Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Departamento de Engenharia Elétrica
Programa de Pós-Graduação em Engenharia Elétrica

Dissertação de Mestrado

Desenvolvimento de um Codificador de Voz Pessoal de Baixa Taxa Baseado em Modelos de Markov Escondidos

Raissa Bezerra Rocha

Campina Grande – PB

Julho de 2012

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Departamento de Engenharia Elétrica
Programa de Pós-Graduação em Engenharia Elétrica

Desenvolvimento de um Codificador de Voz Pessoal de Baixa Taxa Baseado em Modelos de Markov Escondidos

Raissa Bezerra Rocha

Dissertação de Mestrado submetida à Coordenação do Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica da Universidade Federal de Campina Grande como requisito necessário para obtenção do grau de Mestre em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Comunicações

Prof. Dr. Marcelo Sampaio de Alencar
Orientador

Campina Grande – PB, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

R672d Rocha, Raissa Bezerra.
Desenvolvimento de um codificador de voz pessoal de baixa taxa baseado em modelos de Markov escondidos/Raissa Bezerra Rocha. – Campina Grande, 2012.
86f.: il.col.

Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática.
Orientador: Prof. Dr. Marcelo Sampaio de Alencar
Referências.

1. Codificação de Voz. 2. Codificação Fonética. 3. Taxa de Transmissão. 4. Reconhecimento de Fonemas. I. Título.

CDU 621.391(043)

IMPLEMENTAÇÃO DE UM CODIFICADOR DE VOZ COM BAIXA TAXA DE TRANSMISSÃO USANDO MODELOS DE MARKOV ESCONDIDOS

RAÍSSA BEZERRA ROCHA

Dissertação Aprovada em 25.07.2012

MARCELO SAMPAIO DE ALENCAR, Ph.D., UFCG
Orientador

WASLON TERLLIZZIE ARAUJO LOPES, D.Sc., UFCG
Componente da Banca

FRANCISCO MADEIRO BERNARDINO JUNIOR, D.Sc., UPE
Componente da Banca

LUCIANA RIBEIRO VELOSO, D.Sc., UFPB
Componente da Banca

CAMPINA GRANDE - PB
JULHO -2012

A Deus!

Aos meus pais, Wilson e Gláucia, e aos meus amados sobrinhos Mateus Richelle e Wilson Neto.

Agradecimentos

O mais belo agradecimento será sempre muito pouco para traduzir minha gratidão a Ele, que ilumina minha vida com seu infinito amor. Que é meu pai, meu refúgio e caminho. A Ele dedico e sempre dedicarei todas as conquistas da minha vida. Ao meu grande amigo Jesus Cristo, muito obrigada por tudo!

Aos meus pais, Wilson e Gláucia, que acreditaram nesse sonho junto comigo, que me ensinaram o caminho do bem e revestiram a minha vida de amor. Sem dúvida, são as maiores bençãos que a Sabedoria Divina colocou em meu caminho.

A Thiago, por estar sempre ao meu lado, por ser meu amigo, cúmplice e companheiro, e principalmente, por me ter feito feliz nesses oito anos de convivência. O teu amor deu vida à minha vida. Amor, obrigada!

Aos meus familiares e amigos que souberam compreender minha ausência, não me abandonando em nenhum instante desse desafio. Que sempre estiveram ao meu lado, me ajudando, incentivando, me dando força e coragem para ultrapassar as adversidades da vida. Em especial, ao meu irmão Gláucio, que nunca me faltou nos momentos que eu precisei da sua ajuda.

Ao meu orientador, Professor Marcelo Sampaio de Alencar, por te me aceitado como orientanda desde a época da graduação como aluna de Iniciação Científica, me incentivando a procurar a pesquisa e fortalecendo o meu desenvolvimento profissional. Em especial, por todos os conselhos e ensinamentos, pelo carinho, atenção e paciência que sempre teve comigo. Pessoa que terá sempre meu respeito e admiração. Que me serviu e sempre servirá de exemplo e referência em todos os momentos da vida. De coração, muito obrigada!!

Agradeço a todos os meus amigos do Iecom, pela companhia, conselhos, incentivos, ensinamentos e por todos os momentos agradáveis. Sem dúvida, essa jornada teria sido mais difícil sem a companhia de vocês.

Ao apoio da Universidade Federal de Campina Grande (UFCG), do Instituto de Estudos Avançados em Comunicações (Iecom) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

A todos, aceitem meus sinceros agradecimentos.

"Enquanto viver, falarei da Tua bondade,
e levantarei as minhas mãos em oração."
Salmo 63.

Resumo

Esta dissertação apresenta o desenvolvimento de um codificador de voz que tem como principal característica a transmissão do sinal de voz com baixas taxas de *bits*. Desenvolvido principalmente para ser utilizado em sistemas móveis celulares, o codificador proposto é do tipo fonético, que entre as técnicas de codificação de sinais de fala é a que permite obter menor taxa de transmissão.

Sua implementação está dividida no desenvolvimento do emissor e receptor. No emissor, os sinais de fala são segmentados por meio de um reconhecedor de fonemas que utiliza Modelos de Markov Escondidos (HMMs) para modelar o sinal de voz. A cada fonema é atribuído um índice pré-estabelecido e sua duração e energia são estimados. A informação transmitida ao receptor consiste no índice, energia e duração de cada fonema. Deste modo, o codificador consegue uma redução na taxa de transmissão do sinal de voz.

O receptor é constituído em duas etapas. Na primeira, cada usuário do codificador tem que construir um banco de unidades acústicas por meio da pronúncia de frases pré-estabelecidas. Na segunda etapa, é realizada a síntese por concatenação de segmentos como sílabas, fonemas e encontros vocálicos.

Para avaliar o desempenho do codificador foi realizado um teste subjetivo informal baseado no teste ACR (*Absolute Category Rating*). Duas avaliações foram feitas. A primeira utiliza segmentação automática no emissor e receptor e o codificador permitiu a transmissão do sinal de voz com uma taxa de, no máximo, 150 *bits*/s. Os resultados da qualidade dos sinais de voz indicam que os avaliadores classificam a maioria deles como de razoável a boa.

Na segunda avaliação, a segmentação utilizada para formar o banco de unidades acústicas foi realizada de forma manual. Sessenta e dois ouvintes-avaliadores foram questionados a respeito da inteligibilidade e qualidade dos sinais de voz. Os sinais de voz foram codificados com 125 *bits*/s e a maior parte deles apresentaram bons níveis de inteligibilidade e foram considerados sinais de fala de qualidade razoável.

Palavras-Chave: Codificação de voz, codificação fonética, taxa de transmissão, reconhecimento de fonemas.

Abstract

This dissertation presents the development of a voice encoder which has the transmission of voice signals with low bitrates as its main feature. Developed mainly for utilization in mobile cellular systems, the proposed encoder uses the phonetic coding technique, which provides the lowest transmission rate.

Its implementation is divided into the development of the emitter and the receiver. In the emitter, the speech signals are segmented by a phoneme recognizer which utilizes Hidden Markov Models (HMMs) to model the voice signal. A pre-established index is assigned to each phoneme and its duration and energy are estimated. The information transmitted to the receiver consists of the index, energy and duration of each phoneme. This way the encoder achieves a reduction in the voice signal transmission rate.

The receiver is constituted in two steps. In the first, each encoder user has to build an acoustic unit bank by pronunciation of pre-established phrases. The second step is a synthesis by concatenation of segments as syllables, phonemes and vowel meetings.

To evaluate the performance of the encoder, an informal subjective test based on the ACR (Absolute Category Rating) test was used. Two evaluations were done. The first used automatic segmentation in the emitter and receiver, and the encoder allowed transmission of the voice signal with a rate of up to 150 bits/s. The results of the voice signal quality indicate that the evaluators classified most of the samples as average to good.

In the second evaluation the segmentation used to form the acoustic unit bank was done manually. Sixty-two listening evaluators were questioned about the intelligibility and quality of the speech signals. The voice signals were coded using 125 bits/s, and most of them presented good levels of intelligibility and reasonable quality.

Key Words: Voice Coding, phonetic coding, transmission rate, phoneme recognition.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

ACELP	*Algebraic Code Excited Linear Predictive*	Preditivo Linear Algébrico Excitado por Código
ACR	*Absolute Category Rating*	Ordenamento por Categoria Absoluta
ADPCM	Adaptative Differential Pulse Code Modulation	Modulação por Codificação Diferencial de Pulso Adaptativa
APCM	Adaptive Pulse Code Modulation	Modulação por Codificação de Pulso Adaptativa
AMPS	*Advanced Mobile Phone System*	Sistema de Telefonia Móvel Avançado
AMR-NB	*Adaptive Multi-Rate Narrowband*	Multitaxa Adaptativa de Faixa Estreita
AMR-WB	*Adaptive Multirate Wideband*	Multitaxa Adaptativa de Faixa Larga
cdmaOne	*Code Division Multiple Access One*	
CELP	*Code Excited Linear Prediction*	Predição Linear Excitada por Código
CT2	*Cordless Telephone II*	Telefone sem Fio II
CNPq		Conselho Nacional de Desenvolvimento Científico e Tecnológico
DCME	*Digital Circuit Multiplication Equipment*	Equipamento de Multiplicação com Circuito Digital
DECT	*Digital European Cordless Telephone*	Telefone sem Fio Digital Europeu
DPCM	Differential Pulse Code Modulation	Modulação por Codificação Diferencial de Pulso
DTX	*Discontinuous Transmission*	Transmissão Descontínua
DTW	*Dynamic Time Warping*	Alinhamento Dinâmico no Tempo
FFT	*Fast Fourier Transform*	Transformada Rápida de Fourier
GSM	Global System for Mobile Communications	Sistema Global para Comunicações Móveis
HMM	*Hidden Markov Models*	Modelos de Markov Escondidos
HTK	*Hidden Markov Models Toolkit*	
IDCT	*Inverse Discrete Cosine Transform*	Inversa da Transformada Discreta do Cosseno
Iecom	*Institute of Advanced Studies in Communications*	Instituto de Estudos Avançados em Comunicações

IFFT	Inverse Fourier Transform	Transformada Inversa de Fourier
ITU	*International Telecommunication Union*	União Internacional de Telecomunicações
ITU-T	International Telecommunication Union-Telecommunication Sector	
FDMA	*Frequency Division Multiple Acess*	Múltiplo Acesso por Divisão na Frequência
FFT	*Fast Fourier Transform*	Transformada Rápida de Fourier
LD-CELP	*Low Delay Code Excited Linear Prediction*	Predição Linear Excitada por Código com Baixo Atraso
LPC	*Linear Predictive Coding*	Codificação por Predição Linear
MFCC	*Mel Frequency Cepstral Coeficients*	Coeficientes Cepestrais de Frequência Mel
MIPS	*Millions of Instructions per Second*	Milhões de Instruções por Segundo
MLF	*Master Label File*	Arquivo de Rótulo Principal
MMF	*Master Macro File*	Arquivo Macro Principal
MOS	*Mean Opinion Score*	Escore Médio de Opniões
MP-MLQ	*Multi-Pulse Maximum Likelihood Quantization*	Quantização de Máximo Verossimilhança Multi-Pulso
PAM	*Pulse Amplitude Modulation*	Modulação por Amplitude de Pulso
PCM	*Pulse Code Modulation*	Modulação por Codificação de Pulso
PNCC	*Power-Normalized Cepstral Coefficients*	Coeficientes Cepstrais Normalizados em Potência
QCELP	*Qualcomm Code Excited Linear Predictive*	Predição Linear Excitada por Código Qualcomm
RAM	*Random-Access Memory*	
RPE-LTP	*Regular Pulse Excited-Long Term Predictor*	Preditor de Longo Prazo com Excitação Regular por Pulso
ROM	Read-Only Memory	
SCR	*Source Controlled Rate*	Taxa Controlada pela Fonte
SD	*Spectral Distortion*	Distorção Espectral
SG15	*Study Group* 15	Grupo de Estudo 15
SID	*Silence Descriptor*	Descritor de Silêncio
SNR	*Signal-to-Noise Ratio*	Relação Sinal Ruído
SQNR	*Signal-to-Quantization Noise Ratio*	Relação Sinal Ruído de Quantização
SSCH	*Subband Spectral Centroid Histograms*	Histograma de Subbandas para um Centróide Espectral
TDMA	*Time Division Multiple Acess*	Múltiplo Acesso por Divisão no Tempo
UFCG	*Federal University of Campina Grande*	Universidade Federal de Campina Grande

UNESCO	*United Nations Economic, Scientific and Cultural Organization*	Organização das Nações Unidas para a Educação, a Ciência e a Cultura
VBR	*Variable Bit Rate*	Taxa de *bits* variável
VSELP	*Vector Sum Excited Linear Predictive*	Preditivo Linear Excitado por Soma Vetorial
wav	*Waveform Audio Format*	Formato Forma de Onda de Áudio
WCDMA	*Wideband Code Division Multiple Acess*	Múltiplo Acesso por Divisão de Código em Banda Larga
WER	*Word Error Rate*	Taxa de Erro de Palavra
3GPP	*3rd Generation Partnership Project*	Projeto de Parceria para Terceira Geração

Lista de Símbolos

y	Saída do quantzador na modulação Delta.
$u(x)$	Função degrau.
$h(n)$	Filtro digital variante com o tempo.
$s(n)$	Saída de um sistema linear variante com o tempo.
$u(n)$	Ruído aleatório.
G_f	Fator de ganho.
$y(n)$	n-ésima amostra de saída predita do trato vocal.
$x(n)$	Entrada amostrada em um tempo n.
p	Ordem do modelo LPC.
a_k	k-ésimo coeficiente do preditor LPC.
C	Dicionário do codificador CELP.
k	Quantidade de excitações do dicionário do codificador CELP.
k_1	Número de sequências retiradas de amostras passadas do codificador CELP.
k_2	Número de sequências estocásticas que constituem o dicionário fixo do codificador CELP.
$W(z)$	Filtro de ponderação ou perceptivo.
$R(s,d)$	Correlação entre o sinal de voz e as respostas da excitação do dicionário.
$R(d,d)$	Autocorrelação entre as respostas correspondentes a cada excitação do dicionário.
$s_p(n)$	Sinal de voz após a pré-ênfase.
$c(t)$	Janela retangular.
$J(n)$	Janela (Hamming ou Hanning).
S_s	Erros por substituição.
I	Erros por inserção.
D_s	Erros por supressão.
N_s	Número de palavras da sequência de teste.
o_t	Vetor de observação no instante de tempo t.
$b_j(o_t)$	Densidade de probabilidade de saída do estado j no instante de tempo t.
O	Sequência de vetores de observação.
O_t	Vetor formado pelos parâmetros obtidos para cada bloco de amostras do sinal de voz.

N	Número de estados de um modelo HMM.
M	Número de símbolos do alfabeto.
S	Estados do modelo HMM.
$q_t(i)$	Indica estar no estado S_i no tempo t.
A	Distribuição de probabilidade de transição entre os estados.
a_{ij}	Coeficientes da matriz de transição A.
B	Representa a distribuição de probabilidade de observação dos símbolos.
$b_j(k)$	Probabilidade da variável aleatória o_t pertencer ao estado j.
c_{jm}	Coeficiente de ponderação.
u_{jm}	Vetor média.
U_{jm}	Matriz covariância.
λ	Conjuntos dos parâmetros do modelo HMM.
N_g	Distribuição gaussiana.
Π	Distribuição do estado inicial do modelo HMM.
n	Ordem de um modelo HMM.
l	Unidade de treinamento.
$\alpha_t(i)$	Probabilidade de avanço.
$B_t(i)$	Probabilidade de retrocesso.
$\hat{c}(t)$	Fator de normalização.
q_t*	Sequência de estado ótima.
$\delta_t(i)$	Máximo valor de probabilidade em um caminho no algoritmo de Viterbi.
X	Sequência de informação acústica observada.
W	Palavra pronunciada.
$\hat{W}$	Estimativa da palavra pronunciada.
m	Quantidade de blocos do sinal de voz.
N_A	Número de amostras contida em cada bloco do sinal de voz.
$L(z)$	Transformada Z do filtro utilizado na pré-ênfase.
a_p	Fator de pré-ênfase.
E_i	Energia do fonema i.
D_i	Duração do fonema i.
E	Energia do sinal de voz.
N_F	Quantidade de amostras em cada fonema.
t_d	Duração de um fonema.
t_f	Tempo final de um fonema.
t_i	Tempo inicial de um fonema.
N_b	Quantidade de utilizadores do codificador.
E_1	Energia do segmento pronunciado pelo orador.
E_2	Energia do segmento armazenado.
NA_1	Quantidade de amostras do primeiro fonema armazenado utilizado na interpolação.

NA_2	Quantidade de amostras do segundo fonema armazenado utilizado na interpolação.
NA_{d1}	Quantidade de amostras que representa a duração do segmento recebido pelo receptor.
NA_i	Quantidade de amostras utilizadas na interpolação linear.
NA_{d2}	Quantidade de amostras restantes do segundo fonema usado na interpolação.
A_{int}	Amostras resultantes da interpolação.
(Z,X)	Par de processos estocásticos.
N_e	Número de estados de um modelo HMM.
c_{jk}	Coeficiente de ponderação para a k-ésima mistura do estado j.
G	Função densidade de probabilidade gaussiana multidimensional.
D	Dimensão do vetor o_t.
$\|U_{jk}\|$	Determinante da matriz de covariância.
U_{jk}^{-1}	Inversa da matriz de covariância.
$\eta(o_t)$	Conjunto das funções densidade de probabilidade.
K	Número de funções densidade de probabilidade.
v_k	k-ésimo símbolo de saída.
$c_j(k)$	Coeficiente de ponderação das gaussianas.
$f(o_t\|v_k)$	k-ésima função densidade de probabilidade para o vetor de parâmetros de entrada o_t.

CAPÍTULO 1

Introdução

A voz é o principal meio de comunicação do homem. A distância da comunicação falada foi vencida em 1876 quando Alexander Graham Bell inventou o telefone e, desde então, este tipo de comunicação não tem parado de crescer.

Um século após a descoberta do telefone, na década de 1970, tiveram início os sistemas de comunicações móveis celulares com a primeira geração de telefonia celular (1G), cujo precursor foi o sistema americano AMPS (*Advanced Mobile Phone System*). Naquela geração, os serviços oferecidos resumiam-se basicamente à comunicação analógica de voz e envio de pequenas mensagens.

No início da década de 1990, começa a segunda geração telefonia celular (2G), cuja principal evolução em relação à geração anterior foi a transmissão digital dos sinais de voz. Essa evolução só foi possível com o desenvolvimento de técnicas de codificação de voz. Nas últimas décadas, os sistemas de telefonia móvel evoluíram de maneira expressiva para suprir a crescente demanda por novos serviços, maiores taxas de transmissão e aumento da capacidade da rede.

O serviço mais importante oferecido ao usuário de uma rede móvel celular é a transmissão da voz. No entanto, devido à capacidade de transmissão restrita do canal, é preciso minimizar o número de *bits* que devem ser transmitidos, tornando-se necessário pesquisar técnicas que busquem diminuir a taxa de transmissão necessária para a representação do sinal digital.

A compressão de sinal tem o objetivo de reduzir o número de *bits* necessário para representar adequadamente os sinais de voz, imagem, vídeo ou áudio. Desta forma, em sistemas cuja capacidade de armazenamento e largura de banda são limitados, como no caso de sistemas de telefonia móvel, a compressão de sinais torna-se necessária, uma vez que define o número de *bits* que representa cada segundo de fala do sinal a ser transmitido, parâmetro fundamental em aplicações envolvendo transmissão e armazenamento de sinais [1].

Particularmente no caso do sinal de voz, os codificadores têm o objetivo de reduzir o número de *bits* necessários para representar adequadademente o sinal de voz. Deste modo, é possível se comunicar em canais de baixa capacidade, além de multiplexar mais sinais no mesmo canal de comunicação devido à diminuição da largura de banda requerida por usuário, aumentando o número de usuários, o que pode tornar o sistema telefônico mais barato e acessível à população.

Para ser possível sua aplicação em sistemas celulares, o processo de codificação deve ser simples e rápido o suficiente para que possa funcionar em tempo real com processadores relativamente

baratos e de baixo consumo. Outro requisito importante é a qualidade da voz codificada, que deve ser tal que permita não só a inteligiblidade, mas também que se possa reconhecer o interlocutor e perceber outras informações como a entonação e a emoção. Além disso, no desenvolvimento do codificador, devem ser levados em consideração atributos como taxa de *bits* produzida, complexidade e memória necessária, retardo de comunicação e sensibilidade aos erros de canal. Esses atributos são descritos a seguir.

1.1 Atributos dos Codificadores de Voz

1.1.1 Taxa de *Bits*

A principal motivação da compressão de sinais é a minimização da taxa de *bits*, que pode ser medida em *bits* por amostra, *bits* por *pixel* (bpp) e *bits* por segundo, dependendo do contexto. A taxa de *bits* consite no produto da taxa de amostragem e do número de *bits* por amostra. A taxa de amostragem é, geralmente, ligeiramente superior a duas vezes a largura de faixa do sinal, de acordo com o Teorema da Amostragem de Nyquist [2, 3].

Tradicionalmente, os codificadores de voz possuem taxa de *bits* fixas. Entretanto, atualmente as redes de comunicação utilizam protocolos flexíveis, capazes de lidar com codificador com taxa de *bits* variável (VBR – *Variable Bit Rate*). Esses codificadores podem ser usados, por exemplo, em sistemas de multiplexação de vários codificadores em um mesmo canal de transmissão. Para atender a um número maior de usuários, o sistema pede aos codificadores que entrem em modo de operação com menor taxa de *bits*, mesmo que isto cause uma diminuição na qualidade do sinal por sua representação com menor precisão, no entanto sempre mantendo uma qualidade mínima aceitável. De modo inverso, quando o tráfego for baixo, os codificadores podem operar em modo de máxima taxa de *bits*, aumentando a qualidade do sinal transmitido.

1.1.2 Qualidade de Sinal Reconstruído

A largura de banda e a taxa de *bits* afetam diretamente a qualidade do sinal. Além disso, devido ao processo de quantização, que consiste na diferença entre o sinal na entrada do quantizador e o sinal na saída, o sinal reconstruído não é igual ao sinal original. De forma geral, há dois tipos de medidas para avaliação da qualidade de sinais: medidas de qualidade objetivas e medidas de qualidade subjetivas.

As medidas de qualidade subjetivas determinam a qualidade perceptual e são processos comuns na validação dos codificadores. Elas são realizadas com testes de escutas ou de visualização e baseiam-se em comparações entre o sinal original e o sinal processado. Deste modo, um grupo de pessoas classifica subjetivamente a qualidade do sinal processado segundo uma escala pré-determinada. É possível citar como exemplo de medida de qualidade subjetiva o Escore Médio de Opniões (MOS), recomendação P.800 do ITU-T [4], em que os ouvintes são confrontados com vídeo, áudio ou sinais de voz processados pelo codificador em teste e atribuem ao sinal reconstruído um

escorre segundo uma escala pré-determinada. No final, é calculada a média aritmética dos escores obtidos e determinado o valor final da avaliação. Outro exemplo de medida de qualidade subjetiva é o Teste de Preferência, realizado por comparações de pares de sinais [5].

Por outro lado, as medidas de qualidade objetivas baseiam-se em uma comparação matemática direta entre o sinal original e o processado. Essas medidas devem apresentar, no mínimo, duas características. Primeiramente devem apresentar correlação com os resultados da avaliação subjetiva, no sentido de que pequenas e grandes variações das medidas objtivas signifiquem pequenas e grandes variações da qualidade subjetiva dos sinais reconstruídos. Segundo, devem ser matematicamente tratáveis e facilmente implementáveis. Exemplos de medidas objetivas utilizadas para avaliação da qualidade do sinal de voz são: relação sinal ruído (SNR, *Signal-to-Noise Ratio*), a relação sinal ruído segmental (SNRseg, *Segmental Signal-to-Noise Ratio*) e a distorção espectral (SD, *Spectral Distortion*) [2].

1.1.3 Complexidade

A complexidade de um algoritmo de codificação é uma questão importante para o custo final do codificador e para o consumo de potência necessário para o seu funcionamento. Sob a perspectiva dos padrões de codificação de voz, onível máximo de complexidade aceitável é determinada pela aplicação.

Velocidade é comumente medida pelo número de instruções por segundo (MIPS, *millions of instructions per second*) necessárias para implementação em tempo real do algoritmo de codificação de voz. A complexidade é geralmente especificada em termos de MIPS e de número de palavras de RAM (*random-access memory*) necessárias para uma implementação, bem como ROM (*read-only memory*). Codificadores de voz requerendo menos de 15 MIPS são considerados de baixa complexidade. Aqueles que requerem 30 MIPS ou mais são considerados de alta complexidade. Do ponto de vista do projetista de sistema, mais complexidade resulta custos mais elevados e mais gasto de potência. Para aplicações portáteis, mais dispêndio de potência significa tempo reduzido entre recargas de bateria ou utilização de baterias maiores, o que implica maiores custo e peso [2].

1.1.4 Retardo de Comunicação

A importância do retardo em um sistema comunicação depende da aplicação. Em sistemas telefônicos o atraso total tem que ser mantido dentro de um limite severo, de forma que não afete a naturalidade da conversação. Segundo experimentos, uma atraso confortável para o ser humano fica na ordem de 100 ms. As redes de telefonia convencional possuem uma latência de 30 ms ou menos [6]. O aumento de complexidade em um algoritmo de codificação é geralmente associado a um aumento de atraso de processamento no codificador e decodificador. Assim, o retardo produzido por um *codec* impõe certas restrições práticas quanto à utilização em sistemas de comunicações, uma vez que o retardo não deve ultrapassar um determinado limite.

1.1.5 Sensibilidade aos Erros de Canal

A sensibilidade aos erros de canal pode ser considerada um aspecto de qualidade. No caso dos padrões de celular digital, são usados *bits* adicionais para codificação de canal, para proteger os *bits* de informação. Nem todos os *bits* de um codificador de voz têm a mesma sensibilidade aos erros de canal. É comum haver duas ou três classes de *bits* mais sensíveis, às quais se dá uma maior proteção contra erros de canal, enquanto às classes de *bit* menos sensíveis aos erros de canal não se dá proteção.

1.2 Motivação

Os algoritmos de codificação de voz podem ser divididos em duas categorias principais: codificadores de forma de onda, que são caracterizados por seguirem o sinal amostra a amostra, ou utilizando um vetor de amostras (quantização vetorial) e os *vocoders*, que descrevem o sinal de fala de um modo paramétrico, conseguindo uma diminuição na taxa de *bits*, mas também na qualidade do sinal sintetizado. Há também a codificação híbrida, que combina a qualidade dos codificadores de forma de onda com a eficiência dos codificadores paramétricos [2].

Entretanto, existem estratégias para codificação da voz a baixa taxa de *bits*, como os *vocoders* segmentais, caracterizados por particionar o sinal de voz em segmentos, que podem ser sílabas, fonemas, difones, entre outros, por meio de técnicas de reconhecimento de fala e são denominados de codificadores fonéticos. Para alcançar uma baixa taxa de transmissão, esse codificador não realiza a codificação do sinal de voz propriamente dito, mas apenas dos parâmetros que caracterizam cada segmento, com os índices e informações como energia, duração e frequência fundamental dos segmentos, que são as características prosódicas do sinal de voz.

Para sintetizar o sinal de voz, a maior parte dos codificadores fonéticos encontrados na literatura realiza a síntese por formantes ou LPC (*Linear Predictive Coding*) com informações previamente armazenadas no receptor e informações prosódicas recebidas do emissor do codificador. Neste caso, o receptor do codificador é constituído de um sistema de reconhecimento de fala que segmenta o sinal em unidades acústicas como difones, trifones, fonemas, entre outros, e extrai informações dos segmentos, como, por exemplo, coeficientes LPC, para a formação de um livro código. Para sintetizar, o receptor busca no livro código (dicionário) as informações do segmento correspondente ao índice recebido e realiza ajustes prosódicos.

No entanto, para formar o livro código, alguns codificadores fonéticos utilizam apenas um orador, tornando o codificador dependente de orador, ou seja, ele será capaz de gerar um sinal de voz de saída inteligível e reconhecível apenas pelo orador para o qual foi treinado.

Para tornar o codificador independente de locutor, deve ser construido um livro de códigos genérico treinado por uma grande variedade de locutores, o que dificulta a construção de tais sistemas. Esses sistemas devem ser capazes de sintetizar a voz de qualquer locutor, mesmos aqueles que não participaram na formação do livro de códigos. No entanto, mesmo nesses casos, há um comprometimento da inteligibilidade e capacidade de reconhecimento do orador pelo uso de um único livro de código genérico na síntese do sinal de voz para todos os oradores.

O codificador proposto neste trabalho é fonético, e se diferencia dos demais encontrados na literatura pelo fato de ser pessoal. Neste caso, para a síntese do sinal de voz, o receptor, em vez de utilizar um livro de códigos genérico para todos os usuários do codificador, utiliza um banco de segmentos específico para cada usuário. Ou seja, o receptor deve conter vários livros de códigos ou banco de unidades acústicas e, ao realizar a síntese do sinal de voz, o receptor busca o banco de unidades daquele usuário específico. Assim, o sinal de voz final mantém as características originais de cada orador, como o timbre, eliminando a necessidade de métodos de adaptação da voz sintetizada à voz pronunciada pelos oradores.

A ideia do uso de um codificador fonético para ser aplicado em sistemas telefônicos celulares parte do princípio de que os usuários, apesar de possuírem vários números cadastrados na agenda do seu aparelho, se comunicam com uma quantidade restrita de pessoas, compreendendo parentes e amigos mais próximos. Nesses casos, o uso do codificador proposto é uma alternativa aos codificadores altualmente utilizados nos padrões móveis, provendo uma comunicação com baixa taxa de transmissão, aumentando a capacidade dos sistemas móveis, podendo até reduzir o custo da ligação, caso as companhias telefônica façam a cobrança por taxa de transmissão.

1.3 Objetivos

Este trabalho tem como objetivo o desenvolvimento de um codificador de voz com baixa taxa de transmissão para o uso principalmente em sistemas de telefonia móvel.

O sistema de codificação proposto deve ser interpretado como alternativo, a ser utilizado principalmente nas comunicações mais frequentes realizadas pelos usuários de telefonia móvel. Como mencionado, o codificador tem a característica de ser pessoal, devido ao uso de um banco de unidades acústicas específico para cada orador na síntese do sinal de voz.

Ao optar por utilizar o codificador, cada usuário inicialmente deve gravar em seu aparelho telefônico frases pré-estabelecidas, utilizadas para compor o seu banco de unidades acústicas, e em seguida enviá-las aos aparelhos telefônicos dos usuários com quem deseja se comunicar, ou gravá-las diretamente no aparelho telefônico desses usuários.

Para alcançar o objetivo da baixa taxa de transmissão, o codificador proposto é do tipo fonético, para o qual se obtém menores taxas de *bits*. Sua implementação está dividida nas epatas de emissão e recepção.

O emissor do codificador é formado por um reconhecedor de fala que utiliza a técnica HMM (*Hidden Markov Models*) para particionar o sinal de voz em segmentos fonéticos. A cada segmento obtido é abribuido um índice pré-estabelecido, totalizando quarenta índices que representam os trinta e oito fonemas da língua portuguesa, além dos índices atribuídos ao silêncio e pausa entre palavras (*short-pause*). De cada segmento fonético reconhecido obtém-se a sua duração e energia média.

Como usual para codificadores fonéticos, as informações transmitidas ao receptor são os índices e características prosódicas, que neste caso, são a duração e energia média de cada fonema.

O receptor é construído em duas etapas. A primeira consiste na segmentação das frases pré-selecionadas em fonemas, sílabas e encontros vocálicos para compor o banco de unidades. A

segunda etapa está relacionada à síntese por concatenação dos segmentos acústicos armazenados no banco de unidades juntamente com adaptações prosódicas de acordo com as informações recebidas do emissor do codificador.

1.4 Estrutura do Texto

Além deste capítulo introdutório que tem o objetivo de introduzir ao leitor conceitos de codificadores de voz e sua importância, bem como os atributos necessários ao desenvolvimento um codificador de voz para ser aplicado em telefonia móvel, a dissertação está dividida em mais seis capítulos. Apresenta-se a motivação para a realização deste trabalho, em que se descreve as características dos codificadores fonéticos e as diferenças entre os codificadores fonéticos encontrados na literatura e o proposto neste trabalho. Por fim, neste capítulo encontra-se o objetivo que se pretende alcançar, que é o desenvolvimento de um codificador a baixa taxa de *bits*.

O Capítulo 2 apresenta as principais técnicas de codificação de voz, utilizadas nos padrões de comunicações móveis. Descrevem-se os codificadores de forma de onda, paramétricos e híbridos, com ênfase no codificador CELP e suas derivações. Além disso, é apresentada a definição do codificador fonético, bem como uma revisão bibliográfica de vários deles encontrados na literatura. Por fim, esse capítulo aborda os codificadores padronizados pela ITU (*International Telecommunication Union*) e contém uma comparação entre as técnicas apresentadas.

No Capítulo 3 é apresentada a teoria dos Modelos de Markov Escondidos (HMM), voltada para o reconhecimento de fala. Inicialmente, apresenta-se a definição de HMM, bem como sua classificação em relação à distribuição de probabilidade e topologia. Em seguida, explica-se como o HMM está inserido nas etapas de treinamento e reconhecimento do sistema.

O Capítulo 4 descreve em que consiste um sistema de reconhecimento de fala, bem com as etapas necessárias ao seu desenvolvimento, que incluem processamento do sinal de voz, extração de características, construção do modelo acústico e decodificação.

O Capítulo 5 aborda o desenvolvimento do codificador de voz proposto neste trabalho. Inicialmente é feita uma descrição geral do sistema e, em seguida, descrevem-se as etapas realizas na sua implementação.

No Capítulo 6 são apresentados os resultados dos testes subjetivos realizado com o objetivo de avaliar o desempenho do codificador.

O Capítulo 7 apresenta as conclusões deste trabalho, bem como as principais contribuições obtidas.

O Anexo A tem uma sucinta descrição da ferramenta HTK (*Hidden Markov Models Toolkit*) utilizada para construir o sistema de reconhecimento de fala desenvolvido neste trabalho. O Anexo B apresenta os segmentos fonéticos do Português Brasileiro, enquanto os Anexos C e D apresentam respectivamente as frases utilizadas no desenvolvimento do codificador e as publicações obtidas durante a realização desta dissertação.

CAPÍTULO 2

Técnicas de Codificação de Voz

O desenvolvimento de técnicas avançadas de codificação de voz tornou possível e viável a introdução dos sistemas digitais de telefonia móvel. Essas técnicas têm o objetivo de representar o sinal de fala com uma menor taxa de *bits*, diminuindo a largura de banda requerida por usuário. Desta forma, é possível aumentar o número de usuários do sistema, tornando a telefonia celular um sistema mais barato e acessível à população.

No início da pesquisa na área de codificação de voz, há cerca de sessenta anos, os pesquisadores tinham como objetivos o desenvolvimento de um sistema que possibilitasse a transmissão da voz por meio da estreita largura de banda dos cabos telegráficos. Homer Dudley, do *Bell Telephone Laboratories*, criou o primeiro método de codificação e demostrou a redundância existente no sinal de voz [6].

Após a invenção desse primeiro codificador, que analisava a voz em termos da sua frequência fundamental e do seu espectro, novos codificadores foram desenvolvidos apresentando melhor qualidade do sinal sintetizado e menor taxa de transmissão.

Como principais codificadores, é possível mencionar os de forma de onda, cujo principal é o PCM (utilizado, por exemplo, no padrão G.711) e suas variantes, utilizados em praticamente todos os sistemas de telefonia fixa do mundo. Outro tipo de codificador bem difundido é o codificador paramétrico, como o codificador LPC, que apresenta baixa qualidade do sinal final mas utiliza uma pequena largura de banda de transmissão. Além desses, há os codificadores híbridos, dentre os quais se destaca o CELP utilizado, por exemplo, no padrão G.729, e suas variantes.

Este capítulo descreve as principais técnicas de codificação de voz. São apresentadas sucintamente os conceitos sobre os codificadores de forma de onda, paramétricos e híbridos, com ênfase nos principais codificadores derivados dessas técnicas. Além disso, descrevem-se alguns codificadores fonéticos já desenvolvidos e os codificadores padronizados pela ITU (*International Telecommunication Union*). Por fim, é apresentada uma comparação entre as técnicas apresentadas.

2.1 Codificadores de Forma de Onda

Os codificadores de forma de onda têm como principal característica reproduzir a forma de onda, amostra por amostra, ou utilizando um vetor de amostras (quantização vetorial), de maneira

mais eficiente possível. São codificadores de pequena complexidade de implementação e geralmente possuem um baixo retardo de voz.

2.1.1 Codificação: PCM, APCM, DPCM e ADPCM

O codificador de forma de onda PCM está definido na CCITT G.711 e AT&T 43801. Nesta técnica, o sinal de voz é amostrado a uma taxa de 8000 amostras por segundo, que é a frequência de amostragem adotada internacionalmente para aplicação em telefonia fixa.

Cada amostra do sinal é codificada como uma sequência de *bits* que traz a informação do valor quantizado da amplitude da amostra em questão. É possível conseguir uma boa qualidade de voz com 256 níveis distintos, em que cada nível corresponde a um código de 8 *bits*, implicando uma taxa de *bits* de 64 k$bit/$s e uma banda passante de aproximadamente de 64 kHz.

Essa taxa é considerada alta para um sistema de telefonia móvel, em que a banda utilizada por cada usuário deve ser a menor possível. Portanto, são necessárias técnicas de codificação adicionais que diminuam a taxa de *bits* para valores mais adequados ao sistema.

Algumas técnicas adicionais podem ser utilizadas para diminuir a taxa de *bits* nos codificadores de forma de onda. Entre elas convêm destacar o PCM Adaptativo (APCM), o PCM Diferencial (DPCM), o PCM Adaptativo e Diferencial (ADPCM) e a Modulação Delta.

A codificação diferencial, PCM Diferencial, explora o fato de o sinal de voz apresentar correlação significativa entre amostras sucessivas, e tem como objetivo a redução da redundância do sinal de voz, quantizando a diferença de amplitude entre as amostras adjacentes que, por ser relativamente pequena pode ser representada com um número menor de *bits*. Utilizando esse sistema de codificação é possível reduzir a taxa de transmissão para 56 k$bits/$s com a mesma qualidade de um PCM.

Na codificação APCM, o passo de quantização varia com o tempo, de modo a acompanhar as variações de amplitude do sinal de voz, baseando-se em amostras passadas. Assim reduz-se a faixa dinâmica (variação de amplitude) do sinal e consequentemente a taxa final de transmissão.

Os codificadores ADPCM empregam quantização ou predição adaptativas. A predição adaptativa consiste no ajuste dinâmico do preditor de acordo com variações no sinal de voz. Assim codificadores ADPCM apresentam boa qualidade de voz para taxas entre 24 e 48 k$bit/$s [7].

A Modulação Delta é um caso especial da sistema de codificação DPCM, no qual a variação de amplitude de amostra a amostra é quantizada, usando apenas dois níveis de quantização. A saída do quantizador de dois níveis é relacionada com a entrada pela expressão $y = 2du(x) - d$, em que $u(\cdot)$ é a função degrau e d representa o passo de quantização.

2.2 Codificadores Paramétricos

Os codificadores paramétricos, também conhecidos como *vocoders*, apresentam taxa de *bits* menor que 4,8 k$bit/$s, mas o atraso e a complexidade são elevados e a voz soa sintética. Assim, não fornecem a qualidade de voz requerida para a rede telefônica, sendo mais utilizados em aplicações com fins militares [2].

A ideia básica para análise por predição linear é que qualquer amostra do sinal de fala pode ser aproximada por uma combinação linear das amostras anteriores. A minimização da soma das diferenças quadradas entre a amostra de fala atual e a predita linearmente, em um intervalo finito, permite que um único conjunto de coeficientes de predição possa ser determinado. Os coeficientes de predição são os pesos dos coeficientes usados na combinação linear.

O princípio da predição linear está relacionado com o modelo de fala, ilustrado na Figura 2.1, definido como a saída de um sistema linear variante no tempo excitado por impulsos quase periódicos, durante a fala vocalizada, ou ruído aleatório, na fala não vocalizada. O método de predição linear é robusto, fiável e preciso para estimação dos parâmetros que caracterizam o sistemas linear variante no tempo [8, 9].

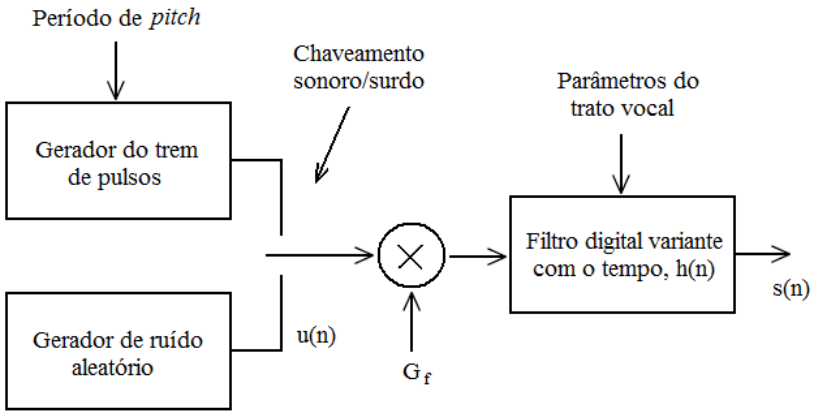

Figura 2.1: Diagrama de blocos para o modelo simplificado de produção de voz [10].

Os coeficientes do preditor devem ser estimados em segmentos de curtos intervalos de tempo, nos quais o sinal de voz pode ser considerado estacionário, e devem ser atualizados periodicamente, devido à natureza variante no tempo do sinal de voz, com o objetivo de se obter uma boa estimativa das propriedades espectrais do sinal de voz. Esses coeficientes podem ser obtidos por meio da análise LPC (*Linear Predictive Coding*), denominados coeficientes LPC, ou por meio de técnicas derivadas dessa análise. Entre os coeficientes utilizados, podem ser citados: coeficientes LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados.

2.2.1 LPC (*Linear Predictive Coding*)

O codificador LPC é o mais importante da família dos codificadores paramétricos. Ele tem a característica de extrair os parâmetros importantes do sinal de voz diretamente da forma de onda no tempo, obtendo um resultado melhor que outros tipos de *vocoders* que obtêm seus parâmetros a partir do espectro de frequência, como é o caso dos *vocoders* de canal e formantes [11].

Um codificador LPC analisa a forma de onda para produzir um modelo do trato vocal e a sua função de transferência. Trata-se de um modelo adaptativo, em que o codificador determina, periodicamente, um novo conjunto de parâmetros correspondentes aos segmentos sucessivos de

voz. Os parâmetros são: natureza da excitação (impulsos ou ruído), contagem do período para a excitação da voz, fator de ganho e coeficientes do preditor (parâmetros do trato vocal).

O sintetizador LPC, que recebe a atualização periódica dos parâmetros do modelo e das especificações de excitação, tem a função de reconstruir o sinal de voz pela passagem da excitação especificada por um modelo matemático do trato vocal, construído a partir do ganho e coeficientes do preditor. Deste modo, a saída do sinal de voz, dada pela Equação 2.1, é representada como a entrada atual do sistema, somada a uma combinação linear da saída predita do trato vocal.

$$y(n) = \sum_{k=1}^{p} a_k y(n - k) + G_f x(n), \tag{2.1}$$

em que, $y(n)$ é a n-ésima amostra de saída, a_k é o k-ésimo coeficiente do preditor, G_f é o fator de ganho, $x(n)$ é a entrada amostrada em um tempo n e p é a ordem do modelo. A maior parte dos LPCs faz a codificação de voz no intervalo de 1,4 a 2,4 kbit/s.

2.3 Codificadores Híbridos

Os codificadores hídridos combinam a qualidade dos codificadores de forma de onda com a eficiência dos codificadores paramétricos. Desta forma, extraem parâmetros do sinal de voz assim como os codificadores paramétricos, e ao mesmo tempo geram a excitação pelo formato de onda da mesma maneira que os codificadores de forma de onda.

Com auxílio de dicionários é possível determinar a melhor excitação, permitindo ao codificador híbrido obter uma boa qualidade do sinal reconstituído que, com taxa de transmissão entre 2 e 16 k$bits$/s, fornece uma qualidade superior à obtida com os codificadores de forma de onda, que apresentam taxas mais elevadas [12].

O melhoramento na qualidade da voz sintetizada deve-se tanto à quantização e codificação dos parâmetros que definem a excitação quanto aos parâmetros do filtro de síntese. Na obtenção dos parâmetros a curtos intervalos de tempo, os codificadores híbridos utilizam um processo conhecido como análise por síntese.

2.3.1 RPE-LTP (*Regular Pulse Excited–Long Term Predictor*)

O RPE-LTP é o codificador utilizado pelo padrão europeu de telefonia móvel, o GSM (*Global System for Mobile communications*). É um esquema híbrido baseado no LPC, em que a excitação consiste em pulsos regularmente espaçados e de amplitude variada e que possui um preditor a longo termo.

O codificador RPE-LTP modifica o codificador RELP (*Residual Excited Linear Predictive*) e incorpora algumas características do codificador MPE-LTP (*Multi-Pulse Excited–Long-Term Prediction*) para alcançar uma taxa de transmissão de 13 k$bits$/s.

A principal modificação proposta no RPE-LTP, consiste na adição de um esquema de predição a longo termo, denominada análise LTP (*Long-Term Prediction*). Além disso, por ter uma excitação com pulsos regularmente espaçados, o codificador RPE-LTP apenas determina a posição do primeiro

pulso e a amplitude de cada um deles, diferentemente do codificador MPE-LTP, em que parte da informação transmitida é referente à posição dos pulsos usados na excitação.

O codificador RPE-LTP é formado por quatro etapas principais de processamento [7]:

- **1º Etapa:** Inicialmente o sinal de voz é pré-enfatizado. Em seguida, é segmentado a curtos intervalos de 20 ms por meio do janelamento de Hamming.

- **2º Etapa:** Um filtro faz a análise STP (*Short-Term Prediction*) do sinal, encontrando os coeficientes preditores do filtro de síntese. Esses parâmetros são utilizados para construir o filtro LPC inverso, que determinará o erro de predição.

- **3º Etapa:** Uma análise LTP do erro de predição é utilizada para determinar o período de *pitch* e um fator de ganho que minimiza o erro de predição, maximizando a correlação cruzada de umas amostras sucessivas. O erro de predição é denominado LTP residual. Esse sinal é ponderado e decomposto em três possíveis sequências de excitação.

- **4º Etapa:** A sequência de maior energia é selecionada para representar o LTP residual, sendo normalizada em relação ao pulso de maior amplitude.

O decodificador RPE-LTP tem a função de decodificar a sequência de pulsos de excitação e processá-la por um filtro de síntese LTP, que faz o uso do *pitch* e do ganho para sintetizar o sinal de longo termo, que passa por um filtro de síntese STP construído a partir dos coeficientes preditores, para obter o sinal original.

2.3.2 CELP (*Code Excited Linear Prediction*)

O codificador CELP é o tipo híbrido mais utilizado. Ele reúne os mesmos princípios da análise LPC para redução de parâmetros a serem transmitidos, ou seja, é feita uma análise de correlação do sinal de voz de curto prazo, entretanto sua fonte de excitação não é do tipo trem de pulsos/ruído como no LPC, mas sim constituído de um esquema de quantização vetorial de dois estágios em paralelo.

Neste codificador, o sinal de voz é particionado em curtos segmentos de 20 *ms* por meio da janela de Hamming, a uma taxa de 8000 amostras/s, resultando em 160 amostras. Em seguida esse bloco é dividido em quatro sub-blocos de 5 *ms* cada.

Para manipular as excitações da entrada do sistema, esse codificador usa dicionários, tanto fixos quanto adaptativos. Como possui um número bem maior de excitações, em relação ao LPC, este codificador é capaz de reconstruir o sinal de voz, após a transmissão, com uma qualidade superior àquele gerado pelo LPC [11].

O dicionário é formado por um conjunto de vetores aleatórios (excitações), com distribuição gaussiana e média zero, com o objetivo de aproximar a estatística destes vetores com as características do sinal de voz de curto prazo. A escolha da sequência ideal a ser utilizada como excitação do filtro de síntese no decodificador é feita com a técnica análise por síntese. A Expressão 2.2 indica que o dicionário armazena k sequências $x(n)$.

$$C = \{[x_0(n)], [x_1(n)], ..., [x_{K-1}(n)]\}. \tag{2.2}$$

No caso do CELP geralmente são utilizados dois dicionários, um com as excitações fixas e o outro com excitações adaptativas. Esse codificador considera um bloco (ou sub-bloco) como tendo uma parte sonora e outra surda. O primeiro estágio consiste no dicionário adaptativo, que é constituído por k_1 sequências retiradas da excitação passada, e tem como objetivo estimar tanto a parte sonora quanto a surda com certa precisão, uma vez que se adapta às características do segmento a ser analisado. Os valores típicos para k_1 são 128 e 256.

O segundo estágio possui um dicionário fixo, constituído por k_2 sequências estocásticas, determinísticas, ou obtidas por meio de um procedimento de treinamento, e é responsável por estimar a parte surda que o adaptativo não conseguiu. Essa estimativa é feita a partir da busca da excitação armazenada no dicionário que gera a resposta mais próxima do sinal a ser estimado. Valores usuais de k_2 são 128, 256, 512 e 1024 [7].

O dicionário adaptativo tem as suas sequências atualizadas a cada bloco de 5 ms, o que é realizado por meio de uma malha de alimentação. Essas atualizações se dão baseadas na soma das melhores excitações dos dicionários fixo e adaptativo, que têm seu conteúdo mais antigo descartado, estando o dicionário inicialmente zerado por convenção.

Deste modo, a sequência de excitação com a qual se deseja reproduzir o segmento de voz é obtida como uma combinação linear de duas sequências de cada dicionário, que são escolhidas por um procedimento em que são testadas diversas excitações possíveis para sintetizar o segmento de voz corrente, escolhendo-se aquela que minimiza a medida de erro ponderado na saída do filtro de ponderação.

O filtro de ponderação ou perceptivo é utilizado no codificador CELP com o objetivo de enfatizar as componentes de mais baixas frequências e menor amplitude, realizando o oposto com as componentes de maior amplitude. Isso se deve a uma das características do ouvido humano que apresenta maior sensibilidade a erros e ruídos de mais baixas amplitudes comparado às componentes de alta amplitude, no domínio da frequência. O filtro é denotado por $W(z)$, e sua função de transferência é dada por [11, 13, 14]

$$W(z) = \frac{A(z)}{A\left(\frac{z}{\gamma}\right)}. \tag{2.3}$$

em que γ representa o fator de percepção, que possui valor típico de 0,8.

Os ganhos, tanto do dicionário fixo quanto do adaptativo, são obtidos por [13]

$$G_f = \frac{R(s,d)}{R(d,d)}, \tag{2.4}$$

em que $R(s,d)$ representa a correlação entre o sinal de voz a ser analisado, também chamado sinal alvo e as respostas correspondentes a cada excitação contida no dicionário em questão. A $R(d,d)$ é a autocorrelação entre as respostas correspondentes a cada excitação contida no dicionário em questão.

Portanto, o que é enviado ao decodificador não é a sequência de excitação em si, mas os ganhos e índices que identificam as duas sequências nos dicionários, que também existem no decodificador, implicando assim uma boa qualidade de voz a uma baixa taxa de *bits*.

A maioria das técnicas de codificação de voz aplicadas na telefonia celular deriva do CELP, se distinguindo deste basicamente por modificações nos dicionários de código que dão origem à excitação.

2.3.3 VSELP (*Vector Sum Excited Linear Predictive*)

O codificador VSELP é baseado no codificador CELP, com aprimoramentos em relação à alta qualidade de voz, complexidade computacional e robustez a erros de canal. Este codificador é utilizado nos sistemas TDMA IS-54B e IS-136 de telefonia móvel celular.

Para alcançar os objetivos, o algoritmo do codificador VSELP faz o uso do dicionário adaptativo, assim como no codificador CELP, entretanto, substitui o dicionário fixo por dois dicionários de código organizados com uma estrutura pré-definida. Esse processo resulta na redução do tempo necessário na busca da palavra-código ótima, além de trazer alta qualidade de voz e uma maior robustez a erros de canal, mantendo baixa complexidade, a uma taxa de 7,95 kbit/s [7].

O VSELP tem sua sequência de excitação formada pela combinação linear de três sequências, obtidas de cada dicionário.

Da mesma forma que no codificador CELP, o dicionário adaptativo do codificador VSELP é responsável pela predição a longo termo, introduzindo a periodicidade encontrada no sinal de voz (*pitch*). Os dicionários fixos, formados a partir de combinações lineares de sete vetores base, possuem 128 vetores de 40 amostras cada.

Após a determinação do atraso ótimo do dicionário adaptativo, é realizada a busca sequencial nos dicionários fixos. Primeiro determina-se a melhor sequência do primeiro dicionário estruturado, levando em conta a sequência já escolhida do dicionário adaptativo. Em seguida, determina-se a melhor sequência do segundo dicionário estruturado, considerando agora as duas sequências previamente escolhidas.

A informação transmitida ao decodificador consiste no atraso do dicionário adaptativo, nos índices dos vetores dos dicionários fixos, nos respectivos ganhos e nos parâmetros LPC do filtro de sintese. Deste modo, o receptor gera a excitação fazendo uma combinação linear das três sequências de excitação especificadas e a utiliza como sinal de entrada no filtro LPC, que introduz a correlação a curto termo, gerando a voz sintetizada [15].

2.3.4 ACELP (*Algebraic Code Excited Linear Predictive*)

O codificador ACELP foi padronizado ITU-T SG15 na recomendação G.723 e é o codificador de voz utilizado pelo padrão IS-136 de telefonia móvel.

O ACELP proporciona qualidade de voz igual ou superior ao codificador ADPCM, cuja taxa de transmissão é de 32 k*bit*/s, e transmite voz com melhor qualidade e de forma menos sensível ao ruído que o VSELP, à mesma taxa de 7,95 k*bit*/s [7].

Baseado no codificador CELP, a diferença básica do codificador ACELP está no dicionário fixo, que neste caso usa uma estrutura algébrica. Essa estrutura proporciona ao codificador sensibilidade reduzida a erros no canal, além de voz codificada de melhor qualidade.

Cada vetor de palavra código do dicionário fixo estruturado algebricamente contém quatro pulsos diferentes de zero, assumindo amplitudes de -1 ou 1. Pela natureza estruturada do dicionário, é possível realizar uma busca mais eficiente, sendo feita por quatro laços em série.

No caso do dicionário adaptativo, a busca é realizada apenas ao redor de uma região limitada por um período de *pitch*, estimado a cada 10 *ms* por duas análises diferentes: *open-loop* e *closed-loop*. Com esse parâmetro, seleciona-se a melhor sequência com uma resolução de 1/3 para o atraso de *pitch* [16].

Assim, por meio da combinação linear das sequências obtidas nos dicionários fixo e adaptativo, é obtida a sequência de excitação. O codificador ACELP também faz o uso da técnica análise por síntese, porém de forma mais eficiente que o codificador CELP.

2.3.5 QCELP (*Qualcomm Code Excited Linear Predictive*)

O codificador de voz QCELP, desenvolvido pela *Qualcomm*, é utilizado pelo sistema IS-95 de telefonia móvel CDMA. Possui a característica de proporcionar taxa de *bits* variável: 9,6/4,8/2,4/1,2 k*bit*/s.

O QCELP se diferencia do codificador CELP na forma como a correlação a longo termo é codificada. Em seu algoritmo, a busca do período de *pitch*, que modela a correlação a longo termo, utiliza dois processos, denominados *open-loop* e *closed-loop*, em contrapartida ao uso de dicionário adaptativo. No entanto, apresenta qualidade do sinal de voz inferior aos demais codificadores, mesmo à taxa máxima [7].

Em seu algoritmo, o sinal de voz passa por um filtro passa-altas que tem a função de retirar as componentes DC do sinal. Em seguida, o sinal é particionado por meio da janela de Hamming.

Após o janelamento, determinam-se os parâmetros do filtro de síntese por meio de uma análise LPC. Paralelamente, é realizado um procedimento para determinação da taxa de dados, feito por meio da análise das características do quadro para decidir se o sinal pode ser codificado a uma taxa reduzida sem afetar a qualidade de voz.

No processo *open-loop* retira-se a correlação a curto termo do sinal de voz ao passá-lo pelo filtro LPC inverso. A seguir, o sinal residual entra no filtro preditor de *pitch*, que tenta retirar a correlação a longo termo, produzindo o *pitch* residual. O filtro tem dois parâmetros, ganho de *pitch* e atraso de *pitch*, que devem ser otimizados para que a energia média do *pitch* residual seja minimizada, resultando em um sinal cujas características lembram as de um ruído branco.

Para reconstrução do sinal de voz, o QCELP utiliza o processo *closed-loop* que realiza o processo inverso ao *open-loop*. Usando ruído branco como fonte de excitação, passando por um filtro de síntese de *pitch* e na sequência por um filtro de síntese de formante, que introduzem respectivamente a correlação a longo e a curto termos, resultando em o que se espera ser, voz sintetizada. A

busca pelos parâmetros ganho de *pitch* e atraso de *pitch* é feita de forma a minimizar a diferença entre o sinal original e o sinal sintetizado obtido a partir de uma excitação que se aproxima de ruído branco. Essa excitação é obtida a partir de um dicionário pseudoaleatório ou de um dicionário gaussiano, para taxas de 1/4 e 1/8 ou 1 e 1/2, respectivamente [7].

A informação cedida ao decodificador consiste nos coeficientes do filtro LPC, nos parâmetros do filtro de síntese de *pitch* (ganho e atraso de *pitch*), e nos parâmetros da excitação. Com essa informação, o decodificador monta os filtros de síntese e utiliza a excitação especificada para sintetizar o sinal de voz.

2.3.6 AMR-NB (*Adaptative Multi-Rate Narrowband*)

O codificador AMR-NB é da família CELP e foi adotado na fase 2+ do GSM. Este codificador opera em oito diferentes taxas de codificação, que variam entre 4,75 e 12,2 kbit/s.

O AMR-NB tem como característica a utilização de detector de atividade vocal, que tem a função de decidir se o quadro transmitido é composto por amostras do sinal de voz ou por silêncio, com base na energia do sinal amostrado. Os trechos de silêncio são codificados a uma taxa denominada SID (*Silence Descriptor*), que reproduz as características do silêncio, produzindo o chamado ruído de conforto. Além disso, este codificador possui mecanismo de substituição e silenciamente de quadros perdidos que diminui os efeitos da perda de pacotes na rede.

A substituição tem o objetivo de atenuar e ocultar os efeitos dos quadros perdidos, enquanto a finalidade de silenciar a saída, no caso de muitos quadros perdidos, é indicar a interrupção do canal ao usuário e evitar a geração de possíveis sons inoportunos com um resultado do procedimento de substituição de quadros.

Sendo assim, os quadros de fala perdidos são substituídos tanto por repetição quanto por extrapolação de bons quadros de fala anteriores para aumentar a qualidade do sinal de voz processado por este codificador.

O codificador AMR-NB realiza a amostragem do sinal de fala a uma taxa de 8 k amostras/s e gera quadros de 20 ms, ou seja, cada quadro com 160 amostras. Cada quadro de 20 ms produz, 95, 103, 118, 134, 148, 159, 204 ou 244 *bits* de informação dependendo da taxa de codificação utilizada. Os *bits* são então separados por categoria e protegidos na etapa de codificação de canal de acordo com a importância que lhe for atribuída.

2.3.7 AMR-WB (*Adaptive Multirate Wideband*)

O codificador AMR-WB foi selecionado em dezembro de 2000 pelo 3GPP (*3rd Generation Partnership Project*) para os padrões GSM e WCDMA (*Wideband Code Division Multiple Acess*) de terceira geração de comunicações móveis, para prestação de serviços de voz de banda larga, e suas especificações foram aprovadas em março de 2001. O ITU-T (*International Telecommunication Union-Telecommunication Sector*), em julho de 2001, também selecionou este codificador para codificação de fala de banda larga em torno de 16 k$bits$/s e foi aprovado como recomendação G.722.2 em janeiro de 2002. A adoção do codificador AMR-WB pela ITU-T é importante pois permite que

o mesmo codificador seja adotado para comunicação sem fio, bem como em serviços de telefonia fixa [17].

O AMR-WB tem seu algoritmo baseado no codificador CELP e fornece nove taxas de codificação, como taxas de 23,85, 23,05, 19,85, 18,25, 15,85, 14,25, 12,65, 8,85 e 6,6 k*bit*/s [18].

Similarmente ao AMR-NB, o codificador AMR-WB também utiliza o sistema de detecção de atividade vocal, em que se detecta se o quadro é composto por voz ou silêncio de acordo com a energia do sinal. Os trechos do silêncio na fala do orador são codificados a uma taxa denominada SID (*Silence Descriptor*), que reproduz as características do silêncio, produzindo um ruído de conforto. Além disso, possui mecanismo de substituição e silenciamento de quadros perdidos que diminui os efeitos da perda de pacotes na rede.

No entanto, após a geração da fala, esse codificador faz a amostragem do sinal a uma taxa de 16 k amostas/s para gerar quadros de 20 ms, em que cada um produz 132, 177, 253, 285, 317, 365, 397, 461 ou 477 *bits* de informação dependendo da taxa de codificação utilizada. Após a codificação, os *bits* são separados em duas categorias (banda de frequência, 50-6400 Hz e 6400-7000 Hz) e protegidos conforme sua importância.

Este codificador usa transmissão descontínua (DTX – *Discontinuous Transmission*) para o padrão GSM e taxa controlada na fonte (SCR – *Source Controlled Rate*) para os sistemas 3G. Em GSM a taxa de *bits* desse modo é de 1,75 k*bit*/s. O modo 12,65 k*bit*/s e os modos acima desse oferecem alta qualidade de voz banda larga. Os dois modos de menor qualidade, 8,85 e 6,6 k*bit*/s devem ser usados somente temporariamente durante condições severas de canal ou durante congestionamentos de redes [18].

2.4 Codificadores Fonéticos

Os codificadores fonéticos são casos particulares dos *vocoders* segmentais, em que os segmentos são fonéticos, difones ou sílabas. Nesses codificadores, a segmentação do sinal de entrada é obtida por meio de técnicas de reconhecimento de sinais de fala, como HMM, ou por técnicas de segmentação, como programação dinâmica, utilizadas quando os segmentos são definidos automaticamente.

Para alcançar o objetivo da baixa taxa de *bits*, os codificadores fonéticos são caracterizados por, em vez de transmitir amostras do sinal de fala, realizarem a quantização em parâmetros que correspondem a segmentos de voz, com os indices dos segmentos fonéticos reconhecidos e informações prosódicas da duração, energia e frequência fundamental.

Os codificores fonéticos têm em seus receptores um livro de códigos que contém informação espectral baseada em coeficientes LPC, utilizada juntamente com a informação prosódica para sintetizar o sinal de saída. Por natureza, resulta da codificação binária uma taxa de *bits* variável em função do número de unidades fonéticas produzidas por segundo para cada orador.

Os codificadores fonéticos permitem a codificação dos sinais de fala com taxa de *bits* extremamente baixas. Em [19], o sinal de voz é codificado com um total aproximadamente de 100 *bit*/s, sendo 60 a 75 *bit*/s utilizados na transmissão da sequência fonética e os demais para transmitir o valor da duração e de uma frequência fundamental por segmento fonético. O sinal usa

síntese LPC. As palavras de códigos são difones e possuem informação sobre a envolvente espectral e ganho. No entanto, esse codificador tem a desvantagem do treino do HMM ter sido produzido por apenas um orador.

Em [20] é proposto um codificador fonético dependende de orador e que utiliza trifones para a síntese. Esse codificador utiliza quarenta e sete segmentos fonéticos e gera dez mil quatrocentos e trinta e sete trifones, obtidos em quinze minutos de fala. A sequência fonética é codificada com 55 *bit*/s, no entanto, os parâmetros prosódicos não foram quantificados. A síntese final é produzida por um filtro LPC.

Em [21] é utilizado um codificador fonético baseado em sílabas. O treino do livro de código ou dicionário é realizado com um orador e a segmentação é feita manualmente. Considerando seis sílabas por segundo, a taxa de *bits* que este codificador oferece é 100 *bit*/s, em que a energia e duração são codificadas com dois *bits* cada, enquanto a frequência fundamental é codificada com três *bits*. Em uma base de quatrocentas e dezessete palavras, foram obtidos quinhentos segmentos fonéticos, referenciando diferentes contextos. O idioma utilizado foi o japonês, no qual existem cerca de cem sílabas.

É proposto em [22] um codificador fonético que utiliza a técnica HMM para dividir o sinal em segmentos fonéticos, produzindo uma média de 12,3 segmentos por segundo. Com o objetivo de tornar o codificador menos dependente do orador, são utilizados 2, 4 ou 8 modelos por segmento fonético, sendo treinados com vozes masculinas e femininas e codificados respectivamente com 7, 8 ou 9 *bits* por segmento. A taxa de reconhecimento foi de 35% e independe do número de modelos por segmentos fonéticos. Apresenta qualidade ruidosa quando se utiliza dois modelos por segmento fonético. Este codificador resulta em uma taxa de 170 *bit*/s.

O codificador proposto em [23] oferece taxa de *bits* de 300 *bit*/s e é independente do orador. Seu projeto inclui a minimização da distorção introduzida pelo ajuste temporal do segmento de código à duração de entrada. Para isso, são propostos livros de códigos cujos segmentos possuem quatro durações distintas. O livro de códigos é treinado com trezentos e noventa e dois oradores, utilizando o *corpus* TIMIT.

Em [24] é proposto um codificador fonético que utiliza trinta e quatro modelos de HMM correspondentes aos fonemas, treinados com apenas um orador. O sinal é sintetizado com os coeficiente *mel-cepstrais* deste orador. A sequência fonética é codificada com 54 *bit*/s. Utilizando modelos bigramas esta taxa cai para 46 *bit*/s. A taxa de *bits*, excluindo a decisão de vozeamento e da frequência fundamental é de cerca de 150 *bit*/s. Para torná-lo independente do orador, os autores adaptam os modelos de HMM aos oradores de entrada, movendo as médias das distribuições de probabilidade de observações, de modo a maximizar as probabilidades de observação de entrada. Essa adaptação acrescenta 100 *bits* na taxa de *bits* total.

O codificador fonético proposto em [25] realiza no emissor a segmentação fonética por meio da técnica HMM, enquanto o receptor utiliza a mesma técnica para extrair parâmetros e construir uma excitação mista para realizar a sintese do sinal de voz. Este codificador produz uma taxa de 265 *bit*/s e consegue bons níveis de inteligibilidade.

O codificador descrito em [26] foi desenvolvido utilizando o procedimento de segmentação baseado em um reconhecedor de segmentos fonéticos, obtendo 52 segmentos fonéticos. Esse codi-

ficador foi desenvolvido em duas etapas, sem e com adapatação ao orador. Na primeira etapa, ou seja, sem adaptação ao orador, a informação transmitida ao receptor inclui o índice do segmento fonético reconhecido e as respectivas informações prosódicas como duração, energia e frequência fundamental e decisão de vozeamento. Na segunda etapa, a adaptação ao orador é inclusa no codificador, embora com um aumento na taxa final de *bits*.

Neste codificador, a atualização dos índices que referência os segmentos fonéticos e as respectivas durações são obtidas uma vez por cada segmento, ao passo que a decisão de vozeamento, o valor da frequência fundamental e da energia são atualizadas 44,4 vezes por segundo, obtendo uma taxa de *bits* média de 443 bit/s. Como a taxa de segmentos fonéticos produzidos por cada orador é variável, o codificador oferece taxa mínima, média e máxima respectivamente de 255, 443 e 524 bit/s.

O receptor tem a função de produzir a transformação da informação fonética em informação paramétrica trama a trama, sendo possível sintetizar o sinal acústico de saída. Para isso, faz uso de blocos de identificação de gênero, seleção da palavra de código, ajuste da duração e interpolação entre segmentos, sendo o sinal sintetizado por meio de um *vocoder* LPC ou um sintetizador harmônico.

A etapa de adptação ao orador, cujo objetivo é aumentar a reconhecimento ao orador, baseia-se na transmissão de informações específicas do orador, que minimizam o erro quadrático médio entre a matriz correspondente ao segmento fonético de entrada e a respectiva matriz da palavra de código. Essa adaptação é realizada apenas nas vogais ou *glides* e acrescenta 116 $bits$/s na taxa de *bits* total, resultando em um codificador com uma taxa média de *bits* de 559 $bits$/s.

2.5 Os Codificadores Padronizados pela ITU.

A ITU (*International Telecommunication Union*) é a parte da UNESCO (*United Nations Economic, Scientific and Cultural Organization*) responsável por estabelecer padrões globais de telecomunicações. Originalmente a ITU era composto pelo CCITT, responsável por estabelecer padrões de telecomunicações, incluindo padrões de codificação de voz, e pelo CCIR, cuja função era estabelecer padrões de rádio. No entanto, em 1993, o ITU foi reorganizado e as organizações CCITT tornaram-se parte do ITU-T (*ITU Telecommunications Standards Sector*), no qual o SG15 (*Study Group* 15) é responsável por formular padrões de codificação de voz.

No decorrer deste capítulo, foram apresentados os tipos de codificação mais utilizados. Para cada um dos tipos, o ITU desenvolveu uma série de recomendações.

2.5.1 G.711

O G.711 foi aprovado em 1972. Ele consiste na codificação PCM e apresenta uma taxa de codificação de 64 k$bits$/s. Há duas variações deste codificador. O PCM adotado no Japão e nos Estados Unidos, utiliza a lei μ na quantização do sinal, enquanto o PCM utilizado no resto do mundo utiliza a lei A no processo de quantização. Ambos os codificadores utilizam oito *bits* na representação do sinal e têm uma relação sinal-ruído efetiva de 35 dB [11, 27].

2.5.2 G.721, G.723, G.726 e G.727 ADPCM

Antes da padronização do G.721, o CCITT padronizou o G.711 32 kbits/s ADPCM, para ser utilizado em equipamentos de mutiplexação de circuitos digitais. Além disso, os *links* frequentemente encontravam o problema de ter o PCM lei μ em um ponto e lei A em outro.

O G.721, padronizado em 1986, foi desenvolvido para aceitar tanto PCM com lei μ quanto PCM com lei A como entradas. Entretanto, não aceita PCM linear como entrada. O G.721 32 k*bits*/s ADPCM foi selecionado para uso em padrões DECT (*Digital European Cordless Telephone*) e CT2 (*Cordless Telephone* II), que utilizam formas de TDMA como esquemas de acesso.

O G.723, padronizado em 1988 e também baseado no ADPCM, foi desenvolvido apenas para aplicações DCME (*Digital Circuit Multiplication Equipment*) e oferece duas taxas de *bits* adicionais: 24 e 40 k*bits*/s [2].

Também baseado no codificador ADPCM, o G.726 [28] é uma recomendação que data de dezembro de 1990. Consiste na união do G.721 e do G.723. Trata-se do codificador com base em forma de onda padronizado mais comum, executando o PCM Lei-A ou Lei-μ ITU-T G.711. Foi concebido para realizar a codificação ADPCM com diversas taxas de codificação, dependendo da aplicação a ser utilizada. Podem ser utilizados de 2 a 5 *bits* para a quantização, sendo que sempre 1 *bit* é reservado para o sinal da amplitude. Desta maneira, taxas de 16 a 40 k*bits*/s podem ser utilizadas.

Semelhante ao G.726, o G.727 inclui as mesmas taxas, mas os quantizadores têm um número par de níveis.

2.5.3 G.728

O programa de trabalho do G.728 [29] foi iniciado pelo CCITT em 1998. Utiliza como tipo de codificação a LD-CELP (*Low Delay Code Excited Linear Prediction*), que é uma variação do sistema CELP original. Neste codificador, os vetores de excitação têm um tamanho de apenas cinco amostras, que correspondem a uma sequência de excitação de 0,625 ms de duração e uma taxa de amostragem de 8000 amostras por segundo, apresentando uma taxa de *bits* de 16 k*bit*/s.

Seu uso ficou restrito à rede telefônica com cobrança. Apresenta um desempenho robusto a sinais com ruído de fundo ou música e erros de *bit* aleatórios, comparados ao G.711 e G.721.

2.5.4 G.729

O G.729 [30] é uma recomendação posterior ao G.728 e também utiliza uma variante do CELP tradicional. Aprovado em 19 de março de 1996, a recomendação ITU G.729 CS-ACELP apresenta uma taxa de transmissão duas vezes menor que o G.728, ou seja, 8 k*bit*/s.

Foi concebido para transmitir sinais de voz com qualidade em ambientes em que baixas taxas de codificação são importantes, como aplicações de comunicação sem fio e circuitos transoceânicos.

Há alguns requisitos para o G.729, como, qualidade não inferior à apresentada pelo G.726 32 k*bits*/s, taxa de erro de *bit* menor que 10^{-3}, a dependência do falante não inferior ao G.726

32 k*bits*/s. Deve ter o objetivo de transmitir música sem nenhum efeito incômodo e capacidade de transmitir tons de sinalização e informação com a menor distorção possível.

O codificador G.729 é muito eficiente em relação à taxa de codificação, entretanto seus requisitos computacionais são elevados. O anexo A da recomendação mantém a operabilidade com o G.729 e reduz sua complexidade computacional, sendo as principais alterações referentes à forma de busca nos dicionários e à forma de operação de cada um dos filtros. O anexo B do G.729 descreve o gerador de ruído de conforto e o detector de voz, utilizados na implementação da compressão de silêncio.

2.5.5 G.723.1

O codificador G.723.1, aprovado em março de 1996, especifica uma determinada codificação a ser utilizada para compressão de sinal de voz ou sinal de áudio de um serviço multimídia qualquer para meios de baixa taxa de transmissão. Os dois tipos de codificadores utilizados são o ACELP com taxa de 5,3 k*bit*/s e o MP-MLQ (Multi-Pulse Maximum Likelihood Quantization) com taxa de 6,3 k*bit*/s.

Este codificador foi selecionado para se tornar o codificador de voz padrão para voz IP. Tem a característica de utilizar detecção de atividade de voz, transmissão descontínua e geração de ruído de conforto.

2.6 Comparação Entre as Técnicas

A eficiência de um codificador de voz pode ser verificada a partir da sua taxa de transmissão, ou taxa de *bits*, e pela qualidade da voz reconstruída. A Figura 2.2 faz uma comparação entre os codificadores de forma de onda, paramétricos e híbridos, com relação aos dois parâmetros.

De modo geral, os codificadores de forma de onda apresentam uma boa qualidade do sinal, entretanto trabalham a uma taxa de transmissão elevada. Por outro lado, os codificadores paramétricos oferecem taxas de *bits* reduzidas, no entanto, geram o sinal de voz com pouca qualidade. Por fim, os codificadores híbridos (G.723.1-ACELP, RPE-LTP (GSM), G.729.1-CS-CELP e G.728L-D-CELP) representam um compromisso entre a taxa de transmissão e qualidade de codificação quando comparados aos codificadores de forma de onda (G.726-ADPCM e G.711-PCM) e paramétricos (LPC).

A Tabela 2.1 resume as técnicas de codificação híbrida aplicadas aos padrões de telefonia celular e suas respectivas taxas de *bits* e MOS (*Mean Opinion Score*). Todas as técnicas são derivadas do codificador paramétrico LPC, com o modelo do trato vocal representado por um filtro que introduz a correlação a curto termo presente no sinal de voz.

A diferença básica entre estas técnicas, que implica em diferentes taxas de *bits* e diferentes qualidade de voz, está na forma como é gerada a excitação, responsável por introduzir a correlação a longo termo e descrever as sucintas diferenças percebidas entre períodos tonais sucessivos do sinal a ser codificado.

O codificador RPE-LTP utiliza um filtro LTP para obter a correlação a longo termo e as diferenças percebidas entre períodos tonais são descritas pela sequência de exitação residual enviada ao

codificador. Nos codificadores VSELP e ACELP o filtro LTP é substituído por um dicionário adaptativo e a sequência residual é obtida a partir de dicionários de código estruturados. No QCELP não há dicionário adaptativo, mas um filtro de síntese de *pitch*. A sequência residual, no entanto, é obtida a partir de dicionários fixos como nos demais codificadores derivados do CELP.

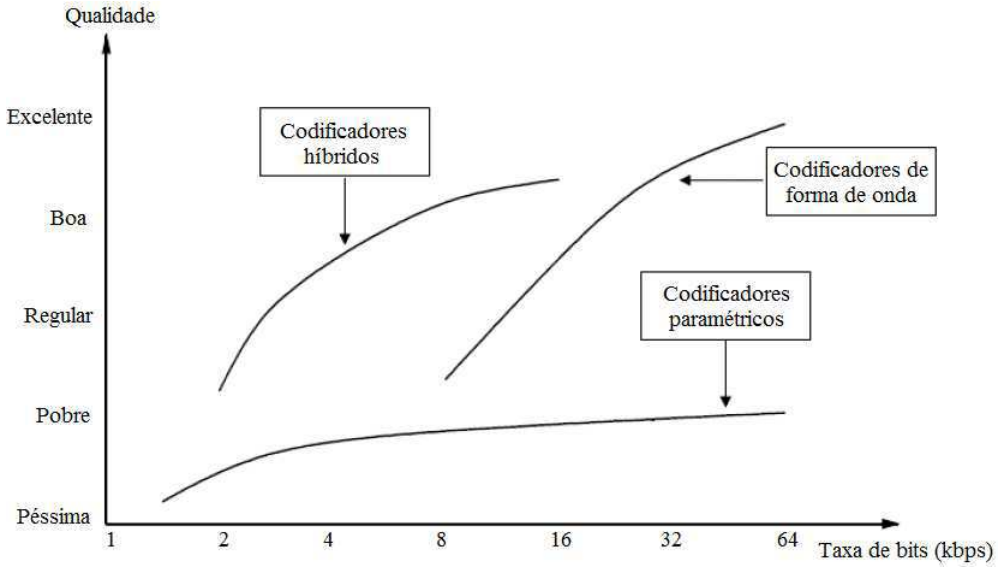

Figura 2.2: Comportamento dos codificadores com relação à taxa de codificação e qualidade [2].

Tabela 2.1: Sistemas móveis, técnicas de codificação aplicadas, suas respectivas taxa de *bits* e MOS [7].

Padrão	**Técnica**	**Taxa de *Bits***	**MOS**
GSM (TDMA/FDMA)	RPE-LTP	13 k*bit*/s	3,8
IS-136/54B (TDMA)	VSELP/ACELP	7,95 k*bit*/s	3,8/3,9
IS-95 (cdmaone)	QCELP	1,2/2,4/4,8/9,6 k*bit*/s	3,45 (Taxa máxima)
WCDMA	AMR-WB	6,6 - 23,85 k*bit*/s	4,14

A Tabela 2.2 apresenta os codificadores especificados pela ITU, com suas respectivas taxas de *bits* e MOS.

Tabela 2.2: Taxa de codificação e pontuação MOS dos codificadores apresentados [11, 31, 6].

Codificador	**Taxa de Codificação (k*bits*/s)**	**MOS**
CELP	8	2,8
G.711	64	4,3
G.726	40/32/24/16	3,8 (32 k*bit*/s)
G.728	16	4,0
G.729	8	3,9
G.723.1	5,3/6,3	3,8

Os codificadores que apresentam menor taxa de *bits* são os fonéticos. Contudo, muitos desses codificadores não apresentam mecanismo de caracterização de orador, como os codificadores apresentados em [19] e [21], que oferecem taxa de *bits* mais baixas (100 *bits*/s). A Tabela 2.3 resume alguns codificadores fonéticos comentados e suas respectivas taxa de *bits*.

Tabela 2.3: Codificadores Fonéticos

Referência	Taxa Total (*bits*/s)
Schwartz (80)	100
Hirata (89)	100
Peterson (90)	300
Ribeiro- (99)	270
Ribeiro+ (99)	560

2.7 Considerações Finais

Neste captítulo foram abordados os conceitos dos principais codificadores de voz. Incialmente foram apresentados os codificadores de forma de onda e suas principais variantes PCM, APCM, DPCM, ADPCM. Em seguida, foram descritos os codificadores paraméticos e seu tipo mais importante, o codificador LPC. Além disso, fala-se sobre os codificadores híbridos e seu principal codificador CELP, cujo algoritmo é a base das maiorias das técnicas de codificação de voz aplicadas em telefonia celular.

Este capítulo também apresenta uma revisão bibliográfica de codificadores fonéticos desenvolvidos com suas respectivas características e taxa de transmissão, necessária ao embasamento teórico do codificador proposto neste trabalho. Por fim, o capítulo compara as técnicas de codificação de voz expostas, realizando um resumo dos codificadores em relação aos padrões de comunicações utulizados, relacionando suas taxas de *bits* e MOS.

CAPÍTULO 3

Modelos de Markov Escondidos Aplicados ao Reconhecimento de Fala

Os primeiros sistemas de reconhecimento automático de fala empregavam métodos baseados em reconhecimento de padrões. A ideia consistia em gerar padrões acústicos de palavras ou de fones e, a partir desses padrões e usando medidas de distância espectral, sequências de palavras ou de fones eram reconhecidas. Além das medidas de distância espectral, métodos baseados em programação dinâmica também foram empregados para alinhar os modelos de padrões acústicos com as informações acústicas que seriam reconhecidas. Esses métodos ainda são empregados, mas combinados com outras técnicas [32].

Entretanto, a maioria dos sistemas de reconhecimento de fala encontrados atualmente na literatura utiliza modelagem estatística para caracterizar os sinais em termos de modelos. Nesses sistemas, os modelos estatísticos têm proporcionado bons resultados. Neste contexto é preciso mencionar os Modelos de Markov Escondidos (HMM–*Hidden Markov Models*). Em sistemas de reconhecimento de fala, os HMMs são utilizados para modelar palavras e até mesmo unidades menores, que são os fones, como é o caso de interesse do presente trabalho.

O som produzido por cada pessoa é uma consequência das características do trato vocal, tamanho da garganta, posição da língua, entre outros fatores. Deste modo, em um reconhecimento de fala, os sons detectados pelo sistema são variações dos sons gerados por essas alterações físicas internas de cada locutor. Os Modelos de Markov Escondidos produzem internamente a voz como uma sequência de estados escondidos, e o som resultante como uma sequência de estados observáveis gerados por uma voz processada que mais se aproxima do estado verdadeiro (escondido).

Considera-se o trato vocal constituído por um número finito de configurações articulatórias ou estados. A cada estado é associado um sinal com características espectrais que o caracterizam. Desta forma, a potência espectral para curtos intervalos do sinal de voz é determinada pelo estado corrente do modelo, enquanto a variação da composição espectral do sinal com o tempo é governada predominantemente pela lei probabilística de transição de estados do canal de Markov básico [10, 33].

Os HMMs possuem várias vantagens, como baixo custo computacional na etapa de reconhecimento, são capazes de treinar vários eventos (fonemas, sílabas, entre outros), não é necessário haver, *a priori*, uma distribuição estatística das entradas para estimação dos parâmetros; as características temporais do sinal de entrada são modeladas inerentemente, pois se considera que as variações estatísticas do sinal de entrada estão implícitas na própria formulação probabilística [10].

A teoria relativa aos Modelos de Markov Escondidos já é bem conhecida e extensivamente documentada. Este capítulo tem o propósito de apresentar alguns conceitos básicos relacionada aos Modelos de Markov Escondidos (HMM – *Hidden Markov Models*), com o objetivo de fornecer uma base teórica ao entendimento do sistema de reconhecimento de fonemas descrito no Captítulo 4. Textos com informações claras e precisas podem ser encontrados em [32, 34].

3.1 Definição e Descrição do Modelo

A teoria de Modelos de Markov Escondidos foi introduzida na literatura na década de 1960 por Baum, entre outros autores [35] [36]. Sua utilização na área de reconhecimento automático de fala se deu na década de 1970, introduzida pelos trabalhos independentes de *Baker*, na *Carnegie Mellon University* [37], e *Jelinek* e colegas, na IBM [38] e, desde então, passou a ser largamente utilizada em diversas aplicações, pois manipulam bem os aspectos estatísticos e as sequências do sinal de voz.

Os Modelos de Markov Escondidos são bastante utilizados em sistemas de reconhecimento de fala pois têm um algoritmo eficiente e robusto para o treinamento e reconhecimento. O treinamento do modelo consiste em modelar o conjunto dos parâmetros acústicos extraídos do sinal de voz (observações) por uma sequência de estados (cadeia de Markov de primeira ordem) de acordo com a variação temporal da voz. Já no reconhecimento, a sequência de observações da elocução de teste é aceita como verdadeira se possuir uma medida de similaridade (verossimilhança) acima de um limiar estipulado com os parâmetros do modelo.

Um HMM consiste em um modelo estatístico baseado na teoria dos processos de Markov, utilizado para modelar processos estocásticos, diferenciando-se pelo fato dos seus estados não serem conhecidos, mas apenas o sinal emitido em cada um dos estados. Deste modo, é definido como um par de processos estocáticos (X,Y), em que X representa uma cadeia de Markov de primeira ordem e não é diretamente observável, enquanto Y é uma sequência de variáveis aleatórias que assumem valores no espaço de parâmetros acústicos (observações).

Assim, um HMM é caracterizado por um conjunto de N estados conectados por transições. Em cada instante de tempo t existe uma mudança de estado do sistema para estados diferentes ou para o mesmo estado, e um símbolo é emitido com uma determinada densidade de probabilidade de saída. A sequência de símbolos emitidos é chamada de sequência de observações, que representa a saída do HMM.

Um HMM é uma máquina de estados finita, que muda o seu estado atual i para o estado futuro j a cada instante de tempo t, e associado ao estado j um vetor de observações o_t é gerado, com densidade de probabilidade de saída $b_j(o_t)$, que pode ser discreta ou contínua. Os estados iniciais e

finais, que são considerados não-eminentes, não possuem probabilidade de saída associada, ou seja, esses estados não representam nenhum som, apenas demarcam o início e o fim do modelo.

A Figura 3.1 ilustra o processo para o reconhecimento de fala, em que o sinal de voz a ser representado pelo HMM consiste em uma sequência de vetores de observações $O = \{O_1, O_2, ...O_T\}$, em que cada vetor O_t é formado pelos parâmetros obtidos para cada bloco de amostra em que o sinal é considerado quase estacionário, que caracteriza o sinal de voz no t-ésimo intervalo de tempo. Assim, cada bloco de amostras do sinal de voz corresponde a um determinado intervalo de tempo.

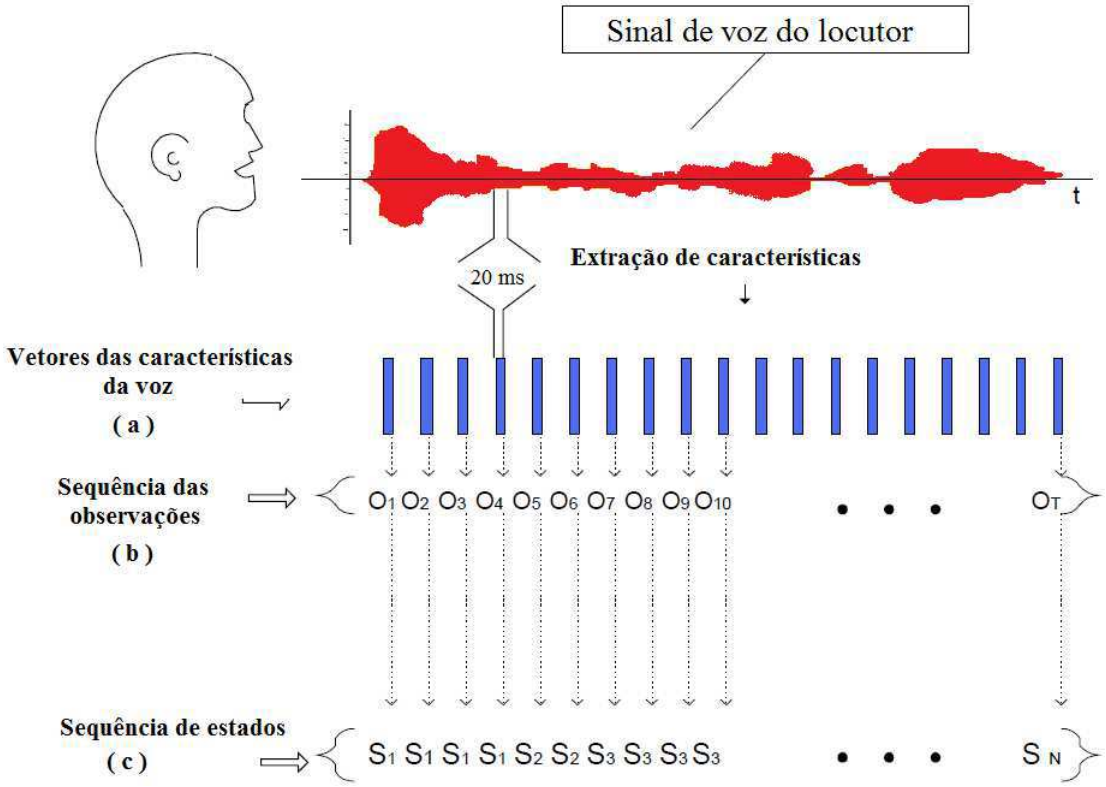

Figura 3.1: (a) Vetores das características da voz do locutor; (b) Sequência das observações; (c) Sequência de estados obtida durante o treinamento [39].

Em resumo, um HMM é constituído pelos seguintes parâmetros: N, M, A, B e Π, definidos como segue [32]:

1. N, o número de estados do modelo, representados como $S = \{S_1, S_2, ..., S_N\}$ e $q_t(i)$ indica estar no estado S_i no tempo t;

2. M, representa o representa número de símbolos no alfabeto, quando o espaço é definido por uma função de densidade probabilidade (fdp) discreta ou número de grupos quando for fdp contínua. Ou seja, é o número de diferentes símbolos de observação por estado.

3. Distribuição de probabilidade de transição entre os estados $A = \left[a_{ij}\right]$, em que

$$a_{ij} = P[q_{t+1} = j | q_t = i], \text{ em que, } 1 \leq i, j \leq N, \tag{3.1}$$

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

$$a_{ij} \geq 0 \text{ para } 1 \leq i, j \leq N, \tag{3.2}$$

$$\sum_{j=1}^{N} a_{ij} \; = \; 1 \text{ para } 1 \leq i \leq N. \tag{3.3}$$

Os valores das probabilidades de transições definem o tipo de topologia do HMM, com as restrições de avanço ou recuo entre estados.

4. B, representa a distribuição de probabilidade de observação dos símbolos, $B \; = \; \left[b_j(k) \right]$. Para o HMM discreto, seus elementos são do tipo $b_j(k) = P[o_t = v_k | q_t = j]$, sendo $1 \leq j \leq N$, $1 \leq k \leq M$, em que $b_j(k)$ é a probabilidade da variável aleatória o_t (observação) pertencer ao estado j e v_k representa o k-ésimo símbolo observado no alfabeto. As condições estocásticas seguintes devem ser satisfeitas:

$$0 \; \leq \; b_j(k) \; \leq 1, 1 \leq j \leq N, 1 \leq k \leq M, \tag{3.4}$$

e

$$\sum_{k=1}^{M} b_j(k) \; = \; 1. \tag{3.5}$$

Para o HMM contínuo, o conjunto de observações pertencentes a cada estado é dividido em M grupos, em que cada qual possui um vetor média e uma matriz covariância associados (gaussiana). A densidade de probabilidade em cada estado é, então, calculada pela soma das M distribuições gaussianas $\mathcal{N}$, ponderadas por c_{mj}, ou seja uma mistura de gaussianas.

$$b_j(o_t) \; = \; \sum_{m=1}^{M} c_{jm} \mathcal{N}(o_t, u_{jm}, U_{jm}), \tag{3.6}$$

em que:

u_{jm} = vetor média,

U_{jm} = matriz covâriancia,

c_{jm} = coeficiente de ponderação, que deve satisfazer as condições estocásticas:

$$c_{jm} \geq 0, 1 \leq j \leq N, 1 \leq m \leq M \tag{3.7}$$

e

$$\sum_{m=1}^{M} c_{mj} = 1, 1 \leq j \leq N. \tag{3.8}$$

5. Distribuição do estado inicial $\Pi = \left[\pi_i\right]$, sendo $\pi_i = P[q_1 = i]$, em que $1 \leq i \leq N$.

Uma forma compacta é utilizada para indicar o conjunto completo de parâmetros do modelo [40, 39]

$$\lambda = (A, B, \Pi). \tag{3.9}$$

3.1.1 Classificação dos HMMs

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

Como mencionado, a cada estado do HMM é associada uma distribuição de probabilidade. De acordo com essa distribuição, que pode ser função densidade de probabilidade (caso contínuo), ou função massa de probabilidade (caso discreto), os HMMs podem ser classificados como contínuos ou discretos. Além dessas duas possibilidades, existe uma outra possibilidade em que há uma combinação do HMM dicreto com o HMM contínuo, resultando em HMM denominado semicontínuo.

HMM discreto

No modelo HMM discreto é definido um dicionário (*codebook*) composto por palavras código ou vetores código. No processo de reconhecimento de fala, em cada quadro do sinal de fala obtém-se um vetor de parâmetros que, após a quantização vetorial, é associado a um dos M possíveis vetores-código. Neste processo, as sequências de observações são formadas por índices de vetores de um dicionário.

O HMM é dito discreto quando o número de possíveis símbolos de saída, M, é finito e a probabilidade de se emitir o símbolo v_k , no estado q_j, é dada por $b_j(k)$, com as seguintes propriedades:

$$b_j(k) \geq 0 \text{ para } 1 \leq j \leq N \text{ e } 1 \leq k \leq M, \tag{3.10}$$

$$\sum_{k=1}^{M} b_j(k) = 1. \tag{3.11}$$

em que:

- N é o número de estados do HMM;
- M é o número de símbolos discretos do modelo;
- $b_j(k)$ é a probabilidade de emitir o símbolo v_k no estado q_j.

HMM contínuo

O HMM é dito contínuo quando sua função densidade de probabilidade for contínua. Usualmente utiliza-se a função densidade de probabilidade modelada por uma mistura de M gaussianas multidimensionais, representada por [32]

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{M} c_{jk} G(\mathbf{o}_t, \mu_{jk}, U_{jk}), \tag{3.12}$$

em que:

- o_t é o vetor de parâmetros de entrada de dimensão D no instante de tempo t;
- M é o número de gaussianas na mistura para cada estado;
- c_{jk} é o coeficiente de ponderação para a k-ésima mistura no estado j;
- G é a função densidade de probabilidade gaussiana multidimensional com vetor média μ_{jk} e matriz de covariância U_{jk} para o componente da k-ésima mistura no estado j.

A função densidade de probabilidade gaussiana multidimensional G é dada por [41]

$$G(o_t, \mu_{jk}, U_{jk}) = \frac{1}{(2\pi)^{D/2}|U_{jk}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(o_t - \mu_{jk})U_{jk}^{-1}(o_t - \mu_{jk})'\right], \tag{3.13}$$

em que:

- D é a dimensão do vetor $\mathbf{o}_t$;
- $|U_{jk}|$ é o determinante da matriz de covariância;
- U_{jk}^{-1} é a inversa da matriz de covariância.

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

$$\sum_{k=1}^{M} c_{jk} = 1,\ 1 \leq j \leq N,\ c_{jk} \leq 0,\ 1 \leq k \leq L. \tag{3.14}$$

A função densidade de probabilidade por sua vez satisfaz a restrição

$$\int_{-\infty}^{\infty} b_j(\mathbf{o}) d\mathbf{o} = 1,\ 1 \leq j \leq N. \tag{3.15}$$

HMM semicontínuo

Um HMM é dito semicontínuo quando agrega as vantagens dos HMM discreto e contínuo, formando um novo modelo intermediário, denominado semicontínuo. Neste caso, a densidade de probabilidade de emissão dos símbolos de saída é dada por [42, 43]

$$b_j(\mathbf{o}_t) = \sum_{v_k \in \eta o_t} c_j(k) f\left(\mathbf{o}_t | v_k\right) \text{ para } 1 \leq j \leq N, \tag{3.16}$$

em que,

- N é o número de estados do modelo;
- $\mathbf{o}_t$ é o vetor de parâmetros de entrada;
- $\eta(\mathbf{o}_t)$ é o conjunto das funções densidade de probabilidade das palavras do dicionário que apresentam os M maiores valores de $f\left(o_t | v_k\right)$, $1 \leq M \leq K$. O valor adequado de M sugerido na literatura está na faixa de 2 a 8 [41];
- K é o número de funções densidade de probabilidade;
- v_k é o k-ésimo símbolo de saída;
- $c_j(k)$ é o coeficiente de ponderação das gaussianas;
- $f\left(\mathbf{o}_t | v_k\right)$ é o valor da k-ésima função densidade de probabilidade para o vetor de parâmetros de entrada $\mathbf{o}_t$.

A operação de quantização vetorial no processo de HMM discreto divide o espaço acústico em diversas regiões usando tipicamente distorção espectral. Essa característica é uma desvantagem do modelo HMM discreto, pois diminui sua precisão. Para contornar esse problema pode-se representar o dicionário por uma mistura de funções densidade de probabilidade, em que a distribuição seja sobreposta em vez de ser dividida. Dessa forma, cada palavra do dicionário é representada por uma das funções densidade de probabilidade.

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

Um HMM semicontínuo pode tornar-se discreto ou contínuo. Quando o valor de M é igual a 1, o HMM semicontínuo torna-se um HMM discreto com um *codebook* formado por funções densidade de probabilidade. Neste caso, usa-se apenas a função $f\left(\mathbf{o}_t|v_k\right)$ que apresentar maior valor para calcular a densidade de probabilidade de emissão de símbolos de saída.

No caso em que o valor de M é igual a K, pode-se considerar o HMM semicontínuo, como um HMM contínuo em que todas as misturas (vetor media μ_{jk} e matriz de covariância U_{jk}) são iguais para todos os estados e todos os modelos. A única variante de um estado para outro são os valores dos coeficientes de ponderação [40].

Topologia dos HMMs

As probabilidades das transições definem o processo de Markov e sua ordem. Quando a transição feita para o estado atual não depender da ocorrência de todos os estados anteriores, mas, somente do estado imediatamente anterior, é caracterizado um Processo de Markov de Primeira Ordem. De acordo com a matriz de transição A, a Cadeia de Markov assume uma certa topologia. A topologia, ou estrutura de um HMM, é determinada pelas transições que ocorrem entre estados.

Deste modo, os HMMs são classificados em ergódicos (totalmente conectados) ou *left-right*, também conhecido como modelo de Bakis [34] [32].

O modelo ergódico, ilustrado na Figura 3.2(a), tem a característica de não restringir nenhuma transição entre os estados, ou seja, a partir de um estado é possível atingir todos os outros estados, fazendo com que sua matriz de transição de estados fique totalmente preenchida. Para esse modelo, tem-se [39]

$$a_{ij} \geq 0, \text{para todos } i \text{ e } j. \tag{3.17}$$

Um HMM do tipo ergódico não permite uma ótima representação da voz, embora permita uma maior flexibilidade na geração das observações. Entretanto, a principal desvantagem deste tipo de modelo está na dificuldade em modelar a sequência temporal dos eventos acústicos em cada estado, além de, no processo de treinamento, aumentar o risco de convergência em um máximo local. Sendo assim, quando este modelo é utilizado para o reconhecimento de voz, as probabilidades de transição de retorno obtidas são próximas a zero [32].

O modelo *left-right* (esquerda-direita), apresentado na Figura 3.2(b), na qual a sequência de estados associada ao modelo tem a propriedade de nenhuma transição ser permitida para estados cujo índice seja menor do que o atual. Em HMM do tipo esquerda direita, a primeira observação é obtida quando a cadeia de Markov se encontra em um estado determinado, chamado de estado inicial. A última observação é gerada enquanto a cadeia de Markov está em um outro estado, chamado estado final ou estado de absorção. Outra propriedade deste tipo de modelo é que uma

vez que a cadeia de Markov deixa um estado, aquele estado não pode ser visitado em um tempo posterior [44]. Ou seja,

$$a_{ij} = 0, \text{para todo } j < i. \tag{3.18}$$

A probabilidade do estado inicial é

$$\pi_i = \begin{cases} 1, & i = 1 \\ 0, & i \neq 1 \end{cases} \tag{3.19}$$

Frequentemente, são impostas restrições nas transições do modelo, atribuindo-se a cada transição um número máximo de estados que pode ser alcançado, ou seja,

$$a_{ij} = 0, \text{para todo } i > j + \Delta. \tag{3.20}$$

No exemplo ilustrado na Figura 3.2(b) o valor de Δ é igual a 2, ou seja, permite a transição até o segundo estado à frente.

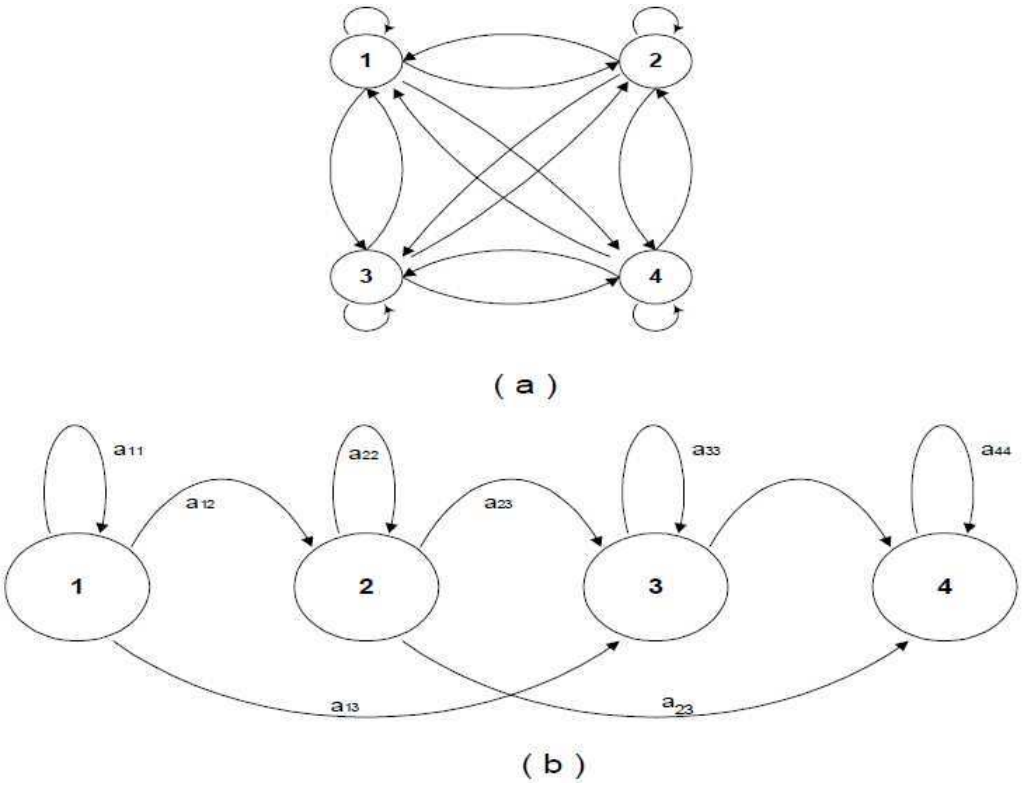

Figura 3.2: (a) Modelo ergódico com quatro estados; (b) Modelo esquerda-direita com quatro estados [39].

Uma vez que a fala possui uma estrutura ineretemente sequencial, para um sistema de reconhecimento de voz, os modelos esquerda direta apresentam melhores resultados comparados aos modelos ergódicos. Além disso, a liberdade adicional de transição de estados presente nos modelos ergódicos não reflete as variações dos parâmetros da fala caracterizados por um vetor de padrões [45] [8]. Desta forma, o modelo esquerda direta foi escolhido para ser utilizado neste trabalho.

Na tarefa de reconhecimento de fala geralmente são adotadas algumas simplificações da teoria de modelos de Markov, que podem ser formalizadas da seguinte maneira [34, 22, 32]:

1. **A suposição de Markov:** o próximo estado do HMM depende somente do estado atual, o modelo resultante torna-se, então, um HMM de primeira ordem. Ou seja,

$$a_{ij} = p\{q_{t+1} = j | q_t = i\}. \tag{3.21}$$

   O próximo estado pode depender de n outros estados anteriores, gerando um modelo denominado HMM de ordem n. Entretando, um HMM de ordem mais elevada implica aumento da complexidade computacional. Nos sistemas de reconhecimento de fala é comum o uso de HMM de primeira ordem, no entanto, alguns estudos têm usado HMM de ordem superior.

2. **A suposição da estacionariedade:** Assume-se que as probabilidades de transição de um estado para outro não se alteram durante o tempo. Matematicamente,

$$a_{ij} = p\{q_{t_1+1} = j | q_{t_1} = i\} = p\{q_{t_2+1} = j | q_{t_2} = i\} \tag{3.22}$$

3. **A Suposição das observações independentes:** Assume-se que uma dada observação corrente é estatisticamente independente das observações anteriores e posteriores, ou seja, não há correlação entre observações adjacentes. Com isso, o HMM desconsidera o efeito da coarticulação. Matematicamente, considere uma sequência de observações, $O = \{O_1, O_2, ..., O_T\}$ e um modelo λ, obtendo

$$P\{O|q_1, q_2, ...q_T, \lambda\} = \prod_{t=1}^{T} P(O_t|q_t, \lambda). \tag{3.23}$$

Nos sistemas de reconhecimento de fala, os HMM são construídos a partir de um conjunto de dados de treinamento para cada segmento de fala, por exemplo, um fonema. Para identificar um determinado segmento de fala, calcula-se a medida de probabilidade associada aos HMMs de referência já armazenados. Neste contexto, um fonema será identificado quando apresentar a maior

probabilidade com o HMM de referência. Desta forma, a modelagem por HMM necessita, para as etapas de treinamento e identificação, da resolução de três problemas básicos, descritos a seguir.

3.2 Modelagem do HMM

A modelagem de um HMM é realizada em três etapas: treinamento, reconhecimento e decodificação.

A etapa de treinamento tem o objetivo de determinar os parâmetros do modelo que maximizem a probabilidade de geração da observação. Para solucionar este problema, utiliza-se o o algoritmo *Forward-Backward*, também conhecido como algoritmo de re-estimação de Baum-Welch [46].

O reconhecimento consiste em determinar qual o modelo, dentre os vários obtidos na etapa de treinamento, que provavelmente gerou uma dada sequência de observação. O algoritmo *Forward* é utilizado na solução deste problema.

Na decodificação determina-se a sequência de estados que provavelmente produziu uma determinada sequência de entrada. Este problema é solucionado com o algoritmo de Viterbi.

A solução desses três problemas permite a elaboração de um sistema de reconhecimento automático da fala utilizando HMM.

3.2.1 O problema do treinamento

O treinamento dos HMMs consiste em ajustar os parâmetros do modelo para satisfazer algum critério de otimização. Ou seja, dado um modelo λ e uma sequência de observações $O = \{O_1, O_2, ..., O_T\}$, ajustar os parâmetros do modelo $\{A, B, \pi\}$ de modo a representar com maior eficiência o sinal que está sendo modelado maximizando $P\{O|\lambda\}$.

O método mais conhecido e utilizado para o treinamento dos HMMs é o algoritmo de Baum-Welch. Este método consiste em um conjunto de equações recursivas, empregando o critério da maximização da verossimilhança, em que o processo de treinamento é repetido enquanto a verossimilhança na interação atual é maior do que a verossimilhança da iteração anterior. Ou seja, o método de Baum-Welch pode ser descrito por meio dos seguintes passos [47, 48]:

1. Para cada l-ésima unidade de treinamento, que pode ser palavras, fonemas ou trifones, atribuir valores iniciais para os parâmetros do modelo $\lambda_l(A, B, \pi)$ e para a probabilidade P_l que representa os modelos HMM de referência para cada uma das L unidades de treinamento;

2. Com base no algoritmo de re-estimação de Baum-Welch, o segundo passo consiste na re-estimação dos parâmetros do modelo para a obtenção de $\bar{\lambda}_l$;

3. No terceiro passo, deve-se calcular a probabilidade $\bar{P}_l$ associada ao modelo $\bar{\lambda}_l$ re-estimado e fazer a comparação com a probabilidade anteriormente calculada;

4. Se $\bar{P}_l - P_l \leq \delta$ (limiar), o processo de re-estimação é finalizado. Caso contrário, retorna-se ao passo 2.

Ao se utilizar a topologia esquerda-direita como modelo de HMM, algumas regras devem ser obdecidas na atribuição dos valores inicial dos parâmetros do modelo, como:

- A restrição $a_{ij} = 0,\ j < i,\ j > i + 2$ deve ser considerada na construção da matriz $A = \left[a_{ij}\right]$, visto que o instante de tempo t não pode ser visitado em um instante de tempo posterior. Esta regra deve ser obedecida até o final do processo de re-estimação;
- O vetor de probabilidade é $\pi_i = 1, \ldots, 0$, uma vez que o modelo esquerda-direita é sempre inicializado no estado 1, sendo desnecessária a sua re-estimação;
- Na construção da matriz $B = \left[b_i(k)\right]$, $b_i(k)$ é inicializado com $1/M$, em que M representa o número de misturas gaussianas, para todos j, k, ou seja, assume-se que todos os símbolos nos estados são igualmente prováveis.

As equações do método de reestimação de Baum-Welch são [33, 32]:

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j\left(O_{t+1}\right) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_i(t)\beta_t(i)},\ 1 \leq i \leq N,\ 1 \leq j \leq N, \tag{3.24}$$

$$\overline{b_j(k)} = \frac{\sum_{t=1, O_t = w_k}^{T} \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)},\ 1 \leq j \leq N,\ 1 \leq k \leq N, \tag{3.25}$$

em que,

1. $\overline{a_{ij}}$ = (número esperado de transições do estado q_i para o estado q_j)/(número esperado de transições no estado q_i).
2. $\overline{b_j(k)}$ = (número esperado de vezes no estado j observando o símbolo w_k)/(número esperado de vezes no estado j).

Com as seguintes condições:

$$\sum_{j=1}^{N} a_{ij} = 1;\ \sum_{k=1}^{M} b_j(k) = 1; \sum_{i=1}^{N} \pi_i = 1,\ a_{ij} \geq 0,\ b_j(k) \geq 0,\ \pi_i \geq 0. \tag{3.26}$$

Cada parâmetro $b_j(\mathbf{o}_t)$, $1 \leq j \leq N$ e $1 \leq t \leq T$, é obtido a partir da comparação (em relação a um dado estado j e variando t), com os valores da matriz $\left[b_j(k)\right]$ referentes ao índice k do símbolo associado ao vetor o_t no mesmo estado j. Atribui-se a $b_j\left(\mathbf{o}_t\right)$ o valor de $b_j(k)$ correspondente ao referido símbolo w_k, no estado j [48].

Para definir o conjunto de equações de re-estimação dos parâmetros do modelo por meio do algoritmo de Baum-Welch é necessário definir dois outros algoritmos, *forward* e *backward* [49].

- **Algoritmo *Forward***

Inicialmente é definida a variável *forward* $\alpha_t(i)$, denominada probabilidade de avanço (*forward probability*), como [32]

$$\alpha_t(i) \;=\; P\left\{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_t, q_t = i|\lambda\right\}, \tag{3.27}$$

que representa a probabilidade da sequência de observações parciais $\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_t$ segundo o tempo crescente (iniciando em $t \;=\; 1$ indo até $t \;=\; T$), dado o modelo λ. O algoritmo pode ser resumido em três passos:

1. Inicialização

$$\alpha_1(i) \;=\; \pi_i b_i\left(\mathbf{o}_1\right),\; 1 \leq i \leq N. \tag{3.28}$$

2. Recursão

$$\alpha_{t+1}(j) \;=\; \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right]\; b_j(\mathbf{o}_{t+1}), \begin{cases} 1 \leq t \leq T-1 \\ 1 \leq j \leq N \end{cases} \tag{3.29}$$

3. Término

$$P\{\mathbf{O}|\lambda\} \;=\; \sum_{i=1}^{N} \alpha_T(i). \tag{3.30}$$

O valor de $P\{\mathbf{O}|\lambda\}$ é uma medida da probabilidade de uma determinada locução formada pela sequência de observações $\mathbf{O}$ ter sido produzida pela sequência de estados $\mathbf{Q} \;=\; \left[q_1, q_2, q_3, ..., q_t, ..., q_T\right]$.

- **Algoritmo *Backward***

De forma similar, inicialmente é definida a variável $\beta_t(i)$, denominada probabilidade de retrocesso (*backward probability*), definida em [32] como

$$\beta_t(i) \;=\; P\{\mathbf{o}_{t+1}, \mathbf{o}_{t+2}, ..., \mathbf{o}_T | q_t = i, \lambda\}. \tag{3.31}$$

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

1. Inicialização

$$\beta_T(i) \;=\; 1,\; 1 \leq i \leq N. \tag{3.32}$$

2. Recursão

$$\beta_t(i) \;=\; \sum_{j=1}^{N} a_{ij} b_j \left(\mathbf{o}_{t+1}\right) \beta_{t+1}(j), \begin{cases} t \;=\; T-1, T-2, ..., 1 \\ 1 \leq i \leq N. \end{cases} \tag{3.33}$$

Os valores calculados de forma recursiva das variáveis de *forward* e *backward* tendem a se tornar bem menores que um à medida que a sequência de observações é processada, ocasionando *underflow*.

Com o objetivo de contornar este problema, utiliza-se um fator de normalização ou o logaritmo dos parâmetros. A normalização consiste em mutiplicar os termos $\alpha_t(i)$ e $\beta_t(i)$ por um fator que é independente de i, mantendo estes termos dentro da faixa de precisão do computador para $1 \leq t \leq T$.

As equações a seguir definem o cálculo do fator de normalização ($\hat{c}(t)$) e dos parâmetros normalizados para o instante de tempo igual a um [41].

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

$$\overline{\alpha_1}(i) \;=\; \alpha_1(i). \tag{3.34}$$

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

$$\hat{c}_1 \;=\; \frac{1}{\sum_{i=1}^{N} \overline{\alpha}_1(i)}. \tag{3.35}$$

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

$$\hat{\alpha_1}(i) = \overline{\alpha_1}(i)\hat{c_1}. \tag{3.36}$$

4. Recursão

$$\overline{\alpha_{t+1}}(j) = \left[\sum_{i=1}^{N} \hat{\alpha}_t(i)\, a_{ij}\right] b_j(\mathbf{o}_{t+1}), \begin{cases} 1 \leq t \leq T-1 \\ 1 \leq j \leq N \end{cases} \tag{3.37}$$

$$\hat{c_t} = \frac{1}{\sum_{i=1}^{N} \overline{\alpha_t}(i)}, \tag{3.38}$$

$$\hat{\alpha_t}(i) = \overline{\alpha_t}(i)\hat{c_t}, \tag{3.39}$$

$$\hat{\beta}_t(i) = \hat{c_t}\overline{\beta_t}(i). \tag{3.40}$$

Para a obtenção de uma boa estimativa dos parâmetros do modelo, uma sequência com uma única observação não é suficiente. Assim, sequências com múltiplas observações devem ser usadas. A sequência de treinamento com múltiplas observações é composta por uma ou mais observações das mesmas palavras.

Os parâmetros são calculados a partir das variáveis *forward* e *backward* normalizadas, do fator de normalização $\hat{c}$, dos vetores de parâmetros acústicos $\mathbf{o}_t$, dos parâmetros que compõem o modelo e dos valores de verossimilhança normalizada definida pela equação 3.41

$$F_t(j,m) = \frac{c_{jm} G(o_t, \mu_{jm}, U_{jm})}{\sum_{k=1}^{N_g} c_{jk} G(o_t, \mu_{jk}, U_j k)}, \tag{3.41}$$

em que N_g é o número de gaussianas na mistura no estado j. A probabilidade de transição de estado, o peso, a média e a matriz de covariância são estimados pelas Equações 3.42–3.45.

- Probabilidade de transição de estados

$$\overline{a_{ij}} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_{d-1}} \hat{\alpha}_t^d(i) a_{ij} b_j(\mathbf{o}_{t+1}^d)\hat{\beta}_{t+1}^d(j)}{\sum_{d=1}^{D}\sum_{t=1}^{T_{d-1}} \hat{\alpha}_t^d(i)\hat{\beta}_t^d(i)/\hat{c}_t^d}. \tag{3.42}$$

- Peso ou coeficiente de ponderação

$$\overline{c_{jm}} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m)/\hat{c}_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)/\hat{c}_t^d}. \tag{3.43}$$

- Média

$$\overline{\mu_{jm}} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m)o_t^d/\hat{c}_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m)/\hat{c}_t^d}. \tag{3.44}$$

- Matriz de covariância

$$\overline{\mathbf{U}_{jm}} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m)(o_t^d-\mu_{jm})(o_t^d-\mu_{jm})'/\hat{c}_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j)\hat{\beta}_t^d(j)F_t^d(j,m)/\hat{c}_t^d}. \tag{3.45}$$

3.2.2 O problema do reconhecimento

O problema do reconhecimento consiste em determinar qual modelo HMM que mais provavelmente gerou uma determinada sequência de observações. Matematicamente, dado um modelo λ e uma sequência de observações $O = \{O_1, O_1, ..., O_T\}$, o reconhecimento busca calcular $P(\mathbf{O}|\lambda)$, ou seja a probabilidade de que as observações tenham sido geradas por aquele modelo.

Em sistemas de reconhecimento de fala, há um modelo de HMM treinado para cada unidade acústica, como palavras, fonemas, trifones, entre outras. Para cada unidade ser reconhecida, a se-

quência de observação é comparada com os modelos treinados por meio do cálculo da probabilidade associada a cada modelo de referência. Para calcular a probabilidade, utilizam-se as equações apresentadas na seção anterior.

3.2.3 O problema da decodificação

O problema da decodificação consiste em encontrar a sequência de estado ótima, dado um modelo λ e uma sequência de observações $O = \{O_1, O_1, ..., O_T\}$. Para isto, utiliza-se o algoritmo de Viterbi.

O algoritmo de Viterbi encontra a sequência de estados ótima q_t*, entre todas as possíveis sequências q, utilizando o seguinte critério [40]

$$q_t* = \arg\max P(q_t = i, \mathbf{O}|\lambda). \tag{3.46}$$

A Figura 3.3 ilustra um estrutura de treliça que relaciona a sequência de estados e intervalos de tempo. É possível observar nesta figura que há mais de um caminho parcial chegando a cada nó ou estado, cada um com determinado comprimento ou valor de probabilidade, para vários instantes de tempo diferentes. É chamado de sobrevivente correspondente a cada nó, aquele segmento de caminho mais curto, ou seja, o que apresenta maior valor de probabilidade. Deste modo, para cada instante de tempo, existe um número de sobreviventes igual ao número de nós na treliça [48].

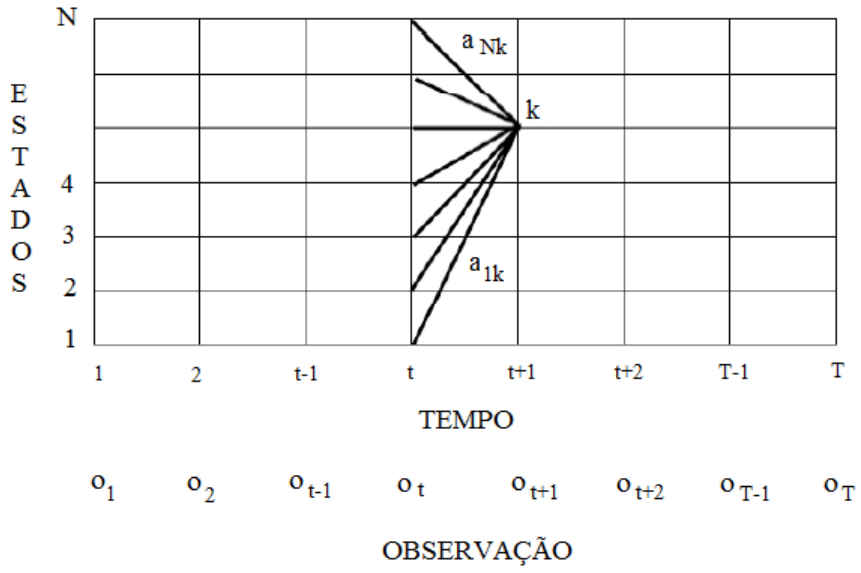

Figura 3.3: Algoritmo de Viterbi [10].

A cadeia de markov deve terminar em um estado bem determinado, existindo apenas um sobrevivente no último instante de tempo. O caminho total (de $t = 1$ até $t = T$) representa o menor caminho percorrido, ou seja, apresenta o maior valor de probabilidade. Percorrendo de volta a sequência de estados desse caminho, determina-se a sequência de estados associada que fornece o caminho mais provável, ou seja, a sequência de estados ótima [10, 48, 8].

Para a aplicação do algoritmo de Viterbi, é necessário inicialmente defini o maior valor da probabilidade em um caminho, no instante de tempo t, ou seja, considerando as t primeiras observações que terminam no estado q_i, tem-se por indução que

$$\delta_t(i) = \max_{q_1,q_2,\ldots,q_t} P[q_1 q_2, \ldots q_{t-1}, q_t = i, \mathbf{o}_1 \mathbf{o}_2 \ldots \mathbf{o}_t | \lambda], \tag{3.47}$$

que pode ser reescrita como

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(\mathbf{o}_{t+1}). \tag{3.48}$$

Para se obter a sequência ótima dos estados e maximizar a Equação 3.48, inicialmente define-se a variável $\psi_t(j)$. O método para se encontrar a sequência de estados ótima é dado por [47, 8]

1. Inicialização – para todos os i

$$\delta_1(i) = \pi_i b_i(\mathbf{o}_1), \text{para } 1 \leq i \leq N, \tag{3.49}$$

$$\psi_t(i) = 0. \tag{3.50}$$

2. Recursão

$$\delta_t(j) = \max_{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(\mathbf{o}_t), 2 \leq t \leq T, 1 \leq j \leq N, \tag{3.51}$$

$$\psi_t(i) = \arg\max_{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij}\right], 2 \leq t \leq T, 1 \leq j \leq N. \tag{3.52}$$

3. Término

$$p^* = \max_{1 \leq i \leq N} \left[\delta_T(i)\right]. \tag{3.53}$$

$$q_T^* = \arg\max_{1 \leq i \leq N} \left[\delta_T(i)\right]. \tag{3.54}$$

4. Sequência de estados ótimos

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right), t = T - 1, T - 2, \ldots, 1. \tag{3.55}$$

3.3 HMM em Reconhecimento de Fala

Um sistema de reconhecimento de fala usa o HMM para modelar segmentos como palavras, fonemas, trifones, difones, entre outros [50, 46, 32]. A escolha do segmento utilizado para o reconhecimento de fala deve ser realizada de acordo com o banco de voz que se tem para treinar e

testar o sistema, em outras palavras, o quão variáveis são as unidades que compõem o banco de voz para o treinamente e teste do sistema. Isso porque o tamanho do banco de voz utilizado para testar e treinar o sistema é um fator importante para se obter resultados confiáveis. Quanto mais dados houver para o treino dos HMMs, mais bem adaptados estarão à língua em questão.

Na escolha de segmento de maior duração, como as palavras, seria necessário para treinar o sistema um banco de voz com várias amostras das mesma palavras para tornar cada HMM treinado apto a reconhecer cada palavra referente ao seu modelo. No entando, os bancos de vozes do português brasileiro são considerados de pequeno porte, dificultando a construção de um sistema de reconhecimento de voz que faz o uso de palavras para treinar seus modelos.

Deste modo, os sistemas de reconhecimento de fala tendem a usar segmentos fonéticos para treinar os modelos HMM, devido à variabilidade de cada fone encontrada no banco de voz. Deste modo, para reconhecer unidades mais longas como palavras ou frases, realiza-se a concatenação dos modelos de fonemas que as formam para construir sua representação.

Os sistemas de reconhecimento de fala podem ser classificados como dependentes ou indepedentes de locutor. No caso dos sistemas dependentes de locutor, os modelos de HMM são treinados com apenas um locutor e estão aptos a reconhecer com uma boa taxa de acerto apenas o locutor utilizado na fase de treinamento. Por outro lado, os sistemas indepedentes de locutor, utilizam vários locutores para o treinamento dos seus modelos e devem está aptos para reconhecer a fala de qualquer locutor, mesmo aqueles que não participaram da etapa do treinamento.

Ao se utilizar segmentos fonéticos, os sistemas de reconhecimento de fala podem ser classificados como independentes ou dependentes de contexto. O sistema independente de contexto é caracterizado por treinar os fones sem levar em consideração o contexto em que está inserido. O dependente de contexto considera que um fonema em um contexto sofre influência dos fonemas vizinhos e realiza a modelagem contínua de trifones [51].

O sistema de reconhecimento de fala desenvolvido neste trabalho é caracterizado por ser dependente de contexto (uso de trifones) e independente do locutor. Realiza a modelagem de fonemas com modelos de HMM contínuos, utilizando a topologia esquerda-direita com cinco estados e uma componente gaussiana por estado. Essas configurações foram escolhidas por apresentar bons resultados no reconhecimento de fala. No entanto, sistemas mais robustos usam uma mistura de componentes gaussianas por estado o que permite que as distribuições de cada estado sejam modeladas com maior precisão. Na prática, reporta-se que com cerca de dez componentes de mistura se obtém bom desempenho [52]. Vários trabalhos na área de voz obtiveram bons resultados utilizando tal estratégia [53, 54].

3.4 Considerações Finais

Os Modelos de Markov Escondidos são amplamentes utilizados em sistemas de reconhecimento de fala para modelar unidades acústicas como palavras, fonemas ou trifones e têm obtido bons resultados [55, 56].

Este capítulo aborda conceitos sobre HMM necessários ao entendimento do funcionamento do sistema de reconhecimento de fonemas desenvolvido neste trabalho. Inicialmente foi apresen-

tada a definição dos Modelos de Markov Escondidos. Foram descritas as classificações dos HMMs segundo dois critérios: quanto à distribição de probabilidade (discreta, contínua ou semicontínua) e quanto à topologia (ergódico e esquerda-direita).

Em seguida, foi explicado o funcionamento do HMM, tanto para a geração de padrões de referência (etapa de treinamento) quanto para a identificação dos padrões de teste (etapa do reconhecimento). Por fim, foi sucintamente abordado como os HMMs se inserem nos sistemas de reconhecimento de fala e as características do sistema de reconhecimento de fala utilizado no reconhecimento de fonemas, descrito no Capítulo 4.

CAPÍTULO 4

Reconhecimento de Fonemas

Este capítulo apresenta os conceitos necessários ao entendimento do reconhecimento de fonemas realizado no desenvolvimento do codificador proposto, descrito no Capítulo 5. Este reconhecimento é realizado por meio da implementação de um reconhecedor de fala que tem o objetivo obter uma sequência de fonemas reconhecidos cujos índices atribuídos são uma das saídas do emissor do codificador. Suas etapas incluem: processamento do sinal de voz, que engloba as etapas de pré-ênfase e segmentação do sinal de voz, extração de características, desenvolvimento do modelo acústico e decodificação.

Na literatura é possível encontrar várias técnicas utilizadas para o reconhecimento de fala, dentre as quais destacam-se: Modelos de Markov Escondidos, Redes Neurais Artificiais, Quantização Vetorial, Análise por Predição Linear e Alinhamento Dinâmico no Tempo (DTW – *Dynamic Time Warping*) [32, 10].

A maioria dos sistemas atuais de reconhecimento de fala contínua é baseada nos princípios de reconhecimento estatístico de padrões, como Modelos de Markov Escondidos (HMMs), Processos Gaussianos, Processos de Poisson, Processos de Markov, entre outros. Na construção destes padrões, os parâmetros extraídos são representados por modelos estatísticos, e a decisão é tomada usando o cálculo das probabilidades associadas aos modelos.

Um sistema de reconhecimento de voz tem como objetivo determinar qual a palavra, a frase ou a sentença que foi pronunciada. Para isso, realiza uma tarefa de reconhecimento de padrões, que inclui duas fases: treinamento e reconhecimento, como ilustrado na Figura 4.1. Com base nos dados de treinamento (fonemas, trifones, palavras ou frases), são gerados os modelos de referência (modelo acústico), aos quais são atribuídos rótulos que identificam cada padrão. Na fase de reconhecimento, a partir dos dados de teste (sinais de voz) são obtidos padrões de teste que, em seguida, são comparados com os modelos gerados durante o treinamento e, utilizando uma regra de decisão, identifica-se o que mais se assemelha ao padrão de entrada desconhecido [34, 57].

Esses sistemas utilizam um modelo estatístico de distribuição conjunta $P(W,X)$ entre a sequência de palavras pronunciadas W e a sequência de informações acústicas observadas X. Deste modo, o sistema de reconhecimento procura uma estimativa $\hat{W}$ da sequência de palavras pronunciadas, a partir da evidência acústica observada X, utilizando a distribuição de probabilidade a *posteriori* $P(W|X)$. Assim, o sistema escolhe a sequência de palavras que maximiniza [58, 56]

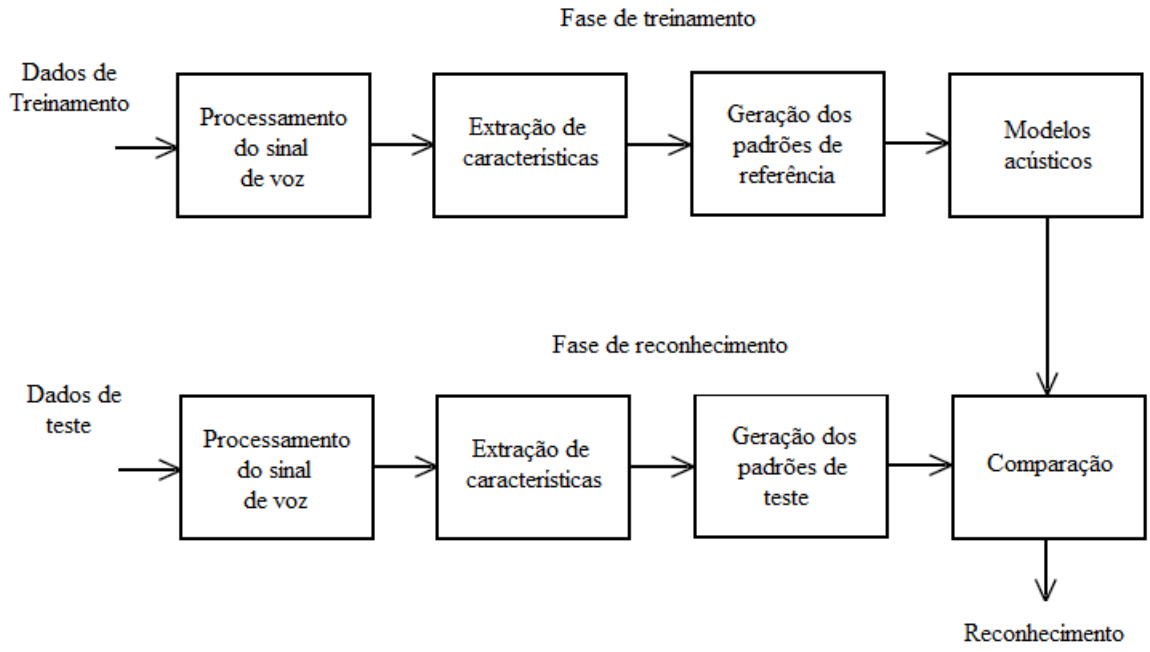

Figura 4.1: Diagrama de blocos de um sistema de reconhecimento de padrões aplicado ao reconhecimento de fala.

$$\hat{W} = \arg_w \max[P(W|X)] = \arg_w \max\left[\frac{P(W)P(X|W)}{P(X)}\right]. \tag{4.1}$$

Após a aplicação do teorema de Bayes, a distribuição a *posteriori* é decomposta na probabilidade a *priori* da sequência de palavras, ou seja, $P(W)$, assim como em $P(X|W)$ que consiste na probabilidade de observar a evidência acústica X quando a sequência W é pronunciada. A distribuição $P(W)$ refere-se às palavras que poderiam ter sido pronunciadas e está associada a um modelo de linguagem, enquanto a probabilidade de uma observação $P(X|W)$ é chamada de modelo acústico.

Um trecho de fala é convertido em uma sequência de vetores acústicos $X = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_T}\}$ que se refere a uma sequência de palavras $W = \{w_1, w_2, ..., w_n\}$. O reconhecimento de fala deve usar a Equação 4.1 para determinar a sequência mais provável, $\hat{W}$, para os vetores acústicos observados, X. Deste modo, cada palavra é convertida em uma sequência de fonemas. Para cada fonema há um modelo estatístico correspondente. A sequência de modelos estatísticos para representar a fala é concatenada, formando um único modelo composto e a probabilidade deste modelo gerar a sequência observada X é calculada, isto é, $P(X|W)$.

4.1 Descrição do Sistema de Reconhecimento de Fonemas

A Figura 4.2 ilustra as etapas necessárias para o reconhecimento de fonemas: extração de parâmetros do sinal de voz (*features*), denominado *Front-End*, modelo acústico que busca modelar, a partir das características do sinal de voz, o sinal acústico e o decodificador que, junto com as etapas citadas, realiza o processo de transcrição do sinal de voz.

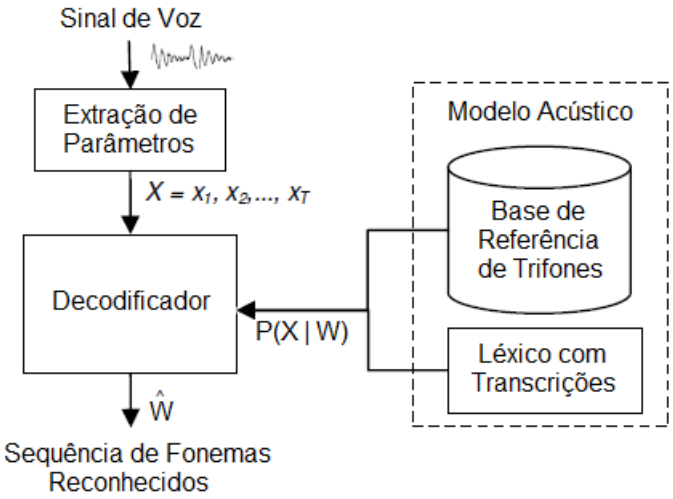

Figura 4.2: Diagrama em blocos de um sistema de reconhecimento de fonemas baseado em modelos estatísticos.

4.1.1 Processamento do Sinal de Voz

No processamento do sinal de voz estão inclusas as etapas de pré-ênfase e janelamento em m blocos, cada uma contendo N_A amostras.

Pré-ênfase

Após a aquisição dos dados, as locuções devem ser pré-enfatizadas. O aparelho fonador humano consiste em um sistema linear, lentamente variante com o tempo. As mudanças de características do aparelho fonador ocorrem em poucos milisegundos, geralmente entre 10 e 30 ms [32], fazendo com que o sinal de voz seja considerado quase estacionário nestes intervalos.

A voz produzida pelo aparelho fonador sofre perdas durante sua passagem pelo trato vocal, inclusive na sua radiação através dos lábios. A distorção provocada pelos lábios produz uma queda na envoltória espectral de, aproximadamente, 6dB/oitava [57].

Além disto, uma das características do sinal de voz é ter a maior parte da sua energia concentrada nas baixas frequências e apresentar baixas amplitudes nas altas frequências, fazendo com que o sinal de voz se torne especialmente vulnerável ao ruído, comprometendo o processo de reconhecimento. As altas frequências também são responsáveis pela produção de sons, especificamentes dos sons surdos. Desta forma, é necessário pré-enfatizar o sinal de voz com o objetivo de acentuar as frequências mais altas e tornar o espectro do sinal de voz mais plano.

Inicialmente, para se obter o modelo do sistema glotal, com os efeitos da radiação dos lábios e da variação da área da glote reduzidos, passa-se o sinal de voz por um filtro de primeira ordem $L(z)$ do tipo

$$L(z) = 1 - a_p z^{-1}. \tag{4.2}$$

O parâmetro a_p é denominado fator de pré-ênfase. Valores típicos de a_p são próximos de 1,0. Neste trabalho foi utilizado $a_p = 0,95$, que corresponde a 20 dB de amplificação para as frequências mais altas. Assim, a pré-ênfase é realizada por meio da fórmula usual, que relaciona a saída da pré-ênfase $s_p(n)$ com a entrada $s(n)$ [34]

$$s_p(n) = s(n) - 0,95s(n-1). \tag{4.3}$$

em que $s_p(n)$ representa a amostra pré-enfatizada e $s(n)$ o sinal original.

Segmentação do Sinal de Voz

Após a etapa da pré-ênfase, inicia a etapa da segmentação do sinal para análise a curtos intervalos. É possível segmentar o sinal de voz em janelas ou quadros de duração definida, desde que esteja dentro do intervalo em que o sinal de voz é considerado quase estacionário. A segmentação é levada a efeito com superposição de 50% entre os quadros, visando reduzir os efeitos da descontinuidade entre segmentos. Neste trabalho, foi utilizada uma janela de 25 ms com deslocamento da janela em análise de 10 ms.

Os tipos de janelas utilizadas para segmentar o sinal são: Janela Retangular, Janela de Hamming e Janela de Hanning, cujas características são mostradas a seguir.

- Janela Retangular

$$c(t) = \begin{cases} 1, & 0 \leq n \leq N_A - 1 \\ 0, & \text{caso contrário.} \end{cases} \tag{4.4}$$

- Janela de Hamming

$$J(n) = \begin{cases} 0,54 - 0,46\cos[2\pi n/(N_A - 1)], & 0 \leq n \leq N_A - 1 \\ 0, & \text{caso contrário.} \end{cases} \tag{4.5}$$

- Janela de Hanning

$$J(n) = \begin{cases} 2a\cos[\pi n/N_A] + b, & 0 \leq n \leq N_A - 1 \\ 0, & \text{caso contrário.} \end{cases} \tag{4.6}$$

sendo $2a + b = 1 (0 \leq a \leq 0,25$ e $0,5 \leq b \leq 1)$.

As Figuras 4.3 e 4.4 ilustram a janela retangular, de Hanning e Hamming, respectivamente no tempo e na frequência. A utilização de uma janela retangular, que particiona o sinal de voz em blocos consecutivos de dimensão igual ao comprimento do bloco, proporciona fugas espectrais indesejáveis alterando o espectro do sinal de voz. A utilização de janelas com sobreposição, como a de Hamming e Hanning diminui estes efeitos pois têm a características de possuir, no domínio da frequência, um lobo principal de amplitude bastante superior aos lobos secundários.

A janela de Hamming tem se mostrado mais eficiente quando comparada aos outros tipos de janela, com uma boa aproximação da janela ideal. Ela mantém as características espectrais do centro do quadro e elimina as transições abruptas das extremidades. Entretanto, atribui um peso muito baixo às amostras das extremidades. Para evitar efeitos indesejáveis a essas amostras, os blocos adjacentes são sobrepostos como o objetivo de cobrir tais amostras por outros blocos. A janela de Hanning assemelha-se a de Hamming porém proporciona um reforço menor nas amostras do centro e uma suavização maior nas amostras da extremidade. Desta forma, a janela de Hamming foi escolhida para ser utilizada neste trabalho [10].

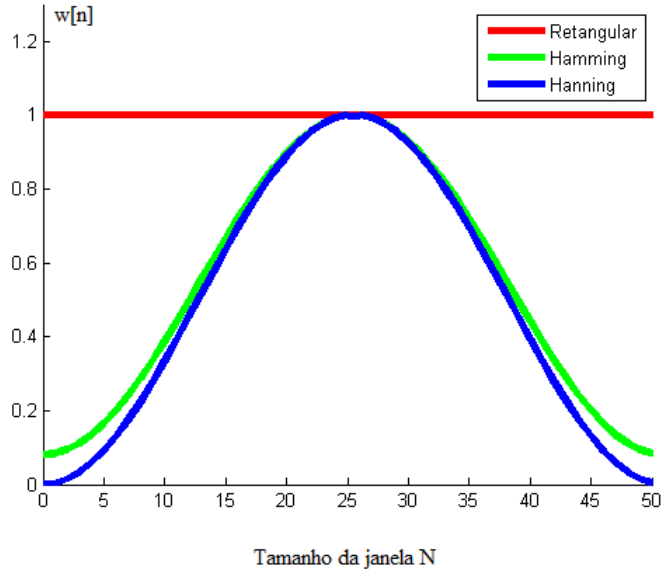

Figura 4.3: Janela retangular, de Hanning e Hamming.

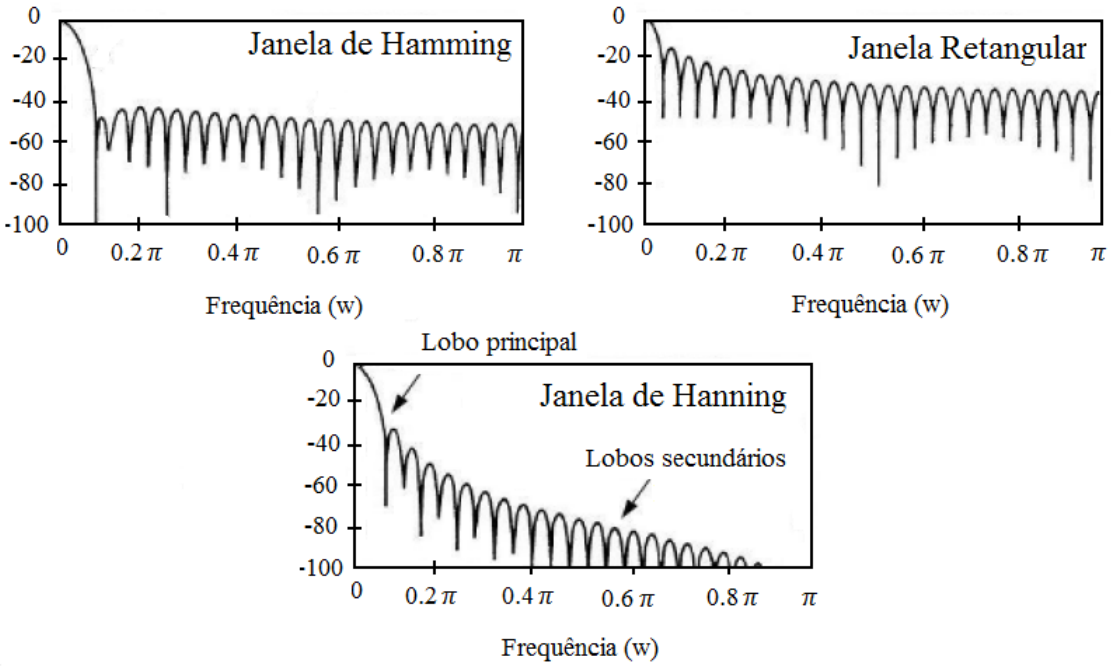

Figura 4.4: Respostas em frequência das janelas retangular, de Hanning e Hamming [48].

4.1.2 Extração de Características

Na etapa de extração de características, denominada *front-end*, busca-se extrair as informações mais relevantes do sinal de voz. Neste trabalho, para cada janela do sinal de voz se extraem os coeficientes MFCC's (*Mel Frequency Cepstral Coeficients*), proposto em 1980 por Davis e Merlmestein [59] e que têm como característica de representar o sinal de voz baseado no comportamento do ouvido humano [60].

O front-end opera em cada janela do sinal de voz (25 milisegundos), convertendo-os em vetores de parâmetros MFCC's. A Figura 4.5 ilustra as etapas necessárias à obtenção dos coeficientes MFCC. Inicialmente, em cada quadro do sinal de fala é aplicada uma Transformada Rápida de Fourier (FFT – *Fast Fourier Transform*), obtendo o espectro em frequência. Em seguida, o espectro passa por um conjunto de filtros triangulares na escala Mel, em que é possível verificar a redução da contribuição das frequências mais elevadas, característica do ouvido humano.

Como passo seguinte, tem-se uma compressão logarítmica, seguida de uma DCT (*Discrete Cosine Transform*) que diminui a correlação entre os elementos do vetor de parâmetros. Como saída, se tem vários coeficientes *cepstrais*, no entanto, escolhe-se os 12 primeiros coeficientes junto com a energia do sinal.

Além disso, adiciona-se aos coeficientes MFCC's as suas derivadas de primeira e segunda ordem adaptando a modelagem acústica que assume que os vetores acústicos estão descorrelacionados dos seus vetores vizinhos, as características dos órgãos do aparato vocal humano que garantem que há continuidade entre sucessivas estimativas espectrais. As componentes de primeira derivada representam a velocidade com que o espectro mel-cepstral varia, enquanto as componentes de segunda derivada consistem na aceleração do espectro mel-cepstral. Desta forma, no final do processo de extração, se obtém, para cada janela do sinal de voz, um vetor com 39 parâmetros.

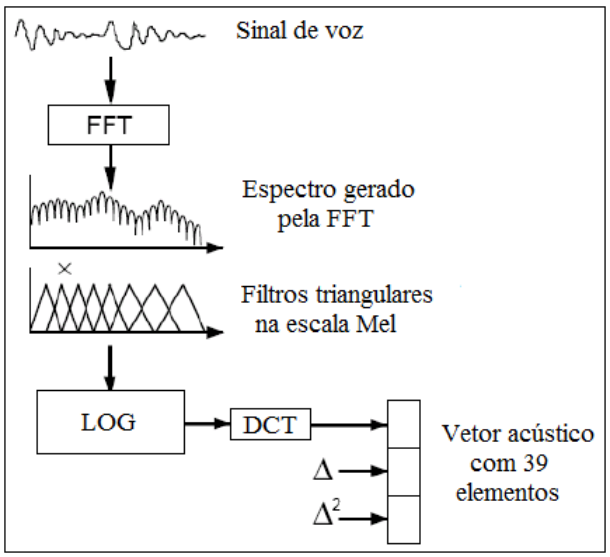

Figura 4.5: *Front-end* com processador baseado em MFCCs [56].

4.2 Modelo Acústico

O algoritmo usa as características acústicas extraídas do sinal de voz para criar um modelo matemático que representa cada tipo de segmento fundamental da fala, que pode ser em nível de palavras, de sentenças, ou mesmo nível fonético.

O treinamento de um modelo acústico capaz de generalizar a descrição de um fonema não é trivial, visto que características individuais como timbre de voz, sotaque e velocidade de fala dificultam essa generalização. Desta forma, torna-se necessário que o treinamento do modelo acústico seja realizado a partir de um grande conjunto de amostras de sentenças faladas por diversas pessoas com diferentes timbres de voz e sotaques. Além disso, é importante que as sentenças utilizadas sejam foneticamente balanceadas [53].

O modelo acústico calcula a verossimilhança da sequência de vetores X, dada uma sequência de palavras W, ou seja, $P(X|W)$. O trabalho proposto busca reconhecer cada fonema da fala contínua. Deste modo, são criados modelos acústicos para cada fonema da língua, e o modelo acústico procura, a partir do vetor que representa o sinal sonoro, inferir qual sequência de fonemas gera aquele vetor.

Atualmente, a solução mais encontrada na literatura para criação do modelo acústico é baseada na extração de amostras sonoras dos coeficientes de frequência *mel-cepstrais*, e, a partir deles, utilizar o HMM para modelar a sequência do sinal sonoro. Assim, cada fonema é representado por um HMM. O estado de saída do modelo de um fonema pode ser unido com o estado de entrada de outro para criar um HMM composto, permitindo que os modelos de fonemas sejam unidos para formar palavras e estas unidas para formar frases completas.

4.2.1 Modelagem Contínua de Trifones

Um modelo acústico que utiliza um HMM por fonema (monofones) supõe que um fonema pode ser seguido por qualquer outro. Entretando, uma das características do trato vocal é que seus articuladores não se movem de uma posição para outra imediatamente na maioria das transições de fonemas. Diante disso, para modelar a fala contínua é importante considerar os efeitos contextuais causados pelas diferentes maneiras que alguns fonemas podem ser pronunciados. Assim, o ideal é modelar e treinar cada um dos diferentes contextos de um mesmo fone com um HMM diferente para obter uma boa discriminação entre eles. Pode-se, dessa maneira, em vez de usar modelos de monofones, usar modelos de trifones.

Em um HMM que emprega trifones, cada fone possui um modelo distinto dependendo dos fones à sua direita e à sua esquerda. Como exemplo, considere a notação x-y+z, em que y representa o fonema central ocorrendo após o fonema x e antes do fonema z. O sinal de subtração ($-$) em x- representa o fone à esquerda e o sinal de adição ($+$) em $+z$ representa o fone à direita.

Há dois tipos de modelos de trifones: os trifones intrapalavra (*internal-word triphones*) e os trifones entre-palavras (*cross-word triphones*). A diferença entre esses modelos é que o primeiro não considera as fronteiras entre palavras, formando um número menor de trifones. O segundo modelo, *cross-word triphones*, produz uma modelagem mais precisa, pois considera as fronteiras entre palavras. Neste caso, o número de trifones aumenta consideravelmente, tornando necessário

um volume maior de dados para o treinamento. Como exemplo, considere a frase ”Muito prazer” em que suas transcrições utilizando monofones e trifones podem ser observadas na Tabela 4.1.

Tabela 4.1: Exemplos de transcrição utilizando monofones e trifones.

Monofones	sil m u j~ t u p r a z e X sil
Trifones intra-palavras	sil m+u m-u+j~ u-j~+t j~-t+u t-u p+r p-r+a r-a+z a-z+e z-e+X e-X sil
Trifones entre palavras	sil sil-m+u m-u+j~ u-j~+t j~-t+u t-u+p u-p+r p-r+a r-a+z a-z+e z-e+X e-X+sil sil

A migração da utilização dos modelos de monofones para trifones provoca um grande aumento no número de modelos. Com o uso dos monofones, seria necessária apenas a criação de trinta e oito modelos, referentes a cada um dos fonemas do Português Brasileiro. Entretanto, com o uso de trifones do tipo *cross-word* o número de modelos passa a ser 54.872 modelos (38^3).

Diante disso, os sistemas de reconhecimento de voz se deparam com o problema da insuficiência de dados para a estimação dos modelos trifones, uma vez que muitos dos trifones terão uma ou nenhuma ocorrência no *corpus*. Outra desvantagem no uso de trifones é um aumento da quantidade de parâmetros que devem ser treinados.

Para solucionar esse problema é comum o uso da união de misturas (*tied-mixture*), ou seja, compartilhar os componentes das misturas de gaussianas entre os estados dos HMMs. Isso porque muitos modelos de trifones possuem características acústicas semelhantes, sendo possível o compartilhamento das distribuições de probabilidade em seus estados [61] [62] [63].

Para o compartilhamento de estados é necessário, em primeiro lugar, que haja a união dos estados (*state-typing*) que são acusticamente comuns. Após as uniões, vários estados passam a compartilhar as mesmas distribuições. A Figura 4.6 ilustra um exemplo de compartilhamento de estados utilizando o monofone **a**, cujo modelo HMM com três estados, cada um com uma componente gaussiana, está apresentado na primeira linha da figura.

Na segunda linha da figura, o exemplo considera vários contextos em que o fonema **a** pode está incluido, formando trifones, cujo fonema central é o **a**. Em seguida, a figura ilustra o compartilhamento dos estados que apresentam características semelhantes. Desta forma, ao se unir dois estados acusticamentes indestinguíveis, por exemplo, o primeiro estado do trifone **p-a+R** com o primeiro estado do trifone **p-a+k**, os parâmetros são copiados de um estado para outro.

A união entre os estados pode ser realizada com a construção de uma árvore binária de decisão para cada fone [64] [65] [66]. Como ilustrado na Figura 4.7, em cada nó da árvore existe uma pergunta cuja resposta é “sim” ou “não”. As perguntas são escolhidas de forma a maximizar a verossimilhança entre os dados de treino e o conjunto resultante da união dos estados, para que existam dados de treino suficientes para estimar de forma robusta os parâmetros das distribuições de probabilidade gaussianas.

O objetivo da árvore é unir os estados que são acusticamente semelhantes. Deste modo, todos os estados de um fonema são posicionados no nó da raiz da árvore, de forma que ao percorrer a árvore os estados vão sendo divididos e, ao final do processo, todos os estados do mesmo nó folha são agrupados.

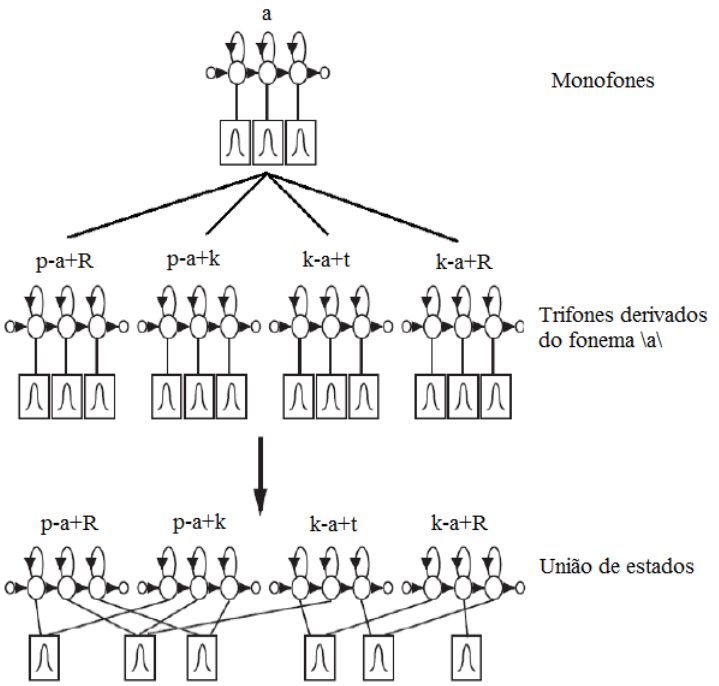

Figura 4.6: União de estados acusticamente indistinguíveis.

A vantagem da utilização de árvores de decisão é que o agrupamento de estados resultante pode estimar as probabilidades para qualquer contexto, ou seja, é possível sintetizar trifones que não aparecem no material de treino, utilizando as distribuições de probabilidades presentes nos nós terminais de seus estados.

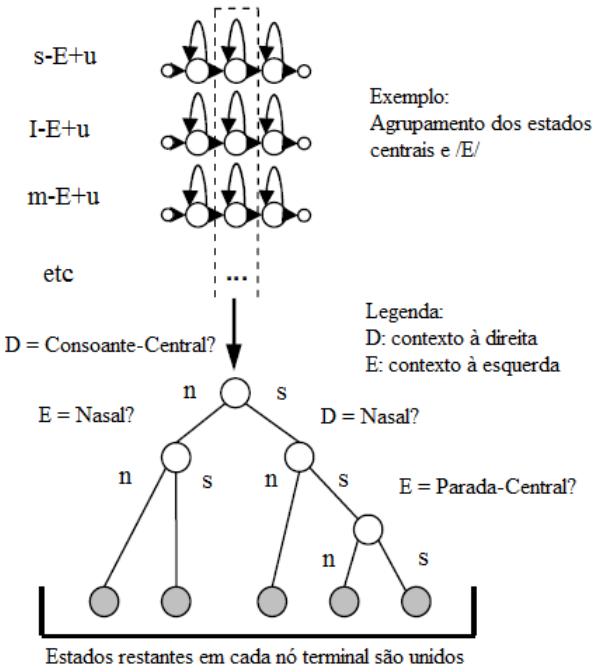

Figura 4.7: Agrupamento baseado em árvores de decisão (Apatado de [53]).

4.2.2 Decodificador

Depois da construção do modelo acústico, o último passo na construção do sistema de reconhecimento de fonemas é o decodificador, que tem a função de transcrever as amostras de voz desconhecidas para sua forma textual.

Para o seu funcionamento, o decodificador requer um dicionário e uma rede de fonemas. O dicionário de fonemas é um simples mapeamento entre um fonema e ele próprio. Isso indica ao decodificador quais devem ser as saídas do sistema. A rede de fonemas indica as transições possíveis que o decodificador pode fazer entre um fonema e outro.

O sistema de reconhecimento de fonemas desenvolvido neste trabalho foi implementado com o *software* HTK (*Hidden Markov Models Toolkit*) que consiste em um conjunto de ferramentas para modelar HMM, descrito sucintamente no Apêndice A. A Tabela 4.2 mostra um exemplo da saída do sistema de reconhecimento de fonemas utilizando o HTK. Por simplicidade, está ilustrada apenas uma palavra (algumas), resultando do reconhecimento realizado na frase "Há algumas coisas que não podem deixar de serem vistas em Paris". A primeira e segunda coluna do exemplo consiste no tempo inicial e final, respectivamente, de cada fonema. O tempo resultante do processo de reconhecimento pelo HTK é da ordem de 100 ns. Deste modo, por exemplo, o primeiro fonema, **a**, está contido entre 0,34 e 0,40 segundos do sinal de voz, o que resulta em um fonema com 60 ms e uma palavra com 0,45 segundos de duração. A terceira e quarta coluna representam respectivamente a transcição fonética da palavra e a verossimilhança com o modelo acústico.

Tabela 4.2: Segmentação automática com o HTK da palavra "algumas"

Tempo inicial	**Tempo final**	**Transcrição fonética**	**Verossimilhança**
3400000	4000000	a	212,618866
4000000	4700000	u	206,416595
4700000	5300000	g	172,741058
5300000	5900000	u	186,433228
5900000	6500000	m	194,832352
6500000	7200000	a	228,597275
7200000	7900000	s	387,630615

Para a análise de resultados do reconhecimento de fonemas, utiliza-se a ferramenta HTK, que realiza um alinhamento entre as sequências fornecidas pelo decodificador e os textos transcritos das frases, e gera uma taxa de erro de fonemas, conforme a WER (*Word Error Rate*), definida em [32],

$$\text{WER} = \frac{S_s + I + D_s}{N_s}, \tag{4.7}$$

em que N_s é o número total de palavras na sequência de teste e S_s, I e D_s são, respectivamente, o número total de erros por substituição (*substitution*), inserção (*insertion*) e supressão (*deletion*) na sequência reconhecida.

4.3 Considerações Finais

Este capítulo apresentou a teoria envolvida nos sistemas de reconhecimento de fala. Inicialmente foi discutido em que consiste um sistema de reconhecimento de fala, seguido da descrição de um sistema de reconhecimento de fonemas, que é o caso de interesse do presente trabalho, que envolve etapas como processamento do sinal de voz, extração de características, construção do modelo acústico e decodificação.

Para implementar um sistema de reconhecimento de fonemas, vários parâmetros devem ser considerados: técnica para construção do modelo acústico, que neste trabalho é a técnica HMM, segmento utilizados para treinar o modelo acústico (fones, trifones, palavras), topologia dos HMMs, número de componentes gaussianas por estados, entre outros parâmetros.

O sistema de reconhecimento de fonemas desenvolvido neste trabalho apresentou uma taxa de reconhecimento de 80% (WER = 20 %) e a descrição dos seus parâmetros está exposta no Capítulo 5.

CAPÍTULO 5

Descrição do Codificador de Voz

Este capítulo apresenta um sistema de transmissão que usa um codificador de voz pessoal. Desenvolvido para ser aplicado principalmente em sistemas móveis celulares, o codificador permite a transmissão do sinal de voz com baixas taxa de transmissão.

Uma das características dos usuários de telefonia móvel é a de armazenarem vários números de telefones celulares em seus aparelhos. No entanto, a sua comunicação com outros usuários no sistema móvel é feita com maior frequência com familiares e amigos mais próximos. Nessa comunicação, o codificador proposto surge como um sistema alternativo, e os usuários podem optar por utilizá-lo para realizarem uma comunicação com baixo custo da ligação, caso as companhias telefônicas façam a cobrança por taxa de transmissão.

O uso desse codificador também possibilita um aumento na capacidade do canal de transmissão, uma vez que com uma menor taxa de transmissão, a largura de banda requerida por cada usuário é menor, sendo possível multiplexar mais usuários em um mesmo canal de comunicação.

5.1 Descrição do Codificador

O codificador de voz proposto neste trabalho é do tipo fonético, visto que, entre as técnicas de codificação, é a que fornece a menor taxa de *bits*, e, tem a característica de utilizar um sistema de reconhecimento de fala com o objetivo de segmentar foneticamente o sinal de voz. Para alcançar uma redução na taxa de transmissão, esse codificador, em vez de codificar amostras do sinal de voz, quantiza parâmetros correspondentes a cada segmento de fala, como índices e informações sobre energia e duração.

Sabendo que no português brasileiro há trinta e oito fonemas, no projeto do codificador de voz é proposta a codificação destes segmentos por meio da atribuição de índices pré-estabelecidos, sendo assim possível codificá-los com, no máximo, seis *bits*, bem como suas informações de energia e duração. No entanto, o codificador tem a característica de fornecer uma taxa de *bits* variável, de acordo com a quantidade de fonemas pronunciada por segundo por cada usuário.

Como mencionado, o codificador encontra sua principal aplicação nos sistemas de telefonia móvel e deve ser interpretado como um sistema opcional aos usuários de telefonia celular. Inicial-

mente para que seja possível o seu uso, o sistema deve formar um banco de unidades acústicas específico para cada usuário, mediante a pronúncia de frases pré-estabelecidas.

A implementação do codificador inclui o desenvolvimento do emissor e receptor. O emissor é constituído por um segmentador e reconhecedor fonético, que converte o sinal acústico em uma sequência de segmentos fonéticos. A informação transmitida ao receptor consiste na sequência dos índices fonéticos além das informações de caráter prosódico, como energia e duração de cada fonema reconhecido.

O receptor realiza a síntese por concatenação de segmentos acústicos para a formação de palavras. Esses segmentos são as sílabas, encontros vocálicos e fonemas. Embora esses segmentos fiquem armazenados no receptor para a síntese, a seleção de unidades acústicas deve ser previamentes definida também no emissor para que seja possível o cálculo da energia e duração dos segmentos constituídos de dois fonemas, como as sílabas e encontros vocálicos.

Um das decisões fundamentais no projeto do codificador, e que interfere diretamente na qualidade do sinal sintetizado, é a escolha das unidades acústicas. Como o banco de voz utilizado para obter estas unidades é pequeno, deve-se utilizar vários dos possíveis segmentos do banco de voz. Deste modo, foram usadas sílabas, fonemas e encontros vocálicos. Assim, é possível garantir as coarticulações presentes entre fonemas que formam as sílabas e os encontros vocálicos. Em situações nas quais não há esses dois segmentos, o receptor faz o uso de fonemas e, para as vogais, faz o uso de algumas variações que são escolhidas de acordo com a posição que ocupam na síntese de uma palavra.

A seguir, são descritas as etapas desenvolvidas na implementação do codificador. Todas as etapas da segmentação e reconhecimento de fonemas descritas neste capítulo foram realizadas utilizando o *software* HTK. As demais etapas implementadas para construção do codificador foram com o MatLab.

5.2 Emissor

O emissor do codificador, cujo diagrama de blocos está ilustrado na Figura 5.1, é constituído das seguintes etapas: segmentação e reconhecimento de fonemas, atribuição de índices aos fonemas, obtenção da energia e duração de cada segmento e codificação das informações por meio de um codificador de Huffman. A seguir, tem-se a descrição de cada uma delas.

Segmentação e Reconhecimento Fonético

A segmentação e reconhecimento de fonemas é o primeiro e o mais importante passo para a implementação do codificador proposto. Seu resultado consiste no tempo inicial e final de cada fonema, necessário para se obter as informações prosódicas e sintetizar o sinal com um bom desempenho. É realizada com a utilização de técnicas de reconhecimento de sinais de fala, obtendo-se uma sequência de fonemas reconhecidos cujos índices, energia e duração formam as saídas do emissor do codificador.

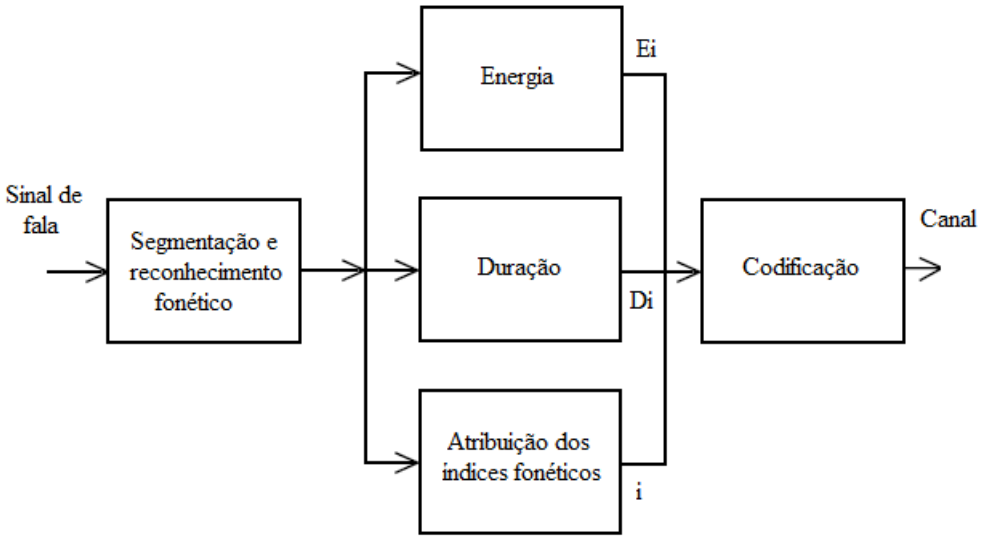

Figura 5.1: Diagrama de blocos do emissor do codificador.

O sistema de reconhecimento de fala desenvolvido neste trabalho é baseado nos Modelos de Markov Escondidos (HMM) e é classificado como independente de locutor, pois usa um banco de voz composto por sinais de fala de vários locutores para treinar seus modelos HMM e dependente de contexto, uma vez que leva em consideração fones à esqueda e à direita de um fone central (trifones). Todos os parâmetros utilizados no seu desenvolvimento são escolhidos por apresentar bons resultados no reconhecimento de fala [55, 56].

Como mencionado, as etapas da segmentação e reconhecimento de fonemas foram realizadas utilizando o *software* HTK. No entanto, a maioria das ferramentas do HTK não possui interface gráfica e a sua utilização é realizada a partir de linhas de comando, com o armazenamento de sequências de comandos e a criação de *scripts*, que neste trabalho foi feita em *bash*, para automatização dos procedimentos. Cada ferramenta possui um conjunto de parâmetros obrigatórios e permite a alteração do seu modo de funcionamento a partir de alguns parâmetros opcionais.

Antes de qualquer operação de processamento é necessário organizar os dados que permitem treinar e testar o sistema. Para o treino, eles são constituídos por um conjunto de sinais de voz e anexos arquivos de texto com suas respectivas anotações. Como não era objetivo do presente trabalho a construção de *corpora* de voz, ou seja, a gravação de um conjunto de sinais de fala e seus arquivos de texto, o *corpus* utilizado para o treinamento e reconhecimento dos HMMs foi obtido de [67]. O *corpus* é composto por 700 frases e 35 locutores com 20 frases cada, sendo 25 homens e dez mulheres, correspondendo a aproximadamente 54 minutos de sinal de fala. Os sinais de voz foram gravados com uma taxa de amostragem de 22050 amostras/s em computadores utilizando microfones comuns. O ambiente não foi controlado e há presença de ruído nas gravações.

Como a quantidade de sinais de voz necessária para o treinamento dos HMMs deve ser maior que a quantidade utilizada para o reconhecimento, o *corpora* de voz foi particionado em dois grupos. O primeiro grupo é composto de 500 frases, pronunciadas com 18 oradores do sexo masculino e sete oradores do sexo masculino, e é utilizado na etapa do treinamento. Para o reconhecimento, utiliza-

se o segundo grupo, com 200 frases, pronunciadas por dez oradores diferentes, sendo sete do sexo masculino e três do sexo feminino.

Como descrito no Capítulo 4, a primeira etapa a ser realizada para o reconhecimento fonético consiste no processamento do sinal de voz. Para esta etapa utilizou-se a ferramenta HCopy do HTK. Inicialmente, para a utilização desta ferramenta é necessário um arquivo de configuração que descreva todos os parâmetros desejados na análise do sinal de voz, como o tamanho e periodicidade dos quadros ao segmentar o sinal, o tipo de janela a ser utilizada, número de coeficientes MFCCs, valor do fator de pré-ênfase, entre outros.

Inicialmente a ferramenta HCopy realiza a pré-enfase do sinal de voz com o coeficiente 0,95, seguida da segmentação do sinal para análise a curtos intervalos. O sinal foi particionado utilizando uma janela de Hamming com 25 ms e deslocamento da janela em análise de 10 ms. Em seguida, inicia a etapa de extração de parâmetros do sinal de voz.

No processo de extração, cujo funcionamento está ilustrado na Figura 5.2, para cada arquivo de entrada é criado um outro correspondente contendo os vetores dos parâmetros. O número de coeficientes extraídos de cada janela do sinal de voz é uma decisão importante para o desempenho do sistema de reconhecimento. Quanto maior for o número de coeficientes melhor será descrito o sinal de voz e, consequentemente, após o treino, mais fiel será o modelo resultante. No entanto, o aumento da dimensão do vetor de características reflete no tempo de treino dos modelos e no número de iterações, que aumentam consideravelmente.

Neste trabalho foram extraídos de cada quadro do sinal de voz (25 ms) os coeficientes MFCCs, a energia e os coeficientes delta e aceleração como representação acústica. Para isso, inicialmente realiza-se uma FFT com 1024 pontos e 26 filtros triagulares. Assim, o vetor de características que representa cada janela do sinal de voz é composto de 13 parâmetros, dos quais 12 são coeficientes cepstrais e um é a energia. Desses parâmetros são extraídos a primeira e a segunda derivada, resultando em um vetor com 39 parâmetros, que são um LogEnergia, 12 MFCC, um ΔLogEnergia, 12 ΔMFCC, um $\Delta\Delta$LogEnergia e 12 $\Delta\Delta$MFCC. Para um maior número de coeficientes, o reconhecimento aumenta ligeiramente, sem que se notem melhorias consideráveis [58].

Depois de serem realizadas as etapas descritas, dá-se início ao desenvolvimento do modelo acústico.

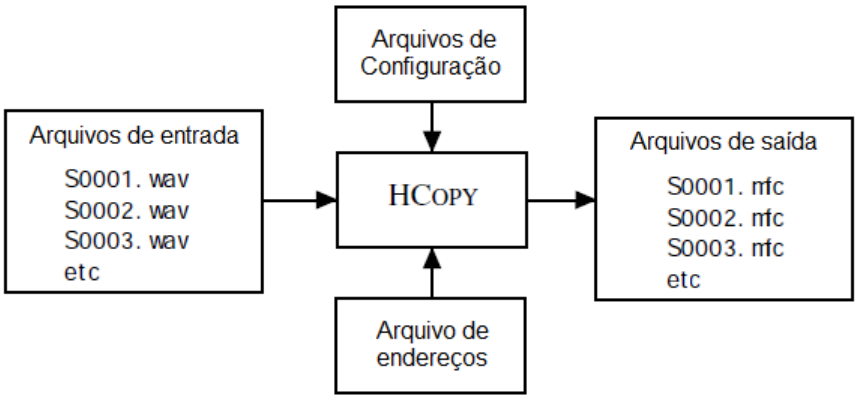

Figura 5.2: Funcionamento da ferramenta HCopy (Adaptado de [53]).

Treino do Modelo Acústico

O procedimento acústico tem como função construir um modelo matemático que represente cada fonema da língua a partir do vetor de características extraídas do sinal de voz, para inferir qual sequência de fonemas gera aquele vetor. O modelo acústico desenvolvido neste trabalho foi baseado nas instruções encontradas em [53] com alguns arquivos obtidos de [67], como dicionário fonético, lista de trifones, árvore de decisão e os *scripts* em Java.

O codificador utiliza modelos de HMM para dividir o sinal em segmentos fonéticos. A estratégia utilizada na construção de modelos acústicos é que os HMMs devem ser refinados gradualmente. Inicialmente, é necessário definir a topologia de cada HMM, ou seja, número de estados, forma das funções de observações e matriz de transição entre os estados. Na primeira fase, criam-se apenas HMMs protótipos, cujos valores especificados são ignorados sendo apenas aproveitada a sua presença como forma de definição da arquitetura. A exceção é feita aos valores das transições de estado que são analisados, mas que podem, inicialmente, serem idênticos, sendo a soma igual a um.

O protótipo definido para os modelos HMM consiste em cinco estados, dos quais três estados emissores de símbolos com uma determinada probabilidade de saída e dois estados não emissores, utilizando a estrutura *left-right* com saltos, como ilustrado na Figura 5.3. Esses parâmetros foram escolhidos por apresentarem bons desempenho no processo de reconhecimento [58, 55]. O sistema é constituído de quarenta modelos HMM, que representam cada um dos trinta e oito fonemas, além dos modelos que representam o silêncio e a pausa entre palavras.

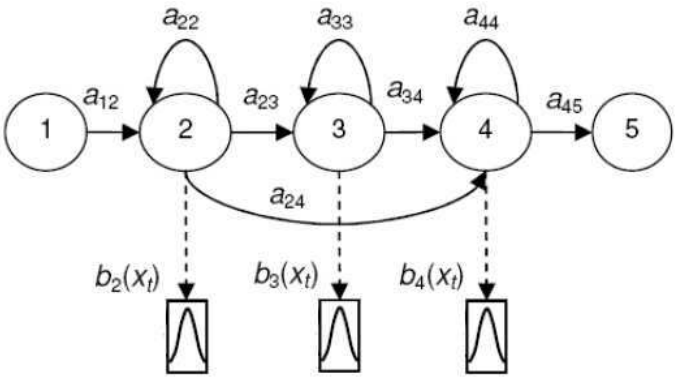

Figura 5.3: Topologia dos modelos de HMM [56].

Após a criação do protótipo, calcula-se as médias e as variâncias globais, utilizando a ferramenta HCompV, que são replicadas para todos os modelos HMMs criados. Assim, todos os modelos iniciais possuem estados (médias e variâncias) e matrizes de transição iguais.

Os modelos HMM são salvos no formato MMF (*Master Macro File*), formado por dois arquivos. O primeiro arquivo, denominado *hmmdefs* armazena todos os fones com os seus respectivos estados e matrizes de variância, enquando o segundo arquivo, *macros*, guarda o tipo paramétrico utilizado e a variável *vFloors* gerada pela ferramenta HCompV, que armazena o valor base de variância, que consiste em 0,01 vezes o valor da variância global.

Antes de iniciar o treinamento dos modelos de HMMs é necessário criar um arquivo denominado MLF (*Master Label File*) que contém todas as transcrições fonéticas dos arquivos de treino. Utiliza-se um *script* em Java, obtido de [67], que tem como entrada a lista de todos os arquivos

com transcrições ortográficas para criar um arquivo MLF com transcrição em nível de palavra. Em seguida, utiliza-se a ferramenta HLEd, associada ao dicionário fonético para realizar a conversão das palavras para sua respectiva representação fonética. Possuindo um conjunto inicial de valores para os modelos e o arquivo MLF, a ferramenta HERest é usada para realizar o treino dos modelos por meio do algoritmo de Baum-Welch.

Inicialmente, a ferramenta HERest segmenta os dados MFCC's uniformemente de acordo com o MLF, realiza a união dos HMMs e executa simultaneamente uma única re-estimação de Baum-Welch sobre todo o conjunto de modelos HMMs. Para as amostras de treino, os modelos correspondentes são concatenados e seguidamente é utilizado o algoritmo *forward/backward* para acumular valores estatísticos representativos para cada HMM na sequência. Após o processamento de todos os dados de treino as estatísticas acumuladas são utilizadas para re-estimar os parâmetros dos modelos.

Até esta etapa do treinamento do modelo acústico, foram criados modelos HMM para os fonemas e um modelo de silêncio, geralmente de longa duração, que se refere às pausas que ocorrem no final de uma frase. Entretanto, é necessário a inclusão de um modelo que represente as pausas curtas que ocorrem entre palavras em fala contínua, denominado de *short-pause* (SP). Este modelo é criado a partir do modelo HMM para o silêncio, fazendo uma cópia do estado central do modelo, seguida por um vínculo do estado central do modelo do silêncio com o estado emissor do modelo *short-pause*, de forma que, durante o treino, eles compartilhem os mesmos parâmetros, como ilustrado na Figura 5.4. Assim, o modelo HMM para o *sp* é composto de um estado emissor e dois não emissores. Desta forma, o modelo possibilita a não emissão de nenhuma observação, representado pela transição entre estados não emissores. Para facilitar a identificação do fim de uma palavra durante o reconhecimento é necessário incluir o *SP* no final de cada palavra contida no dicionário fonético.

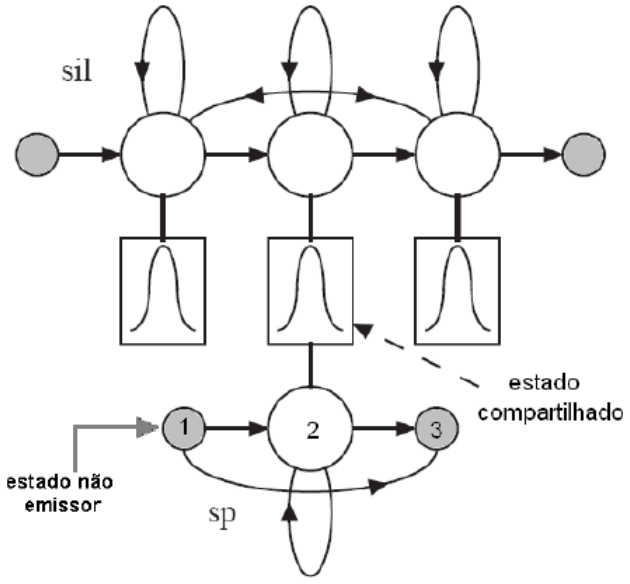

Figura 5.4: Modelo HMM do *short-pause* que compartilha parâmetros com o modelo do silêncio [55].

Utilizar um HMM por fonema supõe que um fonema seja seguido por qualquer outro fonema. Para considerar os efeitos contextuais causados pelas diferentes maneiras como os fonemas podem ser pronunciados é necessário adaptar o modelo acústico às características do trato vocal com a criação de modelos HMMs que representem cada um dos diferentes contextos de um fonema. A criação de modelos trifones é realizada com a ferramenta HLEd do HTK, a partir de uma adaptação dos modelos HMMs monofones, tornado os HMMs dependentes de contexto.

Para isto, é necessário antecipadamente um *script* contendo as modificações a serem realizadas para a expansão de modelos monofones para modelos trifones. Nos testes realizados foram utilizados trifones do tipo *cross-word*, com o objetivo de modelar melhor as transições entre palavras.

A ferramenta HLEd modifica o arquivo MLF utilizado para treino dos modelos de monofones, de forma a representar modelos dependentes de contexto, criando o MLF trifone e a lista de trifones. Em seguida, essa ferramenta clona todos os trifones e realiza o vínculo das matrizes de transição. Os trifones herdam os três estados do seu monofone central e todas as matrizes de transição são vinculadas para cada conjunto de trifones que possuem o mesmo estado central. Após estas modificações, os modelos passam pela reestimação de Baum-Welch com a ferramenta HERest.

Com o objetivo de obter uma estimação mais robusta dos parâmetros dos HMMs, é realizado na construção do modelo acústico um vínculo de estados dos trifones de forma a permitir o compartilhamento de dados.

Como mencionado, o método escolhido foi o vínculo por meio de árvore de decisão fonética. A árvore de decisão fonética utilizada neste trabalho foi obtida de [67].

Utiliza-se a ferramenta HHEd para aplicação da árvore. Tendo como entrada os arquivos de definição dos HMMs e o *script* contendo a árvore, a ferramenta tem como saída um arquivo contendo todos os trifones, juntos com os respectivos rótulos mostrando com quais outros trifones eles compartilham estados.

Modelos acústicos mais robustos possuem como etapa final a implementação de misturas gaussianas nos modelos HMMs. Entretanto, esta etapa não foi realizada no modelo acústico desenvolvido neste trabalho. Desta forma, o modelo acústico proposto apresenta apenas uma componente gaussiana por HMM.

5.2.1 Atribuição de Índices

Da etapa de segmentação e reconhecimento de fonemas obtém-se como saída uma sequência de segmentos fonéticos com seus respectivos tempos iniciais e finais.

Para cada fonema reconhecido é atribuído um índice pré-estabelecido, totalizando quarenta diferentes índices, sendo trinta e oito índices referentes aos fonemas do português brasileiro e dois índices utilizados para referenciar o silêncio e a pausa entre palavras. O índices são atribuídos como apresentado na Tabela 5.1.

Além dos índices dos segmentos fonéticos, deve estar definidos no emissor os índices que são atribuídos a todos os segmentos acústicos que serão posteriormente utilizados na síntese do sinal de voz, como as sílabas e os encontros vocálicos, embora esses índices não sejam transmitidos. Os índices atribuídos às sílabas e aos encontros vocálicos são construídos a partir dos fonemas que os

formam. Essas definições são necessárias para que o emissor do codificador possa calcular a energia e duração correta de cada segmento, como explicado a seguir.

Tabela 5.1: Índices atribuidos aos fonemas.

Fonemas	Índices	Fonemas	Índices
p	1	l	21
b	2	w	22
t	3	λ	23
d	4	i	24
k	5	$\tilde{i}$	25
g	6	e	26
t$\int$	7	$\tilde{e}$	27
dZ	8	é	28
f	9	a	29
v	10	$\tilde{a}$	30
s	11	ó	31
z	12	o	32
$\int$	13	$\tilde{o}$	33
Z	14	u	34
x	15	$\tilde{u}$	35
R	16	I	36
m	17	i~	37
n	18	ê	38
$\tilde{y}$	19	sil	39
r	20	sp	40

5.2.2 Estimação da Energia

A estimação da energia é a próxima etapa realizada pelo emissor do codificador após o reconhecimento de fonemas. A energia do sinal de voz está concentrada na região de frequências mais baixas do espectro e, para sinais que possuem valor médio nulo, como é o caso dos sinais de fala, a energia pode ser definida como a média do quadrado dos valores das amostras. Deste modo, a energia media para cada segmento acústico pode ser obtida por

$$E = \frac{1}{N}\sum_{n=1}^{N} x^2(n), \tag{5.1}$$

em que *x(n)* representa amostras do sinal de voz e *N* a quantidade de amostras em cada fonema.

Os segmentos acústicos utilizados pelo codificador de voz consistem nas sílabas, encontros vocálicos e fonemas. Assim, para calcular a energia, o emissor seleciona três índices por vez e busca se há algum segmento acústico com este índice. Se sim, o emissor deve calcular e enviar a energia deste segmento por meio das amostras contidas no intervalo de tempo dos três fonemas que formam

tal segmento. Se não, o emissor seleciona os dois primeiros índices do conjunto anteriormente selecionado, e novamente busca se há algum segmento com este índice. Isso porque as síbalas podem ser formadas por dois ou três fonemas. Caso não encontre nenhum segmento silábico ou encontro vocálico, o emissor calcula a energia de cada fonema isoladamente.

Como exemplo, considere a sequência de símbolos 5, 29 e 1, que corresponde a **k a p**. Inicialmente, o emissor busca uma sílaba ou encontro vocálico pré-definido com o íncide 5291. Como não há, ele seleciona os dois primeiros índices do conjunto, 5 e 29 e novamente faz uma busca de sílabas ou de encontro vocálicos. Caso a síbala **ca**, cuja transcrição fonética é **k a** e índice atribuído 529, uma vez que 5 corresponde ao índice do fonema **k** e 29 ao índice do fonema **a**, esteja pré-definida, o emissor calcula a enegia da sílaba, por meio das amostras contidas no instante inicial do fonema **k** ao instante final do fonema **a**.

A próxima janela de índices a ser analisada é formada pelo deslizamento de fator um para direita, obtendo um índice posterior ao último índice analisado na janela anterior.

5.2.3 Estimação da Duração

Estimar a duração de cada fonema é fundamental para um bom desempenho do codificador. A duração de cada fonema é obtida na etapa do reconhecimento de fonema, que fornece a fala segmentada com o tempo inicial e final de cada fonema.

Como mencionado, a síntese do sinal de fala é feita com o uso de sílabas, encontro vocálicos e fonemas. Para obter a duração correta de cada segmento acústico, a etapa que estima a duração dos segmentos utiliza o mesmo esquema de análise dos índices que a etapa de estimação de energia. Ou seja, por meio de, inicialmente, um conjunto de três índices, verifica se há alguma sílaba pré-definida com este conjunto de índices. Se não, realiza uma nova busca com os dois primeiros índices do conjunto. Caso encontre alguma sílaba ou encontro vocálico com este conjunto de índices, a duração é estimada por meio da subtração do tempo final do último fonema do tempo incial do segundo fonema.

Voltando ao exemplo anterior, e considerando o fonema **k** inicie no tempo t_i e que o fonema **a** termine no tempo t_f, ao identificar que o índice 529 pertence à sílaba **ca**, o emissor estima a duração desta sílaba, dada por $t_d = t_f - t_i$.

5.2.4 Codificação de Huffman

As informações obtidas do emissor são codificadas com o código de Huffman, escolhido por apresentar melhor desempenho em aplicações estatísticas em relação a outros tipos de codificação [68].

O método de Huffman codifica a partir da ordenação decrescente das frequências de ocorrência dos símbolos, construindo uma árvore estritamente binária (Árvore de Huffman), base para a codificação e decodificação.

A árvore binária é construída recursivamente a partir da junção dos dois símbolos de menor probabilidade, que são então somados em símbolos auxiliares e estes recolocados no conjunto de símbolos. O processo termina quando todos os símbolos foram unidos em símbolos auxiliares, com a probabilidade final unitária, formando uma árvore binária. A árvore é então percorrida, atribuindo-se valores binários de 1 ou 0 para cada aresta, e os códigos são gerados a partir desse percurso.

O processo de Huffman é baseado em duas observações. A primeira é a que os símbolos que ocorrem com maior frequência terão uma palavra - código menor em relação aos símbolos que ocorrem com menor frequência. A segunda observação consiste nos dois símbolos com menor frequência de ocorrência serem codificados com um código com mesmo comprimento.

A Figura 5.5 ilustra um exemplo da árvore de Huffman que forma um código para cinco fonemas distintos. A parte da árvore em que os códigos estão apresentados é denominada raiz da árvore, enquanto a outra extremidade é denominada folhas da árvore de Huffman. As frequências de ocorrência das vogais apresentadas na tabela, assim como as frequências de ocorrência dos fonemas codificados foram obtidos de [69].

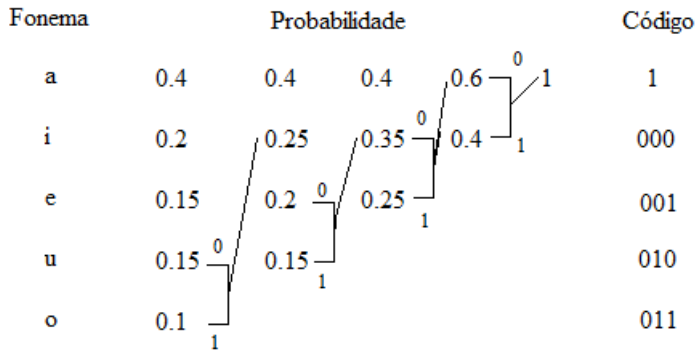

Figura 5.5: Exemplo de codificação de vogais com o código de Huffman.

5.3 Receptor

O receptor do codificador proposto tem a função de converter a sequência de índices fonéticos em um sinal acústico. Para isto, realiza uma síntese por concatenção que gera o sinal de fala a partir da justaposição de segmentos pré-gravadados. Esses segmentos são selecionados a partir de um inventário de unidades previamente construído. Deste modo, a implementação do receptor está dividida em duas etapas: aquisição de um banco de unidades e síntese do sinal de voz.

5.3.1 Primeira Etapa – Obtenção do Banco de Unidades

Esta etapa tem o objetivo de compor bancos de unidades, formados por um conjunto de segmentos obtidos do sinal de fala que são posteriormente utilizados na síntese por concatenação.

Em uma síntese por concatenação, a escolha do tamanho das unidades a serem utilizadas no processo de síntese é uma das decisões mais importantes, pois deve representar um compromisso entre inteligibilidade e naturalidade requerida. Várias são as possibilidades de tamanhos e quantidades que podem ser utilizadas.

Um dos segmentos que podem ser usados na síntese por concatenação são os difones, unidades formadas por uma dupla de fones. Ele inicia na metade do primeiro fone e termina na metade do fone seguinte. Sua vantagem consiste em conter inteiramente as transições entre os fones. No entanto, os difones incluem apenas parte dos vários efeitos coarticulatórios da língua falada, o que justifica o uso, mesmo que parcial, de unidades maiores, como os trifones.

Os trifones são segmentos que incluem um fone inteiro e suas transições à esquerda e à direita. Entretanto, devido à grande quantidade de trifones presentes na língua portuguesa, essas unidades são utilizadas como um complemento, para casos de sons especiais, de bancos baseados em unidades menores [70, 71].

Outras unidades que podem ser utilizadas na síntese por concatenação são as metades dos fones, sílabas, demissílabas, palavras e fonemas. A metade dos fones, se estende desde a fronteira entre fones até o ponto médio, ou se estende a partir deste ponto médio até o final do fone. Entretanto, essa unidade quando utilizada de forma isolada apresenta dificuldade de representação da coarticulação.

As sílabas podem ser consideradas unidades naturais, uma vez que apresentam a coarticulação entre os fonemas que as formam e são mais importantes que as coarticulações presentes nos segmentos intra-sílabas. No entanto, a ausência dessas coarticulações diminui a qualidade do sinal sintetizado. Outra desvantagem desses segmentos é a sua grande quantidade na língua portuguesa, o que dificulta a construção de um banco utilizando esse tipo de segmento [72].

Com base nos mesmos princípios fonológicos das síbalas, as demissílibas são formadas a partir da divisão das sílabas em duas partes parcialmente sobrepostas, com o pico silábico (núcleo) pertencendo a ambas as partes. Como um exemplo, considere a sílaba *tar*, que possui uma demissílaba inicial *ta* e uma demissílaba final *ar*. Uma desvantagem do uso desse tipo de segmento é que nem sempre é possível desprezar a interação que ocorre entre os segmentos pertencentes a sílabas diferentes [73, 74].

Além das unidades de concatenação expostas, a síntese de uma sinal de fala também pode ser realizada com palavras ou frases. A desvantagem dessas tipo de unidades é o elevado número necessário em um sistema de síntese irrestrita, ou seja, aquela síntese que não é restrita a um conjunto de palavras ou frases.

Por fim, mas não menos importante, a síntese por concatenação pode ser realizada utilizando segmentos fonéticos. Sua vantagem consiste na pequena quantidade de unidades presentes na língua portuguesa, sendo obtidas com a utilização de um pequeno banco de voz. Entretanto, a síntese utilizando estas unidades apresenta um comportamento não muito estável oscilando entre falas sintetizadas com uma alto grau de naturalidade e falas sintetizadas com distorções desagradáveis [75]. Isso ocorre pelo fato de que os pontos de coarticulação passam a ser realizados nas fronteiras dos fones, dificultando a representação precisa do efeito de coarticulação, o que requer várias amostras de uma mesma unidade em diferentes contextos (alofones) [74].

Em uma síntese por concatenação é necessário levar em consideração a variação à qual as unidades de concatenação estão sujeitas de acordo com a posição ocupada dentro de uma frase ou com a entoação aplicada. Por exemplo, no caso da palavra **casa**, a pronúncia do primeiro fonema **a** é diferente do segundo fonema **a**. Assim, para manter uma entoação correta e natural, seria necessário considerar todas as variantes do fonema **a** como uma unidade de concatenação [74].

Diante das considerações feitas sobre as unidades de concatenação e das restrinções de base voz para a formação do inventário impostas pelo projeto do codificador, foi decidida a utilização algumas unidades de concatenação que podem ser obtidas no banco de voz, ou seja, sílabas e variações de fonemas. Além disso, para aumentar a qualidade do sinal sintetizado, o sistema também faz o uso de encontro vocálicos, para garantir as coarticulações entre vogais, considerados importantes para uma síntese de boa qualidade no desenvolvimento deste trabalho, ao se concatenar fonemas. O uso de outras unidades de concatenação que poderiam ser obtidas do banco de voz, como difones e trifones, não foi feito devido à presença de poucos desses segmentos presentes no banco de voz em relação à quantidade existente na língua portuguesa.

Na implementação do codificador, os bancos de unidades são formados mediante a pronúncia de frases que podem ser gravadas diretamente em cada aparelho telefônico ou enviadas a outros aparelhos. Ou seja, ao optar por utilizar o codificador, cada usuário pode gravar frases pré-estabelecidas em seu próprio aparelho telefônico e em seguida enviá-las para aparelhos telefônicos de outros usuários, ou gravá-las diretamente no aparelho dos usuários com quem se pretende fazer a comunicação utilizando o codificador.

O projeto do codificador inclui um banco de unidades específicos para cada usuário cadastrado na agenda do aparelho telefônico, como ilustrado na Figura 5.6, em que N_b representa a quantidade de utilizadores de codificador cujo banco de unidades está gravado em um determinado aparelho, atribuindo uma característica pessoal ao codificador desenvolvido. Assim, no uso do codificador, o sistema, ao identificar o usuário que está solicitando uma comunicação, realiza uma busca do banco de unidades específico daquele usuário, eliminando a necessidade de etapas de adaptação ao orador.

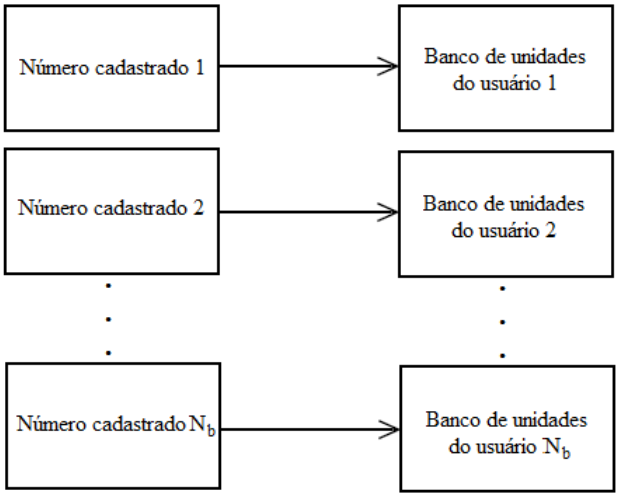

Figura 5.6: Bancos de unidades de cada usuário do codificador.

Para a aquisição dos fonemas, encontros vocálicos e sílabas, as frases das listas número 11 e 16 de [69], apresentadas no Anexo C, foram selecionadas por conterem um maior número de fones e consequentemente uma maior variabilidade deles. Foi observado no desenvolvimento deste trabalho que a variação em diferentes contextos das vogais é mais importante para uma boa qualidade do sinal sintetizado do que a variação das demais unidades fonéticas.

Neste caso, são consideradas algumas variações para cada vogal e a utilização de cada uma delas está de acordo com a posição que ocupa em cada frase. Ou seja, de cada vogal, há uma unidade para ser utilizada, caso seja o primeiro fonema da palavra, outra variação, caso a vogal esteja localizada no meio de uma palavra e, por fim, uma terceira variação, caso o fonema seja o último da palavra. No entanto, a síntese faz uso de pouca ou nenhuma variação de sílabas ou encontros vocálicos, umas vez que não foi possível a sua obtenção no banco de voz.

Inicialmente, o projeto do receptor do codificador incluía uma etapa de segmentação automática, cuja função é partir as frases utilizadas para compor o banco de unidades em fonemas, encontros vocálicos e sílabas. No entanto, devido à taxa de erro obtido no sistema de reconhecimento de fonemas (20%), a duração obtida na etiquetagem deses segmentos não corresponde à duração exata de cada um deles e, consequentemente, as amostras extraídas do sinal de voz neste intervalo de tempo obtido não correspondem de fato às amostra de um determinado segmento.

Por exemplo, na segmentação automática da palavra **casa**, cuja transcrição fonética é **k a z a**, as amostras do segundo fonema da palavras, ou seja, do primeiro **a**, extraidas no intervalo de tempo obtido na segmentação, podem pertencer aos fonemas **k a** ou aos fonemas **a z**. Neste caso, ao colocar essas amostras em outras palavras, o sinal seria sintetizado com distorções pela presença de amostras que não pertencem ao fonema **a**.

Uma vez que a qualidade da fala sintetizada está diretamente relacionada à qualidade do banco de unidades, ou seja, ao correto inventário de unidades, à qualidade do sinal gravado e ao recorte das unidades, foi decidido utilizar a segmentação manual com o objetivo de aumentar a qualidade do sinal sintetizado, pela minimização dos erros ocasionados na segmentação.

A segmentação manual foi realizada por meio do *software* Audacity 2.0.0®. Cada frase foi pronunciada de forma individual e em seguida particionada. Os sinais de voz foram gravados utilizando uma taxa de amostragem de 22050 amostras/s, e armazenados no formato *Waveform Audio Format* (wav). A Figura 5.7 ilustra a interface do Audacity 2.0.0® exemplificando um exemplo de segmentação de uma palavra contida em uma das frases segmentadas. Na figura é ilustrada a forma de onda da palavra **casa** e as delimitações dos seus fonemas, em que é possível observar a exclusão das transições entre fonemas.

Após a etiquetagem manual de todos os segmentos requeridos para a síntese do sinal de voz, os mesmos índices apresentados na Tabela 5.1 são atribuídos aos fonemas. Para o caso dos segmentos de sílabas e encontros vocálicos, os índices a eles atribuídos são determinados de acordo com os fonemas que os compõem, como descrito na próxima sessão.

5.3.2 Segunda Etapa: Síntese do Sinal de Voz

Após a construção do banco de unidades, o receptor faz seu uso para a síntese por concatenação. O diagram de blocos do receptor está ilustrado na Figura 5.8. Sua primeira função consiste

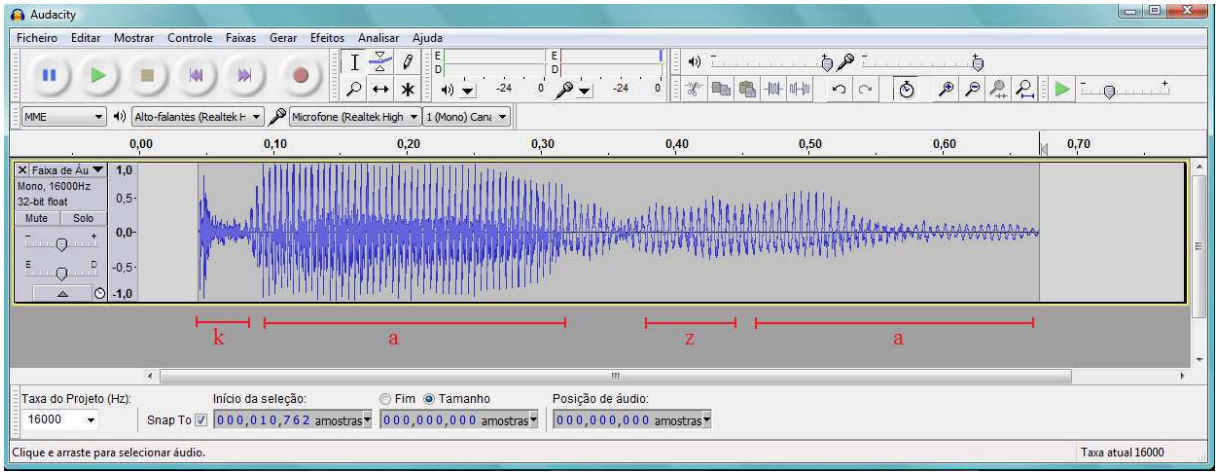

Figura 5.7: Interface do Audacity

na decodificação de Huffman, obtendo as informações dos índices fonéticos, energia e duração de cada segmento.

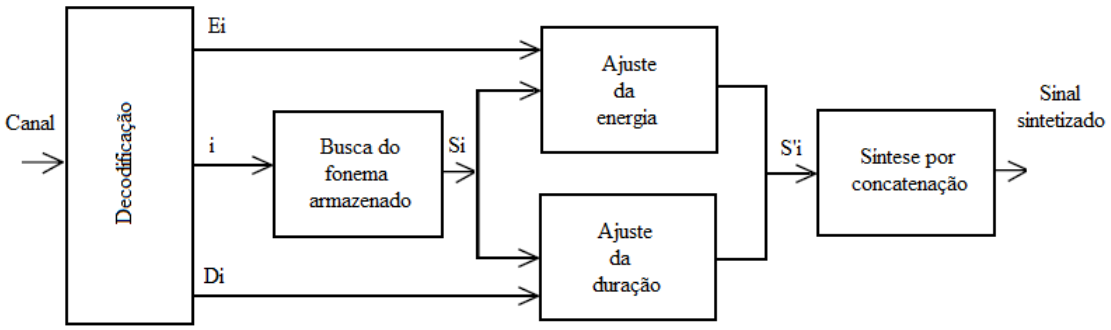

Figura 5.8: Diagrama de blocos do receptor do codificador.

Por meio dos índices dos segmentos fonéticos, o receptor realiza a busca dos segmentos armazenados no banco de unidades. Para cada segmento selecionado na busca, é feito um ajuste da energia média do sinal, realizado por meio de uma relação da energia média do segmento armazenado com a energia recebida. Para o ajuste da duração de cada sílaba e ditongo, é implementada uma interpolação linear e, por fim, os segmentos são concatenadas para a formação do sinal sintetizado. As etapas são descritas a seguir.

Decodificação de Huffman

A decodificação é realizada percorrendo inversamente a árvore de Huffman. Assim, os *bits* recebidos são utilizados para percorrer a árvore de Huffman da raiz até alguma folha, quando se obtém o símbolo decodificado.

Busca do Fonema Armazenado

Como mencionado, o codificador proposto tem a característica de ser pessoal. Isso parte do princípio de que o receptor utiliza um banco de unidade distinto para cada usuário do codificador.

Ao receber uma solicitação de chamada, os aparelhos celulares são programados para identificar qual usuário cadastrado na agenda telefônica está solicitando uma comunicação. Ao realizar esta identificação, o receptor do codificador faz uma busca do banco de unidades daquele determinado usuário. Ao identificar o banco de unidades, o sistema realiza a síntese com os segmentos específicos daquele usuário.

Como mencionado, na etapa da construção do banco de unidades cada segmento fonético é etiquetado utilizando os mesmos índices da etapa de atribuição de índices no emissor do codificador, ou seja, os mesmos índices apresentados na Tabela 5.1. No entanto, além dos segmentos fonéticos, o banco de unidades também é composto por encontros vocálicos e sílabas. Para a etiquetagem desses segmentos, o sistema utiliza os índices correspondentes aos fonemas que os formam. Por exemplo, na etiquetagem do ditongo **au**, o sistema atribui o índice 2934, uma vez que o índice 29 corresponde ao fonema **a**, e o índice 34 corresponde ao fonema **u**.

Os índices dos segmentos fonéticos obtidos após a decodificação de Huffman são utilizados na seleção dos segmentos armazenados no banco de unidades para a realização da síntese. Deste modo, ao identificar o banco de unidades, o receptor, para identificar se o segmento consiste em um fonema, encontro vocálico ou sílaba, inicialmente faz a identificação de três índices conjuntamente para identificar uma possível sílaba. Caso o receptor não identifique uma sílaba por meio desses três índices selecionados, ele realiza a busca de sílaba e encontros vocálicos com os dois primeiro índices recebidos. Ainda não encontrando nenhum segmento com tal índice, realiza a busca do fonema correspontende a cada índice isoladamente.

Como exemplo, considere que o receptor tenha recebido os índices 29 e 34. Inicialmente, ele busca o encontro vocálico ou sílaba indexado por 2934, neste caso, o ditongo **au**. No caso da não identificação de um encontro vocálico ou sílaba com este índice, o receptor realiza a busca do fonema 29, ou seja, do fonema **a** e, em seguida, do fonema 34, ou seja, do fonema **u**.

Como dito, em uma síntese por concatenação é necessário considerar a variação das unidades de concatenação de acordo com a posição ocupada dentro de uma frase ou com a entoação aplicada. Neste trabalho, devido à restrição do banco de voz utilizado para compor o banco de unidades, são consideradas apenas algumas variações de vogais.

Para se conseguir a maior quantidade de coarticulações possíveis, o receptor atribui níveis de privilégios aos segmentos do banco de unidades. Deste modo, sua função é fazer inicialmente a busca de sílabas e encontros vocálicos. Caso esses segmentos não sejam encontrados, o receptor utiliza fonemas para gerar o sinal de saída.

No caso da utilização de fonemas, o receptor é capaz de identificar se o segmento fonético corresponde a uma vogal ou uma consoante, por meio dos índices pré-estabelecidos. Ao reconhecer o índice de uma vogal, o receptor seleciona a variação mais apropriada para a síntese de acordo com a posição que a vogal ocupa na frase.

Para isso, o receptor observa os índices precedentes e posteriores aos índices correspondentes às vogais. Caso o índice anterior ao índice da vogal seja o íncide atribuído ao silêncio ou

short-pause, o sistema seleciona a variação da vogal apropriada para o início de uma palavra. Caso os índices à esquerda e à direita do índice da vogal sejam atribuídos a segmentos quaisquer, a variação pré-estabelecida para ser utilizada no meio de uma palavra é selecionada. No caso em que o índice posterior à vogal seja atribuído ao silêncio ou *short-pause*, a variação escolhida será aquela utilizada no final de uma palavra.

Ajuste da Energia

Após a seleção da unidade de concatenação, a próxima função do receptor consiste no ajuste da energia, para que os segmentos contidos no banco de unidades sejam adaptados aos segmentos pronunciados pelos oradores.

A informação recebida pelo codificador consiste na energia média de cada sílaba, encontro vocálico ou fonema. A implementação do ajuste da energia é feita por meio de uma relação entre a energia média de cada segmento pronunciado e a energia média do segmento presente no inventário de unidades multiplicada pela energia no segmento armazenado.

Matematicamente, a energia do segmento para a síntese é

$$E(t) = \left(\frac{\bar{E_p}}{\bar{E_a}}\right) \times E_a(t) \tag{5.2}$$

em que $\bar{E_p}$ é a energia do segmento pronunciado cuja valor é recebido pelo receptor e $\bar{E_a}$ é a energia do segmento armazenado.

As Figuras 5.9 e 5.10 ilustram exemplos do ajuste da energia média realizado pelo receptor, em que a forma de onda do fonema **a** é apresentada com energias médias distintas, neste caso a segunda forma de onda possui metade da energia média da primeira forma de onda.

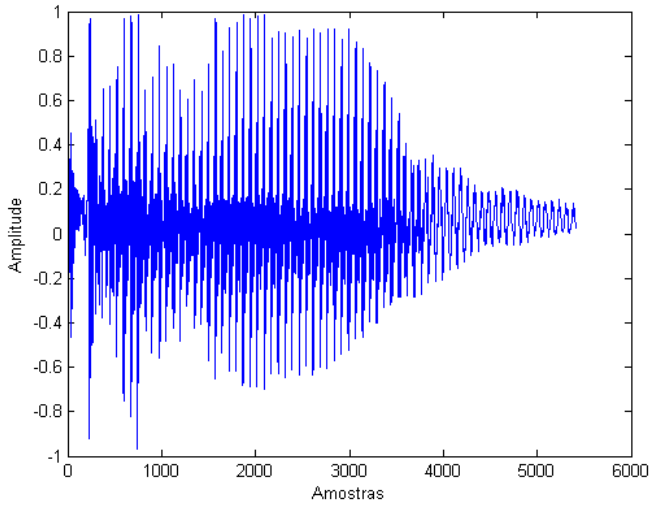

Figura 5.9: Forma de onda do fonema **a**.

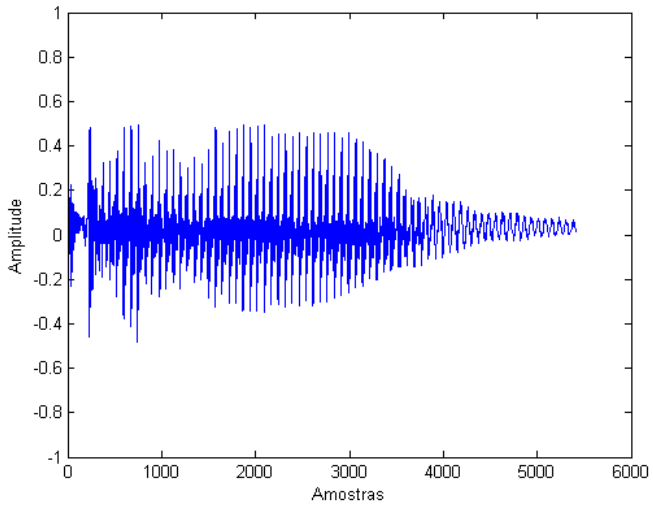

Figura 5.10: Forma de onda do fonema **a** com energia reduzida.

Ajuste da Duração e Concatenação de Segmentos

O ajuste da duração é a próxima função do receptor. A duração de cada segmento é ajustada de acordo com as novas informações recebidas do emissor do codificador.

Para o ajuste da duração e concatenação de segmentos é implementada uma interpolação linear. As informações de duração são recebidas na forma de unidades de milisegundo (ms). Inicialmente, a duração recebida é convertida em quantidade de amostras, fazendo a multiplicação da duração em milisegundo pela taxa de amostragem (22050 amostras/s).

Considere a Figura 5.11, em que as formas de onda em azul e vermelho representam segmentos armazenados, com quantidades de amostras NA_1 e NA_2, respectivamente. A duração recebida, convertida em quantidade de amostras, é representada por NA_{d1} para o segmento em azul. A interpolação linear realiza a concatenação entre segmentos interpolando as últimas NA_i amostras do primeiro segmento com as primeiras NA_i amostras do segundo segmento, em que $NA_i = NA_1 - NA_{d1}$, sendo que NA_{d1} consiste na quantidade de amostras que representa a duração do segmento recebido pelo receptor e a sigla NA_i significa a quantidade de amostras utilizada na interpolação.

A interpolação inicia com as últimas amostras dos segmentos usados com a atribuição de pesos diferentes a cada uma e, em seguida, é feita a soma dessas componentes. Neste caso, a amostra pertencente ao segundo segmento possui um peso muito maior que a última amostra do primeiro segmento. Assim, forma-se a última amostra do segmento resultante da interpolação, que pode ser interpretado como um segmento de transição entre os dois segmentos acústicos que estão sendo concatenados. O peso maior atribuído à última amostra do segundo segmento significa que a amostra reultante da interpolação terá maior influência do segundo segmento.

Seguindo essa sequência, as amostras dos segmentos são interpoladas uma a uma e o peso atribuído a cada uma delas varia de acordo com a proximidade que a amostra resultante da interpolação tem com os segmentos que estão sendo concatenados, de forma que a amostra central do segmento interpolado tenha uma contribuição igual das amostras usadas na interpolação.

A interpolação segue a Equação 5.3, em que $i = 1,2,\ldots,NA_i$, k varia de $1/NA_i$ a 1, com incremento de $1/NA_i$ e $y_{int} = NA_i, NA_i - 1, \ldots, 0$. Desta forma, há um ajuste na duração do segmento armazenado em função de NA_1, assim como uma concatenação entre segmentos.

$$A(NA_1 - i) = NA_1(NA_1 - i) \times k + NA_2(y_{int}) \times (1 - k). \tag{5.3}$$

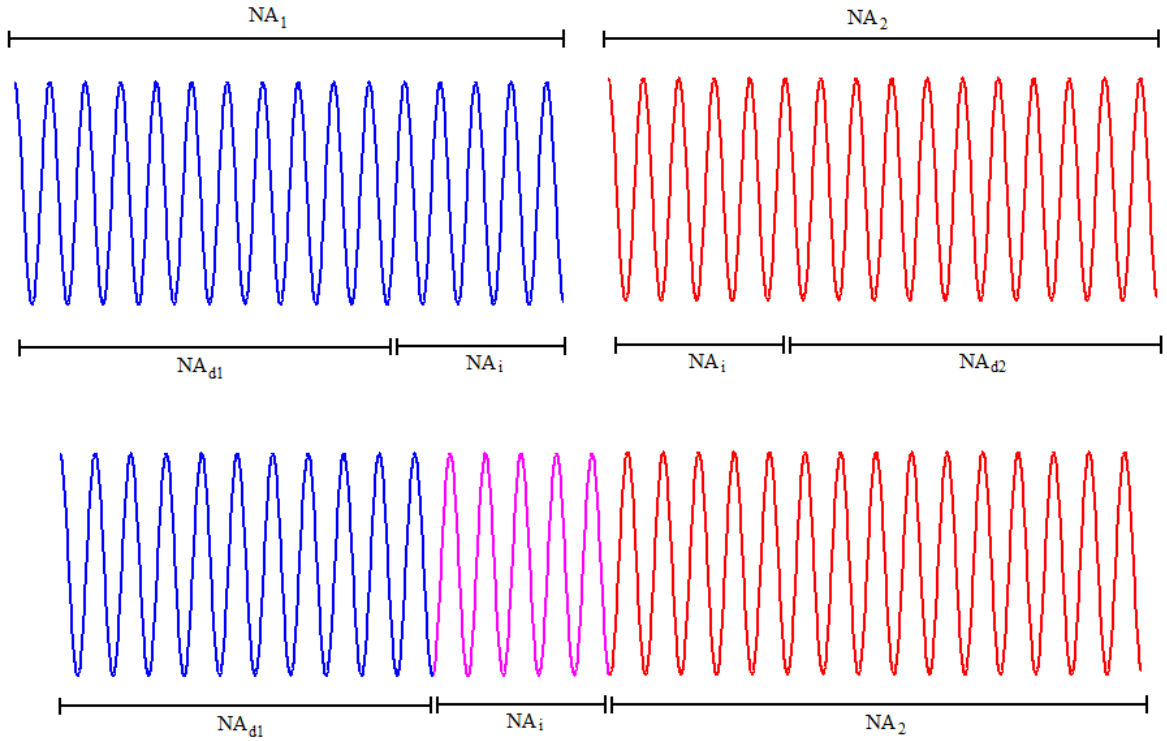

Figura 5.11: Exemplo de interpolação – Ajuste da duração e concatenação de segmentos.

As Figuras 5.12 e 5.13 ilustram a forma de onda da palavra **casa** com e sem interpolação linear entre as sílabas. Deste modo, é possível observar uma diminuição da primeira sílaba (**k a**), além de uma diminuição no tempo total da palavra. O ajuste de duração na segunda sílaba (**z a**) é realizado com as amostras contidas no intervalo de tempo em que há silêncio ou pausa entre palavras.

Síntese por Concatenação

A síntese concatenativa é caracterizada por produzir o som sintetizado por meio da concatenação de segmentos correspondentes a unidades acústicas, previamente gravadas e armazenadas em um banco de unidades. O sintetizador do codificador proposto pode ser representado pelo diagrama de blocos apresentado na Figura 5.14 [76].

Esta síntese é utilizada no codificador de voz por ter a vantagem de o processamento do sinal de fala ser feito com a própria forma de onda, mantendo assim as características originais do sinal, sobretudo o timbre, sem a necessidade de qualquer método de adaptação da voz sintetizada a voz dos oradores.

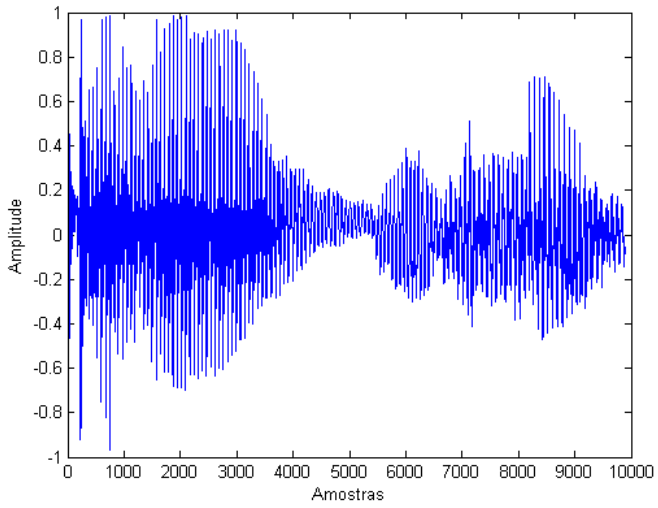

Figura 5.12: Forma de onda da palavra **casa** sem interpolação linear entre síbalas.

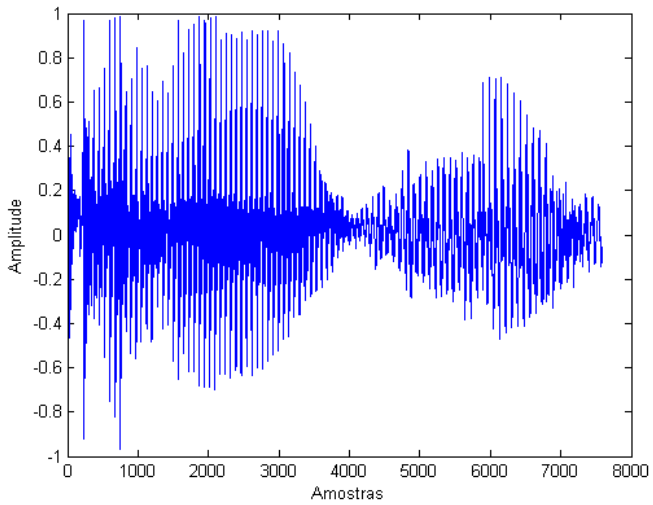

Figura 5.13: Forma de onda da palavra **casa** com interpolação linear entre síbalas.

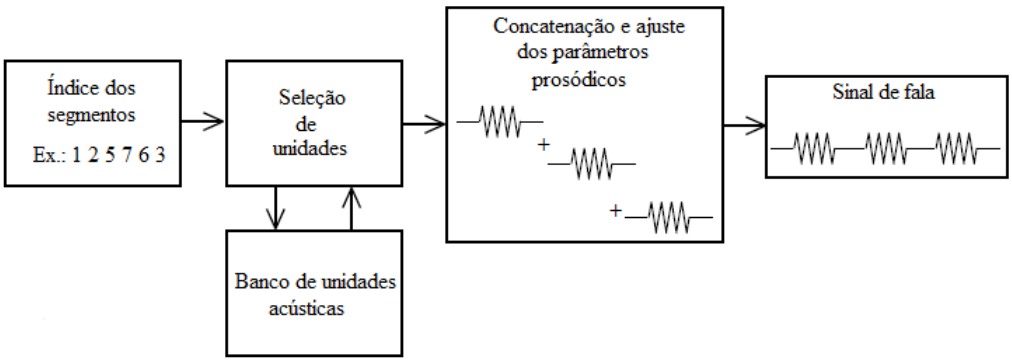

Figura 5.14: Diagrama esquemático do sintetizador concatenativo do codificador de voz.

No entanto, algumas desvantagens podem ocorrer na síntese concatenativa: descontinuidades no envelope espectral, descontinuidades de amplitude, de *pitch* e de fase entre os segmentos. As

descontinuidades espectrais ocorrem quando os formantes de segmentos adjacentes não têm os mesmos valores e estão relacionadas, principalmente, à coarticulação. Esse problema pode ser atenuado com a suavização das bordas dos segmentos [76, 71, 77].

Particularmente no caso da síntese por concatenação do codificador de voz, o sinal de fala é obtido por meio das seguintes etapas:

1ª Etapa – Obenção dos índices dos segmentos fonéticos com a decodificação de Huffman.

2ª Etapa – Busca das unidades acústicas presentes no banco de unidades que correspondem aos índices recebidos pelo receptor. Inicialmente, o receptor analisa conjuntamente três índices e procura selecionar uma sílaba. Caso não encontre, ele seleciona os dois primeiros índices e faz uma nova busca por sílabas ou encontros vocálicos. Não os encontrando, são selecionadas as unidades acústicas que correspondem a cada índice isoladamente.

Neste caso, como os índices dos segmentos acústicos são pré-estabelecidos, o receptor é capaz de reconhecer se o fonema selecionado corresponde a uma vogal ou consoante. Ao indentificar uma vogal, o receptor deve selecionar aquela variação da vogal mais adequada para a síntese, de acordo com o posicionamento que ela ocupa na palavra, que pode ser o primeiro fonema da palavra, um fonema intermediário ou o último fonema da palavra.

Para isso, o receptor observa os índices à esquerda e à direita do índice analisado. Caso à esquerda do índice do fonema o receptor identifique o índice pré-estabelecido ao silêncio ou pausa entre palavras (*short-pause*), ele seleciona aquela variação definida para ser usada com o primeiro fonema de uma palavra. No caso dos índices adjacentes ao índice analisado serem pertencentes a outros segmentos quaisquer, o receptor deve selecionar a variação pré-estabelecida para seu usada no meio de uma palavra.

Caso nenhuma dessas análises tenha sido satisfeita, o receptor analisa o índice à direita do analisado. Se esse índice for o do silêncio ou (*short-pause*), o receptor entende que aquele fonema que esta sendo analisado está posicionado como o último da palavra e seleciona a variação estabelecida para este caso.

3ª Etapa – Ao selecionar o segmento acústico, o receptor realiza o ajuste da energia por meio da relação entre as energias médias do segmento pronunciado e do segmento armazenado.

4ª Etapa – Tem a função de ajustar a duração do segmento acústico conforme a descrição prosódica, além de fazer a concatenação para a formação de palavras.

5ª Etapa – As palavras são concatenadas para a formação das frases.

5.4 Considerações Finais

Este capítulo descreve o projeto do codificador de voz proposto neste trabalho. O codificador é do tipo fonético e reduz a taxa de transmissão realizando a quantização de parâmetros que representam o sinal de voz.

O emissor do codificador é composto por um sistema de reconhecimento de fala que tem o objetivo de obter uma sequência de fonemas com seus respectivos tempos iniciais e finais, necessários para se obter características prosódicas como energia e duração, sendo também possível obter estas informações de sílabas e encontros vocálicos. A cada fonema atribui-se um índice pré estabele-

cido e, juntamente com as informações de energia e duração, quantiza-se esses parâmetros que são enviados ao receptor do codificador.

Para manter as características da voz de cada usuário, o receptor do codificador realiza a síntese por concatenação de unidades acústicas armazenados no banco de unidades de cada utilizador do sistema, atribuindo uma característica pessoal ao codificador proposto. Para compor o banco de segmentos acústicos, frases pré-estabelecidas são segmentadas, a princípio, por um sistema de reconhecimento de fonemas e, em seguida, de forma manual.

O esquema do codificador proposto permite a transmissão do sinal de voz com baixa taxa de *bits*. No entanto, o seu bom desempenho depende de um bom sistema de reconhecimento de fonemas, da qualidade da gravação e da correta escolha dos segmentos acústicos para a realização da síntese.

CAPÍTULO 6

Resultados

Para avaliar o desempenho do codificador proposto, foi realizado um teste subjetivo informal baseado no teste subjetivo ACR (*Absolute Category Rating*), da recomendação P.800 do ITU-T [4].

O método de avaliação subjetiva ACR consiste em uma metodologia de estímulo único, ou seja, os sinais de voz processados pelo codificador são apresentados um por vez e, após cada apresentação, os ouvintes-avaliadores classificam subjetivamente a qualidade do sinal processado. A opinião de cada ouvinte sobre a qualidade absoluta da voz e sobre o esforço exercido para a compreensão da fala é expressa em uma escala de opnião que varia entre um e cinco.

O teste ACR utiliza três escalas de opnião: qualidade de audição, esforço de audição e preferência de sonoridade.

A qualidade de audição é aferida por um sistema de pontuação que define a qualidade de pequenos grupos de sentenças descorrelacionadas, cada uma submetida ao processo de teste. O MOS é calculado pela média dos resultados individuais. A Tabela 6.1 apresenta este esquema de avaliação na coluna "Qualidade".

O esforço de audição avalia os níveis de degradação. Neste caso, preocupa-se mais com a inteligibilidade do sinal, do que com a qualidade, fato aceitável em algumas aplicações, como nas comunicações militares. A Tabela 6.1 apresenta este esquema de avaliação na coluna "Esforço".

A escala de preferência de sonoridade define o grau de sonoridade (volume) percebido pelo ouvinte. Sua escala de pontuação também varia entre cinco e um, em que o maior valor significa uma preferência de sonoridade mais alto do que o ideal, enquanto o menor valor atribui uma preferência de sonoridade mais baixo que o ideal.

A avaliação do codificador foi realizada em duas partes. Na primeira, o codificador foi avaliado de acordo com o seu projeto inicial, ou seja, o de possuir um sistema de reconhecimento de fonemas no emissor e no receptor. Nesta etapa, o banco de unidades e, consequentemente, a síntese do sinal de voz, forem constituidos apenas de segmentos fonéticos.

Na segunda avaliação, a segmentação de frases utilizadas para compor o banco de unidades foi realizada manualmente e a síntese por concatenação faz uso de segmentos como sílabas, fonemas e encontros vocálicos.

Os sinais de fala foram processados por um computador Intel(R) Pentium(R) Dual CPU T3400 @ 2,16 GHz 2,17 GHz, com memória RAM 2,00 GB e sistema operacional de 32 *bits*. Para

a realização dos testes foi utilizado apenas uma oradora, com idade entre 21 a 30 anos e com nível de escolaridade superior completo. E por melhor se adequar à avaliação do codificador, foram utilizadas apenas as duas escala de opnião, ou seja, qualidade de audição e esforço de audição.

Tabela 6.1: Escala de opinião usada no teste ACR.

Valor	Qualidade	Esforço
5	Excelente	Relaxamento completo, nenhum esforço é necessário
4	Boa	Atenção necessária, não é preciso muito esforço
3	Razoável	Um certo esforço é necessário
2	Pobre	Muito esforço é necessário
1	Ruim	Ininteligível, apesar de qualquer esforço empregado

6.1 Primeira Avaliação

Como mencionado no Capítulo 5, o projeto inicial do receptor do codificador incluía a etapa de reconhecimento de fala, com o objetivo de segmentar e reconhecer fonemas para a formação do banco de unidades acústicas.

Na primeira avaliação, o sistema de reconhecimento de fonemas segmenta 20 frases, apresentadas no Anexo C, e armazena no banco de unidades as amostras dos primeiros trinta e oito diferentes fonemas encontrados nas frases segmentadas. Ou seja, o banco de unidades é composto por trinta e oito segmentos correspondentes aos fonemas, sem a presença de suas variações. A duração média de cada fonema foi de 80 ms. A Tabela 6.2 apresenta os fonemas utilizados na síntese por concatenação de cada frase utilizada na primeira avaliação do codificador.

Tabela 6.2: Unidades acústicas das frases utilizadas na síntese da primeira avaliação do codificador.

Frases	Unidades acústicas
Frase 1	a u g u m a s sp k o i z a s
Frase 2	a d e m a~ d a sp p o r sp R e a u
Frase 3	a s e~ t u i s E t e~ t a sp b i L o~ i s
Frase 4	k a d a u~ m a d E l a s
Frase 5	u m e r k a d u sp f i k a d i sp a u t u
Frase 6	a p e r s p e k t i v a sp k o~ t i n u a
Frase 7	u d e z a f i u a g O r a sp E
Frase 8	u Z o g u k o~ t r a a s u E s i a
Frase 9	e s t a i~ s t a l a d u n a k a z a
Frase 10	f u~ s i o n a r i u s d u g o v e r n u

Após a composição do banco de fonemas, dez frases distintas, apresentadas no Anexo C, foram processadas pelo codificador. A taxa de *bits* resultante depende da quantidade de fonemas

que cada sinal de voz possui. Deste modo, as informações dos índices dos segmentos fonéticos, duração e energia de cada fonema são codificadas com três ou quatro *bits*, como apresentado na Tabela 6.3.

Tabela 6.3: Número médio de fonemas, quantidades de *bits* por parâmetros e taxa de *bits* média (1ª Avaliação).

Frases	**Fonemas/s**	**Codificação (*bits*)**	**Taxa de *bits* média (*bit*/s)**
Frase 1	7	Índices: 3	112,5
		Energia: 3	
		Duração: 3	
Frase 2	8	Índices: 3	112,5
		Energia: 3	
		Duração: 3	
Frase 3	10	Índices: 4	150,0
		Energia: 4	
		Duração: 4	
Frase 4	6	Índices: 3	112,5
		Energia: 3	
		Duração: 3	
Frase 5	10	Índices: 4	150,0
		Energia: 4	
		Duração: 4	
Frase 6	10	Índices: 4	150,0
		Energia: 4	
		Duração: 4	
Frase 7	8	Índices: 3	112,5
		Energia: 3	
		Duração: 3	
Frase 8	9	Índices: 4	150,0
		Energia: 4	
		Duração: 4	
Frase 9	9	Índices: 4	150,0
		Energia: 4	
		Duração: 4	
Frase 10	10	Índices: 4	150,0
		Energia: 4	
		Duração: 4	

O teste ACR foi realizado mediante o envio dos sinais de voz processados aos participantes que, antecipadamente instruídos, realizaram individualmente o teste utilizando o sistema de som dos seus próprios computadores. O teste contou com a participação de 12 ouvintes-avaliadores não especialistas na área, cujas informações sobre idade e nível de escolaridade estão apresentadas respectivamente nas Tabelas 6.4 e 6.5, em que é possível observar a predominância de participantes com idades entre 21 e 30 anos, com nível de escolarida superior completo ou com pós-graduação.

Tabela 6.4: Distribuição dos ouvintes-avaliadores por idade (1ª Avaliação).

Idade (anos)	**Quantidade de ouvintes-avaliadores**
10 - 20	1
21 - 30	9
31 - 40	0
41 - 50	0
51 - 60	1
61 - 70	1
Total	**12**

Tabela 6.5: Distribuição dos ouvintes-avaliadores por nível de escolaridade (1ª Avaliação).

Nível de escolaridade	**Quantidade de ouvintes-avaliadores**
Ensino Médio	0
Superior Incompleto	2
Superior Completo	6
Pós-Graduação	4
Total	**12**

A Tabela 6.6 mostra os resultados obtidos nos testes subjetivos e seus respectivos desvios padrão e intervalo de confiança. Essa primeira avaliação teve o objetivo de avaliar a qualidade geral dos sinais de fala processados pelo codificador, o que inclui fatores como inteligibilidade e naturalidade do sinal de voz.

O codificador permite a transmissão do sinal de voz com a taxa de transmissão média de 112,5 e 150,0 *bits/s*, dependendo da quantidade de fonemas pronunciados por segundo em cada frase. Os resultados indicam que cinco dos sinais de voz obtiveram notas acima de três, sendo considerados de qualidade razoável a boa. Por outro lado, três dos sinais de fala alcançaram notas entre dois e três, resultando em sinais de qualidade pobre. Os demais sinais de voz obtiveram notas abaixo de dois, sendo classificados como sinais de qualidade ruim.

As Tabelas 6.7, 6.8 e 6.9 apresentam, respectivamente, os resultados dos testes subjetivos de acordo com a idade dos participantes, seus respectivos desvios padrão e intervalo de confiança. Este teste é necessário para verificar como os participantes com maior idade avaliam os sinais de voz processados pelo codificador, uma vez que a deficiência auditiva em idosos é considerada um problema de saúde pública, em função de sua alta prevalência e das dificuldades que acarreta [78].

Os resultados por idade indicam que os dois participantes com idade superior a 50 anos atribuiram notas aos sinais de voz não muito distintas dos outros participantes com idade inferior, significando que o esforço auditivo foi semelhante para todos os ouvites-avaliadores, independente da tendência daqueles participantes possuirem dificuldade auditiva causada pela idade.

A qualidade dos sinais de fala foi afetada pelos erros ocorridos na segmentação, que obteve uma taxa de erro de 20%, principalmente para a segmentação realizada para a formação do banco

Tabela 6.6: Resultados dos testes subjetivos de qualidade (1ª Avaliação).

Frases	**Pontuação**	**Desvio padrão**	**Intervalo de confiança**
Frase 1	3,4	0,88	$3,4 \pm 0,88$
Frase 2	2,2	1,06	$2,2 \pm 1,06$
Frase 3	1,5	0,97	$1,5 \pm 0,97$
Frase 4	3,1	1,15	$3,1 \pm 1,15$
Frase 5	2,1	1,07	$2,1 \pm 1,07$
Frase 6	1,2	0,98	$1,2 \pm 0,98$
Frase 7	2,0	1,32	$2,0 \pm 1,32$
Frase 8	3,3	1,10	$3,3 \pm 1,10$
Frase 9	3,2	1,01	$3,2 \pm 1,01$
Frase 10	3,5	0,86	$3,5 \pm 0,86$

Tabela 6.7: Resultados dos testes de qualidade por idade dos participantes dos testes subjetivos.

Frases	**10-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**
Frase 1	5,0	3,1	-	-	4,0	4,0
Frase 2	3,0	2,3	-	-	2,0	2,0
Frase 3	1,0	1,4	-	-	2,0	2,0
Frase 4	2,0	3,3	-	-	2,0	3,0
Frase 5	2,0	2,0	-	-	2,0	3,0
Frase 6	1,0	1,3	-	-	1,0	1,0
Frase 7	2,0	2,1	-	-	1,0	2,0
Frase 8	3,0	3,3	-	-	3,0	4,0
Frase 9	3,0	3,3	-	-	3,0	3,0
Frase 10	4,0	3,4	-	-	3,0	4,0

Tabela 6.8: Desvio padrão dos resultados por idade dos participantes dos testes subjetivos.

Frases	**10-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**
Frase 1	0,00	0,86	-	-	0,00	0,00
Frase 2	0,00	0,89	-	-	0,00	0,00
Frase 3	0,00	0,91	-	-	0,00	0,00
Frase 4	0,00	0,90	-	-	0,00	0,00
Frase 5	0,00	0,93	-	-	0,00	0,00
Frase 6	0,00	0,79	-	-	0,00	0,00
Frase 7	0,00	0,89	-	-	0,00	0,00
Frase 8	0,00	0,93	-	-	0,00	0,00
Frase 9	0,00	0,98	-	-	0,00	0,00
Frase 10	0,00	0,87	-	-	0,00	0,00

Tabela 6.9: Intervalo de confiança dos resultados por idade dos participantes dos testes subjetivos.

Frases	**10-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**
Frase 1	$5,0 \pm 0,00$	$3,1 \pm 0,86$	-	-	$4,0 \pm 0,00$	$4,0 \pm 0,00$
Frase 2	$3,0 \pm 0,00$	$2,3 \pm 0,89$	-	-	$2,0 \pm 0,00$	$2,0 \pm 0,00$
Frase 3	$1,0 \pm 0,00$	$1,4 \pm 0,91$	-	-	$2,0 \pm 0,00$	$2,0 \pm 0,00$
Frase 4	$2,0 \pm 0,00$	$3,3 \pm 0,90$	-	-	$2,0 \pm 0,00$	$3,0 \pm 0,00$
Frase 5	$2,0 \pm 0,00$	$2,0 \pm 0,93$	-	-	$2,0 \pm 0,00$	$3,0 \pm 0,00$
Frase 6	$1,0 \pm 0,00$	$1,3 \pm 0,79$	-	-	$1,0 \pm 0,00$	$1,0 \pm 0,00$
Frase 7	$2,0 \pm 0,00$	$2,1 \pm 0,89$	-	-	$1,0 \pm 0,00$	$2,0 \pm 0,00$
Frase 8	$3,0 \pm 0,00$	$3,3 \pm 0,93$	-	-	$3,0 \pm 0,00$	$4,0 \pm 0,00$
Frase 9	$3,0 \pm 0,00$	$3,3 \pm 0,98$	-	-	$3,0 \pm 0,00$	$3,0 \pm 0,00$
Frase 10	$4,0 \pm 0,00$	$3,4 \pm 0,87$	-	-	$3,0 \pm 0,00$	$4,0 \pm 0,00$

de unidades. Isso porque esses erros implicam uma segmentação imprecisa de cada fonema, ou seja, não é realizada exatamente no tempo que inicia e no tempo que termina cada fonema.

Para o funcionamento do emissor, esses erros implicam alterações no valor obtido do cálculo da energia e duração de cada fonema. No receptor, a não correta segmentação forma um banco de unidades cujos fonemas armazenados possuem amostras de fonemas adjacentes a eles na frase segmentada. Assim, ao se utilizar os fonemas na concatenação para a formação de palavras, estas ficam com a qualidade distorcida pela presença de amostras que não pertencem aos fonemas que as formam.

6.2 Segunda Avaliação

A segunda avaliação foi dividida em duas etapas. A primeira etapa teve o objetivo de avaliar apenas a inteligibilidade do sinal produzido pelo codificador, ou seja, se é possível a compreenssão da mesagem falada em cada sinal de fala. A segunda etapa avaliou a qualidade geral do sinal de voz, o que engloba além da inteligibilidade, naturalidade, ausência de ecos, ruídos, entre outros fatores.

Nesta segunda avaliação, com o intuito de garantir as coarticulações entre fonemas e aumentar a qualidade do sinal processado pelo codificador, o banco de unidades acústicas foi formado por sílabas, fonemas e encontros vocálicos. Para minimizar os erros provocados na segmentação automática de segmentos, realizou-se a segmentação manual das unidades com o *software* Audacity 2.0.0®, como explicado no Capítulo 5. Assim, 20 frases foneticamente balanceadas obtidas em [69], apresentadas no Anexo C, foram gravadas com uma taxa de amostragem de 22050 amostras/s, e segmentadas manualmente, obtendo um total de duzentos e três segmentos dintintos, amarmazenados em um arquivo de tamanho 22,7 M*bits*.

Após a construção do banco de unidades, 15 frases foneticamente balanceadas foram escolhidas aleatoriamente de [69] e processadas pelo codificador. A síntese utiliza níveis de prioridade para selecionar os segmentos a serem utilizados na concatenação, sendo selecionados, em primeiro

lugar, as sílabas e encontros vocálicos, e por último, os fonemas. As unidades acústicas usadas na formação das frases estão apresentadas na Tabela 6.10.

Tabela 6.10: Unidades acústicas das frases utilizadas na síntese da segunda avaliação do codificador.

Frases	**Unidades acústicas**
Frase 1	a / ka za / f oi / v e di da / s em /p r e sa
Frase 2	e la / t em / m uin ta / fo mi
Frase 3	di / d ia / a pa gue / a / l u z / se p r i
Frase 4	m eu / ti mi / se / com sa g r ou/ com mu / o / me lh o r
Frase 5	com me r/ k im d im/ e / se p r i / um ma / b o a / pe di da
Frase 6	u / com gr e sso/ v ou ta / a t r a s/ em / s ua / pa la v r a
Frase 7	as / k r i an sa s / com nh ê ce r am / u / fi lh o ti / di / em ma
Frase 8	a / a p r e zem tar ção / f oi / can ce la da/ p ô r/ k au za / du / s om
Frase 9	u ma / ga r ô ta / f oi / prê za / om t em / a / n oi te
Frase 10	u / k l i ma / n ão / e / m ai s / sê co / nu / im te r io r
Frase 11	m uin to / p r a ze r / em / com nh ê ce lu
Frase 12	t r a ba lh ei / m ai s / du / que / po d ia
Frase 13	ô g i / eu / a co r d ei / m uin to / c au mu
Frase 14	s eu / s au do/ b an ca r io r/ es ta / b ai xo
Frase 15	a im da / t em nho / s im co / te le f om nem ma s/ pa ra / da r

Ambas as etapas foram realizadas com 62 ouvintes-participantes, cujas distribuição de idade e formação acadêmica estão apresentadas nas Tabelas 6.11 e 6.12, em que é possível notar uma maior presença de participantes com idade entre 21 e 30 anos e com nível de escolaridade de pós-graduação.

Tabela 6.11: Distribuição dos ouvintes-avaliadores por idade (2ª Avaliação).

Idade (anos)	**Quantidade de ouvintes-avaliadores**
10 - 20	23
21 - 30	30
31 - 40	8
41 - 50	-
51 - 60	-
61 - 70	1
Total	**62**

Para os 15 sinais de voz processados pelo codificador, foram utilizados quatro, três e três *bits* para a codificação, respectivamente, dos índices fonéticos, energia e duração. Desta forma, todos foram codificados com uma taxa de *bits* média de 125,0 *bit*/s, como apresentado na Tabela 6.13.

Tabela 6.12: Distribuição dos ouvintes-avaliadores por nível de escolaridade (2ª Avaliação).

Nível de escolaridade	**Quantidade de ouvintes-avaliadores**
Ensino Médio	9
Superior Incompleto	13
Superior Completo	8
Pós-Graduação	32
Total	**62**

Tabela 6.13: Número médio de fonemas, quantidades de *bits* por parâmetros e taxa de *bits* média.

Frases	**Fonemas/s**	**Codificação (*bits*)**	**Taxa de *bits* média (*bits*/s)**
Frase 1	11	Índices: 4 Energia/Duração: 3	125,0
Frase 2	7	Índices: 4 Energia/Duração: 3	125,0
Frase 3	10	Índices: 4 Energia/Duração: 3	125,0
Frase 4	9	Índices: 4 Energia/Duração: 3	125,0
Frase 5	9	Índices: 4 Energia/Duração: 3	125,0
Frase 6	10	Índices: 4 Energia/Duração: 3	125,0
Frase 7	10	Índices: 4 Energia/Duração: 3	125,0
Frase 8	11	Índices: 4 Energia/Duração: 3	125,0
Frase 9	11	Índices: 4 Energia/Duração: 3	125,0
Frase 10	11	Índices: 4 Energia/Duração: 3	125,0
Frase 11	10	Índices: 4 Energia/Duração: 3	125,0
Frase 12	11	Índices: 4 Energia/Duração: 3	125,0
Frase 13	9	Índices: 4 Energia/Duração: 3	125,0
Frase 14	10	Índices: 4 Energia/Duração: 3	125,0
Frase 15	10	Índices: 4 Energia/Duração: 3	125,0

Primeira etapa

Como mencionado, a primeira epata da segunda avaliação teve o objetivo de avaliar a inteligibilidade dos aúdios processados. Desta maneira, os ouvintes-avaliadores foram instruídos a avaliarem apenas a compreensão da mensagem dita, independentemente da qualidade do sinal de voz.

A Tabela 6.14 apresenta o resultado geral do teste de inteligibilidade e seus respectivos desvios padrão e intervalos de confiança. De acordo com os resultados, quase todos os sinais de fala obtiveram inteligibilidade considerada de razoável a boa. Seguindo a escala mostrada na Tabela 6.1, os participantes fizeram pouco ou nenhum esforço para entenderem a mesagem pronunciada no sinal de fala.

Tabela 6.14: Resultados dos testes subjetivos de inteligibilidade.

Frases	**Pontuação**	**Desvio padrão**	**Intervalo de confiança**
Frase 1	3,85	1,15	$3,85 \pm 1,15$
Frase 2	3,92	0,97	$3,92 \pm 0,97$
Frase 3	2,81	1,49	$2,81 \pm 1,49$
Frase 4	3,35	1,18	$3,35 \pm 1,18$
Frase 5	3,16	1,31	$3,16 \pm 1,31$
Frase 6	2,94	1,42	$2,94 \pm 1,42$
Frase 7	3,11	1,32	$3,11 \pm 1,32$
Frase 8	3,77	0,89	$3,77 \pm 0,89$
Frase 9	3,90	0,98	$3,90 \pm 0,98$
Frase 10	3,40	1,18	$3,40 \pm 1,18$
Frase 11	3,89	1,18	$3,89 \pm 1,18$
Frase 12	4,16	0,77	$4,16 \pm 0,77$
Frase 13	3,81	0,93	$3,81 \pm 0,93$
Frase 14	3,23	1,30	$3,23 \pm 1,30$
Frase 15	3,60	1,08	$3,60 \pm 1,08$

Voltando ao problema de sáude pública da tendência de deficiência auditiva de pessoas com maior idade, é necessário analisar os resultados dos testes por idade para verificar qual foi a inteligibilidade percebida por tais participantes, ou seja, o quanto de esforço os sinais de voz exigiram para que essas pessoas precisassem conseguir entender a mensagem falada.

De acordo com os resultados apresentados na Tabela 6.15, cujos desvios padrão e intervalo de confiança estão apresentados nas Tabelas 6.16 e 6.17, o único ouvinte com idade superior a 50 anos atribuíu uma pontuação considerada de razoável a excelente, não exigindo muito esforço por parte de tal participante para a compreensão dos sinais de fala.

Tabela 6.15: Resultados por idade dos participantes dos testes subjetivos de inteligibilidade.

Frases	10-20	21-30	31-40	41-50	51-60	61-70
Frase 1	3,83	3,73	4,25	-	-	5,00
Frase 2	3,83	3,93	4,00	-	-	5,00
Frase 3	2,74	2,63	3,38	-	-	5,00
Frase 4	3,52	3,17	3,50	-	-	4,00
Frase 5	3,09	2,97	4,00	-	-	4,00
Frase 6	2,96	2,80	3,38	-	-	3,00
Frase 7	3,00	3,07	3,63	-	-	3,00
Frase 8	3,78	3,87	3,50	-	-	3,00
Frase 9	3,87	3,97	3,75	-	-	4,00
Frase 10	3,48	3,27	3,63	-	-	4,00
Frase 11	3,96	3,80	4,00	-	-	4,00
Frase 12	4,09	4,20	4,25	-	-	4,00
Frase 13	3,78	3,77	4,00	-	-	4,00
Frase 14	3,22	3,03	3,88	-	-	4,00
Frase 15	3,70	3,37	4,00	-	-	5,00

Tabela 6.16: Desvio padrão dos resultados por idade dos participantes dos testes subjetivos de inteligibilidade.

Frases	10-20	21-30	31-40	41-50	51-60	61-70
Frase 1	1,17	1,12	1,09	-	-	0,00
Frase 2	1,01	0,93	1,00	-	-	0,00
Frase 3	1,48	1,45	1,41	-	-	0,00
Frase 4	1,06	1,29	1,00	-	-	0,00
Frase 5	1,44	1,20	1,00	-	-	0,00
Frase 6	1,46	1,30	1,73	-	-	0,00
Frase 7	1,35	1,24	1,49	-	-	0,00
Frase 8	0,93	0,81	1,00	-	-	0,00
Frase 9	1,08	0,84	1,20	-	-	0,00
Frase 10	1,35	1,12	0,86	-	-	0,00
Frase 11	1,04	1,33	1,00	-	-	0,00
Frase 12	0,78	0,70	0,97	-	-	0,00
Frase 13	1,06	0,80	1,00	-	-	0,00
Frase 14	1,28	1,30	1,17	-	-	0,00
Frase 15	0,86	1,20	1,00	-	-	0,00

Segunda etapa

A segunda parte da segunda avaliação teve como objetivo avaliar a qualidade geral dos sinais de voz, englobando parâmetros como inteligibilidade, naturalidade, ruído, ecos, entre outros.

Tabela 6.17: Intervalo de confiança dos resultados por idade dos participantes dos testes subjetivos de inteligibilidade.

Frases	**10-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**
Frase 1	$3,83 \pm 1,17$	$3,73 \pm 1,12$	$4,25 \pm 1,09$	-	-	$5,00 \pm 0,00$
Frase 2	$3,83 \pm 1,01$	$3,93 \pm 0,93$	$4,00 \pm 1,00$	-	-	$5,00 \pm 0,00$
Frase 3	$2,74 \pm 1,48$	$2,63 \pm 1,45$	$3,38 \pm 1,41$	-	-	$5,00 \pm 0,00$
Frase 4	$3,52 \pm 1,06$	$3,17 \pm 1,29$	$3,50 \pm 1,00$	-	-	$4,00 \pm 0,00$
Frase 5	$3,09 \pm 1,44$	$2,97 \pm 1,20$	$4,00 \pm 1,00$	-	-	$4,00 \pm 0,00$
Frase 6	$2,96 \pm 1,46$	$2,80 \pm 1,30$	$3,38 \pm 1,73$	-	-	$3,00 \pm 0,00$
Frase 7	$3,00 \pm 1,35$	$3,07 \pm 1,24$	$3,63 \pm 1,49$	-	-	$3,00 \pm 0,00$
Frase 8	$3,78 \pm 0,93$	$3,87 \pm 0,81$	$3,50 \pm 1,00$	-	-	$3,00 \pm 0,00$
Frase 9	$3,87 \pm 1,08$	$3,97 \pm 0,84$	$3,75 \pm 1,20$	-	-	$3,0 \pm 0,00$
Frase 10	$3,48 \pm 1,35$	$3,27 \pm 1,12$	$3,63 \pm 0,86$	-	-	$4,00 \pm 0,00$
Frase 11	$3,96 \pm 1,04$	$3,80 \pm 1,33$	$4,00 \pm 1,00$	-	-	$4,00 \pm 0,00$
Frase 12	$4,09 \pm 0,78$	$4,2 \pm 0,70$	$4,25 \pm 0,97$	-	-	$4,00 \pm 0,00$
Frase 13	$3,78 \pm 1,06$	$3,77 \pm 0,80$	$4,00 \pm 1,00$	-	-	$4,00 \pm 0,00$
Frase 14	$3,22 \pm 1,28$	$3,03 \pm 1,30$	$3,88 \pm 1,17$	-	-	$4,00 \pm 0,00$
Frase 15	$3,70 \pm 0,86$	$3,37 \pm 1,20$	$4,00 \pm 1,00$	-	-	$5,00 \pm 0,00$

O resultado geral de avaliação da qualidade dos sinais de voz está apresentado na Tabela 6.18. Como esperado, a qualidade dos áudios obtive uma pontuação inferior comparada à inteligibilidade, entretanto a maior parte dos áudios obtive qualidade consideradas de razoável a boa.

Tabela 6.18: Resultados dos testes subjetivos de qualidade.

Frases	**Pontuação**	**Desvio padrão**	**Intervalo de confiança**
Frase 1	3,44	1,06	$3,44 \pm 1,06$
Frase 2	3,37	0,95	$3,37 \pm 0,95$
Frase 3	2,63	1,18	$2,63 \pm 1,18$
Frase 4	2,92	1,08	$2,92 \pm 1,08$
Frase 5	2,84	1,05	$2,84 \pm 1,05$
Frase 6	2,69	1,12	$2,69 \pm 1,12$
Frase 7	2,79	1,11	$2,79 \pm 1,11$
Frase 8	3,03	0,98	$3,03 \pm 0,98$
Frase 9	3,13	1,02	$3,13 \pm 1,02$
Frase 10	3,00	1,02	$3,00 \pm 1,02$
Frase 11	3,16	0,97	$3,16 \pm 0,97$
Frase 12	3,40	0,97	$3,40 \pm 0,97$
Frase 13	3,18	0,91	$3,18 \pm 0,91$
Frase 14	2,89	1,02	$2,89 \pm 1,02$
Frase 15	2,94	0,95	$2,94 \pm 0,95$

As Tabelas 6.19, 6.20 e 6.21 apresentam respectivamente os resultados selecionados por idade da qualidade dos sinais de fala e seus respectivos desvios padrão e intervalo de confiança.

Assim, é possível observar que, independentimente da idade dos ouvintes-participantes, a qualidade, para maioria dos sinais de voz, se manteve entre razoável a boa.

Tabela 6.19: Resultados por idade dos participantes dos testes subjetivos de qualidade.

Frases	**10-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**
Frase 1	3,39	3,43	3,63	-	-	3,00
Frase 2	3,48	3,30	3,38	-	-	3,00
Frase 3	2,57	2,57	3,00	-	-	3,00
Frase 4	2,96	2,83	3,13	-	-	3,00
Frase 5	2,74	2,90	2,88	-	-	3,00
Frase 6	2,61	2,73	2,75	-	-	3,00
Frase 7	2,74	2,77	3,00	-	-	3,00
Frase 8	3,09	3,03	2,88	-	-	3,00
Frase 9	3,09	3,13	3,25	-	-	3,00
Frase 10	3,04	2,90	3,25	-	-	3,00
Frase 11	3,04	3,17	3,50	-	-	3,00
Frase 12	3,22	3,47	3,75	-	-	3,00
Frase 13	3,09	3,23	3,25	-	-	3,00
Frase 14	2,74	2,93	3,13	-	-	3,00
Frase 15	2,91	2,90	3,13	-	-	3,00

Tabela 6.20: Desvio padrão dos resultados por idade dos participantes dos testes subjetivos de qualidade.

Frases	**10-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**
Frase 1	0,92	1,15	1,11	-	-	0,00
Frase 2	0,93	0,94	1,11	-	-	0,00
Frase 3	1,21	1,17	1,12	-	-	0,00
Frase 4	1,08	1,04	1,27	-	-	0,00
Frase 5	1,15	1,01	0,93	-	-	0,00
Frase 6	1,17	1,09	1,09	-	-	0,00
Frase 7	1,07	1,17	1,00	-	-	0,00
Frase 8	0,93	1,08	0,78	-	-	0,00
Frase 9	0,97	1,06	1,09	-	-	0,00
Frase 10	1,04	1,08	0,66	-	-	0,00
Frase 11	0,95	1,07	0,50	-	-	0,00
Frase 12	0,83	1,06	0,97	-	-	0,00
Frase 13	0,83	0,99	0,83	-	-	0,00
Frase 14	0,94	1,09	0,93	-	-	0,00
Frase 15	0,88	1,01	0,93	-	-	0,00

Como esperado, os resultados obtidos na segunda avaliação foram superiores ao obtidos na primeira avaliação. O melhor desempenho alcançado na segunda avaliação foi devido a dois fatores:

Tabela 6.21: Intervalo de confiança dos resultados por idade dos participantes dos testes subjetivos de qualidade.

Frases	**10-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**
Frase 1	$3,39 \pm 0,92$	$3,43 \pm 1,15$	$3,63 \pm 1,11$	-	-	$3,00 \pm 0,00$
Frase 2	$3,48 \pm 0,93$	$3,30 \pm 0,94$	$3,38 \pm 1,11$	-	-	$3,00 \pm 0,00$
Frase 3	$2,57 \pm 1,21$	$2,57 \pm 1,17$	$3,00 \pm 1,12$	-	-	$3,00 \pm 0,00$
Frase 4	$2,96 \pm 1,08$	$2,83 \pm 1,04$	$3,13 \pm 1,27$	-	-	$3,00 \pm 0,00$
Frase 5	$2,74 \pm 1,15$	$2,90 \pm 1,01$	$2,88 \pm 0,93$	-	-	$3,00 \pm 0,00$
Frase 6	$2,61 \pm 1,17$	$2,73 \pm 1,09$	$2,75 \pm 1,09$	-	-	$3,00 \pm 0,00$
Frase 7	$2,74 \pm 1,07$	$2,77 \pm 1,17$	$3,00 \pm 1,00$	-	-	$3,00 \pm 0,00$
Frase 8	$3,09 \pm 0,93$	$3,03 \pm 1,08$	$2,88 \pm 0,78$	-	-	$3,00 \pm 0,00$
Frase 9	$3,09 \pm 0,97$	$3,13 \pm 1,06$	$3,25 \pm 1,09$	-	-	$3,0 \pm 0,00$
Frase 10	$3,04 \pm 1,04$	$2,90 \pm 1,08$	$3,25 \pm 0,66$	-	-	$3,00 \pm 0,00$
Frase 11	$3,04 \pm 0,95$	$3,17 \pm 1,07$	$3,50 \pm 0,50$	-	-	$3,00 \pm 0,00$
Frase 12	$3,22 \pm 0,83$	$3,47 \pm 1,06$	$3,75 \pm 0,97$	-	-	$3,00 \pm 0,00$
Frase 13	$3,09 \pm 0,83$	$3,23 \pm 0,99$	$3,25 \pm 0,83$	-	-	$3,00 \pm 0,00$
Frase 14	$2,74 \pm 0,94$	$2,93 \pm 1,09$	$3,13 \pm 0,93$	-	-	$3,00 \pm 0,00$
Frase 15	$2,91 \pm 0,88$	$2,90 \pm 1,01$	$3,13 \pm 0,93$	-	-	$3,00 \pm 0,00$

O primeiro está realcionado à segmentação realizada para formar o banco de unidades acústicas. Ao se realizar a segmentação manualmente, os erros de segmentação foram minimizados, quando comparados aos erros provocados pelo sistema de reconhecimento de fonemas. O segundo motivo foi a utilização de segmentos maiores na síntese por concatenação, como encontros vocálicos e palavras, mantendo as coarticulações entre os fonemas que os formam e proporcionando aos sinais de voz uma melhor inteligibilidade e qualidade.

6.3 Avaliação Geral do Codificador

Para que seja possível sua utilização em sistemas móveis celulares, a implementação de um codificador de voz deve levar em consideração requisitos como: taxa de *bits*, qualidade do sinal sintetizado, custo computacional, memória para armazenamento e retardo de comunicação.

A taxa de *bits* proporcionada pelo codificador desenvolvido é muito inferior às taxas dos codificadores atualmente utilizados nos sistemas de telefonia móvel. Em relação ao codificador AMR-WB utilizado no padrão WCDMA (3G), o codificador proposto oferece uma taxa de *bits* quarenta e quatro vezes menor, se comparado à menor taxa de *bits* proporcionanda pelo AMR-WB (6,6 kbit/s) e à maior fornecida pelo codificador proposto (150 bit/s).

No entanto, o codificador AMR-WB proporciona uma qualidade boa (MOS = 4,14) dos sinais de voz sintetizados, enquanto o codificador proposto apresenta uma qualidade média razoável (MOS = 3,0 para qualidade e MOS = 3,5 para inteligibilidade), ou seja, uma qualidade da voz sintetizada 28% menor.

Em relação à complexidade do algoritmo, o processo que apresenta o maior custo computacional é o treinamento dos modelos HMM, realizado no sistema de reconhecimento de fonemas. Entretando, o processo de treinamento é feito uma única vez na implementação do sistema de reconhecimento de fonemas. Na utilização do codificador é realizado apenas o reconhecimento de fala.

Após feito o reconhecimento de fala, as demais etapas são processos de baixa complexidade que levam menos de um segundo em um computador pessoal, como apresentado nas Tabelas 6.22 e 6.23, para sintetizar uma frase, proporcionando um pequeno retardo na comunicação.

Em relação à memória necessária para utilização do codificador, a parte que exige maior espaço de armazenamento é o banco de unidades acústicas que necessitou de 22,7 M*bits* para armazenar os 203 segmentos acústicos com uma taxa de amostragem de 22050 amostras/s. No entanto, esse valor pode ser diminuido, reduzindo a taxa de amostragem dos sinais de fala.

Uma observação adicional, é que o reconhecimento ao locutor não foi afetado em nenhum dos dois testes. Isso porque a síntese do sinal de voz utilizou segmentos acústicos retirados do sinal de voz do orador, o que mantém as características da voz, como o timbre.

Além disso, para este codificador, o aumento na taxa de *bits* não significa um aumento na qualidade do sinal final. Isso porque, o que se transmite não é o sinal de voz propriamente dito, sendo a quantização realizada nos parâmetros que representam cada segmento acústico. A qualidade do sinal sintetizado está relacionada diretamente com a qualidade do sistema de reconhecimento de fonemas, além da correta seleção das unidades acústicas para a realização da síntese.

Tabela 6.22: Tempo de processamento das frases utilizadas na primeira avaliação.

Frases	**Tempo (s)**
Frase 1	0,11
Frase 2	0,13
Frase 3	0,21
Frase 4	0,17
Frase 5	0,20
Frase 6	0,19
Frase 7	0,12
Frase 8	0,20
Frase 9	0,18
Frase 10	0,22

Tabela 6.23: Tempo de processamento das frases utilizadas na segunda avaliação.

Frases	**Tempo (s)**
Frase 1	0,79
Frase 2	0,60
Frase 3	0,43
Frase 4	0,55
Frase 5	0,49
Frase 6	0,44
Frase 7	0,56
Frase 8	0,43
Frase 9	0,44
Frase 10	0,57
Frase 11	0,50
Frase 12	0,63
Frase 13	0,45
Frase 14	0,61
Frase 15	0,49

6.4 Considerações Finais

Neste capítulo foram apresentados os resultados dos testes subjetivos realizados para avaliar o desempenho do codificador proposto no trabalho.

Foram feitas duas avaliações. A primeira utilizou um reconhecedor fonético no emissor para obter uma sequência de índices fonéticos e no receptor para formar o banco de unidades acústicas necessário na síntese por concatenação que, nesta avaliação, era composto por trinta e oito segmentos, correspondendo aos fonemas do português do brasil.

Na primeira avaliação, os parâmetros que representam o sinal de voz foram codificados com, no máximo, 150 bit/s e 50 por cento dos sinais de voz foram considerados de qualidade razoável a bom.

Com a intenção de melhorar a qualidade final da voz processada pelo codificador, a segunda avaliação foi realizada com sinais sintetizados pela concatenação de segmentos obtidos manualmente, alguns com coarticulação entre os fonemas que os formam. Esses segmentos foram as sílabas, encontros vocálicos e na falta destes, fonemas.

A segunda avaliação teve como objetivo medir a inteligibilidade e qualidade de 15 frases foneticamente balanceadas. Os sinais de fala foram codificados com 125 bit/s e os resultados indicaram bons níveis de inteligibilidade e qualidade razoável.

CAPÍTULO 7

Considerações Finais e Trabalhos Futuros

Um sistema de codificação de voz tem o objetivo de reduzir a quantidade de *bits* necessária para representar o sinal de voz. Esses sistemas possibilitaram a transmissão digital da voz nos sistemas telefônicos, tornando importante a representação digital do sinal de voz de forma eficiente.

Particulamente, no caso dos sistemas de comunicações móveis, que nas últimas décadas teve uma grande evolução, o desenvolvimento de métodos de codificação de voz a baixas taxas de *bits* é relevante, devido às características do sistemas móveis de apresentarem uma capacidade de canal limitada e uma grande adesão de usuários.

A codificação de voz a baixa taxas de *bits* permite aumentar a capacidade dos sistemas de telefonia móvel, por requererem uma menor largura de banda por usuário. Além disso, permite uma comunicação a custos mais acessíveis, caso a cobrança seja realizada em função da taxa de transmissão.

Esta dissertação descreve o desenvolvimento de um codificador de voz pessoal. Classificado como do tipo fonético, foi implementado para ser utilizado principalmente nos sistemas de telefonia móvel e tem como principal característica a codificação do sinal de voz a baixa taxa de *bits*.

O codificador deve ser utilizado principalmente nas comunicações mais frequentes realizadas pelos usuários de telefonia móvel, com o objetivo de proporcionar um aumento da quantidade de usuários do sistema móvel, além de obter um menor custo por ligação.

O codificador tem seu emissor formado por um sistema de reconhecimento de fala, cujo objetivo é obter uma sequência de segmentos fonéticos que são etiquetados com índices pré-estabelecidos e têm suas durações e energia estimadas. O codificador reduz a taxa de *bits* por codificar apenas os parâmetros que representam cada segmento da fala, como índice, duração e energia, em vez de codificar as amostras do sinal de voz.

O receptor é constituído de bancos de unidades acústicas específicos para cada usuário, formado por segmentos de sílabas, encontros vocálicos e fonemas, obtidos inicialmente por um sistema de reconhecimento de fala, e em outra ocasião, manualmente, utilizadas na síntese por concatenação do sinal de voz.

Para avaliar o desempenho do codificador foram realizadas duas avaliações subjetivas informais. A primeira avaliação teve o objetivo de aferir a qualidade geral dos sinais de voz sintetizados pelo codificador utilizando o sistema de reconhecimento automático para obter os fonemas usados na síntese. A segunda avaliação utilizou sinais de voz processadas pelo codificador ao utilizar segmentação manual para obter as unidades acústicas como fonemas, sílabas e econtros vocálicos usados na síntese por concatenação e foi dividida em duas etapas. A primeira etapa foi utilizada apenas para aferir a inteligibilidade dos sinais de fala, enquanto a segunda etapa teve o objetivo de avaliar a qualidade geral dos sinais.

A taxa de *bits* média resultante de cada frase foi obtida em função da quantidade de fonemas pronunciados por segundo. Inicialmente, foram processados dez sinais de voz e o codificador permitiu a codificação dos parâmetros a uma taxa de 112,5 a 150 *bits*/s. A primeira avaliação consistiu em aferir a qualidade geral desses sinais de fala e os resultados mostram que 50 por cento dos sinais de voz foram classificados de qualidades razoável, enquanto os demais foram considerados sinais de fala de qualidade pobre ou ruim.

Para a segunda avaliação, 15 sinais de voz foram processados pelo codificador e sintetizados por meio da concatenação de segmentos obtidos manualmente. A quantidade de fonemas pronunciada por segundo possibilitou a codificação dos sinais de voz com 125 *bit*/s. Esses sinais de fala inicialmente foram avaliados em termos apenas da inteligibilidade. Os resultados indicam que a maior parte dos sinais de voz obteve MOS maior que três, significando que os participantes fizeram pouco ou nenhum esforço para entenderem a mensagem pronunciada nos sinais de voz.

A segunda fase da avaliação teve o objetivo de avaliar a qualidade geral do áudio. Como esperado, os resultados dos MOS obtidos foram inferiores em relação aos MOS obtidos no teste de inteligibidade, no entanto, a maior parte deles foi considerada de qualidade razoável a bom.

No geral, os resultados da segunda avaliação foram superiores aos obtidos na primeira avaliação. Essa melhoria está relacionada ao uso de unidades acústicas maiores, como sílabas e encontros vocálicos, garantindo a coarticulação entre os fonemas que formam essas unidades. Além disso, os segmentos de fala utilizados na síntese por concatenação foram obtidos de forma manual, o que possibilitou a minimização dos erros na segmentação, quando comparados aos segmentos obtidos no sistema de reconhecimento de fonemas, que obteve um erro de 20%.

Os testes foram feitos apenas com um locutor do sexo feminino. No entanto, a realização dos testes com outros locutores fornece resultados semelhantes, pois o que determina a qualidade do sinal de voz sintetizada é a qualidade do sistema de reconhecimento de fonemas, além da correta escolha dos segmentos que formam o banco de unidades acústicas para a realização da síntese.

Por fim, foram avaliados os requisitos de taxa de *bits*, qualidade do sinal sintetizado, custo computacional, memória para armazenamento e retardo de comunicação.

Em relação à taxa de *bits*, o codificador proposto sobressai em relação a todos os codificadores utilizados nos padrões de telefonia móvel. No entanto, apresenta MOS 28% menor que aqueles.

O custo computacional da implementação do codificador é mais relevante na etapa de treinamento dos modelos HMM. Entretanto, no uso do codificador, os modelos HMM são previamente

treinados, realizando apenas as etapas de reconhecimento e demais funções, que representam um baixo custo computacional.

A memória requerida está diretamente relacionada à memória necessária para armazenar o banco de unidades acústicas. Para o armazenamento das duzentos e três segmentos de um usuário foi necessário um espaço de 22,7 M*bits*.

Em relação ao retardo na comunicação, o codificadou levou menos de um segundo para processar cada frase.

Portanto, o codificador proposto apresentou resultados satisfatórios, viabilizando o seu uso, uma vez que permitiu a transmissão do sinal de voz, com uma baixa taxa de *bits* com níveis de razoável a bom para os sinais sintetizados.

7.1 Contribuições

7.1.1 Desenvolvimento de um Sistema de Reconhecimento de Fala

A primeira etapa realizada neste trabalho consistiu no desenvolvimento de um sistema de reconhecimento de fala. Esse sistema faz uso da técnica HMM para modelar o sinal de voz.

Consiste em um sistema dependente de contexto, pois utiliza modelos HMM de trifones para reconhecer cada fonema dos sinais de voz. Além disso, é um sistema independente de locutor, ou seja, capaz de reconhecer o que está sendo falado por qualquer orador, visto que os modelos acústicos que representam cada trifone foram treinados com um banco de voz contendo frases pronunciadas por vários interlocutores.

No caso do presente trabalho, esse sistema foi utilizado para reconhecer apenas fonemas. No entanto, é também é capaz de reconhecer palavras ou frases, por meio da concatenação dos modelos HMM dos fonemas que as formam.

7.1.2 Desenvolvimento de um Codificador de Voz

O desenvolvimento de um codificador de voz é a contribuição mais relevante deste trabalho. O codificador foi projetado para ser utilizado principalmente nos sistemas móveis celulares. Sua principal característica é a possibilidade da codificação do sinal de voz a baixas taxa de transmissão.

O codificador é do tipo fonético e utiliza um sistema de reconhecimento de fala para segmentar foneticamente os sinais de voz dos usuários. Os segmentos fonéticos são etiquetados com índices pré-estabelecidos e suas informações prosódicas de energia e duração são estimadas.

Para reduzir a taxa de transmissão, o codificador quantiza e transmite apenas os parâmetros correspondente a cada segmento, que compreende os índices atribuídos a cada fonema e duração e energia de fonemas, sílabas ou encontros vocálicos.

O receptor armazena segmentos de voz, como fonemas, sílabas e encontros vocálicos, previamente gravados, e realiza a síntese por concatenação após a adaptação com as novas informações de energia e duração recebidas do emissor do codificador.

Como a síntese é realizada por meio de segmentos de voz de cada usuário específico, o codificador não utiliza métodos para aumentar o reconhecimento do orador, uma vez que os segmentos utilizados possuem características de cada orador necessárias ao reconhecimento, como o timbre da voz.

7.1.3 Síntese por Concatenação

A síntese por concatenação realiza a junção em uma dada sequência de segmentos previamente obtidos do sinal de fala. Esses segmentos formam o banco de unidades acústicas e podem ser representados pelos fonemas, sílabas, palavras, ou unidades menores, como difones e trifones.

A síntese por concatenação é utilizada em sistemas de conversão texto-fala [74, 72]. No entanto, esses sistemas utilizam um inventário de unidades formado por um grande conjunto de unidades acústicas, conseguindo uma boa variedade dos segmentos que o compõe, sendo possível escolher, na realização da síntese, o segmento de acordo com a entoação ou com a posição que ocupa em uma palavra.

Entretanto, o esquema do codificador de voz desenvolvido neste trabalho exige que o banco de voz utilizado para compor o inventário de unidades acústicas seja pequeno, visto que são apenas algumas frases pronunciadas pelos oradores, no caso, 20 frases foneticamente balanceadas. Do contrário, limitaria o uso do codificador pela quantidade de frases a serem pronunciadas por cada usuário para obter um banco de unidades composto por um grande conjunto de segmentos. Assim, a síntese por concatenação do codificador de voz recai sobre o problema de ter poucas unidades acústicas formando o inventário dos usuários.

No entanto, foi observado no desenvolvimento deste trabalho que as vogais tem um peso maior em relação às consoantes na qualidade final do áudio sintetizado. Deste modo, para contornar o problema de ter um número restrito de unidades acústicas, esse trabalho utilizou, além de várias possibilidades de segmentos que poderiam ser encontrados no banco de voz, como sílabas e encontros vocálicos, segmentos vocálicos com variações prosódicas. Assim, para cada vogal, foram selecionadas amostras com variações prosódicas e sua utilização na síntese foi de acordo com a posição que ocupava na palavra.

Como mencionado, a síntese por concatenação também fez o uso de sílabas e encontros vocálicos. Esses tipos de segmentos foram escolhidos por manterem as coarticulações entre os fonemas que os formam, fato importante para uma boa qualidade dos sinais de voz sintetizados, visto que as coarticulações presentes entre fonemas são mais importantes que as coarticulações presentes entre sílabas.

Outra observação também feita neste trabalho e que justifica o uso de encontros vocálicos é o fato de que as coarticulações presentes entre as vogais serem importantes para a naturalidade do sinal de voz.

Deste modo, a síntese por concatenação realizada neste trabalho utilizou um pequeno banco de voz, mas que, se utilizado da maneira descrita, sintetiza sinais de voz com bons níveis de inteligibilidade e qualidade razoável.

7.2 Trabalhos Futuros

Sendo o codificador proposto nesse trabalho desenvolvido para ser utilizado em sistemas de telefonia móvel, ele deve proporcionar, além da baixa taxa de transmissão, uma excelente qualidade dos sinais de voz sintetizados. Deste modo, com o objetivo de dar continuidade ao desenvolvimento do codificador proposto e torná-lo mais eficiente, vários aspectos deste trabalho podem ser aperfeiçoados. Alguns desses aspectos são apresentados a seguir.

1. Como a qualidade dos sinais sintetizados pelo codificador está diretamente relacionada ao reconhecimento de fonemas, pretende-se, como trabalho futuro, o aprimoramento do sistema de reconhecimento de fala, que tem uma taxa de erro de 20%, com o objetivo de obter uma segmentação fonética mais precisa. Este melhoramento pode ser alcançado por meio de

   - Melhor treinamento dos HMMs com um maior banco de sinais de voz;
   - Utilização de misturas gaussianas em cada modelo HMM. O aumento do número de misturas contribui para a melhoria da fidelidade dos modelos. Estes benefícios nos resultados fazem sentido pois, ao aumentar as misturas aumenta-se também o número de valores que constituem o modelo e isso poderá melhorar a sua fidelidade.
   - Verificar a taxa de reconhecimento, com a extração de outras características do sinal de voz, além dos coeficientes MFCCs, tais como SSCH (*Subband Spectral Centroid Histograms*) e PNCC (*Power-Normalized Cepstral Coefficients*)
   - Análise de outras técnicas de reconhecimento, como a técnica *Type-2 Fuzzy Hidden Markov Models*, que prometem melhor desempenho que a técnica HMM.

2. Pretende-se verificar o desempenho do codificador em situações de emoções na fala e ambientes ruidosos. Para isso é necessário treinar o reconhecimento com uma base de voz que proporcione tais características;

3. Outro ponto fundamental do desenvolvimento do codificador é a síntese por concatenação, que necessita de uma variedade de unidades, em diferentes contextos prosódicos e textuais para sintetizar o sinal com uma boa qualidade, mesmo sem se preocupar com o ponto correto para a concatenação das unidades. Deste modo, a síntese do codificador precisa de aprimoramentos para encontrar uma boa relação entre quantidade de unidades para sintetizar o sinal e viabilidade do uso no codificador. Além disso, devem ser analisadas as melhores frases para compor o banco de unidades acústicas de maneira a aprimorar a seleção das unidades.

APÊNDICE A

HTK (*Hidden Markov Models Toolkit*)

O HTK é um *kit* de domínio público utilizado para construir e manipular HMMs. A primeira versão do HTK foi disponibilizada em 1989 e foi desenvolvida pelos pesquisadores do Departamento de Engenharia da Universidade de Cambridge [53, 79]. Este *software* está disponível em [79].

O HTK foi desenvolvido para pesquisas em reconhecimento de voz. Todavia, devido a sua característica de modelar séries temporais, tem sido usado em diversas aplicações como na pesquisa de síntese de voz, reconhecimento de caracteres, entres outros.

O *software* consite em uma biblioteca de módulos e ferramentas escritos na linguagem C. As ferramentas proveem funcionalidades para análise de fala, treinamento dos HMMs, teste e análise dos resultados. Além disto, suporta a implementação de HMMs contínuas e discretas, ambas com múltiplas distribuições gaussianas, o que permite a construção de sistemas com alto grau de complexidade. Neste trabalho foi utilizada a versão mais recente, ou seja, o HTK 3.4.1.

APÊNDICE B

Segmentos fonéticos do português brasileiro

Símbolo	Classificação	Exemplos
p	Oclusiva bilabial desvozeada	**P**ato
b	Oclusiva bilabial vozeada	**B**ala, **B**arco
t	Oclusiva alveolar desvozeada	**T**apa, **T**elha
d	Oclusiva alveolar vozeada	**D**ata, **D**ado
k	Oclusiva velar desvozeada	**C**apa, **C**arro
g	Oclusiva velar vozeada	**G**ato
t∫	Africada alveopalatal desvozeada	**T**ia
dZ	Africada alveopalatal desvozeada	**D**ia
f	Ficativa labiodental desvozeada	**F**aca, **F**ala, **F**arelo
v	Ficativa labiodental vozeada	**V**ento, **V**aca
s	Ficativa alveolar vozeada	**S**ala, Ca**ç**a, **C**ebola
z	Ficativa alveolar vozeada	Ca**s**a, **Z**ero
∫	Ficativa alveopalatal desvozeada	**Ch**á, A**ch**a
Z	Ficativa alveopalatal vozeada	**J**á
x	Fricativa velar desvozeada	**R**ata
R	Fricativa velar vozeada	Ca**r**ga
m	Nasal bilabial vozeada	**M**ala, **M**arca
n	Nasal alveolar vozeada	**N**ada, **N**ervo
$\tilde{y}$	Nasal palatal vozeada	Ba**nh**a, Arra**nh**ado
r	Tepe alveolar vozeado	Ca**r**a, P**r**ata
l	Lateral alveolar vozeada	**L**ata, P**l**ana, **L**uz
w	Lateral alveolar vozeada velarizada	Sa**l**ta, Ma**u**
λ	Lateral palatal vozeada	Ma**lh**a, Caval**lh**eiro

Símbolo	Classificação	Exemplos
i	Vogal alta anterior não-arredondada	Vi
$\tilde{i}$	Vogal alta anterior não-arredondada nasal	Vim
e	Vogal média-alta anterior não-arredondada	Ipê
$\tilde{e}$	Vogal média-alta anterior não-arredondada	Tempo
é	Vogal média-baixa anterior não-arredondada	Pé
a	Vogal baixa central não-arredondada	Pá
$\tilde{a}$	Vogal baixa central não-arredondada nasal	Lã
ó	Vogal média-baixa posterior arredondada	Avó
o	Vogal média-alta posterior arredondada	Avô
$\tilde{o}$	Vogal média posterior arredondada nasal	Tom
u	Vogal alta posterior arredondada	Jacu
$\tilde{u}$	Vogal alta posterior arredondada nasal	Jejum
I	Vogal alta anterior não-arredondada	Vi
i~	Vogal alta posterior arredondada	Vim
ê	Vogal média-baixa central	Ipê

APÊNDICE C

Frases utilizadas no desenvolvimento do codificador

Frases utilizadas para construção do banco de unidades:

1. Um casal de gatos come no telhado.
2. A cantora foi apresentar seu grande sucesso.
3. Lá é um lugar ótimo para tomar uns chopinhos.
4. 0 musical consumiu sete meses de ensaio.
5. Nosso baile inicia após as nove.
6. Apesar desses resultados, tomarei uma decisão.
7. A verdade não poupa nem as celebridades.
8. As queimadas devem diminuir este ano.
9. 0 vão entre o trem e a plataforma é muito grande.
10. Infelizmente não comparecí ao encontro.
11. A sensibilidade indicará a escolha.
12. A Amazônia é a reserva ecológica do globo.
13. O ministério mudou demais com a eleição.
14. Novos rumos se abrem para a informatica.
15. 0 capital de uma empresa depende da produção.
16. Se nao fosse ela, tudo teria sido contido.

17. A principal personagem no filme é uma gueixa.
18. Receba seu jornal em sua casa.
19. A juventude tinha que revolucionar a escola.
20. A atriz terá quatro meses para ensaiar seu canto.

Frases utilizadas na primeira avaliação do codificador:

1. Algumas coisas.
2. A demanda por real.
3. Há cento e setenta bilhões.
4. Cada uma delas.
5. O mercado fica de alto.
6. A perspectiva continua.
7. O desafio agora é.
8. O jogo contra a Suécia.
9. Está instalado na casa.
10. Funcionários do governo.

Frases utilizadas na segunda avaliação do codificador:

1. A casa foi vendida em pressa.
2. Ela tem muita fome.
3. De dia apague a luz sempre.
4. Meu time se consagrou como o melhor.
5. Comer quindim é sempre uma boa pedida.
6. O congresso volta atrás em sua palavra.
7. As crianças conheceram o filhote de ema.

8. A apresentação foi cancelada por causa do som.
9. Uma garota foi presa ontem à noite.
10. O clima não é mais seco no interior.
11. Muito prazer em conhecê-lo.
12. Trabalhei mais do que podia.
13. Hoje eu acordei muito calmo
14. Seu saldo bancário está baixo.
15. Ainda tenho cinco telefonemas para dar.

APÊNDICE D

Publicações

Artigos publicados.

1. ROCHA, R. B., ROCHA, G. B., ALENCAR, M. S. **Codificador de Voz Pessoal.** In: XXX Simpósio Brasileiro de Telecomunicações (SBrT'12), 2012, Brasília.

2. ROCHA, R. B., REGIS, C. D. M., ALENCAR, M. S. **Avaliação da Qualidade de Vídeos Transcodificados após a Transmissão.** In: InfoBrasil, 2011, Fortaleza.

3. ROCHA, R. B., SILVA, T. L. V. N., REGIS, C. D. M., ALENCAR, M. S. **Subjective and Objective Evaluation of Transcoded Video Quality after Transmission.** In: International Workshop on Telecommunications, 2011, Rio de Janeiro.

Capítulo de livro publicado.

1. ROCHA, R. B. **Evolução de Longo Prazo.** In: Telefonia Celular Digital. 3 ed. : Érica, 2012, p. 431-447.

Referências Bibliográficas

[1] F. M. B. Junior. Projeto e Avaliação de Dicionários para Quantização Vetorial de Voz e Imagem. Tese de doutorado, Universidade Federal da Paraíba, Campina Grande, Brasil, Dezembro de 2001.

[2] M. S. Alencar. *Telefonia Celular Digital.* Editora Érica, São Paulo, 2012.

[3] B. P. Lathi. *Modern Digital and Analog Communications Systems.* Oxford University Press, São Paulo, 1988.

[4] ITU-T. ITU-T Recommendation P.800, Methods for Objective and Subjective Assessment of Quality. August 1996.

[5] C. D. M. Regis. Avaliação de Técnicas de Redução da Resolução Espacial de Vídeos para Dispositivos Móveis. Dissertação de mestrado, Universidade Federal de Campina Grande, Campina Grande, Brasil, Março de 2009.

[6] M. de S. Freitas. A Qualidade da Voz em Sistemas de Telecomunicações. Dissertação de mestrado, Universidade Federal Fluminense, 2009.

[7] J. L. A. Carvalho e D. Dias. Técnicas de Codificação de Voz Aplicadas em Sistemas Móveis Celulares.

[8] W. C. A. Costa. Reconhecimento de Fala Utilizando Modelos de Markov Escondidos (HMM's) de Densidades Contínuas. Dissertação de mestrado, Universidade Federal da Paraíba, 1994.

[9] J. P. R. Teixeira. Modelização Paramétrica de Sinais para Aplicação em Sistemas de Conversão texto-Fala. Dissertação de mestrado, Universidade do Porto, 1995.

[10] J. M. Fechine. Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística. Tese de doutorado, Universidade Federal da Paraíba, Campina Grande, Brasil, 2000.

[11] R. de M. L. S. Lamas. Avaliação de Codificadores de Voz em Ambiente VoIP. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, Dezembro 2005.

[12] T. de M. Prego. Aperfeiçoamento do Codificador de Voz CELP. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, Agosto 2007.

[13] B. C. Bispo. Otimização do Codificador de voz CELP. Projeto final, Pontifícea Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil, Dezembro de 2005.

[14] R. da S. Maia. Codificação CELP e Análise Espectral da Voz. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, 2000.

[15] A. R. Fiacador. Estudo e Simulação do Codificador de Voz VSELP do Padrão IS-136. Relatório de projeto final, Universidade de Brasília, 1999.

[16] W. Chung and K. Sangwon. Design of a Variable Rate Algorithm for the CS-ACELP Coder. *IEIC Transactions Inf. & Syst.*, E82-D(10):1364 – 1371, Outubro de 1999.

[17] R. Salami, R. Lefebvre, A. Lakaniemi, K. Kontola, S. Bruhn, and A. Taleb. Extended AMR-WB for High-Quality Audio on Mobile Devices. *IEEE Communications Magazine*, pages 90 – 97, May 2006.

[18] B. Bessette, R. Salami, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Järvinen. The Adaptive Multirate Wideband Speech Codec (AMR-WB). *IEEE Transactions on Speech and Audio Processing*, 10(8):620 – 636, November 2002.

[19] R. Schwartz, J. Klovstad, J. Makhoul, and J. Soresen. A Preliminary Design of a Phonetic Vocoder Based on a Diphone Model. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1980.

[20] F. Soong and B. Juang. A Phonetically Labelled Acoustic Segment (PLAS) Approach to Speech Analysis-Synthesis. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1989.

[21] Y. Hitata and S. Nagawa. A 100 bit/s Speech Coding using a Speech Recognition Technique. *Proceedings of the European Conference on Speech Communication and Technology*, 1989.

[22] J. Picone and G. Doddington. A Phonetic Vocoder. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1989.

[23] P. Peterson, P. Jeanrenaud, and J. Vandergrift. Improving Intelligibility of a 300 B/S Segment Vocoder . *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1990.

[24] K. Tokuda, T. Masuko, J. Hiroi, T. Kobayashi, and T. Kitamura. A Very Low Bit Rate Speech Coders Using HMM-Based Recognition/Synthesis Techniques . *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1998.

[25] R. da S. Maia, R. J. da R. Cirigliano, D. Rojtenbeig, and E C. I. Resende Jr. Mixed-Excited Phonetic Vocoding at 265 BPS. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 796 – 799, 2003.

[26] C. E. de M. Ribeiro. Codificação de Fala Baseada em Segmentos Classificados Foneticamente. Tese de doutorado, Universidade Técnica de Lisboa, 1999.

[27] ITU-T Rec. G.711. General Aspects of Digital Transmission Systems Terminal Equipments – Pulse Code Modulation (PCM) of Voice Frequencies. 1993.

[28] ITU-T Rec. G.726. General Aspects of Digital Transmission Systems Terminal Equipments – 40, 32, 24, 16 kbits/s Adaptive Differential Pulse Code Modulation (ADPCM). 1972.

[29] ITU-T Rec. G.728. General Aspects of Digital Transmission Systems Terminal Equipments – Coding of Speech at 16 kbits/s Using Low-Delay Code Excited Linear Prediction. 1972.

[30] ITU-T Rec. G.729. General Aspects of Digital Transmission Systems Terminal Equipments – Coding of Speech at 8 kbits/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP). 1996.

[31] M. S. Nascimento. Medidas de Qualidade da Voz em Redes IP. Dissertação de mestrado, Universidade Federal do Paraná., Curitiba, 2006.

[32] L. R. Rabiner and B. Juang. *Fundamentals on Speech Recognition*. 1996.

[33] L. R. Rabiner S. E. Levinson and M. M. Sondhi. An Introduction to the Application of the Theory of Probabilist Functions of a Markov Process to Automatic Speech Recognition. *The Bell System Technical Journal*, 62(4):1035–1068, April 1983.

[34] J. R. Deller Jr., J. G. Proakis, and J. H. L. Hansen. *Discrete time Processing of Speech Signals*. Macmillan Publishing Co., 1993.

[35] L. E. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Mathematical Statistics*, 1966.

[36] E. L. Baum et alii. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Mathematical Statistics*, 1970.

[37] J. K. Baker. The Dragon System - An Overview. *IEEE Transactions on Acoustics, Speech and Signal Processing.*, pages 24–29, February 1975.

[38] F. Jelinek, L. R. Bahl, and R. L. Mercer. Design of a linguistic statistical decoder for the continuous speech. *IEEE Transactions on Information Theory.*, (21):250–256, May 1975.

[39] E. D. S. Paranaguá. Reconhecimento de Locutores Utilizando Modelos de Markov Escondidos Contínuos. Dissertação de mestrado, Instituto Militar de Engenharia, Rio de Janeiro, Brasil, Maio de 1997.

[40] A. M. Selmini. Sistema Baseado em Regras para o Refinamento da Segmentação Automática de Fala. Tese de doutorado, Universidade Estadual de Campinas, Campinas, Brasil, Agosto de 2008.

[41] X. D. Huang, A. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, 1990.

[42] X. D. Huang and M. A. Jack. Hidden Markov Modelling of Speech based on a Semicontinuous Model. *Eletronics Letters*, 1988(a).

[43] X. D. Huang and M. A. Jack. Performance Comparison between Semicontinuous and discrete Hidden Markov Models of Speech. *Eletronics Letters*, 1988(b).

[44] L. R. Rabiner S. E. Levinson and M. M. Sondhi. An Introduction to the Application of the Theory of Probabilist Functions of a Markov Process to Automatic Speech Recognition. *The Bell System Technical Journal*, 62(4):1035–1068, April 1983.

[45] S. E. Levinson L. R. Rabiner and M. M. Sondhi. On the Application of Vector Quantization and Hidden MarkovModels to Speaker-independent, Isolated Word Recognition. *The Bell System Technical Journal*, 62(4):1075–1105, April 1983.

[46] J. Martins. Avaliação de diferentes técnicas para reconhecimento de fala. Tese de doutorado, 1997.

[47] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[48] D. D. C. da Silva. Reconhecimento de Fala Contínua para o Português Brasileiro em Sistemas Embarcados. Tese de doutorado, Universidade Federal de Campina Grande, Campina Grande, Brasil, Dezembro de 2011.

[49] L. E. et alii. BAUM. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[50] B. H. Juang and L. R. Rabiner. Hidden Markov Models for Speech Recognition. *Technometrics*, 1991.

[51] R. Dias. Normalização de Locutor em Sistema de Reconhecimento de Fala. Tese de doutorado, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, 2000.

[52] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. Large vocabulary continuous speech recognition using HTK. *Proceedings ICASSP*, 19:125 – 128, April 1994.

[53] S. Young et al. *The HTK Book*. Cambridge University Engineering Department, 2009.

[54] R. Teruszkin and F. Vianna. Implementation of a Large Vocabulary Continuous Speech Recognition System for Brazilian Portuguese. *Journal of Communication and Information System.*, 2006.

[55] C. P. A. da Silva. Um *Software* de Reconhecimento de Voz para Português Brasileiro. Dissertação de mestrado, Universidade Federal do Pará, Belém, Brasil, 2010.

[56] R. T. Tevah. Implementação de um Sistema de Reconhecimento de Contínua Com Amplo Vocabulário Para o Português Brasileiro. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, Junho de 2006.

[57] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, New Jersey, 1978.

[58] L. F. M. P. Coelho. Etiquetagem Automática de Sinais de Fala Segmentação e Classificação Fonética. Dissertação de mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, Fevereiro de 2005.

[59] S. Davis and P. Merlmestein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on ASSP*, pages 357–366, August 1980.

[60] J. Picone. Signal Modeling Techniques in Speech Recognition . *Proceedings of the IEEE*, 1993.

[61] P. Woodland and S. Young. The HTK Tied-State Continuous Speech Recognizer. *Proceedings Eurospeech'93*, 1993.

[62] S. Young and P. Woodland. State clustering in hmm-based continuous speech recognition. *Computer Speech and Language*, 1994.

[63] M. Hwang and X. Huang. Shared Distribution Hidden Markov Models for Speech Recognition. *IEEE Transactions Speech and Audio Processing*, 1993.

[64] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny. Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees. *Proc DARPA Speech and Natural Language Processing Workshop*, pages 264–270, February 1991.

[65] A. Kannan, M. Ostendorf, and J. R. Rohlicek. Maximum Likelihood Clustering of Gaussians for Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 2(3):453–455, 1994.

[66] S. J. Young, J. J. Odell, and P. C. Woodland. Tree-Based State Tying for High Accuracy Acoustic Modeling. *Proceedings Human Language Technology Workshop*, pages 307–312, March 1994.

[67] Laboratório de Processamento de Sinais. FalaBrasil – Reconhecimento de Voz para o Português Brasileiro. http://www.laps.ufpa.br/falabrasil/. Visitado em 23 de outubro de 2011.

[68] G. C. da Silva e P. E. D. Pinto. Análise Comparativa de Métodos de Compactação de Dados sem Perda.

[69] A. Alcaim, J. A. Solewicz, and J. A. de Moraes. Frequência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecornunicacoes*, 7(1), Dezembro 1992.

[70] P. Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[71] T. Dutoit. *A Introduction to Text-to-Speech Synthesis.* Academic Publishers, 2011.

[72] K. W. A. X. da Silva. Sistema de Conversão Texto-Fala com Busca Otimizada de Unidades Acústicas em Banco de Voz. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Dezembro 2011.

[73] F. O. Simões. Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil. Tese de mestrado, Unicamp, 1999.

[74] V. L. Latsch. Construção de um Banco de Unidades para Síntese da Fala por Concatenação no Domínio Temporal. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, 2005.

[75] E. da S. Morais. Algoritmo OPWI e LDM-GA para Sistemas de Conversão Texto-Fala de Alta Qualidade Empregando a Tecnologia SCAUS. Tese de doutorado, Unicamp, 2006.

[76] M. L. da C. Neto. Um Modelo para Geração de Prosódia de Palavras em Conversores Texto-Fala para a Língua Portuguesa Falada no Brasil. Tese de doutorado, Universidade Federal de Campina Grande, 2004.

[77] E. A. M. Klabbers. Segmental and Prosodic Improvements to Speech Generation. Tese de doutorado, Technische Universiteit Eindhoven, Netherlands, 2000.

[78] L. F. Millão A. K. Gonçalves B. B. Junior A. F. Vieira E. M. Farias C. R. Martins A. M. P. V. dos Santos P. T. C. Lopes I. A. Martins D. O. da C. Pol e C. J. dos S. Gonçalves A. R. Teixeira, C. de La R. Freitas. Relação entre Deficiência Auditiva, Idade, Gênero e Qualidade de Vida de Idosos. *Arquivos Internacionais de Otorrinolaringologia*, 12:62–70, 2008.

[79] Cambridge University Engineering Department. HTK Speech Recognition Toolkit. http://htk.eng.cam.ac.uk/. Visitado em 9 de novembro de 2011.