Sarah Negreiros de Carvalho

Estudo de um Sistema de Conversão Texto-Fala Baseado em HMM

CAMPINAS
2013

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

SARAH NEGREIROS DE CARVALHO

ESTUDO DE UM SISTEMA DE CONVERSÃO TEXTO-FALA BASEADO EM HMM

Orientador: Prof. Dr. Fábio Violaro

Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para a obtenção do título de Mestra em Engenharia Elétrica. Área de concentração: Telecomunicações e Telemática.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO
DEFENDIDA PELA ALUNA SARAH NEGREIROS DE CARVALHO
E ORIENTADA PELO PROF. DR. FÁBIO VIOLARO

_____

CAMPINAS
2013

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

Carvalho, Sarah Negreiros de, 1985-
C253e Estudo de um sistema de conversão texto-fala baseado em HMM / Sarah Negreiros de Carvalho. -- Campinas, SP: [s.n.], 2013.

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

1. Síntese da voz. 2. Modelos ocultos de Markov (HMM). 3. Sistemas de processamento da fala. I. Violaro, Fábio, 1950-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: Study of a HMM-based text-to-speech system
Palavras-chave em Inglês: Voice synthesis, Hidden Markov models (HMM), Speech processing systems
Área de concentração: Telecomunicações e Telemática
Titulação: Mestra em Engenharia Elétrica
Banca examinadora: Carlos Alberto Ynoguti, Renato da Rocha Lopes
Data da defesa: 18-02-2013
Programa de Pós Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidata: Sarah Negreiros de Carvalho

Data da Defesa: 18 de fevereiro de 2013

Título da Tese: "Estudo de um Sistema de Conversão Texto-Fala Baseado em HMM"

Prof. Dr. Fábio Violaro (Presidente): _____

Prof. Dr. Carlos Alberto Ynoguti: _____

Prof. Dr. Renato da Rocha Lopes: _____

Aos meus pais

Agradecimentos

Agradeço a Deus que dispôs os eventos em minha vida de forma que eu pudesse iniciar e concluir este trabalho.

Agradeço ao meu orientador Prof. Dr. Fábio Violaro por seu profissionalismo, dedicação, paciência e solicitude em me receber em seu laboratório e compartilhar comigo seus conhecimentos e experiências acadêmicas.

Agradeço a todos os professores e funcionários da FEEC que desde a época em que eu cursava a graduação me auxiliaram no percurso de aprendizagem e pesquisa na área de Engenharia Elétrica.

Agradeço aos meus pais, Paulo e Cláudia e as minhas irmãs, Patrícia e Flávia, pelo apoio, incentivo e compreensão.

Pelos bons momentos e ajudas, agradeço a todos os amigos, em especial aos do DECOM/FEEC/UNICAMP: Veruska, Fábio, Diana, Harlei, Alice, Pâmela, Ramiro, Cláudio e Adailton.

Agradeço ao Ranniery Maia, pela atenção e esclarecimentos sobre o HTS.

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão da bolsa de Mestrado.

"*Ultima hominis felicitas est in contemplatione veritatis*"

(S. Tomás de Aquino)

Resumo

Com o contínuo desenvolvimento da tecnologia, há uma demanda crescente por sistemas de síntese de fala que sejam capazes de falar como humanos, para integrá-los nas mais diversas aplicações, seja no âmbito da automação robótica, seja para acessibilidade de pessoas com deficiências, seja em aplicativos destinados a cultura e lazer. A síntese de fala baseada em modelos ocultos de Markov (HMM) mostra-se promissora em suprir esta necessidade tecnológica. A sua natureza estatística e paramétrica a tornam um sistema flexível, capaz de adaptar vozes artificiais, inserir emoções no discurso e obter fala sintética de boa qualidade usando uma base de treinamento limitada. Esta dissertação apresenta o estudo realizado sobre o sistema de síntese de fala baseado em HMM (HTS), descrevendo as etapas que envolvem o treinamento dos modelos HMMs e a geração do sinal de fala. São apresentados os modelos espectrais, de pitch e de duração que constituem estes modelos HMM dos fonemas dependentes de contexto, considerando as diversas técnicas de estruturação deles. Alguns dos problemas encontrados no HTS, tais como a característica abafada e monótona da fala artificial, são analisados juntamente com algumas técnicas propostas para aprimorar a qualidade final do sinal de fala sintetizado.

Palavras-chave: Conversão texto-fala, modelos ocultos de Markov (HMM), Sistema de síntese de fala baseado em HMM (HTS).

Abstract

With the continuous development of technology, there is a growing demand for text-to-speech systems that are able to speak like humans, in order to integrate them in the most diverse applications whether in the field of automation and robotics, or for accessibility of people with disabilities, as for culture and leisure activities. Speech synthesis based on hidden Markov models (HMM) shows to be promising in addressing this need. Their statistical and parametric nature make it a flexible system capable of adapting artificial voices, insert emotions in speech and get artificial speech of good quality using a limited amount of speech data for HMM training. This thesis presents the study realized on HMM-based speech synthesis system (HTS), describing the steps that involve the training of HMM models and the artificial speech generation. Spectral, pitch and duration models are presented, which form context-dependent HMM models, and also are considered the various techniques for structuring them. Some of the problems encountered in the HTS, such as the characteristic muffled and monotone of artificial speech, are analyzed along with some of the proposed techniques to improve the final quality of the synthesized speech signal.

***Keywords**: Text-to-Speech System, Hidden Markov Models (HMM), HMM-Based Speech Synthesis System (HTS).*

Lista de figuras

Lista de tabelas

Lista de Abreviaturas e Siglas

DFT	Discrete Fourier Transform
ESPS	Entropic Signal Processing System
FFT	Fast Fourier Transform
GV	Global Variance
HMM	Hidden Markov Model
HSMM	Hidden Semi-Markov Model
HTK	HMM Toolkit
HTS	HMM-Based Speech Synthesis System
LPC	Linear Predictive Coding
LSP	Line Spectrum Pair
MDL	Minimum Description Length
MGC	Mel-Generalized Cepstral
MGLSA	Mel-Generalized Log Spectrum Approximation
MLSA	Mel Log Spectrum Approximation
MSD	Multi-Space Distribution
NLP	Natural Language Processing
PDF	Probability Density Function
SAMPA	Speech Assessment Methods Phonetic Alphabet
SPTK	Speech Signal Processing Toolkit
TD-PSOLA	Time-Domain Pitch-Synchronous Overlap-Add
TTS	Text-to-Speech
UELS	Unbiased Estimator of Log Spectrum

Sumário

1. Introdução

1.1 Contextualização e Motivação

No início da computação, a interação com as máquinas era predominantemente sob a forma escrita, já que essa era a maneira mais eficiente, econômica e confiável para transmitir e armazenar informações. Programadores "ensinavam" o que as máquinas deveriam fazer através de códigos e linhas de comandos.

Até os anos 1970 os computadores eram máquinas enormes, caras, de difícil operação e que realizavam atividades específicas. Atualmente, o avanço tecnológico computacional permitiu a miniaturização dos processadores, a redução dos custos e a possibilidade de interação gráfica, fatores que contribuíram para a popularização dos computadores e para a sua difusão como equipamentos de uso pessoal nas mais diversas áreas e aplicações. Hoje os computadores são essenciais na vida cotidiana de muitas pessoas, tanto no ambiente de trabalho quanto para atividades de lazer.

Com o intuito de facilitar a relação homem-máquina, a tecnologia busca cada vez mais dar inteligência a elas, em um processo de "humanização" das máquinas. Para isso, torna-se indispensável que estas sejam capazes de entender os diferentes modos de comunicação humana, seja na forma escrita, falada ou visual (feita por meio de fisionomias e gestos). Em particular, a comunicação verbal permite o desenvolvimento de diversas aplicações interessantes, tais como: utilização de computadores por deficientes visuais, leitura de e-mails e textos, sistemas automatizados em centrais de atendimento, acesso a bancos de dados por via telefônica, aplicativos de comunicação celular para deficientes vocais e auditivos, entre outras.

De acordo com Chapanis (1975) [1], a troca de informações feita de modo verbal em situações de interação homem-máquina é cerca de duas vezes mais eficiente do que qualquer outra forma de comunicação.

Considerando o exposto, é muito importante habilitar o computador para o uso da linguagem oral. E, para isto, é necessário que ele seja capaz de entender um comando de voz e esteja apto a se expressar oralmente com o usuário para informar ou indagar alguma informação necessária. Em suma, é preciso que a máquina seja capaz de reconhecer e sintetizar fala.

Este trabalho tem por objetivo estudar o processamento de sinais de fala com foco nas características relacionadas aos sistemas de conversão texto-fala (TTS) baseados em Modelos Ocultos de Markov (HMM). Para isso, procurou-se descrever as diversas etapas de processamento de sinais necessárias para sintetizar fala.

Este trabalho foi desenvolvido utilizando o software livre *HMM-Based Speech Synthesis System* (HTS) [2] elaborado pelo grupo de pesquisa HTS, o qual permite implementar o treinamento dos modelos HMMs a partir de uma base de fala pré-gravada e transcrita foneticamente.

1.2 História da Síntese de Fala

Em 1939, Homer Dudley apresentou o chamado Voder, que pode ser considerado o primeiro sintetizador de fala [3]. Isto indica que desde os primórdios da computação já existia a necessidade de se produzir voz artificialmente.

Mais tarde, em 1950, foi desenvolvido o sintetizador *Pattern Playback* [4] nos laboratórios Haskins. A partir da década de 60, os estudos de sistemas de síntese de fala a partir do texto tiveram um grande impulso, principalmente com o desenvolvimento dos modelos fonte-filtro. Nesta época, também se iniciou o desenvolvimento das primeiras regras de conversão de texto para fonemas, principalmente para a língua inglesa.

Estes avanços permitiram o desenvolvimento de vários sistemas de conversão texto-fala e o aprimoramento dos mecanismos de geração da voz sintética, incorporando melhorias aos sistemas existentes até o momento, tais como a expansão do domínio de síntese para vocabulário ilimitado, a garantia de inteligibilidade do discurso e tentativas de aproximar a voz artificial da voz humana natural.

Em 1995, Tokuda *et al.* [5] propuseram um método de síntese de fala baseado em Modelos Ocultos de Markov. Este método se diferencia dos previamente existentes, por apresentar flexibilidade de adaptação dos modelos de voz, facilidade de realizar alterações prosódicas no discurso, garantia de frases sintetizadas inteligíveis e necessidade de uma base de fala de tamanho limitado, da ordem de 500 a 1000 sentenças para o treinamento, entre outros.

No Brasil, o primeiro trabalho desenvolvido na área de síntese de fala foi o de Campos em 1980 [6], sobre um sintetizador de voz para o idioma Português, capaz de aceitar entradas na forma fonética do Português. No ano de 1985, Esquivel [7] apresentou um sistema TTS no qual sinais adicionais eram acrescentados ao texto para a correta pronúncia de determinados sons. Na Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas (FEEC -UNICAMP), foram realizados diversos estudos sobre o TTS, sendo pioneiro o trabalho desenvolvido por Egashira em 1992 [8,9] usando o processo de síntese concatenativa. Este trabalho, serviu de apoio para a posterior implementação do programa Aiuruetê, software que realiza a conversão texto-fala para o Português falado no Brasil [10,11] e que foi desenvolvido conjuntamente na FEEC e no Instituto de Estudos da Linguagem (IEL) da UNICAMP.

Em 2003, foi implementado por Maia *et al*. [12] uma demonstração disponível online [2] do método de síntese baseado em HMM para a língua Portuguesa, e em 2006 este sistema de demonstração foi aprimorado [13].

Recentemente, existem vários trabalhos que propõem algoritmos considerando as emoções do discurso visando sintetizar a fala da maneira mais natural possível para a língua portuguesa falada no Brasil, como o trabalho desenvolvido por Silva na Universidade Federal do Rio de Janeiro em 2011 [14].

1.3 Estrutura da Dissertação

Esta tese está dividida em 7 capítulos. No Capítulo 2 é introduzido o conceito dos sistemas de síntese de fala e são apresentadas as principais técnicas desenvolvidas para realizar esta tarefa.

No Capítulo 3 é discutido brevemente o mecanismo de produção da fala humana, apresentando as principais ideias que justificam empregar o modelo fonte-filtro no sistema de síntese de fala. Em seguida, são discutidas as diferentes formas de excitação que podem ser aplicadas neste modelo, e é apresentado o filtro *Mel Log Spectrum Approximation* (MLSA).

Uma breve introdução sobre a teoria estatística de HMMs, com um enfoque para os modelos HMM utilizados em síntese de fala, é apresentada no Capítulo 4.

O funcionamento do sistema de síntese via HMM é discutido no Capítulo 5, no qual também são vistos os modelos espectrais, de excitação e de duração que o constituem. São apresentados alguns experimentos realizados, com o intuito de compreender melhor os modelos e os fatores mais determinantes na qualidade final da fala sintética.

No Capítulo 6 são apresentadas algumas técnicas propostas para melhorar o sinal de fala artificial, tais como o Straigth, a pós-filtragem e a variância global.

A conclusão do trabalho é apresentada no Capítulo 7.

2. Introdução aos Sistemas TTS

Um conversor texto-fala, comumente denotado por TTS (*Text-to-Speech*), é um sistema que recebe como entrada um texto corretamente escrito em uma determinada língua e o sintetiza, de modo que na saída deste sistema obtém-se a forma de onda de um sinal de fala correspondente ao texto introduzido na sua entrada.

As principais características almejadas para um sistema TTS são:

- Naturalidade: indica quão parecida a fala sintetizada está da voz humana,
- Inteligibilidade: se refere à facilidade de compreensão pelos ouvintes da fala sintetizada.

Normalmente, um sistema TTS é dividido em dois estágios, conforme ilustra a Figura 2.1, denominados frequentemente de *Front-End*, onde é realizado o processamento do texto, e *Back-End*, que realiza o processamento digital do sinal, isto é, a síntese propriamente dita, a qual dará origem à forma de onda da fala. Nos tópicos 2.1 e 2.2 estão apresentados com mais detalhes cada um desses estágios.

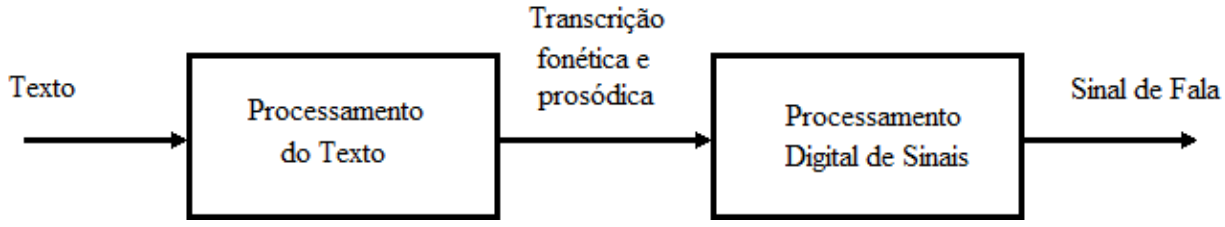

Figura 2.1 – Esquema de um sistema TTS genérico

2.1 Processamento do Texto

No primeiro estágio do sistema TTS é realizado o processamento do texto. O objetivo desta etapa é obter a partir da análise linguística do texto de entrada a sua transcrição fonética e prosódica. É intuitivo notar que este estágio é fortemente dependente das características da língua com a qual se está trabalhando.

Um analisador de textos completo e eficiente deve ser capaz de fornecer não somente dados relativos ao conteúdo semântico do texto, mas também informações prosódicas, como, por exemplo, a duração dos fonemas, mudança nos valores de pitch, ênfases específicas no discurso, entre outras características, que servirão como base para a fase sucessiva de síntese. Para extrair estas características do texto escrito é necessário analisá-lo sob diversas óticas tais como: análise lexical, gramatical, sintática e semântica.

Geralmente, o *Front-End* inicia com um pré-processamento no qual o texto de entrada é convenientemente manipulado e produz um texto na saída. Neste processo escreve-se por extenso todos os caracteres não alfabéticos que aparecem no texto de entrada, de modo que no texto de saída eles apareçam da forma como devem ser pronunciados. Por exemplo, a frase "*Moro na av. Dr. Ruiz n° 54*" deve ser reescrita como "*Moro na avenida doutor Ruiz número cinquenta e quatro*". Para ilustrar, apresentam-se outros exemplos que requerem tratamento de números, datas, siglas, acrônimos e símbolos:

20º	vigésimo
20ºC	vinte graus centígrados
20/10/2010	vinte de outubro de dois mil e dez
20:50	vinte horas e cinquenta minutos
R. Alvorada, nº13	Rua Alvorada, número treze
UNICAMP	Unicampi
maria@decom.com	maria arroba decom ponto com
FFT	éfe éfe tê

Após esta fase de normalização do texto, as palavras são transcritas foneticamente, considerando a acentuação das palavras e sinais de pontuação, além das pausas necessárias na leitura, para que o computador saiba como pronunciar cada fonema considerando as características prosódicas do texto. A expressão "*Leila tem um lindo jardim*" poderia apresentar a seguinte transcrição fonética: "*#lejla te~ u~ li~du ZaXdi~#*", onde # representa o silêncio. O dicionário fonético utilizado no trabalho está apresentado no Apêndice A.

É interessante lembrar que, durante este processo, devem ser tratados alguns problemas próprios da língua, como, por exemplo, a ambiguidade de homógrafos. Homógrafos são palavras

que possuem a mesma grafia, mas que são pronunciadas de forma diferente, dependendo do contexto ou da função sintática exercida. Considere as expressões "*eu gosto de maçã*" e "*o gosto da maçã*"; a palavra *gosto* é escrita do mesmo modo em ambas as sentenças, entretanto na primeira assume a função sintática de verbo, enquanto que na segunda é um substantivo. Como resultado, a pronúncia é feita de forma diferente em cada um dos casos. Outros exemplos:

Tenho sede de água	A sede do congresso
Vou colher o fruto	A colher de pau
Tu leste o artigo	Leste europeu
O governo da união	Ele governa a união

Outra dificuldade está relacionada a diferentes pronúncias relacionadas a um mesmo símbolo fonético. A letra x, por exemplo, corresponde a diferentes fonemas em cada uma das seguintes palavras: "*exame*" – som de z, "*enxame*" – som de ch e "*êxtase*" – som de s.

Todas as particularidades da língua devem ser tratadas para que não ocorram erros de pronúncia na fala artificial. Tendo considerado estas análises para realizar a transcrição fonética do texto, o passo sucessivo é gerar um arquivo que contenha a codificação prosódica e informações contextuais, ou seja, o texto deve ser etiquetado. Um exemplo de etiquetas para o início da frase "*# lejla tem ...*" seria:

y^y-**#**+l=e/M2:y_y/S1:0_@0-_@y+1_@3/S2:y_y/S3:y_y/S4:y_y/S5:y_y/S6:y/W1:0_#0-y_#y+content_#2/W2:y_y/W3:y_y/W4:y_y/W5:0/P1:0_!0-y_!y+8_!5/P2:1_1/U:8_$5_&1

y^#-**l**+e=j/M2:1_3/S1:0_@0-1_@3+0_@2/S2:1_2/S3:1_8/S4:1_3/S5:0_4/S6:e/W1:0_#0-content_#2+content_#1/W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

#^l-**e**+j=l/M2:2_2/S1:0_@0-1_@3+0_@2/S2:1_2/S3:1_8/S4:1_3/S5:0_4/S6:e/W1:0_#0-content_#2+content_#1/W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

l^e-**j**+l=a/M2:3_1/S1:0_@0-1_@3+0_@2/S2:1_2/S3:1_8/S4:1_3/S5:0_4/S6:e/W1:0_#0-content_#2+content_#1/W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

e^j-**l**+a=t/M2:1_2/S1:1_@3-0_@2+0_@3/S2:2_1/S3:2_7/S4:1_3/S5:1_3/S6:a/W1:0_#0-content_#2+content_#1/W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

j^l-**a**+t=e~/M2:2_1/S1:1_@3-0_@2+0_@3/S2:2_1/S3:2_7/S4:1_3/S5:1_3/S6:a/W1:0_#0-content_#2+content_#1/W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

O fone tratado em cada sentença de etiquetas está destacado em vermelho. Nota-se que ele apresenta-se no seu contexto. No exemplo, aparecem os dois fones precedentes e os dois posteriores a ele na frase, juntamente com as informações linguísticas e prosódicas. No Apêndice A é discutido com mais detalhes a simbologia empregada neste trabalho na etiquetagem das sentenças.

Apesar da complexidade do tratamento linguístico, fonético e prosódico, para que o sistema TTS seja autônomo, é necessário que esta etapa do *Front-End* seja executada de forma automática, robusta e rápida. Para a língua inglesa existe um sistema *open source* desenvolvido utilizando o Festival [15] que realiza esta tarefa. Para a língua portuguesa falada no Brasil, ainda não existem sistemas abertos disponíveis, mas existe uma base de fala já etiquetada disponível em [2], a qual foi utilizada para o desenvolvimento deste trabalho.

É intuitivo compreender que esta etapa do *Front-End* é de fundamental importância para o resultado final da fala sintetizada, pois influencia não só a assertividade semântica como também determina a qualidade prosódica do sinal de fala gerado.

2.2 Processamento Digital de Sinais de Fala

Existem diversos métodos para produzir a fala sintética, correspondente ao texto de entrada, a partir do arquivo de etiquetas que contém a sua descrição fonética e prosódica. Os principais tipos são:

- Síntese por Formantes
- Síntese Articulatória
- Síntese Concatenativa
- Síntese Baseada em HMM

Cada tipologia tem suas vantagens e desvantagens características. Normalmente, é a aplicação e o escopo de utilização do sistema de síntese que determina qual delas deve ser adotada.

2.2.1 Síntese por Formantes

A síntese por formantes normalmente utiliza um conjunto de regras, as quais determinam os parâmetros necessários para sintetizar uma dada expressão [16]. Este método não utiliza amostras de fala humana, mas elabora a voz baseado em modelos acústicos, caracterizados geralmente pela frequência de pitch, pelos formantes (ressonâncias do trato vocal na produção de cada som) e os níveis de ruído para excitar o filtro. Para modelar os sons sonoros, sintetizadores por formantes utilizam um sinal periódico que excita o filtro digital construído a partir de várias ressonâncias semelhantes aos formantes produzidos no trato vocal, enquanto que para modelar os sons não sonoros, utiliza-se uma fonte de ruído. O sinal de excitação é gerado por um trem de impulsos com frequência de pitch determinada pela curva melódica. Este trem de impulsos é filtrado por um banco de filtros dispostos paralelamente, os quais modelam os formantes do trato vocal [17].

Por não utilizar amostras de fala humana em tempo de execução, os programas são geralmente leves e podem ser utilizados em sistemas embarcados, onde memória e capacidade de processamento são limitadas.

A principal qualidade desta abordagem de síntese é o fato de gerar um discurso inteligível. Em contrapartida, a naturalidade fica comprometida pelo uso dos vocoders, de modo que as vozes geradas tem uma aparência robótica.

Um dos mais sofisticados sistemas TTS utilizando esta técnica foi desenvolvido por Klatt nos anos 90 [18].

2.2.2 Síntese Articulatória

A síntese articulatória se refere às técnicas computacionais para sintetizar a fala baseadas em modelos do trato vocal humano e no funcionamento do processo articulatório. O objetivo é imitar o mecanismo de produção da fala, simulando assim os movimentos de todos os articuladores relacionados a ela (língua, mandíbula, lábios, palato, glote, etc.), bem como das pregas vocais. Deste modo, para sintetizar a fala, a forma do trato vocal definida pelas posições da articulação é modelada por uma função de transferência que pode, por exemplo, estimar

funções de área ou frequências de formantes, enquanto que o modelo de pregas vocais pode ser usado para gerar sinais de excitação apropriados. Assim, o problema de síntese é transformado no problema de especificar a situação articulatória para cada fonema e modelar com precisão o comportamento dinâmico do aparelho fonador.

A principal dificuldade desta abordagem é obter um modelo preciso da articulação humana no processo de fala. Não se tem conhecimento de sistemas comerciais que empregam esta técnica, mas existem várias pesquisas nesta área, podendo-se citar os trabalhos de Engwall [19] e Mullen *et al.* [20].

2.2.3 Síntese Concatenativa

O processo de síntese concatenativa está baseado na concatenação de segmentos de fala natural pré-gravada. Portanto, é necessário armazenar um banco de dados com uma base de áudio e texto, sendo a qualidade da fala artificial diretamente relacionada à extensão do banco de dados.

A principal vantagem deste método é a qualidade da fala gerada em termos de naturalidade do som. Entretanto, diferenças entre variações naturais da fala e a natureza das técnicas automáticas para segmentação das formas de onda podem acarretar falhas que comprometem a inteligibilidade do discurso.

Para evitar este problema, que surge devido à concatenação de segmentos de fala não naturalmente sequenciais, a maior parte dos sistemas empregados em uso comercial aplica a concatenação de polifones que compreendem grupos de difones e trifones. Empregando os polifones, as transições entre os fones são preservadas e a concatenação é feita entre sinais com conteúdo espectral semelhante. Com esta técnica pode-se obter um sinal de fala, resultante da concatenação dos diversos segmentos, mais natural e agradável. Todavia, esta abordagem exige o uso de uma base de fala extensa que permita extrair os polifones nos diversos contextos.

Uma das técnicas mais utilizadas atualmente e que fornece alta qualidade de síntese é a TD-PSOLA (*Time-Domain Pitch-Synchronous Overlap-Add*) [21]. Ela permite a alteração da duração da fala e da frequência de pitch dos fones, de modo a satisfazer os requisitos das palavras que conterão aqueles fones. Este processamento evita o problema da monotonicidade do discurso sintetizado, melhorando a prosódia da fala. O esquema de funcionamentos desta abordagem [22]

é o seguinte: o sinal de fala é inicialmente submetido a um algoritmo de marcação de pitch, o qual, em segmentos sonoros, marca os picos do sinal que ocorrem distanciados do período de pitch; para os segmentos não sonoros faz-se uma marcação a cada 10ms, aproximadamente. A síntese é realizada por superposição de segmentos janelados, utilizando janelas de Hanning, centrados nas marcações de pitch e extendido da marca de pitch anterior até a marca sucessiva. A modificação da duração é feita cancelando ou replicando algumas das janelas, enquanto que para modificar o período de pitch a superposição entre os segmentos janelados é aumentada ou diminuída. Estes mecanismos de modificação da duração e frequência são ilustrados nas Figuras 2.2 e 2.3, respectivamente.

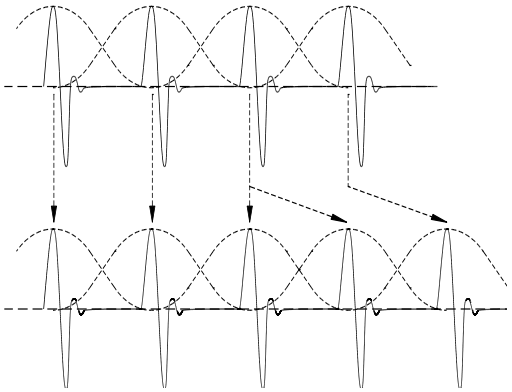

Figura 2.2 – Aumento da duração por duplicação de sinais elementares
Fonte: [22]

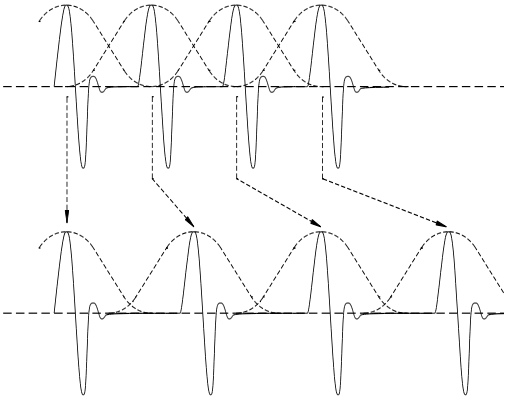

Figura 2.3 – Redução da frequência de pitch
Fonte: [22]

2.2.4 Síntese Baseada em HMM

O sistema de síntese de fala baseado em modelos ocultos de Markov (HMM) foi proposto por Tokuda *et al.* [5] em meados da década de 90. Este método paramétrico e estatístico permite realizar a síntese de fala concatenando fones dependentes de contexto modelados por HMM. Esta é a tecnologia de síntese objeto de estudo deste trabalho e está apresentada no Capítulo 5.

2.2.5 Considerações sobre os Métodos de Síntese

Dos quatro métodos apresentados no presente capítulo, os dois primeiros (síntese por formantes e articulatória) são atualmente pouco utilizados para construir os sistemas de síntese de fala comerciais. Além disso, não foram encontrados trabalhos recentes sobre sistemas TTS sendo desenvolvidos com estas técnicas. O foco dos pesquisadores e desenvolvedores de sistemas de síntese de fala parece estar direcionado para a síntese concatenativa, que é o método comercial utilizado com mais frequência atualmente, e o método de síntese via HMM, que é visto como promissor devido à sua estabilidade, flexibilidade em modificar as características da voz artificial, desempenho e necessidade de uma base de fala para treinamento limitado, da ordem de uma a duas horas de gravação. O tamanho do corpus de treinamento necessário para o sistema de síntese via HMM depende da complexidade da língua, uma vez que os modelos HMM dos fones dependentes de contexto necessitam de amostras representativas no corpus para que seja possível realizar boas estimativas estatísticas, entretanto é possível estimar fones em contextos que não aparecem na base de fala de treinamento, como está explicado na Seção A-3 do Apêndice.

3. Processo de Produção da Fala

3.1 Aparelho Fonador Humano

Para compreender melhor a estrutura de funcionamento da síntese de fala utilizada no HTS, é importante considerar os principais aspectos relacionados à produção de fala. A Figura 3.1 apresenta o aparelho fonador humano responsável pela produção da fala.

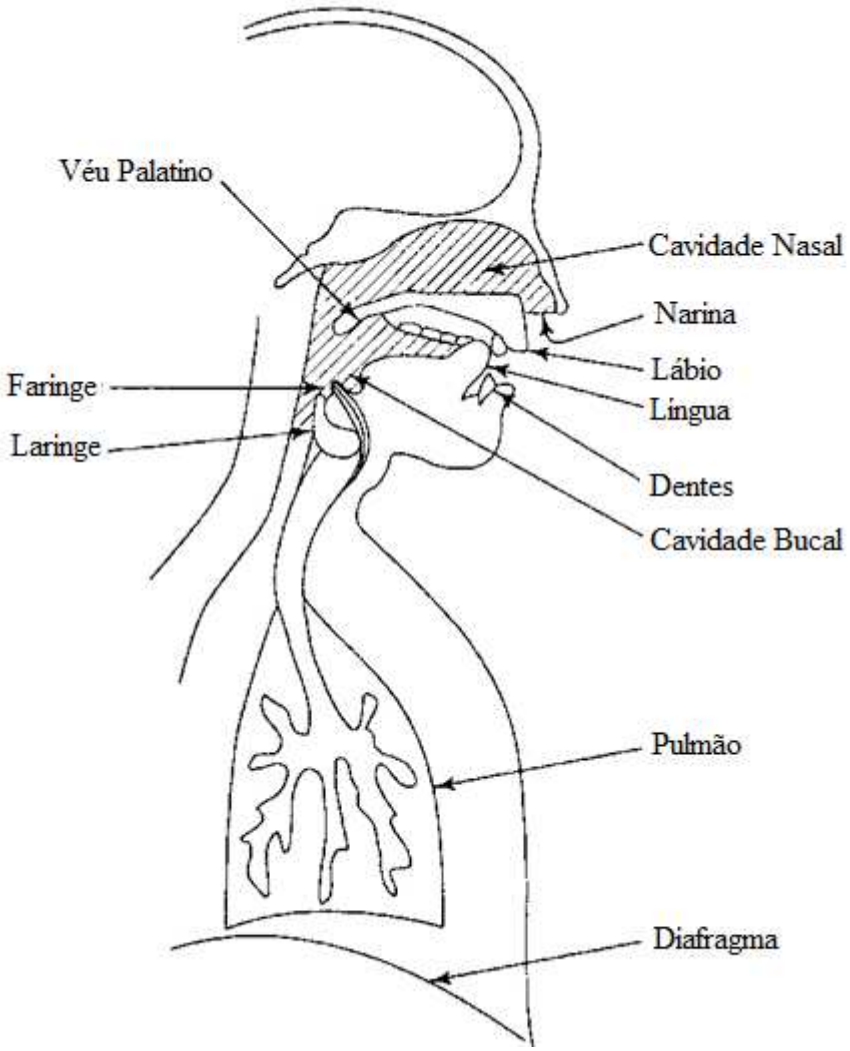

Figura 3.1 – Aparelho fonador humano
Fonte: [23]

O mecanismo da geração do sinal de fala humano funciona da seguinte maneira: o fluxo de ar vindo dos pulmões passa pela laringe e, por nosso comando neural, por meio de ajustes musculares, faz pressões de diferentes graus sobre as cordas vocais, localizadas na região da glote, fazendo-as eventualmente vibrarem em diferentes frequências. Este fluxo de ar é modificado ao longo do tempo pelas cavidades bucais e nasais do trato vocal, de modo que os sons vão sendo articulados adequadamente, e geram os diferentes fonemas da língua, os quais são sucessivamente emitidos através da boca, criando a onda sonora.

3.2 Sons Articulados

Os diferentes sons podem ser classificados basicamente em dois tipos: *sonoros* e *não sonoros,* dependendo do modo como são gerados. Normalmente, as vogais possuem uma natureza de som sonoro, enquanto que as consoantes podem apresentar os dois comportamentos. A distinção entre consoantes não sonoras e sonoras pode ser ilustrada nestes exemplos:

Não sonoro: /p/ pê - /t/ tê

Sonoro: /b/ bê - /d/ dê

Uma forma de perceber a natureza do som é posicionar os dedos na garganta na região das pregas vocais. Ao se pronunciar um som sonoro, é possível sentir a sua vibração. A taxa de vibração (abertura e fechamento das cordas vocais) determina a frequência fundamental, ou frequência de pitch, do som. A frequência fundamental de vibração das cordas vocais depende das características pessoais. Em geral, as pregas vocais são mais longas para os homens e mais curtas para as mulheres, o que determina uma voz masculina mais grave (frequência de oscilação menor) e uma voz feminina mais aguda (frequência de oscilação maior).

A Tabela 3.1 apresenta os intervalos de valores para a frequência fundamental de vibração das cordas vocais que geralmente são observados para cada tipo de locutor.

Tabela 3.1 – Valores indicativos da frequência fundamental dos sons sonoros		
Orador	**f_{0min} (Hz)**	**$f_{0máx}$ (Hz)**
Homem	70	200
Mulher	150	400
Criança	200	600

A Figura 3.2 mostra a forma de onda da vogal /e/. Pode-se verificar que os sons sonoros apresentam um comportamento quase periódico, sendo as variações da forma de onda causadas quer pela variação (movimentação) lenta do trato vocal, quer por diferenças de energia. É importante ressaltar que a variação da frequência de vibração está diretamente relacionada com a emoção do discurso.

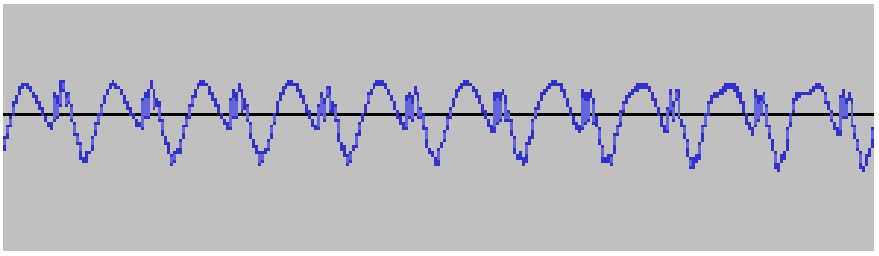

Figura 3.2 – Sinal de voz sonoro – forma de onda da vogal /e/

Os sinais não sonoros ou surdos são produzidos sem vibração das cordas vocais. A rápida passagem do ar pelo trato vocal produz uma turbulência. Este tipo de sinal não apresenta periodicidade. Os sons surdos são gerados pela filtragem deste fluxo de ar através do trato vocal.

A Figura 3.3 apresenta a forma de onda da consoante /s/. É imediato notar a característica ruidosa da forma de onda, típica dos sons não sonoros.

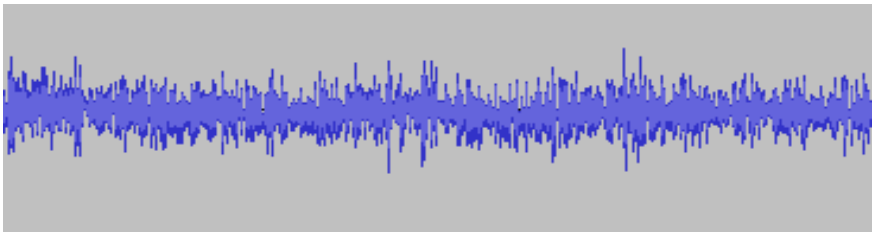

Figura 3.3 – Sinal de voz não sonoro – forma de onda da consoante /s/

Apesar dos sons sonoros e não sonoros apresentarem características tão distintas, é necessário lembrar que a fala não é constituída por uma sequência de sons bem definidos, com uma mudança brusca entre estes. Pelo contrário, a transição entre um par de segmentos fonéticos produz-se de forma gradual, com variações suaves das características de um som para o som sucessivo. Isto se deve ao fato de que a vibração das cordas vocais e o trato vocal são mecanismos mecânicos, cujas movimentações e alterações ocorrem de forma lenta. Por isso, os sinais de fala podem ser considerados estacionários em curtos períodos de tempo (~25ms) [24]. Desta forma, é frequente tratar os sinais de fala usando **janelas** de 20 a 25 ms, com deslocamento da janela de 5 a 10 ms, para analisar as características espectrais e de pitch da fala. Estes deslocamentos originam os **quadros,** que representam o intervalo de tempo em que as análises são atualizadas.

3.3 Modelo Fonte-Filtro

O modelo fonte-filtro é uma das estruturas mais simples e amplamente empregadas para a síntese de fala [24, 25]. Neste modelo, considera-se que a síntese é o resultado de um sinal de excitação submetido a um filtro variante no tempo. A fonte representa o ar expelido pelos pulmões que, ao passar pela glote, pode causar ou não a vibração das pregas vocais, de acordo com os comandos cerebrais, enquanto que o filtro modela todo o trato vocal que modula este fluxo de ar. Assume-se que a fonte e o filtro são modelos independentes.

Existem diversas técnicas para modelar a fonte e o filtro. As principais abordagens empregadas no sistema HTS para gerar a excitação da fonte são: a excitação simples, a excitação mista e o Straight. Estes métodos estão descritos nas Seções 3.3.1, 3.3.2 e 6.1, respectivamente. Para a filtragem utiliza-se o filtro *Mel Log Spectrum Approximation* (MLSA) [26], apresentado na Seção 3.3.4.

Sistemas mais antigos costumavam geralmente adotar o filtro *Linear Prediction Conding* (LPC) [25]. Entretanto, não é possível garantir a estabilidade deste filtro na modelagem estatística empregada, o que obriga a utilização de coeficientes de reflexão derivados dele. Isto se torna um problema, principalmente quando se utilizam filtros de ordem elevada. Outra desvantagem é que,

sendo um filtro só de pólos, ele não é capaz de modelar bem os sons nasais, para os quais o filtro que modela o trato vocal apresenta zeros.

3.3.1 Excitação Simples

Este é o modelo mais simples de vocoder e é baseado no duplo comportamento do som (sonoro/não sonoro). Sabe-se que a excitação dos sons sonoros tem uma natureza periódica e pode ser representada no domínio do tempo por um trem de impulsos distanciados pelo período de pitch, enquanto que a excitação não-sonora possui uma natureza ruidosa, podendo ser representada por uma vasta gama de modelos de ruído gerados por distribuições aleatórias, tais como a distribuição uniforme, gaussiana ou a sequência-m [27], a qual gera uma sequência de tamanho m assumindo valores aleatórios de +1 e -1.

Devido a este duplo comportamento da excitação na fala, o modelo representado na Figura 3.4 comuta convenientemente entre as duas fontes de excitação. Esta comutação, bem como os parâmetros do filtro, mudam a cada quadro, ou seja, a cada intervalo de cerca 5 ms. É razoável assumir que o sistema é invariante no tempo neste período, pois as propriedades gerais do trato vocal e da excitação estão submetidas à inércia mecânica do aparelho vocal. Sob estas condições, a excitação $e(n)$ é filtrada por um sistema linear $h(n)$ para gerar o sinal de fala $x(n)$. O sinal de fala $x(n)$ pode ser calculado a partir da excitação $e(n)$ e da resposta ao impulso do sistema $h(n)$ usando a expressão de convolução discreta: $x(n) = e(n)*h(n)$, ou seja, a fala sintetizada corresponde à convolução temporal do sinal de excitação com a resposta ao impulso do filtro.

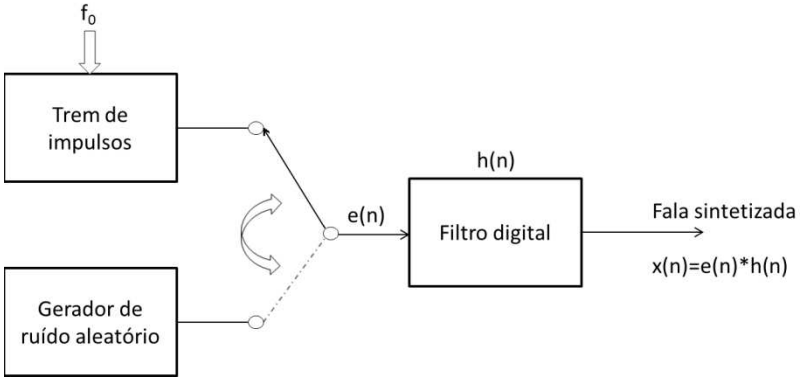

Figura 3.4 – Modelo fonte-filtro

3.3.1.1 Implementação da Excitação Simples no HTS

No HTS, a síntese de fala usando o método de excitação simples pode ser implementada pelo programa auxiliar hts_engine-API [28], que possui todas as rotinas prontas para, a partir dos modelos HMMs, gerar o trem de impulsos e o filtro necessários. O hts_engine-API não necessita das bibliotecas do HTK/HTS [29] sendo, portanto, mais flexível para desenvolver aplicações, apesar de restringir algumas opções de modelos.

A rotina funciona do seguinte modo: durante a síntese o primeiro quadro recebe sempre ruído branco, o que obviamente não causa nenhum prejuízo, pois é natural que uma forma de onda contenha ao menos 5 ms de "silêncio/ruído" no início. Isto permite que no *loop* de leitura das frequências de pitch, o programa conheça sempre dois valores de pitch, o do quadro precedente (*p1*) e o do quadro atual (*p*), para poder gerar a excitação do quadro atual. A rotina responsável por calcular os valores de excitação para a síntese de fala usando o hts_engine-API é a *vocoder.c*. Ela realiza procedimentos diferentes dependendo da frequência de pitch do quadro atual corresponder a um som sonoro ou a um som não sonoro.

Quadro de som sonoro:

O período do quadro, no qual o programa computa os parâmetros do filtro e o sinal de excitação, foi fixado em 5 ms neste trabalho, e a frequência de amostragem utilizada foi de 16 kHz, de forma que cada quadro contém 80 amostras. O script trabalha com o número de amostras, assim o período de pitch associado (em número de amostras) pode ser calculado para os quadros sonoros por:

$$p_1 = \frac{16\ kHz}{\exp(\ln f_0)} = \frac{16\ kHz}{f_0} \tag{3.1}$$

Observe que se utiliza o logaritmo da frequência de pitch ($\ln f_0$) no modelo de excitação, pois este é melhor modelado por uma distribuição Gaussiana.

Da expressão (3.1), se o quadro é não sonoro, p_1 é anulado. A frequência f_0 pode assumir valores entre 130 Hz e 320 Hz, que foram os valores configurados durante a fase de extração de pitch no treinamento, uma vez que se trabalhou com uma base de fala feminina. Deste modo, p_1 pode assumir valores no intervalo entre 50 e 123 amostras. Assim sendo, se p_1 for diferente de

zero, trata-se de um quadro sonoro e é necessário gerar um trem de impulsos que tenha a mesma frequência de pitch do quadro correspondente e determinar a amplitude de cada impulso. Esta amplitude é dada por $\sqrt{p_1}$ de modo a assegurar variância unitária (o ganho é associado ao filtro). No algoritmo, este valor é ligeiramente adaptado levando em consideração os valores de pitch dos quadros vizinhos, de forma a não gerar transições abruptas entre quadros sonoros e não sonoros. Para maiores detalhes, pode-se consultar o código fonte *vocoder.c* [28].

Quadro de som não-sonoro:

Para os quadros não sonoros, p_1 é feito igual a zero e o algoritmo gera um ruído branco. Este ruído pode ser de vários tipos, como, por exemplo, gaussiano, sequência m, uniforme, entre outros. Todos estes ruídos brancos possuem variância unitária. Neste trabalho, foram analisados os dois primeiros tipos citados (gaussiano e sequência-m). A síntese originada por ambos foi considerada equivalente no sentido acústico qualitativo/subjetivo. Contudo, a forma de onda apresenta-se claramente diferente, uma vez que o ruído branco gaussiano produz uma variação maior na amplitude do sinal gerado, enquanto que a sequência-m assume sempre os valores +1 e -1 aleatoriamente.

3.3.2 Excitação Mista

A excitação mista considera o fato de que as duas naturezas do som (sonoro/não sonoro) não aparecem completamente separadas durante a produção natural da fala. Assim, sons sonoros podem possuir componentes ruidosas mesmo que em níveis menos intensos, principalmente considerando o caso das consoantes vozeadas. Ao introduzir esta característica no processo de excitação, deseja-se aprimorar a naturalidade da voz, eliminando um pouco o aspecto robótico da fala artificial gerada com a excitação simples.

A ideia da excitação mista é dividir o sinal em bandas. Considerando gravações com 16 kHz de taxa de amostragem, têm-se as seguintes bandas: 0 – 1; 1 – 2; 2 – 4; 4 – 6; 6 – 8 kHz. Para cada banda é gerado um sinal de excitação, que combina o ruído branco e o trem de impulsos, atribuindo um peso adequado para cada componente. Na excitação mista todos os

filtros são de fase linear com comprimento $N = 31$ e com atraso de $(N-1)/2 = 15$ amostras, considerando o caso de frequência de amostragem de 16 kHz.

A Figura 3.5 ilustra o processo de geração da excitação mista para um sinal amostrado à frequência de 8 kHz. Em cada gráfico da figura vê-se a divisão das 5 bandas de frequências: 0 – 0,5; 0,5 – 1; 1 – 2; 2 – 3; 3 – 4 kHz. Os dois gráficos superiores correspondem aos dois tipos de excitação, impulsiva e ruidosa. Para diferenciar qual das duas excitações é dominante na faixa de frequências em questão, utilizou-se a combinação das cores branco e cinza. Observe que elas se alternam nos dois gráficos de excitação, sendo que a cor de fundo branco indica que aquela faixa de frequências possui componentes principais do respectivo tipo de excitação, ou seja, no exemplo da Figura 3.5 as faixas 0 – 0,5 e 1 – 2 kHz possuem excitação predominantemente impulsiva (periódica), enquanto que as outras faixas de frequências possuem excitação ruidosa. Esta diferenciação indica o tipo de filtragem que cada faixa de frequências dos sinais de excitação (sonoro e não sonoro) deve sofrer. Usando filtros do tipo banda-passante adequados, obtém-se o sinal resultante da excitação mista, ilustrado no último quadro da Figura 3.5.

No HTS usa-se o modelo de mistura multi-bandas [30] que tende a representar melhor a voz humana. Uma dificuldade deste método é determinar com precisão os pesos das componentes sonoras e não sonoras para cada banda. Uma atribuição errada compromete muito a qualidade final do sinal de fala sintetizado.

Para utilizar esta técnica de excitação não é possível sintetizar fala com a ferramenta hts_engine-API e deve-se usar a rotina HMGenS desenvolvida para o HTS. Isto exige os módulos das bibliotecas do HTK/HTS em fase de síntese, o que dificulta o processo de implantação desse método para aplicações embarcadas. Todavia, esta ferramenta apresenta muitas funcionalidades, o que permite construir os modelos com maior flexibilidade, sendo adequada para pesquisas e desenvolvimento de novos modelos.

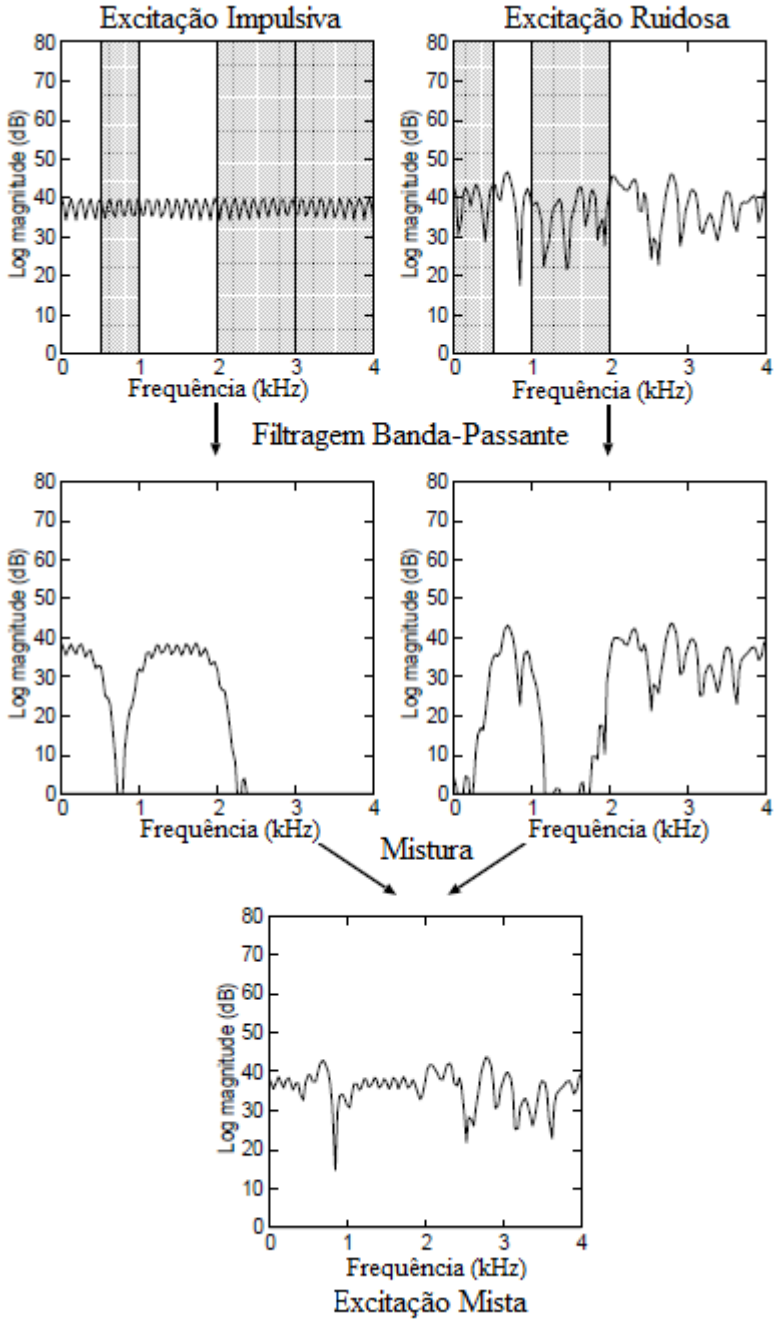

Figura 3.5 – Mistura multi-bandas
Fonte: adaptado de [30]

3.3.3 Comparações dos Métodos de Excitação

Neste trabalho, os modelos HMM foram treinados utilizando a base de fala disponível para a língua portuguesa [2]. E foram sintetizadas frases considerando ambas as técnicas de excitação descritas: excitação simples e mista.

A pesquisa subjetiva realizada com ouvintes não especializados em síntese de fala mostrou que eles não tiveram preferência por uma ou outra excitação, considerando as amostras muito semelhantes. Entretanto, de acordo com a literatura, a utilização do modelo de excitação mista agrega um ganho de qualidade perceptível acusticamente. No experimento descrito por Yoshimura (2002, p. 66) em [30], a preferência dos ouvintes é cerca de 50% maior para as frases sintetizadas utilizando o método de excitação mista. Uma possível explicação para o resultado diferente obtido no experimento realizado neste trabalho, é que as amostras da base de fala contém um pouco de ruído de fundo, o que pode ter atrapalhado a obtenção correta das ponderações de ruído/impulso nos quadros de excitação, degradando o sinal de fala sintético.

As Figuras 3.6 e 3.7 apresentam os espectrogramas de um dos sinais de fala sintetizados e avaliados subjetivamente pelos ouvintes. Eles correspondem à parte sublinhada da sentença *"Apenas os ônibus circularão pela pista bairro-centro nos dois sentidos",* utilizando a excitação simples e a excitação mista, respectivamente. Nota-se que os sinais, além de acusticamente parecidos, também são espectralmente semelhantes. Desta forma, o aumento da complexidade do modelo com o uso da excitação mista acabou não sendo justificado no sistema TTS implementado e escolheu-se utilizar a excitação simples para realizar as demais análises deste estudo.

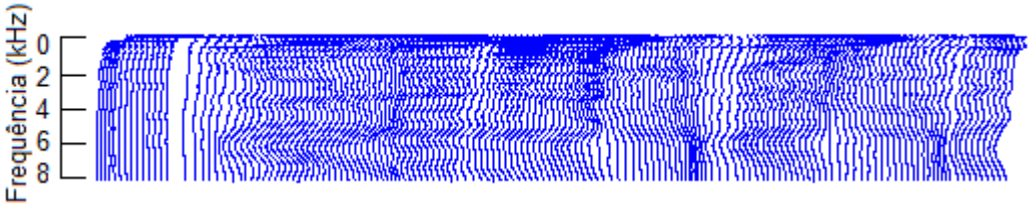

Figura 3.6 – Espectrograma – excitação simples

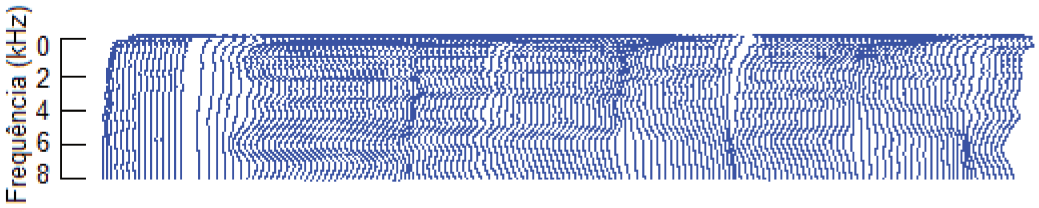

Figura 3.7 – Espectrograma – excitação mista

3.3.4 Filtro MLSA

Como o sistema auditivo humano opera em uma escala de frequências não linear, é natural adotar modelos que envolvem o uso de escalas de frequência não lineares tais como as escalas mel e Bark para representar a percepção auditiva. A escala mel foi proposta em 1937 como resultado de uma série de experimentos usados para estabelecer a escala de percepção baseada na percepção dos tons. Conforme apresentado na Figura 3.8, esta escala representa com maior precisão as baixas frequências, nas quais o ouvido humano é mais sensível, e modela com menor precisão as altas frequências. O uso da escala mel é praticamente padronizado para aplicações de processamento de fala, devido à sua facilidade de aproximar muito bem a sensibilidade auditiva humana. Acrescido a isto, temos o fato que a natureza do espectro do sinal de fala, composto por formantes (picos de ressonância) e anti-formantes (vales de antirressonâncias), faz com que ele possa ser bem representado por um filtro com pólos e zeros. Combinando estas duas características, pode-se pensar em modelar o envelope espectral do sinal de fala, ou seja, o trato vocal, usando um filtro do tipo pólos-zeros com coeficientes mel-cepstrais [31].

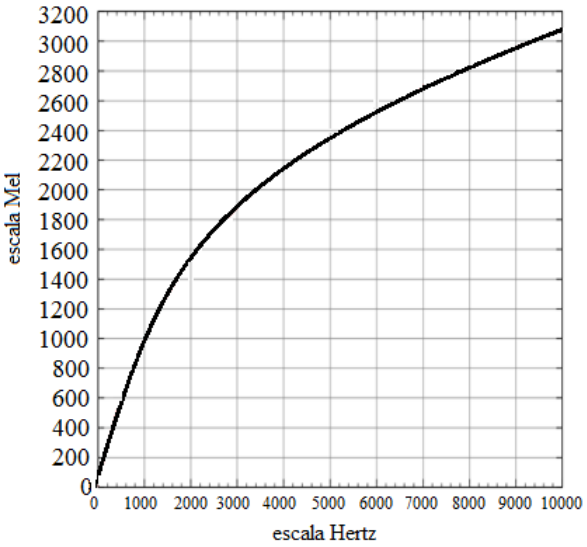

Figura 3.8 – Gráfico da escala mel versus a escala Hertz

Um filtro que apresenta estas características é o *Mel Log Spectrum Approximation* (MLSA) que é empregado no HTS. Sua transformada Z é dada por:

$$H(z) = \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m} \tag{3.2}$$

Note que usando a exponencial garante-se sempre a estabilidade do filtro. Além disso, $H(z)$ é um sistema de fase mínima [32]. Garantir a estabilidade do filtro é essencial para o funcionamento do modelo, uma vez que os coeficientes vão ser determinados por modelos estatísticos.

Os parâmetros $c(m)$ são os coeficientes mel-cepstrais, M é a ordem do filtro e $\tilde{z}^{-1}$ é dado por:

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{3.3}$$

com $|\alpha| < 1$ e $z = e^{j\omega}$, onde ω é a frequência normalizada em radianos. Em particular, se $\alpha = 0$ tem-se que $\tilde{z}^{-1} = z^{-1}$ de modo que os coeficientes $c(m)$ corresponderão aos coeficientes cepstrais. O parâmetro α corresponde ao fator de frequência e deve ser escolhido de forma apropriada segundo a frequência de amostragem com a qual se trabalha.

A Tabela 3.2 mostra os valores de α adotados em algumas frequências de amostragem típicas, de modo a aproximar adequadamente a escala de frequência auditiva.

Tabela 3.2 – Valores usuais de α segundo a frequência de amostragem

Frequência de Amostragem	8kHz	16kHz	48kHz
Escala mel	0,31	0,42	0,55
Escala Bark	0,42	0,55	-

Fazendo $z = e^{j\omega}$ e $\tilde{z} = e^{j\beta}$, onde β é a frequência distorcida (*warping frequency*) na escala mel, obtém-se:

$$\beta(\omega) = \arctan \frac{\left(1-\alpha^2\right) sen\,\omega}{\left(1+\alpha^2\right)\cos\omega - 2\alpha} \tag{3.4}$$

$\beta(\omega)$ fornece uma boa aproximação para a escala de frequência auditiva quando o valor de α é escolhido adequadamente.

A Figura 3.7 apresenta um exemplo das escalas de frequência considerando a frequência de amostragem de 16 kHz e usando a escala mel com α =0,42.

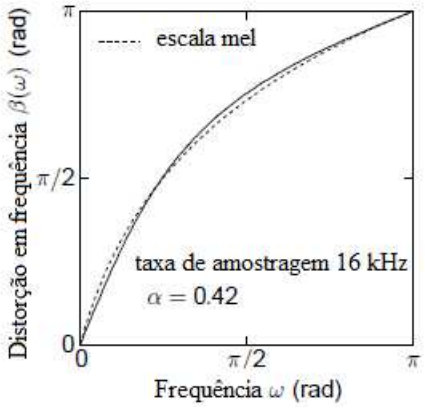

Figura 3.9 – Distorção em frequência (*frequency warping*) da escala mel
Fonte: [32]

Pode-se demonstrar que, através da técnica de análise cepstral denominada UELS (*Unbiased Estimation of Log Spectrum*) [33], os parâmetros mel-cepstrais podem ser encontrados minimizando-se E a fim de se obter as estimativas da potência espectral estimada $\left|H(e^{j\omega})\right|^2$ [32].

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[e^{R(\omega)} - R(\omega) - 1\right]d\omega \tag{3.5}$$

onde

$$R(\omega) = \ln I_N(\omega) - \ln\left|H(e^{j\omega})\right|^2 \tag{3.6}$$

com $I_N(\omega)$ sendo o períodograma modificado de um processo $x(n)$ estacionário no sentido amplo,

$$I_N(\omega) = \frac{\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\omega n}\right|^2}{\sum_{n=0}^{N-1} w^2(n)} \tag{3.7}$$

e $w(n)$ a janela de comprimento N. Decompondo o filtro $H(z)$, tal que:

$$H(z) = K.D(z) \tag{3.8}$$

e considerando $H(z)$ causal com $d[0] = 1$, tem-se:

$$K = \exp\sum_{m=0}^{M}(-\alpha)^m c(m) \tag{3.9}$$

e

$$D(z) = \exp\sum_{m=1}^{M} c_1(m)\tilde{z}^{-m} = \exp(\mathrm{F(z)}) \tag{3.10}$$

onde,

$$F(z) = \sum_{m=1}^{M} c_1(m)\tilde{z}^{-m} \tag{3.11}$$

sendo a relação entre os coeficientes $c(m)$ e $c_1(m)$ dada por:

$$c_1(m) = \begin{cases} c(0) - \sum_{m=0}^{M}(-\alpha)^m c(m), & m = 0 \\ c(m), & 1 \le m \le M \end{cases} \tag{3.12}$$

Como visto anteriormente, o filtro $H(z)$ é de fase mínima e estável, logo $D(z)$ também será um filtro de fase mínima e estável. Isto permite a sua utilização no modelo estatístico dos coeficientes do filtro. Pode-se estabelecer a seguinte relação:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left|H(e^{j\omega})\right|^2 d\omega = \ln K^2 \tag{3.13}$$

e

$$E = \frac{\varepsilon}{K^2} - \frac{1}{2\pi}\int_{-\pi}^{\pi}\ln I_N(\omega)d\omega + \ln K^2 - 1 \tag{3.14}$$

com

$$\varepsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{I_N(\omega)}{\left|D(e^{j\omega})\right|^2}d\omega \tag{3.15}$$

Assim, tem-se que a minimização de E com respeito a $\boldsymbol{c}$ implica na minimização de ε com respeito a $\boldsymbol{c_1}$ e na minimização de E com respeito a K. Calculando a derivada de E com respeito a K e igualando o resultado a zero, obtém-se:

$$K = \sqrt{\varepsilon_{\min}} \tag{3.16}$$

onde $\varepsilon_{\min}$ é o valor mínimo de ε. É possível demonstrar que a minimização da expressão (3.15) implica na minimização da energia residual.

Como E é convexo com respeito a $\boldsymbol{c}$, existe somente um ponto de mínimo. Consequentemente, o problema da minimização de E pode ser solucionado aplicando-se o algoritmo de Newton-Raphson.

Para construir o filtro MLSA é necessário aproximar a função de transferência do tipo exponencial por uma função racional. Para isso, utilizam-se os aproximantes de Padé [26] de ordem L, onde:

$$e^{\omega} \approx R_L(\omega) = \frac{1+\sum_{l=1}^{L} A_{L,l}(\omega)^l}{1+\sum_{l=1}^{L} A_{L,l}(-\omega)^l} \tag{3.17}$$

$$A_{L,l} = \frac{1}{l!}\frac{\binom{L}{l}}{\binom{2L}{l}} \tag{3.18}$$

sendo $\binom{L}{l}$ o binômio de Newton:

$$\binom{L}{l} = \frac{L!}{(L-l)!l!} \tag{3.19}$$

Fazendo $\omega = F(z)$ resulta que a expressão (3.17) pode ser reescrita como:

$$D(z) = e^{F(z)} \approx \frac{1+\sum_{l=1}^{L} A_{L,l}\left[c_1(0)+c_1(1)\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}+c_1(2)\left(\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\right)^2+\cdots\right]^l}{1+\sum_{l=1}^{L} A_{L,l}\left[-c_1(0)-c_1(1)\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}-c_1(2)\left(\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\right)^2-\cdots\right]^l} \tag{3.20}$$

assim,

$$D(z) \approx \frac{1+\sum_{l=1}^{L} A_{L,l}[F(z)]^l}{1+\sum_{l=1}^{L} A_{L,l}[-F(z)]^l} = \frac{Y(z)}{X(z)} \tag{3.21}$$

Desta forma, a função $D(z)$ pode ser aproximada com boa precisão fazendo:

$$D(z) = \exp F(z) \approx R_L(F(z)) \tag{3.22}$$

Este filtro não somente trabalha na escala mel, mas também na escala Bark, bastando substituir os parâmetros do filtro passa-tudo. A ordem dos aproximantes de Padé usada pelo sistema geralmente é 4 ou 5 [26].

A estrutura do filtro MLSA usando aproximantes de Padé de ordem 4 é representada na Figura 3.10.

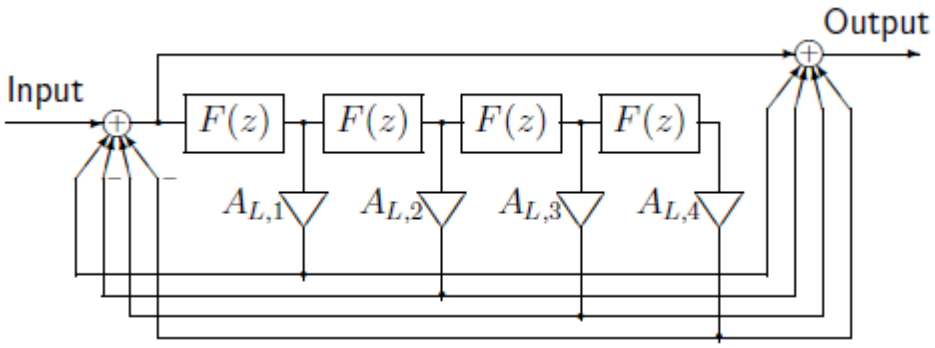

Figura 3.10 – Estrutura do filtro $R_4(F(z)) \approx D(z)$
Fonte: [32]

Vê-se que esta estrutura apresenta *zero delay loop,* o que torna o sistema não implementável devido à realimentação. Para evitar este problema faz-se a seguinte transformação:

$$F(z) = \sum_{m=1}^{M} c_1(m)\tilde{z}^{-m} = \sum_{m=1}^{M} b(m)\Phi_m(z) \tag{3.23}$$

assim, (3.22) pode ser escrita como:

$$D(z) = \exp\left(\sum_{m=1}^{M} b(m)\Phi_m(z)\right) \tag{3.24}$$

onde,

$$b(m) = \begin{cases} c_1(m) & m = M \\ c_1(m) - \alpha b(m+1) & m = 1,2,\ldots,M-1 \end{cases} \tag{3.25}$$

e,

$$\Phi_m(z) = \begin{cases} 1 & m = 0 \\ \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\tilde{z}^{-(m-1)} & m \geq 1 \end{cases} \tag{3.26}$$

Desta forma propõe-se o filtro com a estrutura apresentada na Figura 3.11.

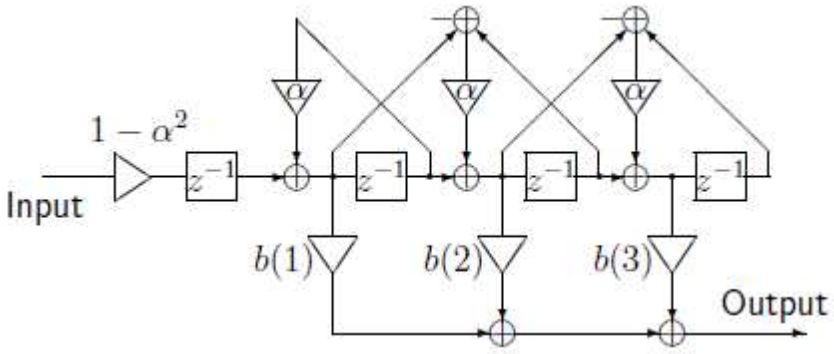

Figura 3.11 – Estrutura do filtro $F(z)$
Fonte:[32]

3.3.4.1 Análises Espectrais

Como foi visto, o sistema HTS utiliza o filtro MLSA para modelar o trato vocal. Todavia, também poderia ser empregado o filtro *Mel Generalized Log Spectrum Approximation* (MGLSA) [34] que trabalha com coeficientes generalizados. A função de transferência $H(z)$ do filtro de síntese MGLSA é dada pela expressão:

$$H(z) = \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\tilde{z}^{-m}\right)^{1/\gamma}, & 0 < \gamma \leq -1 \\ \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\tilde{z}^{-m}, & \gamma = 0 \end{cases} \tag{3.27}$$

onde M é a ordem do filtro e $\tilde{z}^{-1}$ é dado como em (3.3). As variáveis α e γ são responsáveis por regular as mudanças na escala de frequência e a amplitude espectral, respectivamente. A variável α assume valores no intervalo]-1,1[e permite modificar a precisão com que são tratadas as baixas frequências na análise espectral. O parâmetro γ assume valores na faixa [-1,1] e ele influência nas ponderações utilizadas para descrever os picos e os vales do espectro. Observe que quando γ = 0, a expressão (3.27) do filtro MGLSA se torna igual à do filtro MLSA apresentado em (3.2).

A Figura 3.12 apresenta os tipos de análises suportados para o projeto do filtro. O sistema HTS-2.2 utiliza o pacote de ferramentas *Speech Signal Processing Toolkit* (SPTK) [35] para realizá-las. Observa-se que o SPTK permite adotar qualquer análise do conjunto mel-cepstral

generalizado, através de uma abordagem unificada [36], possibilitando todas as possíveis combinações dos parâmetros α e γ.

Na Seção 5.3.1 é discutida a capacidade de algumas dessas análises modelarem o trato vocal humano, bem como são apresentadas considerações a respeito da ordem do filtro a ser utilizado e da estabilidade do sistema.

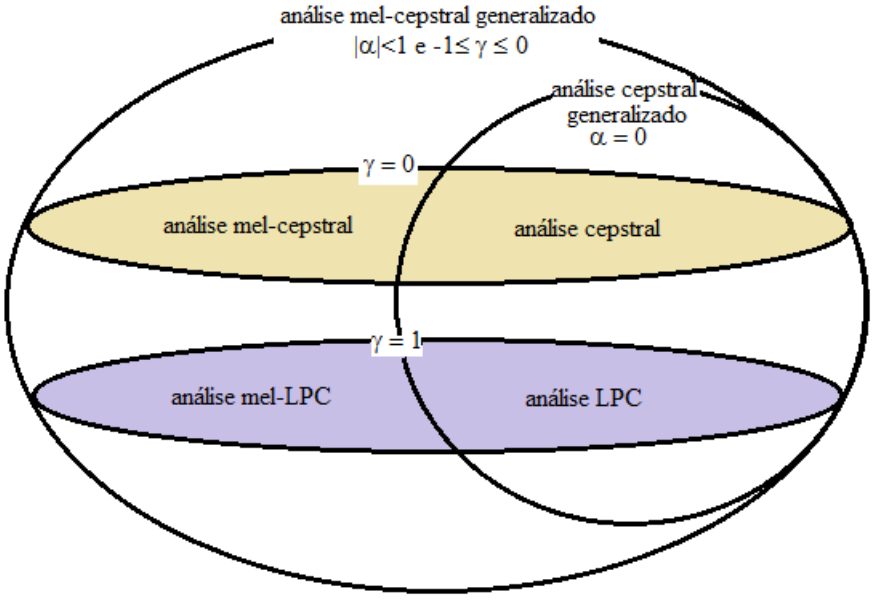

Figura 3.12 – Possíveis análises espectrais realizadas no SPTK.

4. Modelos Ocultos de Markov

A teoria básica de modelos ocultos de Markov (HMM) foi introduzida na literatura por Baum [37] na década de 60. Nos anos 70 ela começou a ser empregada em sistemas de reconhecimento de voz [29] e, em meados dos anos 90, foi proposta para sistemas de síntese de fala [5]. Os HMM são modelos estatísticos amplamente utilizados quando se manipulam fontes de sinais sequenciais. Em processamento de sinais de fala, utilizam-se os modelos HMM para representar as subunidades acústicas. Neste trabalho de síntese de fala, os HMM são utilizados para modelar fones dependentes de contexto.

4.1 Definição de HMM

O HMM é um processo duplamente estocástico, sendo que um dos processos estocásticos não é observável, mas ele pode ser observado através de um conjunto de outros processos estocásticos que produzem a sequência das observações [38]. O processo estocástico observável é dado por um conjunto de estados finitos, onde cada estado é geralmente associado com uma distribuição de probabilidade multidimensional e as transições entre os estados são regidas estatisticamente. No processo estocástico oculto um evento pode ser observado em qualquer estado, de modo que é possível analisar somente as observações geradas sem ver em qual estado ela ocorre, pois os estados são ocultos para o observador.

Os HMMs podem ser vistos como máquinas de estados finitos, onde a cada unidade de tempo ocorre uma transição de estado e, a cada estado emite-se um vetor acústico com uma função densidade de probabilidade associada [38].

A Figura 4.1 ilustra um HMM de 5 estados. Os processos observáveis consistem de um conjunto de vetores de saídas ou observações, sendo que cada um pode ser emitido por cada estado e segue uma função densidade de probabilidade (PDF). Os processos ocultos são formados pelos 5 estados do HMM, estes estados se relacionam através das probabilidades de transição, sendo que a probabilidade de passar do estado i para o estado j é dada por a_{ij}. A cada instante de tempo t existe uma mudança de estado (que pode ser para o mesmo estado) e um símbolo é

emitido com uma determinada densidade de probabilidade de saída $b_i(\boldsymbol{o})$, onde $\boldsymbol{o}$ é o símbolo (vetor de parâmetros) emitido no estado *i,* também denominado de sequência de observações.

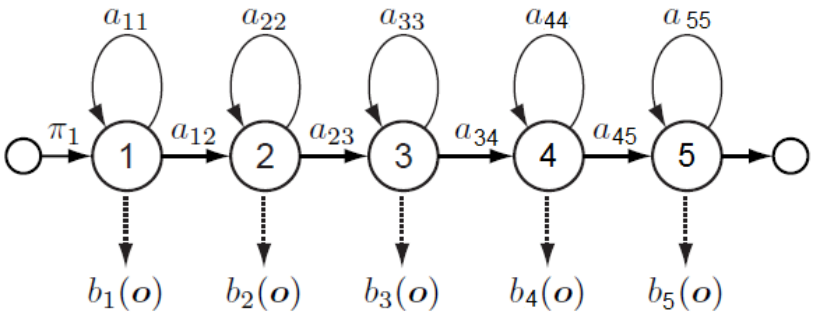

Figura 4.1 – HMM de 5 estados com topologia *left-to-rigth*

Um HMM de *N* estados é definido pelas probabilidades de transição de estado $A = \{a_{ij}\}$ com $1 \leq i,j \leq N$, pelas densidades de probabilidades de saída $B = \{b_j(\boldsymbol{o})\}$ e pela distribuição de probabilidades iniciais dos estados $\Pi = \{\pi_i\}$. A notação adotada para indicar o conjunto de parâmetros do modelo é:

$$\lambda = (A, B, \Pi) \tag{4.1}$$

A probabilidade de saída $b_j(\boldsymbol{o})$ pode ter uma distribuição discreta ou contínua, dependendo do tipo de observação que o HMM modela. Baseado no tipo de distribuição do vetor de observação pode-se classificar os HMM em:

- Discreto: o vetor de observação é discreto e, portanto, utiliza-se um alfabeto finito;
- Contínuo: o vetor de observação é contínuo e utiliza-se uma PDF para descrevê-lo;
- Semicontínuo: combina as vantagens do HMM discreto e do HMM contínuo, construindo dessa forma um modelo intermediário.

Para maiores detalhes sobre os modelos HMMs, sugere-se a consulta dos materiais referenciados em [37] e [38].

Neste trabalho, a probabilidade de saída dos HMMs com distribuição contínua foi modelada por uma mistura de *K* distribuições Gaussianas:

$$b_j(\boldsymbol{o}) = \sum_{k=1}^{K} w_{jk}\, \mathcal{N}(\boldsymbol{o}|\mu_{jk} U_{jk}) \tag{4.2}$$

onde w_{jk} , μ_{jk} e U_{jk} são, respectivamente, o peso, o vetor de médias e a matriz de covariância da componente *k* da mistura no estado *j*. Para satisfazer as restrições estocásticas, deve-se ter:

$$\sum_{k=1}^{K} w_{jk} = 1 \tag{4.3}$$

e

$$\int_{\boldsymbol{o}} b_j(\boldsymbol{o})\, d\boldsymbol{o} = 1 \tag{4.4}$$

com $w_{jk} \geq 0$, $1 \leq j \leq N$ e $1 \leq k \leq K$. A distribuição Gaussiana $\mathcal{N}(\boldsymbol{o}|\mu_{jk} U_{jk})$ é definida como:

$$\mathcal{N}\left(\boldsymbol{o}\middle|\mu_{jk} U_{jk}\right) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{U}|}} \exp\left(-\frac{1}{2}(\boldsymbol{o} - \mu_{jk})^T U_{jk}^{-1} (\boldsymbol{o} - \mu_{jk})\right) \tag{4.5}$$

onde *d* é a dimensão de $\boldsymbol{o}$, $|\boldsymbol{U}|$ é o determinante da matriz de covariância e U_{jk}^{-1} é a matriz de covariância inversa.

Os HMM são usados para resolver basicamente três tipos de problemas que aparecem nas aplicações do mundo real [37]:

Problema 1: Dada uma sequência de observações $\boldsymbol{O}$ = {o_1, o_2, ..., o_T} e um modelo $\lambda = (A, B, \Pi)$, como calcular $P[\boldsymbol{O}|\lambda]$, a probabilidade de observar esta sequência dado o modelo?
Problema 2: Dada a sequência de observações $\boldsymbol{O}$ = {o_1, o_2, ..., o_T} e o modelo $\lambda = (A, B, \Pi)$, como escolher a sequência de estados $\boldsymbol{q}$ = {q_1, q_2, ... , q_T} que seja ótima, ou seja, a que melhor explica a observação?
Problema 3: Como ajustar os parâmetros do modelo $\lambda = (A, B, \Pi)$ para maximizar $P[\boldsymbol{O}|\lambda]$?

Em síntese de fala, encontram-se estes três problemas para serem solucionados ao longo das fases de treinamento e síntese dos modelos HMM.

4.2 Número de Estados do HMM

Na Figura 4.2 vê-se um sinal de fala hipotético correspondente a um fonema. A ideia é descrever matematicamente esta forma de onda por meio de um HMM. Observa-se também que o sinal muda de comportamento ao longo do tempo. Por isso, utiliza-se cada estado do HMM para representar um segmento do fonema, de forma que cada estado apresenta uma densidade de probabilidade dos parâmetros que melhor modelam aquele trecho do sinal.

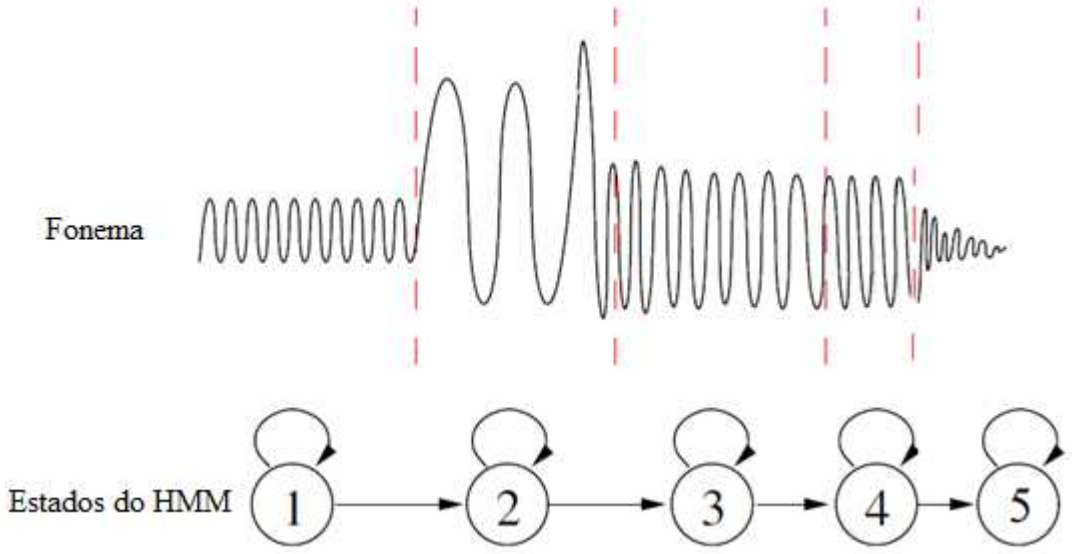

Figura 4.2 – Fonema modelado por HMM
Fonte: adaptado de [40]

É intuitivo deduzir que, quanto mais estados o HMM apresentar, mais perfeitamente os fonemas podem ser modelados. Entretanto, isto resulta em um aumento da complexidade matemática do modelo e em uma estimativa mais pobre dos parâmetros, pois deve-se lembrar que, em síntese de fala, trabalha-se com fones dependentes de contexto, e é necessário ter uma base de fala para o treinamento grande o suficiente para modelar bem todos os estados de todos os fonemas, de forma que ao aumentar o número de estados do HMM é necessário aumentar a base de fala de treinamento para conseguir modelos precisos.

Para compreender a influência do número de estados HMM na qualidade final da síntese, foram treinados HMMs com diferentes números de estados: 2, 3, 5, 7, 9. Este experimento utilizou uma base de fala com 613 frases gravadas em língua portuguesa falada no Brasil. As

frases sintetizadas foram avaliadas subjetivamente por seis ouvintes não especializados. Em geral, os ouvintes não foram capazes de perceber uma melhora nos modelos que utilizavam 7 e 9 estados, provavelmente devido à limitação da base de treinamento. Os modelos que utilizavam de 3 a 7 estados para o HMM, mostraram-se suficientes para obter uma boa qualidade de síntese, mesmo sendo possível distinguir pequenas diferenças de qualidade entre as frases sintetizadas. Foi possível notar uma degradação perceptível quando se utilizaram HMM de 2 estados. O sistema HTS permite construir modelos com no mínimo 2 estados.

Na literatura de síntese de fala encontra-se como padrão modelos HMM com 5 estados e, por isso, este foi o valor adotado ao longo dos experimentos deste trabalho.

4.3 HMM em Síntese de Fala

A topologia do HMM que normalmente se adota para modelar sequências de parâmetros de fala é a *left-to-rigth*, pois os sinais de fala evoluem intrinsicamente de maneira sucessiva no tempo.

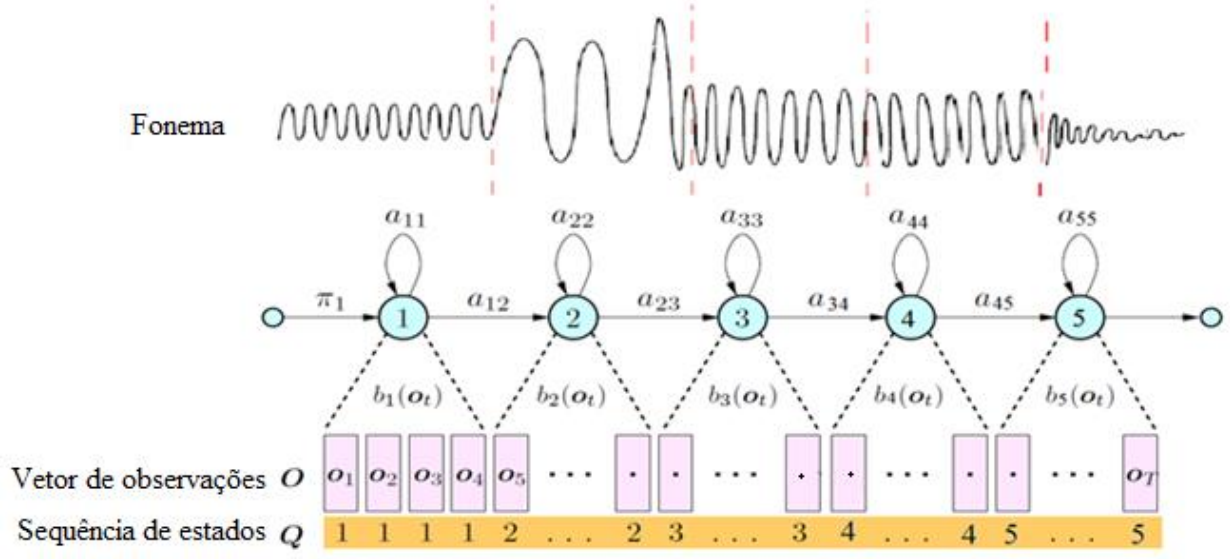

Figura 4.3 – Fonema modelado por HMM com vetor de observações e sequência de estados
Fontes: adaptado de [39,40]

A Figura 4.3 apresenta o modelo HMM de um fonema. A probabilidade de transição de estado a_{ij} determina o número de quadros que um determinado estado ocupa. Entretanto, como é apresentado na Seção 5.5, para controlar a estrutura temporal de forma eficaz, os HMMs possuem as densidades de duração explicitadas. E utilizam-se distribuições Gaussianas para construir o modelo de duração das durações para cada estado HMM. Cada quadro do sinal de fala possui um vetor de observações $\boldsymbol{O}$ com os parâmetros espectrais e de excitação. Deste modo, o modelo HMM de cada fonema possui informações sobre a excitação, espectro e duração dos estados.

5. Síntese usando HMM

Existem diversas formas de implementar um sistema TTS, conforme é apresentado no Capítulo 2. Neste trabalho, a técnica estudada foi a síntese baseada em modelos ocultos de Markov. Este tipo de síntese surgiu em meados da década de 90 para superar algumas limitações existentes na síntese concatenativa, tais como a necessidade de uma biblioteca extensa de polifones pré-gravados para a geração da fala artificial com qualidade e a dificuldade de adaptação das características do locutor e de inserção de emoção no discurso. Esta técnica se baseia em modelos estatísticos HMM e foi proposta por um grupo de pesquisadores, no Instituto de Nagoya e no Instituto de Tecnologia de Tóquio, coordenados por Keiichi Tokuda [5]. O sistema desenvolvido por eles é denominado HTS do acrônimo H Triple S em referência ao nome *HMM-Based Speech Synthesis System.*

O HTS é um software livre que está disponível online [2]. Ele necessita de uma base de fala relativamente pequena para treinamento dos modelos HMM quando comparada ao corpus necessário para a síntese concatenativa. Por ser de natureza estatística e paramétrica, este método é intrinsicamente flexível no que tange à alteração das características da voz, adaptações prosódicas e inserção de emoções no discurso. As técnicas mais comumente utilizadas para este propósito são a de adaptação [41], eigenvoices [42], interpolação [43,44], ou regressão múltipla [45]. Além disso, é possível garantir a inteligibilidade do discurso da fala sintetizada já que o HTS concatena modelos HMM e não sinais de fala pré-gravados. A principal desvantagem do HTS é a falta de naturalidade da voz sintética causada pela perda de variabilidade dos parâmetros estatísticos dos fones durante a fase de treinamento. Este excesso de suavização dá origem a uma fala sintética abafada e monótona. Técnicas como a pós-filtragem [30], o Straight [46], a utilização de parâmetros dinâmicos e da variância global (GV) [47,48] foram propostas visando aprimorar o sistema de modo a obter uma fala sintética mais natural e expressiva.

5.1 Introdução ao HTS

O HTS é uma ferramenta capaz de trabalhar com HMM. Ele permite realizar formulações probabilísticas, gerar parâmetros de interesse, construir modelos acústicos com HMM, efetuar análises espectrais, construir filtros de síntese, ou seja, realizar todos os procedimentos necessários para sintetizar fala.

O sistema HTS foi construído baseado no software *Hidden Markov Model Toolkit* (HTK) [29] já existente desde 1989. O HTK é um programa de código aberto que permite construir e manusear modelos HMM necessários para o reconhecimento de fala. O HTS funciona adaptando algumas rotinas e funções do HTK de modo a permitir a síntese de fala via HMM.

Neste trabalho, utilizou-se a versão HTS-2.2 [2], que foi disponibilizada em julho de 2011. Apesar deste programa fornecer as ferramentas necessárias para trabalhar com os modelos estatísticos, treinando os modelos HMM com fones dependentes de contexto, ele não inclui um analisador de texto. Por isso, a etapa do *Front-End* deve ser realizada por outros programas auxiliares compatíveis, como, por exemplo, o Festival [49] e o MARY [50]. Além dos analisadores de texto, para executar o HTS é necessário instalar alguns programas que servem de suporte à sua execução. O primeiro programa a ser instalado é o HTK, o qual sucessivamente é adaptado para trabalhar com síntese de fala através da aplicação de um patch. Outro programa necessário é o HDecode [29] que, juntamente com o HTK, formam a plataforma de base do HTS. Para executar algumas funções como a extração de parâmetros do corpus de treinamento, pode ser útil utilizar as rotinas já prontas do SPTK [35]. O hts_engine-API [28] fornece rotinas voltadas para desenvolvedores de síntese de fala e possui um processamento mais leve, permitindo que em fase de síntese não sejam necessárias às bibliotecas do HTK. Existe também uma versão voltada para processamento embarcado, denominado flite + hts_engine [49], adaptada para trabalhar com uma versão mais leve do Festival. Esta opção, por enquanto, é disponível apenas para a língua inglesa.

5.2 Síntese de Fala com o HTS

A estrutura de funcionamento do sistema TTS baseado em HMM pode ser dividida em duas etapas: **treinamento** e **síntese**, de acordo com o esquema em blocos apresentado na Figura 5.1. O objetivo da primeira fase é construir os modelos HMM, ou seja, parametrizar os sons da fala matematicamente. Estes modelos permitirão que o sinal de fala artificial seja gerado na segunda fase. As Seções 5.2.1 e 5.2.2 descrevem os detalhes de cada etapa.

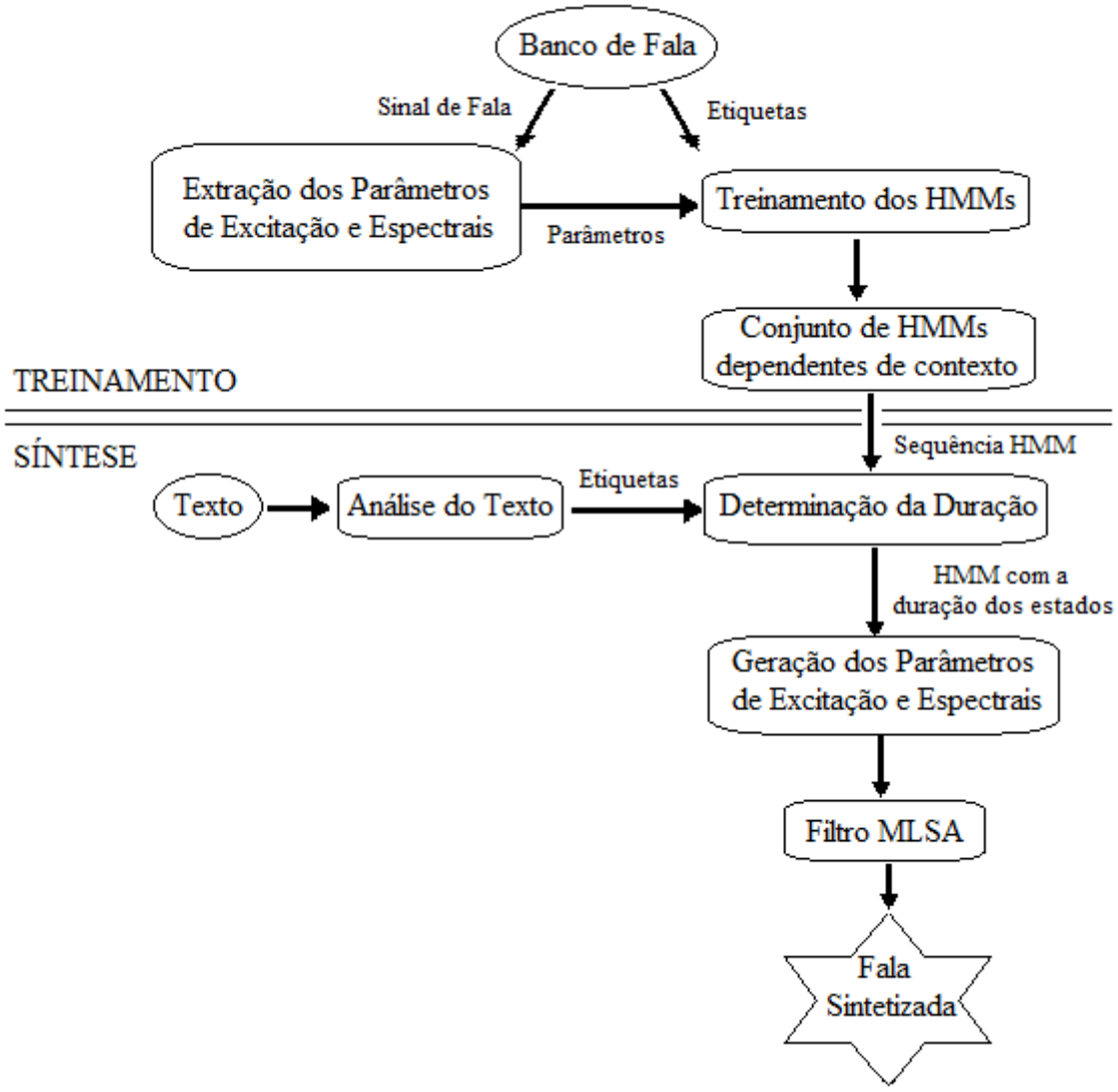

Figura 5.1 – Esquema da síntese via HMM

5.2.1 Treinamento

Primeiramente, é necessário ressaltar que para utilizar o sistema HTS é necessário ter um corpus com frases gravadas em formato *raw,* devidamente transcritas e etiquetadas (ver Apêndice A). O corpus deve ser foneticamente balanceado e diversificado de forma a abranger vários contextos dos fonemas da língua. Quanto maior a base de fala, mais precisos serão os modelos e melhor será a qualidade da voz sintética gerada. Além das gravações e arquivos de transcrição, também são necessários os arquivos de questões que permitem a interpretação das etiquetas e a construção das árvores de decisões. No Apêndice A estão apresentados mais detalhes destes arquivos.

Para realizar o treinamento dos modelos HMM para a língua portuguesa falada no Brasil, utilizaram-se as 613 frases gravadas e etiquetadas que estão disponíveis online [2] em *Speaker dependent training demo Normal Brazilian Portuguese* [13]. As transcrições e etiquetas disponibilizadas contêm alguns erros que prejudicam a construção dos modelos HMM e, consequentemente, a qualidade final do sinal de fala sintetizado. Todavia, estes erros requerem conhecimentos específicos de linguística e, portanto, não foram corrigidos, pois fugiam do escopo deste trabalho. As gravações foram amostradas a 48 kHz, usando 16 bits por amostra e um canal mono. Entretanto, sabendo que a audição humana é limitada, não sendo capaz de ouvir sons acima de 22 kHz [51], decidiu-se dizimar a frequência de amostragem dessas gravações para 16 kHz, de forma a reduzir o número de amostras em um terço, sem incorrer em uma significativa perda de qualidade da fala sintética.

A partir de todos os arquivos de entrada, inicia-se o treinamento com a aquisição dos coeficientes espectrais e da frequência de pitch ($ln\ f_0$) das gravações presentes no banco de dados de fala de um dado locutor. A extração dos parâmetros espectrais é feita janelando o sinal de fala de cada gravação. Tipicamente utilizam-se as janelas de Blackman, Hanning ou Hamming. Neste trabalho utilizou-se janelas de Blackman de 25 ms, com deslocamento de 5 ms. Portanto, os coeficientes espectrais foram calculados a cada quadro de 5 ms.

Depois de obtidos os parâmetros espectrais de todas as frases, inicia-se a extração das frequências de pitch. Utiliza-se o script *getf0.tcl* que necessita do aplicativo ActiveTcl-Tk® com Snack [52]. O método de extração de pitch utilizado é a função de correlação cruzada normalizada [53] com a técnica *Entropic Signal Processing System* (ESPS) [54]. O script tem

como parâmetros de entrada a duração do quadro (5 ms), a frequência de amostragem do sinal (16 kHz) e os limites máximos e mínimos de frequência de pitch válidos, ou seja, o intervalo de frequências de sons considerados sonoros. Como a voz das gravações utilizadas provêm de uma locutora feminina, os valores limites da frequência de pitch foram estabelecidos em 130 Hz – 320 Hz. Este script fornece na saída a sequência de valores de pitch extraída do sinal de fala a cada quadro. Pode-se optar para que os valores estejam na escala de frequência em Hertz ou na forma do logaritmo natural da frequência (ln f_0). Neste trabalho escolheu-se trabalhar com ln f_0 pois, a distribuição de ln f_0 é melhor aproximada pelo modelo da distribuição Gaussiana [55]. Uma vez extraídos os valores de pitch, o programa é capaz de classificar os quadros em sonoro ou não sonoro. O critério adotado foi: os quadros sonoros são aqueles cujo valor do pitch pertence ao intervalo determinado, ou seja, quando f_0 estiver entre 130 Hz e 320 Hz. Os outros quadros são classificados como não sonoros e recebem $f_0 = 0$ ou em escala logarítmica, ln $f_0 = -10^{10}$, que tende a menos infinito.

Os parâmetros espectrais e de pitch extraídos de cada frase são agrupados em arquivos com extensão *cmp,* que podem ser lidos usando o comando *HList* do HTK com a seguinte sintaxe:

HList -C /Demo /data/hlist.conf -h /Demo/data/cmp/portuguese_f001_001.cmp

sendo */Demo* o patch onde estão localizados os arquivos do *demo Normal Brazilian Portuguese.*

De posse desses parâmetros e dos arquivos de etiquetas, inicia-se a determinação dos modelos HMM, de acordo com o processo esquematizado na Figura 5.2. As funções utilizadas pelo HTS estão descritas e especificadas no livro do HTK [56]. Observa-se que, primeiramente, são feitas as iterações e estimativas considerando os 35 fones independentes de contexto (ver Apêndice A-2). Em seguida, são verificados todos os contextos destes fones e geram-se os modelos HMM para cada fone dependente de contexto. O script que realiza este processo de treinamento é o *Training.pl*, que executa as seguintes etapas:

1.Treinamento dos HMMs independentes de contexto

- HCompV: calcula a variância geral. Este valor é calculado levando em conta todos os quadros de todos os arquivos *cmp* (que contém as informações de pitch e dos parâmetros

espectrais) e se cria um arquivo denominado *init.mmf* com 1% dos valores das variâncias gerais para cada parâmetro dos arquivos *cmp*. Estes valores determinam o piso inferior das variâncias ao se estimar os valores dos coeficientes espectrais e dos valores de pitch. HInit: estima os parâmetros usando o algoritmo de Viterbi [56] e a segmentação uniforme para cada estado do HMM que modela o fone. No final deste processo tem-se a estatística relacionada às probabilidades de saída de cada estado $b_i(o)$ e as informações referentes ao estado inicial, respeitando π=1, já que os modelos são *left-to-right* e iniciam sempre no primeiro estado.

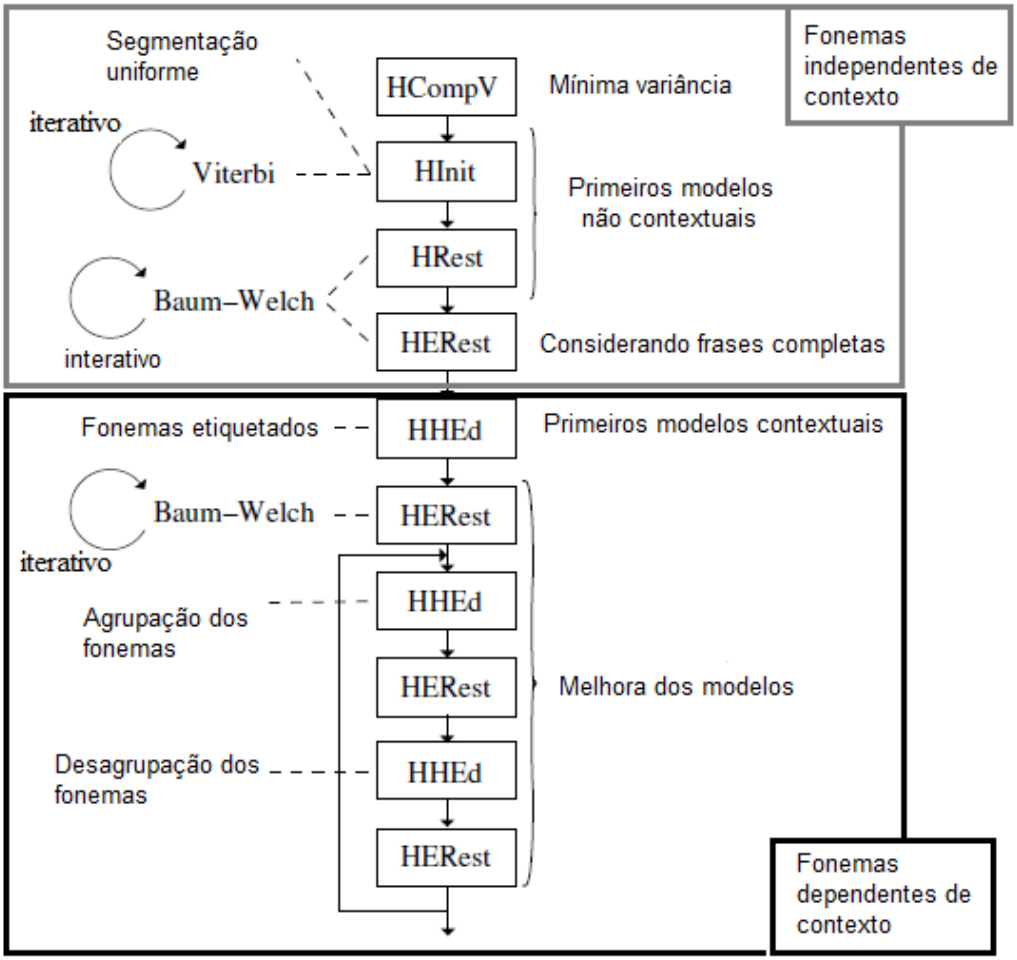

Figura 5.2 – Esquema de treinamento do HTS
Fonte: adaptado de [40]

- HRest: os parâmetros são re-estimados usando o algoritmo de Baum-Welch [56], gerando as primeiras estimativas das probabilidades de transição de estados.

- HERest: nesta etapa considera-se toda a frase de maneira global, ao invés de considerar cada fone separadamente como nas fases anteriores. Para cada sentença constrói-se um HMM concatenando-se os modelos obtidos anteriormente. Depois, aplica-se o algoritmo de Baum-Welch na frase completa para obter as probabilidades dos estados. Em seguida reestimam-se os parâmetros. O processo realiza 5 iterações ao todo.
- HHEd: gera os HMMs dependentes de contexto a partir dos HMMs independentes de contexto por simples cópia dos modelos. O número de HMMs dependentes de contexto está relacionado ao número de diferentes contextos determinados pelas etiquetas.

2. Modelamento dos HMMs dependentes de contexto

- HERest: aplicando-se novamente o algoritmo de Baum-Welch em cada frase da base de dados se reestima, com uma iteração, as estatísticas de saída para os parâmetros espectrais e de pitch, e constrói-se também as estatísticas de duração, baseando-se nas informações de duração contextual fornecidas nos dados de entrada. Aqui é necessário lembrar que a base de treinamento está segmentada em nível de fones. O modelo de duração resulta independente das probabilidades de transição a_{ij}.
- HHEd: como os modelos contextuais são muito variados, é natural que nem todos os casos estejam presentes na base de treinamento. Porém, é razoável considerar que, mesmo em contextos diferentes, alguns fones podem apresentar parâmetros semelhantes, de forma que a partir dos modelos existentes, podem-se inferir os modelos ausentes. Para isso é utilizada a técnica de *clustering* [32] e constroem-se árvores de decisão levando em conta o arquivo de questões (*questions)* e as etiquetas das frases (ver Apêndice A). Como os modelos de espectro, frequência de pitch e duração são afetados diferentemente pelo contexto, são construídas três diferentes árvores.
- HERest: novamente os parâmetros são reestimados para cada frase com o algoritmo de Baum-Welch. São considerados os modelos dependentes de contexto e são realizadas 5 iterações. Como os parâmetros estão agrupados, as alterações comprometem todos os modelos. Por exemplo, considerando que o primeiro estado do fone *'a'* nos contextos 'e^m-**a**+r=t' e 'a^m-**a**+g=e' estão agrupados, a estimação dos parâmetros deste agrupamento afetará todas as frases que contenham estes contextos. Deste modo, estes agrupamentos permitem relacionar melhor as semelhanças entre os fones em diferentes

contextos e melhorar a qualidade da estimação dos HMMs usando uma base de dados limitada.

- HHEd: desagrupa os parâmetros dos modelos espectrais, de pitch e de duração
- HERest: novamente são refeitas as estimativas considerando os modelos desagrupados.

Na sequência são repetidos os quatro últimos passos de forma análoga, ou seja, os modelos são reagrupados, estimados, desagrupados e estimados.

Considerando um corpus com 613 frases, a fase de treinamento dos modelos HMM consome em geral cerca de 4 horas empregando um computador de bom desempenho, com 8 GB de memória RAM , processador Intel i7 – 2600 com clock de 3,40 GHz.

Ao final do processo de treinamento obtêm-se os arquivos *mgc.pdf, lf0.pdf* e *dur.pdf,* que contém as funções densidade de probabilidade dos modelos HMM dependentes de contexto. Estes arquivos, juntamente com os arquivos das árvores de decisão *tree-mgc.inf, tree-lf0.inf, tree-dur.inf,* permitem realizar a síntese de fala usando o hts_engine-API. O conteúdo dos arquivos com extensão *inf* pode ser visualizado com um programa comum de editoração de texto; já os arquivos com extensão *pdf* estão em formato binário e podem ser lidos usando o seguinte comando do SPTK:

./swab +f nome.pdf | *./dmp +i* | *less* -> cabeçalho

./swab +f nome.pdf | *./dmp +f* | *less* -> parâmetros

No momento da síntese, a duração dos estados é determinada pela sua respectiva árvore de decisão. Deve-se considerar que cada estado HMM apresenta uma árvore de decisão para o pitch e outra para o espectro, conforme ilustra a Figura 5.3. Sabendo que cada nó final da árvore de decisão remete a um dado valor na PDF, consegue-se compreender a ordem de grandeza das estruturas e a complexidade associada ao sistema de síntese.

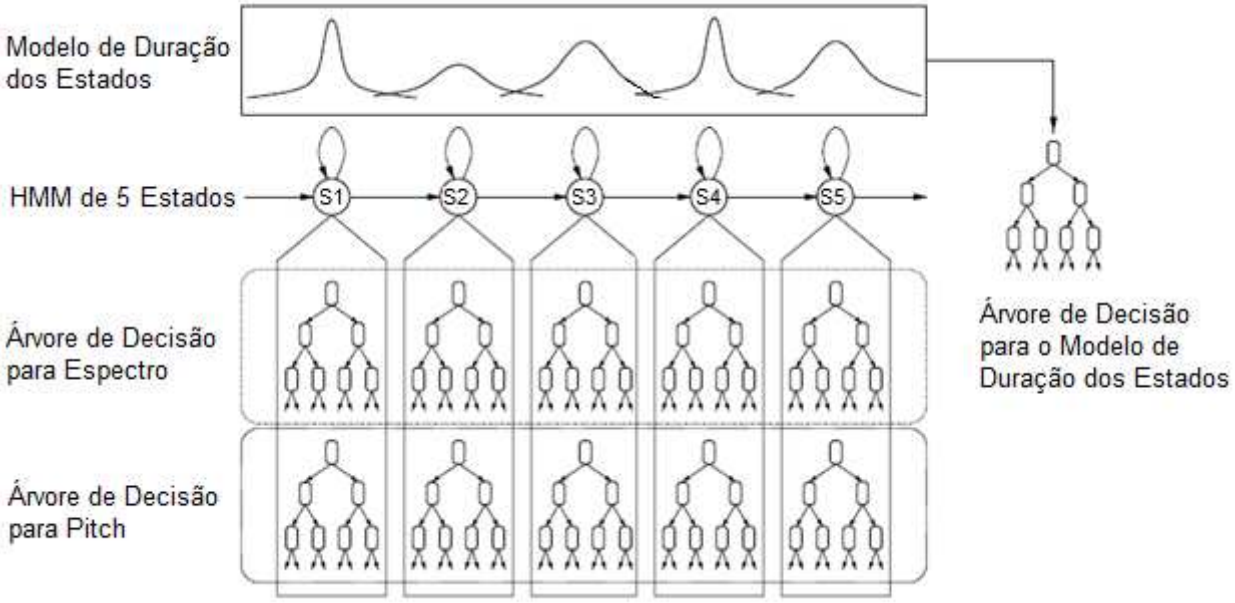

Figura 5.3 – Árvores de decisão associadas ao HMM
Fonte: adaptado de [57]

Para compreender um pouco a ordem de grandeza dos vetores dos HMM, apresenta-se a estrutura dos arquivos *pdf* gerados no treinamento dos HMM com 5 estados, considerando que o modelo espectral usa vetores de ordem 24, que totalizam 75 coeficientes entre estáticos e dinâmicos (ver Seções 3.2 e 5.2) e que o modelo de excitação tem ordem 1 acrescido dos dois coeficientes dinâmicos, com tamanho total de 3.

O cabeçalho do arquivo com informações dos coeficientes espectrais *mgc.pdf* é:

mgc.pdf

0 0 => mistura de apenas uma distribuição
1 1 => apenas uma distribuição
2 75 => tamanho do vetor
3 167 => número de nós do estado 1 do HMM
4 199 => número de nós do estado 2 do HMM
5 194 => número de nós do estado 3 do HMM
6 188 => número de nós do estado 4 do HMM
7 195 => número de nós do estado 5 do HMM

O número total de nós considerando os 5 estados HMM é igual a 943. Cada nó tem associado um vetor de 75 posições para média e outro de igual tamanho para a variância. Portanto, existe um total de 141450 entradas na PDF que modela o espectro, pois 141450 = 943 x 75 x 2.

Analogamente, para o *lf0.pdf*, tem-se:

lf0.pdf

0 1 => mistura de duas distribuições

1 3 => tamanho do vetor da distribuição discreta

2 3 => tamanho do vetor da distribuição contínua

3 364 => número de nós do estado 1 do HMM

4 612 => número de nós do estado 2 do HMM

5 706 => número de nós do estado 3 do HMM

6 515 => número de nós do estado 4 do HMM

7 357 => número de nós do estado 5 do HMM

O número total de nós é 2554, para os 5 estados HMM. Cada nó tem associado um vetor de 3 posições para média ($\ln f_0$, $\Delta \ln f_0$ e $\Delta^2 \ln f_0$) e outro de mesma dimensão para a variância. Considerando que existe uma mistura de duas distribuições, uma discreta e uma contínua, para modelar as componentes sonoras e não sonoras da excitação (ver Capítulo 5.4) tem-se, portanto 2 médias e 2 variâncias, uma para cada distribuição. Assim, o total de entradas da PDF que descreve o modelo de excitação é: 2554 x 3 x 4 = 30648.

O cabeçalho do modelo de duração tem a seguinte forma:

dur.pdf

0 0 => mistura de apenas uma distribuição

1 5 => HMM com 5 estados

2 5 => tamanho do vetor

3 283 => número de nós

O modelo de duração considera cada posição do vetor referente a um estado do HMM. Assim o número total de entradas da PDF, considerando que cada estado terá uma média e uma variância associada é: 283 x 5 x 2 =2830.

5.2.2 Síntese

O procedimento de síntese de fala de um dado texto pode ser representado esquematicamente na Figura 5.4, onde foram considerados, apenas para facilitar a ilustração, modelos HMM de três estados. Entretanto, as simulações realizadas neste estudo empregaram geralmente HMM com cinco estados. Conforme se observa na Figura 5.4, o sistema de síntese pode ser subdividido em quatro fases: primeiramente o texto a ser sintetizado deve ser transcrito foneticamente e adequadamente etiquetado, considerando os fatores contextuais utilizando o mesmo padrão empregado nas frases do corpus de treinamento (ver Apêndice A). Com base nas informações fonéticas contextuais destes arquivos é possível definir a sequência dos modelos HMMs dependentes de contexto que descreve os fonemas. Na segunda fase, são determinadas as durações dos estados para a sequência HMM, a partir da distribuição PDF da duração dos estados (*dur.pdf)* e da sua árvore de decisão (*tree-dur.inf)*. O critério utilizado é maximizar a probabilidade de saída [58].

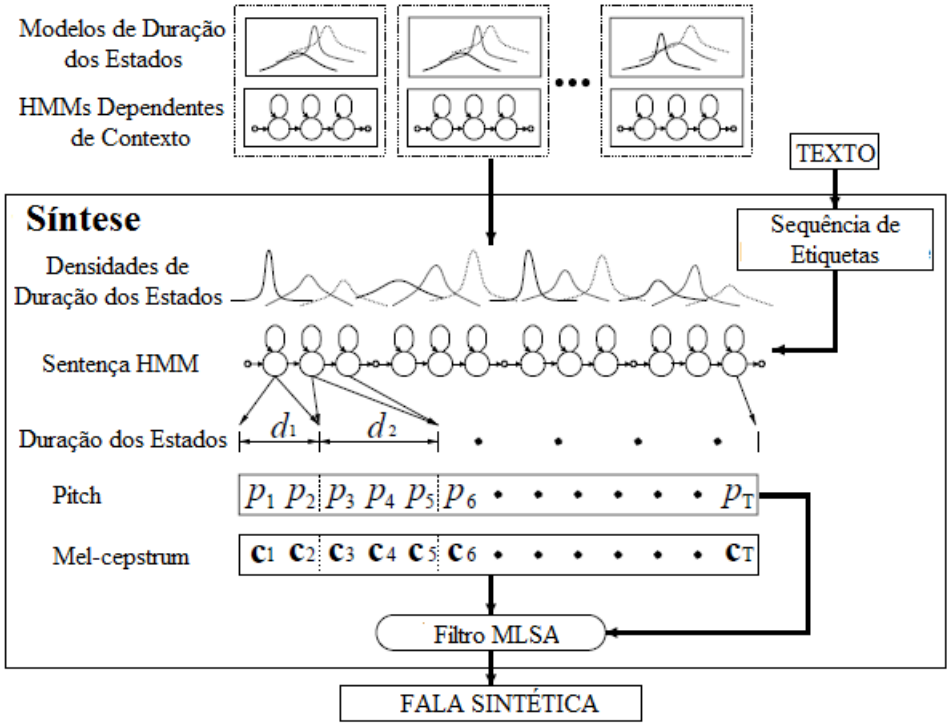

Figura 5.4 – Esquema de síntese
Fonte: adaptado de [57]

Assim, após seguir as questões contextuais da árvore de decisão da duração de estados, o algoritmo encontrará no nó final de cada fone dependente de contexto uma informação do tipo *'dur_s2_19'.* Isto indicará que o valor desejado no modelo de duração (*dur.pdf)* corresponde ao nó 19 do estado 1 '*s2*'. Uma vez associado o modelo de duração a todos os estados de todos os fonemas a serem sintetizados, passa-se para a terceira etapa, na qual se obtém os modelos espectrais e de excitação para cada estado de cada fonema do texto utilizando as respectivas árvores de decisão e arquivos *pdf. D*esta forma, o algoritmo consegue determinar a trajetória dos coeficientes mel-cepstrais e dos valores de *ln f0*, incluindo as decisões de sonoro/não sonoro, através do algoritmo de geração dos parâmetros de fala, baseado no critério de maximização da verossimilhança, conforme é descrito na próxima seção. Finalmente, na quarta fase, a forma de onda é sintetizada a partir dos parâmetros espectrais e de excitação usando o filtro MLSA.

5.2.3 Geração dos Parâmetros de Fala

Para obter a sequência de parâmetros espectrais e de pitch, necessários para realizar a síntese de fala, a partir dos modelos HMM treinados, adota-se o critério da máxima verossimilhança [30].

Considere um HMM λ com uma sequência de estados $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ e um vetor de observações $\boldsymbol{O} = \{o_1, o_2, \ldots, o_T\}$, sendo T o número total de quadros da sentença que se deseja sintetizar. Deseja-se encontrar a sequência de parâmetros $\overline{\boldsymbol{O}}$ que maximiza $P[\boldsymbol{O}|\lambda]$ com respeito a $\boldsymbol{O}$, ou seja:

$$\overline{\boldsymbol{O}} = \underbrace{argmax}_{O} P[\boldsymbol{O}|\lambda] \tag{5.1}$$

Este problema não pode ser resolvido analiticamente, mas considerando que:

$$P[\boldsymbol{O}|\lambda] = \sum_{\boldsymbol{q}} P[\boldsymbol{q}, \boldsymbol{O}|\,\lambda] \tag{5.2}$$

pode-se aproximar o problema, de modo a encontrar a sequência dos parâmetros de fala que maximizam esta probabilidade com respeito a $\boldsymbol{O}$ e a $\boldsymbol{q}$, ou seja:

$$\bar{\boldsymbol{O}} = \underbrace{max}_{\boldsymbol{q}} \underbrace{argmax}_{\boldsymbol{O}} P[\boldsymbol{q}, \boldsymbol{O}|\lambda] \tag{5.3}$$

Se não houver uma restrição que ligue os valores dos parâmetros do quadro *t* independente dos quadros *t+1* e *t-1*, a solução desta maximização será o vetor de médias dos estados, ou seja, serão os valores médios das Gaussianas que modelam cada estado do HMM (veja a Seção 5.3 para maiores detalhes sobre as distribuições PDF adotadas para o modelo espectral). Adotar os valores médios causa uma descontinuidade abrupta nas transições dos quadros de diferentes estados gerando um sinal de fala de baixa qualidade acústica. Nas duas seções seguintes (5.2.3.1 e 5.2.3.2) são discutidos dois modos de reduzir estes saltos.

5.2.3.1 Interpolação Linear dos Parâmetros

Com o objetivo de suavizar as transições entre os estados do HMM e entre diferentes HMMs associados aos diferentes fones dependentes de contexto, pensou-se em realizar uma interpolação linear dos parâmetros. Para aplicá-la, utilizou-se a ferramenta de síntese hts_engine-API e alterou-se um trecho da função *gstream.c* de modo que os parâmetros que fazem fronteira entre as transições dos estados fossem recalculados e permitissem que a transição ocorresse de maneira mais suave. A Figura 5.5 ilustra qualitativamente a ideia da interpolação empregada.

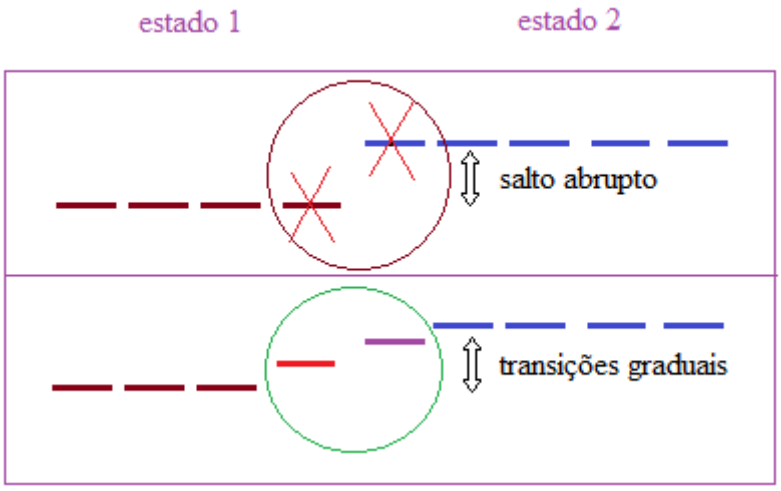

Figura 5.5 – Esquema de interpolação linear dos parâmetros

Uma análise acústica subjetiva demonstrou que a técnica de interpolação linear implica em uma melhora perceptível na qualidade da fala artificial. As Figuras 5.6 e 5.7 mostram o espectrograma da parte sublinhada da frase *"Apenas os ônibus circularão pela pista bairro-centro nos dois sentidos*" dos sinais de fala sintetizados usando os coeficientes estáticos nos dois casos, com e sem interpolação linear dos parâmetros. Das figuras, resulta evidente que as transições ocorrem de forma mais suavizada quando se utiliza a interpolação linear dos parâmetros. Entretanto, apesar da considerável melhora na qualidade sonora do sinal de fala sintético, com uma técnica que exige baixa complexidade computacional, esta solução não alcança os níveis de qualidade do método proposto na literatura que utiliza os coeficientes dinâmicos [5]. Conforme é apresentado na próxima seção, este método é custoso computacionalmente, todavia a qualidade da síntese é inquestionavelmente melhor.

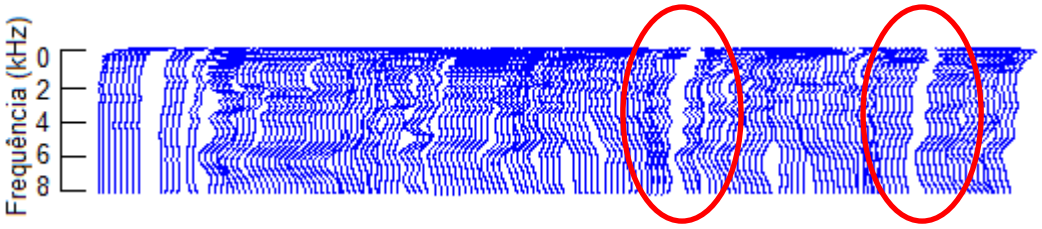

Figura 5.6 – Espectrograma usando parâmetros estáticos

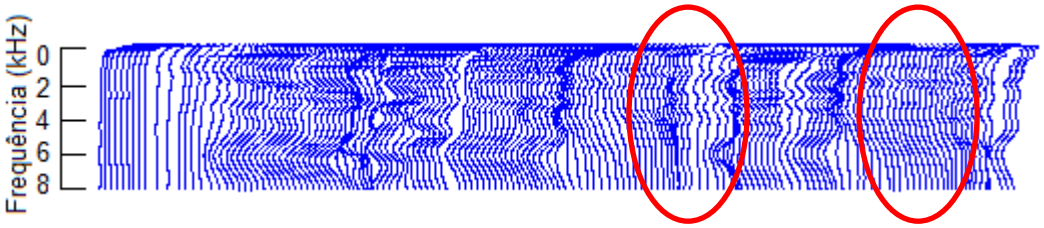

Figura 5.7 – Espectrograma usando interpolação linear

5.2.3.2 Coeficientes Dinâmicos

Os coeficientes dinâmicos foram propostos para resolver o problema da perda de qualidade do sinal sintetizado devido às mudanças bruscas entre os parâmetros de um estado para outro do HMM. Eles introduzem as ideias de primeira e segunda derivada na sequência de coeficientes, as quais permitem saber como variam os coeficientes dos quadros vizinhos, servindo assim de restrição a respeito da variação dos coeficientes estáticos.

Seja o vetor dos coeficientes estáticos dado por:

$$\boldsymbol{c} = \begin{bmatrix} c_1(0) \\ \vdots \\ c_1(M) \\ \vdots \\ c_T(0) \\ \vdots \\ c_T(M) \end{bmatrix}$$

Define-se a derivada primeira dos coeficientes:

$$\Delta c_t^{(i)} = \frac{c_{t+1}^{(i)} - c_{t-1}^{(i)}}{2} \tag{5.4}$$

onde $c_t^{(i)}$ representa as componentes do vetor $\boldsymbol{c}$, com $t=1,...,T$ e $i= 0, ...,M$, sendo T o número total de quadros da sentença que se deseja sintetizar e M a ordem do filtro.

A derivada segunda dos coeficientes estáticos é dada por:

$$\Delta^2 c_t^{(i)} = \frac{\Delta c_{t+1}^{(i)} - \Delta c_{t-1}^{(i)}}{2}$$

$$\Delta^2 c_t^{(i)} = \frac{\left(\frac{c_{t+2}^{(i)} - c_t^{(i)}}{2}\right) - \left(\frac{c_t^{(i)} - c_{t-2}^{(i)}}{2}\right)}{2}$$

$$\Delta^2 c_t^{(i)} = \frac{c_{t+2}^{(i)} - 2c_t{}^{(i)} + c_{t-2}^{(i)}}{4}$$

que devido à simetria, pode ser calculado considerando os termos mais adjacentes a t com a seguinte expressão:

$$\Delta^2 c_t^{(i)} = c_{t+1}^{(i)} - 2c_t{}^{(i)} + c_{t-1}^{(i)} \tag{5.5}$$

Desta forma, têm-se os vetores de parâmetros dinâmicos com a seguinte estrutura:

$$\Delta \boldsymbol{c} = \begin{bmatrix} \Delta c_1(0) \\ \vdots \\ \Delta c_1(M) \\ \vdots \\ \Delta c_T(0) \\ \vdots \\ \Delta c_T(M) \end{bmatrix} \qquad \Delta^2 \boldsymbol{c} = \begin{bmatrix} \Delta^2 c_1(0) \\ \vdots \\ \Delta^2 c_1(M) \\ \vdots \\ \Delta^2 c_T(0) \\ \vdots \\ \Delta^2 c_T(M) \end{bmatrix}$$

Considera-se agora que o vetor de observações é da forma: $o_t = \{ c_t, \Delta c_t, \Delta^2 c_t \}$. Em termos matriciais, tem-se:

$$\boldsymbol{O = Wc} \tag{5.6}$$

com

$$\boldsymbol{O} = \begin{bmatrix} \boldsymbol{c} \\ \Delta \boldsymbol{c} \\ \Delta^2 \boldsymbol{c} \end{bmatrix} \qquad \boldsymbol{W} = \begin{bmatrix} w_1^{(0)} & w_2^{(0)} & \dots & w_T^{(0)} \\ w_1^{(1)} & w_2^{(1)} & \dots & w_T^{(1)} \\ w_1^{(2)} & w_2^{(2)} & \dots & w_T^{(2)} \end{bmatrix}$$

onde os valores da matriz $\boldsymbol{W}$ são constantes que relacionam as matrizes $\boldsymbol{O}$ e $\boldsymbol{c}$, de forma que as expressões (5.4) e (5.5) sejam obedecidas. Assume-se que $c_t = 0$ para $t < 1$ e para $t > T$.

Para facilitar o entendimento, a matriz $\boldsymbol{W}$ é explicitada considerando um caso hipotético no qual $M = 1$ e $T = 3$.

$$\begin{bmatrix} c_1(0) \\ c_1(1) \\ c_1(2) \\ c_2(0) \\ c_2(1) \\ c_2(2) \\ c_3(0) \\ c_3(1) \\ c_3(2) \\ \Delta c_1(0) \\ \Delta c_1(1) \\ \Delta c_1(2) \\ \Delta c_2(0) \\ \Delta c_2(1) \\ \Delta c_2(2) \\ \Delta c_3(0) \\ \Delta c_3(1) \\ \Delta c_3(2) \\ \Delta^2 c_1(0) \\ \Delta^2 c_1(1) \\ \Delta^2 c_1(2) \\ \Delta^2 c_2(0) \\ \Delta^2 c_2(1) \\ \Delta^2 c_2(2) \\ \Delta^2 c_3(0) \\ \Delta^2 c_3(1) \\ \Delta^2 c_3(2) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hdashline 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ -0.5 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & -0.5 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & -0.5 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & -0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.5 & 0 & 0 & 0 \\ \hdashline -2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -2 \end{bmatrix} . \begin{bmatrix} c_1(0) \\ c_1(1) \\ c_1(2) \\ c_2(0) \\ c_2(1) \\ c_2(2) \\ c_3(0) \\ c_3(1) \\ c_3(2) \end{bmatrix}$$

Assim, o problema posto de maximizar a probabilidade $P[\boldsymbol{q}, \boldsymbol{O}|\lambda]$ para uma dada sequência $\boldsymbol{q}$ com respeito a $\boldsymbol{c}$ pode ser reformulado, considerando que:

$$P[\boldsymbol{q}, \boldsymbol{O}|\lambda] = P[\boldsymbol{q}|\lambda].P[\boldsymbol{O}|\boldsymbol{q}, \lambda] \tag{5.7}$$

Como $P[\boldsymbol{q}|\lambda]$ não depende do vetor de observação $\boldsymbol{O}$, conclui-se que, para uma dada sequência $\boldsymbol{q}$, maximizar $P[\boldsymbol{q}, \boldsymbol{O}|\lambda]$ com respeito a $\boldsymbol{c}$ equivale a maximizar $P[\boldsymbol{O}|\boldsymbol{q}, \lambda]$ com respeito a $\boldsymbol{c}$. Considerando que os parâmetros espectrais são modelados por uma Gaussiana de dimensão $3(M+1) = 75$, esta probabilidade pode ser escrita como:

$$P[\boldsymbol{O}|\boldsymbol{q},\lambda] = \frac{1}{(2\pi)^{\frac{3MT}{2}}|\boldsymbol{U}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{O}-\boldsymbol{\mu})^T.\boldsymbol{U}^{-1}.(\boldsymbol{O}-\boldsymbol{\mu})} \tag{5.8}$$

onde $\boldsymbol{\mu} = \{\mu_{s_1}, \mu_{s_2}, \ldots, \mu_{s_T}\}$ é o vetor coluna dos valores médios das observações de um estado no instante *t* e $\boldsymbol{U} = \{U_{s_1}, U_{s_2}, \ldots, U_{s_T}\}$ é a matriz de covariância das observações para um estado no instante *t*. Assumindo que as componentes do vetor $\boldsymbol{c}$ são independentes entre si, a matriz $\boldsymbol{U}$ resulta diagonal, simplificando os cálculos.

Aplicando a função logaritmo em (5.8), tem-se:

$$ln\, P[\boldsymbol{O}|\boldsymbol{q},\lambda] = -\frac{1}{2}[(\boldsymbol{O}-\boldsymbol{\mu})^T.\boldsymbol{U}^{-1}.(\boldsymbol{O}-\boldsymbol{\mu})] - \frac{1}{2}ln|\boldsymbol{U}| - \frac{3MT}{2}\ln(2\pi) \tag{5.9}$$

Assim, para maximizar $P[\boldsymbol{O}|\boldsymbol{q},\lambda]$ com respeito a $\boldsymbol{c}$, basta calcular:

$$\frac{\partial P[\boldsymbol{O}|\boldsymbol{q},\lambda]}{\partial \boldsymbol{c}} = \frac{\partial P[\boldsymbol{W}.\boldsymbol{c}|\boldsymbol{q},\lambda]}{\partial \boldsymbol{c}} = 0 \tag{5.10}$$

que resulta em:

$$(\boldsymbol{W}^T\boldsymbol{U}^{-1}\boldsymbol{W})\boldsymbol{c} = \boldsymbol{W}^T\boldsymbol{U}^{-1}\boldsymbol{\mu} \tag{5.11}$$

A solução de (5.11) determina a sequência dos parâmetros $\boldsymbol{c}$ que maximiza $P[\boldsymbol{O}|\boldsymbol{q},\lambda]$. Entretanto, para resolver o problema, diretamente a partir desta equação, é necessário um número de operações muito grande, da ordem de T^3M^3, lembrando que $\boldsymbol{M}$ é a ordem do filtro MLSA e $\boldsymbol{T}$ é o número de quadros da frase, sendo que cada quadro tem 5 ms. Todavia, pode-se aplicar o método de Cholesky e, com algumas manipulações matemáticas convenientes, consegue-se reduzir o número de operações para algo da ordem de TM, tornando este método viável para síntese de fala em dispositivos móveis, por exemplo.

De forma análoga à empregada para os parâmetros espectrais, pode-se gerar os parâmetros do modelo de excitação para os quadros que apresentam sons sonoros. Os quadros com sons não sonoros são modelados por ruído.

A Figura 5.8 ilustra o vetor de observações composto de dois blocos, um contendo os parâmetros espectrais e outro as informações dos parâmetros de excitação, ambos com os coeficientes estáticos e dinâmicos. Esta é a estrutura do vetor de saída de cada estado do modelo HMM.

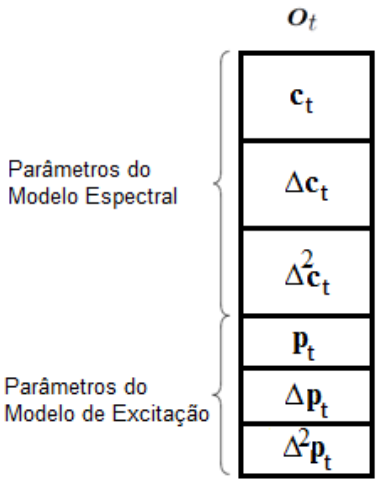

Figura 5.8 – Vetor de saída de cada estado do HMM

5.2.3.3 Simulações com os Coeficientes Dinâmicos

Para compreender a influência na qualidade da fala sintética dos coeficientes dinâmicos nos modelos espectrais e de excitação, foram feitas algumas análises. Realizou-se o treinamento de quatro diferentes modelos, realizando todas as possíveis combinações, conforme consta na Tabela 5.1.

Tabela 5.1 – Experimento com os parâmetros estáticos e dinâmicos

Caso	Parâmetros de Excitação	Parâmetros Espectrais
A	Estático + Dinâmico	Estático + Dinâmico
B	Estático	Estático + Dinâmico
C	Estático + Dinâmico	Estático
D	Estático	Estático

Foram sintetizadas frases usando os modelos gerados em cada um dos casos de treinamento. As Figuras 5.9, 5.10, 5.11 e 5.12 apresentam os espectrogramas gerados para um

mesmo trecho de sinal de fala sintetizado referentes à parte sublinhada da sentença *"Apenas os ônibus circularão pela pista bairro-centro nos dois sentidos"* para cada um dos casos apresentados na Tabela 5.1. Comparando os trechos circulados em cada um dos espectrogramas, é possível notar que ocorre uma sucessão mais suave para os espectrogramas do caso A e B, do que para os casos C e D. Uma análise subjetiva demonstrou que o caso A é o que produz o sinal de fala mais agradável. Os ouvintes consideraram as frases sintetizadas no caso D como as mais desagradáveis, pois era possível distinguir acusticamente as transições dos estados. O caso B apresenta uma qualidade ligeiramente inferior ao encontrado no caso A, enquanto que o caso C apresenta uma ligeira melhora em relação ao caso D. Assim, é imediato concluir que os coeficientes dinâmicos agregam uma melhoria na qualidade da síntese, sendo que eles exercem uma maior influência no modelo espectral que no modelo de excitação.

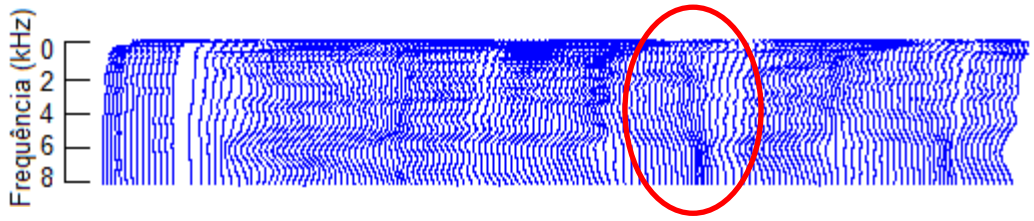

Figura 5.9 – Espectrograma – caso A

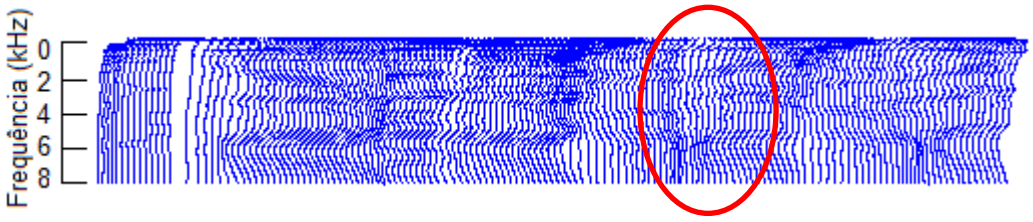

Figura 5.10 – Espectrograma – caso B

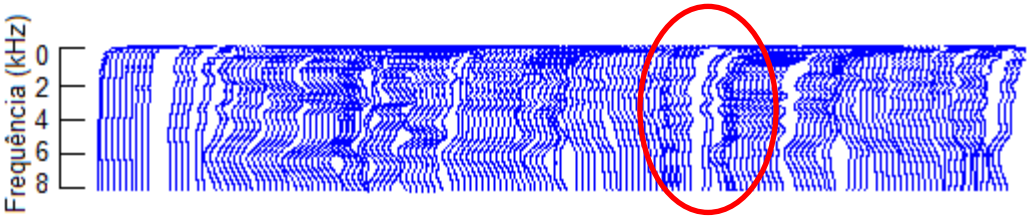

Figura 5.11 – Espectrograma – caso C

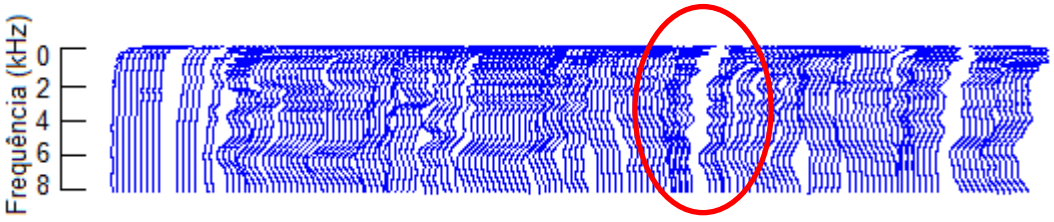

Figura 5.12 – Espectrograma – caso D

5.3 Modelo Espectral

As características do modelo espectral afetam a qualidade da fala sintética uma vez que ele é responsável por modelar as características do trato vocal, conforme se verificou nos testes descritos na seção 5.2.3.3. Assim, no HTS o espectro geralmente é modelado utilizando os coeficientes mel-cepstrais ou cepstrais, juntamente com os parâmetros dinâmicos correspondentes à primeira e segunda derivada, respectivamente os parâmetros delta e delta-delta.

Para modelar a densidade de probabilidade de saída dos coeficientes estáticos e dinâmicos, utiliza-se uma mistura de funções de densidade de probabilidade Gaussianas. Neste trabalho utilizou-se a mistura de uma única Gaussiana no modelo e consideraram-se os coeficientes independentes entre si de forma a se ter uma matriz de covariância diagonal.

Os coeficientes são escolhidos estatisticamente para projetar o filtro e, a fim de obter fala sintética de boa qualidade, ou seja, que não seja degradada e não sofra *clipping,* é necessário assegurar a estabilidade do filtro projetado. Conforme apresentado no Capítulo 3, o filtro utilizado pelo HTS é o MLSA, que devido à sua estrutura garante sempre a estabilidade.

5.3.1 Experimentos com o Filtro MLSA

Para compreender as influências dos parâmetros utilizados na construção do filtro MLSA, foram realizados alguns experimentos. Primeiramente, foram treinados HMMs cujo modelo espectral utilizava diferentes tipos de coeficientes para projetar o filtro e utilizando estes modelos foram sintetizadas algumas frases. A Tabela 5.2 mostra as principais combinações de ordem e tipo de análises testadas, considerando que a frequência de amostragem utilizada foi de 16 kHz. As análises mel-cepstral e cepstral fornecem diretamente os coeficientes do filtro MLSA, que é sempre estável. Quando se trabalha com os outros tipos de análises, deve-se fazer uma transformação desses coeficientes para coeficientes do tipo LSP de forma a garantir a estabilidade do filtro que será projetado. Consultando a literatura da área de síntese de fala, constatou-se que o tipo de análise utilizada com mais frequência é a mel-cepstral com ordem 24 (para frequência de amostragem de 16 kHz). De acordo com Kim [59], a análise MGC-LSP de ordem 18 proporciona qualidade equivalente à mel-cepstral de ordem 24.

Tabela 5.2 – Análises testadas para construir o filtro		
Tipo de Análise	**Ordem do Filtro**	**(α; γ)**
LSP	12	(0; -1)
LSP	16	(0; -1)
LSP	18	(0; -1)
Mel-Cepstral Generalizado	24	(0,42; -1/3)
Mel-Cepstral Generalizado	24	(0,42; -1/2)
Mel-Cepstral	16	(0,42; 0)
Mel-Cepstral	20	(0,42; 0)
Mel-Cepstral	24	(0,42; 0)
Cepstral	16	(0; 0)
Cepstral	24	(0; 0)

Testes subjetivos, utilizando as combinações propostas na Tabela 5.2, revelaram a preferência dos ouvintes pelas frases sintetizadas utilizando os filtros projetados com os coeficientes mel-cepstrais e cepstrais. Por isso, foram realizados testes adicionais levando em conta estes dois tipos de análises, com o intuito de compreender a influência que a ordem do filtro tem sobre a qualidade final da síntese. Para isto, foram utilizadas gravações do som sonoro *'a'* e do som não sonoro *'s'* da língua portuguesa. A taxa de amostragem utilizada foi de 16 kHz com 16 bits por amostra. Para a extração dos coeficientes cepstrais e mel-cepstrais o sinal gravado foi janelado usando janela de Blackman de 25 ms com deslocamento de 5 ms. Configurando as variáveis $\alpha = 0$ e $\gamma = 0$, obteve-se os coeficientes cepstrais e, fazendo $\alpha = 0{,}42$ e $\gamma = 0$, obteve-se os coeficientes mel-cepstrais. O comando utilizado foi o *'mcep'* [60] do SPTK, que aplica uma função de custo baseada no método UELS [33], e utiliza o algoritmo de Newton-Raphson para minimizá-la.

5.3.1.1 Considerações sobre a Ordem do Filtro

A expressão do filtro MLSA é dada em (3.2) e reportada novamente aqui para facilitar o entendimento. Observe que um filtro de ordem M apresenta $M+1$ coeficientes.

$$H(z) = \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m} \tag{5.12}$$

A ideia inicial é verificar como varia a qualidade da síntese ao variar a ordem *M* do filtro. Para isso foram projetados filtros com diversas ordens usando os coeficientes mel-cepstrais e cepstrais.

Considerando primeiramente os coeficientes mel-cepstrais, vê-se na Figura 5.13 o espectro $H(e^{j\omega})$ do som sonoro *'a'*, juntamente com os filtros obtidos considerando três casos: $M = 12$, $M = 24$ e $M = 48$. Analogamente, a Figura 5.14 apresenta as curvas obtidas para o som não sonoro *'s'*. Observa-se que $H(z)$ tenta acompanhar o espectro do sinal quando a ordem do filtro aumenta ($M = 48$), principalmente para as baixas frequências. Isto se deve à interferência que o modelo espectral sofre do sinal de excitação. Este problema é verificado onde ocorrem oscilações periódicas na curva do espectro da DFT. Estas oscilações são maiores em baixa frequência e, por isso, verifica-se uma interferência mais intensa nesta região.

Pode-se inferir que, sendo os parâmetros extraídos diretamente da fala, se a ordem dos coeficientes mel-cepstrais é excessiva, o filtro MLSA projetado modelará não somente o filtro associado ao trato vocal, mas incorporará também a excitação do sinal, prejudicando o modelo espectral. Isto não deve ocorrer, pois a ideia do sistema de síntese via HMM é modelar separadamente excitação e filtro.

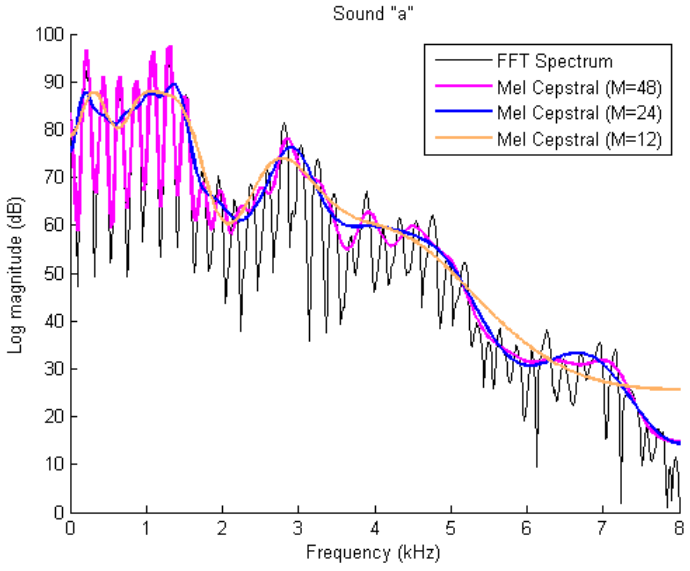

Figura 5.13 – Envelope espectral do som *'a'* e filtros MLSA com coeficientes mel-cepstrais.

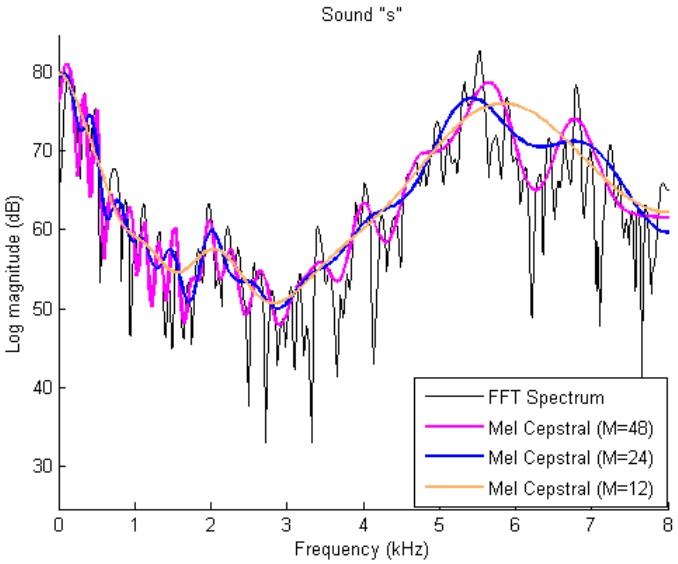

Figura 5.14 – Envelope espectral do som *'s'* e filtros MLSA com coeficientes mel-cepstrais.

Por outro lado, se a ordem da análise mel-cepstral utilizada é pequena, por exemplo, $M = 12$, nota-se que $H(z)$ não é capaz de modelar todas as formantes do sinal de fala e, consequentemente, é impossível gerar um filtro que modele bem o som desejado.

Observa-se que o som sonoro, que possui as principais componentes em baixa frequência, é mais sensível à ordem do filtro, uma vez que as interferências com o sinal de excitação são mais intensas nas baixas frequências. Já os sons não sonoros são mais imunes a estes problemas, uma vez que apresentam natureza ruidosa.

As simulações demonstraram que os filtros construídos com coeficientes mel-cepstrais de ordem entre $M = 20$ e $M = 24$ (usando frequência de amostragem de 16 kHz) são os que conseguiram representar mais fielmente o trato vocal.

Agora, considerando os coeficientes cepstrais extraídos do som sonoro *'a'*, foram projetados filtros MLSA usando ordens $M = 96$, $M = 48$, $M = 24$ e $M = 12$. A Figura 5.15 apresenta o espectro do sinal janelado e dos diferentes filtros calculados. Um filtro de baixa ordem ($M = 12$), usando coeficientes cepstrais, também se mostrou insuficiente para modelar todas as formantes do sinal de fala. Observa-se que para $M = 48$ não ocorre o mesmo grau de interferência com o sinal de excitação, conforme observado no caso dos coeficientes mel-cepstrais. Mas, aumentando ainda mais a ordem do filtro, torna-se possível observá-lo. Para $M = 96$ ele se apresenta nitidamente.

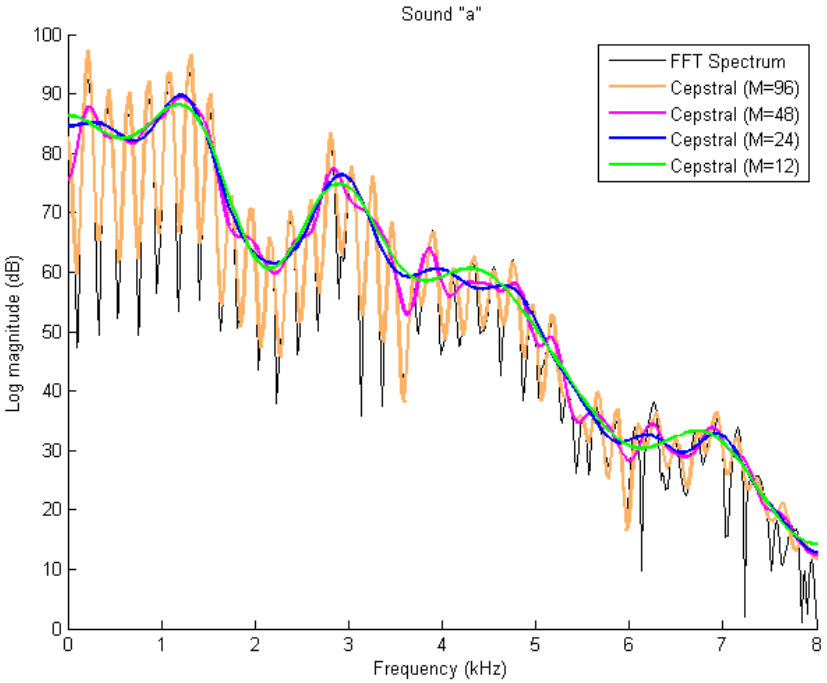

Figura 5.15 – Envelope espectral do som *'a'* e filtros MLSA usando coeficientes cepstrais.

5.3.1.2 Comparações dos Filtros

Análises acústicas subjetivas demonstraram que a fala sintetizada, utilizando o filtro MLSA de ordem 24 projetado com os coeficientes mel-cepstrais, apresenta uma qualidade melhor que a fala sintetizada com o mesmo filtro de ordem 24 utilizando coeficientes cepstrais. Este resultado pode ser também visualizado na Figura 5.16, na qual se observa que o filtro MLSA de ordem 24 projetado com os coeficientes mel-cepstrais é capaz de modelar a envoltória do espectro melhor, principalmente para baixas frequências, do que o filtro construído usando os coeficientes cepstrais. Disso infere-se que, usando a mesma ordem 24, os dois filtros projetados não apresentaram um desempenho equivalente. Com base na Figura 5.15, entende-se que é necessário usar uma ordem superior a 24 para modelar melhor o $H(z)$ do som *'a'* com coeficientes cepstrais. As simulações realizadas utilizando Matlab® revelaram que uma ordem adequada seria por volta de 36, ou seja, projetar o filtro MLSA usando 37 coeficientes cepstrais na expressão (5.12).

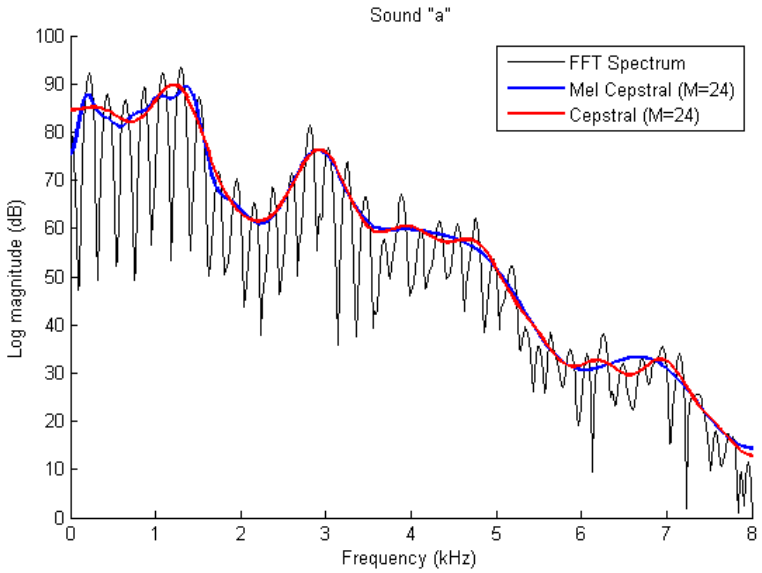

Figura 5.16 – Filtros de ordem 24 com coeficientes mel-cepstral e cepstral do som 'a'

A Figura 5.17 permite comparar a capacidade de seguir o envelope espectral do som *'a'* pelos filtros MLSA projetados, um usando coeficientes mel-cepstrais de ordem 24 e o outro usando os coeficientes cepstrais de ordem 36. Observa-se que ambos os filtros seguem bem a envoltória do espectro do sinal de voz, apresentando comportamento equivalente para o som sonoro *'a'*, cujas principais componentes estão nas baixas frequências.

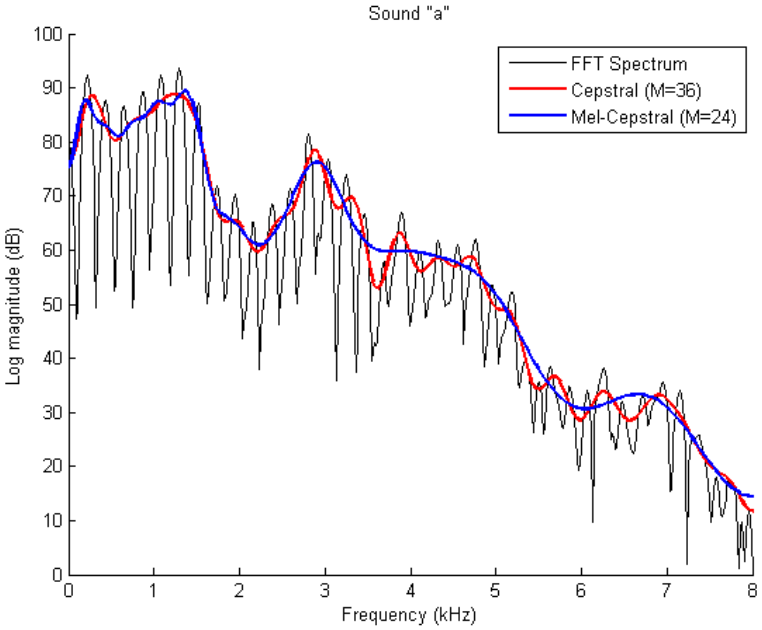

Figura 5.17 – Filtros com coeficientes cepstrais e mel-cepstrais do som *'a'*

A Figura 5.18 apresenta as mesmas curvas considerando o som não sonoro *'s'*. Nota-se que o filtro MLSA usando os coeficientes cepstrais seguiu melhor o envelope espectral do sinal principalmente nas altas frequências. Entretanto, os sons não sonoros possuem natureza ruidosa, e é praticamente imperceptível, em termos de qualidade acústica, a perda causada pelo filtro projetado com coeficientes mel-cepstrais. Os ouvidos humanos são mais sensíveis aos sons sonoros [61], que possuem as principais componentes em baixa frequência. Por isso, no projeto dos filtros, é necessária uma maior preocupação com estas componentes de frequências.

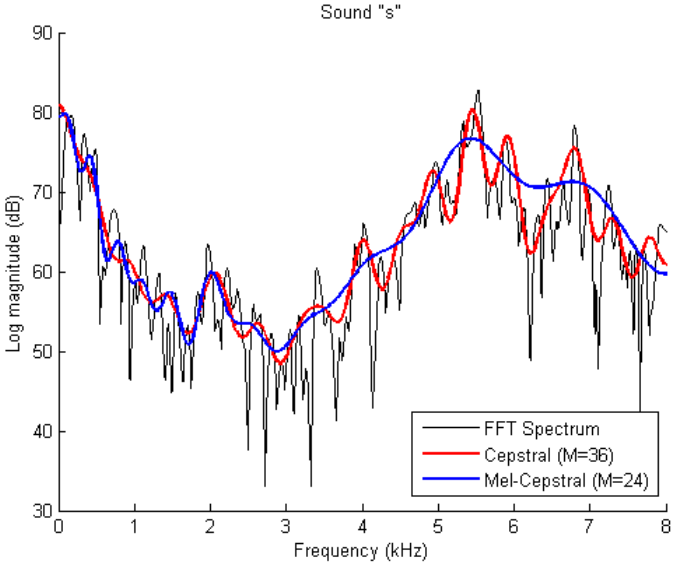

Figura 5.18 – Filtros com coeficientes cepstrais e mel-cepstrais do som *'s'*.

Com o objetivo de conhecer qual tipo de análise permite construir o filtro que melhor modela o trato vocal, foi realizado um teste subjetivo. Três diferentes frases foram sintetizadas usando as mesmas condições (conforme descrito em 5.3.1). Os modelos HMM gerados possuíam a mesma estrutura. A única diferença foi a configuração do filtro MLSA, um construído com ordem 36 usando os coeficientes cepstrais e outro com ordem 24 usando os coeficientes mel-cepstrais. Foi solicitado a dez pessoas não especializadas em síntese de fala, que escutassem as frases e escolhessem aquela mais agradável de cada um dos três pares. A Figura 5.19 mostra o resultado.

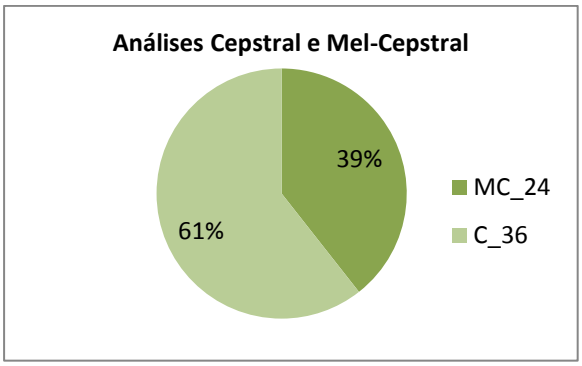

Figura 5.19 – Análise de preferência usando diferentes filtros MLSA

Observa-se que, neste estudo preliminar, as frases sintetizadas com o HTS usando o $H(z)$ construído com os coeficientes cepstrais foram preferidas pelos ouvintes, entretanto foi necessário aumentar em mais de 40% a ordem do filtro MLSA.

5.3.1.3 Conclusões do Experimento

As simulações e estudos realizados permitiram compreender melhor as diferenças entre as análises cepstrais e mel-cepstrais e suas implicações na construção do filtro MLSA. Foi evidenciada a capacidade da análise mel-cepstral de modelar mais precisamente as baixas frequências, que são as principais componentes dos sons sonoros. Considerando uma mesma ordem, a análise mel-cepstral é capaz de representar melhor as formantes nas baixas frequências que a análise cepstral. As simulações mostraram que, com taxa de amostragem de 16 kHz, é necessário utilizar no projeto do filtro uma ordem $M = 36$ quando se utilizam os coeficientes cepstrais e uma ordem $M = 24$ quando se utilizam os coeficientes mel-cepstrais, de modo a se obter um filtro MLSA capaz de modelar bem o trato vocal. O teste subjetivo realizado demonstrou que para conseguir uma qualidade praticamente equivalente do sinal de fala sintetizado utilizando o filtro projetado com coeficiente mel-cepstrais é necessário utilizar um filtro MLSA de ordem cerca de 40% superior quando se usam os coeficientes cepstrais.

5.4 Modelo de Excitação

A excitação no sinal de fala pode ser de dois tipos: ruidosa, que modela os sons não sonoros, e um trem de impulsos, que consegue representar os sons sonoros, conforme discutido no Capítulo 3. Devido a esta dupla natureza do som, para construir o modelo de excitação é necessário, primeiramente, caracterizar o tipo do som em sonoro/não sonoro e, para isto, usa-se um HMM discreto. Nos casos em que o som é classificado como sonoro, é necessário modelar a sua frequência de pitch, e utiliza-se um HMM contínuo. Os sons não sonoros não possuem uma frequência de pitch associada e, por isso, não precisam ser modelados. Para adequar esta exigência híbrida do modelo de excitação, utiliza-se uma estrutura HMM baseada na distribuição de probabilidade multiespacial (MSD) [62,63], a qual permite misturar grandezas discretas e contínuas.

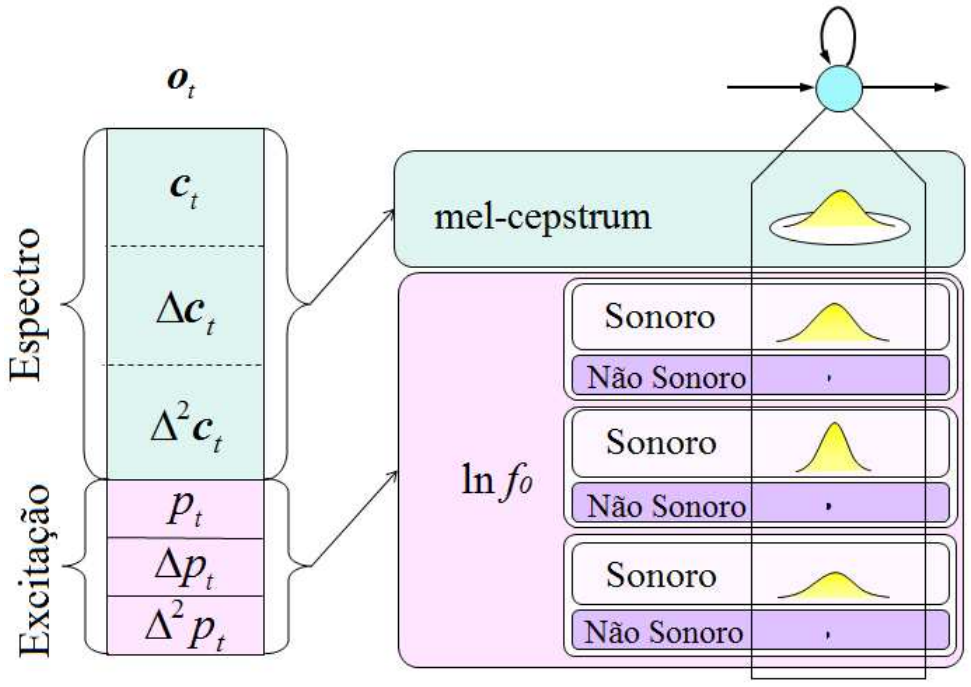

Figura 5.20 – Estrutura dos modelos de cada estado do HMM
Fonte: adaptado de [39]

O modelo de excitação dos sons sonoros é composto pelo logaritmo da frequência fundamental (ln f_0) e pelos respectivos coeficientes dinâmicos, delta e delta-delta. Para os sons não sonoros usa-se apenas um valor discreto, de forma que o modelo MSD contém uma mistura de duas componentes, uma Gaussiana para os parâmetros contínuos e uma distribuição discreta.

A Figura 5.20 ilustra a estrutura utilizada para modelar um estado do HMM contendo as informações espectrais e de excitação.

5.5 Modelo de Duração

Este modelo fornece a informação de duração de cada fonema dependente de contexto, determinando o tempo de permanência em cada estado do HMM. Para treinar os modelos, é necessário que os arquivos da base de fala, que apresentam a transcrição fonética e etiquetas das sentenças, contenham também a informação da duração de cada fonema. Esses dados são utilizados para determinar a distribuição Gaussiana multivariada [58] que modela a duração dos estados de cada HMM.

Usando a topologia HMM do tipo *left-to-right* sem saltos, típico para a síntese de fala, tem-se que a dimensão da densidade da duração para o HMM é igual ao número de estados do HMM de cada fone dependente de contexto.

Para controlar a estrutura temporal dos HMMs as densidades de duração dos estados devem ser explicitadas. A função densidade de probabilidade escolhida para modelá-las foi uma Gaussiana.

Adotando este modelo, pode-se verificar que a probabilidade relacionada à sequência de estados $\mathbf{q} = \{q_1, q_2, \ldots, q_T\}$ pode ser dada por:

$$P(\boldsymbol{q}|\lambda, T) = \prod_{k=1}^{K} p_k(d_k) \tag{5.13}$$

onde $p_k(d_k)$ é a probabilidade de se ter exatamente d_k quadros no estado k e K é o número total de estados visitados durante T quadros, sendo T o número total de quadros da sentença, ou seja:

$$\sum_{k=1}^{K} d_k = T \tag{5.14}$$

Maximizando (5.13) com a restrição de (5.14) obtém-se:

$$d_k = m_k + \rho \,.\, {\sigma_k}^2 \tag{5.15}$$

com $1 \leq k \leq K$, onde m_k e ${\sigma_k}^2$ são respectivamente, a média e a variância da densidade de duração do estado *k* e, ρ é dado por:

$$\rho = \frac{\left(T - \sum_{k=1}^{K} m_k\right)}{\sum_{k=1}^{K} {\sigma_k}^2} \tag{5.16}$$

As equações (5.15) e (5.16) mostram como é possível controlar a duração da fala sintetizada alterando ρ. Assim, se $\rho = 0$ a fala sintetizada usa os valores médios e não ocorre variação da duração. Se $\rho > 0$ a fala se torna mais lenta e se $\rho < 0$ a fala se torna mais rápida. Observe que a variabilidade da duração dos estados depende da variância do estado *k*, de modo que as alterações das durações de cada estado não ocorrem de maneira uniforme.

O fator de duração ρ não pode ser indicado diretamente entre os parâmetros de entrada do HTS, pois não é uma variável necessária para o treinamento dos modelos HMM. Entretanto, o hts_engine-API, que realiza o processo de síntese da fala, permite a configuração da taxa de velocidade da fala através da variável de entrada *r*. Esta variável pode receber valores entre 0 e 10 e posteriormente é mapeada no valor equivalente de ρ da seguinte maneira:

$$\begin{cases} \text{se } r = 1\,, & \text{então } \rho = 0; \\ \text{se } r \neq 1, & \text{então } \rho = \dfrac{\frac{\sum_{k=1}^{K} m_k}{r} - \sum_{k=1}^{K} m_k}{\sum_{k=1}^{K} {\sigma_k}^2} \end{cases} \tag{5.17}$$

Desta forma, verifica-se que, quando *r* tende a zero, ρ tende a mais infinito e, à medida que *r* se aproxima de 10, ρ assume valores negativos cada vez menores e a fala vai ficando mais rápida. Mas, é necessário lembrar que cada estado do HMM deve ter uma duração mínima de 1 quadro (5 ms), de modo que o número de fonemas da frase determina a duração mínima da mesma. Exemplificando, a palavra "*olá*", necessita de 3 HMMs, um para cada fone. Se for usado HMM de 5 estados por fone, serão necessários um total de 15 estados (sem considerar o silêncio inicial e final). Assim esta palavra terá uma duração mínima de 15 estados (75 ms) para ser pronunciada. Entretanto, é óbvio que neste caso extremo a qualidade da síntese se degrada, pois nem todos os fonemas podem ser bem modelados com apenas um quadro.

6. Melhorando a Qualidade da Voz Sintética

6.1 Straight

Teoricamente, o envelope espectral pode ser representado mais perfeitamente com o aumento da ordem dos parâmetros do filtro. Entretanto, como é visto na Seção 5.3.1.1, o aumento da ordem do filtro acaba por comprometer o modelo espectral, uma vez que incorpora a ele as informações indesejadas do modelo de excitação. Devido a esta interferência, a ordem máxima dos parâmetros para projetar o filtro acaba sendo limitada e isto compromete o desempenho do modelo do trato vocal.

Para estimar de forma mais precisa o envelope espectral, Kawahara [64] propôs em 2001 o sistema Straight. A Figura 6.1 apresenta este vocoder em uma estrutura de blocos.

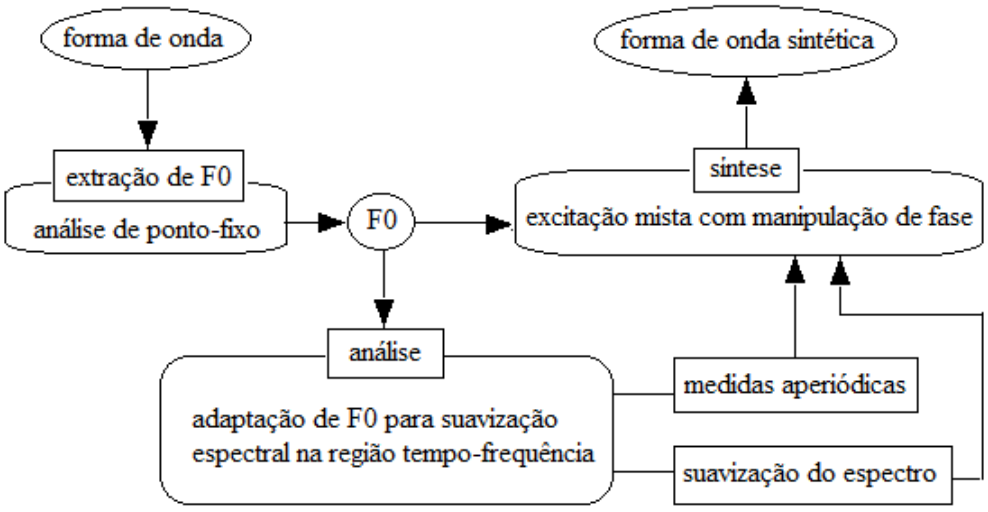

Figura 6.1 – Diagrama de blocos do vocoder Straight
Fonte: adaptado de [65]

O Straight é um sistema de síntese, modificação e análise de sinais de fala implementado com um vocoder. Ele se baseia em três principais componentes: extração de f_0, análise das medidas espectrais e de aperiodicidade e síntese de fala. Primeiramente, são extraídas as frequências de pitch f_0 usando a análise de ponto fixo [66]. Com estes valores, o Straight realiza

a análise espectral com uma adaptação das frequências f_0 (f_0-*adaptive spectral analysis*), combinando com um método de reconstrução da superfície na região tempo-frequência, de modo a remover a periodicidade do sinal. As medidas de aperiodicidade no domínio da frequência também são extraídas [64]. Como consequência, o Straight gera um envelope espectral com efeito reduzido de f_0 e uma medida de aperiodicidade.

Na fase de síntese utiliza-se a excitação mista, que é realizada fazendo a ponderação do trem de impulsos com manipulação de fase e do ruído Gaussiano. As ponderações de cada componente (sonora/ não sonora) são realizadas no domínio da frequência usando as medidas de aperiodicidade extraídas. O Straight sintetiza a forma de onda com o espectro amortecido e a excitação mista usando o processo baseado em FFT [65,66].

Utilizando os demos disponíveis em [2] para a língua inglesa, *Speaker dependent training demo English/STRAIGHT demo* e *Speaker dependent training demo English/Normal demo,* foram treinados os modelos HMM utilizando a técnica do Straight e a técnica sem Straight com excitação simples, respectivamente, considerando para ambos os casos a mesma base de dados e as configurações de *default*. Testes acústicos subjetivos, realizados com ouvintes não especializados, revelaram que não houve melhora da qualidade da síntese que compensasse o aumento computacional introduzido pela metodologia de síntese; por isso, não foram realizados estudos complementares sobre esta técnica no decorrer deste trabalho.

6.2 Pré-Ênfase e De-Ênfase

Normalmente, os TTS que utilizam vocoder possuem uma etapa de pré-processamento na qual é aplicado um filtro de pré-ênfase nos sinais de fala do corpus. Este filtro serve para minimizar os efeitos da radiação dos lábios e da variação da área da glote nas gravações.

Com a ideia de aperfeiçoar o processamento de síntese via HMM e facilitar a extração dos parâmetros espectrais, foi aplicado este procedimento de pré-ênfase nas gravações da base de fala. Realizou-se uma filtragem do tipo:

$$H_{pré-ênfase}(z) = 1 - 0.95z^{-1} \tag{6.1}$$

que apresenta o comportamento esquematizado qualitativamente na Figura 6.2.

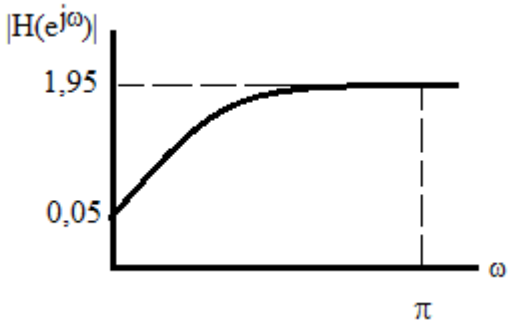

Figura 6.2 – Filtro de pré-ênfase

Após isto, seguiu-se com o processamento de treinamento habitual dos modelos HMM e realizou-se a síntese de algumas sentenças. Estes sinais sintetizados foram submetidos a um filtro de de-ênfase com o intuito de cancelar os efeitos da pré-ênfase na fala sintética. O filtro de de-ênfase correspondia à função de transferência inversa da empregada no filtro de pré-ênfase:

$$H_{de-ênfase}(z) = \frac{1}{1-0.95z^{-1}} \tag{6.2}$$

que apresenta o comportamento esquematizado qualitativamente na Figura 6.3.

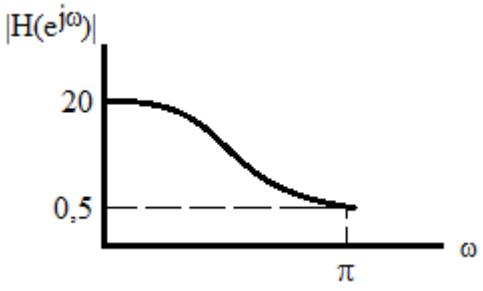

Figura 6.3 – Filtro de de-ênfase

Os resultados com a análise-síntese não foram satisfatórios, pois o uso desse mecanismo degradou o sinal de fala artificial. Isto se deve ao fato que o filtro de de-ênfase não é capaz de cancelar todos os efeitos produzidos pelo filtro de pré-ênfase, pois os modelos espectrais usam critérios estatísticos para estabelecerem os parâmetros a serem usados. Dessa forma o filtro

escolhido para a síntese não é necessariamente o mesmo que sofreu a pré-ênfase na análise. Por isso, a técnica de pré-ênfase/de-ênfase não é plausível de ser empregada nos sistemas de síntese baseados em HMMs.

6.3 Pós-Filtro

A técnica do pós-filtro é utilizada em muitos vocoders para melhorar a qualidade do sinal sintetizado [30]. Ela consiste em submeter o sinal sintetizado a uma nova filtragem. O pós-filtro é calculado de maneira análoga ao filtro MLSA, acrescentando um coeficiente β em seu cálculo. Esta variável expressa a intensidade da pós-filtragem. Retomando as expressões do filtro MLSA descritas no Capítulo 3, (3.24) e (3.25), tem-se:

$$D(z) = \exp\left(\sum_{m=1}^{M} b(m)\Phi_m(z)\right) \tag{6.3}$$

$$b(m) = \begin{cases} c_1(m) & m = M \\ c_1(m) - \alpha b(m+1) & m = 1,2,\dots,M-1 \end{cases} \tag{6.4}$$

Multiplicando os coeficientes $c(m)$ dessas expressões por β, obtêm-se as expressões do pós-filtro:

$$D_\beta(z) = \exp\left(\beta.\sum_{m=1}^{M} \bar{b}(m)\Phi_m(z)\right) \tag{6.5}$$

com

$$\bar{b}(m) = \begin{cases} b(m) & 2 \le m \le M \\ -\alpha b(2) & m = 1 \end{cases} \tag{6.6}$$

onde $b(m)$ é dado por (6.4). Note que o coeficiente $c(1)$ assume valor zero, para evitar que todo o espectro seja enfatizado. Observe que, quando β = 0 a pós filtragem não é realizada e, quando β > 0, as formantes do espectro são enfatizadas. A Figura 6.4 ilustra o efeito da pós-filtragem, a curva pontilhada corresponde ao sinal antes da pós-filtragem e a curva sólida representa o sinal

após a pós-filtragem utilizando β = 0,5. É possível perceber que os picos e vales do sinal, ou seja, as formantes e anti-formantes são enfatizadas quando se emprega o pós-filtro.

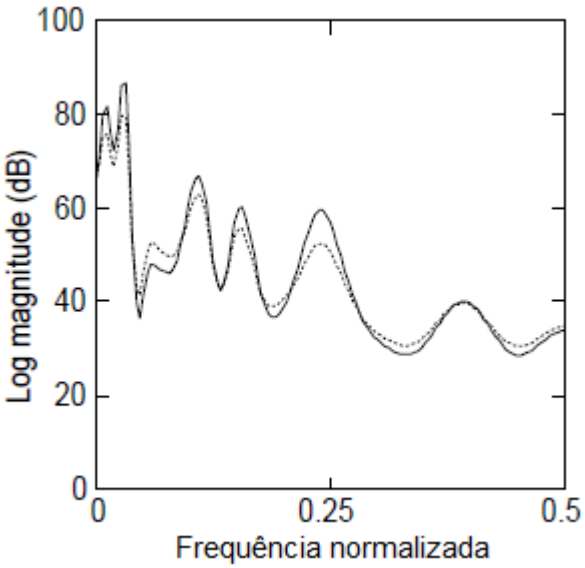

Figura 6.4 – Efeito da pós-filtragem
Fonte: adaptado de [30]

Para ativar esta técnica no HTS, deve-se configurar a variável que controla a pós-filtragem *PSTFILTER* com o valor (β+1) desejado. Por default ela assume o valor de 1,4; o que corresponde a um aumento de 40% no valor dos coeficientes do filtro MLSA.

A literatura afirma que esta técnica melhora um pouco a qualidade da fala sintética, entretanto, das análises acústicas subjetivas realizadas neste trabalho constatou-se que o uso da variância global, que será apresentado na próxima seção, mostrou-se mais eficaz.

6.4 Variância Global

Os parâmetros espectrais e de excitação necessários para sintetizar fala são gerados através de um algoritmo que leva em conta os coeficientes estáticos e dinâmicos, conforme apresentado na Seção 5.2.3.2. Entretanto, os parâmetros gerados às vezes resultam muito suavizados, porque o processo estatístico para modelar os HMM tende a eliminar os detalhes das estruturas espectrais. Embora esta suavização seja necessária para reduzir os erros na etapa de

geração do espectro, ela acaba degradando a naturalidade da fala sintetizada que se apresenta com uma característica abafada para o ouvinte, uma vez que os detalhes eliminados são necessários para produzir síntese com alta qualidade.

Uma técnica proposta por Toda e Tokuda [47,48] para reduzir a suavização excessiva dos parâmetros é denominada variância global e denotada por GV do termo em inglês *Global Variance*. A GV é a variância dos coeficientes estáticos calculada em uma sentença e a sua PDF é modelada por uma distribuição Gaussiana com uma matriz de covariância diagonal.

A Figura 6.4 apresenta um esquema dos processos de treinamento e síntese via HMM usando GV. Durante a síntese de fala, os parâmetros são calculados levando em conta os parâmetros estáticos e dinâmicos, conforme descrito na Seção 5.2.3. Após isso, a trajetória gerada é convertida, de modo que sua GV se iguale à média da distribuição Gaussiana. Usando a trajetória convertida como valor inicial, os parâmetros que maximizam uma função objetiva, que é definida pela soma ponderada do logaritmo da probabilidade de saída da sequência dos parâmetros de fala e das suas GV, são iteravamente otimizados com o método de Newton-Raphson [48].

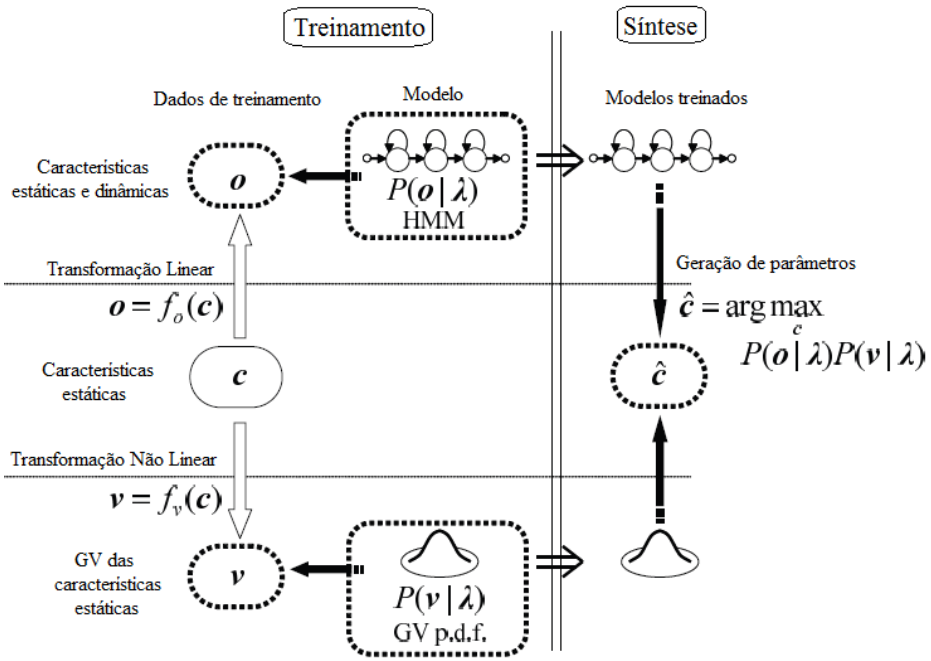

Figura 6.4 – Esquema de síntese usando GV
Fonte: [48]

Neste trabalho foi realizado um experimento para compreender a diferença na qualidade final do sinal de fala sintetizado utilizando a técnica da GV. Ouvintes não especializados em processamento de sinais de fala foram convidados a ouvir frases sintetizadas com e sem o uso da GV. Os resultados indicaram a preferência deles para as frases sintetizadas com a GV, pois a fala sintética utilizando esta ferramenta é sensivelmente mais agradável por apresentar uma melhoria na qualidade prosódica do discurso, parecendo mais natural e menos monótona. Apesar desta leve diferença percebida acusticamente, nas Figuras 6.5 e 6.6 pode-se notar a semelhança espectral apresentada pelos sinais de fala sintetizados com e sem a utilização da técnica da GV, referentes à parte sublinhada da frase *"Apenas os ônibus circularão pela pista bairro-centro nos dois sentidos"*.

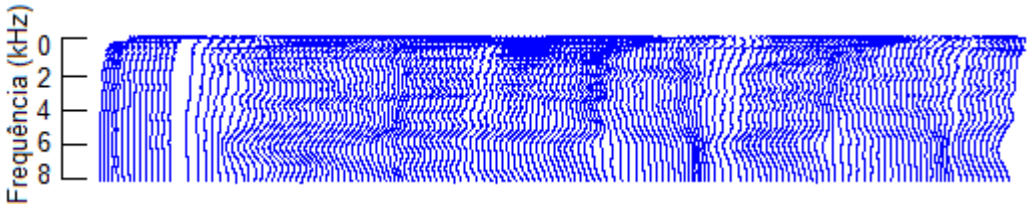

Figura 6.5 – Espectrograma – sem GV

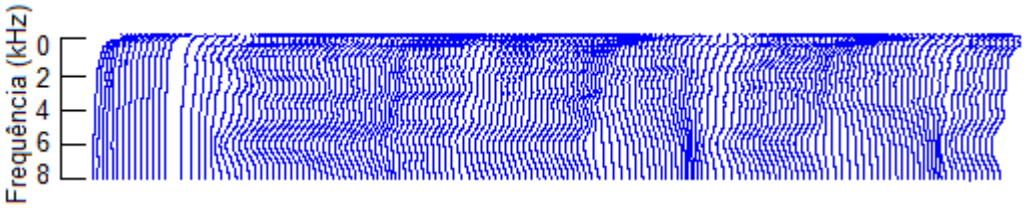

Figura 6.6 – Espectrograma – com GV

7. Conclusões

Ao longo deste estudo verificou-se a abrangência e interdisciplinaridade do sistema de síntese de fala baseado em HMM. O enfoque do trabalho poderia ser dado em uma das diversas subpartes do sistema TTS, tais como: o desenvolvimento de mecanismos para a melhoria da qualidade da síntese de fala, técnicas de adaptação da voz artificial, inserção de emoções no discurso e melhoria da prosódia, aprimoramento dos modelos HMM com propostas de novos filtros, sistemas de excitação ou controle de duração, entre outras. Cada uma dessas partes contempla outros inúmeros desmembramentos que permitem um estudo especializado e pontual. Contudo, neste trabalho de Mestrado, optou-se por realizar um panorama geral de todo o sistema de conversão texto-fala, de modo a compreender e descrever as suas principais características.

A síntese de fala via HMM pode ser dividida em duas etapas. A primeira realiza o treinamento dos modelos HMM dependente de contexto. Nesta fase o sistema aprende como deve representar os fonemas da língua nos diversos contextos. A segunda etapa é a de síntese propriamente dita, onde o conversor texto-fala é capaz de gerar a forma de onda da fala sintética a partir dos modelos HMM. É interessante ressaltar que, através das técnicas de árvore de decisão e agrupamento de contexto, o sistema é capaz de extrapolar seus conhecimentos conseguindo converter em sinal de fala qualquer texto de entrada, levando em conta obviamente o idioma para o qual os modelos HMMs foram treinados.

Cada HMM possui as informações de cada fone dependente de contexto necessárias para executar a síntese de fala, ou seja, informações espectrais, de excitação e de duração dos fonemas. Diversas simulações foram realizadas com o intuito de compreender como cada parâmetro interfere na qualidade da fala sintetizada.

No que tange ao modelo espectral, foram analisados diferentes modos de projetar o filtro MLSA utilizado para simular o trato vocal. Concluiu-se que as análises mel-cepstral e cepstral são as melhores opções para construir o filtro, pois garantem a estabilidade do sistema mesmo quando os parâmetros do filtro são calculados através de estimações estatísticas. Esta estabilidade, que é essencial para se obter sinal de fala de boa qualidade, não pode ser assegurada quando se usa a técnica LPC ou a análise mel-cepstral generalizada. Nestes casos é necessário trabalhar com coeficientes transformados LSP. Além do tipo de coeficiente utilizado no projeto,

foi verificado que a ordem do filtro MLSA deve ser adequada, de forma que seja possível modelar todos os formantes do sinal, sem gerar interferência com o modelo de excitação. Foi encontrado que a ordem a ser utilizada, considerando a frequência de amostragem de 16 kHz, é de 24 para a análise mel-cepstral e de 36 para a análise cepstral.

No estudo do modelo de excitação, a principal dificuldade obtida foi na obtenção de material de treinamento etiquetado em língua portuguesa falada no Brasil. A base de fala disponibilizada online [2] apresenta alguns erros na transcrição fonética, o que comprometeu a qualidade prosódica da síntese. Utilizando esta base de fala, foram comparadas as diferentes técnicas propostas para o modelo de excitação: a excitação simples (trem de impulsos/ ruído), a excitação mista, e a técnica do Straight. Neste trabalho optou-se por utilizar a excitação simples, que apresentou melhor desempenho considerando o compromisso entre custo computacional e melhora na qualidade da síntese. Constatou-se que a determinação dos limites de frequência configurados durante a extração do pitch, usando o método ESPS [19] para caracterizar os sons sonoros e não sonoros, é de fundamental importância para a qualidade da síntese final. Aumentando os limites, sons não sonoros acabam sendo classificados como sonoros e este erro, mesmo tratando-se de poucos milissegundos, causa uma considerável degradação na forma de onda gerada.

Notou-se que a utilização dos parâmetros estáticos para modelar o espectro e a excitação não são suficientes, uma vez que causam transições abruptas entre os diferentes estados do HMM, gerando um sinal de fala de baixa qualidade. A proposta de interpolação linear dos parâmetros gerou uma síntese de melhor qualidade, com baixo aumento de complexidade computacional. Entretanto, constatou-se que a utilização dos parâmetros dinâmicos (delta e delta-delta) para modelar o espectro e a frequência de excitação contribui grandemente para a naturalidade e inteligibilidade da fala artificial, apesar do considerável aumento dos cálculos na fase de síntese, sendo necessário realizar cerca de 5000 operações a mais para uma frase de 1s com filtro MLSA de ordem 24.

Outro fator relevante para a qualidade final da fala sintetizada é o número de estados com os quais se modelam os HMMs. Verificou-se que, aumentando o número dos mesmos, ocorre uma melhora na pronúncia dos fonemas, à medida que os parâmetros que caracterizam cada estado resultam mais precisos. Entretanto, para garantir esta precisão, é necessário que o material de treinamento seja suficientemente abrangente e contemple várias amostras de cada fone. Das

simulações realizadas, utilizando sempre a mesma base de fala para a língua portuguesa com 613 frases, constatou-se que modelos de 5 e 7 estados são suficientes, não sendo mais perceptível a melhora da qualidade da síntese quando se usam modelos com mais estados.

Verificou-se que, acrescentando a técnica da variância global (GV) no processo de síntese do sinal de fala consegue-se melhorar a variabilidade das trajetórias geradas, pois leva-se em consideração a variância global dos coeficientes estáticos da sentença além das restrições dadas pelos parâmetros estáticos e dinâmicos no momento da determinação dos parâmetros de síntese. Isto implica em uma melhora audível da fala sintetizada, com um aprimoramento na prosódia e uma diminuição da característica robótica da voz sintética.

Dada a complexidade do sistema como um todo, a variedade de domínios de conhecimento envolvidos e a qualidade da síntese de fala obtida até o momento, compreende-se a dificuldade encontrada para aprimorar os modelos existentes.

Tendo presente sempre a base de fala em português falado no Brasil disponível em [2], com taxa de amostragem de 16 kHz, a configuração que proporcionou síntese de melhor qualidade utilizou os seguintes parâmetros:

- Janelamento do sinal para a extração dos parâmetros utilizando janelas de Blackman de 25 ms com deslocamento de 5 ms;
- Projeto do filtro MLSA usando análise cepstral com $\alpha = 0$; $\gamma = 0$ e $M = 36$ ou análise mel-cepstral com $\alpha = 0{,}42$; $\gamma = 0$ e $M = 24$;
- Limites de frequência na extração de pitch dos sons sonoros de 130 Hz a 320 Hz, tendo presente que a base de fala foi gravada com voz feminina;
- Número de estados do HMM configurado com 5 ou 7;
- Utilização dos coeficientes dinâmicos nos modelos do espectro e da excitação;
- Uso da GV;
- Os modelos espectrais e de duração foram modelados com uma única mistura de Gaussiana;
- O modelo de excitação utiliza o MSD, modelando a frequência de pitch dos sons sonoros com uma distribuição Gaussiana e caracterizando com uma distribuição discreta os sons não sonoros.

Referências

[1] Chapanis, A., "*Interactive human communication*", Scientific American, 232, p.36-42, 1975.

[2] *HMM-Based Speech Synthesis System (HTS)*. Acessado em dezembro/2012. Disponível em: http://hts.sp.nitech.ac.jp/

[3] Dudley, H., Riesz, R.R., e Watkins, S.A., "*A Synthetic Speaker*", J. Franklin Inst. 227, 739-764, 1939.

[4] Cooper, F.S., Liberman, A.M., e Borst, J.M., "*The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception os Speech*", Proc. Natl. Acad. Sci. (US) 37, 318-325, 1951.

[5] Tokuda, K., Kobayashi, T. e Imai, S., "*Speech parameter generation from HMM using dynamic features*", in Proceedings of ICASSP, pp.660–663, 1995.

[6] Campos, G.L. "Síntese de Voz para o Idioma Português", Tese de Doutorado - Escola Politécnica, Universidade de São Paulo, São Paulo, 1980.

[7] Esquivel, A.S. "Um Sistema de Síntese de Voz". In: Congresso Nacional de Informática, 18. São Paulo, 1985. Anais. São Paulo, Sucesu, pp. 776-82, 1985.

[8] Egashira, F.; Violaro, F. "Síntese de Voz a Partir de Texto". Campinas, Faculdade de Engenharia Elétrica da Universidade Estadual de Campinas, (Publicação FEE 01/93), 1993.

[9] Egashira, F. e Violaro, F., "Conversor Texto-Fala para a Língua Portuguesa", 13º Simpósio Brasileiro de Telecomunicações, Águas de Lindóia, SP. pp.71-76, Setembro/1995.

[10] Barbosa, P.A, Violaro, F., Albano E.C., Simões, F., Aquino, P., Madureira, S. e Françoso, E., - "*Aiuruetê: A High-Quality Concatenative Text-to-Speech System for Brazilian Portuguese with Demisyllabic Analysis-Based Units and a Hierarchical Model of Rythim Production"*, Eurospeech´99, 6th European Conference on Speech Communication and Technology, Budapest, Hungria. Vol. 5, pp. 2059-2062, Setembro/1999.

[11] Simões, F.O, Violaro, F., Barbosa, P.A., Albano, E.C., "Um Sistema de Conversão Texto-Fala para o Português Falado no Brasil" - Revista da Sociedade Brasileira de Telecomunicações, ISSN 0102-986X, Vol. 15, nº2, pp. 70-77, Dezembro/2000.

[12] Maia, R., Zen, H., Tokuda, K., Kitamura, T., Resende Jr., F., "*An HMM-Based Brazilian Portuguese Speech Synthesizer and its characteristics"*- Journal of Communication and Information Systems, v. 21, p. 58-71, 2006.

[13] Maia, R., Zen, H., Tokuda, K., Kitamura, T., Resende Jr., F., "*Towards the development of a Brazilian Portuguese text-to-speech system based on HMM"*- In: Proc. Eurospeech, pp. 2465–2468, 2003.

[14] Silva, D.C., "Algoritmos de Processamento da Linguagem e Síntese de Voz com Emoções Aplicados a um Conversor Texto-Fala Baseado em HMM", Tese de Doutorado. COPPE Instituto Alberto Luiz Coimbra de Pós Graduação e Pesquisa de Engenharia UFRJ, Rio de Janeiro, 2011.

[15] *Festival Source Distribution – version 2.1 – disponibilized November 2010*. Acessado em dezembro de 2012. Disponível em: http://www.cstr.ed.ac.uk/projects/festival/download.html

[16] Gomes, L.C.T., "Sistema de conversão texto-fala para a língua portuguesa utilizando uma abordagem de síntese por regras", Dissertação de Mestrado, Faculdade de Engenharia Elétrica da Universidade Estadual de Campinas, Campinas, 1998.

[17] Klatt, D., "*Software for a cascade/parallel formant synthesizer,*" Journal of the Acoustical Society of America, vol. 67, pp. 13-33, 1980.

[18] Klatt, D. H. e Klatt, L. C., "*Analysis, synthesis, and perception of voice quality variations among female and make talkers*", Journal of the Acoustical Society of America, vol. 87, no. 2, pp. 820-857, 1990.

[19] Engwall, O., "*Speech production: Models, Phonetic Processes and Techniques"*, Chap. Assessing MRI measurements: Effects of sustentation, gravitation and coarticulation., pp. 301 – 314. New York: Psychology Press, 2006.

[20] Mullen, J., Howard, D., e Murphy, D., "*Waveguide Physical Modeling of Vocal Tract Acoustics: Flexible Formant Bandwith Control from Increased Model Dimensionality*". IEEE Transactions on Audio, Speech and Language Processing, 14(3), 964 – 971, 2006.

[21] Toma, S.A., Târşa, G.I., Oancea, E., Munteanu, D.P., Totir, F., Anton, L., "*A TD-PSOLA based method for speech synthesis and compression",* Communications (COMM), 2010 8th International Conference on , vol., no., pp.123-126, 10-12 Junho 2010 doi: 10.1109/ICCOMM.2010.5509044.

[22] Violaro, F. e Boeffard, O., "*A Hybrid Model for Text-to-Speech Synthesis*," IEEE Transactions on Speech and Audio Processing, vol. 6, no. 5, pp. 426–434, September 1998.

[23] Deller, J.R., Proakis, J.G. e Hansen, J.H.L., "*Discrete-Time Processing of Speech Signals*", Macmillan, 1993. ISBN 0-02-328301-7

[24] Dutoit, T., *"An Introduction to Text-to-Speech Synthesis"*, Kluwer Academic Publishers, Londres, 1997. ISBN 0-7923-4498-7.

[25] Rabiner, L.R. e Shafer, R.W., "*Digital Processing of Speech Signals"*, Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1978. ISBN 0-13-213603-1.

[26] Imai, S., Sumita, K. e Furuichi, C., *"Mel Log Spectrum Approximation (MLSA) Filter for Speech Synthesis"*, Electronics and Communications in Japan, Vol. 66-A, n. 2, 1983. ISSN 0424-8368/83/0002-0010.

[27] Kerr, W., *"Creating M-sequences",* Acessado em dezembro/2012. Disponível em: https://cfn.upenn.edu/aguirre/wiki/public:m_sequences

[28] *hts_engine API*. Acessado em dezembro/2012. Disponível em: http://hts-engine.sourceforge.net/

[29] *HTK*. Acessado em dezembro/2012. Disponível em: http://htk.eng.cam.ac.uk/

[30] Yoshimura, T., Kitamura, T., Tokuda, K., *"Simultaneous Modeling of Phonetic and Prosodic Parameters, and Characteristic Conversion for HMM-Based Text-to-Speech Systems"*, Tese de Doutorado, Department of Electrical and Computer Engineering- Nagoya Institute of Technology, Japão, Janeiro/2012.

[31] Fukada, T., Tokuda, K., Kobayashi, T. e Imai, S., "*An adaptive algorithm for mel-cepstral analysis of speech*", in Proc.ICASSP, pp.137–140, 1992.

[32] Masuko, T. *"HMM-Based Speech Synthesis and Its Applications"*, Tese de PhD, Instituto de Tecnologia de Tóquio, Novembro/2002.

[33] Imai, S. e Furuichi, C., "*Unbiased estimator of log spectrum and its application to speech signal processing*", Proc. of EURASIP, pp.203-206, Sep. 1988.

[34] Kobayashi, T., Imai, S., e Fukuda, Y., "*Mel generalized-log spectrum approximation (MGLSA) filter", Journal of IEICE (Japanese Edition)*, vol. J68-A, no. 6, pp. 610–611, 1985.

[35] *Speech Signal Processing Toolkit (SPTK) – Version 3.5 – Dezembro 2011*. Acessado em Dezembro de 2012. Disponível em: http://sp-tk.sourceforge.net/

[36] Tokuda, K., Kobayashi, T. e Imai, S., "*Generalized cepstral analysis of speech: unified approach to LPC and cepstral method*", Proc. ICSLP-90, pp.37–40, Nov. 1990

[37] Rabiner, L., Juang, B.H., *"Fundamentals of Speech Recognition",*Prentice Hall, Inc. Englewood Cliffs, New Jersey, 1993. ISBN 0-13-015157-2.

[38] Rabiner, L., Juang, B.H., *"An Introduction to Hidden Markov Models"*, IEEE ASSP Magazine, Janeiro/1986. 0740-7467/86/0100-0004.

[39] Tokuda, K., Zen, H., *"Fundamentals and recent advances in HMM-based speech synthesis"*, Tutorial at Interspeech 2009.

[40] Bande, I.F., Cávez, A.B., "*Síntesis de voz mediante Modelos Ocultos de Markov*", Projeto de Final de Curso, Engenharia de Telecomunicações, Universidade Politécnica de Catalunya-Espanha, 2008.

[41] Tamura, M., Masuko, T., Tokuda, K. e Kobayashi, T., *"Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR",* in *Proc. ICASSP*, pp. 805–808, 2001.

[42] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "*Eigenvoices for HMM-based speech synthesis"*, Proc. EUROSPEECH, 2002.

[43] Tachibana, M., Yamagishi, J., Masuko, T. e Kobayashi, T., *"Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing", IEICE Trans. Inf.& Syst.*, vol. E88-D, no. 11, pp. 2484–2491, 2005.

[44] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. e Kitamura, T.,*"Speaker interpolation in HMM-based speech synthesis system"*, in *Proc. Eurospeech*, pp. 2523–2526, 1997.

[45] Nose, T., Yamagishi, J. e Kobayashi, T., *"A style control technique for speech synthesis using multiple regression HSMM",* in *Proc. Interspeech*, pp. 1324–1327, 2006.

[46] Banno, H., Hata, H., Morise, M., Takahashi, T., Irino, T., Kawahara, H., *"Implementation of Real time Straight Speech Manipulation System: Report of its first implementation",* The Acoustical Society of Japan, 28, 3, 2007. doi:10.1250/ast.28.140.

[47] Toda, T. e Tokuda,K., *"Speech parameter generation algorithm considering global variance for HMM-based speech synthesis"*, Proc. of Eurospeech, 2005.

[48] Toda, T., *"Modeling of Speech Parameter Sequence Considering Global Variance for MM-Based Speech Synthesis"*, in Hidden Markov Models, Theory and Applications, P. Dymarski, Ed. InTech, 2011.

[49] *The Festival Speech Synthesis System*. Acessado em dezembro/2012. Disponível em: http://www.cstr.ed.ac.uk/projects/festival/

[50] *Mary: Text to Speech.* Acessado em dezembro/2012. Disponível em: http://mary.dfki.de/

[51] Painter, T., Spanias, A., *"Perceptual Coding of Digital Audio"*, in Proceedings of the IEEE, Vol. 88, n. 4, Abril/2000.

[52] *ActiveTcl*. Acessado em dezembro/2012. Disponível em: http://www.activestate.com/activetcl

[53] Samad, S.A, Hussain, A. e Fah, L.K., *"Pitch Detection of Speech Signals using the Cross-Correlation Technique"*, Proceedings of TENCON, 1:283–286. IEEE, 0-7803-6355-8, 2000.

[54] Talkin, D., "*A robust algorithm for pitch tracking (RAPT)*", In Speech Coding and Synthesis, Elsevier:495-518, 1995.

[55] Sönmez, M., Heck, L., Weintraub, M., Shriberg, E., *"A Lognormal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition",* In: Proc. Fifth European Conf. on Speech Communication and Technology (Eurospeech 1997), Rhodos, Greece, pp. 1391–1394, Setembro/1997.

[56] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P., *"The HTK Book (for HTK Version 3.4)*", Cambridge, United Kingdom: Entropic Ltd, 1999.

[57] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. e Kitamura, T., "*Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis*", Proc. EUROSPEECH, vol.5, pp.2347–2350, 1999.

[58] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. e Kitamura, T., "*Duration modeling for HMM-based speech synthesis*", in Proc. ICSLP, pp.29–32, 1998.

[59] Kim, S.J., Kim, J.J., e Hahn, M., "*Implementation and evaluation of an HMM-Based Korean speech synthesis system"*, IEICE trans. Inf.&Syst., vol.E89-D, no.3, pp1116-1119, 2006.

[60] *Reference Manual for Speech Signal Processing Toolkit Ver. 3.4.1*., pp. 111-112, Abril/201. Acessado em Dezembro/2012. Disponível em: http://sp-tk.sourceforge.net

[61] G. Fant, *Speech sound and features*, MIT Press, Cambridge, 1973.

[62] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "*Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling"*, Proc. ICASSP, 1999.

[63] Miyazaki,N., Kobayashi, T., Tokuda, K., Masuko, T., "*Multi-Space Probability Distribution HMM",* IEICE Transactions on Information and Systems, pages 1579-1589, 2000.

[64] Kawahara, H., Estill, J. e Fujimura, O.,*"Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight"*, Proc. of MAVEBA, pp.13–15, 2001.

[65] Oppenheim A. e Johnson, D., *"Discrete representation of signals"*, Proc. of IEEE, pp.681–691, 1972.

[66] Kawahara, H., Katayose, H., Cheveign'e, A. e Patterson,R., *"Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity",* Proc. of Eurospeech, pp.2781–2784, 1999.

[67] Tokuda, K., Zen, H. e Black, A.W., *"An HMM-Based Speech Synthesis System Applied to English"*, In: Proc. IEEE Speech Synthesis Workshop, 2002.

[68] Tabela SAMPA. Acessado em dezembro/2012. Disponível em: http://www.phon.ucl.ac.uk/home/sampa/portug.htm

[69] Zen, H., Tokuda, K., Kitamura, T., "*Decision tree based simultaneous clustering of phonetic contexts, dimensions, and state positions for acoustic modeling*", In: Proc. Eurospeech, pp. 3189–3192, 2003.

[70] Shinoda, K., Watanabe, T., *"Acoustic Modeling Based on The MDL Principle for Speech Recognition,"* Proc.EUROSPEECH, pp. 99-102, 1997.

[71] Odell, J., *"The Use of Context in Large Vocabulary Speech Recognition*", Ph.D. Thesis, University of Cambridge, 1995.

Apêndice A

A-1) Formato das Etiquetas no HTS

Para realizar a síntese utilizando o software HTS é necessário gerar um arquivo de etiquetas que contenha não só a transcrição fonética do texto, mas também as informações contextuais de cada fone, bem como as características prosódicas obtidas durante o processamento do *Front-End*. Um exemplo do formato de etiquetação utilizado por este software e implementado nos demos para a língua portuguesa [12,13] e para a língua inglesa [67] disponibilizados no site [2], possui a estrutura apresentada na Figura A.1. Neste exemplo, cada fone é etiquetado com 46 parâmetros, os quais fornecem informações de posicionamento e de tonicidade relativos aos fones, às sílabas, às palavras, a cada grupo fônico e à frase completa.

$m_1\hat{}m_2-m_3+m_4=m_5$ /M2:$m_6\_m_7$
/S1:$s_1\_@s_2-s_3\_@s_4+s_5\_@s_6$ /S2:$s_7\_s_8$ /S3:$s_9\_s_{10}$ /S4:$s_{11}\_s_{12}$ /S5:$s_{13}\_s_{14}$ /S6:s_{15}
/W1:$w_1\_\#w_2-w_3\_\#w_4+w_5\_\#w_6$ /W2:$w_7\_w_8$ /W3:$w_9\_w_{10}$ /W4:$w_{11}\_w_{12}$ /W5:w_{13}
/P1:$p_1\_!p_2-p_3\_!p_4+p_5\_!p_6$ /P2:$p_7\_p_8$
/U:$u_1\_\$u_2\_\&u_3$

m_1	the phoneme identity before the previous phoneme
m_2	the previous phoneme identity
m_3	the current phoneme identity
m_4	the next phoneme identity
m_5	the phoneme after the next phoneme identity
m_6	position of the current phoneme identity in the current syllable (forward)
m_7	position of the current phoneme identity in the current syllable (backward)
s_1	whether the previous syllable stressed or not (0: not stressed, 1: stressed)
s_2	the number of phonemes in the previous syllable
s_3	whether the current syllable stressed or not (0: not stressed, 1: stressed)
s_4	the number of phonemes in the current syllable
s_5	whether the next syllable stressed or not (0: not stressed, 1: stressed)
s_6	the number of phonemes in the next syllable
s_7	position of the current syllable in the current word (forward)
s_8	position of the current syllable in the current word (backward)
s_9	position of the current syllable in the current phrase (forward)
s_{10}	position of the current syllable in the current phrase (backward)
s_{11}	the number of stressed syllables before the current syllable in the current phrase
s_{12}	the number of stressed syllables after the current syllable in the current phrase
s_{13}	the number of syllables, counting from the previous stressed syllable to the current syllable in this utterance
s_{14}	the number of syllables, counting from the current syllable to the next stressed syllable in this utterance
s_{15}	name of the vowel of the current syllable
w_1	part-of-speech classification of the previous word
w_2	the number of syllables in the previous word
w_3	part-of-speech of classification of the current word
w_4	the number of syllables in the current word
w_5	part-of-speech classification of the next word
w_6	the number of syllables in the next word
w_7	position of the current word in the current phrase (forward)
w_8	position of the current word in the current phrase (backward)
w_9	the number of content words before the current word in the current phrase
w_{10}	the number of content words after the current word in the current phrase
w_{11}	the number of words counting from the previous content word to the current word in this utterance
w_{12}	the number of words counting from the current word to the next content word in this utterance
w_{13}	interrogation flag of the current word
p_1	the number of syllables in the previous phrase
p_2	the number of words in the previous phrase
p_3	the number of syllables in the current phrase
p_4	the number of words in the current phrase
p_5	the number of syllables in the next phrase
p_6	the number of words in the next phrase
p_7	position of the current phrase in this utterance (forward)
p_8	position of the current phrase in this utterance (backward)
u_1	the number of syllables in this utterance
u_2	the number of words in this utterance
u_3	the number of phrases in this utterance

Figura A.1 – Estrutura de etiquetação no HTS
Fonte: [2]

Retomando o exemplo apresentado no Capítulo 2, o início da frase *"#lejla te~"* apresenta a seguinte etiquetagem:

y^y-#+l=e/M2:y_y/S1:0_@0-_@y+1_@3/S2:y_y/S3:y_y/S4:y_y/S5:y_y/S6:y/W1:0_#0-y_#y+content_#2/
W2:y_y/W3:y_y/W4:y_y/W5:0/P1:0_!0-y_!y+8_!5/P2:1_1/U:8_$5_&1

y^#-l+e=j/M2:1_3/S1:0_@0-1_@3+0_@2/S2:1_2/S3:1_8/S4:1_3/S5:0_4/S6:e/W1:0_#0-content_#2+content_#1/
W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

#^l-e+j=l/M2:2_2/S1:0_@0-1_@3+0_@2/S2:1_2/S3:1_8/S4:1_3/S5:0_4/S6:e/W1:0_#0-content_#2+content_#1/
W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

l^e-j+l=a/M2:3_1/S1:0_@0-1_@3+0_@2/S2:1_2/S3:1_8/S4:1_3/S5:0_4/S6:e/W1:0_#0-content_#2+content_#1/
W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

e^j-l+a=t/M2:1_2/S1:1_@3-0_@2+0_@3/S2:2_1/S3:2_7/S4:1_3/S5:1_3/S6:a/W1:0_#0-content_#2+content_#1/
W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

j^l-a+t=e~/M2:2_1/S1:1_@3-0_@2+0_@3/S2:2_1/S3:2_7/S4:1_3/S5:1_3/S6:a/W1:0_#0-content_#2+content_#1/
W2:1_5/W3:1_4/W4:0_1/W5:0/P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1

⋮

Considerando a quarta sentença de etiquetas destacada em azul no exemplo, tem-se a seguinte correspondência ao formato de etiquetas adotado:

*m*1^*m*2-*m*3+*m*4=*m*5 /	**l^e-j+l=a/**
M2:*m*6_*m*7/S1:*s*1_ @*s*2-*s*3_@*s*4+*s*5_@*s*6 /	**M2:3_1/S1:0_@0-1_@3+0_@2/**
S2:*s*7_*s*8 /S3:*s*9_*s*10 /S4:*s*11_*s*12 /S5:*s*13_*s*14	**S2:1_2/S3:1_8/S4:1_3/S5:0_4**
/S6:*s*15/W1:*w*1_#*w*2-*w*3_#*w*4+*w*5_#*w*6/W2:*w*7_*w*8	**/S6:e/W1:0_#0-content_#2+content_#1/W2:1_5**
/W3:*w*9_*w*10 /W4:*w*11_*w*12 /W5:*w*13/	**/W3:1_4/W4:0_1/W5:0/**
P1:*p*1 !*p*2-*p*3_!*p*4+*p*5_!*p*6 /P2:*p*7_*p*8/U:*u*1_$*u*2_&*u*3	**P1:0_!0-8_!5+0_!0/P2:1_1/U:8_$5_&1**

Note que o símbolo m_3 representa o fone tratado na sentença. No caso do exemplo, trata-se do fone ***'j'***, enquanto que os dois fones à sua esquerda são m_1= *'l'* e m_2= *'e'* e os dois fones à sua direita são m_4 = *'l'* e m_5= *'a'*. Desta forma fica explicito o contexto em que o fone m_3= '*j*' aparece.

Os arquivos de etiquetas de todas as frases pertencentes à base de fala de treinamentos se encontram na pasta '*data/labels/full'*. O programa varre todos estes arquivos analisando os rótulos m_1, m_2, m_3, m_4 e m_5 de modo a gerar uma lista (*'data/lists/full'*), a qual contém todos os contextos para cada fone presente na base de fala. Como é impossível contemplar todos os

possíveis contextos em uma base de fala limitada, o programa trabalha com árvores de decisão e agrupamento de contexto para permitir a síntese de fones em contextos que não pertencem ao corpus de treinamento. Maiores detalhes sobre este procedimento são apresentados na Seção A.3. As demais etiquetas presentes na Figura A.1 servem para detalhar informações prosódicas e são utilizadas para a construção de árvores de decisão através de um sistema de perguntas binárias. É imediato intuir que, quanto mais etiquetas forem fornecidas, mais ricas serão as informações contextuais e prosódicas do texto e, consequentemente, melhor será a qualidade melódica da fala sintetizada. Entretanto, é necessário haver mais material de treinamento para que os diferentes contextos apareçam na base de fala e possam ser modelados com precisão. Além disso, aumenta-se a complexidade no processo de análise do texto durante a etapa de *Front-End*.

A-2) Unidades Acústicas Adotadas

Para gerar o conjunto de etiquetas, conforme apresentado na seção precedente, é necessário que o texto tenha sido transcrito foneticamente para gerar os rótulos de m_1 a m_5. Esta transcrição fonética é realizada considerando as unidades individuais da fala que são denominadas **fonemas**. Existem alfabetos fonéticos que permitem representar estes fonemas utilizando uma simbologia padronizada. Neste trabalho, baseou-se nas regras estabelecidas pelo *Speech Assessment Methods Phonetic Alphabet* (SAMPA) [68]. A Tabela A.1 mostra o conjunto das 35 unidades acústicas da língua portuguesa falada no Brasil, adotadas neste trabalho, seguida de exemplos para melhor ilustrar o som correspondente.

Observe que poderiam ser propostas mais unidades acústicas, como, por exemplo, um *E* para contemplar o som *'é',* da palavra *'céu', 'mel',* ou ainda o som *'T'*, para representar o típico som do '*t'* de *'tia'* falado no interior do estado de São Paulo. Aumentando a biblioteca de fones possíveis, melhora-se o modelo do fonema correspondente, entretanto aumenta-se a complexidade do sistema de transcrição fonética e torna-se necessário mais material de treinamento no corpus.

Tabela A.1 – Unidades acústicas básicas do Português		
Classe	**Símbolo**	**Exemplo**
Consoantes		
Plosivas		
	p	**p**ai
	b	**b**arco
	t	**t**enho
	d	**d**oce
	k	**c**om
	g	**g**rande
Fricativas		
	f	**f**alo
	v	**v**erde
	s	**c**éu
	z	ca**s**a
	S	**ch**apéu
	Z	**j**oia
Nasais		
	m	**m**ar
	n	**n**ada
	J	vi**nh**o
Líquidas		
	l	**l**anche
	L	traba**lh**o
	r	ca**r**o
	R	**r**ua
	X	casa**r**
Vogais		
	a	lu**a**
	e	faz**e**r
	i	b**i**co
	O	f**o**rte
	o	c**o**r
	u	f**u**t**u**ro
	a~	**an**dar
	e~	**en**tão
	i~	f**im**
	o~	b**om**
	u~	**um**
Semi Vogais		
	j	pa**i**
	j~	mu**i**to
	w	fáci**l**
	w~	cã**o**

A-3) Árvore de Decisão

Existem inúmeros fatores contextuais, prosódicos e fonéticos que afetam os modelos espectrais, de frequência de pitch e a duração dos fones, tais como a posição do fone na frase, o tipo de fone, a sílaba tônica na palavra, entre outros. Estes fatores são considerados na construção da etiquetagem das frases para permitir a construção de modelos precisos e obter fala sintetizada de boa qualidade prosódica.

Entretanto, em uma base de fala de dimensão limitada, é impossível que apareçam os infinitos casos dos fonemas contextuais de uma língua. Uma das vantagens que a técnica de síntese baseada em HMM apresenta é a necessidade de uma base de fala de tamanho reduzido, contendo cerca de 500 a 1000 frases para realizar o treinamento dos modelos. Esta limitação do corpus só é possível porque o HTS utiliza a técnica da árvore de decisão, baseada em agrupamento de contexto (*clustering*) para as distribuições do espectro, pitch e duração dos estados [69]. Com esta técnica é possível estimar os parâmetros dos modelos HMM para os fonemas que aparecem em contextos que não estavam presentes na base de fala, a partir dos modelos existentes nela e que se assemelham a ele.

Como cada fator contextual influencia de forma diferente as distribuições de espectro, pitch e duração dos estados, são necessárias três diferentes árvores de decisão independentes entre si para agrupar os diferentes contextos. A árvore de decisão é uma estrutura binária construída com base nas etiquetas e em questões pré-determinadas que relacionam os fones (*state tying*). Cada questão é definida por 3 campos:

QS – descrição da questão – conjunto de padrões

Por exemplo, o comando:

QS "LL_Anterior_vowel" { e^*,i^*,e~^*,i~^*}

questiona se o símbolo presente à esquerda-esquerda (LL), ou seja m_1, pertence ao conjunto dos símbolos {e^*,i^*,e~^*,i~^*}, onde '*' representa qualquer caractere.

Para realizar o treinamento dos modelos HMM é necessário especificar arquivos de questões. Estes documentos estão presentes em '*data/questions_utt_qst001*' e '*data/questions_qst001'* e devem conter tantas questões quantas forem necessárias de forma a contemplar todos os casos possíveis de acordo com as etiquetas dependentes de contexto

apresentadas na Figura A.1. O primeiro arquivo contempla questões somente sobre os *labels* que fazem referência ao número de sílabas, palavras e frases da sentença, ou seja, '*U:u1 $u2 &u3'*. O segundo arquivo aborda questões sobre os demais *labels* do conjunto de etiquetas, conforme o padrão apresentado na Figura A.1.

O algoritmo utilizado para construir a árvore de decisão utiliza o critério *Minimum Description Length* (MDL) [70] de forma a escolher as questões necessárias dentre aquelas presentes nos arquivos. A função HHEd do HTK permite gerar a árvore de decisão a partir das questões. Para gerar uma voz de qualidade média são necessárias questões que avaliem a tonicidade, a posição relativa e o número de ocorrências de determinado fone. Questões adicionais que considerem a prosódia da palavra, fronteiras da prosódia e *Part-of-Speech,* permitem melhorar a qualidade prosódica da voz artificial.

As três árvores geradas no processo de treinamento dos modelos podem ser encontradas nos seguintes endereços:

- Árvore de duração: *trees\qst001\ver1\dur*
- Árvores de pitch e espectro: *trees\qst001\ver1\cmp*

As Figuras A.2, A.3 e A.4 apresentam exemplos de árvores de decisão para cada modelo. Note que os ramos finais das figuras seguem com questões, exceto aqueles preenchidos com preto, que indicam o nó final e contém o índice a ser usado para consultar a PDF do referido modelo.

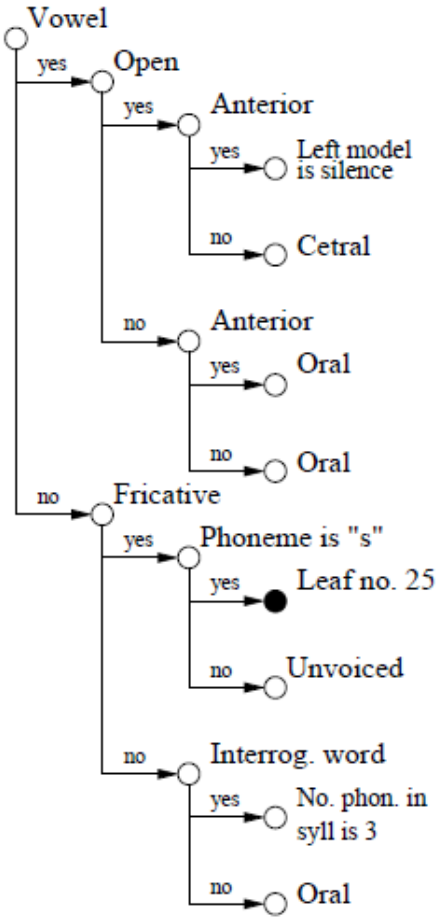

Figura A.2 – Questões sobre atributos fonéticos – árvore para espectro
Fonte: [13]

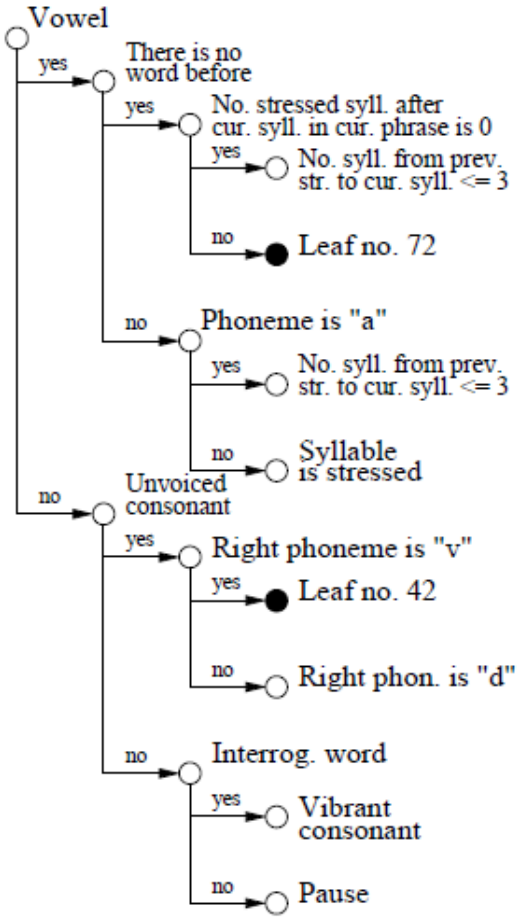

Figura A.3 – Questões sobre atributos linguísticos – árvore para pitch
Fonte: [13]

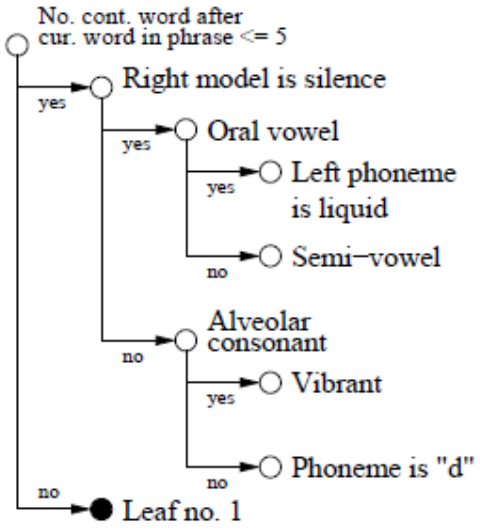

Figura A.4 – Questões linguísticas para pausa – árvore de duração
Fonte: [13]

A técnica de agrupamento de contexto é adotada para tratar os casos de fonemas em contextos que não apareceram no corpus de treinamento. A Figura A.5 apresenta um diagrama ilustrativo do procedimento de *clustering* em modelos HMM de 3 estados. Neste exemplo, têm-se os modelos HMM do fone '*a*' da língua inglesa, em diversos contextos, tais como '*kab*' e '*tan*'. Para originar os modelos HMM que não existem, realiza-se uma estimativa com base nos modelos existentes e na árvore de decisão. É razoável supor que o 2º estado do HMM do modelo do '*a*' nos diversos contextos seja semelhante ao 2º estado do fone '*a*' no contexto '*wat*', por exemplo. Analogamente, podem-se estimar todos os outros estados de todos os fones que aparecem em contextos não tratados durante a fase de treinamento.

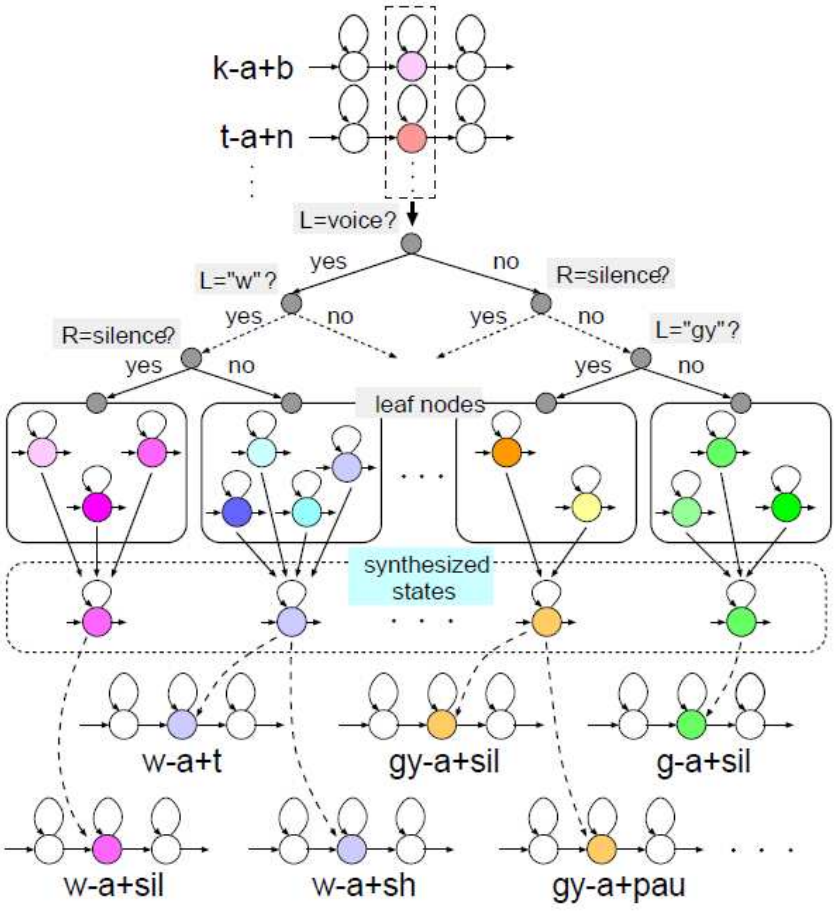

Figura A.5 – Agrupamento de contexto para o som '*a*'
Fonte: [71]