apresentados ao longo dos capítulos 3 e 4.

No capítulo 3 foi apresentada uma abordagem teórica sobre o problema de segmentação de um sinal, onde foi utilizada a programação dinâmica como uma solução para o problema. Dada a similaridade observada com os métodos de reconhecimento de dígitos conectados, os métodos TLDP e LBDP foram revisados e aplicados ao caso de segmentação do sinal de voz, onde se concluiu que o método TLDP, apesar de ser considerado como uma proposta menos eficiente no reconhecimento de dígitos, se mostrou mais adequado à proposta de segmentação

Foi observado ainda que o método de segmentação proposto trata-se de um processo de *quantização vetorial*, ou uma técnica de *clustering*, onde a ordem das amostras é considerada. E sob este ponto de vista, foi observado um caso de super-segmentação, ou seja, quando o número de segmentos a serem encontrados são especificados além do número de segmentos fonéticos. Utilizando este processo, foi verificado que as marcas de fronteira, inseridas pelo método, tornam-se mais coincidentes com os segmentos fonéticos.

O problema de obtenção das marcas de pitch, foi abordado no Capítulo 4, onde foi estabelecido um método semi-automático para a obtenção das marcas de pitch, coincidentes com os GCIs, reduzindo a possibilidade de erros de fase na concatenação. O método proposto representa um método de baixo custo, robusto e preciso, alternativo ao uso do *eletroglotógrafo*. O método utiliza a gravação simultânea do sinal de voz e do sinal obtido por um disco piezoelétrico em contato com a pele, localizado na base do pescoço, no qual foi necessária a montagem de um circuito de pré-amplificação de dois canais utilizando a entrada line-in da placa de som para a aquisição dos sinais. A precisão na localização dos GCIs fornecida pelo método, foi verificada como satisfatória na concatenação de segmentos, apresentando uma concatenação livre de erros de fase.

Foi observado que parte do tempo destinado à construção do banco é usado na manipulação de diferentes aplicativos, um para gravar, outro para recortar, etc. Assim,

todas as tarefas relacionadas à construção do banco foram implementadas em um aplicativo chamado de *Editor*, no qual foram incluídas as técnicas semi-automáticas descritas nos Capítulos 3 e 4.

A técnica de segmentação apresentada no capítulo 3 foi modificada no *Editor* de modo a explorar a robustez do método de inserção dos GCIs em detectar os trechos sonoros e surdos. Além disso, a transcrição fonética dos logotomas foi utilizada para a etiquetagem dos segmentos. Conclui-se que o método quando usado no recorte de logotomas apresenta resultados úteis para auxiliar na identificação e recorte das unidades.

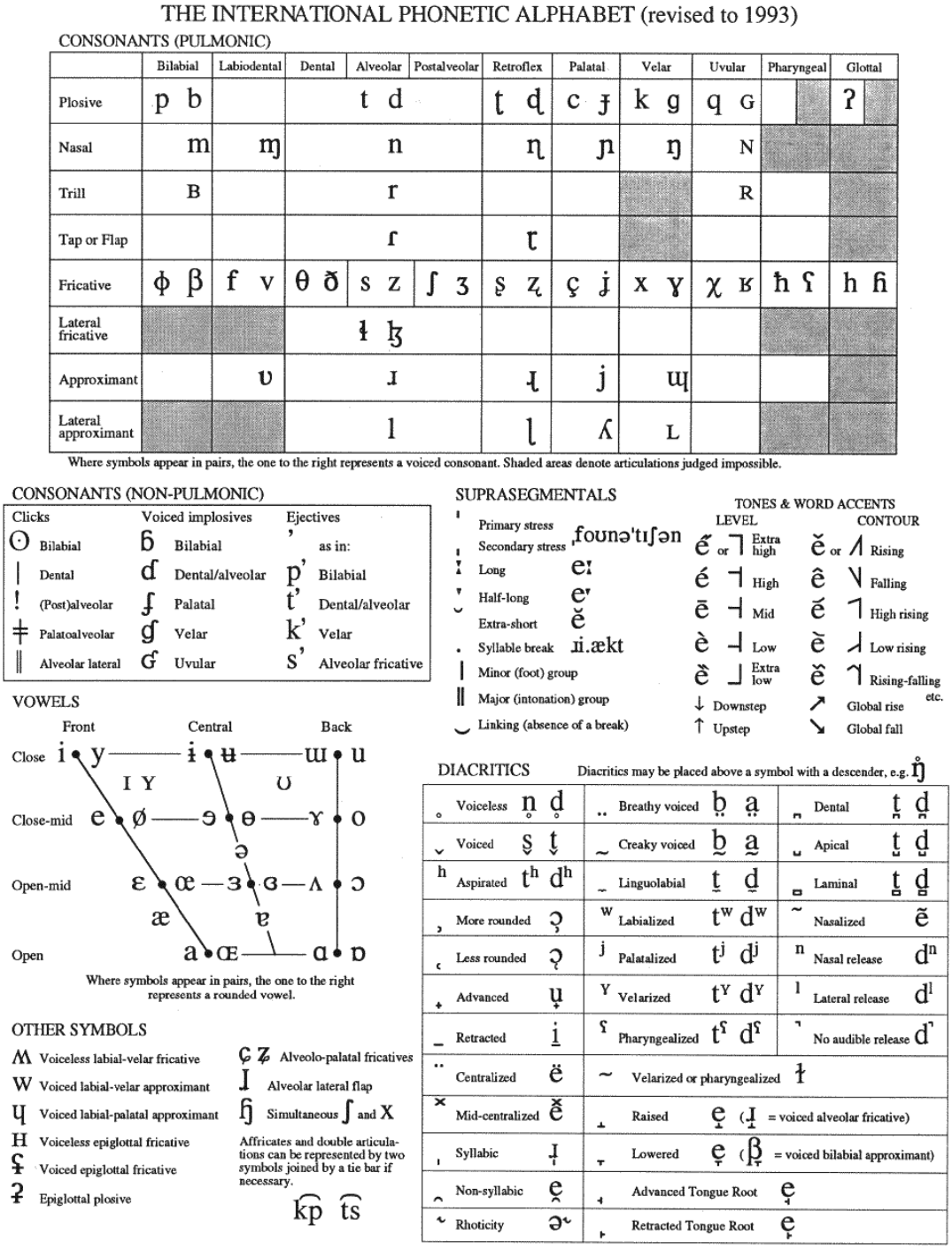
Sendo a construção do banco um processo recursivo, também foi implementado no *Editor* o método TD-PSOLA para manipulação das durações e do contorno de pitch, no qual através de um *tranplante de prosódia* é possível, em tempo de construção do banco, avaliar as unidades recortadas em um ambiente prosódico natural.

Futuramente, sugere-se a construção completa de um banco de unidades, conforme descrita nesta tese, considerando os processos de redução do inventário sendo as gravações anotadas para uma consulta futura na construção de novas vozes.

Sugere-se ainda que sejam experimentados novos sensores para captação da vibração das cordas vocais, assim como outros circuitos que possam ser mais adequados a este tipo de instrumentação. Neste sentido, sugere-se ainda que na marcação dos GCIs seja verificada a utilidade e precisão em posicionar os GCIs a partir da busca direta pelos mínimos locais encontrados no sinal do microfone de contato.

Após a construção do primeiro banco de unidades, as outras etapas no desenvolvimento de um sistema de conversão texto-fala podem ser prosseguidas, como a conversão grafema-fonema e os módulos de determinação de regras prosódicas. Para a construção de novas vozes, ou para a abordagem de um sistema de concatenação por seleção automática de unidades, que requerem grande quantidade de corpus etiquetado, sugere-se a exploração dos métodos de segmentação e etiquetagem baseados no DTW.

ANEXO A: Alfabeto fonético internacional



THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	ʼ	as in:
|	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
!	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən	
ˌ	Secondary stress		
ː	Long	eː	
ˑ	Half-long	eˑ	
̆	Extra-short	ĕ	
.	Syllable break	ɹi.ækt	
|	Minor (foot) group		
‖	Major (intonation) group		
‿	Linking (absence of a break)		

TONES & WORD ACCENTS

LEVEL		CONTOUR	
é or ˥	Extra high	ě or ˩˥	Rising
é ˦	High	ê ˥˩	Falling
ē ˧	Mid	ě ˧˥	High rising
è ˨	Low	è ˩˧	Low rising
ȅ ˩	Extra low	ȅ ˥˩˧	Rising-falling
↓	Downstep	↗	Global rise
↑	Upstep	↘	Global fall

etc.

VOWELS

Where symbols appear in pairs, the one to the right represents a rounded vowel.

DIACRITICS

Diacritics may be placed above a symbol with a descender, e.g. n̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	i̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
◌˞	Rhoticity	ə˞	◌̙	Retracted Tongue Root	e̙			

OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ʃʒ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

k͡p t͡s

ANEXO B: Segmentos fonéticos do português brasileiro.

Segmentos consonantais

Símbolo	Classificação	Exemplo	Transcrição
p	Oclusiva bilabial desvozeada	**p**ata	[ˈpata]
b	Oclusiva bilabial vozeada	**b**ala	[ˈbala]
t	Oclusiva alveolar desvozeada	**t**apa	[ˈtapa]
d	Oclusiva alveolar vozeada	**d**ata	[ˈdata]
k	Oclusiva velar desvozeada	**c**apa	[ˈkata]
g	Oclusiva velar vozeada	**g**ata	[ˈgata]
tʃ	Africada alveopalatal desvozeada	**t**ia	[ˈtʃia]
dʒ	Africada alveopalatal vozeada	**d**ia	[ˈdʃia]
f	Fricativa labiodental desvozeada	**f**aca	[ˈfaka]
v	Fricativa labiodental vozeada	**v**aca	[ˈvaka]
s	Fricativa alveolar vozeada	**s**ala, ca**ç**a	[ˈsala], [ˈkasa]
z	Fricativa alveolar vozeada	ca**s**a	[ˈkaza]
ʃ	Fricativa alveopalatal desvozeada	**ch**á, a**ch**a	[ˈʃa], [ˈaʃa]
ʒ	Fricativa alveopalatal vozeada	**j**á	[ˈʒa]
x	Fricativa velar desvozeada	**r**ata	[ˈxata]
ɣ	Fricativa velar vozeada	ca**r**ga	[ˈcaɣga]
m	Nasal bilabial vozeada	**m**ala	[ˈmala]
n	Nasal alveolar vozeada	**n**ada	[ˈnada]
ɲ ou ỹ	Nasal palatal vozeada	ba**nh**a*	[ˈbaɲa] ou [ˈbãỹa]
ɾ	Tepe alveolar vozeado	ca**r**a, p**r**ata	[ˈcaɾa],[ˈpɾata]
l	Lateral alveolar vozeada	**l**ata, p**l**ana	[ˈlata], [ˈplana]
ɬ ou w	Lateral alveolar vozeada velarizada	sa**l**ta**	[ˈsaɬta], [ˈsawta]
ʎ ou lʲ	Lateral palatal vozeada	malha*	[ˈmaʎa] ou [ˈmalʲa]

*A primeira forma ocorre na fala de poucos falantes do português brasileiro.

** A segunda forma é um segmento com características de uma vogal do tipo [u].

Vogais tônicas orais e nasais

Símbolo	Classificação	Exemplo	Transcrição
i	Vogal alta anterior não-arredondada	vi	[ˈvi]
ĩ	Vogal alta anterior não-arredondada nasal	vim	[ˈvĩ]
e	Vogal média-alta anterior não-arredondada	ipê	[iˈpe]
ẽ	Vogal média-alta anterior não-arredondada nasal	tempo	[ˈtẽpʊ]
ɛ	Vogal média-baixa anterior não-arredondada	pé	[ˈpɛ]
a	Vogal baixa central não-arredondada	pá	[ˈpa]
ã	Vogal baixa central não-arredondada nasal	lã	[ˈlã]
ɔ	Vogal média-baixa posterior arredondada	avó	[aˈvɔ]
o	Vogal média-alta posterior arredondada	avô	[aˈvo]
õ	Vogal média posterior arredondada nasal	tom	[ˈtõ]
u	Vogal alta posterior arredondada	jacu	[ʒaˈku]
ũ	Vogal alta posterior arredondada nasal	jejum	[ʒeˈʒũ]

Vogais postônicas finais

Símbolo	Classificação	Exemplo	Transcrição
ɪ	Vogal alta anterior não-arredondada	vi	[ˈvi]
ʊ	Vogal alta posterior arredondada	vim	[ˈvĩ]
ə	Vogal média-baixa central	ipê	[iˈpe]

Ditongos

Crescentes início em [ɪ]	**Crescentes início em [ʊ]**
[ɪɪ] se reduz a [ɪ] série [ˈsɛrɪ]	[ʊʊ] se reduz a [ʊ] vácuo [ˈvakʊ]
[ɪə] área [ˈarɪə]	[ʊə] mágoa [ˈmagʊə]
[ɪʊ] áereo [ˈaɛrɪʊ]	[ʊɪ] tênue [ˈtenʊɪ]
[ɪo] gracioso [gɾaˈsɪosʊ]	

Decrescentes orais término em [ɪ]	**Decrescentes orais término em [ʊ]**
[aɪ] pai [ˈpaɪ]	[aʊ] saudade [saʊˈdadʃɪ]
[eɪ] seita [ˈseɪtə]	[eʊ] judeu [ʒuˈdeʊ]
[ɛɪ] papéis [paˈpɛɪs]	[ɛʊ] bedéu [beˈdɛʊ]
[oɪ] afoito [aˈfoɪtʊ]	[oʊ] Moscou [mosˈkoʊ]
[ɔɪ] mói [ˈmɔɪ]	[ɔʊ]* volta [ˈvɔʊtə]
[uɪ] cuida [ˈcuɪdə]	[iʊ] fugiu [fuˈʒɪʊ]

*Ocorre nos casos de vocalização do “l”, transcrito aqui por [ʊ] ao invés de [w].

Decrescentes nasais término em [ɪ] e [ʊ]
[ãɪ]* mãe [ˈmãɪ]
[ũɪ]* cuida [ˈcuɪdə]
[õɪ]* afoito [aˈfoɪtʊ]
[ẽɪ] item [ˈitẽɪ]
[ãʊ] pão [ˈpãʊ]

* Ocorrem sempre em sílabas tônicas

ANEXO C: Caracteres fonéticos SAMPA

CODIGO*	SAMPA	IPA	CODIGO*	SAMPA	IPA
10	p	p	30	l	l
11	b	b	31	L	ʎ
12	t	t	32	i	i
13	d	d	33	i~	ĩ
14	k	k	34	e	e
15	g	g	35	e~	ẽ
16	tS	tʃ	36	E	ɛ
17	dZ	dʒ	37	a	a
18	f	f	38	a~	ã
19	v	v	39	O	ɔ
20	s	s	40	o	o
21	z	z	41	o~	õ
22	S	ʃ	42	u	u
23	Z	ʒ	43	u~	ũ
24	X	x	44	@	ə
25	G	ɣ	45	@~	ə̃
26	m	m	46	I	ɪ
27	n	n	47	I~	ɪ̃
28	J	ɲ	48	U	ʊ
29	r	ɾ	49	U~	ʊ̃

50	_	**silêncio**
51	!	indefinido

*os códigos foram determinados de maneira que tivessem sempre 2 algarismos.

ANEXO D: Dynamic time warping (DTW)

Sejam dois padrões *X* e *Y*, representados pelas seqüências $(x_1, x_2, ..., x_{T_x})$ e $(y_1, y_2, ..., y_{T_y})$. Considerando i_x e i_y como notação dos índices de *X* e *Y* no tempo, onde $i_x = 1, 2, ..., T_x$ e $i_y = 1, 2, ..., T_y$, então a dissimilaridade entre *X* e *Y* é definida considerando alguma função de distorção $d(i_x, i_y)$ (RABINER, 1993).

Na técnica de linearização e normalização temporal, a dissimilaridade entre *X* e *Y* é definida por: $d(X,Y) = \sum_{i_x}^{T_x} d(i_x, i_y)$ onde $i_y = \frac{T_y}{T_x} i_x$

No entanto, um esquema mais geral envolve o uso de duas funções *warping*, ϕ_x e ϕ_y, na qual mapeiam os índices dos dois padrões, i_x e i_y, a um eixo temporal *k*, "normal", isto é:

$$i_y = \phi_y(k) \qquad k = 1, 2, ..., T$$

$$i_x = \phi_x(k) \qquad k = 1, 2, ..., T$$

Deste modo, um padrão global de medida de dissimilaridade pode ser definido baseado no par de funções *warpings*, como sendo a distorção acumulada ao longo de todas a seqüência (RABINER, 1993).

$$d_\phi(X,Y) = \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)) \qquad \text{(D.1)}$$

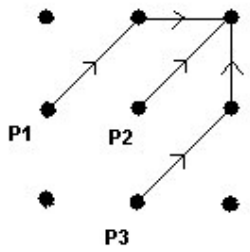
É necessário que as funções *warpings* sejam monotonicamente não decrescentes para que a ordem temporal seja mantida. Obviamente existem inúmeras possibilidades de funções *warpings*. Uma escolha natural é definir a dissimilaridade $d(X,Y)$ como o mínimo de $d_\phi(X,Y)$, tal que:

$$d(X,Y) \triangleq \min_{\phi} d_\phi(X,Y) \qquad \text{(D.2)}$$

Na realidade existem várias escolhas da medida de distorção ou de dissimilaridade. O mais comum é a medida do erro quadrático (RABINER, 1993).

O ponto fundamental é que encontrar o melhor alinhamento entre um par de padrões é equivalente a encontrar o melhor caminho através de uma grade, mapeando um padrão ao outro. Encontrar o melhor caminho requer resolver um problema de minimização (RABINER, 1993).

Seja um caminho *P* como uma seqüência de movimentos, cada um especificado por um par de incremento nas coordenadas. $P{:}(p_1,q_1)(p_2,q_2)...(p_k,q_k)...(p_T,q_T)$



Por exemplo:

P1:*(1,1)(1,0)*

P2:*(1,1)*

P3:*(1,1)(0,1)*

Para um caminho que começa em *(1,1),* normalmente $p_1 = q_1 = 1$ (como se o caminho tivesse origem sempre em *(0,0)*) e portanto:

$$\phi_x(\mathrm{k}) = \sum_{i=1}^{k} p_i \qquad \phi_y(\mathrm{k}) = \sum_{i=1}^{k} q_i$$

Além disso

$$T_x = \sum_{k=1}^{T} p_k \qquad T_y = \sum_{k=1}^{T} q_k$$

Alguns tipos de imposições locais a estes caminhos são mostrados na Figura D-1. A definição destas imposições locais é puramente heurística. O que possuem em comum é a especificação do passo na forma *(1,1)* para preservar o alinhamento temporal linear.

Retornando ao problema de minimização da função de dissimilaridade, podemos reescrever que para duas seqüências de tamanho T_x e T_y:

$$\begin{aligned} d(X,Y) &\triangleq D(T_x,T_y) \\ &= \min_{\phi_x,\phi_y} \sum_{k=1}^{T} d(\phi_x(k),\phi_y(k)) \end{aligned} \tag{D.3}$$

deste modo, o caminho parcial ligando o ponto *(1,1)* e (i_x,i_y) será:

$$D(i_x, i_y) \triangleq \min_{\phi_x, \phi_y} \sum_{k=1}^{T'} d(\phi_x(k), \phi_y(k)) \quad \text{onde} \quad \phi_x(T') = i_x \quad e \quad \phi_y(T') = i_y$$

Aplicando o principio de Bellman (RABINER, 1993), a recursão por programação dinâmica resulta em:

$$D(i_x, i_y) = \min_{(i_x', i_y')} \left[D(i_x', i_y') + \zeta((i_x', i_y'), (i_x, i_y)) \right] \tag{D.4}$$

sendo

$$\zeta((i_x', i_y'), (i_x, i_y)) = \sum_{l=0}^{L_s} d(\phi_x(T' - l), \phi_y(T' - l)) \tag{D.5}$$

onde L_s é o número de passos do caminho do ponto (i_x', i_y') ao ponto (i_x, i_y) e

$$\phi_x(T' - L_s) = i_x' \quad e \quad \phi_y(T' - L_s) = i_y' \tag{D.6}$$

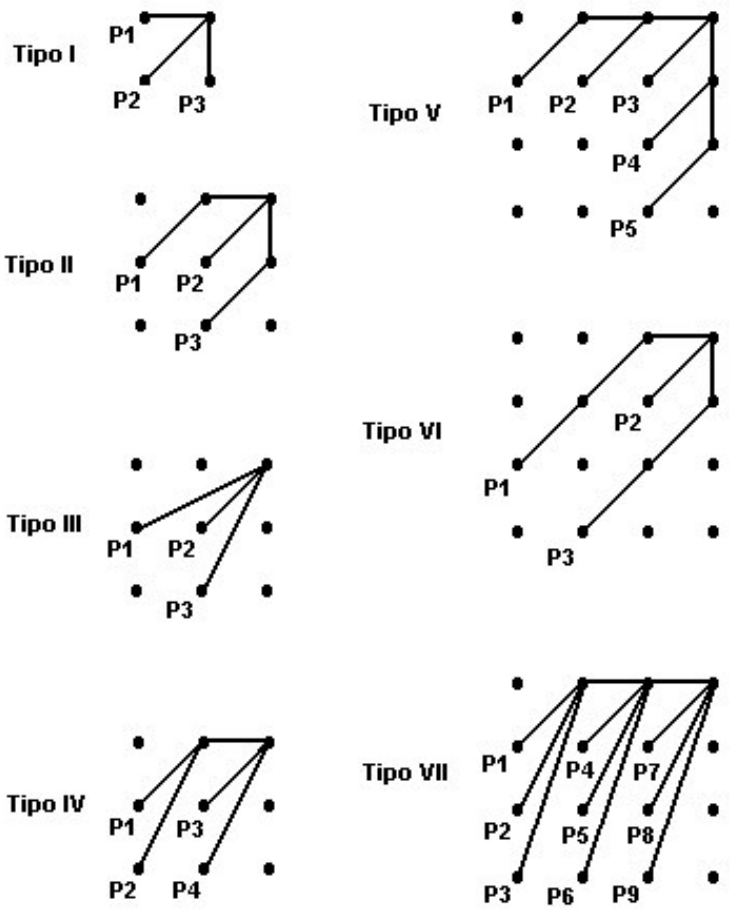


Figura D-1: Tipos de imposição locais à função de alinhamento temporal.

Foi implementado o DTW, para alinhamento de dois sinais, traçados na Figura D-2.

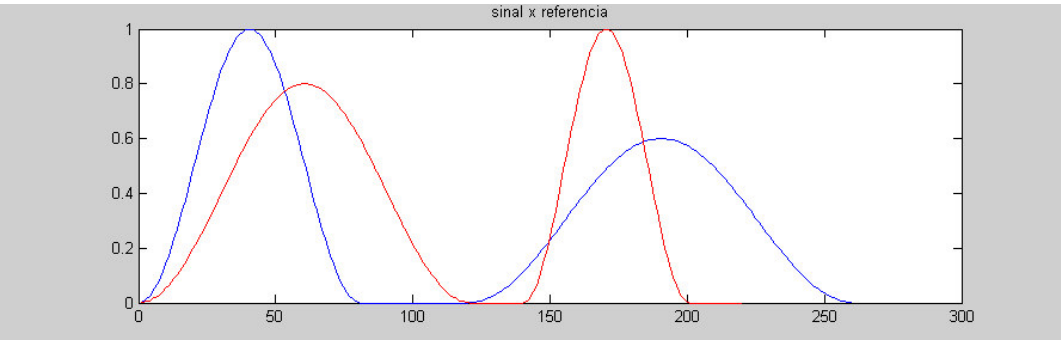


Figura D-2: Sinais de exemplo para aplicação do DTW.

A matriz de dissimilaridade, foi obtida pelo erro médio quadrático. Na Figura D-3, à esquerda, é mostrada uma imagem referente a matriz de dissimilaridade local. Os pontos de maior similaridade aparecem em preto. À direita tem-se a matriz de dissimilaridade (acúmulo das dissimilaridades locais), onde os pontos claros representam os pontos de menor dissimilaridade acumulada. Deste modo, a linha em vermelho indica o caminho de menor dissimilaridade, encontrado pelo algoritmo.

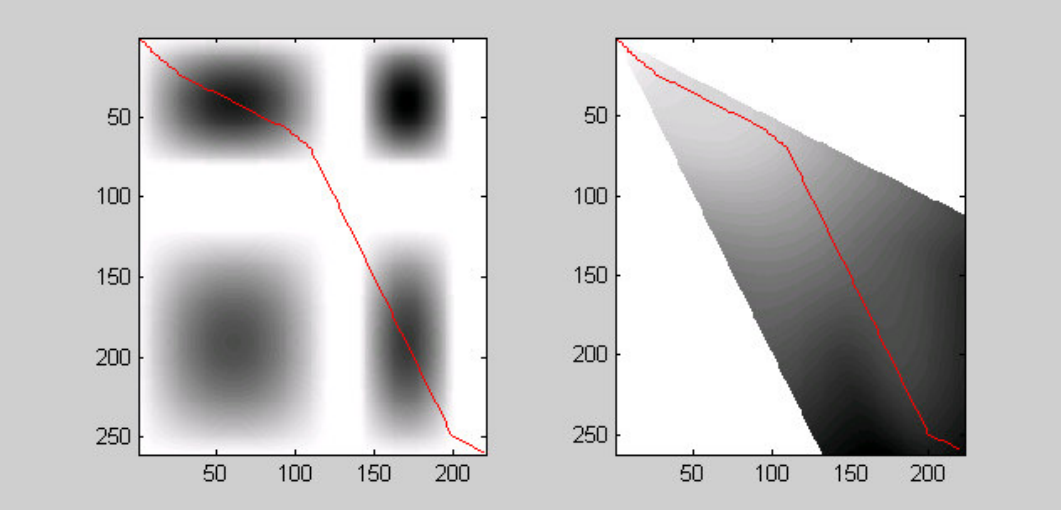


Figura D-3: A esquerda a matriz de dissimilaridade local e a direita a matriz de dissimilaridade acumulada. A linha em vermelho indica o camiho de menor similaridade.

Na Figura D-4, o sinal em vermelho, tomado como referência, foi modificado pela curva gerada pelo DTW para observarmos a aproximação realizada.

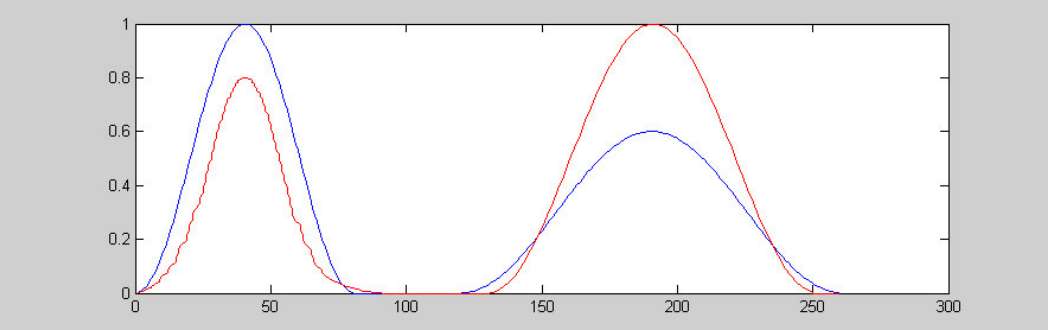


Figura D-4: Sinal de teste (azul) e sinal de referência (vermelho) modificado pela curva de warping.

Referências

ADELL, J., BONAFONTE, A., 2004, "Towards phone segmentation for concatenative speech synthesis", In *Proceedings of 5th ISCA Speech Synthesis Workshop*, Pittsburg, pp.139-144, Junho.

ALBANO, E. C., AQUINO, P. A., 1997, "Linguistic criteria for building and recording units for concatenative speech synthesis in Brazilian Portuguese", In *Proceedings Eurospeech 97*, v.2, pp. 725-728.

ALCAIM, A., SOLEWICZ, J.A., MORAES, J.A., 1992, "Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro", In *Revista da Sociedade Brasileira de Telecomunicações*, v. 7, n.1, pp. 23-41, Dezembro.

ANANTHAPADMANABHA T.V., YEGNANARAYANA, B., 1979, "Epoch Extration from linear prediction residual for identification of closed glottis interval", In *IEEE Transaction on Acoustic, Speech and Signal Processing*, v. 27, n. 4, pp. 309-319, Agosto.

ASKENFELT, A., GAUFFIN, J., SUNDBERG, J., et al., 1980, "A comparison of contact microphone and electroglottograph for the measure of fundamental frequency", In *Journal of Speech and Hearing Research*, v. 23, n. 2, pp. 258-273, Junho.

BARBOSA, P. A., VIOLARO, F., ALBANO, E.C., et al., 1999, "Aiuruetê: A high-quality concatenative text-to-speech system for Brazilian Portuguese with demisyllabic analisys-based unit and a hierarchical model of rhythm production", In *Proceedings of Eurospeech 99*, v. 5, 1059-1062.

BARNER, K. E., GALLANT, J.A., 1994, "Nonlinear estimation of EGG signals with application to speech pitch detection", In *Proceedings of the 16th Annual International Conference of the IEEE*, *Engineering in Medicine and Biology Society*, v. 2, pp. 1340-1341, Novembro.

BISHOP, M.C., 1995, *Neural Networks for pattern Recognition*, Oxford, Oxford University Press.

BISOL, L., 1999, *Introdução a estudos de fonologia do Português brasileiro*, Porto Alegre, Pontifícia Universidade Católica do Rio Grande do Sul.

BLACK, A.W., LENZO, K.L., 2003, *Building Synthetic Voices*, FestVox 2.0.

BIGORNE D., BOEFFARD O., CHERBONNEL B., et al., 1993, "Multilingual PSOLA text-to-speech system", In *Proceedings of ICASSP*, v. 2, pp. 187-190.

BHASKARARAO, P., 1991, "Use of triphones for demissylable-based speech synthesis", In *Proceedings of ICASSP*, v. 1, pp. 517-520.

CALLOU, D., LEITE Y., 1990, *Iniciação à Fonética e Fonologia*, Rio de Janeiro, Jorge Zahar Editora Ltda.

CHANGXUE MA, KAMP Y. K., WILLEMS L. F., 1994, "Frobenius norm approach to glottal closure detection from the speech signal", In *IEEE Transaction on Speech and Audio Processing*, v.2, pp. 258-265.

CHARPENTIER, F., MOULINES, E., 1989, "Pitch-synchronous wave form processing techniques for text-to-speech synthesis using diphones", In *Proceedings of Eurospeech 89*, v. 2, pp. 13-19.

CHENG, Y. M., O'SHAUGHNESSY, D., 1989, "Automatic and reliable estimation of glottal closuse instants and period", In *IEEE Transactions on Acoustics, Speech, and Signal processing.,* v. 37. n. 12, Dezembro.

CITTADINNI, R., POULAN, F., 2002, "TS971 based electret condenser microphone amplifier", *Application Note AN1534*, STMicroeletronics, Março.

DAVIS, S. B., MERMELSTEIN, P., 1980, "Comparison of parametric representation for monosyllabic word recognition in continously spoken sentences", In *IEEE Transactions on Acoustic, Speech and Signal Processing*, v. 28, n. 4, pp. 357-366, Agosto.

DELLER, JR. J. R., HANSEN J. H. L., PROAKIS, J. G. , 1993, *Discrete-Time Processing of Speech Signal*, IEEE Press.

DONOVAN, R.E., 1996, *Trainable Speech Synthesis*, Ph.D. thesis, Cambridge University Engineering Department, UK.

DJURIC, P. M., KAY, S.M., BOUDREAUX-BARTELS, G.F., 1992, "Segmentation of nonstationary signals", In *Proceedings of ICASSP*, v. 5, pp. 161-164.

DUTOIT, T. , 1997, *An Introduction to Text-To-Speech Synthesis*, London, Kluwer Academic Publishers.

GÓMEZ, J., CASTRO, M.J., 2002, "Automatic segmentation of speech at the phonetic level", *Joint IAPR International Workshops SSPR 2002 and SPR 2002*, pp. 672-680, Agosto.

GRACIARENA, M., FRANCO, H., SONMEZ, K., et al., 2003, "Combining standard and throat microphones for robust speech recognition",In *IEEE Signal Processing Letters*, v. 10, n. 3, Março.

HAMON, C., MOULINES, E., CHARPENTIER, F., 1989, "A diphone synthesis system based on time-domain prosodic modification of speech", In *Proceedings of ICASSP*, pp. 238-241.

HART, D., 1973, “Unsupervised learning and clustering”, In *Pattern Classification and Scene Analysis*, chapter 6, Wiley.

HOSON, J.P., 2000, *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information,* Ph.D. thesis, Oregon Graduate Institute of Science an Technology, Oregon, USA.

HUNT, A. J., BLACK, A.W., 1996, “Unit selection in a concatenative speech synthesis system using a large speech database”, In *Proceedings ICASSP*, v. 1, pp. 373-376.

KAFKA, S.G., PACHECO, F. S., SEARA, I.C. et al., 2002, “Utilização de segmentos transicionais homorgânicos em síntese de fala concatenativa”, *XIV Congresso Brasileiro de Automática*, Natal, Setembro.

KAY, S. HAN X., 2001, “Optimal segmentation of signals based on dynamic programming and its application to image denoising and edge detection”, *Unpublished notes.*

KRAFT V., ANDREWS, J. R., 1992, “Design, evaluation and acquisition of a speech database for german synthesis-by-concatenation”, In *Proceedings of SST 92*, Brisbane, pp. 724-729.

KRISHNAMURTHY, A.K., CHILDERS, D. G., 1986, “Two-channel speech analysis”, In *IEEE Transactions on Acoustics, Speech, and Signal processing*, v. 34, n. 4, pp. 730-743, Agosto.

LATSCH, V. L., 2002, *Conversor texto-fala compatível com a Microsoft Speech API*, Projeto de Fim de Curso, DEL/UFRJ, Rio de Janeiro, Maio.

LATSCH, V. L., NETTO, S.L. 2005, “Obtenção de marcas de pitch em sinais de voz para síntese por concatenação temporal”, *IX Convenção Nacional, Sociedade de Engenharia de Áudio*, São Paulo, Abril.

LENZO, K. A., BLACK A. W., 2002, “Diphone collection and synthesis”.

LJOLJE, A., RILEY, M. D., 1991, “Automatic segmentation and labelling of speech”, In *Proceedings of ICASSP*, pp. S473-S476, Toronto, Maio.

LJOLJE, A., RILEY, M. D., 1993, “Automatic Segmentation of Speech for TTS”, In *Proceedings of Eurospeech 93*, v. 2, pp. 1445-1448, Setembro.

LJOLJE, A., HIRSCHBERG, J., VAN SANTEM, P. H., 1996, “Automatic speech segmentation for concatenative inventory selection”, In: Spring-Verlag, *Progress In Speech Synthesis*, New York, pp. 304-331.

MALFRÈRE, F., T. DUTOIT, 1997 “High-quality speech synthesis for phonetic Speech segmentation”, In *Proceedings of Eurospeech 97*, v. 5, pp. 2631-2634, Rhodes, Greece.

MYERS, C., RABINER, L.R., 1981, “A level building dynamic time warping algorithm for connected word recognition”, In *IEEE Transactions on Aocustics, Speech, and Signal processing*, v. 29, n. 2, Abril.

NAKAGIMA, S., HAMADA, H., 1988, “Automatic generation of synthesis units base don Context-Oriented-Clustering”, In *Proceedings of ICASSP*, v. 51, S14.2, pp. 659-662.

RABINER, L., 1993, *Fundamental of Speech Recognition*, New Jersey, Prentice Hall.

SHARMA, M., MAMMONE, R., 1996, “Blind speech segmentation: Automatic segmentation of speech without linguistics knowledge”, In *Proceedings of ICSLP 96*, Philadelphia, PA, USA Outubro.

SETHY, A., NARAYANAN, S. , 2002, “Refined speech segmentation for concatenative speech synthesis”, In *Proceedings of ICSLP 02*, Denver.

SILVA, T.C. , 2003, *Fonética e Fonologia do Português: Roteiro de Estudos e Guia de Exercícios*, São Paulo, Editora Contexto.

SIMÕES, F. O., 1999, *Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil*, Tese de M. Sc., UNICAMP, Campinas, Brasil.

SLANEY, M., 1998, *Auditory Toolbox, Version2, Technical Report No: 1998-2010*, Interval Research Corporation.

SOLEWICZ, J.A., MORAES, J.A., ALCAIM, A., 1994, "Text-to-speech system for brazilian portuguese using a reduced set of synthesis units", *International Symposium on Speech, Image Processing and Neural Networks*, pp. 579 - 582, Hong Kong, Abril,

SOLIMAR S. S., 2001, "Sistema de Conversão Texto-Fala Usando Unidades Silábicas", *Projeto de fim de curso*, DEL/UFRJ, Rio de Janeiro, Brasil.

SMITS, R. , YEGNANARAYANA, B. , 1995, "Determination of instants of significant excitation in speech using group delay function**",** In *IEEE Transactions on Speech, Audio and Signal Processing,* v. 3, pp. 325-333.

STRUBE, H. W., 1974, "Determination of the instants of glottal closure from the speech wave", In *Journal of Acoustic Society American,* v. 56, pp. 1625-1629.

SVENDSEN, T., SOONG, F., 1987, "On the automatic segmentation of speech signal", In *Proceedings of ICASSP*, pp. 77-88, Abril.

TEIXEIRA, J. P., FREITAS, D., BRAGA, D., BARROS, M. J., LATSCH, V., 2001, "Phonetic events from the labeling the european portuguese database for speech synthesis, FEUP/IPB-DB", In *Proceedings of Eurospeech 01*, Aalborg.

VAN HEMERT, J. P., 1991, "Automatic segmentation of speech", In *IEEE Transactions on Acoustics, Speech, and Signal processing*, v. 39, n. 4, Abril.

VAN SANTEM, J. P. H., SPROAT, R. W., 1999, "High-accuracy automatic segmentation", In *Proceeding sof Eurospeech 99,* Budapest, Hungary.

YOSHIDA, Y., NAKAJIMA, S., HAKODA, K., et al., 1996, "A new method of generating speech synthesis units based on phonological knowledge and clustering technique". In *Proceedings of ICSLP 96*, v. 3, pp 1712-1715, Outubro.

WONG, D.Y., MARKEL, J.D., GRAY, A.H., JR., 1979, "Least square glottal inverse filtering from the acoustic speech waveform", In *IEEE Transactions on Acoustics, Speech, and Signal processing*, v. 27, pp. 350-355.