UNIVERSIDADE FEDERAL DE SANTA CATARINA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

DETECÇÃO E TRATAMENTO DE CLIQUES NATURAIS EM BANCOS DE FALA VISANDO SÍNTESE CONCATENATIVA DE ALTA QUALIDADE

Dissertação submetida à
Universidade Federal de Santa Catarina
como parte dos requisitos para a
obtenção do grau de Mestre em Engenharia Elétrica.

MONIQUE VITÓRIO NICODEM

Florianópolis, janeiro de 2006.

DETECÇÃO E TRATAMENTO DE CLIQUES NATURAIS EM BANCOS DE FALA VISANDO SÍNTESE CONCATENATIVA DE ALTA QUALIDADE

Monique Vitório Nicodem

'Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Engenharia Elétrica, Área de Concentração em *Comunicações e Processamento de Sinais*, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina.'

_____

Prof. Rui Seara, Dr.
Orientador

_____

Prof. Alexandre Trofino Neto, Dr.
Coordenador do Programa de Pós-Graduação em Engenharia Elétrica

Banca examinadora:

_____

Prof. Rui Seara, Dr.
Presidente

_____

Prof. Sidnei Noceti Filho, Dr.

_____

Prof[a]. Izabel Christine Seara, Dr[a].

_____

Prof. Walter Pereira Carpes Jr., Dr.

Aos meus pais, Mário e Gilza,
à minha irmã, Michelle,
ao meu avô materno, Gil,
e ao meu esposo, Erni.

AGRADECIMENTOS

A Deus, por estar sempre presente me conduzindo em todos os caminhos.

Ao professor Rui Seara, pela amizade, motivação e orientação.

À Daiana, pelo auxílio na realização de experimentos constantes neste trabalho.

Ao Fernando, pela amizade e pela constante disposição em ajudar.

À Izabel, à Sandra e à Simone, pela alegria, amizade e pelo auxílio no entendimento de diversos conceitos das áreas de lingüística e fonética, essenciais para o desenvolvimento deste trabalho.

A todos os colegas de trabalho do LINSE que de alguma forma contribuíram para o desenvolvimento deste trabalho.

Ao CNPQ, pelo suporte financeiro.

À minha família e à família de meu esposo, pelo carinho, motivação e compreensão.

Aos meus pais, Mário e Gilza, à minha irmã, Michelle, e ao meu avô materno, Gil, pelo convívio diário, por todo o carinho e pelo auxílio, sempre dispensados.

Ao meu esposo, Erni, pelo amor, compreensão, motivação e alegria, sempre presentes.

Resumo da Dissertação apresentada à UFSC como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

DETECÇÃO E TRATAMENTO DE CLIQUES NATURAIS EM BANCOS DE FALA VISANDO SÍNTESE CONCATENATIVA DE ALTA QUALIDADE

Monique Vitório Nicodem

Janeiro/2006

Orientador: Prof. Rui Seara, Dr.
Área de Concentração: Comunicações e Processamento de Sinais.
Palavras-chave: Detecção de cliques, melhoria de sinais de fala, síntese concatenativa de fala.
Número de Páginas: 89 p.

RESUMO: O presente trabalho estuda certas degradações produzidas involuntarimente pelo próprio aparelho fonador humano. Tais degradações são aqui nomeadas cliques naturais ou cliques involuntários. Os cliques naturais manifestam-se na forma de estalos de pequena intensidade. Esses estalos podem prejudicar consideravelmente a qualidade da fala sintética produzida por sistemas de síntese concatenativa, especialmente quando tais estalos estão associados a descontinuidades resultantes do processo de concatenação. O presente trabalho objetiva melhorar a qualidade do banco de fala e, conseqüentemente, a qualidade da fala sintética. Para tal, são abordadas técnicas visando a atenuação e/ou supressão dos cliques naturais presentes em um banco de fala. A redução dos efeitos audíveis dos cliques é obtida considerando-se dois estágios de processamento: a detecção e o tratamento propriamente dito. Técnicas de detecção baseadas em filtragem inversa, derivada de quarta ordem, modelagem do aparelho auditivo humano, decomposição *wavelet* e redes neurais artificiais são aqui analisadas. Técnicas de processamento baseadas em extrapolação, mascaramento, suavização e *pruning* são também estudadas. A avaliação das técnicas tanto de detecção quanto de tratamento é realizada tomando-se como referência um banco de fala, com duração de 45 minutos, gravado em um estúdio apropriado por um locutor profissional. Tal banco foi extraído do sistema de síntese concatenativa de fala desenvolvido no LINSE (Laboratório de Circuitos e Processamento de Sinais do Departamento de Engenharia Elétrica da UFSC). Testes perceptuais baseados em escuta são realizados visando avaliar as técnicas estudadas. Os resultados dos testes indicam uma significativa melhoria na qualidade da fala gravada. O trabalho aqui desenvolvido aplica-se não somente a bancos de fala visando sistemas de síntese, mas também pode ser considerado para tratar qualquer fala gravada.

Abstract of Dissertation presented to UFSC as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering.

REDUCING THE NATURAL CLICK EFFECT WITHIN DATABASE FOR HIGH QUALITY CORPUS-BASED SPEECH SYNTHESIS

Monique Vitório Nicodem

January/2006

Advisor: Prof. Rui Seara, Dr.
Area of Concentration: Communications and Signal Processing.
Keywords: Click detection, speech enhancement, concatenative speech synthesis.
Number of Pages: 89 p.

ABSTRACT: This work studies certain degradations produced in an involuntary way by the human vocal tract. Such degradations here named natural clicks or involuntary clicks are perceived as small cracks. These natural clicks can degrade the quality of synthetic speech produced by concatenative speech synthesis systems specially when such cracks are associated with possible discontinuities generated in the concatenation process. The current research work intends to improve the quality of speech databases and as a consequence the synthetic speech quality. For such a task, techniques for suppression or attenuation of natural clicks existing in a speech corpus are presented. The reduction of click audible effects is obtained considering two processing steps: detection and treatment. Detection techniques based on inverse filtering, fourth-order derivative, human vocal tract modeling, wavelet decomposition, and artificial neural networks are here discussed. Processing techniques based on extrapolation, masking, smoothing, and pruning are also considered. The evaluation of the techniques for detection and even for treatment is made by taking as a reference a 45 minute long speech database recorded in an appropriate studio by a professional speaker. Such a database has been collected from the concatenative speech synthesis system developed at LINSE (Laboratory of Circuits and Signal Processing of the Electrical Engineering Department at UFSC). Perceptual experiments based on listening test are adopted aiming to evaluate the investigated techniques. Experimental results point out a considerable improvement of recorded speech quality. The work here developed has applications not only in speech databases aiming synthesis systems but it can also be considered to improve any recorded speech.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A fala consiste no principal meio de comunicação entre as pessoas [2], [3]. Esse meio de comunicação e seu mecanismo de produção despertam um evidente interesse na comunidade científica. A fala é produzida a partir da emissão de uma corrente de ar pelos pulmões. Essa corrente alcança as cordas vocais (pregas de ligamentos que se estendem pela laringe). As pregas vocais podem realizar movimentos de abertura ou fechamento. Os articuladores (mandíbula, língua, lábios, dentes e véu palatino) assumem um determinado posicionamento. O ar escoa pelas cavidades nasal e/ou oral. Essa configuração do aparelho fonador humano, ilustrada na Fig. 1.1, é a responsável pela geração dos sons da fala humana.

Na tentativa de reproduzir o mecanismo de produção da fala humana surgiram os primeiros dispositivos mecânicos capazes de produzir fala sintética. Os dispositivos evoluíram de mecânicos para elétricos. Com o desenvolvimento dos computadores, surgiram os sintetizadores baseados em algoritmos computacionais. Os mais recentes tipos de sintetizadores têm permitido a produção de fala a partir de um texto tomado como entrada. Esse processo é denominado conversão texto-fala (TTS - *text-to-speech*).

Os sistemas de conversão texto-fala transformam um texto de entrada em informações lingüísticas equivalentes a partir da realização de uma etapa de processamento lingüístico. A informação lingüística é convertida em fala considerando-se o uso de técnicas de processamento de sinais. Dentre as técnicas existentes (articulatória, por formantes, concatenativa), a concatenativa tem conduzido a uma fala sintética de melhor qualidade.

A síntese concatenativa é baseada na concatenação de segmentos de fala armazenados em um *corpus* (banco) previamente gravado. Dentre os fatores que exercem influência sobre o desempenho de um sistema de síntese, está a qualidade do *corpus*. Tal *corpus* é gravado por um locutor experiente em um estúdio profissional. Entretanto, a fala humana pode apresentar eventuais degradações causadas por variações abruptas de freqüência e amplitude de vibração das cordas vocais (*jitter* e *shimmer*, respectivamente) como também por cliques.

Cliques são eventuais descontinuidades existentes no sinal de fala produzidas pelo próprio aparelho fonador humano. Manifestam-se na forma de estalos de pequena intensi-

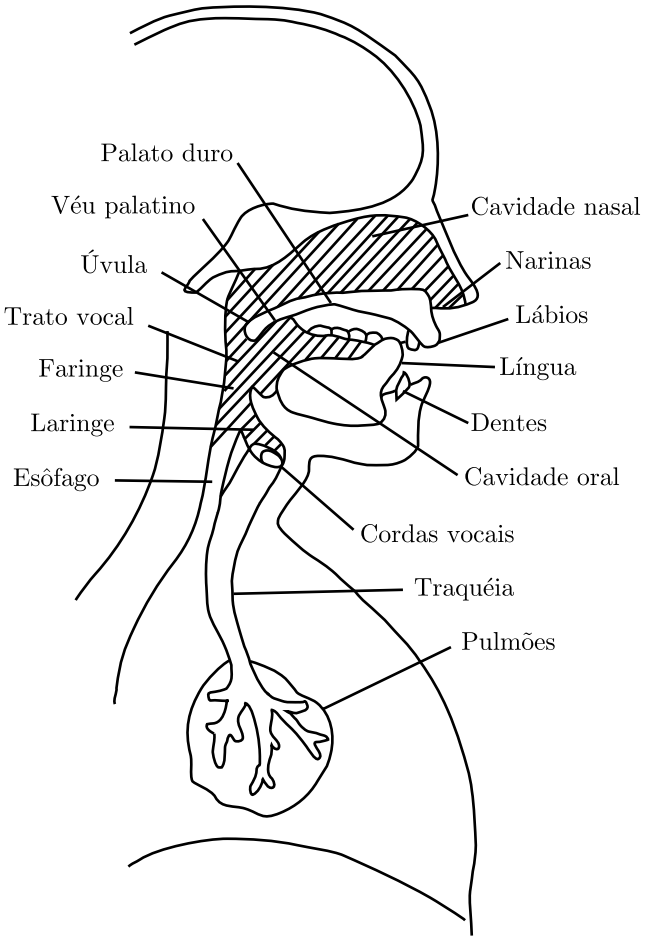

Fig. 1.1: Aparelho fonador humano.

dade, praticamente imperceptíveis na fala corrente. Porém, quando associados a distúrbios provenientes do processo de concatenação (descontinuidades), podem prejudicar significativamente a qualidade da fala sintética.

Para se obter uma melhoria de qualidade, a proposta do presente trabalho é eliminar ou reduzir o efeito audível de cliques existentes em *corpora* desenvolvidos para compor sistemas concatenativos de síntese de fala. Tal eliminação e/ou redução é realizada de maneira *off-line*, não afetando a complexidade computacional da síntese propriamente dita. Por outro lado, se a fala sintética fosse obtida através do tratamento de descontinuidades, nesse caso, a complexidade seria então afetada. A Fig. 1.2 apresenta um diagrama com o objetivo de contextualizar o tema cliques em síntese de fala.

É importante destacar que a redução dos cliques é aqui obtida considerando-se duas etapas: determinação da localização temporal dos cliques (detecção) e processamento (tratamento).

Para se determinar a localização temporal dos cliques, realiza-se uma etapa de detecção. Essa etapa pode ser efetuada de forma manual, o que certamente conduz aos melhores

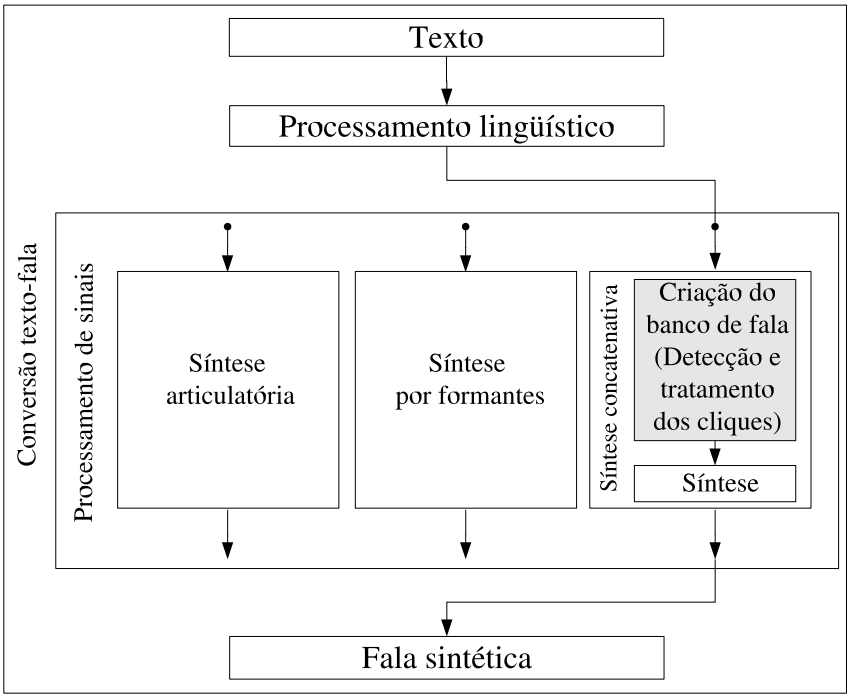

Fig. 1.2: Situando o tema Cliques em síntese de fala.

resultados. Entretanto, em bancos de fala de longa duração (da ordem de dezenas de horas), a detecção manual é inviável, pois demanda um tempo considerável (cerca de dias) para analisar poucos minutos de gravação. Além disso, tal procedimento é exaustivo, inclusive para pessoas treinadas.

O presente trabalho descreve possíveis técnicas de detecção automática de cliques e estabelece uma comparação entre as técnicas. É importante mencionar que, em nosso conhecimento, técnicas para detecção automática de cliques naturais não são sequer apresentadas na literatura da área.

Após a etapa de detecção automática, os cliques devem ser processados (tratados). Para tal, o trabalho em questão apresenta possíveis técnicas de tratamento, realizando comparativos entre as técnicas. Essa comparação é realizada considerando-se testes baseados em escuta. Dessa forma, a avaliação da qualidade da fala está intimamente relacionada à percepção de qualidade sonora e aos fatores que tornam a fala sob certos aspectos mais "agradável". Assim, é importante abordar, nesta seção, conceitos relevantes de percepção acústica, como audibilidade e mascaramento.

1.1 Audibilidade

A audibilidade relaciona-se à percepção da intensidade sonora e dentre os fatores que a influenciam, os mais evidentes são a pressão sonora e a freqüência do som [4], [5].

Na década de 30, dois pesquisadores dos Laboratórios Bell, Harvey Fletcher e Wilden Munson, realizaram experimentos relacionados à percepção da intensidade sonora. Ouvintes escutavam um tom puro de freqüência de 1 kHz e ajustavam um segundo tom para um nível sonoro até que os dois tons fossem percebidos como iguais em intensidade. Os resultados dos experimentos de Fletcher e Munson são resumidos em suas curvas com nível de audibilidade constante (desde o limiar de audição de 0 dB ao limiar da dor de 120 dB, ambos na freqüência de 1 kHz) apresentadas na Fig. 1.3 [6]. Como exemplo de leitura das curvas, tem-se que, enquanto é necessária uma intensidade sonora de 10^{-12} W/m^2 para atingir o Limiar de Audição a 1 kHz, a 90 Hz é necessária uma intensidade de 10^{-8} W/m^2, ou seja, um nível sonoro 40 dB acima da referência [4].

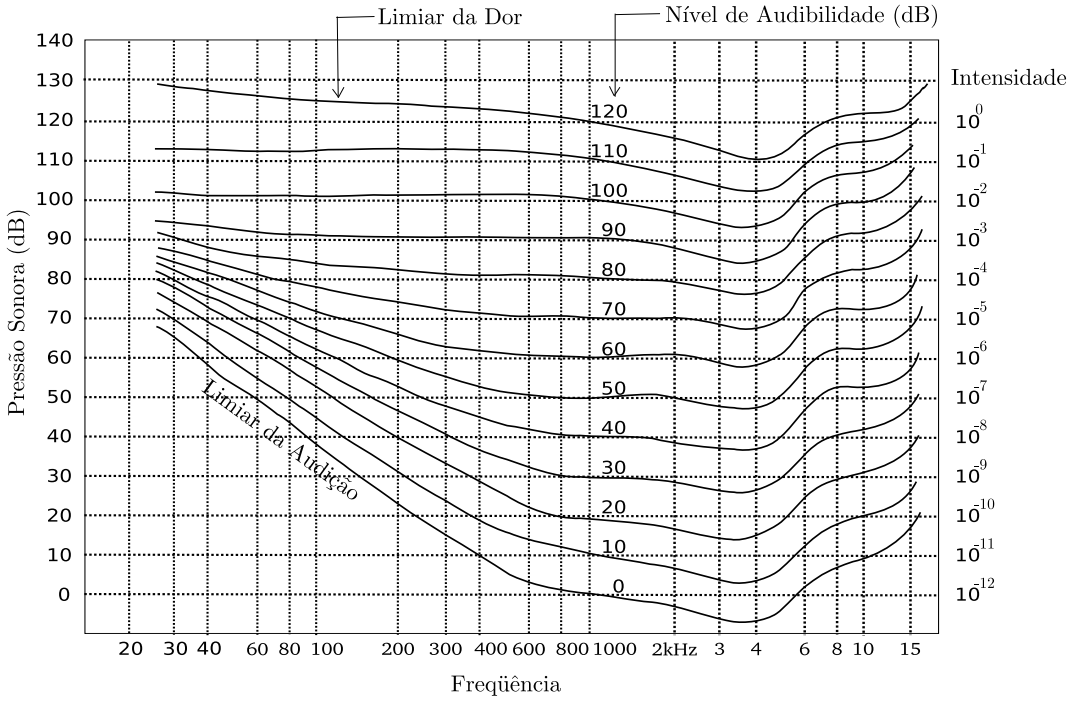

Fig. 1.3: Diagrama de Audibilidade de Fletcher-Munson.

A partir das curvas de Fletcher-Munson, conclui-se que a sensibilidade auditiva humana assume menores valores em baixas e altas freqüências, sendo máxima entre 3 kHz e 4 kHz.

Em 1956, D. W. Robinson e R. S. Dadson melhoraram as curvas de Fletcher e Munson, realizando experimentos similares em câmaras anecóicas. As curvas de igual intensidade de Robinson e Dadson são atualmente padronizadas pela *International Standards Organization* (ISO). Essas curvas, apresentadas na Fig. 1.4, são freqüentemente referenciadas como curvas de Fletcher-Munson [6].

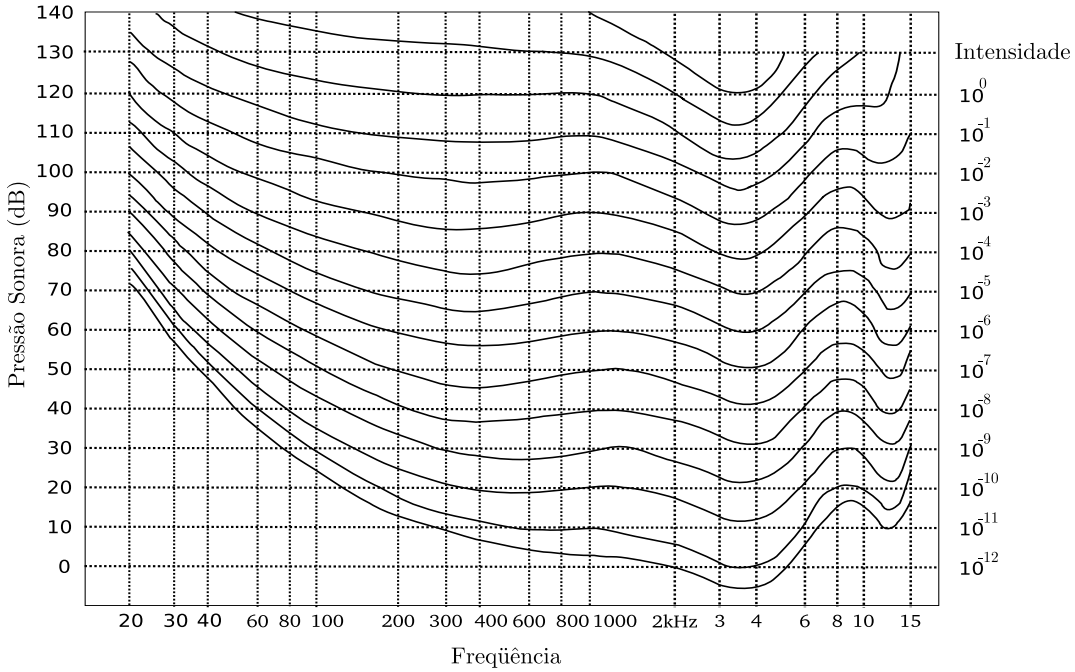

Fig. 1.4: Curvas de Robinson e Dadson.

1.2 Mascaramento

Em diversas situações do cotidiano, é possível perceber sons tornando-se menos audíveis ou inclusive inaudíveis na presença de outros sons mascaradores. Como exemplo, tem-se a música proveniente do rádio de um carro que em alguns casos mascara o ruído proveniente do motor.

A nomenclatura mascaramento foi definida pela *American Standards Association* como segue:

a) processo no qual o limiar de audibilidade de um som é aumentado pela presença de outro som (mascarador);

b) quantidade de aumento do limiar de audibilidade de um som na presença de outro mascarador. A unidade freqüentemente usada é decibel (dB) [7].

É importante mencionar que o mascaramento não ocorre em casos de coincidência temporal entre sinais mascarador e mascarado (mascaramento simultâneo). O mascaramento pode também ocorrer com um sinal mascarador localizado ligeiramente antes ou depois do sinal mascarado. Esse tipo de mascaramento é denominado não-simultâneo [7].

As primeiras experiências com o fenômeno de mascaramento foram publicadas por Wegel e Lane, em 1924 [4], [7]. Esses pesquisadores determinaram o limiar para detecção de um tom puro com freqüência ajustável na presença de um tom puro mascarador com

freqüência e intensidade fixas. O gráfico do limiar de mascaramento em função da freqüência do sinal é conhecido como padrão de mascaramento ou audiograma de mascaramento. Outros experimentos sobre mascaramento foram posteriormente realizados utilizando um tom puro e um ruído de banda estreita. Tal ruído representa ora o sinal mascarador ora o sinal mascarado [4], [7].

A Fig. 1.5 mostra uma adaptação dos padrões de mascaramento obtidos por Egan e Hake, em 1950. Nesse caso, o sinal mascarador é um ruído de banda estreita (90 Hz) centrado em 410 Hz e o sinal mascarado é um tom puro com freqüência variando aproximadamente entre 100 Hz e 5000 Hz [4].

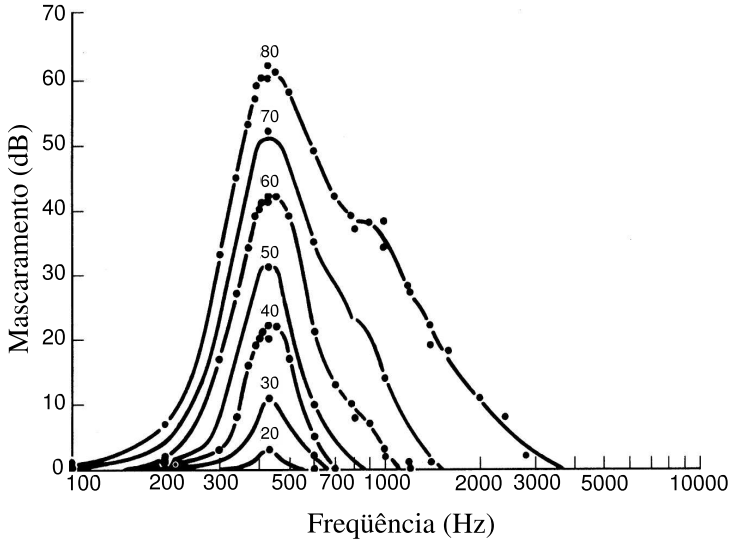

Fig. 1.5: Padrões de mascaramento.

A observação da Fig. 1.5 permite concluir que os maiores níveis de mascaramento ocorrem quando a freqüência do tom coincide com a freqüência central da banda do ruído mascarador (410 Hz).

1.3 Objetivos do Trabalho

O presente trabalho tem como objetivo geral a apresentação de uma técnica automática de supressão e/ou atenuação de cliques naturais presentes em bancos de fala integrantes de sistemas de síntese concatenativa de fala. Para alcançar tal intento, esse trabalho tem como objetivos específicos:

a) a apresentação de técnicas já mencionadas pela literatura, utilizadas em outras aplicações, que permitam a detecção e o tratamento de cliques naturais presentes em bancos de fala;

b) o desenvolvimento de outras técnicas de detecção e tratamento mais adequadas para a aplicação considerada.

A utilização de técnicas para redução do número de cliques naturais em bancos de fala proporciona uma melhoria da qualidade da fala sintetizada. É desejável que as técnicas de detecção apresentadas localizem o maior número possível de cliques naturais (verdadeiros positivos), podendo inclusive indicar erroneamente a existência de cliques (falsos positivos). É evidente que tais indicações errôneas são toleradas até o ponto em que a técnica de tratamento aplicada sobre falsos cliques não prejudique a qualidade da fala gravada. É também desejável que não haja a inserção de degradações perceptíveis no processo de tratamento.

1.4 Organização da Dissertação

O presente trabalho mostra a seguinte estrutura de apresentação. O Capítulo 2 apresenta um breve histórico dos processos de síntese de fala, as principais aplicações de conversores texto-fala, o princípio de funcionamento desses conversores, noções de síntese concatenativa incluindo as etapas envolvidas na concepção de um *corpus* de fala.

O Capítulo 3 apresenta a descrição e a modelagem dos cliques naturais produzidos pelo aparelho fonador humano. Esse capítulo se baseia em uma análise obtida de forma manual com o objetivo de determinar, por um procedimento de escuta, a localização temporal dos cliques presentes em um banco de fala com 45 minutos de duração.

As técnicas propostas para detecção automática de cliques são descritas no Capítulo 4. Algumas abordagens consideradas são inspiradas na detecção de ruídos impulsivos em gravações antigas de áudio.

No Capítulo 5, as técnicas de detecção automática são avaliadas quanto ao desempenho. Tal avaliação é baseada em duas figuras de mérito propostas no presente trabalho.

O Capítulo 6 descreve alguns possíveis métodos para reduzir e/ou suprimir os efeitos audíveis causados por cliques naturais. Dentre os métodos descritos, podem-se enumerar, por exemplo, métodos baseados em conceitos de mascaramento ou extrapolação.

Uma avaliação comparativa dos métodos de tratamento de cliques naturais é apresentada no Capítulo 7. Tal comparação serve como suporte no sentido de se verificar a metodologia mais adequada para supressão dos cliques naturais.

1.5 Contribuições Originais

É importante mencionar as contribuições originais do trabalho aqui desenvolvido. Em nosso conhecimento, o assunto “cliques naturais” não tem sido apresentado na literatura tanto das áreas de fonética e fonologia quanto da área de processamento de sinais de fala.

Em função da originalidade do tema, a filosofia de adotar ferramentas de detecção automática e de tratamento de cliques naturais também consiste em uma característica inovadora do presente trabalho.

Deve-se enfatizar também que, apesar deste trabalho estar voltado para uma aplicação em síntese de fala, o tratamento de cliques naturais pode ser aplicado sobre qualquer fala gravada.

Capítulo 2

Síntese de Fala

Um sistema de síntese de fala consiste em um mecanismo utilizado para produzir artificialmente fala sintética similar à humana. Tal mecanismo pode adotar como entrada um conjunto de informações. Essas informações podem consistir em representações lingüísticas (transcrição fonética, por exemplo), ou ainda, em uma mensagem na forma textual. Nos casos em que a informação de entrada é fornecida nos moldes de um texto, tais sistemas são chamados conversores texto-fala (TTS – *text-to-speech*).

2.1 Histórico e Desenvolvimento da Síntese de Fala

A possibilidade de converter diretamente texto em fala só é atualmente possível em função de todo um desenvolvimento científico realizado com o decorrer dos anos. Tal desenvolvimento é apresentado de maneira sucinta nesta seção, baseando-se nas referências [2], [3], [8]–[13].

2.1.1 Síntese com Equipamentos Mecânicos

Os primeiros esforços para produzir fala sintética foram realizados pelo professor Christian Gottlieb Kratzenstein. Em 1779, Kratzenstein apresentou seus experimentos com ressoadores acústicos. Em seus experimentos, cada ressoador imitava a configuração assumida pelo trato vocal humano na produção de uma vogal ([a], [e], [i], [o], [u]). Além disso, cada ressoador dispunha de uma palheta vibrante com a função de interromper a corrente de ar de maneira similar às cordas vocais. A Fig. 2.1 ilustra a estrutura básica dos ressoadores criados por Kratzenstein.

Doze anos após a introdução dos ressoadores de Kratzenstein, Wolfgang Ritter von Kempelen divulgou sua máquina falante. Essa máquina, apresentada através da publicação do livro: “O Mecanismo da Fala Humana e a Construção de uma Máquina Falante”, foi resultado de mais de 20 anos de pesquisa. Ela possibilitava a produção de palavras inteiras e

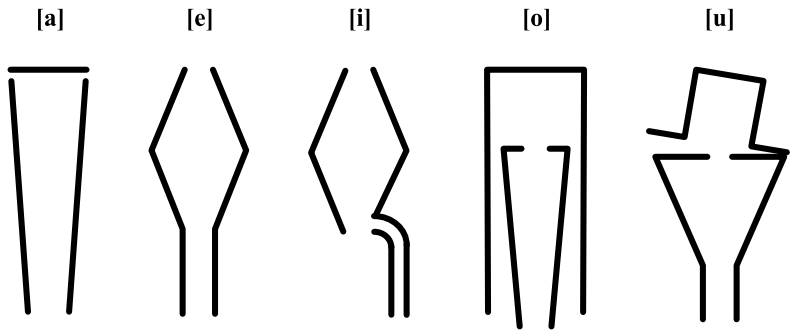

Fig. 2.1: Ressoadores de Christian Gottlieb Kratzenstein.

de pequenas sentenças. Continha um fole, uma palheta vibrante e um tubo de couro (ressoador), que simulavam, respectivamente, o pulmão, as cordas vocais e o trato vocal humanos. A Fig. 2.2 ilustra, através de um desenho esquemático, a máquina falante desenvolvida por von Kempelen.

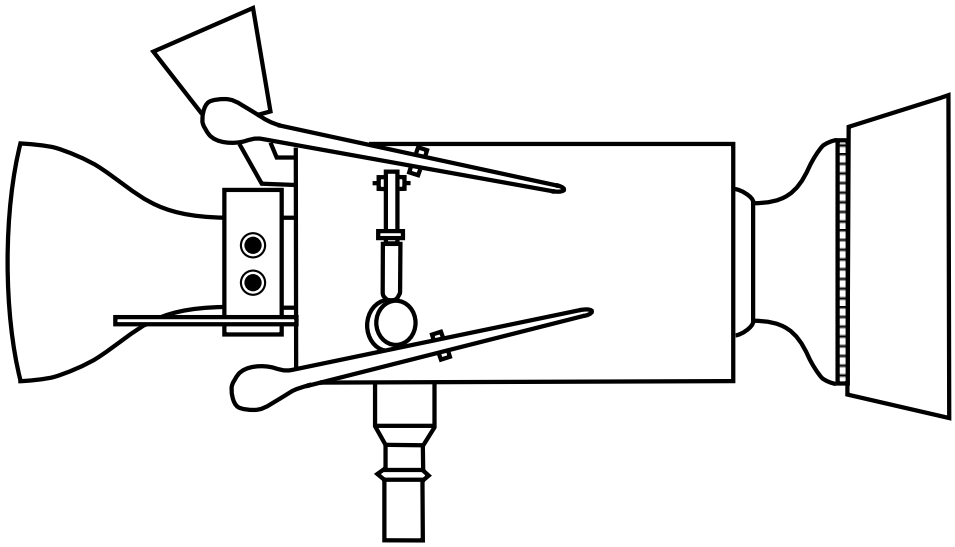

Fig. 2.2: Máquina falante desenvolvida por von Kempelen.

O trabalho de von Kempelen não obteve muito crédito na época devido a um acontecimento que marcou negativamente seu criador. O pesquisador, enquanto trabalhava em sua máquina falante, construiu uma máquina automática de jogar xadrez para a imperatriz Maria Theresa. Entretanto, descobriu-se que a máquina apresentava, na realidade, um jogador de xadrez experiente em seu interior.

Baseado nas especificações de von Kempelen, Charles Wheatstone construiu uma versão melhorada de tal máquina. Essa versão foi apresentada, em 1835, no encontro Dublin da Associação Britânica para o Avanço da Ciência. A máquina tinha um fole, para simular um pulmão, uma palheta vibrante, representando as cordas vocais, e um tubo de couro flexível (similar ao trato vocal), cuja área de seção transversal poderia variar produzindo diferentes sons vozeados. A máquina de Wheatstone possibilitava produzir vogais, a maioria

das consoantes, palavras e até algumas pequenas sentenças. A Fig. 2.3 apresenta um desenho esquemático da máquina falante produzida por Wheatstone.

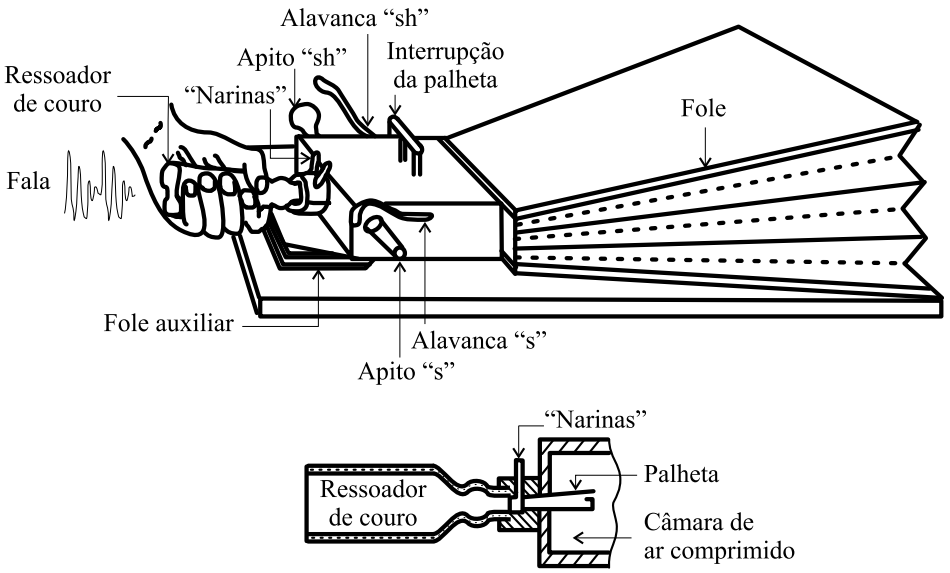

Fig. 2.3: Máquina falante de Charles Wheatstone.

Joseph Faber exibiu em Londres, em 1846, outra máquina falante, esta chamada "Euphonia". Tal máquina incluía uma modelagem da cavidade faringeal e da língua. Com esse incremento científico, era possível não apenas falar como também cantar. Faber demonstrou a habilidade de canto em sua máquina com a música "*God Save the Queen*". A Fig. 2.4 mostra a máquina falante de Faber.

Fig. 2.4: Máquina falante "Euphonia" de Joseph Faber [10].

Outros experimentos para produção de fala sintética utilizando equipamentos mecânicos e semi-eletrônicos foram realizados até a década de 60, porém não houve notáveis

avanços científicos.

2.1.2 Síntese com Equipamentos Eletrônicos

A síntese completamente eletrônica foi iniciada com um equipamento criado por J. Q. Stewart, em 1922. Nesse equipamento, dois circuitos ressonantes, excitados por uma cigarra elétrica, modelavam as duas freqüências de ressonância (formantes) mais baixas do trato vocal. Tal modelagem permitia sintetizar apenas sons vocálicos, não gerando qualquer consoante.

Em 1939, um sintetizador capaz de gerar sons vocálicos e consonantais foi apresentado na Feira Mundial de Nova York. O engenheiro eletricista Homer Dudley apresentou o VODER (Voice Operating Demonstrator), desenvolvido nos Laboratórios Bell. O VODER apresentava 14 chaves que controlavam a estrutura de ressonância do trato vocal, uma barra que selecionava entre uma fonte vozeada ou de ruído e um pedal para o pé direito que permitia variar a freqüência fundamental (*pitch*) dos sons vozeados. O VODER também apresentava uma caixa de controle de ressonância composta por dez filtros passa-faixa cujos níveis de amplitude eram controlados por operadores treinados. Embora o treinamento de um operador fosse longo (da ordem de um ano), técnicos altamente treinados conseguiam produzir fala inteligível. A Fig. 2.5 ilustra o princípio de funcionamento do VODER.

H. K. Dunn apresentou, em 1950, um sistema de síntese cuja qualidade superava a do VODER. O "*Electrical Vocal Tract*", mostrado na Fig. 2.6, era composto por uma fonte de energia vibratória, para simular as cordas vocais, um modelo de linha de transmissão (indutores e capacitores), para modelar o trato vocal, e filtros passa-baixa que forneciam o atraso correspondente ao de uma onda sonora atravessando o trato vocal humano.

Em 1951, nos Laboratórios Haskins, Franklin S. Cooper e seus colegas desenvolveram o sintetizador "*Pattern Playback*". Esse equipamento realizava a função inversa de um espectrógrafo, permitindo converter os padrões de um espectrograma em fala.

Outras pesquisas nas áreas de síntese e modelagem da fala foram realizadas com o desenvolvimento dos sistemas computacionais. Com a popularização dos computadores, atualmente a maioria dos sistemas de síntese utilizam técnicas essencialmente computacionais e convertem não espectrogramas, mas textos em fala. Alguns sistemas (*softwares*) de conversão texto-fala estão inclusive disponíveis para *download* na Internet, tais como: Festival, Flite, FreeTTS, MBROLA. Outros sistemas apresentam apenas demonstrativos disponíveis, dentre os quais pode-se enumerar: Rhetorical rVoice, Loquendo TTS, AT&T Natural Voices, Microsoft Mandarin Chinese TTS e IBM TTS.

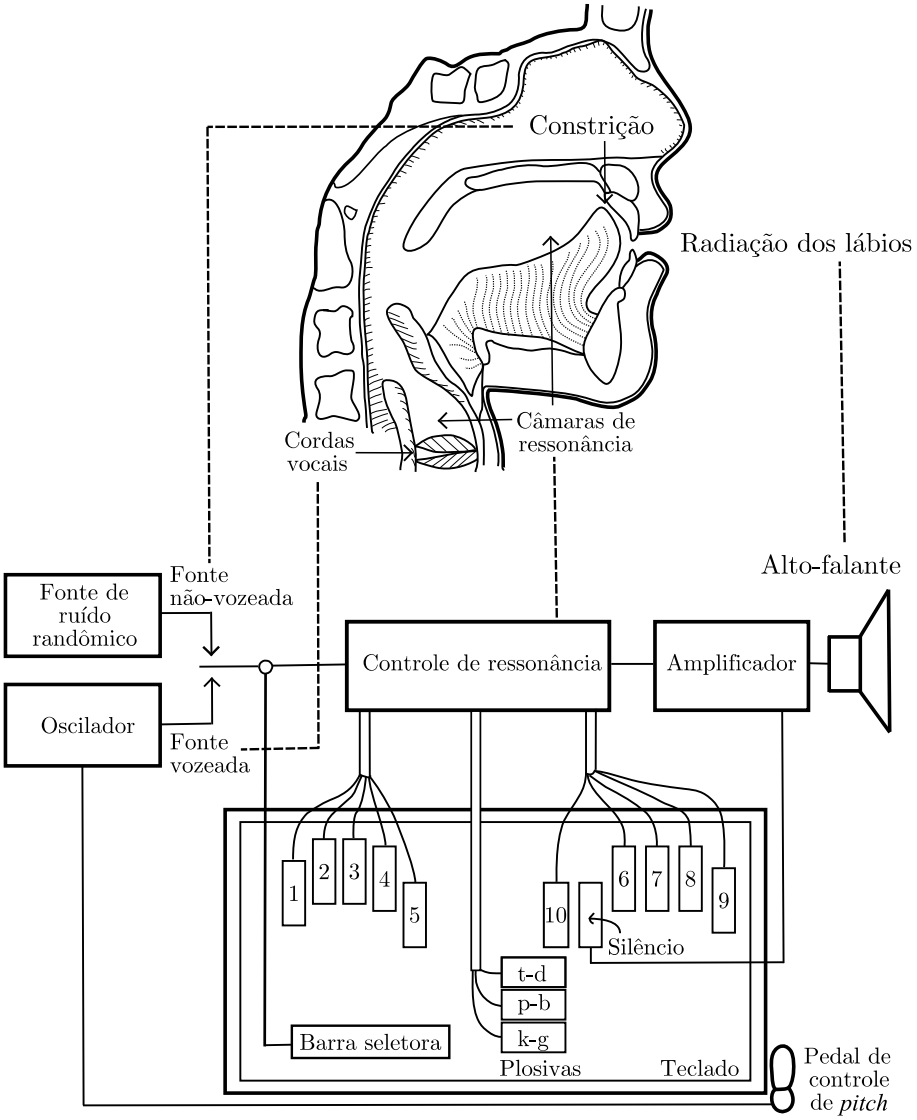

Fig. 2.5: Princípio de funcionamento do VODER.

2.2 Aplicações de um Conversor Texto-Fala

O desenvolvimento de ferramentas computacionais, a melhoria da qualidade e a popularização dos sistemas de conversão texto-fala possibilitaram o aumento do número de aplicações. Em algumas delas, esses sistemas são indispensáveis. Em outras, a fala sintética apenas substitui tecnologias mais simples, oferecendo vantagens significativas. Dentre as situações nas quais a fala sintética é preferida em relação a soluções mais simples (como, por exemplo, a fala pré-gravada), podem-se elencar [14]:

a) em textos imprevisíveis e dinâmicos. Um sistema TTS é vantajoso para sintetizar fala utilizando como texto de entrada mensagens de curta duração e com conteúdo de informação significativamente variável;

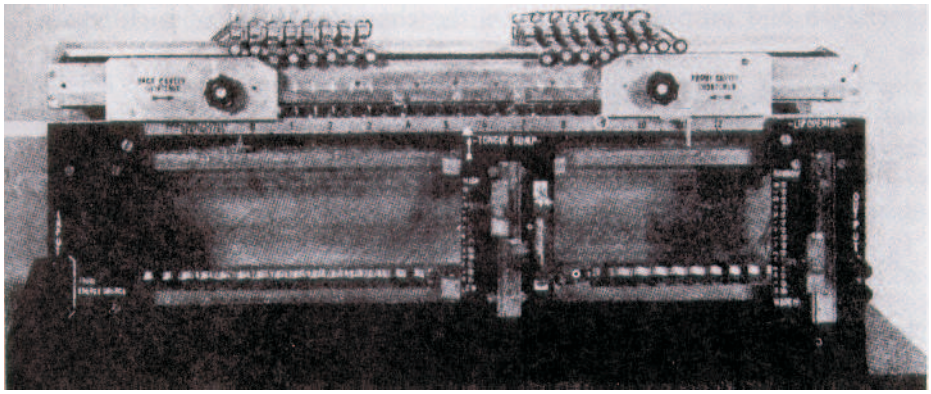

Fig. 2.6: “Electrical Vocal Tract” de H. K. Dunn.

b) para bancos de dados de longa duração. Em grandes bancos de dados, não é adequado armazenar informação na forma de fala gravada em função dos altos custos de gravação e armazenamento. Grandes bancos de dados são raramente estáveis, favorecendo a utilização da tecnologia texto-fala;

c) para saída relativamente estável, com custos críticos de provimento e tempo de resposta. Um exemplo dessa situação é um sistema de anúncios telefônicos no qual a maioria deles permanecem inalterados, mas podem surgir necessidades de novas mensagens que mantenham a mesma voz. Dessa maneira, a síntese de fala, quando comparada com gravações que levam um tempo significativo para serem prepararadas, simplifica consideravelmente a geração de mensagens fixas para novos serviços;

d) quando a consistência da voz é desejada. Muitos serviços requerem a mesma voz para todos os anúncios. Nesse caso, a voz sintética é preferida, pois o anúncio pode ser alterado a qualquer momento sem depender da disponibilidade do locutor;

e) quando se deseja estreita largura de banda. Utiliza-se uma estreita largura de banda com a transmissão na forma de texto e a conversão para fala no receptor.

Alguns exemplos de possíveis aplicações dos conversores texto-fala são [8], [15]–[19]:

a) Auxílio a deficientes. Os conversores texto-fala podem auxiliar a aprendizagem da fala de deficientes auditivos submetidos a implante coclear, permitem o acesso de deficientes visuais a informações textuais e possibilitam a comunicação falada a deficientes vocais.

b) Educação lingüística. O aprendizado da pronúncia de palavras pertencentes ao vocabulário de idiomas sob estudo é facilitado com o uso de um sistema TTS.

c) Livros e brinquedos falantes. O mercado de livros e brinquedos está adotando a cada dia mais as ferramentas de síntese de fala em virtude da atratividade oferecida por tais sistemas.

d) Monitoração vocal. Em alguns casos, informações orais são mais eficientes do que escritas. A idéia é utilizar sintetizadores de fala em sistemas de controle e medidas como também para alertar sobre perigos iminentes.

e) Comunicação homem-máquina. O desenvolvimento de sistemas TTS de alta qualidade é um passo necessário (bem como o melhoramento de reconhecedores de fala) para obter meios de comunicação mais completos entre homens e máquinas. Um exemplo dessa comunicação é o uso da síntese aliada ao reconhecimento (diálogo) para realizar o preeenchimento automático de formulários.

f) Possibilidade de comunicação homem-veículo. Um veículo poderia, por exemplo, informar excesso de velocidade, alta temperatura do motor, iminência de contato com outro veículo, rota mais segura e econômica entre lugares.

g) Pesquisa fundamental e aplicada. Um sistema TTS é uma ferramenta multidisciplinar extremamente útil para auxiliar no desenvolvimento de experimentos por lingüistas, foneticistas, fisiologistas, psicólogos, engenheiros e analistas de sistemas.

h) Sistemas de tradução. Tais sistemas utilizam as tecnologias de síntese e reconhecimento para tradução tendo como entrada e saída a fala em diferentes idiomas. Um exemplo de conversor fala-fala é o DARPA-Babylon.

i) Serviços de telecomunicações. Dentre as diversas aplicacões na área de telecomunições, podem-se enumerar:

   - solicitação de pedidos;
   - reservas e informações sobre vôos e hotéis;
   - leitura da agenda de compromissos, *e-mail*, fax e conteúdo da *internet*;
   - solicitação de informações como resultados de eventos esportivos, hora local ou clima (temperatura, direção e velocidade do vento, pressão barométrica);
   - acesso a serviços de comunicação e entretenimento;
   - para sistemas automáticos de interceptação (AIS - *Automatic Intercept System*), nos quais uma ligação telefônica realizada para número inativo (ou fora de serviço) é interceptada e uma mensagem apropriada é fornecida ao usuário;
   - consulta de informações bancárias (saldo, cotações, seguros).

2.3 Visão Geral dos Sistemas de Conversão Texto-Fala

Os sistemas de conversão texto-fala são adotados nas aplicações que realizam a transformação automática de qualquer texto em fala. Tais sistemas apresentam essencialmente dois módulos principais: processamento lingüístico e processamento de sinais [15].

O módulo de processamento lingüístico é responsável por dar ao texto de entrada uma representação lingüística detalhada [3]. Tal módulo usualmente envolve as etapas de pré-processamento do texto de entrada (com a separação em blocos de análise, identificação e expansão de algarismos arábicos e romanos, abreviaturas, siglas, dentre outros), análise morfossintática, separação silábica, determinação da tonicidade, transcrição ortográfico-fonética e modelagem prosódica. Esse módulo depende fortemente do idioma a que se propõe o sistema de conversão.

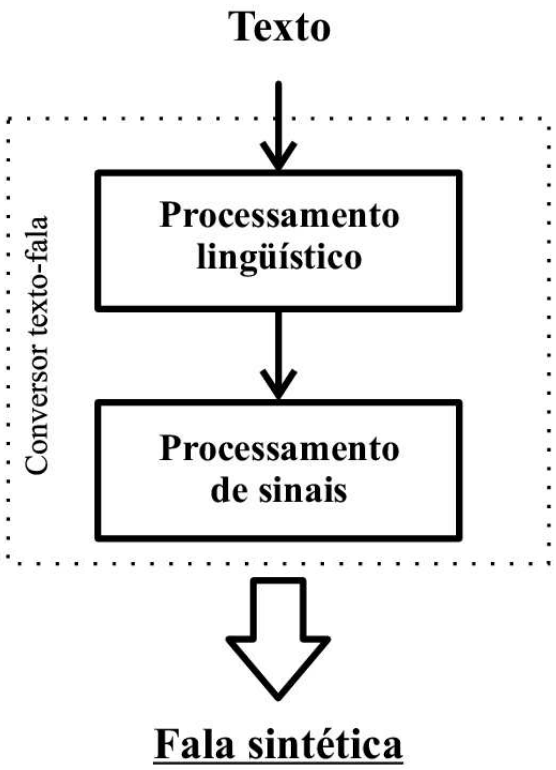

Fig. 2.7: Etapas de um sistema de conversão texto-fala [20].

A etapa de processamento de sinais, por sua vez usualmente independente do idioma, transforma a informação lingüística obtida no processamento anterior em fala sintetizada. De forma geral, essa etapa pode ser implementada por três diferentes métodos: articulatório, por formantes e concatenativo.

O método articulatório realiza uma espécie de modelagem do aparelho fonador humano, obtendo resultados satisfatórios na síntese. Entretanto, constitui-se em um dos métodos mais difíceis de implementação e com maior esforço computacional [3]. Devido ao alto

esforço computacional requerido, não tem alcançado ainda tanto sucesso quanto os outros métodos de síntese de fala.

A síntese por formantes, ou paramétrica, baseia-se no modelo fonte-filtro de produção da fala, ilustrado na Fig. 2.8. Sendo assim, a síntese é realizada através da combinação entre a excitação (vozeada ou não) e o filtro que modela o trato vocal [3]. Por muito tempo, essa técnica de síntese foi dominante (décadas de 70 e 80).

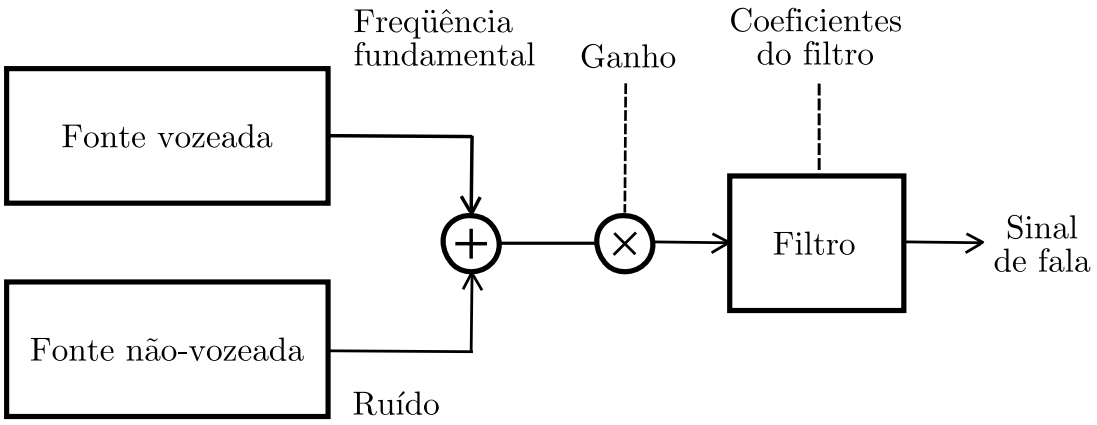

Fig. 2.8: Modelo fonte-filtro de produção de fala [3].

Atualmente o método concatenativo é o mais empregado nos sistemas de conversão texto-fala [21]. Essa popularização da técnica concatenativa de síntese deve-se tanto à sua relativa simplicidade de implementação quanto à produção de fala sintética de melhor qualidade comparativamente às demais técnicas. O presente trabalho aplica-se a tal técnica visto que seu desempenho é influenciado pela presença ou ausência de cliques naturais.

2.4 Síntese Concatenativa de Fala

Nos sistemas de síntese concatenativa, a fala sintética é obtida através do encadeamento (concatenação) de segmentos de fala selecionados em um banco de unidades previamente gravado. A qualidade da fala produzida por esses sistemas depende fortemente da qualidade do banco, dos procedimentos de seleção de segmentos e da técnica adotada para concatenação.

2.4.1 Criação de um Banco de Unidades

A primeira etapa de um procedimento de síntese concatenativa consiste na criação de um banco de unidades. Para tal, é necessário definir quais unidades utilizar durante a

síntese propriamente dita: fone, difone, demissílaba, trifone[1], ou até uma palavra inteira. Essa definição deve satisfazer uma certa relação de compromisso entre tamanho do *corpus* e reprodução dos efeitos de coarticulação[2]. O uso de unidades longas (tal como palavra) implica em reproduzir melhor os efeitos de coarticulação entre fonemas. Entretanto, quanto maior a unidade, maior deverá ser o tamanho do *corpus*.

Esta relação de compromisso entre tamanho do *corpus* e reprodução dos efeitos de coarticulação é mais adequadamente satisfeita nos sistemas TTS mais recentes que utilizam unidades de comprimento variável, selecionando ora fone, ora difone, ora uma palavra inteira [23]. Dessa maneira, unidades longas (especialmente palavras comumente utilizadas na fala corrente) podem ser selecionadas na síntese se estiverem presentes no corpus e apresentarem as características prosódicas desejadas.

Após a escolha das unidades, há a definição do texto a ser gravado. O texto deve ser tal que o *corpus* de fala resultante da gravação englobe a maior parte dos contextos fonéticos (à direita e à esquerda) e variações prosódicas da língua considerada. Nesse sentido, o uso de grandes *corpora* pode ser útil.

Consideram-se grandes *corpora* os bancos de fala com duração aproximada de 10 horas ou mais [23]. Um caso extremo é um banco constituinte do sistema TTS Ximera, em desenvolvimento na ATR, que apresenta 110 horas de duração e foi gravado no idioma japonês por um locutor do sexo masculino [24].

O *corpus* de fala geralmente é gravado por um único locutor, seja homem ou mulher. Esse *corpus* pode ser um banco de dados já existente que tenha sido desenvolvido para uso em outras aplicações, como reconhecimento de fala, ou pode ser desenvolvido e gravado especialmente para uma aplicação de síntese de fala [25]. É importante mencionar que a gravação do banco deve ser realizada em ambiente controlado, sem ruído, com um locutor experiente, de maneira a manter a mesma voz e as mesmas configurações de gravação por todo o banco.

A etapa que sucede a gravação do banco consiste em segmentar e especificar a localização temporal de cada unidade de fala. Isso é o que se chama rotulagem ou anotação fonética. Além da anotação fonética, características prosódicas podem também ser utilizadas para diferenciar os segmentos de fala.

O *corpus* de fala pode ser ou não codificado, sendo posteriormente armazenado de tal forma que possa ser acessado durante a síntese. Há várias formas de armazenamento: formas de onda, parâmetros resultantes da análise da fala como coeficientes LPC (*linear*

[1]Em fonética e fonologia, a unidade fone consiste no menor segmento sonoro distintivo da fala. O difone é o segmento que se inicia na metade de um fonema se prolongando até a metade do fonema seguinte. Demissílaba corresponde à metade inicial ou final de uma sílaba. O trifone, por sua vez, pode ser considerado como um fonema com um contexto específico anterior e posterior [8].

[2]Por coarticulação entende-se a coordenação de diversos movimentos articulatórios para a realização de uma mesma unidade fônica. Pode-se citar como exemplo de coarticulação a nasalização de vogais próximas a consoantes nasais [22].

predictive coding), dentre outros.

Um diagrama geral das etapas para a concepção de um banco de unidades é mostrado na Fig. 2.9.

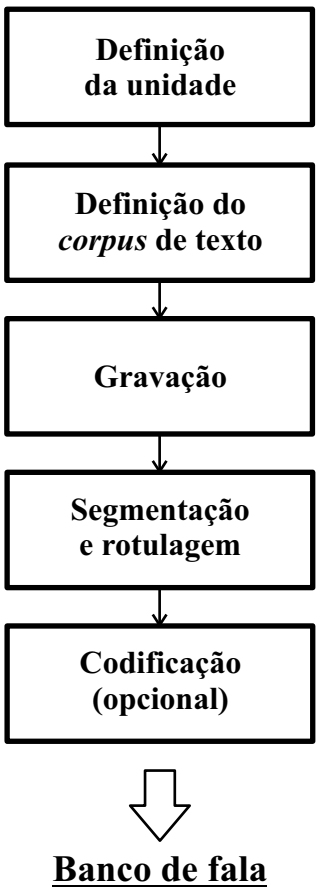

Fig. 2.9: Etapas para a obtenção de um banco de unidades.

2.4.2 Síntese do Sinal de Fala Usando Técnica Concatenativa

Durante a síntese propriamente dita do sinal de fala, cujo diagrama simplificado é apresentado na Fig. 2.10, há a seleção das melhores unidades (em termos de identidade fonética, contexto fonético e características prosódicas) existentes no banco. Para essa etapa, ferramentas de seleção automática de unidades vêm sendo comumente adotadas [21], [26].

A seleção automática de unidades constitui-se em um procedimento aplicado em grandes *corpora* com a finalidade de se obter a seqüência ótima de segmentos que atenda aos requisitos de prosódia desejada e minimize as descontinuidades de concatenação.

Para se encontrar um conjunto de unidades que minimize as descontinuidades espectrais, utiliza-se uma função objetivo. Essa função serve como uma medida numérica que avalia a qualidade da concatenação entre segmentos de fala. A avaliação da qualidade é re-

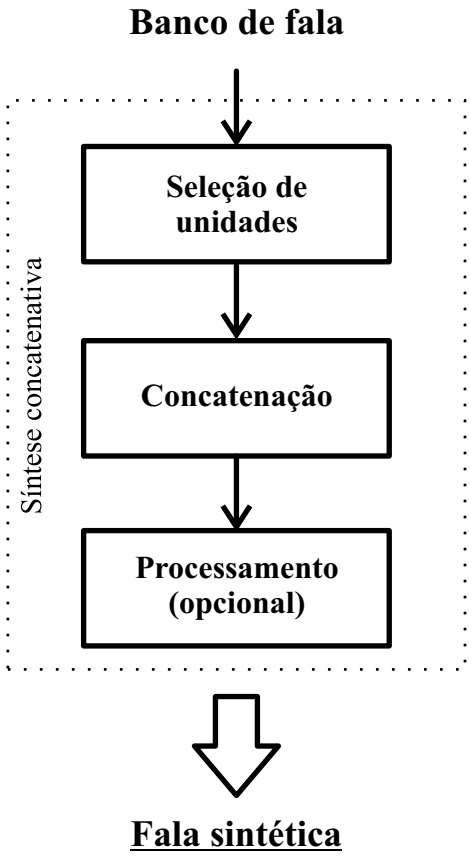

Fig. 2.10: Síntese de fala usando técnica concatenativa.

alizada através da utilização dos conceitos de custo alvo e custo de concatenação entre duas unidades [27].

O custo alvo $C^t(t_i,u_i)$ representa uma estimativa da diferença entre o segmento alvo t_i e a unidade existente no banco de dados u_i. Já o custo de concatenação representa uma estimativa da qualidade da união entre duas unidades consecutivas (u_{i-1} e u_i), conforme ilustrado na Fig. 2.11 [26].

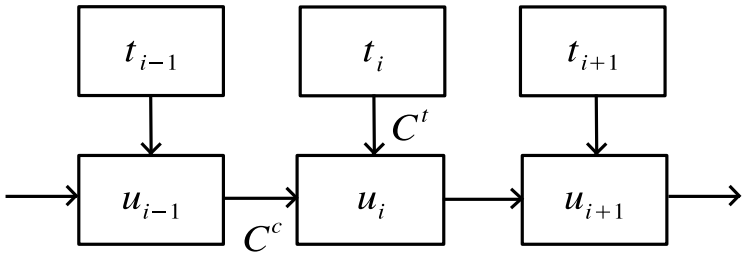

Fig. 2.11: Custos alvo e de concatenação.

A função objetivo, utilizada para obter a seqüência ótima de unidades, apresenta algumas diferenças entre os sistemas de síntese de fala, mas, em essência, corresponde a

uma soma (ponderada ou não) dos custos alvo e de concatenação.

Dada uma seqüência alvo, $t_1^n = (t_1, ..., t_n)$, realiza-se a seleção de um conjunto de unidades, $u_1^n = (u_1, ..., u_n)$, que mais se aproxime do alvo.

O custo alvo $C^t(t_i,u_i)$ corresponde a uma soma ponderada das diferenças entre os elementos dos vetores (contendo parâmetros resultantes de análise LPC, energia, *pitch*, dentre outros) que caracterizam as unidades alvo e candidata. As diferenças representam os p subcustos alvo $C_j^t(t_i,u_i)$, $j = 1,...,p$. Para cada subcusto alvo, um peso correspondente a w_j^t é associado. Dessa maneira, o custo alvo pode ser obtido através de (2.1). Assim,

$$C^t(t_i,u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i,u_i). \tag{2.1}$$

O custo de concatenação $C^c(u_{i-1},u_i)$ corresponde a uma soma ponderada das diferenças entre os elementos dos vetores que caracterizam as unidades anterior e corrente. As diferenças representam, nesse caso, os q subcustos de concatenação $C_j(u_{i-1},u_i)$, $j = 1,...,q$. Para cada subcusto de concatenação, um peso w_j^c é associado. Assim,

$$C^c(u_{i-1},u_i) = \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1},u_i). \tag{2.2}$$

O custo total para sintetizar uma seqüência composta por n unidades é dado por (2.3).

$$C(t_1^n,u_1^n) = \sum_{i=1}^{n} C^t(t_i,u_i) + \sum_{i=2}^{n} C^c(u_{i-1},u_i) \tag{2.3}$$

O custo total, dado por (2.3), representa a função a ser minimizada objetivando selecionar a seqüência de unidades a ser utilizada na síntese.

A etapa seguinte à minimização dos custos (seleção automática) consiste em concatenar segmentos individuais em um conjunto contínuo. Após a etapa de concatenação de unidades, pode ocorrer um processamento visando alterar parâmetros prosódicos do sinal de fala, tais como *pitch* e duração. Atualmente, muitos sistemas de síntese concatenativa não têm considerado tal processamento, pois, em alguns casos, esse processamento pode causar algum prejuízo à qualidade da fala sintetizada.

2.4.3 Fala Sintética de Alta Qualidade

Cada etapa do processo de síntese concatenativa exerce certa influência sobre o resultado final da fala sintética obtida. Uma fala sintética é considerada como de alta qualidade quando apresenta tanto inteligibilidade quanto transparente naturalidade. Por inteligibilidade

entende-se a fala sintética na qual há o perfeito entendimento daquilo que está sendo falado. Naturalidade, por sua vez, diz respeito à fala sintética que se assemelha ao máximo à fala humana.

Alguns fatores são especialmente importantes para a obtenção de uma fala sintética de alta qualidade, dentre os quais podem-se enumerar:

a) a "qualidade" do banco de fala;

b) a eficácia do algoritmo adotado para seleção automática de unidades;

c) o procedimento adotado para concatenar segmentos de fala em um conjunto integrado;

d) as técnicas de processamento utilizadas na alteração de parâmetros prosódicos.

Dentre os fatores mencionados, a qualidade do banco de fala desenvolvido para síntese possui especial importância no presente trabalho. A qualidade do banco é determinada por fatores como:

a) a duração do *corpus*. Em *corpora* de fala de longa duração, as unidades básicas de síntese (fone, difone, etc.) estão disponíveis em maior quantidade. Assim sendo, a variedade na escolha das unidades de síntese é aumentada, propiciando uma melhor escolha de tais unidades. Dessa forma, uma maior naturalidade de fala sintética pode ser alcançada [28];

b) o texto escolhido para gravação. O texto escolhido para gravação deve procurar englobar o maior número possível de variações fonéticas e prosódicas disponíveis na língua. Essa variabilidade possibilita encontrar unidades consecutivas de síntese com maior similaridade, fator que reduz as descontinuidades resultantes do processo de concatenação. Tais descontinuidades, similares a estalos para a percepção humana, consistem em mudanças abruptas de um ou mais parâmetros acústicos (como, por exemplo, as freqüências dos formantes – F_n, onde n representa o n-ésimo formante) do sinal de fala. Tais mudanças podem ocorrer em virtude de diferentes contextos fonéticos, segmentação incorreta, variabilidade acústica e prosódia diferente [27], [29]. A Fig. 2.12 mostra um espectrograma contendo uma descontinuidade típica em F_n. A freqüência do formante F_n muda abruptamente de 180Hz para 170 Hz em um curto período de tempo (5 ms);

c) o ambiente de gravação. A gravação de um *corpus*, desenvolvido para aplicações em síntese de fala, deve acontecer em um estúdio adequado, evitando, dessa maneira, a existência de ruídos indesejados na gravação. Um monitoramento por especialistas também é recomendado;

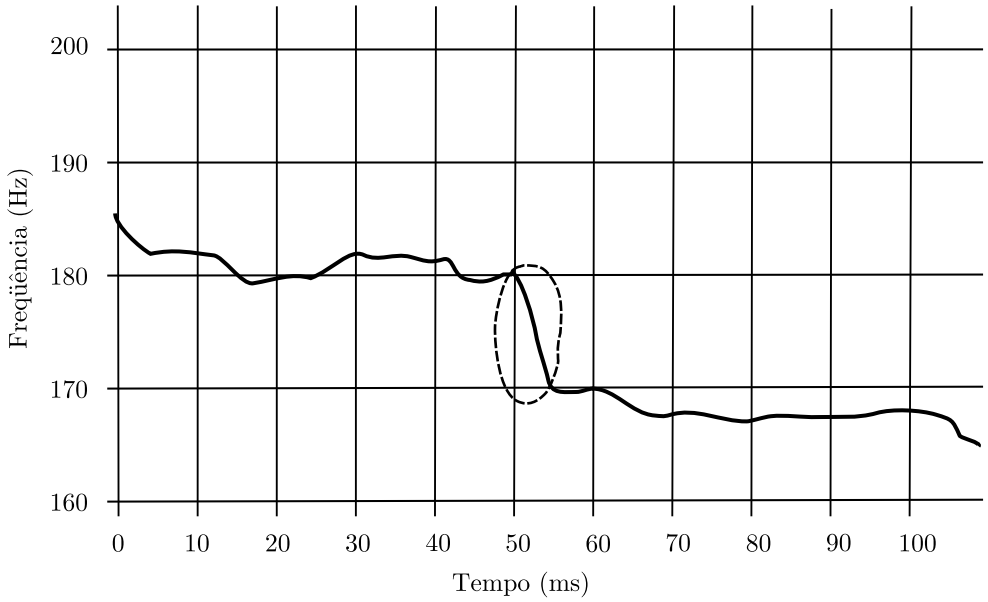

Fig. 2.12: Descontinuidade típica em F_n.

d) o locutor. Outro fator determinante da qualidade da fala sintética é a própria voz do locutor. É desejável que o locutor escolhido produza o menor número possível de degradações audíveis causadas involuntariamente pelo seu próprio aparelho fonador, tais como perturbações em freqüência e amplitude de vibração das cordas vocais (*jitter* e *shimmer*, respectivamente) bem como cliques naturais.

A interação na síntese entre descontinuidades resultantes da concatenação e cliques naturais, sob análise no presente trabalho, pode prejudicar consideravelmente a qualidade da fala sintética [30].

A Fig. 2.13 ilustra a interação entre cliques e descontinuidades em um trecho (amostrado à taxa de 16 kHz) de fala sintética contendo três cliques naturais e uma descontinuidade. Esse trecho [a d ʊ s a l aˈ] é retirado da síntese da frase "A força também lembra da promessa de campanha do presidente Lula de dobrar o poder de compra do salário mínimo em quatro anos". Na figura, os cliques são envolvidos por elipses. Os pontos de concatenação de segmentos são indicados por setas. Na região vizinha ao primeiro ponto de concatenação, os formantes são indicados por linhas horizontais. Percebe-se, nesse ponto, uma mudança abrupta de valor do primeiro e terceiro formantes o que, por si só, conduz a uma degradação audível. Essa degradação associada à degradação causada pelos três cliques existentes no segmento reduz consideravelmente a qualidade da fala sintética.

Para se minimizar a degradação causada pela interação entre cliques e descontinuidades, uma solução apropriada seria o tratamento tanto de cliques quanto de descontinuidades presentes na fala sintética. Entretanto, o tratamento das descontinuidades implica em um aumento da complexidade computacional do processo de síntese propriamente dito. O

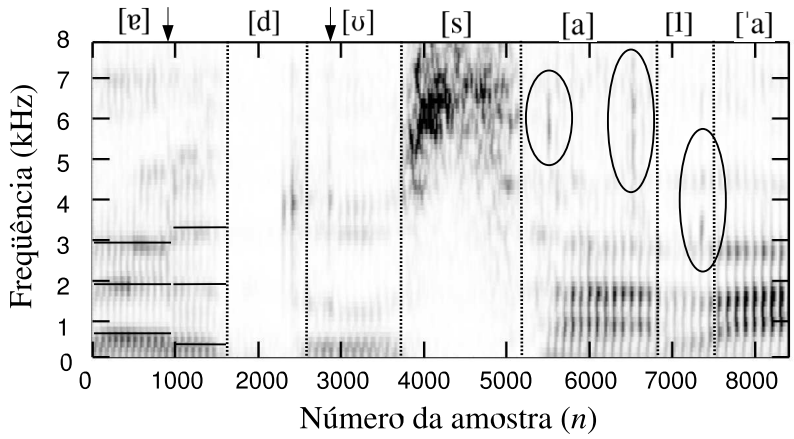

Fig. 2.13: Espectrograma de um trecho de fala sintética contendo cliques naturais e descontinuidades resultantes do processo de concatenação.

tratamento dos cliques naturais, por sua vez, não tem qualquer influência na complexidade, pois é realizado de maneira *off-line* sobre o banco de fala. Dessa forma, o processamento de cliques naturais é aqui preferido no que tange a melhora de qualidade da fala sintética.

2.5 Conclusões

Este capítulo apresenta uma introdução à síntese concatenativa de fala. Primeiramente, uma breve descrição da evolução dos sistemas de síntese é apresentada. Nessa descrição, alguns sistemas mecânicos, elétricos e computacionais baseados em conversão texto-fala são mostrados. As aplicações desses conversores bem como seus módulos de funcionamento (processamento lingüístico e de sinais) são também descritos. O texto concentra-se no módulo de processamento de sinais, o qual é comumente realizado adotando-se um procedimento concatenativo. A síntese concatenativa, técnica de processamento de sinais mais adotada nos sistemas no estado-da-arte em síntese de fala, e os fatores que conduzem a uma fala sintética de alta qualidade são também relatados. Dentre esses fatores, a qualidade do banco desenvolvido para um sistema de síntese apresenta especial importância para este trabalho. Isto porque os cliques naturais exercem significativa influência sobre a qualidade do banco de fala. Uma descrição de tais cliques é apresentada de maneira mais detalhada no Capítulo 3.

Capítulo 3

Cliques Naturais

Os cliques naturais consistem em degradações audíveis presentes em bancos de fala. Manifestam-se na forma de estalos de pequena intensidade. São praticamente imperceptíveis na fala corrente, sendo facilmente percebidos por um ouvinte experiente escutando elocuções previamente gravadas. São produzidos de maneira involuntária pelo próprio aparelho fonador humano, o que nos leva a nomeá-los cliques involuntários.

É importante mencionar que tais cliques não representam elementos de degradação para todos os idiomas existentes. Em alguns idiomas africanos, dentre os quais podem-se enumerar *!Xóõ*, *!Xũ*, Nama, Zulu e *Xhosa*, e no idioma australiano *Damin* [31], cliques são classificados como fonemas consonantais, sendo produzidos de forma voluntária e carregando, portanto, informações úteis da língua [32]–[35].

Além dos cliques consonantais e dos naturais, investigados neste trabalho, existem também os cliques emergentes, descritos em [20], [36]. Esses cliques ocorrem exatamente na transição entre dois fonemas consonantais, sendo freqüentemente confundidos com fonemas plosivos[1], tais como [p] e [t]. Um exemplo é o aparecimento do "fonema [p]" entre os fonemas [m] e [n] na pronúncia da palavra inglesa *damnation*. Nesse caso, na transição entre os fonemas [m] e [n], há duas constrições simultâneas no trato vocal (uma labial[2] e outra apical [3]). Essa configuração, ilustrada na Fig. 3.1, cria uma pequena cavidade entre os lábios e a língua que se expande e, com a liberação abrupta da constrição labial, há a produção de um clique audível.

O conhecimento sobre cliques consonantais e emergentes [20], [34]–[36] pode ser útil para tentar explicar o mecanismo de produção dos cliques naturais, ainda não explicado satisfatoriamente na literatura. A produção de cliques consonantais envolve um fluxo de ar

[1] As consoantes plosivas /b, d, g, p, t, k/ são sons transitórios da fala produzidos pela formação de uma cavidade de ar e liberação abrupta de uma constrição total no trato vocal [13].

[2] Por constrição labial entende-se o estreitamento do trato vocal utilizando os lábios superior e inferior como articuladores [35].

[3] A constrição apical refere-se a um estreitamento do trato vocal no qual a parte anterior da língua se aproxima de um outro articulador (dente, palato, dentre outros) [35].

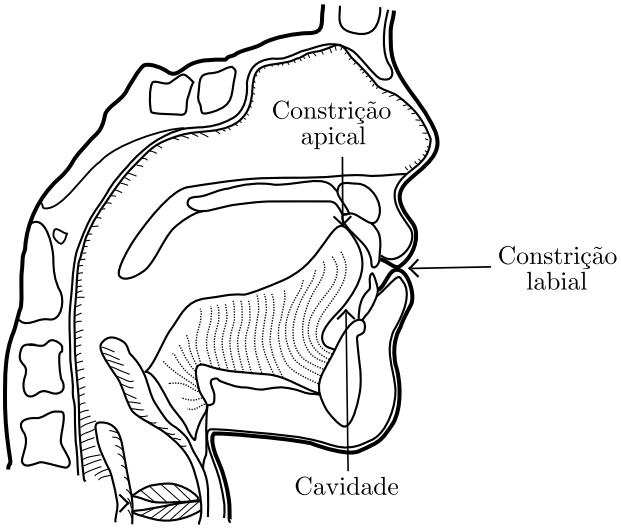

Fig. 3.1: Configuração do trato vocal durante a produção de um clique emergente [20].

ingressivo atravessando uma constrição total ou parcial formada entre a língua e um ponto articulatório. Os cliques emergentes, conforme exemplificado anteriormente, surgem a partir da formação de uma cavidade entre dois pontos de articulação. Quando essa cavidade sofre expansão de volume, o relaxamento de um dos pontos articulatórios leva à produção de um clique. Dessa maneira, acreditamos que os cliques involuntários sejam originados por mecanismos similares.

Para se entender como os cliques involuntários comportam-se em termos temporais, a Fig. 3.2 mostra um segmento de fala vozeado contendo um clique facilmente percebido pelo ouvido humano. Tal clique, em destaque, está inserido no início de um fone [d] presente no banco de fala de um sistema de síntese. Outro exemplo de clique é o mostrado na Fig. 3.3. Esse clique está presente em um segmento em que há uma pausa de respiração. Como os cliques apresentam características similares a de um ruído adicionado ao sinal de fala, parece ser mais fácil notar a existência do clique localizado em um segmento vozeado. Nesse segmento vozeado existe uma maior correlação entre amostras contíguas. Assim qualquer característica ruidosa é evidenciada e, portanto, o clique é da mesma forma evidenciado.

O comportamento dos cliques em termos espectrais é analisado através das Figs. 3.4 e 3.5. A Fig. 3.4 apresenta o espectrograma de um segmento de fala vozeado contendo um clique natural. Tal segmento corresponde a um fone [i] seguido do início de um fone [a]. A Fig. 3.5 ilustra o comportamento espectral de um clique localizado em um fone não-vozeado. O segmento de fala apresentado inclui, respectivamente, os fones [v], [ʊ], [s] e uma região de pausa. O clique localiza-se no fim do fone [s]. A existência de um clique é observada nos espectrogramas por uma linha vertical em evidência. Essa linha evidencia um conjunto de freqüências inexistentes nas amostras contíguas aos cliques.

Para verificar como os cliques naturais se comportam em termos audíveis, o arquivo

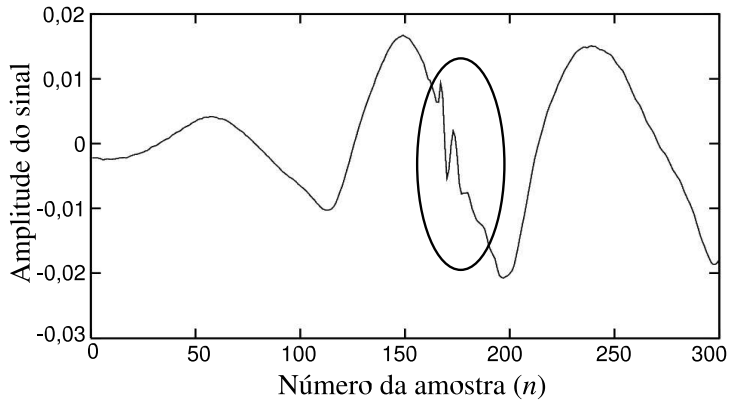

Fig. 3.2: Segmento de fala vozeado contendo clique involuntário.

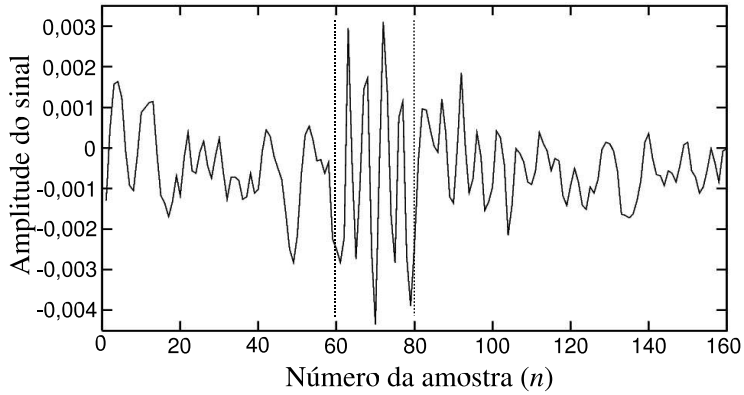

Fig. 3.3: Segmento de fala não-vozeado contendo clique involuntário.

"Frase01_original.wav" é apresentado no CD anexado a este trabalho. Nesse arquivo, cliques naturais estão localizados nos instantes de tempo destacados na Tabela 3.1.

3.1 Marcação Manual

Os cliques naturais ou involuntários não foram ainda satisfatoriamente explorados na literatura da área. Em nosso conhecimento, este é um dos primeiros trabalhos que tratam do tema cliques involuntários (excetuando-se os artigos publicados em [30] e [37]).

Para se obter uma descrição estatística dos cliques naturais e perceber seus efeitos audíveis, um banco de fala é adotado como referência. Esse banco apresenta 45 minutos de duração e contém um considerável número de cliques naturais. Tal banco, gravado por uma locutora experiente em um estúdio profissional, é analisado por uma ouvinte experiente responsável por anotar cada clique audível percebido. Dados como duração do clique natural, fonema de ocorrência, fonemas anterior e posterior ao fonema degradado, freqüências limites

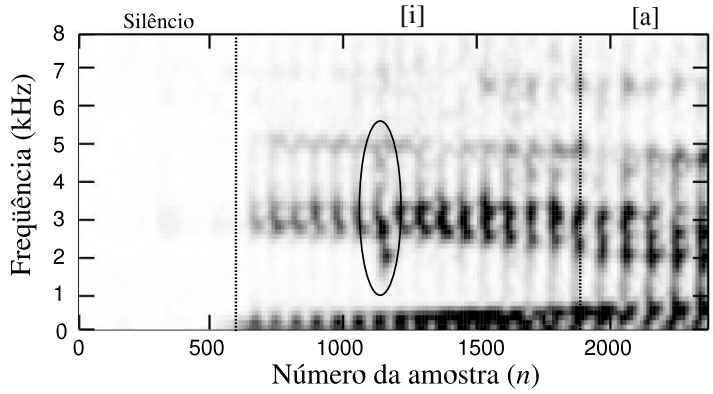

Fig. 3.4: Espectrograma de um segmento de fala vozeado contendo um clique involuntário.

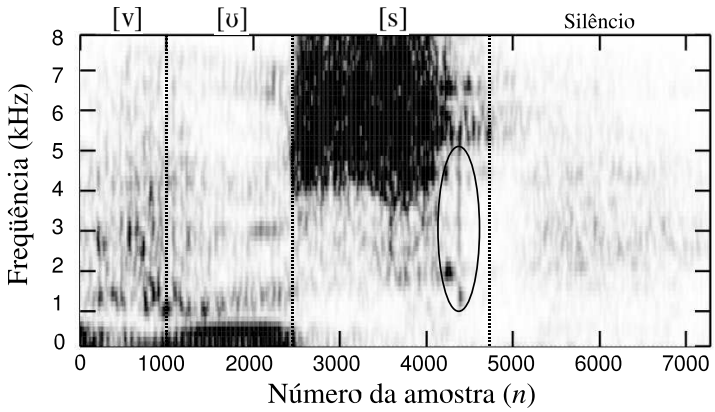

Fig. 3.5: Espectrograma de um segmento de fala não-vozeado apresentando um clique natural.

de banda, freqüência dominante[4], dentre outros, também são considerados.

É importante mencionar ainda que o processo de marcação de dados sobre cliques possibilita a avaliação do desempenho de técnicas automáticas de detecção. A existência de um banco com cliques rotulados permite comparar a localização temporal dos cliques naturais encontrados e anotados no processo manual com os detectados via algoritmos computacionais.

O procedimento de anotação manual é extremamente trabalhoso, demandando um considerável tempo de realização (cerca de 400 horas para o banco de 45 minutos).

[4]Entende-se por freqüência dominante a freqüência na qual ocorre um máximo na curva de densidade espectral de potência.

Tabela 3.1: Localização temporal dos cliques presentes no arquivo Frase01_original.wav

Clique	Tempo inicial	Tempo final
01	0,623886	0,624698
02	2,059753	2,061840
03	3,533954	3,535737
04	4,411446	4,412239
05	4,921266	4,923492
06	5,105262	5,106453
07	5,354491	5,355735
08	6,008077	6,009748
09	6,313066	6,315330
10	6,398797	6,400536
11	6,662912	6,664126

3.2 Conclusões Obtidas a partir da Marcação Manual

A marcação manual, apesar de ser um procedimento exaustivo, possibilitou obter uma descrição estatística dos cliques presentes no banco sob análise. Essa descrição demonstra sua importância, pois permite que pesquisadores sem acesso a um banco de referência anotado manualmente possam gerar cliques artificialmente, visando testar procedimentos de detecção e tratamento.

Algumas conclusões obtidas são aqui relatadas:

a) os cliques naturais aparecem em maior quantidade quando a locutora está cansada, fato evidenciado por uma respiração mais ofegante;

b) o número de cliques encontrados no banco é de 3024. Tal número implica em uma taxa de cliques para a locutora em questão de 67,20 cliques por minuto (3024 cliques/45 minutos), valor que não pode ser desconsiderado. Deve-se destacar que a taxa varia bastante no próprio banco. Existem segmentos com 8 cliques em um segundo de duração como também segmentos sem nenhum clique. A Fig. 3.6 apresenta um histograma mostrando o número de ocorrências de cada taxa de cliques (em cliques por segundo). É importante destacar que essa taxa pode variar muito entre locutores. Para outro locutor profissional analisado, observa-se uma taxa de cliques de 101 a cada minuto (para uma gravação de dois minutos). A escolha do locutor de um banco utilizado para síntese deve, portanto, levar em consideração o número médio de cliques que ele produz a cada minuto.

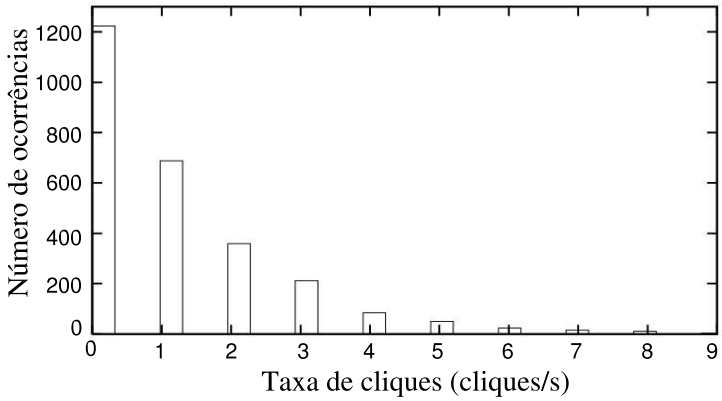

Fig. 3.6: Histograma da taxa de cliques para o banco sob análise.

c) a duração média dos cliques é de 2,17 ms. Um histograma da duração dos cliques localizados no banco e uma curva de densidade de probabilidade (modelagem considerada) correspondente são apresentados na Fig. 3.7. A análise da curva permite concluir que a duração dos cliques naturais pode ser representada por uma variável aleatória com densidade de probabilidade qui-quadrado com 7 graus de liberdade;

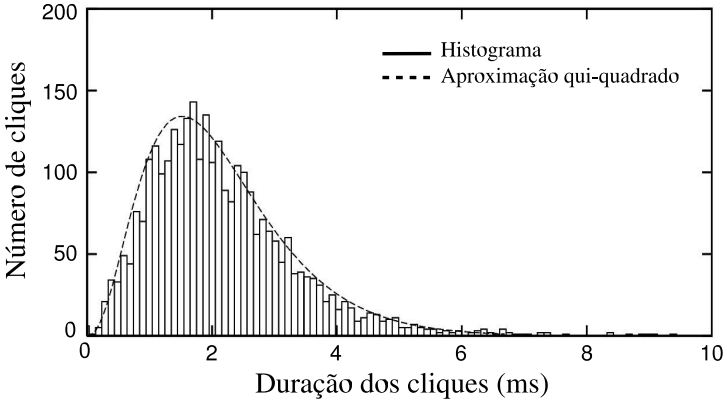

Fig. 3.7: Histograma da duração para os 3024 cliques naturais sob análise.

d) o *locus* (freqüência dominante) médio dos cliques é de 3642,4 Hz. Um histograma do *locus* é apresentado na Fig. 3.8. Tal histograma apresenta uma distribuição que não pode ser modelada por uma função densidade de probabilidade conhecida;

e) o valor médio da freqüência limite inferior de banda para os 3024 cliques sob análise é de 2473,8 Hz. O histograma respectivo é ilustrado na Fig. 3.9. Nesse caso, o histograma também apresenta uma distribuição que não pode ser modelada por uma função densidade de probabilidade conhecida. Verifica-se ainda que um número considerável de cliques apresenta freqüência limite inferior de banda de aproximadamente 1 kHz;

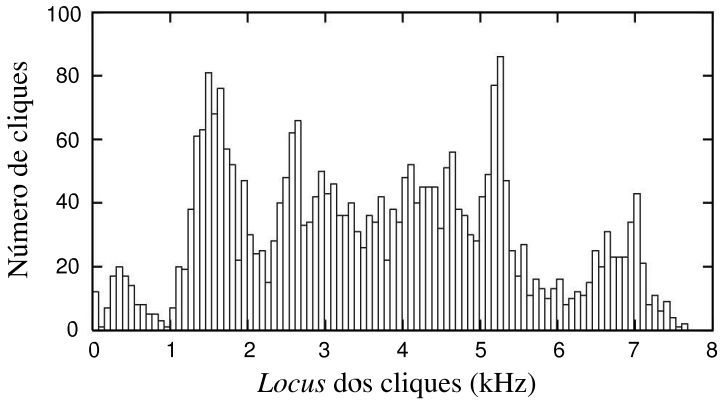

Fig. 3.8: Histograma do *locus* para os 3024 cliques naturais sob análise.

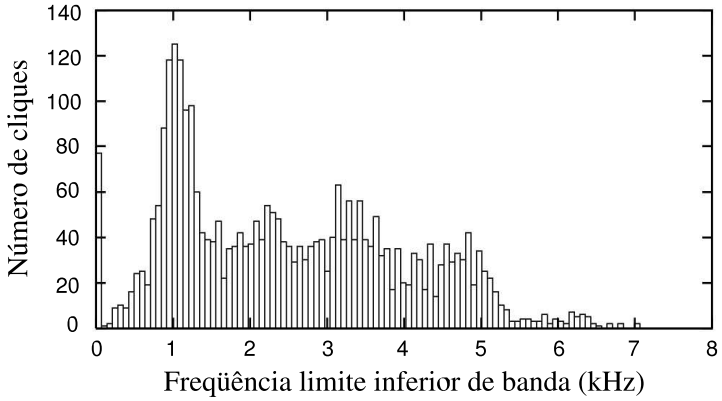

Fig. 3.9: Histograma da freqüência limite inferior de banda para os 3024 cliques naturais sob análise.

f) o valor médio da freqüência limite superior de banda para os 3024 cliques naturais é de 6937,2 Hz. O histograma da freqüência limite superior é apresentado na Fig. 3.10. Percebe-se, em tal histograma, que uma grande percentagem dos cliques apresenta freqüência limite superior de banda igual ou muito próxima a 8 kHz;

g) excetuando o fone [b], todos os demais apresentaram ao menos um clique audível. O fonema com maior ocorrência de cliques foi o fonema [t]. A Tabela 3.2 apresenta os segmentos de fala responsáveis por cerca de 50% dos cliques existentes no banco analisado[5]. No Anexo 1, é apresentada uma tabela contendo o número de cliques existentes em cada fonema do banco de fala analisado;

[5]Na Tabela 3.2, pausas curtas representam segmentos de silêncio com duração superior a 90 ms e inferior a 300 ms. Quando a duração de um segmento excede 300 ms, tem-se uma pausa longa. Os segmentos inicial e final de uma sentença nos quais não há fala são sempre considerados como silêncio. É importante mencionar que o som correspondente aos demais fones apresentados pode ser escutado no *site*: http://web.uvic.ca/ling/resources/ipa/charts/IPAlab/IPAlab.htm. Tal *site* apresenta o som correspondente a cada símbolo fonético definido pela Associação Internacional de Fonética (IPA).

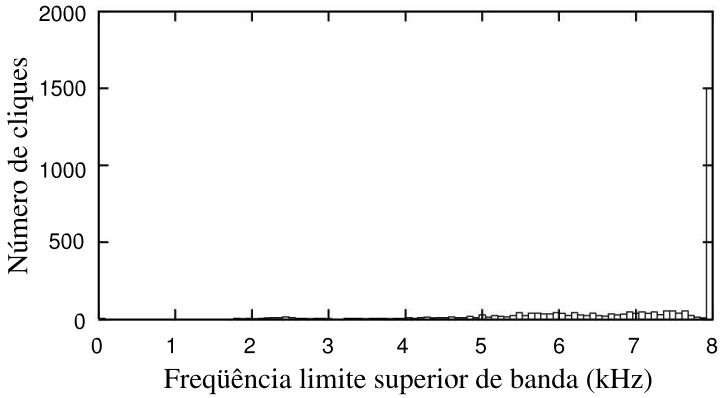

Fig. 3.10: Histograma da freqüência limite superior de banda para os 3024 cliques naturais sob análise.

Tabela 3.2: Segmentos de fala responsáveis por 51,75% dos cliques existentes no banco

Fone	Número de cliques	Porcentagem
[t]	189	6,25
Pausa longa	166	5,49
[a]	161	5,32
Silêncio	160	5,29
[n]	154	5,09
[l]	137	4,53
[m]	122	4,03
[r]	107	3,54
[k]	97	3,20
[S]	93	3,08
Pausa curta	91	3,01
[ɾ]	90	2,98

h) a maior porcentagem de cliques ocorre na transição entre os fones [s] (final de palavra) e [k] (0,96%).

i) cliques localizados em segmentos de fala de menor energia são mais facilmente percebidos em virtude do fenômeno de mascaramento que ocorre em regiões com maior energia. Tal fato é confirmado por uma maior audibilidade de cliques localizados nas regiões de silêncio, pausas curtas e pausas longas. Percebe-se ainda, analisando a Tabela 3.2, que 13,79% dos cliques se encontram nessas regiões de silêncio e pausas (longas ou curtas).

3.3 Modelagem

A Fig. 3.11 ilustra um segmento de fala vozeado contendo um clique involuntário encontrado no banco sob análise. O segmento considerado nessa figura, selecionado dentre gravações que compõem o *corpus* de fala considerado, é degradado pela presença de um clique localizado entre as amostras 130 e 145.

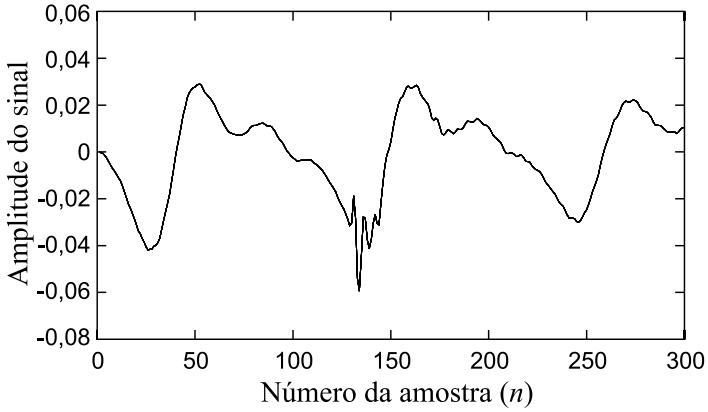

Fig. 3.11: Segmento de fala vozeado contendo um clique involuntário.

O clique em questão pode ser considerado como um sinal espúrio (ruído) adicionado ao sinal de fala ideal (isento de cliques). Tal padrão é sempre encontrado quando cliques involuntários estão presentes em gravações.

Dessa forma, para modelar cliques involuntários é adotada uma representação similar à de um ruído aditivo intermitente incorporado ao sinal de fala ideal. Tal representação é mostrada na Fig. 3.12. Assim, um sinal $y(n)$, contendo cliques involuntários, é modelado por

$$y(n) = x(n) + i(n)r(n), \tag{3.1}$$

onde $x(n)$ é o sinal de fala ideal, $i(n)$ denota um processo de chaveamento assumindo valores $\{0, 1\}$, o qual indica respectivamente a ausência ou presença de clique, e $r(n)$ representa os cliques que degradam o sinal de fala ideal. A modelagem aqui proposta é inspirada na representação de ruídos impulsivos presentes em gravações de áudio antigas [38]. O objetivo de realizar tal modelagem consiste em possibilitar a geração de cliques artificiais semelhantes aos naturais (involuntários). A geração artificial, por sua vez, facilita a avaliação das ferramentas de detecção de cliques à medida que se conhecem *a priori* suas localizações.

3.4 Conclusões

Neste capítulo, uma descrição sobre cliques naturais (involuntários) é apresentada. Realiza-se também uma diferenciação entre esses tipos de cliques (naturais) e os emergentes

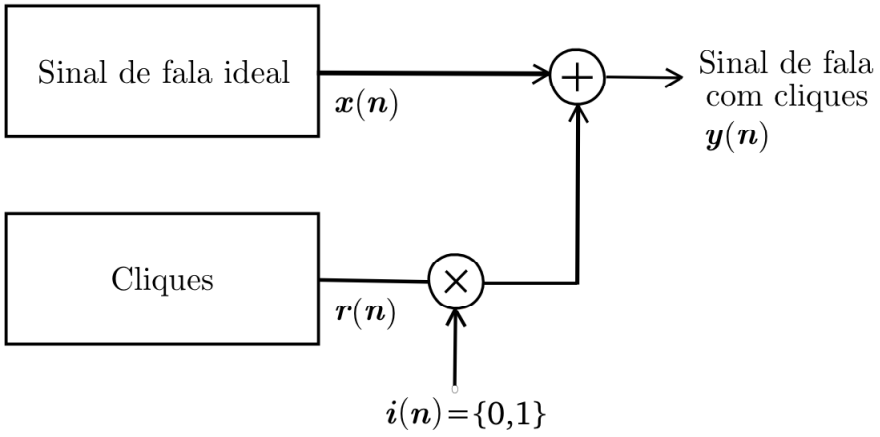

Fig. 3.12: Modelagem de um sinal de fala contendo cliques involuntários.

e/ou consonantais. A percepção dos cliques em termos audíveis é ilustrada também através do CD disponível em anexo. Segmentos e espectrogramas de sinais com cliques são também apresentados. Uma marcação manual dos cliques presentes em um banco de fala com 45 minutos de duração é realizada. Essa marcação tem permitido obter uma breve descrição estatística dos cliques bem como avaliar o desempenho de técnicas automáticas de detecção, apresentadas no capítulo seguinte. A avaliação de técnicas automáticas é facilitada através da modelagem e geração artificial de cliques (também relatadas neste capítulo), especialmente para os pesquisadores sem acesso a um banco de referência obtido manualmente.

Capítulo 4

Detecção de Cliques

Uma etapa necessária para eliminar e/ou atenuar o efeito audível (estalo) de eventuais cliques naturais presentes em bancos de fala consiste em determinar suas localizações ao longo do tempo em um sinal de fala. Essa etapa, denominada detecção, deve idealmente apresentar um desempenho que possibilite encontrar todos os cliques (verdadeiros positivos) sem indicar falsos cliques (falsos positivos). Entretanto, em aplicações práticas, falsos cliques são obtidos a uma taxa que não pode ser desprezada e verifica-se, por experimentação, que não há prejuízos à qualidade da fala quando um tratamento apropriado é aplicado sobre falsos cliques (desde que falsas detecções não ocorram a uma taxa excessiva).

Tanto no que se diz respeito a detecções incorretas quanto a falsas, a técnica que forneceria um melhor desempenho de detecção certamente seria a manual realizada por um ouvinte experiente. Entretanto, tal técnica é extremamente trabalhosa e demanda um considerável tempo (da ordem de dias) para a análise de poucos minutos de gravação. Por esse motivo, pesquisas foram iniciadas visando conceber métodos de detecção automática.

Neste capítulo são apresentadas as técnicas pesquisadas, dentre as quais podem-se citar:

a) análise temporal;

b) análise tempo-freqüência;

c) modelagem do aparelho auditivo humano;

d) análise multirresolução;

e) ou ainda redes neurais artificiais.

Dentre as técnicas de análise temporal, a filtragem inversa e a análise baseada na derivada são atualmente adotadas para detectar ruídos impulsivos (cliques) presentes em gravações de áudio antigas [4], [39]. É importante mencionar que esses ruídos impulsivos apresentam amplitudes consideravelmente superiores às amplitudes dos cliques naturais, conforme

pode-se observar comparando as amplitudes relativas dos cliques apresentados nas Figs. 3.2 e 4.1. Essa diferença de amplitude relativa implica em uma dificuldade consideravelmente superior para detectar cliques naturais, não tão evidentes quanto aqueles presentes em gravações de áudio antigas.

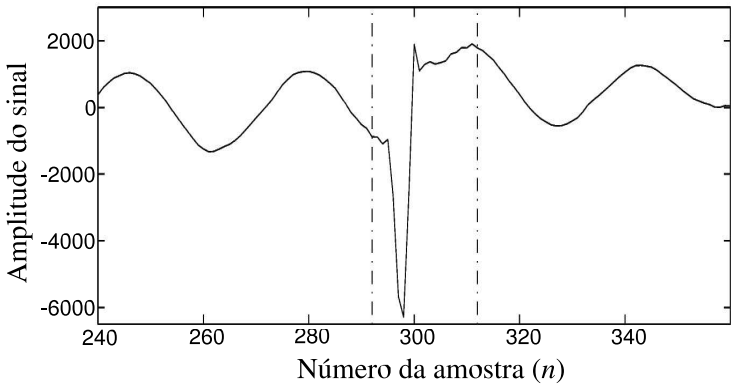

Fig. 4.1: Clique existente em gravação de áudio [4].

O presente trabalho ainda apresenta neste capítulo uma técnica de análise tempo-freqüência, baseada em filtragem por sub-bandas e análise do erro de predição para cada sub-banda. Outras técnicas são apresentadas ora visando imitar as características do aparelho auditivo humano, ora realizando análise por decomposição *wavelet*. Uma análise baseada em redes neurais também é discutida.

4.1 Análise Temporal

4.1.1 Detecção por Filtragem Inversa

A detecção por filtragem inversa é baseada na análise do erro de predição considerando-se uma modelagem autorregressiva (AR). O presente trabalho pretende avaliar o desempenho de tal método para detectar cliques naturais presentes em sinais de fala.

Neste método, um segmento considerado estacionário do sinal de fala ideal (isento de cliques) $x(n)$ é modelado como um processo AR de ordem p. Conhecidos os coeficientes $a(j)$ do modelo autorregressivo, cada amostra do sinal nos instantes $n = p, p+1, \ldots, N-1$ é dada por

$$x(n) = \sum_{j=1}^{p} a(j)x(n-j) + e(n), \tag{4.1}$$

onde, a primeira parcela passível de predição é representada por uma combinação linear das p amostras anteriores ao instante n e a segunda é chamada erro de predição ($e(n)$) ou excitação [4].

Como o sinal ideal $x(n)$ não é conhecido, os coeficientes do modelo AR são estimados considerando-se como referência o sinal com cliques $y(n)$. Este, por sua vez, é filtrado pelo filtro inverso com função de transferência

$$H(z) = 1 - \sum_{j=1}^{p} a(j) z^{-j}, \tag{4.2}$$

de modo a se obter o sinal de excitação (ou erro de predição) correspondente

$$e(n) = y(n) - \sum_{j=1}^{p} y(n-j)a(j). \tag{4.3}$$

Substituindo-se (3.1) em (4.3), tem-se

$$e(n) = e_x(n) + i(n)r(n) - \sum_{j=1}^{p} i(n-j)r(n-j)a(j), \tag{4.4}$$

com

$$e_x(n) = x(n) - \sum_{j=1}^{p} x(n-j)a(j), \tag{4.5}$$

onde $e_x(n)$ é a parcela da excitação correspondente apenas ao sinal ideal $x(n)$ (sem cliques). Tal parcela pode ser modelada por um ruído branco Gaussiano com variância $\sigma^2_{e_x}$ se o sinal $x(n)$ tiver média zero e se a estimativa dos parâmetros do modelo autorregressivo for adequada. A segunda parcela de (4.4) representa o sinal de clique propriamente dito. A última parcela, por sua vez, corresponde a uma combinação linear das amostras degradadas por cliques. Analisando essa última parcela, verifica-se que o sinal de clique exerce influência sobre o erro de predição não apenas nas correspondentes amostras com cliques como também nas p amostras posteriores ao final de um clique [4].

O erro de predição $e(n)$ assume valores pequenos para a maioria das amostras de um sinal de fala. Entretanto, quando há uma mudança abrupta de magnitude esse erro de predição tende a aumentar. Tal fato é verificado, por exemplo, em consoantes plosivas e em cliques naturais. Dessa maneira, a análise do erro de predição pode ser útil para indicar a existência de cliques naturais em um sinal de fala.

Um procedimento comumente adotado por detectores baseados em filtragem inversa supõe que os parâmetros do modelo AR e a variância do erro de predição σ^2_e do sinal com cliques sejam conhecidos. Os parâmetros do modelo AR podem ser estimados através da minimização do erro $e(n)$ em relação a cada um dos coeficientes. Existem duas técnicas principais de estimação: o método da autocorrelação, em que é realizado um "janelamento" do sinal de fala e o erro é minimizado no intervalo $0 \leq n \leq (N + p - 1)$, e o método da covariância em que é feito um janelamento do erro, calculado no intervalo $0 \leq n \leq (N - 1)$,

onde N é o comprimento da janela e p, a ordem do preditor. O método da autocorrelação é mais eficiente em termos computacionais do que o método da covariância, pois este último requer um número maior de operações para a resolução das equações matriciais. Além disso, no método da covariância, a estabilidade do filtro preditor não é garantida [8], [40].

Após a estimação dos parâmetros do modelo AR, o erro de predição é obtido e submetido a um limiar. Nesse caso, se $|e(n)|$ superar um dado limiar definido por $k\sigma_e$, então atribui-se o valor unitário a $i(n)$. Caso contrário, o valor nulo é atribuído.

O valor de k serve como um parâmetro que ajusta o limiar de detecção. Tal parâmetro deve assumir um valor que estabeleça um compromisso entre não-detecção (k de valor alto) e falsa detecção (k de valor baixo). Um dos maiores problemas relacionados à técnica de detecção por filtragem inversa é, justamente, a obtenção de limiares que realizem um compromisso satisfatório entre índices de falsa detecção e não-detecção de cliques. Dadas a característica altamente não-estacionária do sinal de fala e a diversidade de amplitude e duração dos cliques em situações reais, a obtenção de tais limiares não é trivial.

Após a obtenção das amostras possivelmente degradadas por cliques é comum se adotar o critério de união de distúrbios adjacentes. Esse critério promove uma união forçada entre os distúrbios detectados e que estejam separados por até n amostras [4].

As Figs. 4.2 e 4.3 apresentam, respectivamente, um segmento de fala com clique e seu correspondente valor absoluto do erro do predição (bem como o limiar). A análise dessas figuras permite concluir que, quando há ocorrência de um clique, há aumento do valor absoluto do erro de predição.

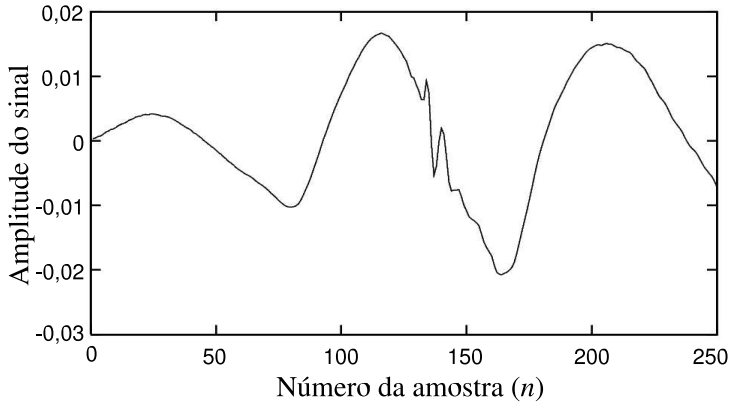

Fig. 4.2: Segmento de fala com clique.

4.1.2 Detecção por Análise Baseada na Derivada

Outra abordagem considerada visando detectar cliques naturais consiste no método baseado na derivada, o qual apresenta resultados similares aos da filtragem inversa para detectar ruídos impulsivos em gravações de áudio antigas [41]. Essa técnica considera como

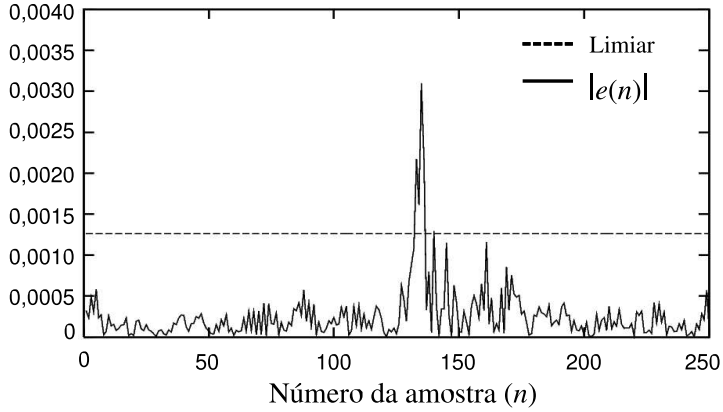

Fig. 4.3: Valor absoluto do erro de predição e limiar para $k = 3$.

corrompidas as amostras cujo valor absoluto da derivada de quarta ordem de um sinal supera um dado limiar, obtido através de média-móvel. O presente trabalho pretende avaliar sua adequação à detecção de cliques naturais presentes em bancos de fala.

A derivada $d(n)$ de um sinal de fala $y(n)$ degradado por cliques é obtida pela divisão da diferença entre duas amostras sucessivas pelo período de amostragem T_s. Assim,

$$d(n) = \frac{y(n+1) - y(n)}{T_s}. \tag{4.6}$$

Para melhorar a detectabilidade de picos, a derivada deve ser aplicada várias vezes. Verifica-se que, na prática, a quarta derivada é a mais adequada para se encontrar as menores degradações audíveis em um sinal [41]. O sinal de detecção efetivamente submetido a um limiar é dado por

$$g(n) = \frac{|y(n-2) - 4y(n-1) + 6y(n) - 4y(n+1) + y(n+2)|}{(T_s)^4}. \tag{4.7}$$

O limiar é obtido por um procedimento de média-móvel, com a seguinte formulação:

$$l(n) = \frac{k}{2i+1} \sum_{m=n-i}^{n+i} g(m), \tag{4.8}$$

onde k é um fator de escala ajustável (obtido experimentalmente) e o comprimento da média é $2i + 1$.

A Fig. 4.4 apresenta $g(n)$ e $l(n)$ para o mesmo sinal de fala considerado na Fig. 4.2. Os valores de $g(n)$ e limiar são obtidos considerando-se $k = 5$, $i = 40$ e uma freqüência de amostragem de 16 kHz. A partir da análise da figura, conclui-se que a derivada de quarta ordem apresenta seu valor aumentado nas amostras correspondentes à região com clique.

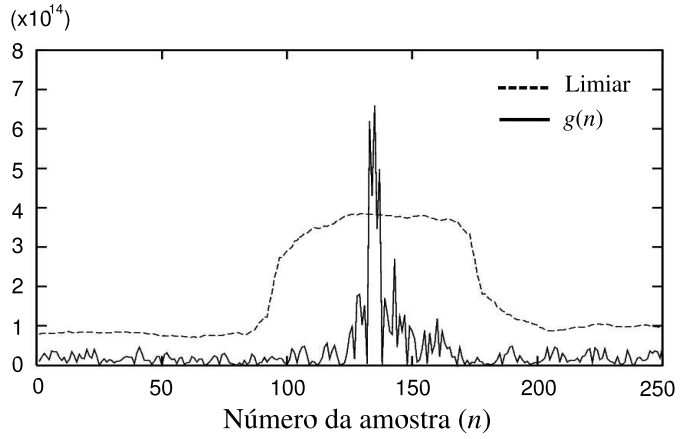

Fig. 4.4: Sinal $g(n)$ e limiar para $k = 5$ e $i = 40$.

4.2 Detecção Baseada na Modelagem do Aparelho Auditivo Humano

O ouvido humano é capaz de perceber com relativa facilidade a existência de um clique natural, o que nos leva a concluir que nosso ouvido possui algumas propriedades responsáveis por enfatizar tais tipos de cliques. O objetivo de um procedimento de detecção baseado na modelagem do sistema auditivo humano consiste em tentar imitar essa capacidade humana de perceber os cliques naturais com relativa clareza.

4.2.1 Sistema Auditivo Humano

A Fig. 4.5 ilustra o sistema auditivo periférico humano. Tal sistema divide-se essencialmente em três partes: ouvido externo, médio e interno.

O ouvido externo, composto pelo pavilhão auditivo (orelha) e pelo canal externo, é responsável pela recepção das ondas sonoras. As ondas recebidas são guiadas pelo canal auditivo até o órgão inicial do ouvido médio, o tímpano. A membrana timpânica vibra. Essa vibração é transmitida à cóclea pela ação conjunta de três ossículos: o martelo, a bigorna (*incus*) e o estribo (todos pertencentes ao ouvido médio).

A cóclea é preenchida por fluidos pouco compressíveis (endolinfo e perilinfo). Apresenta em sua base a janela oval. A vibração mecânica proveniente do estribo é transmitida pela janela oval para os fluidos internos à cóclea, dividida em seu comprimento por duas membranas: a membrana de Reissner e a basilar. Cada freqüência de onda sonora transmitida excita uma determinada região da membrana basilar. Sons de alta freqüência produzem maior excitação (deslocamento) da membrana basilar em regiões próximas à janela oval. O contrário ocorre para sons de baixa freqüência. Dessa forma, a cóclea se comporta como um analisador de espectro. Ao longo da membrana basilar existem sensores chamados de célu-

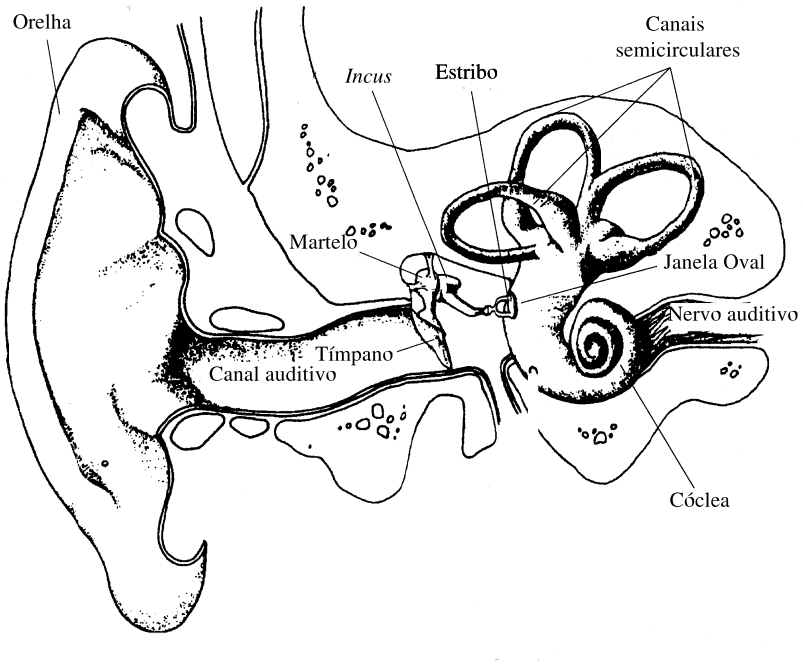

Fig. 4.5: Estrutura do sistema auditivo periférico humano [7].

las capilares internas, responsáveis em transformar o deslocamento mecânico da membrana basilar em informação a ser transmitida pelo nervo auditivo [7], [42].

É importante mencionar que o ouvido médio do sistema auditivo humano pode ser modelado como um filtro passa-altas [42]. Além disso, o deslocamento da membrana basilar em um certo ponto pode ser visualizado como um sinal de saída de um filtro passa-faixa cuja resposta em freqüência tem um pico de ressonância em uma freqüência que caracteriza tal ponto. Essa freqüência de ressonância é denominada freqüência característica. O logaritmo da freqüência característica é aproximadamente proporcional à distância ao longo da membrana.

4.2.2 Modelagem do Aparelho Auditivo Humano

Na tentativa de reproduzir de forma aproximada o processo de percepção dos sons efetuado pelo ouvido humano, uma codificação comumente considerada é a MFCC (*Mel frequency cepstrum coefficients* [13]). Nessa codificação, coeficientes MFCC são obtidos adotando-se o seguinte procedimento:

a) Pré-ênfase: Nessa etapa, o sinal de fala é transformado usando-se um filtro de pré-ênfase, cujo objetivo é enfatizar o conteúdo de energia em altas freqüências. Para tal, considera-se a seguinte equação de diferenças:

$$s'(n) = s(n) - a_1 s(n-1), \tag{4.9}$$

onde $s(n)$ representa o sinal de fala e a_1 (fator de pré-ênfase), uma constante que usualmente assume valores compreeendidos entre 0,9 e 1. É possível observar que a pré-ênfase reproduz a característica em freqüência do ouvido médio (filtro passa-altas).

b) Segmentação: Durante a segmentação, o sinal $s'(n)$ é ponderado, através do produto, por um conjunto de funções janela deslocadas, de maneira a obter sucessivos quadros do sinal de fala.

c) Transformação de Fourier: Essa transformação é aqui obtida considerando-se o uso da transformada rápida de Fourier (FFT – *fast Fourier transform*). Os quadros resultantes de (b) são transformados para o domínio da freqüência através da referida transformação.

d) Aplicação do logaritmo: A aplicação do logaritmo sobre o sinal no domínio da freqüência é realizada visando transformar a operação de produto (entre o sinal de excitação e a resposta impulsiva do trato vocal) em uma operação de soma (análise homomórfica).

e) Banco de filtros distribuídos conforme a escala Mel: Nessa etapa, aplica-se um banco de filtros distribuídos de acordo com a escala Mel. A freqüência em Mel representa a escala de freqüência de um som percebida pelo ouvido humano. A função que mapeia a freqüência real (f_{Hz}) na freqüência Mel (f_{Mel}) é dada por

$$f_{Mel} = \frac{1000}{\log 2}\left[1 + \frac{f_{Hz}}{1000}\right]. \tag{4.10}$$

Uma forma de se aplicar os filtros digitais distribuídos ao longo da escala Mel consiste, primeiramente, em mapear as freqüências (em Hz) para a escala de freqüências percebidas (em mels), e após aplicar um banco de filtros linearmente espaçados nesse domínio (mel). Seja $H_i(\omega)$ a resposta em freqüência do *i*-ésimo filtro do banco e $S(k,m)$ o *k*-ésimo elemento da FFT do *m*-ésimo quadro de $s(n)$. A a saída do *i*-ésimo filtro é dada por

$$Y(i) = \sum_{k=0}^{N/2} \log|S(k,m)|H_i\left(k\frac{2\pi}{N'}\right), \tag{4.11}$$

onde N é o comprimento do quadro m e N' é o número de pontos adotados para computar a FFT. Note que a adoção de um banco de filtros busca simular o comportamento da membrana basilar humana (conjunto de filtros passa-faixa).

f) Transformada inversa do Cosseno (ICT): Tal transformada é adotada com o objetivo de descorrelacionar as saídas do banco de filtros, sendo calculada por

$$c(n) = \sum_{k=1}^{M} \log|Y(k)|\cos\left[\frac{n\pi}{M}\left(k - \frac{1}{2}\right)\right], \tag{4.12}$$

para $0 \leq n < p$, onde $c(n)$ é o *n*-ésimo coeficiente mel-cepstral e M, o número de filtros adotados no banco de filtros.

4.2.3 Localização dos Cliques

Através de critérios *ad hoc*, verifica-se que o cálculo da distância entre vetores de coeficientes MFCC consecutivos serve como indicativo para obter a localização temporal de cliques naturais.

Na técnica proposta, um sinal de fala $s(n)$ é inicialmente segmentado em quadros. Para cada quadro q_k, uma distância $d(q_k,q_{k+1})$ entre vetores de parâmetros MFCC consecutivos é obtida. Algumas medidas de distância adotadas para calcular a distância entre vetores de coeficientes são a absoluta e a Euclidiana.

A distância absoluta entre dois vetores **x** e **y** é

$$Ab(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \,. \tag{4.13}$$

A distância Euclidiana é

$$Eu(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{4.14}$$

Ambas as distâncias serão avaliadas quanto ao desempenho para detectar cliques naturais.

Após o cálculo da distância, uma métrica de distância relativa $d_r(q_k)$ é obtida. Tal métrica varia com a medida de distância escolhida. A métrica $d_r(q_k)$, aqui proposta, para a distância absoluta é dada por

$$d_r(q_k) = 2d(q_k,q_{k+1}) - d(q_{k-1},q_k) - d(q_{k+1},q_{k+2}). \tag{4.15}$$

Para a distância Euclidiana, em função do termo quadrático existente nessa medida, adota-se a métrica

$$d_r(q_k) = 2\sqrt{d(q_k,q_{k+1})} - \sqrt{d(q_{k-1},q_k)} - \sqrt{d(q_{k+1},q_{k+2})}. \tag{4.16}$$

O objetivo dessa transformação consiste em verificar o quanto a distância entre dois quadros consecutivos se destaca em relação à distância entre os demais quadros situados na vizinhança.

Finalmente, $d_r(q_k)$ é submetido a um limiar estipulado. Quando o valor de $d_r(q_k)$ supera o limiar, considera-se que q_k e/ou q_{k+1} apresentam um clique natural. Dessa maneira, conclui-se pela existência de um clique dentro do intervalo entre o início de q_k e o fim de q_{k+1}. Para se obter uma localização mais precisa do clique, realiza-se uma busca (em segmentos vozeados), dentro do intervalo considerado, da região com maior número de variações do

sinal da derivada de $s(n)$. Essa região corresponde à região com clique.

A Fig. 4.6 apresenta o sinal $s(n)$ contendo um clique natural. Tal sinal é submetido a técnica de detecção aqui considerada. A Fig 4.6 ilustra a métrica de distância relativa e o limiar estipulado, nesse caso igual a 0,2. Observa-se que, na região com clique, há um aumento do valor da métrica de distância relativa, conforme esperado.

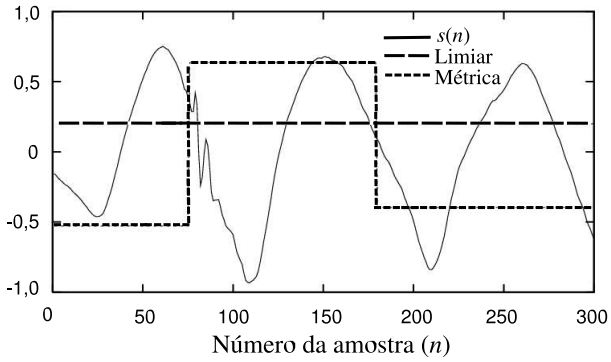

Fig. 4.6: Métrica de distância relativa, limiar e o sinal $s(n)$ sob investigação.

4.3 Detecção Baseada em Análise do Erro de Predição em Sub-bandas

Outra técnica de detecção de cliques pesquisada baseia-se na análise do erro de predição para sub-bandas distintas de um sinal de fala. Nesse caso, realizar-se-ia uma análise tanto no domínio do tempo quanto no domínio da freqüência, o que não se verifica nas técnicas baseadas em filtragem inversa e em derivada de quarta ordem, apresentadas anteriormente, ambas baseadas essencialmente em informações temporais.

A técnica proposta é motivada pelo conceito de que um clique é evidenciado quando em uma sub-banda específica sua energia supera a energia de suas amostras contíguas. A Fig. 4.7 ilustra tal fenômeno. Nessa figura, é mostrado o espectrograma de um segmento de fala amostrado à taxa de 16 kHz, com 3200 amostras. Esse segmento é extraído de um fone [ã] contendo o clique involuntário (entre as amostras 2608 e 2623) indicado pela seta.

Através de uma análise da Fig. 4.7 é possível identificar o clique com relativa facilidade na sub-banda de freqüências que se estende aproximadamente de 2 kHz a 5 kHz. A identificação é facilitada porque, na referida sub-banda, a energia do clique se destaca em relação à energia do sinal analisado nas regiões temporalmente próximas ao clique. O que implica em se dizer que, relativamente ao sinal de clique, a região sob análise é de baixa energia. A Fig. 4.8 mostra o diagrama em blocos da técnica de detecção proposta.

O método de detecção envolve as seguintes etapas:

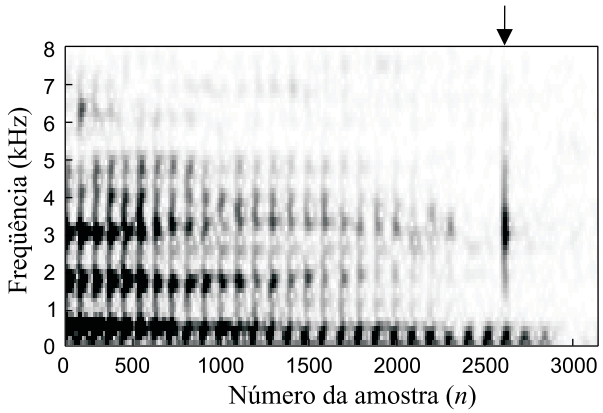

Fig. 4.7: Espectrograma de um segmento de fala contendo um clique involuntário.

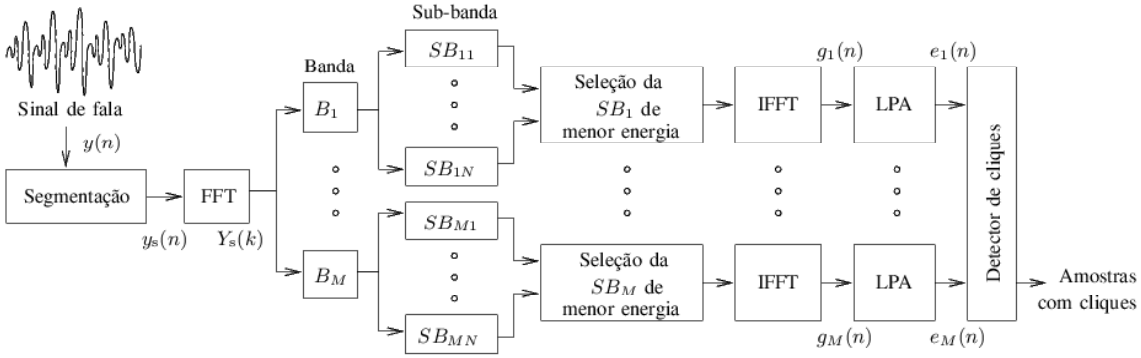

Fig. 4.8: Procedimento proposto para detecção de cliques involutários em segmentos de fala.

a) Segmentação: A primeira etapa da detecção consiste em segmentar o sinal de fala multiplicando cada quadro por uma janela de Hanning.

b) Classificação vozeado/não-vozeado: Nessa etapa, há a leitura de um arquivo de dados, obtidos na etapa de rotulagem do banco, contendo a classificação de um segmento de fala como vozeado/não-vozeado.

c) Transformação para o domínio da freqüência: Cada segmento de fala é, por sua vez, transformado para o domínio da freqüência com o auxílio da transformada de Fourier discreta (DFT). A DFT é aqui obtida através de um algoritmo de transformada rápida de Fourier (FFT).

d) Divisão em M bandas: O segmento de fala, agora no domínio da freqüência, é dividido em M bandas B_1,B_2,...,B_M.

e) Subdivisão em N sub-bandas: Cada uma das M bandas é ainda subdividida em N sub-bandas SB_{ij} para $i = 1,\ldots,M$ e $j = 1,\ldots,N$.

f) Seleção da sub-banda de menor energia: A próxima etapa consiste em selecionar para cada banda a sub-banda de menor energia, desde que essa energia não ultrapasse um limiar estipulado. Dessa forma, são atribuídos valores nulos aos coeficientes da FFT não correspondentes à sub-banda selecionada.

g) Submissão da energia a um limiar: Se a energia de uma dentre as sub-bandas de menor energia ultrapassar o limiar de energia, considera-se uma região de alta energia, em que um possível clique estaria mascarado. Nesse caso, o processo de detecção para a banda considerada é interrompido.

h) Transformada de Fourier discreta inversa (IDFT): Cada sub-banda de menor energia é transformada para o domínio do tempo via um algoritmo FFT inverso (IFFT), resultando em até M sinais de baixa energia.

i) Análise preditiva linear: Cada sinal de baixa energia é submetido a uma análise preditiva linear (LPA), técnica usualmente considerada para restauração de sinais de áudio [38]. Tal análise engloba uma estimação dos parâmetros de um modelo autorregressivo (AR) e um cálculo do erro de predição normalizado $e_i(n)$. O valor do erro é obtido comparando-se o sinal real $g_i(n)$ com a sua estimativa $\widehat{g}_i(n)$, para $i = 1, \ldots, M$.

j) Submissão dos sinais de erro a um limiar: Quando o valor absoluto do erro de predição para um dos M sinais superar um limiar estipulado, conclui-se pela existência de um clique.

k) Visando a melhoria de desempenho da detecção, todo o procedimento é considerado também para o sinal de fala reverso.

Um ponto essencial no processo de detecção consiste em obter os limiares adequados para o detector de cliques. O presente trabalho propõe uma abordagem de maior simplicidade para obtenção do limiar. O seguinte procedimento é aplicado sobre um segmento de fala sob análise:

i) em cada segmento, 1% das amostras de maior amplitude do valor absoluto do erro de predição normalizado é descartado;

ii) a amostra de maior amplitude do sinal resultante após (i) é armazenada. Tal valor é multiplicado por $1{,}5$ para se definir o requerido limiar.

A Fig. 4.9 ilustra um limiar de detecção obtido considerando-se o procedimento anterior, como também o valor absoluto do erro de predição normalizado para um trecho de sinal de fala que apresenta um clique.

É importante mencionar a existência de outros procedimentos para a obtenção de tal limiar. Em detecção de distúrbios impulsivos provenientes do meio de gravação,

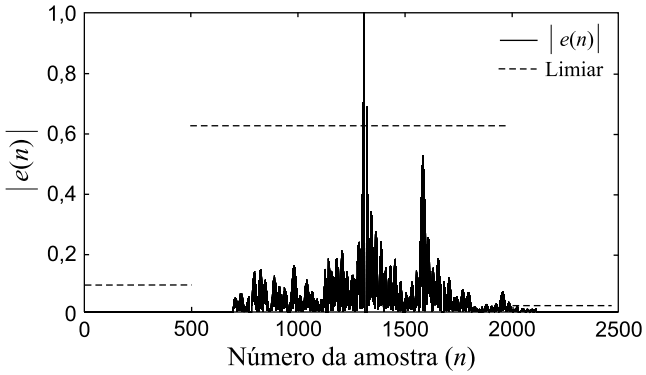

Fig. 4.9: Valor absoluto do erro de predição normalizado e limiar de detecção.

considera-se a estimativa do desvio padrão do sinal de excitação [38] ou, ainda, a mediana do valor absoluto do erro de predição [4], [43]. Entretanto, tais procedimentos se mostraram menos eficazes do que o apresentado anteriormente para o tipo de aplicação em questão.

Destacamos ainda que a técnica proposta se restringe ao tratamento de cliques involuntários existentes em segmentos de fala vozeados, como também não-vozeados de baixa energia. Essa restrição consiste em uma solução aceitável, visto que a percepção de cliques em segmentos não-vozeados de alta energia é atenuada através do efeito de mascaramento existente em tais segmentos.

4.4 Detecção Baseada em Decomposição *Wavelet*

Para detecção de cliques naturais, pretende-se ainda verificar o desempenho de ferramentas de decomposição *wavelet*, freqüentemente adotadas para identificar transitórios existentes em sinais diversos [44]–[47].

4.4.1 Transformada *Wavelet*: Fundamentos Básicos

A análise de dados de acordo com escalas variáveis no domínio do tempo e da freqüência é a idéia fundamental da utilização da teoria de *wavelets* [48]. Tal teoria, desde a década de 80, tem despertado um crescente interesse na comunidade científica. Suas aplicações vão desde análise e compressão de sinais, astronomia, acústica, geofísica, matemática, até a análise de sinais biométricos.

O termo "*wavelet*" foi originalmente introduzido por Jean Morlet e Alex Grossman na década de 80. Esses pesquisadores iniciaram utilizando a palavra francesa "*ondelette*", que significa "onda pequena". Tal palavra foi convertida para o inglês resultando em *wavelet* [49].

Morlet adotou a teoria de *wavelets* para análise de dados sísmicos, para os quais a transformada de Fourier não se fazia adequada visto que tais dados apresentavam conteúdos de freqüência que se alteravam rapidamente ao longo do tempo (não-estacionariedade) [48].

De fato, a transformada de Fourier é somente apropriada para análise de sinais periódicos ou com características estacionárias no tempo [45]. Isso porque essa transformada não permite analisar localmente o conteúdo de freqüência de um sinal. Dessa maneira, eventos que ocorrem em um curto período de tempo afetam a transformada como um todo. Os sinais apresentados nas Figs. 4.10 e 4.11 ilustram essa idéia. Apesar do segundo ser uma réplica revertida do primeiro, ambos apresentam a mesma transformada de Fourier, não sendo possível localizar via DFT onde ocorre o sinal espúrio em ambos os sinais. A magnitude da transformada (obtida via DFT) de tais sinais é apresentada na Fig. 4.12.

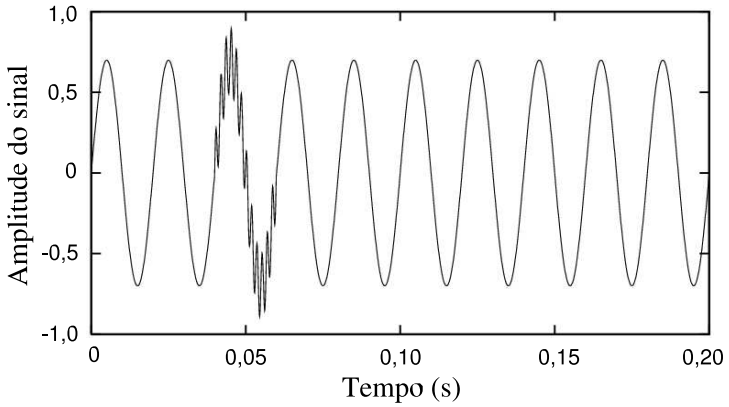

Fig. 4.10: Sinal original no tempo.

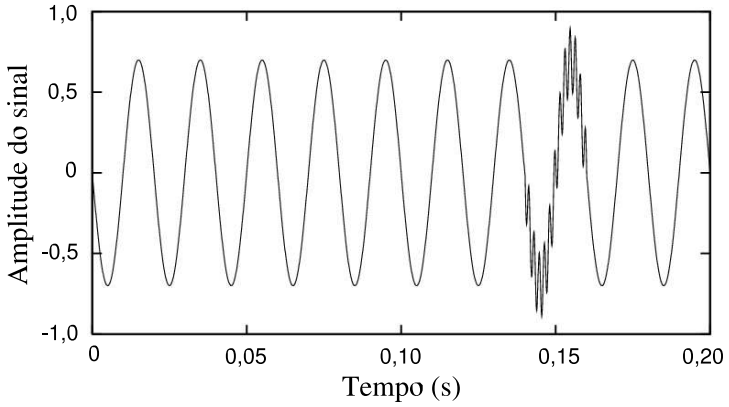

Fig. 4.11: Sinal revertido no tempo.

A transformada de Fourier de um sinal $x(t)$ é definida por

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt, \tag{4.17}$$

onde $e^{-j\omega t}$ é a base da transformada [50]. Deve-se notar que tal transformada baseia-se na integração de todo o sinal para calcular a função que representa o seu espectro de freqüência $X(j\omega)$ [48].

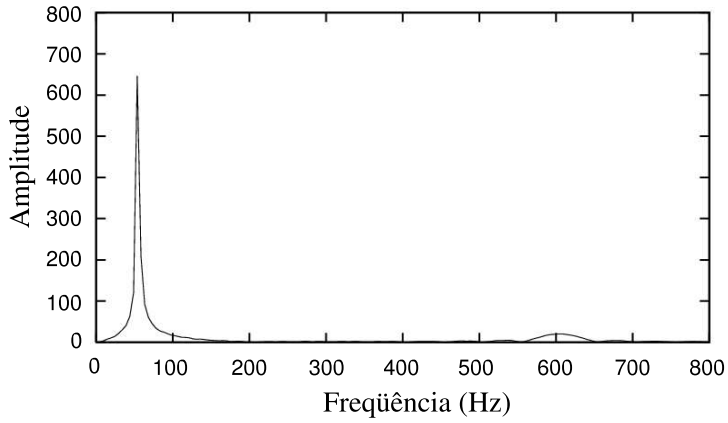

Fig. 4.12: Magnitude da transformada de Fourier dos sinais original e revertido.

A transformada de Fourier de curto tempo, proposta por Dennis Gabor, por sua vez, permite uma análise em freqüência com uma melhor localização temporal. Nesse caso, uma janela de ponderação é deslocada no domínio do tempo e a transformada é obtida para cada posição da janela. Sendo $w(t)$ a janela de observação, define-se a transformada de Fourier de curto tempo como

$$X(\omega,\tau) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j\omega t}dt. \tag{4.18}$$

A transformada de Fourier de curto tempo (STFT), apesar de proporcionar uma melhoria na localização temporal, pode não ser adequada para analisar determinados tipos de sinais. O problema da STFT remonta ao conhecido Princípio da Incerteza de Heisenberg [44]. Tal princípio, originalmente adotado para analisar partículas móveis, pode ser aplicado à informação tempo-freqüência de um sinal qualquer. Esse princípio determina que não se podem obter de forma simultânea localizações precisas no tempo e na freqüência. Esse problema está relacionado ao tamanho da janela de observação, tendo em vista que seu tamanho permanece constante para todas as freqüências.

As Figs. 4.13 e 4.14 mostram espectrogramas, referentes ao sinal da Fig. 4.10, que ilustram o Princípio da Incerteza de Heisenberg. O primeiro espectragrama é obtido considerando-se uma janela com duração de 40 ms, levando a uma adequada resolução em freqüência (1/40 ms = 25 Hz). Esse tamanho de janela (longa duração) conduz a uma baixa resolução no tempo. O segundo espectrograma, por sua vez, considera uma janela com duração de 10 ms, o que conduz a uma melhor resolução no tempo e pior resolução em freqüência (1/10 ms = 100 Hz).

Muitos sinais requerem uma melhor resolução tempo-freqüência, necessitando-se variar o tamanho da janela para apresentar maiores resoluções no tempo e na freqüência [48].

Assim sendo, Morlet introduziu uma transformada, a *wavelet*, na qual o tamanho da janela é variável, permitindo que eventos de alta freqüência sejam localizados com maior resolução temporal [48].

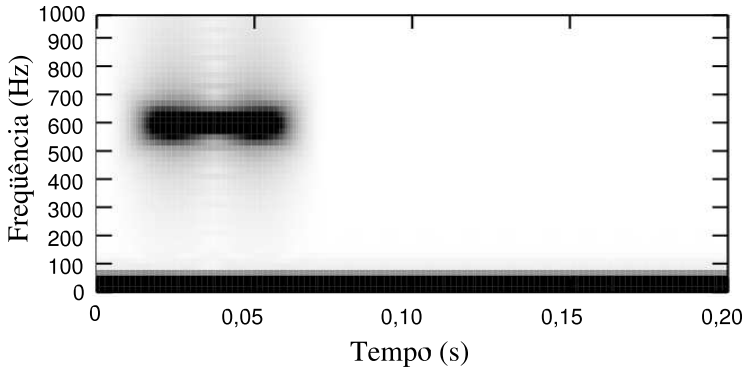

Fig. 4.13: Espectrograma para STFT com janela de 40 ms.

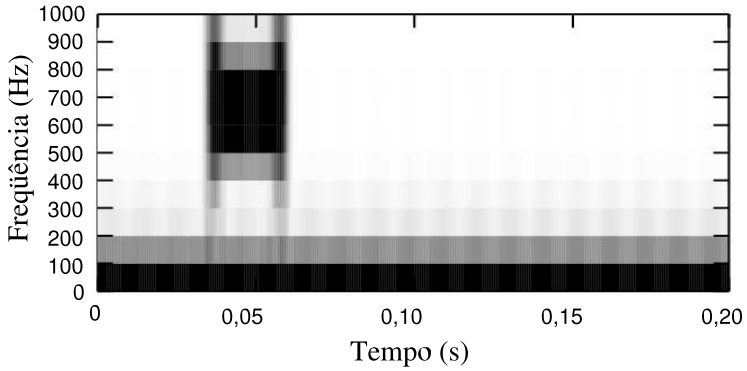

Fig. 4.14: Espectrograma para STFT com janela de 10 ms.

A transformada *wavelet* contínua (*continuous wavelet transform* - CWT) de um sinal $x(t)$, definido em um espaço vetorial de funções quadráticas integráveis $L^2(\Re)$, é dada por

$$X(a,b) = \langle x(t), \psi_{ab}(t) \rangle = \int_{-\infty}^{\infty} x(t)\psi_{ab}^*(t)dt, \tag{4.19}$$

onde a família de funções $\psi_{ab}^*(t)$ representa versões escaladas e transladadas da função $\psi(t)$, denominada *wavelet* mãe. Tal família de funções é obtida escalando ψ por a e transladando por b, obtendo-se

$$\psi_{ab}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \tag{4.20}$$

com os parâmetros de escalamento $a \in \Re^+$, de translação $b \in \Re$ e * denotando o complexo conjugado de $\psi_{ab}(t)$. Essas funções são denominadas *wavelets* filhas ou apenas *wavelets*. O fator $1/\sqrt{a}$ serve para manter a energia das *wavelets* filhas igual à da *wavelet* mãe.

É importante mencionar que existem algumas restrições sobre a *wavelet* mãe $\psi(t)$. Ela deve ser oscilatória, ter média zero, apresentar suporte compacto e energia finita (rápido decaimento). Existem infinitos tipos de funções com essas características. Algumas dessas funções são difundidas na literatura, tais como as *wavelets* de Haar e de Morlet, dentre outras.

A *wavelet* de Haar, por exemplo, é definida como sendo

$$\psi_{Haar}(t) = \begin{cases} 1, & 0 < t \leq 1/2 \\ -1, & 1/2 < t \leq 1 \\ 0, & \text{para os demais valores de t.} \end{cases} \tag{4.21}$$

Tal wavelet é real e descontínua, sendo mostrada na Fig. 4.15. A wavelet de Morlet real (Fig. 4.16), por sua vez, é contínua e é representada pela seguinte expressão:

$$\psi_{Morlet}(t) = e^{-t^2/2}\cos(5t). \tag{4.22}$$

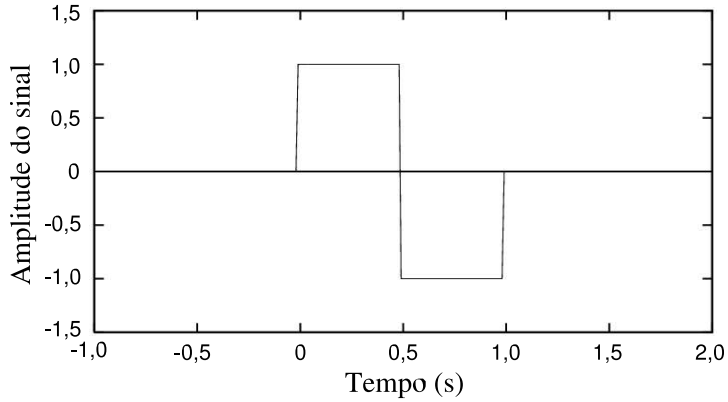

Fig. 4.15: *Wavelet* de Haar.

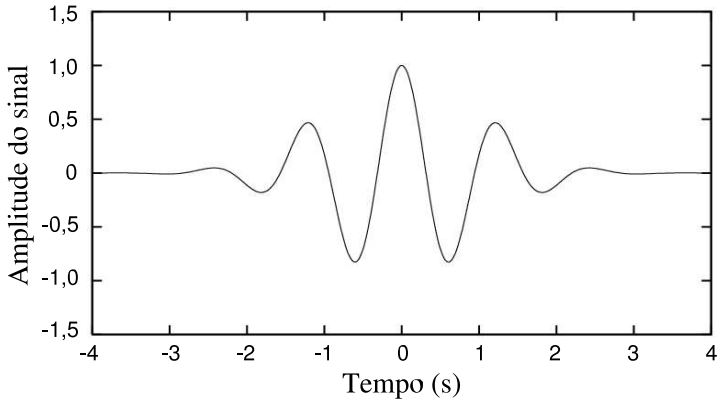

Fig. 4.16: *Wavelet* de Morlet.

O cálculo da CWT é realizado considerando-se, por exemplo, em uma primeira etapa, $a = 1$. Realiza-se então o produto do sinal $x(t)$ pela *wavelet* filha $\psi_{ab}(t)$, com essa *wavelet* localizada no ponto $t = 0$. Calcula-se a integral de tal produto na região para qual $\psi_{ab}(t)$ for diferente de zero. A *wavelet* é agora transladada de um fator b. Calcula-se novamente o produto de $x(t)$ pela nova *wavelet* (transladada) e, posteriormente, a integral de tal

produto. Dessa forma, outro valor é obtido no plano tempo-escala. O procedimento é repetido até que todo o sinal tenha sido analisado para a escala considerada. Posteriormente, o parâmetro de escalamento a é aumentado de um pequeno valor, repetindo-se os passos anteriormente descritos. Quando o processo estiver completo para todos os valores desejados de a, a CWT do sinal estará determinada.

Esta operação de CWT, descrita para calcular a transformada *wavelet* em uma escala considerada, corresponde em essência a uma operação de produto interno. Dessa maneira, assim como o produto interno, a CWT pode ser interpretada como uma medida de similaridade entre o sinal $x(t)$ e cada uma das *wavelets* filhas.

É importante mencionar a existência da transformada *wavelet* discreta (*Discrete Wavelet Transform* - DWT). A principal diferença entre a CWT e a DWT consiste no fato da primeira operar sobre um conjunto contínuo de escalas e translações, enquanto que a DWT opera sobre um conjunto discreto de escalas e translações. Este trabalho adota, no cálculo da transformada *wavelet*, a CWT visto que a análise é efetuada sobre um conjunto contínuo de escalas, permitindo assim se obter uma melhor resolução em freqüência.

4.4.2 Detecção de Cliques através da CWT

Para a detecção de cliques via CWT, deve-se inicialmente definir a *wavelet* mãe $\psi(t)$. A escolha da *wavelet* mãe é aqui realizada (lembrando que a CWT pode ser interpretada como uma medida de similaridade entre o sinal $x(t)$ e cada uma das *wavelets* filhas) verificando-se qual função *wavelet* apresenta maior similaridade (quanto à forma do sinal) com os cliques presentes em sinais de fala. Dentre as *wavelets* clássicas apresentadas na literatura, verifica-se uma maior similaridade dos cliques com a função Morlet. Assim, tal *wavelet* é adotada no presente trabalho como *wavelet* mãe.

Seja o segmento de fala vozeado extraído de um fone [a] do banco de fala, contendo um clique, apresentado na Fig. 4.17. A partir do espectrograma desse sinal (Fig. 4.18), verifica-se que o clique altera o conteúdo de freqüências na faixa que se estende aproximadamente de 4500 Hz a 7000 Hz. Utilizando-se (4.23), equação que transforma a freqüência f na escala respectiva a, verifica-se que essa faixa de freqüências corresponde às escalas de 1,82 a 2,83. Assim a pode ser obtido por

$$a = \frac{5 f_s}{2\pi f}, \tag{4.23}$$

onde f_s é a freqüência de amostragem utilizada.

A CWT é obtida para o segmento de fala sob análise considerando-se as escalas 1, 2, 3, 4 e 5. O valor absoluto dos coeficientes obtidos são mostrados nas Figs. 4.19, 4.20, 4.21, 4.22 e 4.23, respectivamente.

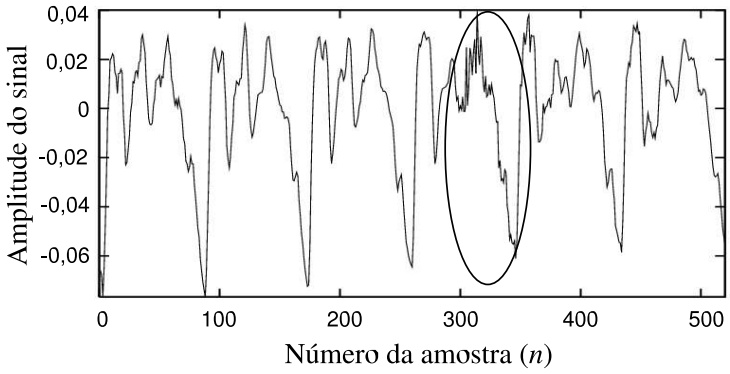

Fig. 4.17: Segmento vozeado contendo um clique natural.

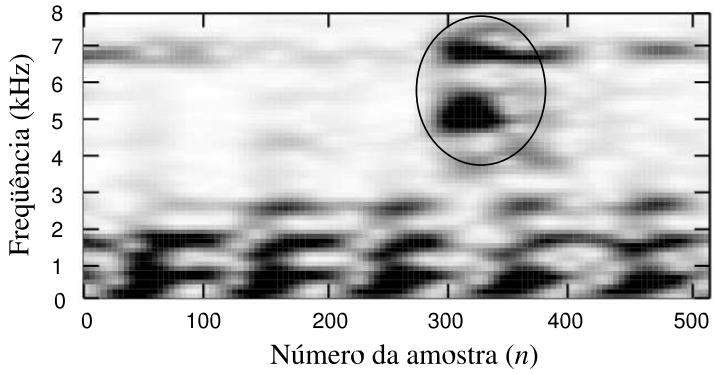

Fig. 4.18: Espectrograma do segmento vozeado contendo um clique natural.

Através da análise destas figuras, verifica-se que, para a escala 2, a CWT atinge seu valor máximo exatamente na região temporal em que o clique se localiza. É interessante observar também que, em escalas correspondentes a baixas freqüências, qualquer clique estará dificilmente enfatizado em função da existência das freqüências dos formantes nessas regiões. Dessa maneira, para se detectarem cliques naturais, deve-se obter, em princípio, a CWT para escalas correspondentes às altas freqüências (baixas escalas). Em um passo subseqüente, a CWT é submetida a um estipulado limiar.

Propõe-se aqui o uso de um limiar baseado na média móvel do valor absoluto da CWT. Essa média, ponderada por um fator estipulado, é aqui considerada como limiar, pois verifica-se, por experimentação, que tal limiar é o mais adequado (dentre os citados em seções anteriores) para perceber maiores variações de amplitude da CWT em regiões com cliques. A Fig. 4.24 apresenta a CWT para a escala 2,5 e o respectivo limiar, para uma média tomada considerando-se 201 pontos e uma constante multiplicativa igual a 7. A análise dessa figura possiblita concluir que o uso da CWT associada a um adequado limiar consiste em um indicativo da existência de um clique natural.

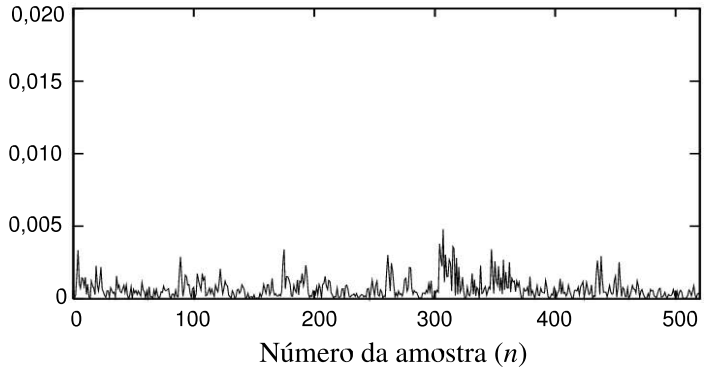

Fig. 4.19: Valor absoluto dos coeficientes *wavelet* para a escala 1.

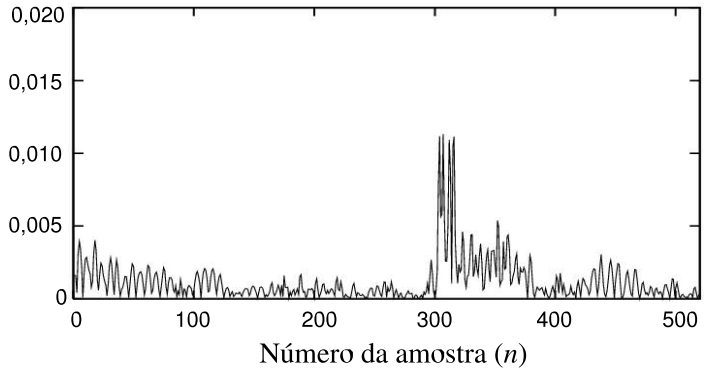

Fig. 4.20: Valor absoluto dos coeficientes *wavelet* para a escala 2.

4.5 Detecção Baseada em Redes Neurais Artificiais

A detecção de cliques utilizando-se redes neurais artificiais (RNA) é também uma técnica aqui discutida quanto a sua eficácia para detecção de cliques naturais. Essa análise é motivada pelo desempenho satisfatório das redes neurais na detecção de distorções impulsivas presentes em gravações de áudio antigas [51].

Rede neurais artificiais são inspiradas nos neurônios biológicos e nos sistemas nervosos. Apresentam essencialmente três topologias: redes diretas (*feedforward*), redes com ciclos e redes simétricas. Dentre essas topologias, a mais popular é a rede direta em função da facilidade de uso e da grande difusão de métodos de aprendizagem para tais redes. Um método de aprendizagem bastante usado nas redes diretas é o de retropropagação do erro (*backpropagation*). Pela popularidade e facilidade de uso, restringiremo-nos à detecção com uma rede neural direta e com algoritmo de aprendizagem de retropropagação do erro.

As redes diretas são comumente representadas em camadas de neurônios. Os neurônios que recebem sinais de excitação compõem a camada de entrada. Neurônios que têm sua saída como saída da rede pertencem à camada de saída. Os demais neurônios compõem uma ou mais camadas intermediárias [52].

O neurônio consiste na unidade de processamento fundamental para a operação

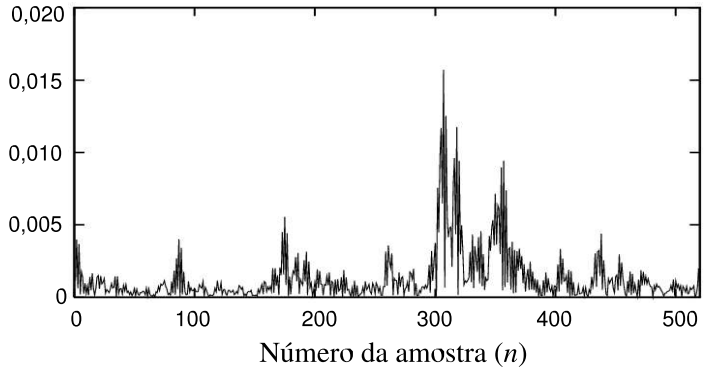

Fig. 4.21: Valor absoluto dos coeficientes *wavelet* para a escala 3.

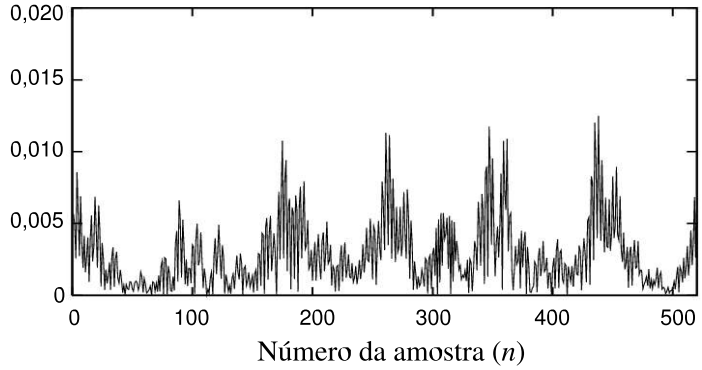

Fig. 4.22: Valor absoluto dos coeficientes *wavelet* para a escala 4.

de uma rede. Sua modelagem básica é apresentada na Fig. 4.25. Um sinal de entrada x_j conectado ao neurônio k é multiplicado pelo peso sináptico w_{kj}. Os sinais de entrada (para $j = 1, \ldots, m$) ponderados são somados e uma função de ativação restringe a amplitude da saída de um neurônio.

Uma função de ativação freqüentemente utilizada na construção de redes neurais artificiais é a sigmoidal [53]. Um exemplo de função sigmóide é a logística, definida por

$$\varphi(v) = \frac{1}{1 + e^{-av}}, \tag{4.24}$$

onde a é o parâmetro que varia a inclinação da sigmóide.

O presente trabalho analisa o desempenho (para detectar cliques) de uma RNA direta, com função de ativação logística e com duas camadas de pesos adaptativos. A utilização de duas camadas permite resolver funções não-linearmente separáveis [52]. A Fig. 4.26 apresenta, a título de exemplo, uma RNA direta com duas camadas de pesos adaptativos.

A rede neural em questão é treinada com o objetivo de separar um quadro do sinal de fala em duas classes: com ou sem cliques. Dessa maneira, a camada de saída utiliza um único neurônio, o que já é suficiente, cujo valor de saída próximo de "1" indica um sinal com clique e próximo de "0", sem clique.

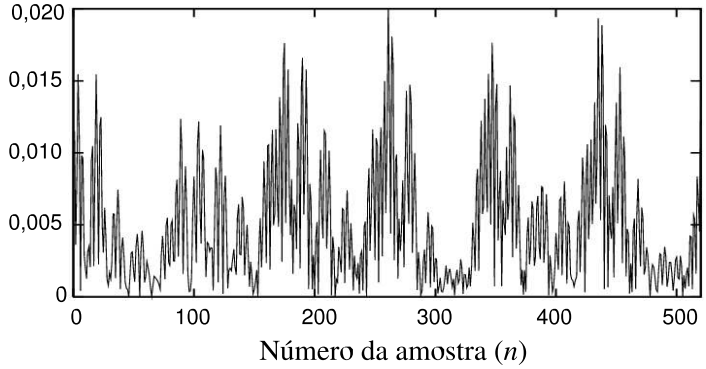

Fig. 4.23: Valor absoluto dos coeficientes *wavelet* para a escala 5.

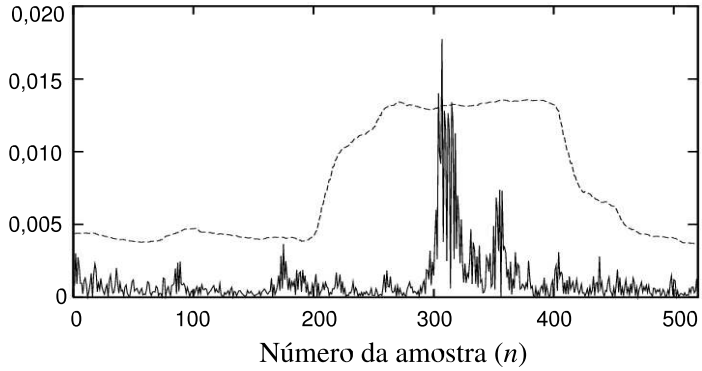

Fig. 4.24: Limiar de detecção e valor absoluto dos coeficientes *wavelet* para a escala 2,5.

O número de neurônios na camada de entrada é determinado pelo conjunto de parâmetros adotados como entrada da rede. A entrada poderia ser o sinal no tempo, a transformada de Fourier do sinal, ou até mesmo coeficientes *wavelet*. A definição dos parâmetros de entrada da rede é aqui realizada considerando-se os resultados obtidos com as técnicas de detecção anteriores. Caso a técnica baseada em filtragem inversa apresente melhor desempenho na detecção, certamente coeficientes LPC são os parâmetros adequados de entrada para a rede. De maneira similar, caso a técnica baseada em decomposição *wavelet* apresente resultados satisfatórios, coeficientes *wavelet* podem ser adotados na entrada da RNA. A RNA funcionaria, na aplicação *wavelet*, como um limiar, pois seria responsável por decidir, a partir dos coeficientes *wavelet*, se um segmento de fala contém ou não clique.

4.6 Conclusões

Este capítulo apresenta um conjunto de técnicas pesquisadas visando a detecção automática de cliques naturais. Técnicas baseadas em filtragem inversa, derivada de quarta ordem, modelagem do aparelho auditivo humano, análise do erro de predição em sub-bandas, análise multirresolução e redes neurais são aqui descritas. Todas essas técnicas são comparadas quanto ao desempenho no Capítulo 5. O objetivo final da detecção é localizar temporal-

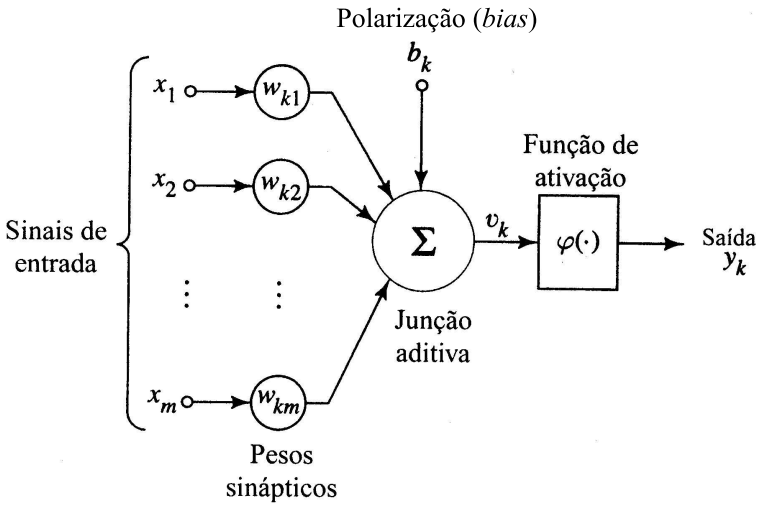

Fig. 4.25: Modelo não-linear de um neurônio [53].

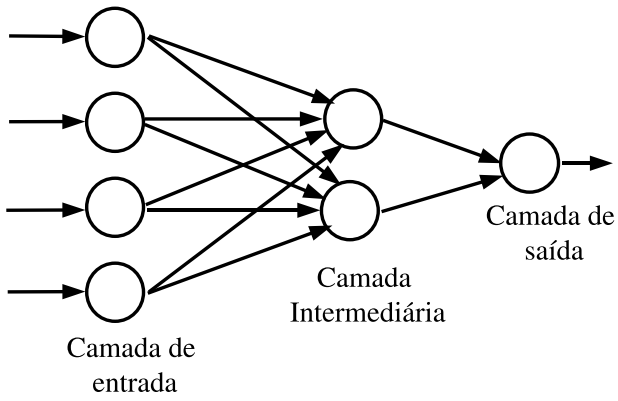

Fig. 4.26: Camadas de neurônios em uma RNA direta.

mente os cliques naturais com vistas a um posterior processamento para torná-los inaudíveis.

Capítulo 5

Comparação entre Técnicas de Detecção

A fim de que as técnicas de detecção apresentadas no capítulo anterior sejam comparadas quanto ao desempenho, é essencial o estabelecimento de critérios adequados de avaliação. Tais critérios são definidos levando em consideração o objetivo deste trabalho: encontrar técnicas de detecção que localizem o maior número possível de cliques naturais e conduzam ao menor número de detecções incorretas.

Neste sentido, a primeira medida proposta como figura de mérito é o índice de detecção correta (IDC) definido por

$$\text{IDC} = \frac{n_\text{c}}{n_\text{t}} \times 100, \tag{5.1}$$

onde n_c representa o número de cliques detectados corretamente e n_t, o número total de cliques naturais existentes no banco de fala sob análise.

Outra medida proposta consiste no índice de detecção incorreta (IDI) definido por

$$\text{IDI} = \frac{n_\text{i}}{n_\text{t}} \times 100, \tag{5.2}$$

onde n_i representa o número de cliques detectados erroneamente.

Os valores de n_c e n_i são determinados através da comparação entre as localizações temporais dos cliques marcados manualmente (marcação discutida na Seção 3.1) e obtidos em cada técnica automática de detecção. Essa comparação é realizada para todos os cliques manualmente marcados no *corpus* de fala com 45 minutos de duração previamente gravado e anotado.

É importante mencionar que a técnica de detecção admitida como mais eficaz é aquela que estabelece uma melhor relação de compromisso entre IDC e IDI (alto IDC e baixo IDI).

5.1 Formatos de Áudio e Plataformas Utilizadas

O banco de fala adotado para testes comparativos das técnicas de detecção de cliques considera freqüência de amostragem de 16 kHz para amostras com resolução de 16 bits. Os arquivos que compõem tal banco apresentam-se no formato ".wav".

Os algoritmos de detecção (exceto o baseado em RNA) são implementados no *software* Octave, usando aritmética de ponto flutuante. Para realização dos experimentos com redes neurais, utiliza-se a biblioteca *fast artificial neural networks* implementada em ANSI C e descrita detalhadamente em [54].

5.2 Experimentos com Detecção via Filtragem Inversa

A primeira técnica de detecção avaliada é baseada em filtragem inversa. Analisando-se os conceitos apresentados no capítulo anterior, verifica-se que alguns parâmetros devem ser definidos anteriormente à aplicação da técnica, tais como:

a) o tipo de janela utilizada para segmentação do sinal de fala;

b) o grau de recobrimento entre quadros consecutivos;

c) a ordem do modelo autorregressivo (número de coeficientes) adotado para representar o filtro inverso;

d) o comprimento de um quadro do sinal de fala;

e) a constante multiplicativa k, que determina o limiar de detecção;

f) a técnica de minimização do erro de predição (autocorrelação ou covariância).

Tais parâmetros são aqui definidos utilizando-se critérios *ad hoc*. Após essa definição, o *corpus* de fala com 45 minutos de duração é segmentado seja com 80 ou com 180 amostras, utilizando-se uma janela de Hanning, com recobrimento (*overlap*) de 50%. Um modelo autorregressivo é obtido para cada quadro de análise. O erro de predição é minimizado utilizando-se a técnica de autocorrelação. Tal erro é submetido a um limiar. Após detectar as amostras possivelmente degradadas, adota-se o critério de união de distúrbios adjacentes. Esse critério considera como um único distúrbio (clique) os cliques naturais localizados a uma distância de até o dobro do número de amostras de um quadro do sinal de fala.

Os valores de IDC e IDI obtidos com a técnica considerada, para diferentes ordens do modelo AR e fatores de ponderação k (que compõem o limiar de detecção) são mostrados na Tabela 5.1.

Tabela 5.1: Resultados experimentais da técnica de detecção de cliques via filtragem inversa

Tamanho do quadro	Ordem do modelo AR	Fator k	IDC	IDI
80	10	3	91,86	4409,00
80	10	4	66,44	1775,00
80	10	5	32,72	651,50
80	10	6	11,79	156,30
80	10	7	2,71	43,57
80	10	8	0,37	11,20
80	10	9	0	0
80	10	10	0	0
80	20	3	92,36	4704,00
80	20	4	73,71	2418,00
80	20	5	40,83	1317,00
180	10	3	93,04	2539,00

Analisando-se os resultados apresentados nesta tabela, verifica-se que a técnica utilizada para detecção de distúrbios impulsivos em gravações de áudio antigas não é adequada para detectar cliques naturais. O número de cliques detectados erroneamente chega, em alguns casos, a superar em 50 vezes o número de cliques corretamente detectados.

5.3 Experimentos com Detecção Baseada na Derivada

Outro procedimento de detecção avaliado é o baseado na derivada de quarta ordem. Alguns parâmetros a serem definidos antes da aplicação da técnica são:

a) o comprimento do filtro de média (número de amostras consideradas na média) utilizado para determinar o limiar de detecção,

b) o fator de ponderação do limiar, obtido por média móvel.

Neste caso, considera-se o critério de união de distúrbios adjacentes localizados a até 360 amostras de distância e o limiar é obtido através do produto entre um fator de ponderação e a média de 81 amostras. Os valores obtidos de IDC e IDI para a referida técnica são mostrados na Tabela 5.2.

Analisando-se os resultados experimentais apresentados nesta tabela, verifica-se que a técnica adotada para detectar ruídos impulsivos em gravações de áudio é, assim como a abordagem AR, inadequada para detectar cliques naturais. O número de cliques detectados erroneamente chega, em alguns casos, a superar em até 130 vezes o número de cliques corretamente detectados. O desempenho dessa técnica é ainda pior do que a técnica utilizando modelagem autorregressiva.

Tabela 5.2: Resultados experimentais da técnica de detecção baseada na derivada

Fator k	IDC	IDI
3	90,46	11883,4
4	90,74	5609,2
5	78,45	3094,6
6	59,46	1262,7
7	39,86	559,9
8	25,02	257,4
9	14,83	140,8
10	8,35	85,7

5.4 Experimentos com a Técnica Baseada na Modelagem do Aparelho Auditivo Humano

Para se realizar a detecção baseada na modelagem do aparelho auditivo humano, um conjunto de coeficientes MFCC é inicialmente obtido. Para tal, realiza-se uma pré-ênfase com fator de 0,97 sobre o banco de fala. Posteriormente, esse banco é segmentado utilizando-se uma janela de Hamming considerando-se quadros com comprimento de 2,5 vezes o período de *pitch*. Por fim, realiza-se uma análise de 12^a ordem, com 24 filtros compondo o banco de filtros, resultando em 12 coeficientes MFCC para cada quadro de análise.

Em uma etapa posterior à obtenção dos coeficientes MFCC, calculam-se as distâncias Euclidiana e absoluta e suas respectivas métricas de distância relativa. As métricas resultantes são normalizadas pelos seus valores máximos. A Tabela 5.3 mostra os valores obtidos de IDC e IDI para a técnica considerada. Os resultados são apresentados para seis diferentes limiares fixados.

Os resultados desta técnica de detecção indicam a inexistência de melhoria significativa em relação às técnicas anteriores. Entretanto, deve-se destacar que tal técnica obtém uma localização mais precisa dos cliques naturais. A aplicação de técnicas de processamento *ad hoc* após a detecção dos cliques resultaria em uma melhor qualidade da fala processada (em relação as técnicas anteriormente analisadas).

5.5 Experimentos com a Técnica Baseada em Análise do Erro de Predição em Sub-bandas

Um conjunto de cliques gerado artificialmente (105), com distribuição estatística (amplitude, duração e fator de amortecimento) semelhante à dos cliques naturais, é adicionado a um sinal de fala isento de cliques. Tal sinal, amostrado à taxa de 16 kHz, apresenta um

Tabela 5.3: Resultados experimentais da técnica baseada na modelagem do aparelho auditivo humano

Distância	Limiar	IDC	IDI
Absoluta	0,05	66,01	4322,29
Absoluta	0,10	58,66	2346,03
Absoluta	0,15	49,07	1298,11
Absoluta	0,20	39,81	758,96
Absoluta	0,25	30,62	476,06
Absoluta	0,30	23,05	312,50
Euclidiana	0,05	66,63	4985,48
Euclidiana	0,10	60,78	2932,37
Euclidiana	0,15	52,64	1631,68
Euclidiana	0,20	43,88	925,69
Euclidiana	0,25	35,32	549,54
Euclidiana	0,30	27,78	351,09

minuto de duração. Esse sinal é segmentado utilizando-se uma janela de Hanning. Cada quadro tem duração de 100 ms com 50 ms de recobrimento (*overlap*). Uma FFT de 4096 pontos é aplicada a cada segmento. Obtém-se então uma divisão em M bandas e N sub-bandas. A escolha dos valores de M, N e limiar de energia (etapa de treinamento) é então feita através de busca exaustiva, visando-se obter o maior número possível de cliques detectados corretamente.

O desempenho da abordagem proposta é avaliado separadamente para segmentos vozeados e não-vozeados. Para os vozeados, os valores de M, N e limiar de energia que proporcionam maior IDC são, respectivamente, iguais a 4, 3 e $0{,}45$. O IDC nesse caso é de $91{,}80\%$. Para segmentos não-vozeados de baixa energia, os valores de M, N e limiar de energia que levam ao maior IDC são, respectivamente, 4, 3 e $0{,}25$. Nesse caso, o IDC obtido é $81{,}82\%$.

O banco de fala com 45 minutos de duração, anotado manualmente, amostrado à taxa de 16 kHz e contendo cliques naturais, é também usado para avaliar a técnica de detecção proposta. Para tal, utilizam-se os mesmos valores de M, N e limiar de energia previamente obtidos. Para o banco considerado, o IDC para segmentos vozeados é agora 55,85%. Para segmentos não-vozeados, o IDC é de 45,22%. O IDI obtido para o caso em questão é de, respectivamente, 681,37% e 731,05%.

Esta técnica, se comparada as anteriormente apresentadas, já estabelece uma melhor relação de compromisso entre IDC e IDI. O IDI é relativamente baixo (quando comparado às demais técnicas) para o IDC considerado.

5.6 Experimentos com a Técnica Baseada em Decomposição *Wavelet*

Para se avaliar a eficácia da técnica de detecção baseada em uma ferramenta CWT, adota-se como referência, em todos os experimentos, o banco de fala com 45 minutos de duração.

Em uma primeira etapa, a CWT é obtida considerando-se como *wavelet* mãe a função Morlet e uma escala $a = 2$, correspondente à freqüência de 6366,2 Hz[1]. Os coeficientes *wavelet* são submetidos a um pré-definido limiar. Tal limiar é obtido através do produto entre o fator de ponderação 6 e a média móvel de 201 coeficientes. Nesse caso, o IDC obtido é 68,74%, e o IDI, 575,53%. Esses valores, comparados com as técnicas anteriores, estabelecem uma melhor relação de compromisso entre IDC e IDI.

Os coeficientes *wavelet* são também calculados para uma escala $a = 2{,}5$. Tal escala corresponde a freqüência de 5093,0 Hz. A partir da análise estatística do banco, verifica-se que 84,33% dos cliques têm algum conteúdo energético sobre essa freqüência. Nesse caso, considerando-se uma constante multiplicativa de 6 para o limiar, o IDC é também 68,74%, e o IDI, 614,75%. A Fig. 5.1 mostra o gráfico que indica a porcentagem de cliques presentes no banco de fala de testes que apresentam conteúdo energético sobre uma freqüência qualquer considerada.

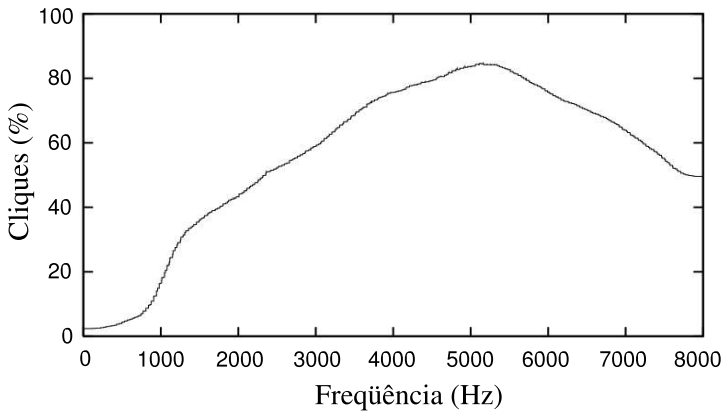

Fig. 5.1: Porcentagem de cliques que possuem conteúdo energético na freqüência considerada.

Outro experimento realizado, considera o maior valor do *i*-ésimo coeficiente *wavelet* nas escalas 2,0 e 2,5. Novamente uma constante multiplicativa de 6 é adotada no cálculo do limiar. O IDC obtido é de 70,13% e o IDI, 617,06%.

[1]É importante mencionar que, a partir da análise estatística do banco, percebe-se que 71,79% dos cliques apresentam algum conteúdo energético sobre a freqüência de 6366,2 Hz.

Tabela 5.4: Resultados experimentais da técnica baseada em análise *wavelet*

Escala	Freqüência	Fator (limiar)	IDC	IDI
2	6366,2	6	68,74	575,53
2,5	5093,0	6	68,74	614,75
2 e 2,5	6366,2 e 5093,0	6	70,13	617,06
1,7 à 2,7	4715,7 à 7489,6	6	76,72	1181,3
1,7 à 2,5	5093,0 à 7489,6	6	79,69	1168,7
1,7 à 2,5	5093,0 à 7489,6	7	87,46	2479,3

Considerando-se o maior coeficiente nas escalas de 1,7 (7489,6 Hz) à 2,7 (4715,7 Hz) com passo de 0,1, obtém-se o IDC de 76,72% e o IDI de 1181,3%.

Este valor de IDC é ainda superado pelo IDC de 79,69%, considerando-se o valor máximo dentre os coeficientes nas escalas de 1,7 a 2,5 com passo de 0,1. Nesse caso, o IDI é de 1168,7%. Com uma constante multiplicativa do limiar igual a 7, o IDC é 87,46%, e o IDI, 2479,3%. Os resultados obtidos para diferentes configurações da detecção baseada em análise *wavelet* são resumidos na Tabela 5.4.

As escalas superiores a 2,7 (4715,7 Hz) não são analisadas quanto ao desempenho, pois em baixas freqüências (maiores escalas) estão contidas as freqüências formânticas. Essas freqüências, por apresentarem alto conteúdo energético, mascaram possíveis cliques (com menor energia) existentes no segmento sob análise.

É importante mencionar que o valor de IDI igual a 1168,7%, apesar de ser um valor alto, quando realizado um procedimento de tratamento específico (descrito no capítulo seguinte) não há prejuízos à qualidade da fala gravada. Um valor de IDI de até 1500% (valor adquirido experimentalmente) ainda é aceitável em termos de não prejudicar a qualidade da fala gravada, quando aplicado um procedimento corretivo de extrapolação (posteriormente descrito). Valores superiores a 1500% resultam em uma deterioração da qualidade da fala gravada e são, portanto, inaceitáveis.

Desta maneira, comparando-se os valores de IDC e IDI obtidos para todas as técnicas analisadas, verifica-se uma melhor relação de compromisso IDC/IDI para a técnica baseada no valor máximo considerando-se os coeficientes *wavelet* das escalas de 1,7 a 2,5 (com passo de 0,1) e adotando-se, como *wavelet* mãe, a função de Morlet e posteriormente um limiar obtido via média móvel com constante multiplicativa igual a 6.

5.7 Experimentos com a Técnica Baseada em Redes Neurais

Após verificar-se a eficácia de detecção da técnica baseada em decomposição *wavelet*, uma RNA é testada quanto à capacidade de exercer a função de obter um limiar de detecção. Essa RNA adota como parâmetros de entrada os coeficientes *wavelet* máximos nas escalas de 1,7 a 2,5 (com passo de 0,1).

O número de neurônios na entrada da rede é definido a partir da duração média dos cliques (2,17 ms). Para uma freqüência de amostragem de 16 kHz, tem-se uma duração média de 35 amostras. Dessa forma, adotam-se 35 neurônios na camada de entrada com um neurônio na camada de saída.

Considera-se função de ativação logística e uma camada oculta de neurônios. O número de neurônios adotados na camada oculta, determinado por critérios *ad hoc*, varia de 10 a 15.

O treinamento é realizado tomando-se como referência a metade do banco anotado manualmente. O banco de fala é segmentado em quadros com duração de 35 amostras (janela retangular) e são extraídos os coeficientes *wavelet* a serem apresentados à entrada da rede neural. Adota-se o algoritmo de treinamento de retropropagação do erro.

Para os diversos testes efetuados, a rede não apresentou um desempenho satisfatório. Mesmo com um número considerável de apresentações da época, o erro não é reduzido e a rede não converge. O provável motivo de tal desempenho consiste na inexistência de um padrão dentre os coeficientes *wavelet* máximos. Sabe-se que o desempenho de uma rede neural atrai as aplicações voltadas à classificação de padrões. Na inexistência de um padrão (ou padrões), o desempenho da rede é quase sempre insatisfatório.

5.8 Conclusões

Neste capítulo, um conjunto de técnicas de detecção automática de cliques naturais é avaliado quanto ao desempenho. Uma técnica de detecção deve idealmente localizar todos os cliques naturais presentes em um sinal de fala, sem indicar falsos cliques. Na prática, verifica-se que com o aumento do número de detecções corretas para uma técnica de detecção considerada, há também o aumento do número de detecções incorretas. Para se estabelecer um compromisso satisfatório entre detecções corretas e incorretas, duas figuras de mérito são aqui definidas: o IDC (índice de detecção correta) e o IDI (índice de detecção incorreta). Por critérios *ad hoc* tem-se verificado que um IDI da ordem de até 1500% não causa danos perceptuais à qualidade da fala com a realização de um processamento apropriado. A técnica que tem apresentado melhor desempenho (IDI de até 1500% e IDC máximo) é a baseada em

decomposição *wavelet*, considerando-se os coeficientes *wavelet* de escalas correspondentes a altas freqüências.

Após a etapa de detecção, um processamento deve ser realizado visando suprimir os cliques detectados. No Capítulo 6, um conjunto de técnicas de processamento são apresentadas para reduzir e/ou eliminar o efeito audível causado por cliques.

Capítulo 6

Processamento dos Cliques Detectados

Após a etapa de detecção, um processamento deve ser realizado visando eliminar e/ou atenuar o efeito audível (estalo) de eventuais cliques presentes no banco de fala.

Neste capítulo, são apresentadas algumas possíveis formas de processamento que serão posteriormente avaliadas quanto ao desempenho na redução dos efeitos audíveis de cliques presentes em sinais de fala. Dentre as abordagens apresentadas, a técnica baseada em extrapolação é atualmente uma das mais consideradas para suprimir ruídos impulsivos (cliques) presentes em gravações de áudio antigas. O presente trabalho pretende avaliar a possibilidade de se aplicar tal técnica bem como outras inéditas para processar cliques naturais.

6.1 *Pruning* dos Fonemas com Cliques

Uma técnica de processamento dos cliques proposta pelo presente trabalho é realizar a exclusão de fonemas (*pruning*) presentes no banco de fala que contenham amostras degradadas por um ou mais cliques.

Um *pruning* apropriado de unidades tem sido proposto por diversos autores como ferramenta para reduzir o tamanho de bancos de fala. Tal procedimento auxilia na remoção de unidades atípicas que podem ter sido alvo de erros de anotação (rotulagem) ou de pobre articulação na gravação. Essa abordagem também permite remover unidades tão comuns que não tenham distinção significativa entre os diversos fonemas candidatos a serem usados na síntese [55], [56].

O presente trabalho propõe o uso da informação da existência de cliques como mais um critério para excluir fonemas, reduzindo, por conseguinte, o tamanho do banco e conseqüentemente a complexidade computacional da seleção automática durante a síntese propriamente dita.

É importante mencionar que, para as técnicas de detecção automática aqui consi-

deradas, a adoção de um procedimento de *pruning* se tornaria inviável. Essa inviabialidade deve-se a um grande número de detecções incorretas que ocorrem nas técnicas de detecção propostas. Entretanto, caso seja realizado um procedimento de detecção manual de cliques, o tratamento através de *pruning* se constituiria em uma solução apropriada.

6.2 Mascaramento

Outra maneira de se reduzir o efeito audível de eventuais cliques é simplesmente através do procedimento de mascaramento (apresentado na Seção 1.2). Tal abordagem consistiria em, por exemplo, adicionar um outro sinal à fala de maneira a reduzir a audibilidade dos cliques. Um exemplo seria a adição de uma música de fundo à fala sintetizada. A utilização de tal fundo mascara consideravelmente os distúrbios audíveis originados tanto por descontinuidades resultantes da concatenação quanto por cliques naturais. Tal procedimento pode ser útil em determinadas aplicações; entretanto, existem aplicações nas quais um sinal mascarador não é desejado.

6.3 Suavização dos Cliques

A suavização de cliques consiste em outra abordagem de processamento que objetiva reduzir a audibilidade dos cliques. O presente trabalho propõe, para redução de audibilidade, a ponderação de um segmento de fala com clique por uma função janela da seguinte forma:

$$h(n) = 1 - \alpha_1 w_{\mathrm{h}}(n) + \alpha_2 w'_{\mathrm{h}}(n), \qquad 0 \leq n \leq 4P, \tag{6.1}$$

onde $w_{\mathrm{h}}(n)$ é uma janela de Hanning com $4P + 1$ coeficientes, $w'_{\mathrm{h}}(n)$, um sinal com P amostras iniciais e finais nulas, tendo $2P + 1$ amostras centrais correspondendo a uma outra função janela de Hanning; $0 \leq \alpha_1 \leq 1$ e $0 \leq \alpha_2 \leq 1$ são parâmetros que ajustam o peso de cada uma das janelas envolvidas em $h(n)$, respectivamente. É ainda importante considerar $\alpha_1 > \alpha_2$ para o presente caso.

A Fig. 6.1 ilustra uma janela de suavização para $\alpha_1 = 1$, $\alpha_2 = 0{,}1$ e $P = 180$.

6.4 Extrapolação

O presente trabalho pretende avaliar o desempenho da técnica de extrapolação no processamento das amostras consideradas como degradadas por cliques naturais. A técnica aqui adotada baseia-se no trabalho de Kauppinen [57]. Em tal trabalho, o autor apresenta,

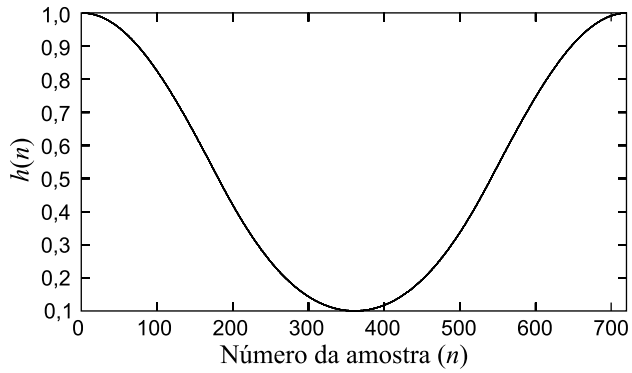

Fig. 6.1: Janela de suavização com $\alpha_1 = 1$, $\alpha_2 = 0{,}1$ e $P = 180$.

como aplicação da técnica de extrapolação, a eliminação de ruídos impulsivos presentes em gravações de áudio antigas.

O trabalho de Kauppinen, que permite reconstruir desejados segmentos de sinais de áudio, apresenta excelente eficácia na extrapolação, sendo, portanto, adequado para reconstruir também segmentos de fala degradados por cliques.

A técnica sugerida por Kauppinen inicia-se com a modelagem de um sinal através da estimação de parâmetros de um modelo autorregressivo via minimização do erro de predição. Considerando-se **h'** o vetor de coeficientes do filtro de predição correspondente ao segmento do sinal $x(n)$ que precede a seção degradada, a equação de extrapolação no sentido direto (*forward*) é dada por

$$x'(n) = \sum_{k=1}^{M} h'(k)x(n-k), \tag{6.2}$$

onde $x'(n)$ representa a *n*-ésima amostra extrapolada no sentido direto. Após calcular-se o valor da primeira amostra de (6.2) e considerar-se tal amostra como conhecida, essa equação pode ser reutilizada para se obter a próxima amostra, e assim por diante.

Considerando-se **h"** o vetor de coeficientes de um filtro de predição relativo ao segmento de $x(n)$ sucessor à seção degradada, a equação de extrapolação no sentido reverso (*backward*) é expressa por

$$x''(n) = \sum_{k=1}^{M} h''(k)x(n+k), \tag{6.3}$$

onde $x''(n)$ representa a *n*-ésima amostra extrapolada no sentido reverso.

Após a extrapolação em ambos os sentidos, as amostras consideradas como degradadas são substituídas por uma média ponderada dos sinais extrapolados nos sentidos direto e reverso de acordo com

$$\widehat{x}(n) = w(n)x'(n) + (1 - w(n))x''(n), \tag{6.4}$$

onde $w(n)$, função de ponderação, deve satisfazer à condição $0 \leq w(n) \leq 1$.

Kauppinen propõe o uso de uma função de ponderação definida por

$$w(n) = \begin{cases} 1 - \frac{1}{2}(2u(n))^a, & u(n) \leq 1/2 \\ \frac{1}{2}(2 - 2u(n))^a & u(n) > 1/2, \end{cases} \tag{6.5}$$

com

$$u(n) = \frac{n - n_s}{n_e - n_s}, \tag{6.6}$$

onde n_s e n_e são respectivamente as amostras inicial e final da região a ser restaurada [58].

A Fig. 6.2 ilustra as funções $w(n)$ e $1 - w(n)$ para extrapolação nos sentidos direto e reverso, respectivamente, considerando-se diferentes valores de a.

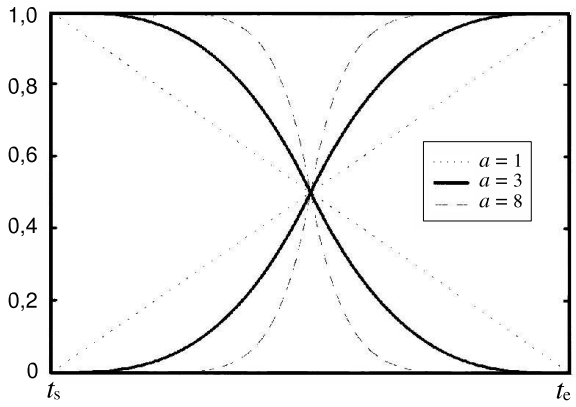

Fig. 6.2: Funções de ponderação $w(n)$ e $1 - w(n)$.

Visando-se ilustrar o procedimento proposto por Kauppinen, a Fig. 6.3 mostra um segmento de um sinal de fala contendo um clique natural. A região com clique é submetida ao procedimento de extrapolação. O segmento extrapolado no sentido direto é apresentado na Fig. 6.4, o reverso, na Fig. 6.5. O sinal resultante é então mostrado na Fig. 6.6. Os espectrogramas correspondentes são apresentados nas Figs. 6.7 e 6.8. Verifica-se, por análise temporal e espectral, que o clique foi eliminado. Assim, tal clique torna-se imperceptível em termos audíveis.

6.5 Conclusões

Neste capítulo, técnicas de processamento de cliques naturais baseadas em mascaramento, *pruning* de fonemas com cliques, suavização e extrapolação são descritas. O objetivo dessas técnicas consiste em restaurar segmentos de fala considerados como degradados por cliques visando eliminar os sons de estalos que os caracterizam. Essas técnicas de processamento aqui discutidas são comparadas quanto à eficácia no Capítulo 7.

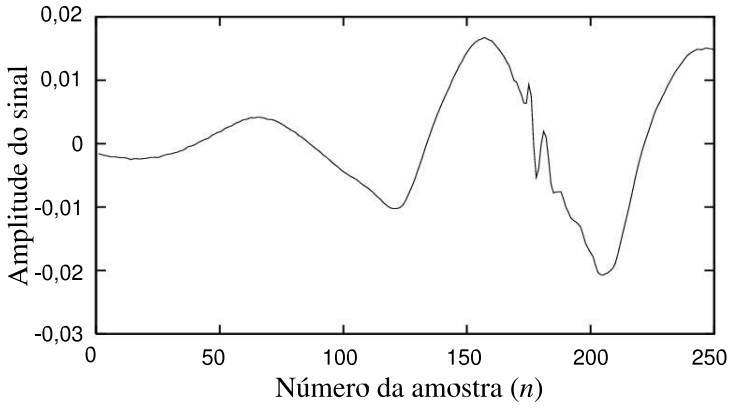

Fig. 6.3: Segmento de um sinal de fala com clique antes da extrapolação.

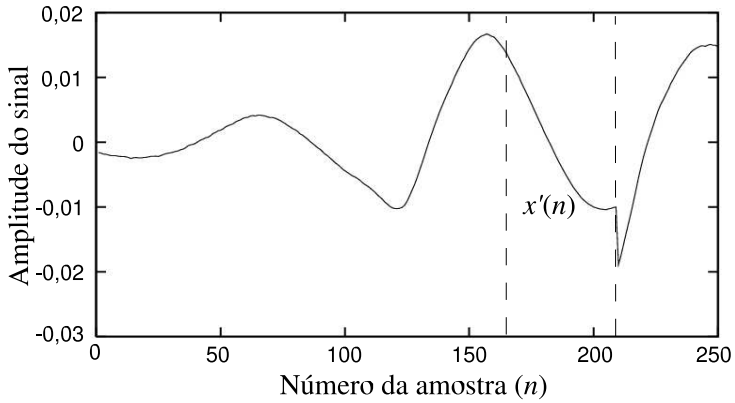

Fig. 6.4: Segmento do sinal extrapolado no sentido direto.

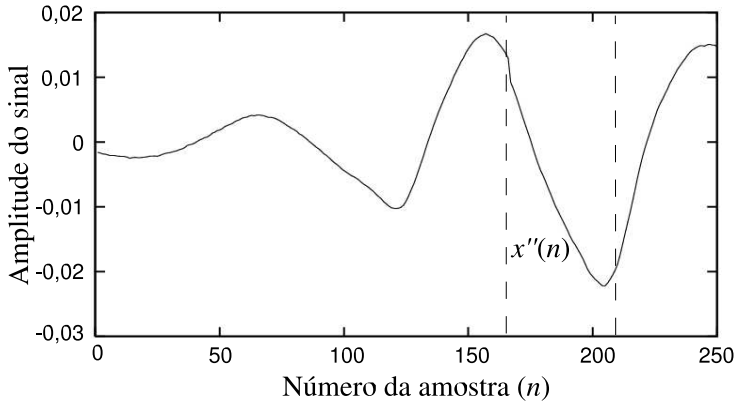

Fig. 6.5: Segmento do sinal extrapolado no sentido reverso.

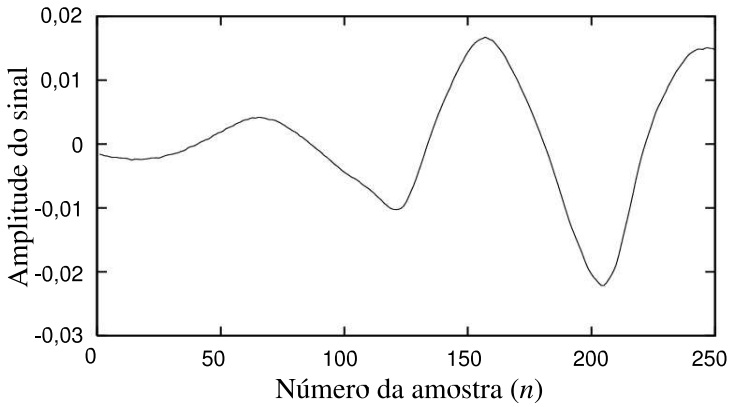

Fig. 6.6: Segmento após o processo completo de extrapolação.

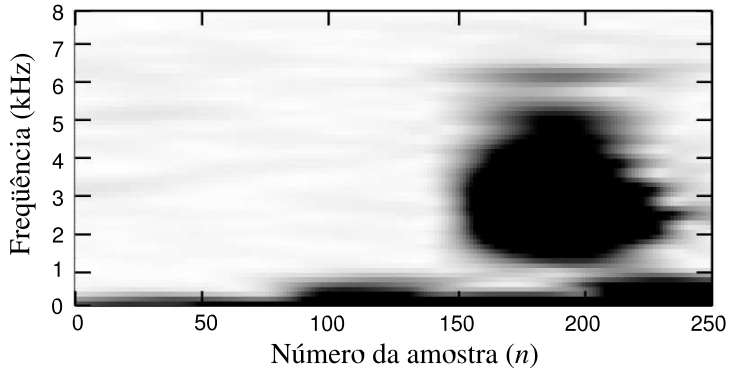

Fig. 6.7: Espectrograma do segmento do sinal de fala antes da extrapolação.

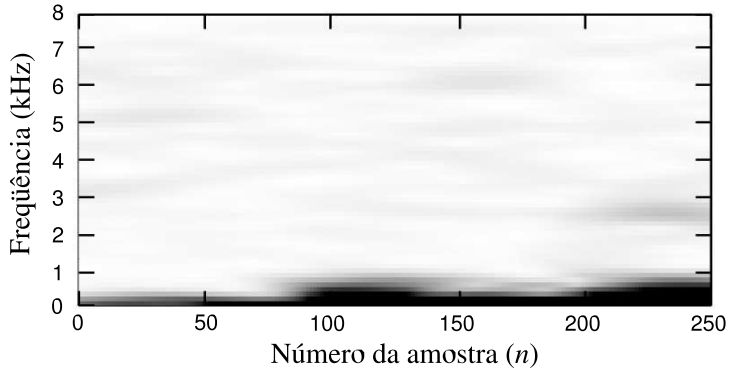

Fig. 6.8: Espectrograma após a extrapolação.

Capítulo 7

Experimentos para Avaliar as Técnicas de Processamento

O desempenho de algumas técnicas para processamento de cliques naturais apresentadas no Capítulo 6 é avaliado neste capítulo. Dentre as técnicas apresentadas, o *pruning* de fonemas com cliques e o mascaramento não são aqui avaliados quanto ao desempenho. A técnica de *pruning* não é avaliada neste trabalho, pois resultaria na eliminação de grande parte dos fonemas no banco, considerando-se o grande número de detecções incorretas das técnicas de detecção automática propostas. A técnica de mascaramento não é também analisada visto que, em grande parte das aplicações em síntese de fala, a adição de um sinal mascarador é indesejável. Dessa forma, o presente capítulo se restringe à avaliação das técnicas de processamento baseadas em suavização e extrapolação.

7.1 Suavização dos Cliques

Para se avaliar a técnica de suavização de cliques, propõe-se o uso de uma medida perceptual definida por

$$\text{IS} = \frac{n_\text{i}}{n_\text{c}} \times 100, \tag{7.1}$$

onde IS representa o índice de suavização, n_i é o número de cliques detectados transformados em inaudíveis e n_c, o número total de cliques detectados.

Para segmentos vozeados, o procedimento de suavização é realizado utilizando-se $\alpha_1 = 1$, $\alpha_2 = 0{,}1$ e $P = 60$. Para fala não-vozeada, considera-se $\alpha_1 = 1$, $\alpha_2 = 0$ e $P = 180$. Após a suavização, o sinal de fala (contendo cliques naturais) é avaliado por dois ouvintes especialistas em processamento de fala utilizando um fone de ouvido de alta qualidade. Cada ouvinte apontou os instantes onde um som de clique era audível. Posteriormente, tais

instantes foram comparados com os instantes obtidos pela técnica automática de detecção de cliques. O IS conjunto obtido na avaliação é igual a $93{,}75\%$, indicando um desempenho satisfatório.

Tal técnica, apesar de reduzir a audibilidade dos cliques deteriora um pouco a qualidade da fala, atribuindo-lhe uma característica tipo "robotizada". Visando superar tal problema, realizaram-se estudos e experimentos com a técnica de extrapolação.

7.2 Extrapolação

Para se avaliar a técnica de extrapolação apresentada no Capítulo 6, realizam-se testes subjetivos baseados em escuta. Para tal, adotam-se dois testes consagrados na literatura: o ACR (*absolute category rating*) [1] e o CCR (*comparative category rating*) [1].

7.2.1 Teste ACR [1]

Neste método, os ouvintes julgam a qualidade das sentenças de acordo com a escala de notas (escores) apresentada na Tabela 7.1. A média dos escores de todos os ouvintes para uma determinada condição é denominada *mean opinion score* (MOS) [1].

Tabela 7.1: Escala de escores para o teste ACR

Qualidade da fala	Escore
Excelente	5
Boa	4
Satisfatória	3
Pobre	2
Péssima	1

O experimento deve adotar como condições de referência a unidade de ruído modulado (*modulated noise reference unit* – MNRU), conforme exigido em [1]. O uso de uma referência, tal como a MNRU, possibilita a comparação de experimentos em diferentes laboratórios ou em diferentes instantes de tempo no mesmo laboratório. A referência MNRU introduz degradações controladas nos sinais de fala e é definida pela seguinte expressão:

$$y(n) = x(n)[1 + 10^{-\mathrm{SNR}/20} r(n)] \tag{7.2}$$

onde $x(n)$ é o sinal inalterado, $r(n)$, o ruído randômico, SNR, a razão sinal-ruído, e $y(n)$, o sinal de fala corrompido por um ruído modulado [59]. Dessa forma, o sinal $y(n)$ consiste no sinal de referência MNRU, a ser utilizado durante os testes de escuta.

7.2.2 Teste CCR [1]

Em um teste CCR, os ouvintes escutam um par de sentenças, sendo uma processada e outra não, em ordem aleatória. A qualidade da segunda sentença é comparada com a da primeira de acordo com escala de notas apresentada na Tabela 7.2. A média dos escores de todos os ouvintes é denominada *comparison mean opinion score* (CMOS) [1].

Tabela 7.2: Escala de escores para o teste CCR

Qualidade da 2^a comparada com a 1^a	Escore
Muito melhor	3
Melhor	2
Pouco melhor	1
Iguais	0
Pouco pior	-1
Pior	-2
Muito pior	-3

É importante mencionar que cuidados devem ser tomados na análise de um experimento CCR. Como parte dos pares de sentenças são apresentados na ordem processada/não-processada e outra parte na ordem contrária, uma média simples de todos os escores resultaria em um valor CMOS aproximadamente nulo. Deve-se, portanto, antes de calcular o valor médio dos escores, multiplicar o escore por -1 caso a ordem de apresentação de um par de sentenças seja processada/não-processada, mantendo o mesmo escore para ordem reversa. O objetivo de alterar o valor do escore consiste em comparar sempre a qualidade da sentença processada em relação à original.

7.2.3 Experimentos

Um conjunto de 10 sentenças é submetida à técnica de detecção de cliques naturais baseada em decomposição *wavelet*. Após a etapa de detecção, as sentenças são processadas segundo a técnica de extrapolação. Tais sentenças apresentam duração média de 6 s e possuem no total 176 cliques naturais. A técnica de extrapolação é aplicada sobre segmentos de fala com duração de 61 amostras tomando como referência 300 amostras dos segmentos anterior e posterior ao restaurado. Os coeficientes de um filtro preditor são obtidos tanto para o segmento antecessor quanto para o sucessor considerando-se uma análise AR de ordem 150. Cada sentença original é também processada visando produzir os sinais de referência MNRU com SNR de 24, 32 e 40 dB.

Para o teste ACR, um conjunto de 23 ouvintes escuta cada uma das 10 sentenças processadas de 4 maneiras distintas (3 tipos de MNRU mais extrapolação) e as 10 sentenças

originais, totalizando 50 sentenças. Os escores MOS obtidos para a técnica de extrapolação, para o sinal original e para MNRU com diferentes SNR são mostrados na Tabela 7.3.

Tabela 7.3: MOS resultante para as diferentes técnicas avaliadas por 23 ouvintes

Técnica	MOS
MNRU - SNR = 24 dB	2,19
MNRU - SNR = 32 dB	3,30
MNRU - SNR = 40 dB	4,05
Original	4,15
Extrapolação	4,31

O conjunto de 23 ouvintes engloba quatro pessoas conhecedoras do assunto "cliques" e 19 sem conhecimento (ouvintes ingênuos ou inexperientes). Para o subgrupo de pessoas conhecedoras (experientes), obtiveram-se os resultados apresentados na Tabela 7.4. Para os demais ouvintes, o MOS resultante é apresentado na Tabela 7.5.

Tabela 7.4: MOS resultante para as diferentes técnicas avaliadas por quatro ouvintes experientes

Técnica	MOS
MNRU - SNR = 24 dB	2,35
MNRU - SNR = 32 dB	3,37
MNRU - SNR = 40 dB	4,00
Original	3,98
Extrapolação	4,48

Tabela 7.5: MOS resultante para as diferentes técnicas avaliadas por ouvintes inexperientes

Técnica	MOS
MNRU - SNR = 24 dB	2,15
MNRU - SNR = 32 dB	3,28
MNRU - SNR = 40 dB	4,06
Original	4,18
Extrapolação	4,27

Para o teste CCR, as sentenças originais e processadas via extrapolação são comparadas de acordo com a Tabela 7.2. Verifica-se que as sentenças processadas apresentam qualidade entre igual e um pouco superior em relação às originais (CMOS = 0,25). Para os ouvintes conhecedores do assunto cliques, tem-se um CMOS de 1,15. Para ouvintes ingê-

nuos, o CMOS obido é de 0,06. A Fig. 7.1 apresenta um histograma dos valores de CMOS obtidos para os 23 ouvintes considerados.

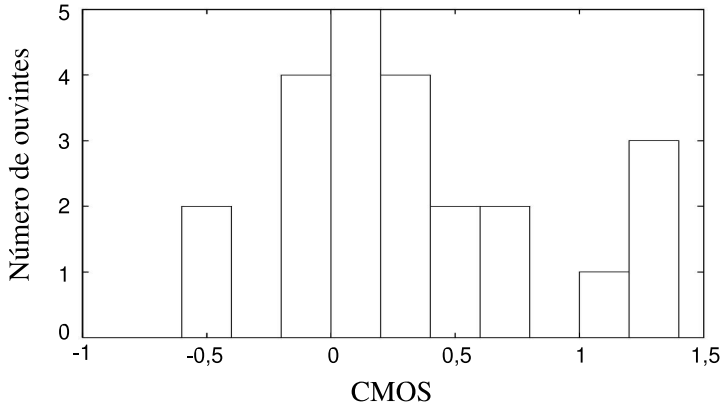

Fig. 7.1: Histograma de CMOS.

As 50 sentenças adotadas nos experimentos são apresentadas no CD disponível em anexo. Dentre as sentenças adotadas nos experimentos está a Frase01_original.wav, mencionada no Capítulo 3. Percebe-se, escutando tal arquivo, a exclusão de 10 dentre os 11 cliques presentes na sentença original (o único clique não eliminado é o de número 6). Assim sendo, verifica-se a eficácia dos procedimentos de detecção via análise *wavelet* e tratamento via extrapolação propostos.

É importante destacar que os ouvintes são instruídos para avaliar a qualidade da fala como um todo, não apenas quanto ao número de cliques. Dessa forma, os testes avaliam os diversos parâmetros que exercem influência sobre a qualidade da fala, incluindo a naturalidade. Assim, os escores MOS e CMOS, advindos dos testes, comprovam que o procedimento proposto não afeta a naturalidade da fala.

Conclui-se, desta forma, que uma alternativa eficaz para eliminar cliques naturais presentes em bancos de fala desenvolvidos para sistemas de síntese concatenativa consiste em se adotar as ferramentas *wavelet* de detecção e um processamento baseado em extrapolação.

7.3 Conclusões

Este capítulo estabelece um comparativo entre métodos de processamento de cliques naturais, concentrando-se nas técnicas de suavização e extrapolação. A suavização pode atribuir uma característica do tipo “robotizada” à fala e, por isso, não é vantajosa. A extrapolação, por sua vez, elimina o efeito audível dos cliques sem causar degradações audíveis à qualidade da fala, conforme pode-se observar nas frases processadas por extrapolação e contidas no CD disponível em anexo. Tais frases, tratadas por extrapolação, são avaliadas

através dos testes ACR e CCR, ambos baseados em procedimento de escuta. Os escores MOS e CMOS, advindos dos testes, indicam também um desempenho satisfatório na eliminação de cliques naturais de todo o procedimento englobando detecção via análise *wavelet* e processamento por extrapolação.

Capítulo 8

Comentários e Conclusões Finais

Em um sistema de conversão texto-fala, ocorre inicialmente a transformação de um texto tomado como entrada em informações lingüísticas. Esta é a etapa de processamento lingüístico. Após tal etapa, a informação lingüística obtida é convertida em fala considerando-se técnicas de processamento de sinais. Dentre as técnicas existentes, a concatenativa é a que tem produzido fala sintética de melhor qualidade, tanto em termos de naturalidade quanto de inteligibilidade.

Na síntese concatenativa, ocorre a conexão de segmentos de fala armazenados em um banco previamente gravado. A qualidade desse banco exerce considerável influência sobre o resultado final da síntese. Tal *corpus* é gravado em um estúdio apropriado, por um locutor treinado. Entretanto, a fala humana, inclusive a de um locutor profissional, pode apresentar eventuais degradações causadas tanto por variações de *pitch* e energia quanto por cliques naturais.

Cliques são descontinuidades que ocorrem na fala humana, manifestando-se na forma de estalos de pequena amplitude. São praticamente imperceptíveis na fala corrente; entretanto, podem prejudicar significativamente a qualidade da fala sintética, especialmente quando associados a descontinuidades resultantes do processo de concatenação.

Uma contribuição do presente trabalho consiste em estudar cliques naturais (involuntários), nunca anteriormente explorados. A pesquisa começa tentando-se explicar como tais cliques são produzidos. É apresentada também uma descrição estatística dos cliques naturais existentes em um banco de fala, com 45 minutos de duração, gravado por um locutor profissional. Essa descrição é relevante, pois permite que outros pesquisadores, interessados neste estudo, possam realizar a geração artificial de cliques. Tal geração pode ser útil para se estudar técnicas de detecção sem a necessidade de realizar previamente a anotação manual de todo um banco de fala, tarefa essa que demanda um considerável tempo de execução.

O presente trabalho ainda apresenta como contribuição o desenvolvimento de um processo para melhorar a qualidade de bancos de fala desenvolvidos para síntese concatenativa. Tal melhoria é alcançada através da eliminação e/ou atenuação dos possíveis cliques

involuntários existentes no banco. Para tal, propõe-se inicialmente a determinação da localização temporal dos cliques e posteriormente, um processamento. Tal procedimento é realizado de maneira *off-line*, não afetando a complexidade computacional da síntese propriamente dita.

Um conjunto de técnicas de detecção são propostas e avaliadas quanto ao desempenho para detecção. Estudaram-se técnicas baseadas em filtragem inversa, derivada de quarta ordem, modelagem do aparelho auditivo humano, transformada *wavelet* e redes neurais artificiais. Dentre as técnicas consideradas, a baseada em decomposição *wavelet* foi a que apresentou a melhor eficácia para se obter a localização de cliques naturais.

Outras técnicas foram avaliadas visando um processamento dos cliques detectados. A técnica de suavização, apesar de reduzir e/ou eliminar a audibilidade dos cliques, conduz a uma certa degradação do sinal de fala, tornando-a com um aspecto "robotizado". A técnica de extrapolação, por sua vez, apresenta capacidade de eliminar a audibilidade dos cliques sem degradar a qualidade do sinal de fala.

A partir de um teste perceptual realizado sobre os bancos de fala original e processado, verifica-se que houve uma melhoria na qualidade do sinal de fala relativo ao sinal original quando realizada a detecção baseada em *wavelet* seguida de um procedimento de extrapolação. Através de um experimento CCR (*comparison category rating*) obteve-se um valor de CMOS (*comparison mean opinion score*) de 0,25 e através de um ACR (*absolute category rating*), obteve-se um MOS (*mean opinion score*) de 4,15 para o banco original e de 4,31 para o banco processado. Esses escores indicam a melhoria da qualidade do banco de fala previamente gravado.

Uma sugestão para trabalhos futuros consiste em aplicar o procedimento proposto para todo um banco de fala desenvolvido para síntese, objetivando realizar um teste perceptual que permita analisar a melhoria na fala sintética propriamente dita. Nesse caso, adotar-se-ia também a técnica de detecção *wavelet* associada à extrapolação. O objetivo seria comparar a qualidade da fala sintética antes e após a realização do procedimento proposto. Outro estudo que deve ser ainda realizado é a análise de outros bancos de fala oriundos de diferentes locutores visando verificar se a estatística descrita para o locutor considerado varia muito entre locutores.

É importante mencionar que tal técnica, apesar de aqui ser aplicada para melhoria de bancos de fala desenvolvidos para aplicações de síntese, pode ser também adotada para síntese de áudio, melhoria de bancos de fala e/ou de gravações desenvolvidas para aplicações diversas, incluindo músicas gravadas aproveitando a voz de cantores profissionais, dentre outros. Sugere-se, portanto, visando obter a melhor qualidade possível em uma gravação, adotar a técnica proposta para excluir cliques presentes em tal gravação.

Deve-se ainda enfatizar que o procedimento proposto não causa danos à naturalidade da fala gravada, apesar de ser aplicado sobre distúrbios produzidos naturalmente pelo

aparelho fonador humano. Podemos sim pensar em fala gravada com qualidade superior à fala natural. Uma maneira de obter tal melhoria de qualidade consiste em realizar o tratamento das degradações conhecidas como cliques naturais.

Referências Bibliográficas

[1] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," ago. 1996.

[2] J. L. Flanagan, "Voices of men and machines," *Journal of the Acoustical Society of America*, vol. 51, no. 5, pp. 1375–1387, maio 1972.

[3] S. Lemmetty, "Review of speech synthesis technology," Master's thesis, Helsinki University of Technology, Departament of Electrical and Communications Engineering, Espoo, Finland, 1999.

[4] P. A. A. Esquef, "Restauração de sinais de áudio degradados por ruído impulsivo," Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 1999.

[5] E. Zwicker e H. Fastl, *Psycho-acoustics: Facts and Models*, 2^a ed. Germany: Springer, 1999.

[6] S. Somers. The mysterious loudness control: What does it do? [Online]. Disponível em: http://www.extron.com/company/archive.asp?id=loudnesscontrol_ts. Acesso em 29 jul. 2005

[7] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5^a ed. San Diego, USA: Academic Press, 2003.

[8] F. S. Pacheco, "Técnicas de processamento de sinais para alteração de parâmetros prosódicos aplicadas a um sistema de conversão texto-fala para a língua portuguesa falada no brasil," Dissertação de mestrado, UFSC, Florianópolis, SC, 2001.

[9] B. H. Juang e T. Chen, "The past, present, and future of speech processing," *IEEE Signal Processing Magazine*, vol. 15, no. 3, pp. 24–48, maio 1998.

[10] H. Traunmüller. Wolfgang von Kempelen's speaking machine and its successors. [Online]. Disponível em: http://www.ling.su.se/staff/hartmut/kemplne.htm. Acesso em 8 jul. 2005

[11] D. H. Klatt, "Review of text-to-speech conversion for English," *Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, set. 1987.

[12] P. Rubin e L. Goldstein. Pattern Playback. [Online]. Disponível em: http://www.haskins.yale.edu/haskins/MISC/PP/pp.html. Acesso em 11 jul. 2005

[13] J. R. Deller Jr., J. H. L. Hansen e J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York, USA: IEEE Press, 2000.

[14] J. H. Page e A. P. Breen, "The Laureate text-to-speech system - architecture and applications," *BT Technology Journal*, vol. 14, no. 1, pp. 57–67, jan. 1996.

[15] T. Dutoit, "High-quality text-to-speech synthesis: an overview," *Journal of Electrical & Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis*, vol. 17, no. 1, pp. 25–37, 1997. [Online]. Disponível em: http://tcts.fpms.ac.be/synthesis/introtts.html

[16] M. Kitai, K. Hakoda, S. Sagayama *et al.*, "ASR and TTS telecommunications applications in Japan," *Speech Communication*, vol. 23, no. 1-2, pp. 17–30, out. 1997.

[17] L. R. Rabiner, "Applications of voice processing to telecommunications," *Proceedings of the IEEE*, vol. 82, no. 2, pp. 199–228, fev. 1994.

[18] M. Contolini, K. Stoimenov e J.-C. Junqua, "Voice technologies for telephony services," in *Proc. IEEE Consumer Communications and Networking Conference (CCNC'04)*, Las Vegas, USA, jan. 2004, pp. 662–664.

[19] Y. Sagisaka, "Speech synthesis from text," *IEEE Communications Magazine*, vol. 28, no. 1, pp. 35–55, jan. 1990.

[20] J. J. Ohala, "A probable case of clicks influencing the sound patterns of some european languages," *Phonetica*, vol. 52, no. 3, pp. 160–170, 1995.

[21] C.-H. Wu e J.-H. Chen, "Automatic generation of synthesis units and prosodic information for chinese concatenative synthesis," *Speech Communication*, vol. 35, pp. 219–237, 2000.

[22] J. Dubois, M. Giacomo, L. Guespin, C. Marcellesi, J.-B. Marcellesi e J.-P. Mevel, *Dicionário de Lingüística*. São Paulo: Cultrix, 1973.

[23] F.-C. Chou, C.-Y. Tseng e L.-S. Lee, "A set of corpus-based text-to-speech synthesis technologies for Mandarin chinese," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 481–494, out. 2002.

[24] H. Kawai, T. Toda, J. Ni *et al.*, "Ximera: a new TTS from ATR based on corpus-based technologies," in *Proc. 5th ISCA Speech Synthesis Workshop (SSW5'04)*, Pittsburg, USA, jun. 2004, pp. 179–184.

[25] K. Ng. (1998) Survey of data-driven approaches to speech synthesis. [Online]. Disponível em: http://citeseer.ist.psu.edu/ng98survey.html. Acesso em 15 jul. 2005

[26] A. J. Hunt e A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, vol. I, Atlanta, GA, maio 1996, pp. 373–376.

[27] X. Huang, A. Acero e H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River: Prentice Hall, 2001.

[28] Y. Stylianou, "Assessment and correction of voice quality variabilities in large speech databases for concatenative speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'*, Phoenix, USA, mar. 1999, pp. 377–380.

[29] J. Lindh, "Acoustic and perceptual analysis of discontinuities in two TTS concatenation systems," in *Proc. Fonetik 2004*, Estocolmo, SWE, maio 2004.

[30] M. V. Nicodem, R. Seara e F. S. Pacheco, "Reducing the natural click effect within database for high quality corpus-based speech synthesis," in *Proc. The Eighth International Symposium on Signal Processing and Its Applications (ISSPA'05)*, Sydney, AU, ago. 2005, pp. 607–610.

[31] E. Pennisi, "The first language?" *Science*, vol. 303, pp. 1319–1320, fev. 2004.

[32] W. J. Hardcastle e J. Laver, *The Handbook of Phonetic Sciences*. Cambridge, USA: Blackwell, 1997.

[33] P. Ladefoged e A. Traill, "Clicks and their accompaniments," *Journal of Phonetics*, vol. 22, pp. 33–64, 1994.

[34] K. N. Stevens, *Acoustic Phonetics*. Cambridge, USA: MIT Press, 1998.

[35] J. Clark e C. Yallop, *An Introduction to Phonetics and Phonology*, 2^a ed. Oxford, UK: Blackwell, 1995.

[36] J. J. Ohala. Emergent stops. Unpublished. [Online]. Disponível em: http://trill.berkeley.edu/users/ohala/papers

[37] M. V. Nicodem, R. Seara e F. S. Pacheco, "Detecção e suavização de cliques naturais em bancos de fala visando síntese concatenativa de alta qualidade," in *Anais do XXII Simpósio Brasileiro de Telecomunicações (SBrT'05)*, Campinas, SP, set. 2005, pp. 429–433.

[38] S. J. Godsill e P. J. Rayner, *Digital Audio Restoration - A Statistical Model Based Approach*. London, UK: Springer-Verlag, 1998.

[39] S. V. Vaseghi e P. J. W. Rayner, "Detection and supression of impulsive noise in speech communication systems," *IEE Proceedings*, vol. 137, no. 1, pp. 38–46, fev. 1990.

[40] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, abr. 1975.

[41] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio signals," in *Proc. IEEE International Conference on Digital Signal Processing (DSP'02)*, vol. 2, Pine Mountain, USA, jul. 2002, pp. 967–970.

[42] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 115–132, jan. 1994.

[43] P. A. A. Esquef, M. Karjalainen e V. Välimäki, "Detection of clicks in audio signals using warped linear prediction," in *Proc. IEEE International Conference on Digital Signal Processing (DSP'02)*, vol. 2, Santorini, Greece, jul. 2002, pp. 1085–1088.

[44] P. M. da Silveira, "Identificação e localização de faltas utilizando análise por decomposição wavelet para relés de linhas de transmissão," Dissertação de mestrado, UFSC, Florianópolis, SC, 2001.

[45] C. S. Burrus, R. A. Gopinath e H. Guo, *Introduction to Wavelets and Wavelet Transforms: a primer*. Upper Saddle River, USA: Prentice-Hall, 1998.

[46] S. Mallat e W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 617–643, mar. 1992.

[47] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego: Academic Press, 1997.

[48] O. D. Filho, "Utilização da transformada wavelet para caracterização de distúrbios na qualidade da energia elétrica," Dissertação de mestrado, USP, São Carlos, SP, 2003.

[49] Wikipedia. Wavelet. [Online]. Disponível em: http://en.wikipedia.org/wiki/Wavelet_transform. Acesso em 31 out. 2005

[50] A. V. Oppenheim, A. S. Willsky e S. H. Nawab, *Signals & Systems*. Upper Saddle River: Prentice-Hall, 1996.

[51] A. Czyzewski, "Some methods for detection and interpolation of impulsive distortions in old audio recordings," in *Proc. IEEE Workshop on Apliccations of Signal Processing to Audio and Acoustics (ASSP'95)*, New York, USA, out. 1995, pp. 139–142.

[52] F. M. de Azevedo, L. M. Brasil e R. C. L. de Oliveira, *Redes Neurais com Aplicações em Controle e em Sistemas Especialistas*. Florianópolis: Bookstore, 2000.

[53] S. Haykin, *Redes Neurais*, 2^a ed. Porto Alegre: Bookman, 2001.

[54] S. Nissen e E. Nemerson. Fast artificial neural network library. [Online]. Disponível em: http://leenissen.dk/fann/. Acesso em 02 nov. 2005

[55] A. Black e P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH'97)*, vol. 2, Rhodes, Greece, set. 1997, pp. 601–604.

[56] Y. Zhao, M. Chu, H. Peng e E. Chang, "Custom-tailoring TTS voice font - keeping the naturalness when reducing database size," in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH'03)*, Geneva, set. 2003, pp. 2957–2960.

[57] I. Kauppinen e K. Roth, "Audio signal extrapolation - theory and applications," in *Proc. 5th International Conference on Digital Audio Effects (DAFx'02)*, Hamburg, Germany, set. 2002, pp. 105–110.

[58] I. Kauppinen e J. Kauppinen, "Reconstruction method for missing or damaged long portions in audio signal," *Journal of the Audio Engineering Society*, vol. 50, no. 7-8, pp. 594–602, jul. 2002.

[59] ITU-T Recommendation P.810, "Modulated noise reference unit (MNRU)," fev. 1996.

Anexo 1

Tabela 8.1: Número de cliques existentes em fones e em pausas do banco

Fone	Número de cliques	Porcentagem do número total de cliques
[t]	189	6,25
Pausa longa	166	5,49
[a]	161	5,32
Silêncio	160	5,29
[n]	154	5,09
[l]	137	4,53
[m]	122	4,03
[r]	107	3,54
[k]	97	3,20
[S]	93	3,08
Pausa curta	91	3,01
[ɾ]	90	2,98
Pausa para vírgula	88	2,91
[i][1]	87	2,88
[i][2]	81	2,68
[u][3]	81	2,68
Pausa para respiração	75	2,48
Bucal	73	2,41
[ˈa]	67	2,22
Pausa de ponto final	65	2,15
[d]	51	1,69

[1]Considera-se para essa estatística apenas os fones [i] que correspondem ao grafema "e" em final de palavra.

[2]Nesse caso, a estatítica exclui os cliques que ocorrem em fones [i] substituindo o grafema "e" no final de palavras.

[3]Para tal estatística são analisados somente os fones [u] correspondentes ao grafema "o" em final de palavra.

[ʤ]	51	1,69
[s]	51	1,69
[ʧ]	45	1,49
[e]	44	1,46
[ˈi]	41	1,36
[ˈã]	32	1,06
[ĩ]	31	1,03
[ˈũ]	31	1,03
[w̃]	30	0,99
[o]	30	0,99
[z]	29	0,96
[ˈe]	28	0,93
[ẽ]	28	0,93
[ʎ]	27	0,89
[j̃]	27	0,89
[õ]	24	0,79
[g]	24	0,79
[w]	23	0,76
[ˈẽ]	20	0,66
[ˈo]	20	0,66
[ɛ]	18	0,60
[ɔ]	17	0,56
[ŋ]	15	0,50
[j]	15	0,50
[u]	14	0,46
[ˈĩ]	12	0,40
[ˈõ]	9	0,30
[ˈu]	9	0,30
[f]	9	0,30
[v]	7	0,23
[ã]	5	0,17
Pausa de ponto e vírgula	5	0,17
[ũ]	4	0,13
[p]	4	0,13
Pausa de dois pontos	4	0,13
Pausa de pergunta	2	0,07

[ʃ]	2	0,07
[ʒ]	1	0,03
[R]	1	0,03
[b]	0	0