Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Coordenação de Pós-Graduação em Ciência da Computação

Reconhecimento Automático de Palavras Isoladas, Independente de Locutor, para Sistemas Embarcados

Maria de Lourdes do Nascimento Neta

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande – Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação

Área de Concentração: Ciência da Computação

Linha de Pesquisa: Redes de Computadores e Sistemas Distribuídos

Elmar Uwe Kurt Melcher (Orientador)

Joseana Macêdo Fechine Régis de Araújo (Orientadora)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

N244r Nascimento Neta, Maria de Lourdes do.

Reconhecimento automático de palavras isoladas, independente de locutor, para sistemas embarcados / Maria de Lourdes do Nascimento Neta. – Campina Grande, 2012.

91 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática.

Orientadores: Prof. Dr. Elmar Uwe Kurt Melcher, Profª. Drª. Joseana Macêdo Fechine Régis de Araújo.

Referências.

1. Reconhecimento de Palavras Isoladas. 2. Sistemas Embarcados. 3. Baixo Consumo Energético. I. Título.

CDU 004.383.3(043)

ECONHECIMENTO AUTOMÁTICO DE PALAVRAS ISOLADAS, INDEPENDENTE DO LOCUTOR, PARA SISTEMAS EMBARCADOS "

MARIA DE LOURDES DO NASCIMENTO NETA

DISSERTAÇÃO APROVADA EM 16/05/2012

ELMAR UWE KURT MELCHER, Dr.
Orientador(a)

JOSEANA MACÊDO FECHINE RÉGIS DE ARAÚJO, D.Sc
Orientador(a)

JOSÉ EUSTAQUIO RANGEL DE QUEIROZ, D.Sc
Examinador(a)

MARCOS RICARDO ALCÂNTARA MORAIS, D.Sc
Examinador(a)

CAMPINA GRANDE - PB

Resumo

O presente trabalho insere-se na área de Reconhecimento da Fala e nele propõe-se um **sistema de reconhecimento automático de palavras isoladas independente do locutor,** para sistemas embarcados dependentes de bateria que apresentam o requisito de baixo consumo. Considerando os critérios de baixo consumo e visando uma implementação em *hardware*, optou-se pelo uso de técnicas simples para o reconhecimento, a saber: (i) uso de coeficientes cepstrais, obtidos a partir dos coeficientes LPC, na composição do vetor de características; (ii) uso da quantização vetorial, na obtenção de padrões; e (iii) regra de decisão, baseada na distância euclidiana. O sistema proposto foi implementado em *software* e validado a partir de uma base de dados composta de 1.232 sentenças de treinamento e 770 sentenças de teste, proporcionando uma taxa de reconhecimento de 96,36%. Comparando-se com modelagens mais complexas, que utilizam Modelos de Markov Escondidos de Densidades Contínuas, modelo linguístico e coeficientes mel cepstrais, que proporcionam uma taxa de reconhecimento de 100%, a técnica proposta se mostra adequada, dado à pequena redução do desempenho para uma redução significativa da complexidade e, consequentemente, do consumo, em uma implementação em hardware.

Palavras-chave: Reconhecimento de Fala de Palavras Isoladas, Sistemas Embarcados, Baixo Consumo Energético.

Abstract

In this dissertation, a speaker-independent speech recognition system for isolated words is presented. Here, we focus on battery-dependent embedded systems that have the requirement of lower power consumption. Considering this requirement and targeting a hardware implementation, we chose to use simpler techniques for recognition such as cepstrum coefficients obtained from LPC coefficients to compose the feature vector. It was also used quantization vectors to generate patterns and decision rule based on Euclidean distance. The proposed system was implemented in software and validated with a database composed of 1,232 training sentences and 770 testing sentences. It was achieved a recognition rate of 96.36%. Compared to more complex modeling that achieved recognition rate of 100% using continuous densities Hidden Markov Models, linguistic models and mel-frequency cepstrum coefficients (MFCC), the proposed technique is highly satisfactory. A significant reduction in complexity and, consequently, in power consumption, necessary for hardware implementation, is achieved by paying the price of only a small reduction in performance.

Keywords: Isolated Word Speech Recognition, Embedded Systems, Low Power Consumption.

"Feliz aquele que transfere o que sabe
e aprende o que ensina."
Cora Coralina

"Porque és precioso a meus olhos,
porque eu te aprecio e te amo,
permuto reinos por ti, entrego nações
em troca de ti. Fica tranquilo, pois
estou contigo, (...)."
Is 43, 4-5b

Dedicatória

Este trabalho é dedicado a Deus, em primeiro lugar, aos meus pais, José Nivaldo e Maria do Socorro, aos meus irmãos Edmar, Sandra, José Luís e Sérgio e às minhas cunhadas Christiane e Alânia.

Agradecimentos

Gostaria de agradecer a Deus, por seu amor incondicional e sem limites e por ter me ajudado em todos os momentos.

À intercessão de Nossa Senhora em minha vida.

À minha família, pelo amor, apoio, orações, revisões do texto e incentivos constantes.

Aos professores Elmar e Joseana, pela orientação, dedicação e apoio, principalmente nos momentos mais difíceis do trabalho.

A todos os membros da COPIN, em especial ao professor Nazareno, pelo parecer favorável à defesa, apresentado à Câmara Superior de Pós-Graduação.

Às secretárias da SODS, Socorro e Rossana, pela gentileza e palavras de incentivos.

Aos amigos e colegas que contribuíram com elocuções para a construção da base de dados empregada neste trabalho: Bruno Vitorino, Fabrício, Gabriela Marques, Helder, Henrique, Isabela, Karina Medeiros, Laércio, Natasha, Savyo e Sérgio Espínola.

A toda a equipe LAD, que participou do projeto SPVR, em especial a Adalberto e Jorgeluis, pelos esclarecimentos prestados ao longo deste trabalho.

Aos amigos e colegas da pós-graduação, pelo incentivo, amizade e contribuições dadas ao trabalho.

Às amigas Anne Caroline, Daniella, Rute, Henza e Ana Karina, pela amizade e palavras de incentivo.

Ao CETENE/LINCS, pelo incentivo dado à pós-graduação.

Ao CNPq, pela bolsa.

Conteúdo

Lista de Siglas e Abreviaturas

A/D	Analógico/Digital
CDE	Comparação por Distância Euclidiana
DV	Detector de Voz
FIR	*Finite Impulse Response*
GPS	*Global Positioning System*
HMM	*Hidden Markov Model*
LBG	*Linde-Buzo-Gray*
LD	Limiar de Decisão
LPC	*Linear Prediction Coding*
PDSV	Processamento Digital de Sinais de Voz
PE	Pré-ênfase
QA	Quantidade de Acertos
QD	Quantidade de Desconhecidos
QE	Quantidade de Erros
QV	Quantização Vetorial
RD	Regra de Decisão
SRF	Sistemas de Reconhecimento de Fala
SRL	Sistemas de Reconhecimento de Locutor
SRV	Sistemas de Resposta Vocal
TF	Tempo de Fim
TI	Tempo de Início

Lista de Figuras

Lista de Tabelas

1 Introdução

A voz é um meio de comunicação prático e usual entre humanos. Por meio da voz, se consegue transmitir quantidades significativas de informação em intervalos de tempo curtos – a maioria das pessoas consegue falar facilmente 200 palavras por minuto, porém, poucas conseguem digitar, em um teclado, mais de 60 palavras por minuto (LEE, HAUPTMANN e RUDNICKY, 1990).

Outra característica da comunicação vocal, diz respeito à facilidade de se identificar particularidades da pessoa que transmite a mensagem, tais como: grupo sociocultural, estado emocional, estado de saúde, região onde habita, etc (RABINER e SCHAFER, 1978).

Com o crescente desenvolvimento tecnológico, a comunicação homem-máquina tem se tornado cada vez mais habitual. Segundo Rabiner e Schafer (1978) e Vidal (2006), esta comunicação pode se tornar mais fácil e produtiva com o uso de interfaces vocais. Fato evidenciado no trabalho de Sherwani et al. (2008). Nesse trabalho, é apresentada uma comparação entre interfaces de voz e de toque em um serviço de informação por telefone. Como resultado, as interfaces de voz se mostraram mais produtivas, inclusive para usuários com grau de escolaridade baixo.

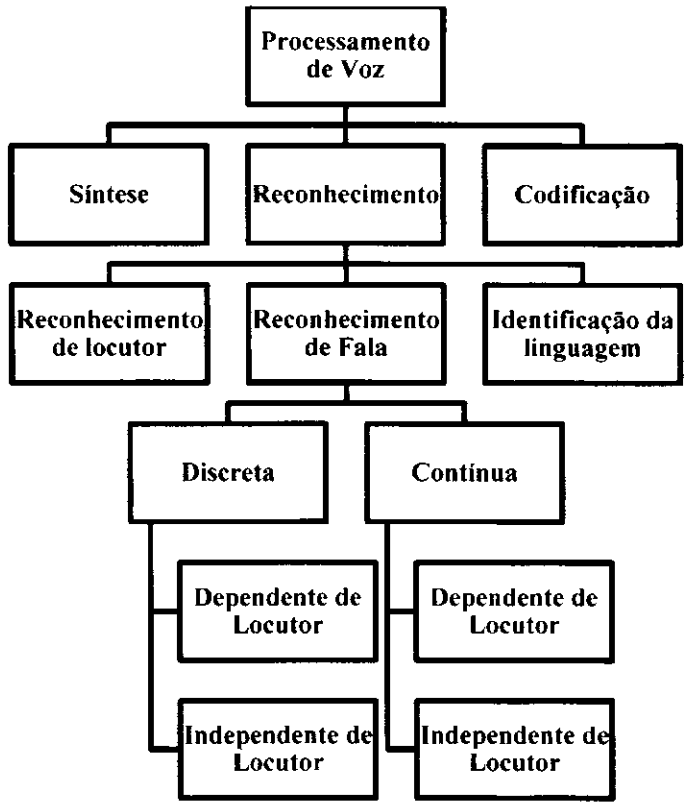
Porém, as interfaces de voz, na interação com as máquinas, ainda são pouco usuais se comparadas às interfaces de toque. Uma das razões apontadas para esta constatação diz respeito às expectativas elevadas dos usuários dos dispositivos eletrônicos em relação às aplicações de voz (GOMES, 2007). Eles idealizam aplicações que reconhecem discursos com a complexidade de um diálogo natural entre humanos (diálogo irrestrito) e com a mesma taxa de acerto de um ser humano.

O estudo da comunicação vocal homem-máquina se divide nas seguintes áreas (O'SHAUGHNESSY, 2000):

- Resposta Vocal (síntese)
- Reconhecimento de Locutor
- Reconhecimento de Fala

O objeto de estudo deste trabalho está inserido na área de Reconhecimento de Fala. Na Figura 1.1, está apresentada uma classificação geral do processamento de voz, com ênfase no reconhecimento de fala (CAMPBELL, 1997).

Figura 1.1: Classificação geral da área de processamento de sinais de voz.



Adaptado de (CAMPBELL, 1997).

Sistemas de Resposta Vocal (SRV) são projetados para responder a solicitações de informação utilizando-se de mensagens faladas. Nesses sistemas, também chamados de sistemas de síntese de voz, a comunicação se faz no sentido máquina-homem (RABINER e SCHAFER, 1978).

Esses sistemas utilizam amostras de voz armazenadas, tais como, frases, palavras, fonemas, ou segmentos mais curtos, para compor a mensagem de saída. A composição de saída é formada com base em regras linguísticas dos idiomas (regras gramaticais e fonéticas).

Sintetizadores de voz estão presentes em sistemas automáticos de informação de preços, de voos, de produtos, em sistemas de auxílio aos portadores de necessidades especiais, nos automóveis com mensagens de alerta, em dispositivos GPS (*Global Positioning System*),

em serviços de leitura de correio eletrônico via telefone, dentre outros (RABINER e SCHAFER, 2010).

Os sistemas de reconhecimento de locutor (SRL), por sua vez, têm como objetivo reconhecer um locutor por meio da voz. Esses sistemas são classificados como sendo de verificação ou identificação de locutor.

As aplicações destinadas à verificação de locutor respondem se um dado locutor é quem alega ser, enquanto que, as aplicações de identificação são capazes de identificar quem é o locutor, dentro de um universo de locutores (MÜLLER, 2007). Nesses sistemas, a comunicação vocal é feita no sentido homem-máquina.

Os SRL são úteis na realização de operações de autenticação pela voz nas áreas de segurança e criminalística. Na área de segurança, esses sistemas atuam na restrição ao acesso, à informação confidencial ou a conteúdo, por exemplo (SHIRALI-SHAHREZA, SAMETI e SHIRALI-SHAHREZA, 2008). Na área de criminalística, auxiliam no reconhecimento de indivíduos, uma vez que, as características vocais são únicas para cada indivíduo (MÜLLER, 2007).

Nos sistemas de reconhecimento de fala (SRF), da mesma forma que nos SRL, a comunicação vocal acontece no sentido homem-máquina. Esses sistemas têm por objetivo reconhecer uma determinada elocução de uma sentença ou "entender" um texto falado (CHEN, 2005).

Existe uma diferença sutil entre reconhecimento e entendimento. No reconhecimento, é identificada toda a sentença pronunciada, enquanto que, no entendimento procura-se identificar os vocábulos-chave pronunciados na(s) sentença(s). Devido a essa diferenciação, o reconhecimento exato de palavras manipula um vocabulário limitado, com um número pequeno de usuários e a pronúncia das palavras acontece de forma pausada. Em se tratando do entendimento, trata-se, normalmente, da voz contínua com grande vocabulário (RABINER e SCHAFER, 1978).

De uma forma geral, os sistemas de reconhecimento automático de fala são classificados como pertencentes às seguintes categorias: Sistemas de Reconhecimento de Palavras Isoladas (fala discreta), Sistemas de Reconhecimento de Palavras Conectadas (fala

contínua), os quais podem ser dependentes ou independentes do Locutor (O'SHAUGHNESSY, 2000).

A diferença entre os sistemas de reconhecimento de palavras isoladas e aqueles de palavras conectadas reside no fato de que nos primeiros é exigida uma pausa curta antes e depois das sentenças que devem ser reconhecidas, enquanto, nos últimos não é necessária pausa, sendo a voz pronunciada de forma mais natural pelo usuário. Contudo, este tipo de comunicação possui algumas limitações, em virtude da complexidade da voz humana, que depende de fatores, tais como entonação, velocidade da fala, fusão do último e primeiro fonema de palavras consecutivas, estado emocional do usuário, etc. (O'SHAUGHNESSY, 2000).

Os sistemas dependentes de locutor são treinados para reconhecer as características específicas da voz de seus usuários. Desta forma, somente os usuários cadastrados são reconhecidos por esses sistemas. Os sistemas independentes de locutor são "insensíveis aos usuários", ou seja, não são dependentes às características específicas da voz dos locutores (CHEN, 2005).

Algumas das aplicações de reconhecimento de fala mais comuns são atendimento telefônico automático, acesso a menus de celulares ou de outros equipamentos eletrônicos, transcrição de fala para texto e aplicações de comando-controle, dentre outros.

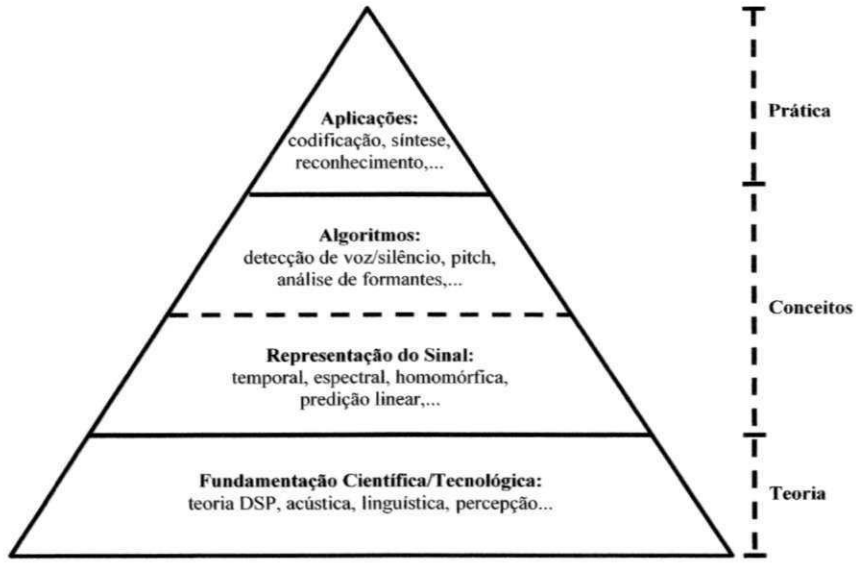
Tanto os sistemas de reconhecimento de locutor quanto aqueles de reconhecimento de fala podem apresentar dificuldade no processo de reconhecimento, pois, estão sujeitos a fatores como, ruído ambiental, problemas na captação da voz devido à qualidade e dependência de microfones, relevância da base de dados de voz utilizada no treinamento desses sistemas e estado de saúde vocal do locutor, dentre outros (CHEN e MASI, 2000; VAREJÃO, 2001; NEVES et al., 2008).

1.1 Motivação

Em se tratando de interfaces de reconhecimento de fala, ainda não é tecnicamente possível corresponder às expectativas dos usuários em um reconhecimento irrestrito (GOMES, 2007),

embora já se tenha uma base teórica e conceitual sólida sobre Processamento Digital de Sinais de Voz (PDSV), conforme ilustrado na Figura 1.2.

Figura 1.2: Pirâmide tecnológica – teoria, conceitos e prática.



Adaptado de (RABINER e SCHAFER, 2010).

Segundo Rabiner e Schafer (2010, p. 2), os conceitos e fundamentos básicos da área de PDSV, conhecidos há algum tempo, continuarão sendo a base das aplicações das próximas décadas e um crescimento mais significativo da prática (aplicações) é esperado (Figura 1.2). Desta forma, as aplicações na área de PDSV ainda utilizarão os conceitos e teorias conhecidos até o momento.

Os sistemas de processamento de voz são o resultado de um processo de conversão de teoria e conceitos em prática. Esse processo envolve a avaliação de requisitos, metas e prós e contras da aplicação alvo, como também a habilidade de produzir implementações (RABINER e SCHAFER, 2010).

Atualmente, o mercado vivencia o crescimento de dispositivos eletrônicos controlados por sistemas embarcados, tais como, telefones celulares, *tablets*, GPS e etc. A maioria destes dispositivos é móvel e agrega várias funcionalidades que antes eram exclusividade dos PC, tais como, navegação na web, acesso a e-mails e envio de mensagens instantâneas pela internet, visualização e edição de arquivos, dentre outros.

Porém, com o crescente número de funcionalidades, a navegação pelos menus está se tornando mais difícil, uma vez que, tais dispositivos geralmente apresentam dimensões físicas reduzidas e o espaço para digitação é limitado (JIANG, MA e CHEN, 2010). Dessa forma, outras formas de interação são necessárias para facilitar a navegação. A interação pela fala, seja ela por comandos (palavras isoladas) ou frases, torna mais fácil e produtiva esta forma de comunicação (RABINER e SCHAFER, 1978; VIDAL, 2006; SHERWANI et al., 2008).

O aumento de funcionalidades de tais dispositivos móveis, também, acarreta em uma redução na durabilidade da carga de suas baterias (JIANG, MA e CHEN, 2010). Assim sendo, *hardware* e *software* dedicados, que consumam menos recursos energéticos, são desejáveis para compor tais dispositivos.

Para a implementação de um projeto de *hardware*, rigorosas metodologias devem ser seguidas para evitar que falhas surjam somente depois que o *hardware* prototipado tenha sido integrado em algum sistema. No processo de desenvolvimento de hardware, a detecção de problemas funcionais e comportamentais deve ser feita na etapa de verificação funcional[1] (SILVA, 2007; OLIVEIRA, 2010). Essa etapa deve ser realizada a partir da comparação de dois modelos, o modelo sendo desenvolvido (*hardware*) e o modelo ideal que reflete a especificação (modelo de referência) – a comparação dos modelos ocorre em um ambiente de simulação denominado *testbench* (BERGERON, 2003).

1.2 Objetivos

1.2.1 Objetivo Geral

O presente trabalho objetiva o desenvolvimento de um sistema de reconhecimento automático de palavras isoladas, independente de locutor, a ser utilizado como modelo de referência para futuras implementações em *hardware*.

---

[1] Processo usado para demonstrar que o objetivo do projeto é preservado em sua implementação (BERGERON, 2003).

1.2.2 Objetivos Específicos

- Construir uma base de dados, composta de palavras isoladas (comandos), para as fases de treinamento e de reconhecimento do sistema.
- Analisar e adequar as técnicas ao reconhecimento de fala, implementada em *software*, tendo em vista obter um modelo com a configuração mais adequada a uma implementação em *hardware*, com a exigência de redução do consumo energético.
- Realizar uma análise comparativa do modelo proposto com o modelo implementado no *software* de reconhecimento de voz Sphinx[2], o qual utiliza técnicas que representam o estado da arte na área de Reconhecimento de Fala (WALKER et al., 2004; VOJTKO, KOROSI e ROZINAJ, 2008).

1.3 Estrutura

O restante da dissertação encontra-se organizado conforme descrição a seguir.

- Capítulo 2 – Neste capítulo, são apresentados os conceitos relacionados ao mecanismo de produção da voz humana, à modelagem matemática para a produção da voz, como também, à tarefa de reconhecimento de padrões de fala, com a descrição das fases de treinamento e de reconhecimento (teste), e das etapas envolvidas em cada uma delas.
- Capítulo 3 – As principais características do sistema de reconhecimento automático de palavras isoladas, independente de locutor são apresentadas. Neste capítulo, também é descrita a base de dados utilizada, como também o método para a determinação dos limiares do módulo de detecção de voz.
- Capítulo 4 – São apresentados os resultados obtidos a partir das configurações do sistema simulado neste trabalho. Além disso, são descritas considerações para uma implementação em *hardware* do SRF.
- Capítulo 5 – Neste capítulo, são apresentadas as considerações finais e as sugestões para os trabalhos futuros.

---

[2] http://cmusphinx.sourceforge.net/

2 Fundamentação Teórica

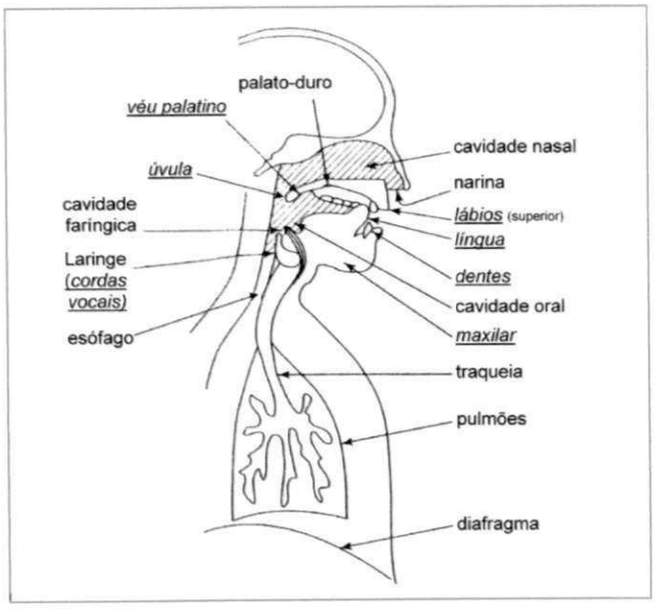
2.1 O Mecanismo de Produção da Voz

Os sinais de voz são compostos por uma sequência de sons, que resultam da ação de certos órgãos sobre a corrente de ar vinda dos pulmões. Para sua produção, três condições são necessárias (CUNHA, 2007): a corrente de ar, obstáculos encontrados por essa corrente e uma caixa de ressonância.

Estas condições são criadas pelo aparelho fonador humano (Figura 2.1), que é constituído das seguintes partes (CUNHA, 2007):

- pulmões, brônquios e traqueia – órgãos respiratórios que fornecem a corrente de ar;
- laringe – onde se localizam as cordas vocais ou dobras vocais, que produzem a energia sonora;
- cavidades supralaríngeas (cavidades faríngea, oral e nasal), que funcionam como caixas de ressonância.

Figura 2.1: Aparelho fonador humano.

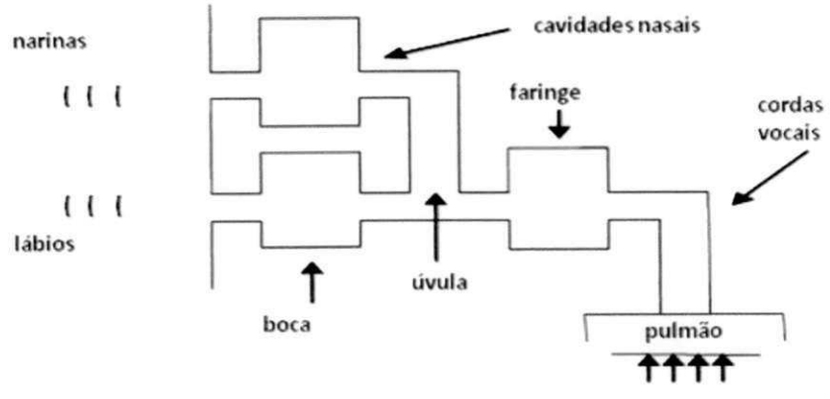


Adaptado de (MAIA, 2010).

A produção da voz inicia-se com a compressão dos pulmões pelo diafragma, que causa uma pressão nos brônquios que expelem o ar através da traqueia. A traqueia faz a interligação entre os pulmões (esquerdo e direito) e a laringe. Chegando à laringe, o fluxo de ar segue pela cavidade faríngea para as cavidades oral e/ou nasal, acabando por sair pela boca e/ou narinas (MAIA, 2010).

A produção da voz também pode ser vista como um sistema de filtragem acústica (Figura 2.2). O filtro principal desse sistema corresponde ao trato vocal e ao trato nasal, cuja fonte de excitação é um sinal que simula o efeito do ar oriundo dos pulmões. O trato vocal começa na abertura entre as dobras vocais e termina nos lábios. O comprimento médio do trato vocal em homens, mulheres e crianças é cerca de 17 cm, 14 cm e 10 cm, respectivamente. Ao longo do trato vocal, a área da seção transversal pode variar de 0 cm^2 (completamente fechado) a até 20 cm^2. O trato nasal começa na úvula e termina nas narinas, sendo o seu comprimento médio de 12 cm num homem adulto. Quando a úvula é abaixada, o trato nasal é acusticamente acoplado ao trato vocal para produzir sons nasais. Quando a úvula encontra-se levantada, a ligação fica completamente fechada e o fluxo de ar atravessa apenas o trato vocal. O sistema de filtragem acústica descrito apresenta uma impedância de radiação. Esta impedância representa o efeito dos lábios na produção da fala (RABINER e SCHAFER, 1978; FECHINE, 2000; MAIA, 2010).

Figura 2.2: Modelo acústico do aparelho fonador.



Adaptado de (RABINER e JUANG, 1993).

Em virtude das limitações dos órgãos humanos de produção de voz e do sistema auditivo, a comunicação humana típica é limitada a 8 kHz (RABINER e SCHAFER, 1978; MALKIN, 2006).

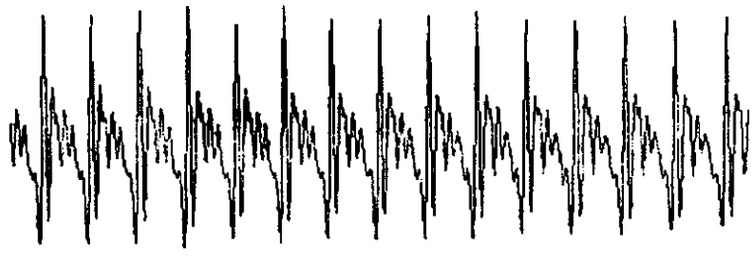
2.1.1 Tipos de Excitação: Classificação dos Sons da Voz

De acordo com o modo de excitação, os sons da voz podem ser classificados em três classes distintas: sons sonoros, sons surdos e sons explosivos (RABINER e SCHAFER, 1978).

2.1.1.1 Sons sonoros

Os sons sonoros acontecem quando o fluxo de ar oriundo dos pulmões passa pela laringe e as dobras vocais interrompem esse fluxo de forma quase periódica, excitando assim o trato vocal (MAIA, 2010). O trato vocal, desta forma, atua como um ressonador, modificando o sinal de excitação e produzindo frequências de ressonância. Essas frequências são denominadas formantes e caracterizam os diferentes sons sonoros (RABINER e SCHAFER, 1978). Na Figura 2.3, é apresentada a forma de onda do som sonoro /a/.

Figura 2.3: Forma de onda do fonema /a/.

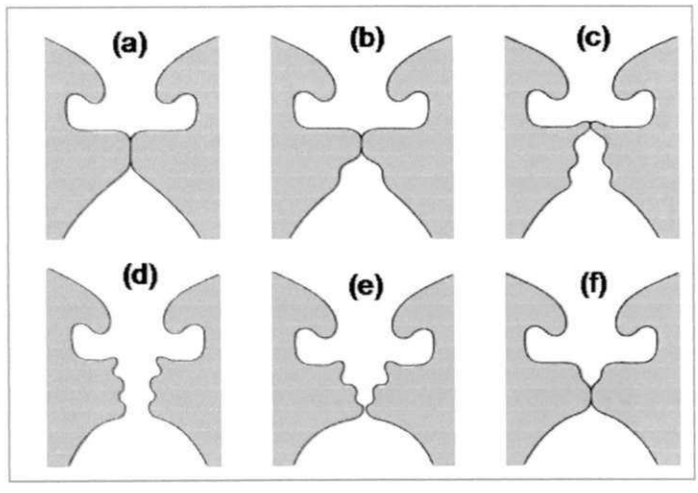


Fonte: (FECHINE, 1994).

As dobras vocais localizam-se na parte inferior da laringe e o espaço entre elas recebe o nome de glote. Estando a glote completamente fechada, o fluxo de ar vindo dos pulmões é interrompido e a pressão subglótica vai aumentando até que se consegue vencer a resistência das dobras vocais, que começam a separar-se. Quando as dobras vocais se afastam, ocorre a

liberação do ar pressionado, gerando um pulso de ar de curta duração. O escoamento do ar reduz a pressão subglótica e possibilita uma nova aproximação das dobras vocais. A pressão subglótica, então, começa novamente a aumentar e o ciclo é repetido (RABINER e SCHAFER, 1978; MAIA, 2010). Na Figura 2.4, tem-se a representação do ciclo vibratório das dobras vocais.

Figura 2.4: Ciclo vibratório das dobras vocais. (a) Glote fechada; (b) e (c) Aumento da pressão subglótica e separação das dobras vocais; (d) Dobras vocais afastadas e liberação do ar; (e) e (f) Diminuição da pressão subglótica e aproximação das dobras vocais.



Fonte: (MAIA, 2010).

O tempo decorrido entre duas sucessivas aberturas da glote chama-se de período fundamental (T_0) e a frequência de abertura, chamada frequência fundamental (F_0), é dada pela Equação 2.1.

$$F_0 = \frac{1}{T_0}. \tag{2.1}$$

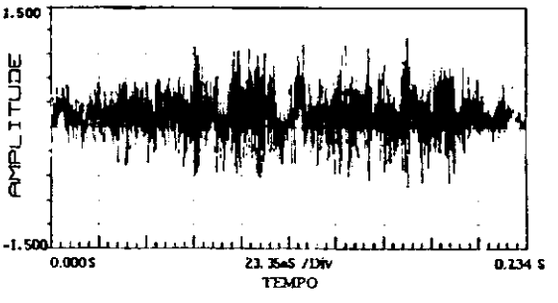
A frequência fundamental dos sons sonoros em homens adultos está situada na faixa entre 80-120 Hz, já em mulheres adultas um valor típico é de 240 Hz, enquanto que para as crianças pode-se encontrar valores em torno de 350 Hz (FELLBAUM, 1984). Valores de frequência tão distintos são explicados pelo fato da frequência fundamental variar de acordo com parâmetros tais como: o comprimento, largura e tensão das dobras vocais, como também, com a porção membranosa dessas, com a cartilagem da tireoide e com a largura da cavidade laríngea (GUIMARÃES, 2007).

2.1.1.2 Sons Surdos

A produção dos sons surdos (ou fricativos) acontece devido à obstrução em algum ponto do trato vocal, que ao receber o fluxo de ar vindo dos pulmões, produz turbulências. O som produzido dessa forma tem características ruidosas, com concentração relativa de energia nas mais altas componentes de frequência do espectro do sinal de voz (RABINER e SCHAFER, 1978). As diferenças entre os diferentes sons surdos dependem do tipo de obstrução no trato vocal (MAIA, 2010).

Durante o processo de produção dos sons surdos, não ocorre vibração das dobras vocais, ou seja, a glote permanece aberta. Na Figura 2.5, é apresentada a forma de onda do som surdo /s/.

Figura 2.5: Forma de onda do fonema /s/.

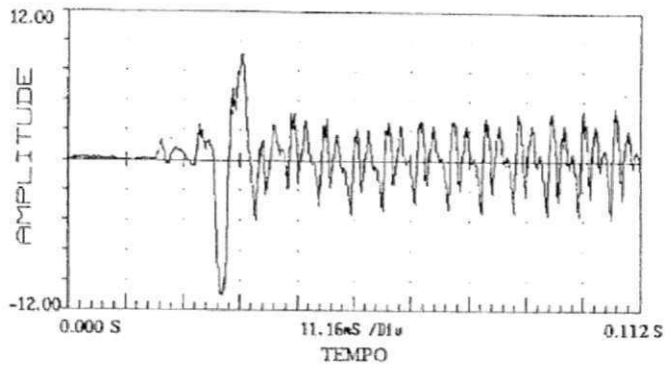


Fonte: (FECHINE, 2000).

2.1.1.3 Sons Explosivos

A geração dos sons explosivos (ou oclusivos) é antecedida de um período de silêncio. Nesse período, a boca encontra-se fechada, e o ar vindo dos pulmões faz pressão nesta cavidade. Com o aumento da pressão, a oclusão é rompida de forma brusca, gerando assim, um pulso que excita o aparelho fonador (RABINER e SCHAFER, 1978; FECHINE, 2000). Na Figura 2.6, é apresentada a forma de onda do som explosivo /p/ na palavra aplausos.

Figura 2.6: Forma de onda do fonema /p/.



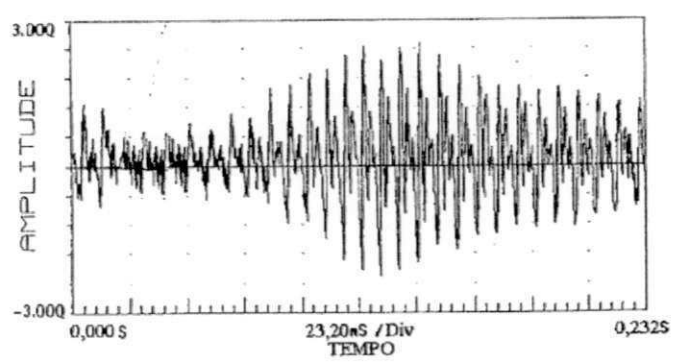
Fonte: (FECHINE, 2000)

2.1.1.4 Sons com Excitação Mista

Os sons constituídos por uma primeira fase fricativa ou oclusiva, seguida de uma fase sonora, são caracterizados como sons de excitação mista.

Os sons fricativos sonoros são o resultado da combinação de excitação turbulenta e vibração das dobras vocais. Nos períodos em que a pressão glótica é máxima, o escoamento de ar, através das obstruções do trato vocal, torna-se turbulento, gerando assim, a componente fricativa do som. A componente sonora, por sua vez, é formada quando a pressão glótica cai abaixo de um dado valor, neste ponto, ocorre à transição de ondas de pressão turbulentas para ondas de pressão de caráter mais suave e periódico (RABINER e SCHAFER, 1978). Na Figura 2.7, é apresentada a forma de onda do som fricativo sonoro /z/.

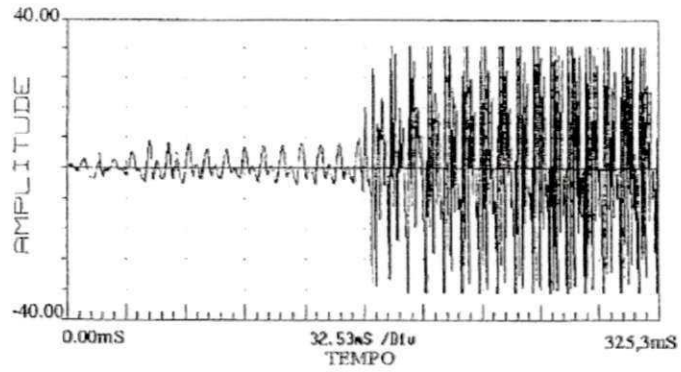
Figura 2.7: Forma de onda do fonema /z/.



Fonte: (FECHINE, 2000)

Os sons oclusivos sonoros são o resultado da combinação de silêncio, período de aumento de pressão na cavidade oral, ruptura brusca de pressão, seguida de uma estabilização de pressão com interrupção periódica do fluxo de ar. Na Figura 2.8, é mostrada a forma de onda do som oclusivo sonoro /b/.

Figura 2.8: Forma de onda do fonema /b/.



Fonte: (FECHINE, 2000)

2.2 Modelo para Produção da Voz

Para a determinação de um modelo apropriado para a representação dos sons da voz, os seguintes efeitos devem ser considerados (RABINER e SCHAFER, 1978):

- Variação da configuração do trato vocal ao longo do tempo;
- Perdas próprias por condução de calor e fricção nas paredes do trato vocal;
- Maciez das paredes do trato vocal;
- Radiação do som pelos lábios;
- Junção nasal;
- Excitação do som no trato vocal, etc.

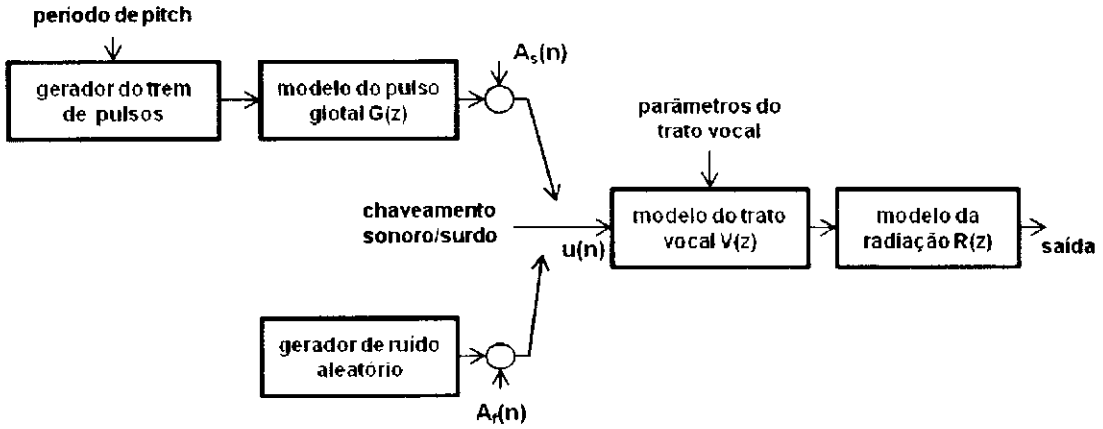
Um modelo para a produção da voz, que leva em conta efeitos de propagação e de radiação do som pelos lábios, pode ser obtido a partir de valores adequados para excitação e parâmetros do trato vocal.

A teoria acústica oferece uma técnica simplificada para modelar os sinais de voz. Nessa técnica, a excitação é separada do trato vocal e da radiação, sendo os efeitos da

radiação e o trato vocal representados por um sistema linear variante no tempo. A excitação, por sua vez, é representada por um gerador de trem de pulsos glotais (pulsos quase-periódicos) ou de sinal aleatório (ruído). Os parâmetros do gerador e do sistema linear são escolhidos de forma a se obter na saída o sinal de voz esperado (RABINER e SCHAFER, 2010).

Na Figura 2.9, é apresentada uma representação do modelo de produção da voz descrito anteriormente, sendo, $u(n)$ o sinal de excitação, e $A_s(n)$ e $A_f(n)$ o controle da intensidade da excitação do sinal sonoro e do ruído, respectivamente.

Figura 2.9: Modelo discreto da produção da fala.



Adaptado de (RABINER e SCHAFER, 1978).

Comutando-se os geradores de excitação, altera-se o modo de excitação. Uma das formas de modelagem do trato vocal consiste em combinar o pulso glotal e os modelos de radiação em um sistema simples. No caso de análise por predição linear, é conveniente fazer essa combinação em conjunto com os componentes do trato vocal, representando-os por meio de uma função de transferência, Equação 2.2 (RABINER e SCHAFER, 1978).

$$H(z) = G(z)V(z)R(z). \tag{2.2}$$

Em que:

$G(z)$ – Transformada-z do modelo do pulso glotal;

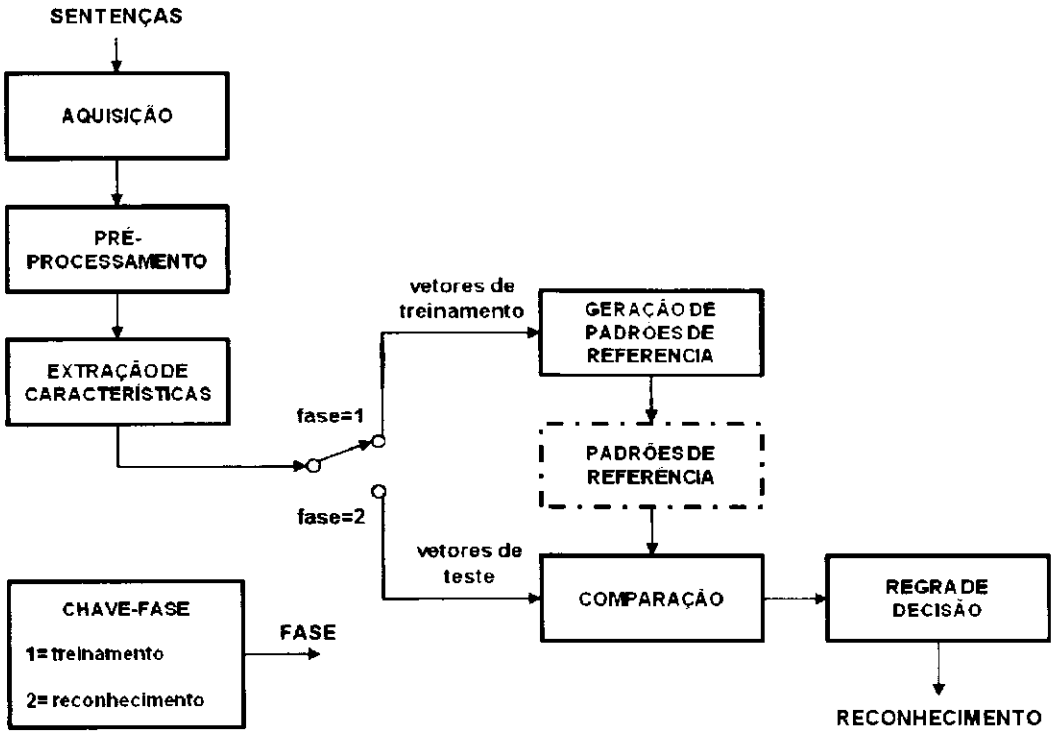
$V(z)$ – Transformada-z do modelo do trato vocal;

$R(z)$ – Transformada-z do modelo de radiação;

2.3 Reconhecimento de Padrões da Fala

O reconhecimento da fala é caracterizado como sendo uma tarefa de reconhecimento de padrões, neste caso padrões da fala, e possui duas fases: treinamento e reconhecimento (teste) (O'SHAUGHNESSY, 2000). Na Figura 2.10, é apresentada uma representação da tarefa de reconhecimento de padrões da fala.

Figura 2.10: Representação da tarefa de reconhecimento de padrões da fala.



A fase de treinamento produz padrões de referência, a partir de sentenças de voz (palavras ou frases) de treinamento. Esses padrões modelam as características comuns das sentenças de voz com o objetivo de identificá-las. Em SRF, independentes de locutor, o treinamento não ocorre, necessariamente, em tempo real (treinamento *offline*), uma vez que, esta fase geralmente demanda um esforço computacional significativo (O'SHAUGHNESSY,

2000; O'SHAUGHNESSY, 2003; HERBIG, GERL e MINKER, 2011). A maioria das estratégias de treinamento (LINDE, BUZO e GRAY, 1980; RABINER, 1989; BOURLARD e MORGAN, 1994) atribui valores iniciais aos padrões de referência e, a partir de processos exaustivos de re-estimação, determina os padrões ótimos. Além disso, uma vez estimados, os padrões de referência não necessitam ser atualizados, dado que o conjunto de sentenças de treinamento permaneça inalterado (CIPRIANO, 2001).

Na fase de reconhecimento (classificação), um conjunto de características de teste, obtido a partir de sentenças teste, é comparado com os padrões de referência armazenados. Essa comparação, e o uso de uma regra de decisão, identificam o padrão que mais se assemelha às características da sentença de voz de teste (voz desconhecida), proporcionando, dessa forma, o reconhecimento (RABINER e SCHAFER, 2010). A maioria dos SRF usa a métrica taxa de acertos (ou taxa de erros[3]) para aferir desempenho (O'SHAUGHNESSY, 2000; CIPRIANO, 2001; YUANYUAN, JIA e RUNSHENG, 2001; AMUDHA, VENKATARAMANI e MANIKANDAN, 2008; ZHOU e HAN, 2009; DIAS, 2011). A taxa de acertos corresponde ao percentual de sentenças de teste que foram associadas corretamente aos seus respectivos padrões de referência.

As etapas de aquisição, pré-processamento e extração de características são comuns às duas fases mencionadas.

2.3.1 Aquisição

A etapa de aquisição do sinal de voz compreende a captura e a conversão A/D (*Analógico/Digital*). A captura do sinal é realizada, usualmente, por um microfone, que converte a vibração sonora em sinal elétrico (sinal analógico contínuo no tempo e em amplitude).

A conversão A/D consiste na amostragem do sinal analógico, $s_a(t)$, a cada T segundos, e na quantização das amostras, para se obter o sinal digital $s(n)$ (CIPRIANO, 2001; RABINER e SCHAFER, 2010), dado por:

---

[3] A taxa de erros corresponde ao percentual de sentenças de teste que foram associadas a outros padrões de referência que não as representavam.

$$s(n) = s_a(n.T), n = 0, 1, 2, ... \quad (2.3)$$

Um sinal contínuo, de largura de banda finita, cuja frequência máxima é f_{max}, pode ser preservado e recuperado caso a sua frequência de amostragem f_s atenda ao critério de Nyquist Equação 2.4.

$$f_s \geq 2.f_{max} \quad (2.4)$$

Na telefonia, adota-se uma frequência de amostragem de 8 kHz para uma faixa de 300-3400 Hz. Com essa frequência de amostragem, garante-se uma inteligibilidade[4] de 85% do sinal de voz na rede telefônica, porém, uma perda de qualidade (naturalidade) do sinal é percebida (O'SHAUGHNESSY, 2000), uma vez que, a comunicação humana típica é limitada a uma frequência de 8 kHz (RABINER e SCHAFER, 1978). Uma frequência de amostragem intermediária de 11025 Hz garante a inteligibilidade com um ganho na energia do sinal (qualidade).

O ideal seria uma frequência de amostragem de 16 kHz. Porém, um dos objetivos deste trabalho é desenvolver um modelo de referência para SRF dependentes de bateria e um número maior de amostras de voz a serem processadas acarretaria impactos sobre o consumo de energia. Devido a essa razão, muitos SRF no contexto embarcado utilizam frequências de amostragem inferiores a 16 kHz (CIPRIANO, 2001; YUANYUAN, JIA e RUNSHENG, 2001; ZHOU e HAN, 2009; DIAS, 2011).

2.3.2 Pré-processamento

A etapa de pré-processamento é responsável pelo tratamento do sinal de voz com relação ao ambiente de gravação e ao canal de comunicação utilizado (DIAS, 2006). Esta etapa é dividida nas seguintes subetapas: normalização, detecção de voz, pré-ênfase, segmentação e janelamento (O'SHAUGHNESSY, 2000; SINGH e GARG, 2005; RABINER e SCHAFER, 2010).

---

[4] A inteligibilidade caracteriza-se como o percentual de palavras perfeitamente reconhecidas em uma conversação.

2.3.2.1 Normalização

A etapa de normalização atua na redução da variabilidade do sinal de voz em relação ao ambiente de gravação (ruído de fundo, canal de comunicação). A variabilidade causada pelas diferenças de intensidade de voz dos locutores também é restringida (O'SHAUGHNESSY, 2000), ou seja, a amplitude do sinal de voz dos locutores é limitada a uma dada faixa de valores.

2.3.2.2 Detecção de Voz

Esta etapa consiste na detecção do início e fim da fala, com o objetivo de eliminar as ocorrências de ruído de fundo e de silêncio do discurso. Esta eliminação promove impactos significativos na eficiência computacional dos SRF, uma vez que, os períodos de inatividade de voz deixam de ser processados, proporcionando uma redução no consumo de energia, como também, um aumento na eficiência do reconhecimento (O'SHAUGHNESSY, 2000; FECHINE et al., 2010).

O'Shaughnessy (2000, p. 388) relata os resultados de um experimento realizado em um sistema de reconhecimento automático de dígitos isolados. Os resultados obtidos neste experimento indicaram que a detecção correta do início e fim da voz proporcionou uma taxa de erro de 7%. Com um erro de detecção de, aproximadamente, 60 ms, a taxa de erro foi elevada para 10% e com uma falha de 130 ms na detecção, obteve-se um erro de 30%.

Muitos métodos de detecção de voz utilizam o parâmetro de energia por segmento do sinal (vide Equação 2.5), para identificar os trechos de voz ativa (YING, MITCHELL e JAMIESON, 1993; PETRY, ZANUZ e BARONE, 1999; FECHINE et al., 2010). Na Equação 2.5, E_{seg} é a energia do segmento, $s(n)$ é a n-ésima amostra do sinal de voz e N_A é o número de amostras por segmento do sinal.

$$E_{seg} = \sum_{n=0}^{N_A-1} [s(n)]^2 \cdot \qquad (2.5)$$

Outros métodos de detecção de início e fim utilizam, por exemplo, a taxa de cruzamento por zero, o período da frequência fundamental (*pitch*), a análise espectral e a análise cepstral, para determinar a atividade de voz (LI et al., 2002).

2.3.2.3 Pré-ênfase

A etapa de pré-ênfase realiza uma filtragem com filtro FIR (*Finite Impulse Response*) de primeira ordem no sinal de voz. O objetivo desta filtragem é atenuar as componentes de baixa frequência do sinal, minimizando, desta forma, os efeitos da variação da glote e da impedância de radiação, causada pelos lábios no processo de produção da voz.

A função de transferência da pré-ênfase é dada pela Equação 2.6, sendo α o fator de pré-ênfase, que usualmente assume valores entre 0,9 e 1,0 (RABINER e SCHAFER, 1978).

$$H(z) = 1 - \alpha.z^{-1}, 0 \leq \alpha \leq 1. \tag{2.6}$$

O sinal de saída da pré-ênfase $s_p(n)$ está relacionado ao sinal de entrada $s(n)$ pela Equação 2.7 (CIPRIANO, 2001; DIAS, 2006):

$$s_p(n) = s(n) - \alpha.s(n-1). \tag{2.7}$$

Em se tratando de uma implementação em *hardware* que utiliza notação em ponto fixo[5], é comum, o fator de pré-ênfase assumir o valor $15/16$, vide Equação 2.8 (CIPRIANO, NUNES e BARONE, 2003), o que significa a realização de uma divisão por uma potência de 2. Operações de divisão ou de multiplicação por potência de 2 podem ser realizadas a partir de simples operações de deslocamento.

$$s_p(n) = s(n) - \frac{15}{16}.s(n-1) = s(n) - s(n-1) + \frac{s(n-1)}{16}. \tag{2.8}$$

[5] Formato numérico em que o ponto decimal ocupa uma posição fixa.

2.3.2.4 Segmentação e Janelamento

O sinal de voz é fundamentalmente não estacionário e não periódico (LAMAS, 2005). Porém, em intervalos de tempo curtos (de 10 a 30 ms), os segmentos de voz possuem caráter estacionário[6] (RABINER e JUANG, 1993; CIPRIANO, 2001; LAMAS, 2005).

A divisão em segmentos é feita a partir da multiplicação do sinal de voz por uma função janela no domínio do tempo. Os tipos de janelas usualmente utilizados são: Janela Retangular, Janela de Hamming e Janela de Hanning.

O efeito da janela Retangular (vide Equação 2.9) consiste na divisão do sinal de voz em quadros consecutivos de mesmo tamanho N_A. Porém, essa divisão abrupta, causa, no domínio da frequência, fugas espectrais que alteram o espectro do sinal (HAYES, 1999; FECHINE, 2000).

$$J(n) = \begin{cases} 1, & 0 \leq n \leq N_A - 1 \\ 0, & caso\ contrário. \end{cases} \tag{2.9}$$

No janelamento de Hamming, Equação 2.10, as características espectrais do centro do quadro são mantidas e as transições abruptas das extremidades são eliminadas (HAYES, 1999; FECHINE, 2000).

$$J(n) = \begin{cases} 0{,}54 - 0{,}46\cos[2\pi n/(N_A - 1)], & 0 \leq n \leq N_A - 1 \\ 0, & caso\ contrário. \end{cases} \tag{2.10}$$

A janela de Hanning (Equação 2.11) comparada a de Hamming, produz um reforço menor nas amostras do centro do quadro e uma suavização maior nas amostras das extremidades (HAYES, 1999; FECHINE, 2000).

$$J(n) = \begin{cases} 0{,}5 - 0{,}5\cos[2\,\pi n/(N_A - 1)], & 0 \leq n \leq N_A - 1 \\ 0, & caso\ contrário. \end{cases} \tag{2.11}$$

Como nas janelas de Hamming e Hanning atribuem-se pesos baixos às amostras da extremidade do quadro, é comum o uso de superposição de quadros adjacentes, a fim de garantir que a variação dos parâmetros entre as janelas adjacentes seja mais gradual e que a

---

[6] Para curtos intervalos de tempo, os segmentos de voz apresentam propriedades estatísticas invariantes.

análise das amostras localizadas nos extremos das janelas não seja prejudicada (SILVA, 2009).

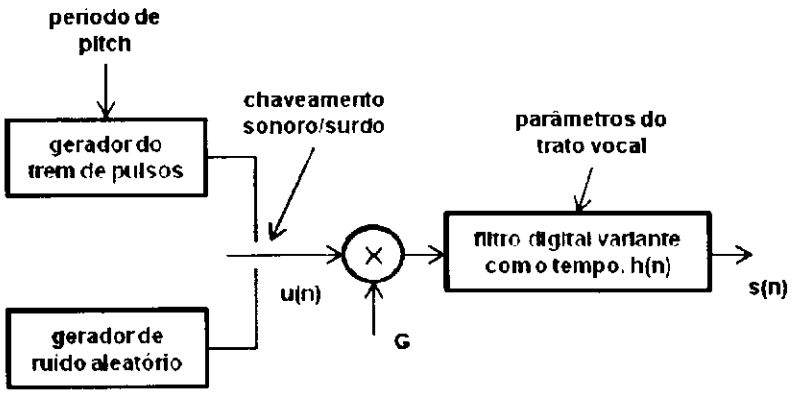
2.3.3 Extração de Características

Na etapa de extração de características, são obtidos os parâmetros que possibilitarão a geração de um padrão da fala. A análise por predição linear ou análise LPC (*Linear Prediction Coding*) é uma das técnicas mais usadas para estimação desses parâmetros (KIM et al., 1996; NAKAMURA et al., 2001; YUANYUAN, JIA e RUNSHENG, 2001; RABINER e SCHAFER, 2010). A importância desta técnica reside na sua boa representação dos parâmetros da fala, como também, na sua facilidade de implementação em *software* (RABINER e SCHAFER, 2010).

A ideia básica dessa análise consiste em se obter uma estimativa da voz amostrada por meio de uma combinação linear de amostras de voz passadas e de valores presentes e passados de uma entrada hipotética de um sistema, sendo, a saída deste sistema o sinal de voz (RABINER e SCHAFER, 2010).

A predição linear está relacionada ao modelo de produção da voz (descrito na Seção 2.2). Na Figura 2.11, está representada uma simplificação do modelo de produção da voz.

Figura 2.11: Simplificação do modelo de produção da voz.



Adaptado de (RABINER e SCHAFER, 2010).

Nesse modelo, o sinal de voz é modelado como a saída de um sistema linear variante no tempo excitado por pulsos quase periódicos (excitação sonora), ou ruído aleatório (excitação surda). As técnicas de predição caracterizam-se como um método preciso para estimação dos parâmetros do sistema linear variante com o tempo (RABINER e SCHAFER, 2010).

O principal problema da predição linear é determinar um conjunto de coeficientes do preditor a partir do sinal de voz, a fim de se obter uma estimativa precisa das propriedades espectrais do sinal de voz. Como o sinal de voz é variante no tempo, os coeficientes do preditor devem ser estimados em curtos intervalos de tempo (de 10 até 30 ms) (RABINER e SCHAFER, 1978; O'SHAUGHNESSY, 2000; CIPRIANO, 2001; LAMAS, 2005).

Esses coeficientes podem ser obtidos diretamente a partir da análise LPC, chamados coeficientes LPC, ou por meio de outras técnicas derivadas dessa análise (DIAS, 2006).

Os coeficientes mais utilizados são: LPC, cepstrais, cepstrais ponderados, delta cepstrais, delta cepstrais ponderados, mel cepstrais, (FURUI, 1981; FECHINE, 2000; VAREJÃO, 2001, LEE et al., 2003; AMUDHA, VENKATARAMANI e MANIKANDAN, 2008; SHIRALI-SHAHREZA, SAMETI e SHIRALI-SHAHREZA, 2008).

Os coeficientes utilizados neste trabalho para determinar os coeficientes do preditor foram os coeficientes LPC e os cepstrais. Esses foram escolhidos por representarem com precisão os parâmetros da fala e apresentarem uma complexidade inferior, em termos de algoritmo, em relação a técnicas que incorporam propriedades de percepção auditiva humana, tal como, os coeficientes mel cepstrais (O'SHAUGHNESSY, 2000; RABINER e SCHAFER, 2010). A redução de complexidade também apresenta como consequência a redução no consumo energético, dado o objetivo de uma implementação futura do SRF em hardware.

2.3.3.1 Coeficientes LPC

No modelo simplificado de produção da voz (Figura 2.11), a função de transferência do trato vocal, $H(z)$, para curtos intervalos de tempo, é definida pela Equação 2.12 (RABINER e SCHAFER, 2010).

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}}, S(z) = U(z)H(z). \quad (2.12)$$

Em que:

$S(z)$ – Transformada-z da sequência de voz $s(n)$;

$U(z)$ – Transformada-z do sinal de excitação $u(n)$;

a_i – coeficientes LPC;

p – ordem de predição (número de coeficientes);

G – parâmetro de ganho.

No domínio do tempo, as amostras de voz, $s(n)$, são relacionadas com a excitação sonora ou surda, $u(n)$, e com o parâmetro de ganho G, pela Equação 2.13 (RABINER e SCHAFER, 2010).

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n). \quad (2.13)$$

Um preditor linear para $s(n)$, considerando-se as suas p amostras anteriores, com coeficientes de predição, a_i, é definido pela Equação 2.14.

$$\tilde{s}(n) = \sum_{i=1}^{p} a_i s(n-i). \quad (2.14)$$

Vários métodos podem ser utilizados na resolução da Equação 2.14. Dentre eles, podem-se citar: o método da autocorrelação, o método da covariância, a formulação do filtro inverso, a formulação da estimação espectral, a formulação da máxima verossimilhança e a formulação do produto interno (RABINER e SCHAFER, 2010).

O método da autocorrelação é baseado na minimização do erro de predição, $e(n)$, associado à predição de $s(n)$ pela estimativa $\tilde{s}(n)$. O erro de predição é dado pela Equação 2.15.

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i). \tag{2.15}$$

Para se obter uma solução para o problema de minimização do erro, deve-se selecionar um segmento do sinal de voz por meio de uma janela de comprimento finito e igual a N_A. O comprimento da janela deve estar compreendido em intervalos de curta duração (de 10 até 30 ms), de forma a garantir a estacionariedade do segmento de voz selecionado (RABINER e SCHAFER, 1978).

Nas próximas equações, $x(n)$, corresponderá ao segmento selecionado e ponderado pela janela.

A Equação 2.13, modificada pela janela, será descrita como:

$$x(n) = \sum_{i=1}^{p} a_i x(n-i) + Gu(n) \tag{2.16}$$

Pela Equação 2.14, tem-se que a predição do segmento $x(n)$ é definida pela Equação 2.17.

$$\tilde{x}(n) = \sum_{i=1}^{p} a_i x(n-i). \tag{2.17}$$

Dado que $\tilde{x}(n)$ é a aproximação de $x(n)$ e que a_i é o i-ésimo coeficiente da predição linear; $\tilde{x}(n)$ é normalmente denominada a estimativa ou a predição de ordem p da amostra $x(n)$.

O erro de predição de cada amostra, $e(n)$, é definido pela Equação 2.18.

$$e(n) = x(n) - \tilde{x}(n) = x(n) - \sum_{i=1}^{p} a_i x(n-i). \tag{2.18}$$

E o erro quadrático, ε, em todo o segmento é dado por:

$$\varepsilon = \sum_{n=-\infty}^{\infty} e(n)^2 = \sum_{n=-\infty}^{\infty} [x(n) - \sum_{i=1}^{p} a_i x(n-i)]^2. \tag{2.19}$$

Como o segmento de voz é nulo para $n < 0$ e para $n > N_A$, o erro de predição (Equação 2.19) é nulo para $n < 0$ e $n > N_A + p - 1$. Logo, obtém-se a Equação 2.20 a partir da Equação 2.19.

$$\varepsilon = \sum_{n=0}^{N_A+p-1} e(n)^2 = \sum_{n=0}^{N_A+p-1} [x(n) - \sum_{i=1}^{p} a_i x(n-i)]^2. \tag{2.20}$$

O erro mínimo para o conjunto de coeficientes a_i é obtido fazendo-se:

$$\frac{\partial(\varepsilon)}{\partial(a_i)} = 0 \quad , 1 \leq i \leq p \tag{2.21}$$

Com a substituição da Equação 2.20 em 2.21 e a realização de p derivadas parciais, obtém-se:

$$\sum_{i=1}^{p} a_i R_r(|j-i|) = R_r(j), \;\; 1 \leq j \leq p \tag{2.22}$$

com

$$R_r(i) = \sum_{n=0}^{N_A-p-1} x(n)x(n+i). \tag{2.23}$$

As Equações 2.22 e 2.23, equações de Wiener-Hopf, podem ser mais bem visualizadas na forma matricial indicada na Equação 2.24 (VIEIRA, 1989; FECHINE, 2000; RABINER e SCHAFER, 2010).

$$\begin{bmatrix} R_r(0) & R_r(1) & \cdots & R_r(p-1) \\ R_r(1) & R_r(0) & \cdots & R_r(p-2) \\ R_r(2) & R_r(1) & \cdots & R_r(p-3) \\ \cdots & \cdots & \cdots & \cdots \\ R_r(p-1) & R_r(p-2) & \cdots & R_r(0) \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_r(1) \\ R_r(2) \\ R_r(3) \\ \cdots \\ R_r(p) \end{bmatrix} \tag{2.24}$$

Os coeficientes a_i do preditor são determinados pela solução das Equações 2.22 e 2.23 ou pela Equação 2.24.

A matriz de autocorrelação (vide Equação 2.24) é uma matriz simétrica e, assim, pode ser resolvida de forma eficiente por algoritmos recursivos, como por exemplo, o algoritmo de

Levinson-Durbin (MARPLE, 1987; FECHINE, 2000; OLIVEIRA, 2001; CARVALHO, 2007).

2.3.3.2 Coeficientes Cepstrais

Conforme dito anteriormente, no modelo simplificado de produção da voz (Figura 2.11), o sinal de voz corresponde à saída de um sistema linear variante no tempo (trato vocal), excitado por pulsos quase periódicos ou ruído aleatório. Porém, para pequenos segmentos do sinal, o sistema linear é invariante no tempo (RABINER e SCHAFER, 2010).

Sistemas lineares e invariantes no tempo apresentam resposta à excitação correspondente à convolução do sinal de entrada a partir de sua resposta impulsiva. Logo, a convolução de um sinal de entrada, $u(n)$, pela resposta impulsiva do sistema, $h(n)$, resulta no sinal $s(n)$, conforme Equação 2.25 (PETRY, ZANUZ e BARONE, 2000; O'SHAUGHNESSY, 2000; CARDOSO, 2009) – A operação de convolução é realizada no domínio do tempo.

$$s(n) = u(n) \otimes h(n) = \sum_{k=-\infty}^{\infty} u(k)h(n-k). \tag{2.25}$$

Efetuando-se, uma mudança de domínio (temporal para espectral), por meio da aplicação da transformada de Fourier discreta na Equação 2.25, a operação de convolução é transformada em uma multiplicação (Equação 2.26), sendo $U(e^{jw})$ e $H(e^{jw})$, as transformadas de Fourier discretas dos sinais $u(n)$ e $h(n)$, respectivamente.

$$S(e^{jw}) = U(e^{jw})H(e^{jw}). \tag{2.26}$$

Aplicando-se o logaritmo na Equação 2.26, é possível separar os espectros de excitação e de resposta impulsiva do trato vocal em duas parcelas, conforme Equação 2.27.

$$\log[S(e^{jw})] = \log[U(e^{jw})] + \log[H(e^{jw})]. \tag{2.27}$$

Com a aplicação da transformada inversa na Equação 2.27, obtém-se o cepstro ou coeficientes cepstrais do sinal de voz (Equação 2.28) (FECHINE, 2000; PETRY, ZANUZ e BARONE, 2000; CARDOSO, 2009). Portanto, o cepstro é definido como sendo a transformada inversa de Fourier do logaritmo da magnitude espectral de um sinal a curtos intervalos de tempo (BOGERT, HEALY e TUKEY, 1963).

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|S(e^{jw})|^2 e^{jwn} dw, \ para - \infty \leq n \leq \infty \tag{2.28}$$

Sabe-se que a parcela do sinal correspondente à resposta ao impulso do trato vocal varia mais lentamente que a parcela de excitação. Sendo assim, os dois sinais podem ser linearmente separados depois da aplicação da transformada inversa de Fourier (O'SHAUGHNESSY, 2000).

Os coeficientes cepstrais também podem ser obtidos, recursivamente, a partir dos coeficientes LPC, $a(1), a(2), \ldots, a(p)$, conforme Equação 2.29 (PAPAMICHALIS, 1987; FECHINE, 2000).

$$\begin{aligned} c(1) &= a(1) \\ c(n) &= a(n) + \sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) a(j)c(n-j), \quad 1 < n \leq p \end{aligned} \tag{2.29}$$

Os coeficientes cepstrais são muito utilizados como método de extração de características nos sistemas de reconhecimento de fala, pelo fato desses coeficientes separarem bem a parcela do sinal correspondente ao trato vocal da parcela de excitação (RABINER e JUANG, 1993; O'SHAUGHNESSY, 2000; CAEIROS, MIYATAKE e MEANA, 2009), dado que a informação do trato vocal é considerada mais importante que a excitação para o reconhecimento de voz (MILNER, 2008).

Contudo, os métodos de extração de características que incorporam propriedades de percepção auditiva humana aos modelos de produção da voz apresentam melhores taxas de reconhecimento nos SRF (O'SHAUGHNESSY, 2000; MILNER, 2008). Como exemplo de métodos com características de percepção auditiva pode-se destacar os coeficientes mel

cepstrais. Estes coeficientes são obtidos a partir dos coeficientes cepstrais[7] com a aplicação de filtros digitais espaçados segundo uma escala acusticamente definida, escala mel (PETRY, ZANUZ e BARONE, 2000). Embora os coeficientes mel cepstrais sejam mais indicados para o reconhecimento de fala, eles acarretam num aumento de complexidade no desenvolvimento do modelo em *hardware*. Este aumento da complexidade deve-se ao cálculo da FFT, do logaritmo e da transformada inversa de Fourier, como também da aplicação do banco de filtros na escala mel. Para maiores detalhes de como são obtidos os coeficientes mel cepstrais vide Anexo A.

2.3.4 Geração de Padrões

Como mencionado anteriormente, o reconhecimento de fala é uma tarefa de reconhecimento de padrões e possui duas fases: treinamento e reconhecimento (teste). Na fase de treinamento, são gerados padrões de referência a partir das características extraídas das sentenças de treinamento. Os padrões gerados são armazenados para posterior comparação.

Os métodos frequentemente utilizados para a construção dos padrões são: Quantização Vetorial (O'SHAUGHNESSY, 2000), Modelos de Markov Escondidos (HMM) (VAREJÃO, 2001; AMUDHA et al., 2007; DIAS, 2011), Alinhamento Dinâmico no Tempo (KIM et al., 1996; YUANYUAN, JIA e RUNSHENG, 2001; PHADKE et al., 2004) e Redes Neurais (BENZEGHIBA e BOULARD, 2002; AMUDHA, VENKATARAMANI e MANIKANDAN, 2008). É possível ainda construir padrões fazendo a combinação de dois ou mais métodos (NAKAMURA et al., 2001; OLIVEIRA, 2001; REZENDE, 2005).

A técnica utilizada neste trabalho para a geração dos padrões de referência foi a Quantização Vetorial. Esta técnica é indicada por remover a redundância, que geralmente existe, entre os sucessivos vetores de observação (O'SHAUGHNESSY, 2000). Além disso, trata-se de uma técnica bastante eficiente quando o vocabulário em questão é limitado e formado por palavras isoladas (CIPRIANO, 2001). O estado da arte da área de reconhecimento de fala recomenda a utilização de HMM de densidades contínuas para a geração dos padrões (WALKER et al., 2004). Contudo, a complexidade dos seus algoritmos

[7] Neste caso, os coeficientes cepstrais são obtidos a partir do método da FFT.

dificulta uma implementação em *hardware* (JIANG, MA e CHEN, 2010). Uma solução com HMM discreto é mais comum para as implementações em *hardware*, proporcionando um melhor desempenho do que a QV. Porém, a quantização vetorial é usualmente utilizada na obtenção do conjunto discreto de vetores de observação usado na geração de padrões do HMM, o que acarreta em um aumento de complexidade em relação ao uso apenas da QV e, consequentemente, de consumo em uma implementação em *hardware*.

2.3.4.1 Quantização Vetorial

A quantização vetorial (QV) é um método de compressão de dados que mapeia um conjunto de vetores de observação, X, com o objetivo de gerar um conjunto de M vetores-código, C. Os conjuntos X e C são definidos formalmente por (SOONG e JUANG, 1987; CIPRIANO, 2001):

$$X = \{x_t \in \mathcal{R}^p | t = 1 \dots T\}. \tag{2.30}$$

$$C = \{y_i \in \mathcal{R}^p | i = 1 \dots M\}. \tag{2.31}$$

O conjunto C corresponde ao dicionário (*codebook*) do quantizador que possui M níveis e os vetores, x_t e y_i, são p-dimensionais. Portanto, tem-se um quantizador vetorial p-dimensional de M-níveis (ou M-partições).

Para realizar o mapeamento do conjunto de observações em um dicionário, o algoritmo LBG (*Linde-Buzo-Gray*) é comumente utilizado. Os passos para geração do dicionário pelo algoritmo LBG são os seguintes (FECHINE, 2000; CIPRIANO, 2001):

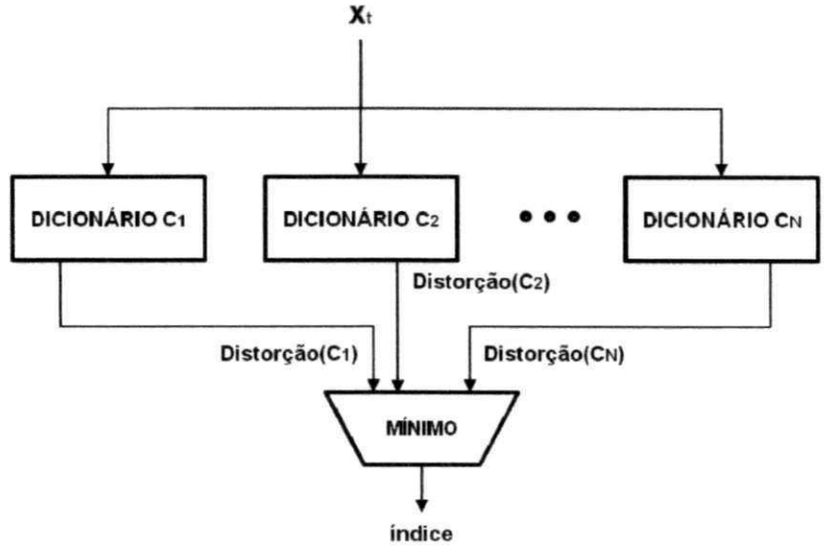
1. Definir o número de níveis do quantizador, M;
2. Escolher de forma aleatória M vetores de observação, $x_t \in X$, para compor um dicionário inicial, C;
3. Calcular a distorção entre cada vetor de observação $x_t \in X$ e os vetores-código $y_i \in C$, codificando $x_t \in X$ pelo índice i do vetor-código $y_i \in C$ que apresenta a menor distorção, ou seja, se classifica cada vetor de observação $x_t \in X$ em uma das M partições do dicionário;

4. Definir novos vetores-código para as M partições de C, tal que, o novo valor do vetor-código y_i corresponda à média dos vetores de observação, $x_t \in X$, que foram classificados na partição i;
5. Calcular o valor da distorção percentual e comparar com o limiar. Se este valor for maior que o limiar, retorna-se ao passo 3, caso contrário finaliza-se o processo com o dicionário, C, definido.

No reconhecimento de fala, a QV é usada nas duas fases: treinamento e reconhecimento. Na primeira, atua na geração de dicionários-modelo (padrões de referência). Na segunda, o vetor de teste é utilizado para o cálculo da distorção em relação aos dicionários-modelo (Figura 2.12). O padrão "vencedor", portanto, será o que apresentar a menor distorção. Uma das técnicas mais utilizadas para a construção do dicionário do quantizador e para a obtenção da distorção do vetor teste em relação ao dicionário é baseada na distância euclidiana (Equação 2.32) (O'SHAUGHNESSY, 2000).

$$d(x_t, y_i)^2 = \sum_{j=1}^{p} (x_{t,j} - y_{i,j})^2, 1 \leq t \leq T, 1 \leq i \leq M. \quad (2.32)$$

Figura 2.12: Reconhecimento de fala baseado em QV.



Adaptado de (CIPRIANO, 2001).

2.4 Discussão

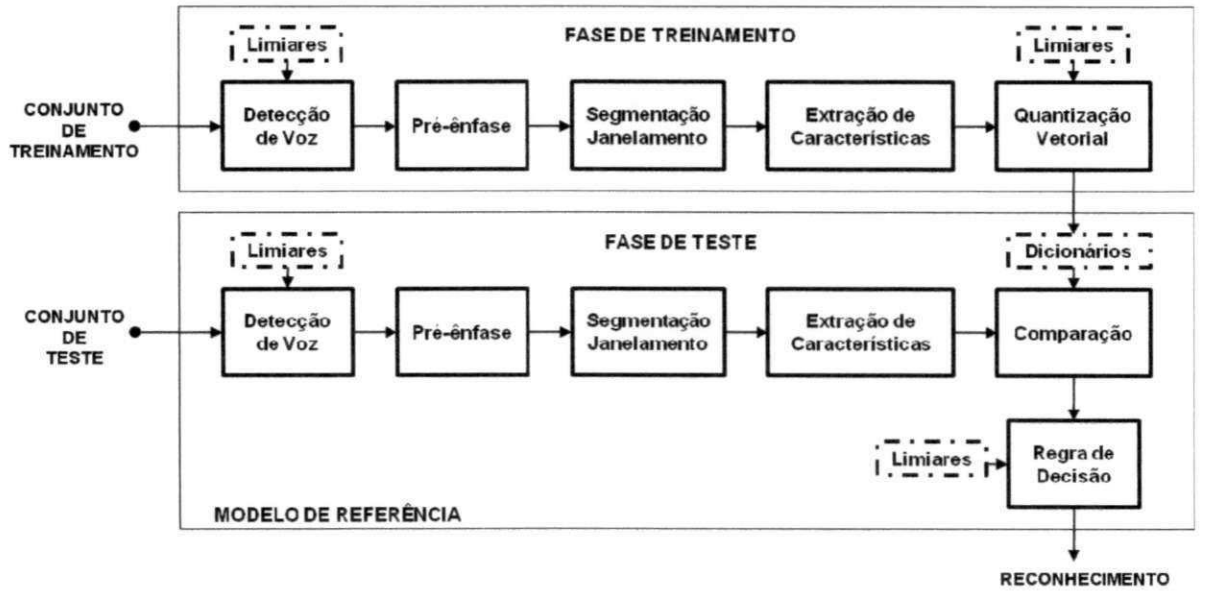
Neste capítulo, foi descrito o mecanismo de produção da voz humana, destacando o funcionamento do aparelho fonador humano com suas formas de excitação. O aparelho fonador também foi apresentado como um sistema mecânico de filtragem acústica. A representação matemática desse sistema acústico também foi descrita. Além disso, foi apresentada a tarefa de reconhecimento de fala com a descrição das fases de treinamento e de reconhecimento (teste) e das etapas envolvidas em cada uma dessas fases.

3 Descrição do Sistema

No Capítulo 2, Seção 2.3, foram descritas as principais características de um sistema de reconhecimento de fala. Neste capítulo, será feita uma descrição do sistema de reconhecimento de palavras isoladas e independente de locutor desenvolvido. Na Figura 3.1, está representado o diagrama em blocos do sistema, que é constituído das seguintes etapas:

- Pré-processamento: etapa composta por normalização, detecção de voz, pré-ênfase, segmentação e janelamento;
- Extração de características: coeficientes LPC e cepstrais;
- Quantização vetorial;
- Comparação: distorção obtida a partir da distância euclidiana;
- Regra de Decisão.

Figura 3.1: Diagrama em blocos do sistema de reconhecimento de palavras isoladas.



Como mencionado anteriormente, a tarefa de reconhecimento de fala é dividida em duas fases: treinamento e teste. Neste trabalho, optou-se por realizar o treinamento *offline*, visto que, esta fase demanda esforços computacionais significativos para gerar os padrões de referência, que não necessitam ser atualizados com frequência, dado que a base de dados de treinamento permanecerá constante (CIPRIANO, 2001; O'SHAUGHNESSY, 2003). Os

softwares de treinamento e reconhecimento (modelo de referência) foram implementados na linguagem de programação C++. Sendo que, os módulos de pré-processamento, extração de características (coeficientes LPC) e quantização vetorial, utilizados neste trabalho, foram codificados pela equipe Brazil-IP CG (FECHINE et al., 2010).

Neste trabalho, é feita uma normalização da amplitude[8] dos sinais de entrada para que todos os valores de amplitude estejam compreendidos na faixa [-1:1]. Na Figura 3.1, a subetapa de normalização foi omitida, uma vez que, em uma implementação em *hardware* a normalização pode ser realizada no módulo de captura e conversão A/D do sinal de voz. Na implementação em *software* (modelo de referência), a normalização antecede a subetapa de detecção de voz.

3.1 Base de Dados

A base de dados é constituída por dois conjuntos distintos, denominados conjunto de treinamento e conjunto de teste. Esses conjuntos são utilizados nas fases de treinamento e de teste, respectivamente, dos SRF.

Na fase de treinamento, o conjunto de treinamento tem por finalidade fornecer insumos para a geração dos padrões de referência. Na fase de teste, o conjunto de vozes de teste possibilita a avaliação do desempenho do sistema. Essa avaliação dá-se por meio de métricas, tais como, taxa de acertos e taxa de erros.

A base de dados utilizada para a validação deste trabalho foi gravada com uma frequência de amostragem de 11.025 Hz utilizando-se o *software* livre Audacity[9]. O ambiente de gravação não foi controlado. Desta forma, algumas amostras podem conter ruído. Outras características da gravação estão apresentadas na Tabela 3.1.

O vocabulário da base de dados compreende as palavras da língua inglesa *go*, *help*, *no*, *repeat*, *start*, *stop* e *yes*. A escolha do idioma inglês foi motivada visando à compatibilidade

---

[8] A normalização na amplitude dos sinais de entrada restringe a variabilidade da intensidade da voz dos locutores.

[9] http://audacity.sourceforge.net/

com o software Sphinx. A documentação do Sphinx não descreve como realizar o mapeamento dos fonemas da língua inglesa para os fonemas de outros idiomas. Além disso, a escolha por essas palavras foi motivada pelo uso em aplicações de comando e controle, visto que elas podem ser utilizadas como comandos para computadores, celulares, GPS, dentre outros. O vocabulário da base de dados comercial TI46[10] também emprega esses comandos.

Tabela 3.1: Características da gravação

Microfone	Leson MC-200; Resposta em frequência: 70 Hz a 12 kHz; Sensibilidade a 1 kHz: -49 dB; Impedância: 600 Ω
Placa de Som	*Realtek High Definition Audio*
Formato	wave
Número de Canais	1 canal (mono)
Resolução	16 bits por amostra

A base de dados foi gerada a partir de 11 locutores, sendo 4 mulheres e 7 homens. Cada locutor pronunciou cada palavra do vocabulário 16 vezes, para o conjunto de treinamento, perfazendo um total de 1.232 elocuções. Para o conjunto de teste, os locutores pronunciaram 10 vezes cada palavra, obtendo-se um total de 770 elocuções.

Devido às características da implementação do sistema proposto e de comparações com outros softwares de reconhecimento, o conjunto de treinamento possui duas variantes. Na primeira, as elocuções de treinamento de uma mesma palavra foram concatenadas em um único arquivo, gerando, portanto, 7 arquivos *wave*, cada um contendo 176 elocuções. Mais detalhes desses arquivos são apresentados na Tabela 3.2.

Tabela 3.2: Características do conjunto de treinamento.

Nome (wav)	Tamanho
go_tr	6,3MB
hp_tr	6,2MB
no_tr	7,1MB
rp_tr	6,6MB
sp_tr	6,7MB
st_tr	6,9MB
ys_tr	7,4MB

[10] http://www.ldc.upenn.edu/

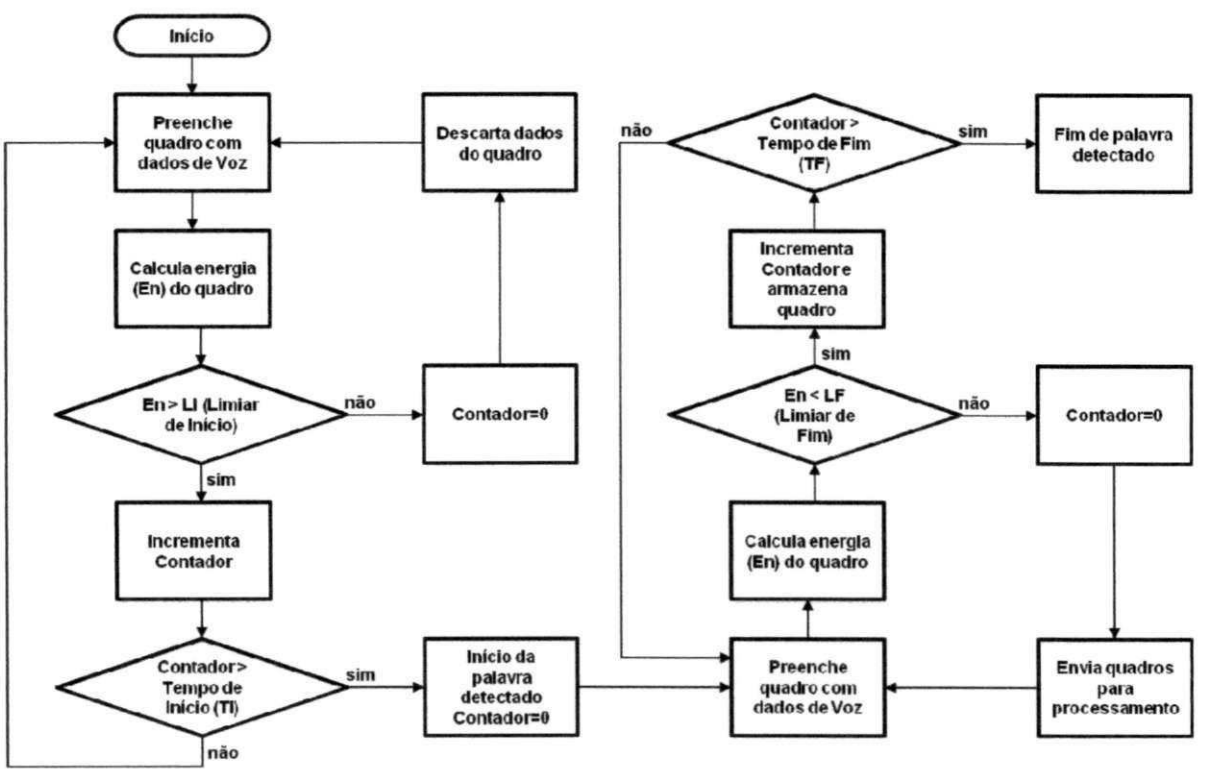
Na segunda versão do conjunto de treinamento, cada elocução encontra-se em um arquivo separado para atender o formato de entrada do Sphinx3, software de referência na área de reconhecimento de voz (WALKER et al., 2004; VOJTKO, KOROSI e ROZINAJ, 2008).

3.2 Detecção de Voz

O método de detecção de voz adotado neste trabalho utiliza o parâmetro de energia por segmento (Equação 2.5), com N_A igual a 110 amostras, o que equivale a um trecho de aproximadamente 10 ms do sinal de voz para uma frequência de amostragem de 11.025 Hz.

Para a detecção de início e fim das palavras, foi utilizado o algoritmo de detecção de voz apresentado por (PETRY, ZANUZ e BARONE, 1999), ilustrado na Figura 3.2.

Figura 3.2: Algoritmo para detecção de limites de palavra.



Adaptado de (PETRY, ZANUZ e BARONE, 1999).

Nesse algoritmo, a energia do sinal é calculada para cada quadro de tamanho fixo (segmento de voz de 110 amostras). O início de uma possível palavra é considerado a partir do primeiro quadro que apresenta uma energia maior que um limiar de início (LI) de voz. Para que os quadros seguintes constituam um início de palavra, a energia desses deve permanecer acima do limiar de início durante um período de tempo, chamado de tempo de início (TI)[11]. Durante o período de início, caso algum quadro apresente uma energia menor que o limiar de início (LI), as amostras de voz armazenadas são descartadas e a procura pelo início da palavra recomeça (PETRY, ZANUZ e BARONE, 1999).

Com o início da palavra detectado, inicia-se a procura pelo fim da palavra. O método de detecção do final da palavra é semelhante ao método de detecção de início. A diferença reside na procura por quadros com energia inferior a um limiar de fim (LF). Dessa forma, os quadros de voz devem permanecer com energia inferior ao limiar de fim por um dado período de tempo, chamado de tempo de fim (TF)[12], para que se considere atingido o fim da palavra. Os quadros inseridos no tempo de fim (TF) não são considerados componentes da palavra (PETRY, ZANUZ e BARONE, 1999).

Para a determinação dos valores dos limiares de energia (LI e LF) e de tempo (TI e TF), os seguintes passos foram seguidos:

1. Determinação de medidas estatísticas da energia do conjunto de treinamento, descrito na Seção 3.1. Esses dados estão apresentados na Tabela 3.3;
2. Escolha de medidas estatísticas da energia do conjunto de treinamento para avaliação dos limiares de energia. As medidas escolhidas foram a mediana (Q2) e o terceiro quartil (Q3);
3. Escolha de limiares de tempo. A faixa escolhida para avaliação está compreendida em um período de 1 a 30 quadros, cada quadro representando 10 ms do sinal de voz;
4. Simulação dos limiares definidos. Na Tabela 3.4, estão representados os limiares avaliados.
5. Análise das métricas de desempenho do sistema para os limiares definidos. A descrição da análise encontra-se no Capítulo 4.

---

[11] Corresponde ao número de quadros sucessivos para a detecção do início de uma palavra. Cada um desses quadros deve apresentar uma energia maior que a energia do limiar de início (LI).

[12] Corresponde ao número de quadros sucessivos para a detecção do final da palavra. Cada um desses quadros deve apresentar uma energia menor que a energia do limiar de fim (LF).

Tabela 3.3: Dados estatísticos da energia do conjunto de treinamento.

Medida Estatística	Energia por Segmento
Mínimo	0,000000
Primeiro Quartil (Q1)	0,000623
Mediana (Q2)	0,001002
Terceiro Quartil (Q3)	0,007760
Máximo	30,848365
Média	0,353537
Variância	1,945835

Tabela 3.4: Limiares de energia e de tempo.

Arranjo	LI	LF	TI	TF
A-I	Q2	Q2	[1:30]	[1:30]
A-II	Q2	Q3	[1:30]	[1:30]
A-III	Q3	Q2	[1:30]	[1:30]
A-IV	Q3	Q3	[1:30]	[1:30]

3.3 Pré-ênfase

Neste trabalho, foi utilizado um fator de pré-ênfase de 15/16 (que corresponde a 0,9375), uma vez que, esse fator facilita a implementação em ponto fixo e como consequência simplifica a implementação em *hardware*, já que, uma operação de divisão ou multiplicação por potência de 2 pode ser realizada por deslocamento de *bits* (CIPRIANO, 2001; DIAS, 2006). Detalhes do processo de pré-ênfase foram descritos na Seção 2.3.2.3.

3.4 Segmentação e Janelamento

Para garantir as condições de estacionariedade do sinal, foram utilizados segmentos de aproximadamente 20 ms, o que equivale a 220 amostras por quadro a uma frequência de 11.025 Hz. A função janela escolhida foi a de Hamming com superposição de 50%. Esta janela mantém as características espectrais do centro do quadro e minimiza as descontinuidades nas extremidades. A superposição de 50%, por sua vez, garante que a

variação dos parâmetros de janelas adjacentes seja mais gradual, uma vez que, a função de Hamming atribui pesos baixos às amostras da extremidade do quadro (HAYES, 1999).

3.5 Extração de Características

Para cada segmento de voz janelado é extraído um conjunto de p coeficientes, a partir da análise por predição Linear, obtendo-se um conjunto de coeficientes LPC ou cepstrais. Os coeficientes cepstrais formam vetores de características mais adequados ao desenvolvimento de aplicações de reconhecimento de fala (RABINER e JUANG, 1993) do que os coeficientes LPC (CAEIROS, MIYATAKE e MEANA, 2009), conforme descrito na Seção 2.3.3. Na Seção 4.1, é apresentada uma análise comparativa de desempenho destes coeficientes com o objetivo de evidenciar este fato, como também, para que se possa mensurar quão bons são os coeficientes cepstrais.

A escolha da ordem de predição p (número de coeficientes) depende da frequência de amostragem f_s do sinal de voz conforme Equação 3.1 (RABINER e SCHAFER, 1978). Assim, sendo p uma aproximação, foi escolhido p=12 dado que f_s=11.025 Hz.

$$p \cong \frac{f_s}{1000}. \tag{3.1}$$

O algoritmo utilizado para o cálculo dos coeficientes LPC foi o de Levison-Durbin. O cálculo dos coeficientes cepstrais foi realizado pelo algoritmo de recursão (Equação 2.29) apresentado por (PAPAMICHALIS, 1987), uma vez que, esta forma de obtenção dos coeficientes cepstrais é mais simples do que a obtida pelo método da FFT.

3.6 Quantização Vetorial

Foi utilizado um quantizador vetorial p-dimensional de 64 níveis (ou 64 partições). Esta quantidade de níveis é suficiente para uma boa representação do sinal de voz sem geração de um grande volume de dados (FECHINE e AGUIAR-NETO, 1993). O algoritmo utilizado

para a geração dos 7 dicionários (que representam as características das 7 palavras do vocabulário descrito na Seção 3.1) foi o LBG descrito na Seção 2.3.4.1.

3.7 Comparação

Na fase de reconhecimento, a comparação dos vetores de teste p-dimensionais com os 7 dicionários armazenados é feita por meio do cálculo da distância euclidiana (vide Equação 2.32). Desse modo, são enviadas 7 distâncias (ou distorções) para o módulo de decisão (vide Figura 3.1).

3.8 Regra de Decisão

Conforme Seção 3.7, o módulo de comparação envia ao módulo de decisão 7 distâncias (ou distorções) de um dado vetor de teste em relação aos dicionários armazenados. O objetivo do módulo de decisão é indicar qual dos dicionários mais se assemelha às características dos vetores de teste, que correspondem à palavra de entrada desconhecida.

Para alcançar este objetivo, se faz necessária à definição de regras de decisão. Neste trabalho, foram definidas 4 regras e o parâmetro de decisão utilizado foi a distância média, vide Equação 3.2.

$$m_i = \frac{1}{Q}\sum_{j=1}^{Q} distância(C_i), 1 \leq i \leq 7 \tag{3.2}$$

Em que:

m_i – Distância média dos vetores teste, que compõem a palavra desconhecida, em relação ao dicionário de índice i;

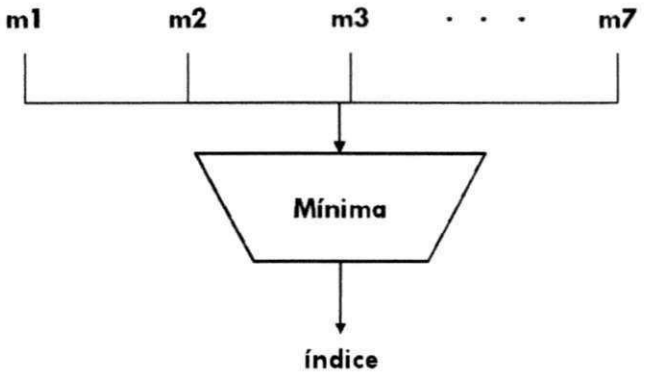
Q – Número de quadros da palavra desconhecida;

C_i – Dicionário (*codebook*) de índice i;

$distância(C_i)$ – Distância de um dado vetor de teste de índice j em relação ao dicionário C_i;

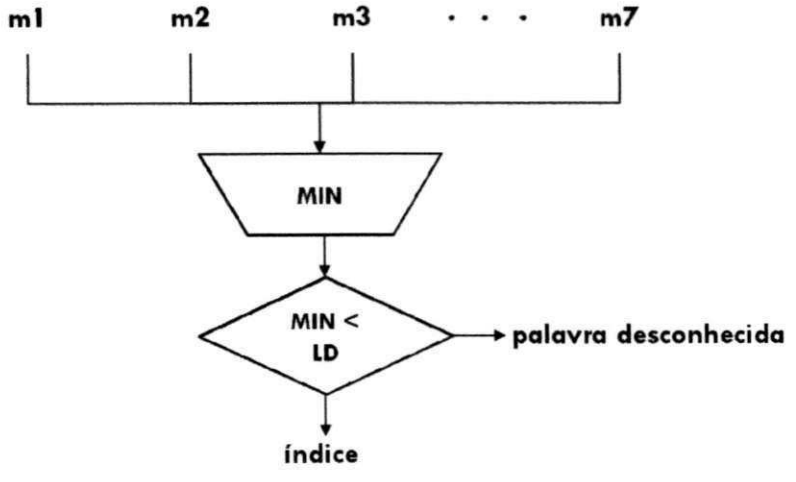
Na Figura 3.3, é apresentada uma representação da primeira regra de decisão (RD-I). Esta regra atribui a palavra desconhecida, voz de entrada, o índice *i* do dicionário que apresenta a menor distância dentre as 7 distâncias médias calculadas.

Figura 3.3: Regra de decisão I.



A segunda regra de decisão (RD-II) também utiliza a menor distância média, porém, para caracterizar o reconhecimento esta distância deve ser menor que um dado limiar. Caso esta distância seja maior ou igual ao limiar, a resposta do sistema será "palavra desconhecida" e será dada uma nova oportunidade ao usuário do sistema de repetir o comando de voz (ver Figura 3.4).

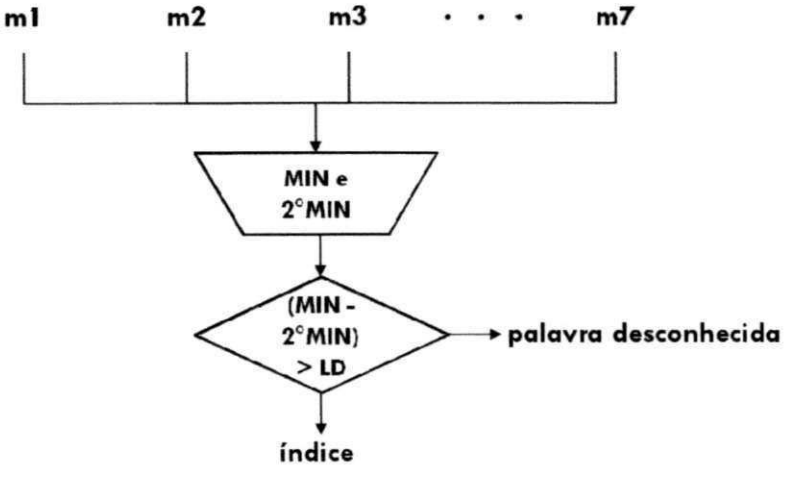
Figura 3.4: Regra de decisão II.



A adição de um limiar nas regras de decisão evita a ocorrência de falsos reconhecimentos (erros), uma vez que, sem o limiar de decisão, palavras não cadastradas no sistema são reconhecidas. O limiar de decisão também atua no universo de palavras cadastradas, quando há a ocorrência de um valor de distância média mínima acima dos limites de tolerância, que pode ser ocasionado por problemas na pronuncia do comando pelo o usuário, pela ocorrência de ruído de fundo no momento da elocução, dentre outros.

Na Figura 3.5, tem-se a representação da terceira regra de decisão (RD-III). Nesta regra, a decisão do reconhecimento é baseada na diferença entre a segunda menor distância média e a menor. Caso esta diferença seja maior que um dado limiar, a resposta do sistema será o índice da palavra associada à menor distância média. A resposta do sistema será desconhecida, para uma diferença menor ou igual ao limiar. O objetivo de RD-III consiste em estabelecer uma tolerância quando as palavras cadastradas no sistema possuem pronúncias similares.

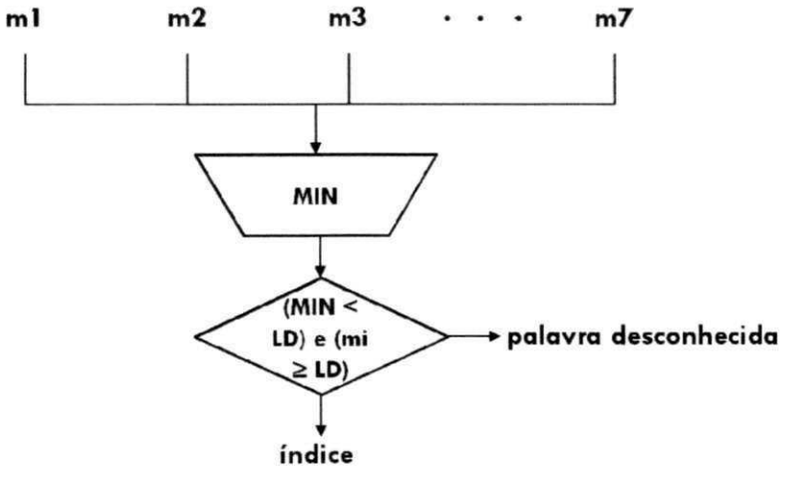
Figura 3.5: Regra de decisão III.



A quarta regra de decisão (RD-IV), tal como RD-I e RD-II, utiliza apenas o parâmetro da distância média mínima para tomada de decisão acerca do reconhecimento. Nesta regra, a menor distância média deve ser menor que um dado limiar. Porém, as demais distâncias devem ser maiores ou iguais ao limiar para a caracterização do reconhecimento. Caso esta condição não seja atendida, a resposta do sistema será de palavra desconhecida e o usuário do sistema poderá repetir o comando de voz (vide Figura 3.6). A regra RD-IV é indicada para

aplicações de reconhecimento de voz críticas em que a ocorrência de falsos reconhecimentos não é admitida.

Figura 3.6: Regra de decisão IV.



As quatro regras descritas nesta seção atendem boa parte das aplicações de reconhecimento de voz, contudo ainda é possível formular mais regras de decisão.

No Capítulo 4, será apresentada uma análise comparativa dessas 4 regras, com o objetivo de determinar qual regra é a mais indicada para o sistema de reconhecimento de palavras desenvolvido.

3.9 Discussão

Neste capítulo, foram apresentadas as principais características do sistema de reconhecimento de palavras isoladas, independente de locutor. Pode-se destacar a apresentação da base de dados com 1.232 elocuções de treinamento e 770 elocuções de teste. Com relação aos módulos do SRF proposto, foi utilizado o algoritmo de (PETRY, ZANUZ e BARONE, 1999) para a detecção de voz, um fator de pré-ênfase de 0,9375 (15/16), segmentos de aproximadamente 20 ms e uma janela de Hamming com superposição de 50%, o uso de 12 coeficientes para as características de cada segmento, um quantizador vetorial de 64 níveis para representar cada palavra, a utilização do algoritmo LBG para geração dos dicionários, a

comparação de vetores de teste feita a partir da distância euclidiana, como também a apresentação das regras para o módulo de decisão.

4 Apresentação e Análise dos Resultados

Neste capítulo, são apresentados e discutidos os resultados obtidos a partir das configurações do sistema simuladas neste trabalho. O estudo dessas configurações tem como objetivo propor uma configuração que maximize as respostas corretas do sistema de reconhecimento de palavras.

Ao todo, foram utilizados 3.600 arranjos de detecção de voz (Tabela 3.4), 2 tipos de características (coeficientes LPC e coeficientes cepstrais) e 4 regras de decisão (Seção 3.8). Os modelos dos módulos de pré-ênfase, segmentação, janelamento, quantização vetorial e comparação (distância euclidiana) se repetem para todos os experimentos, conforme exposto no Capítulo 3.

As configurações foram agrupadas em 8 processamentos[13], conforme descrito na Tabela 4.1. A métrica utilizada para aferir o desempenho de cada configuração foi a quantidade de erros[14] (QE). Os resultados de cada processamento estão apresentados no formato de uma matriz de ordem 30 e encontram-se no Apêndice A. As linhas e colunas dessas 8 matrizes representam os limiares de tempo TI e TF, respectivamente (para mais detalhes dos limiares do detector de voz vide Seção 3.2).

Tabela 4.1: Descrição dos processamentos.

Processamento	Limiares do DV	Extração de Características	Regra de Decisão
P1	A-I	LPC	RD-I
P2	A-II	LPC	RD-I
P3	A-III	LPC	RD-I
P4	A-IV	LPC	RD-I
P5	A-I	CEP	RD-I
P6	A-II	CEP	RD-I
P7	A-III	CEP	RD-I
P8	A-IV	CEP	RD-I

[13] Cada processamento consiste em um conjunto formado por 900 elementos, no qual cada elemento corresponde a uma configuração do sistema de reconhecimento automático de palavras isoladas.

[14] Número de sentenças de teste que foram associadas a outros padrões de referência que não as representavam.

Inicialmente, todos os processamentos foram submetidos à regra de decisão I. Duas razões são apontadas para tal escolha. A primeira e principal razão reside no fato de que as demais regras são derivadas de RD-I. O outro motivo está relacionado à redução de iterações computacionais, visto que os 8 processamentos correspondem a 7.200 configurações (cada processamento com 900 configurações[15]). Se fossem avaliadas as 4 regras, esse número seria elevado para 28.800 configurações.

As regras RD-II, RD-III e RD-IV foram avaliadas apenas para a configuração de extração de características e limiares do detector de voz que apresentou a menor quantidade de erros.

É importante destacar, que todas as simulações dos processamentos foram realizadas usando a base de dados descrita na Seção 3.1. Esta base também foi utilizada para comparação do modelo (sistema) desenvolvido com o software de reconhecimento de voz Sphinx3. A ferramenta Sphinx foi escolhida para a realização da comparação, por utilizar modelos que representam o estado da arte na área de Reconhecimento de Fala (WALKER et al., 2004; VOJTKO, KOROSI e ROZINAJ, 2008). O Sphinx na sua versão 3 apresenta modelagem linguística, extração de características por coeficientes mel cepstrais e classificação/geração de padrões por HMM de densidades contínuas.

4.1 Análise Estatística

Nesta seção, apresenta-se a análise estatística dos processamentos apresentados na Tabela 4.1. Um dos objetivos dessa análise é identificar os limiares de detecção de voz que minimizam a métrica QE.

Na Tabela 4.2, tem-se um resumo das medidas estatísticas dos processamentos. As medidas utilizadas foram: medidas de tendência central (média e mediana, Q2); valores extremos (mínimo e máximo), medidas de espalhamento (quartis: Q1, Q2, e Q3) e dispersão (desvio padrão, σ).

---

[15] Cada processamento levou em média 36 horas para ser computado em uma máquina com processador Intel Core 2 Duo (1,8GHz, 2M L2 Cache, 800MHz FSB) e memória RAM de 4GB.

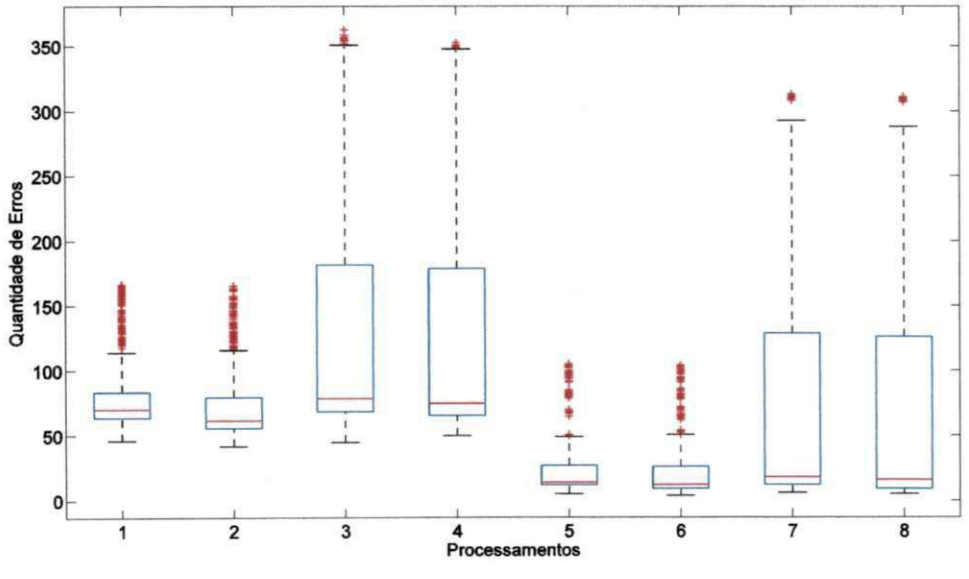
Tabela 4.2: Medidas resumo da métrica QE para os 8 processamentos.

Processamento	Mínimo	Q1	Mediana	Q3	Máximo	Média	σ
P1	46	64	70,5	84	167	81,26	28,16
P2	42	56	62	80	166	74,35	29,12
P3	45	69	79	182	363	133,03	91,30
P4	50	66	75	179	353	129,1	89,06
P5	5	12	14	27	105	26,34	26,25
P6	4	9	12	26	104	24,22	26,39
P7	6	12	18	128,5	313	78,60	94,29
P8	5	9	16	126	311	76.13	94,22

Na Tabela 4.2, as colunas "Mínimo" e "Máximo" representam a menor e a maior métrica QE, respectivamente, dentre as 900 configurações de um dado processamento; as colunas "Q1", "Q2" e "Q3" correspondem à realização QE que ocupa a posição um quarto, um meio e três quartos, respectivamente, das 900 configurações ordenadas de um dado processamento; as colunas "Média" e "σ" apresentam, respectivamente, a média aritmética e o desvio padrão das métricas QE das 900 configurações de um dado processamento.

A partir das informações da Tabela 4.2, que também podem ser visualizadas na forma de diagrama esquemático (Figura 4.1), percebe-se que o processamento com extração de características LPC que apresenta a menor quantidade de erros é P2.

Figura 4.1: Diagrama esquemático para os 8 processamentos.



O processamento P2 apresenta as menores medidas de tendência central, valores extremos e espalhamento, quando comparado aos demais processamentos com coeficientes LPC. Nesta comparação, o processamento P2 apresenta apenas uma medida de dispersão (σ=29,12) um pouco maior que o processamento P1 (σ=28,16).

Para os processamentos com extração de características cepstrais, o processamento P6 apresenta a menor quantidade de erros, quando comparado às medidas de mínimo (4), mediana (12), terceiro quartil (26), máximo (104) e média (24,22) (Tabela 4.2). Para a medida Q1, o processamento P6 é equivalente ao processamento P8 (ambos apresentam Q1=9). Contudo, o primeiro quartil de P6 ainda apresenta um valor menor que os demais processamentos. Em relação ao desvio padrão, o processamento P6, com σ=26,39, é significativamente menos disperso que P7 (σ=94,29) e P8 (σ=94,22), e ligeiramente mais disperso que P5 (σ=26,25).

Conforme descrito na Tabela 4.1, os processamentos P2 e P6 foram simulados com o mesmo arranjo de limiares de detecção de voz (A-II). Sendo, portanto, os limiares de energia desses processamentos iguais a 0,001002 (limiar de início) e 0,007760 (limiar de fim) (vide Tabela 3.3 e Tabela 3.4). Logo, tendo em vista os processamentos analisados, esses limiares de energia são os mais indicados para detectar os trechos de atividade de voz e com isso contribuir para a redução da métrica QE.

Após a definição dos limiares de energia, tem-se uma análise da influência dos limiares de tempo no reconhecimento. Conforme descrito na Seção 3.2, os limiares TI e TF foram avaliados para períodos de tempo de 10 ms a 0,3 s (ou 1 a 30 quadros). Nas Figuras 4.2 e 4.3, apresenta-se uma comparação dos processamentos com extração de características LPC, sob a influência da variação de TI e TF, respectivamente, na quantidade de erros.

Na comparação apresentada na Figura 4.2, constata-se que, os processamentos P1, P2, P3 e P4 apresentam um acréscimo significativo da quantidade de erros à medida que TI assume valores acima de 20 quadros. Em se tratando do limiar de tempo de fim (Figura 4.3), não se percebe variações expressivas de QE nos processamentos quando da variação de TF.

Essas observações do comportamento da métrica QE, nos processamentos com extração de características LPC, quando da variação dos limiares de tempo, também são observadas para os processamentos com extração de características cepstrais (vide Figura 4.4 e Figura 4.5).

Figura 4.2: Comparação dos processamentos com extração de características LPC sob a influência da variação de TI na métrica QE.

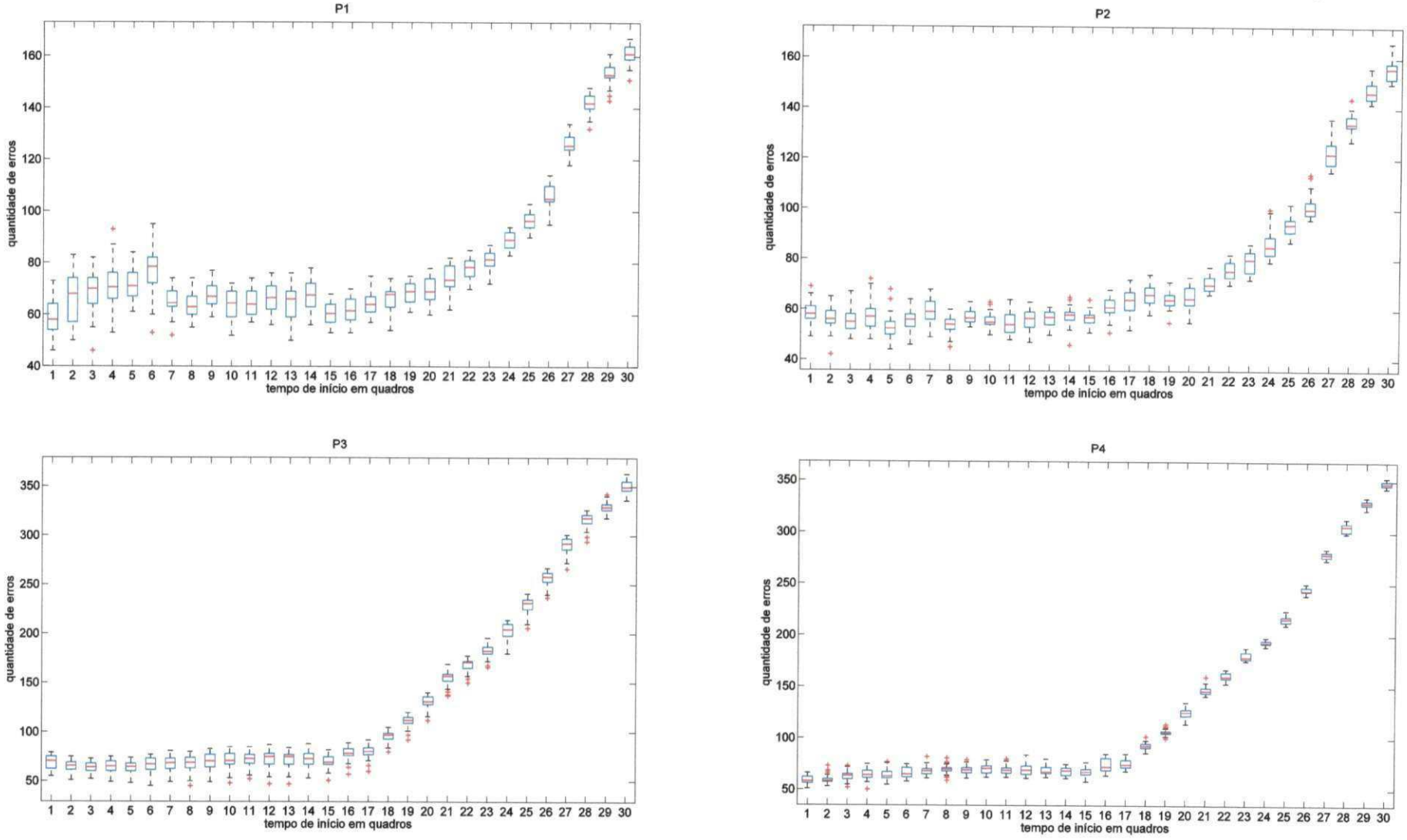


Figura 4.3: Comparação dos processamentos com extração de características LPC sob a influência da variação de TF na métrica QE.

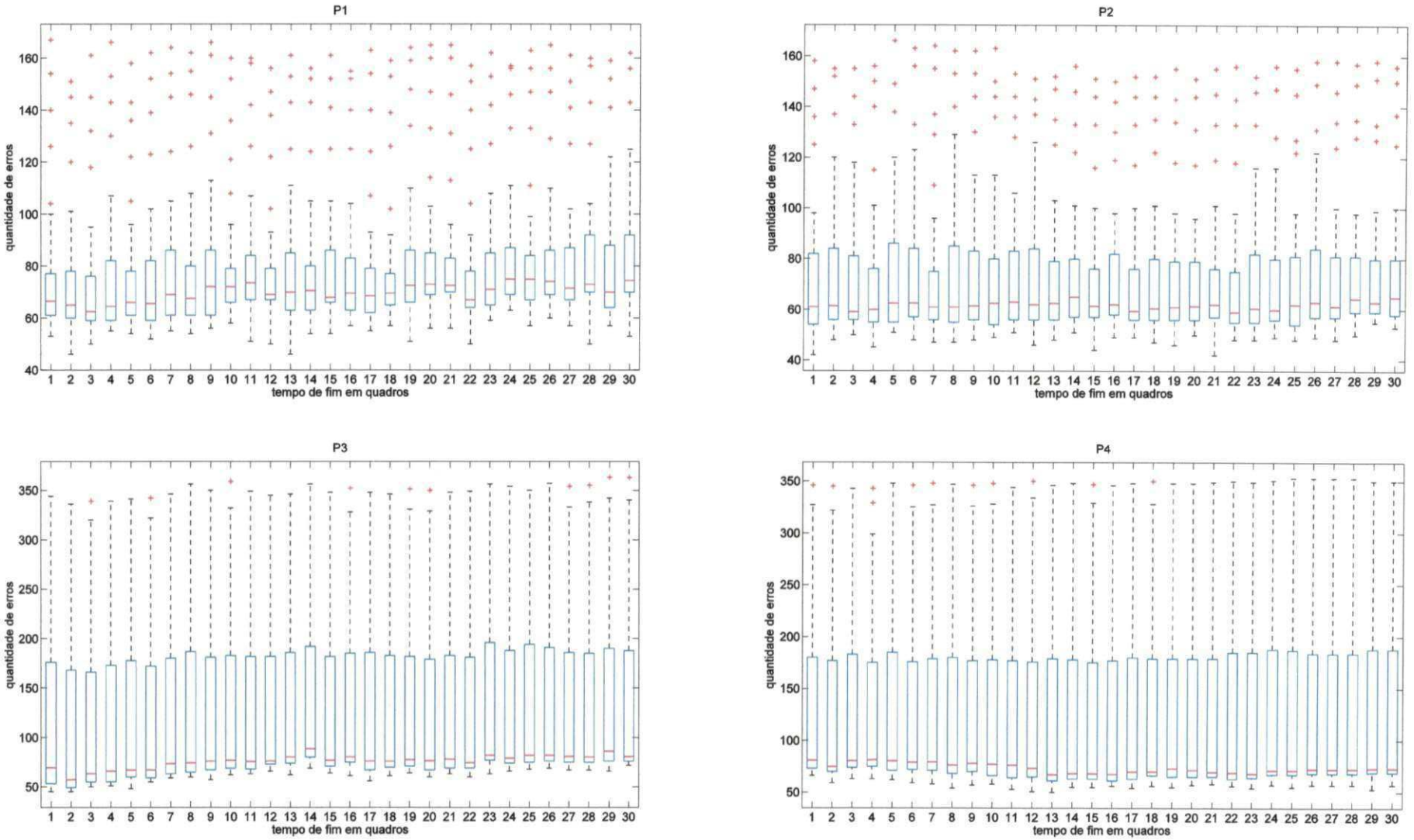


Figura 4.4: Comparação dos processamentos com extração de características cepstral sob a influência da variação de TI na métrica QE.

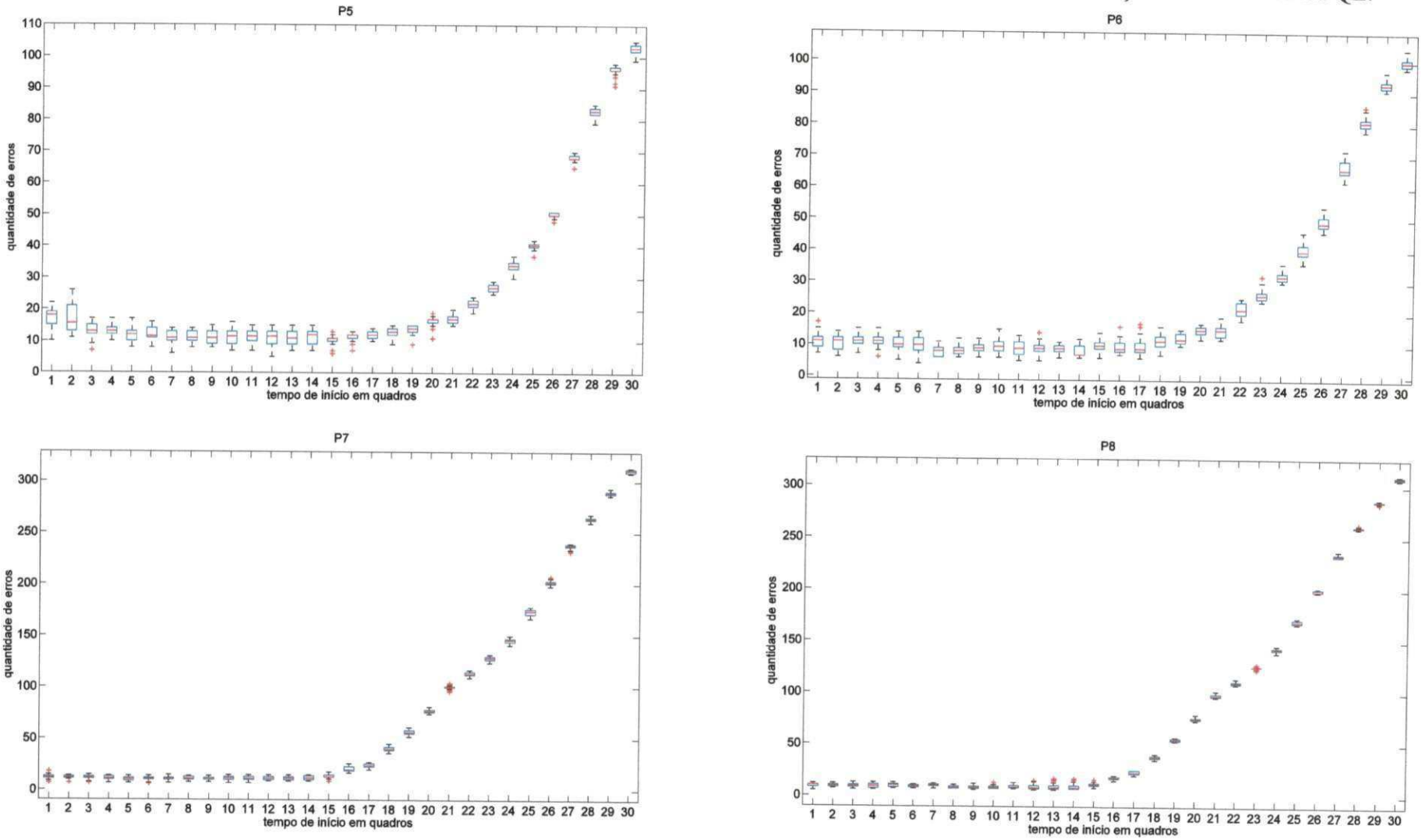
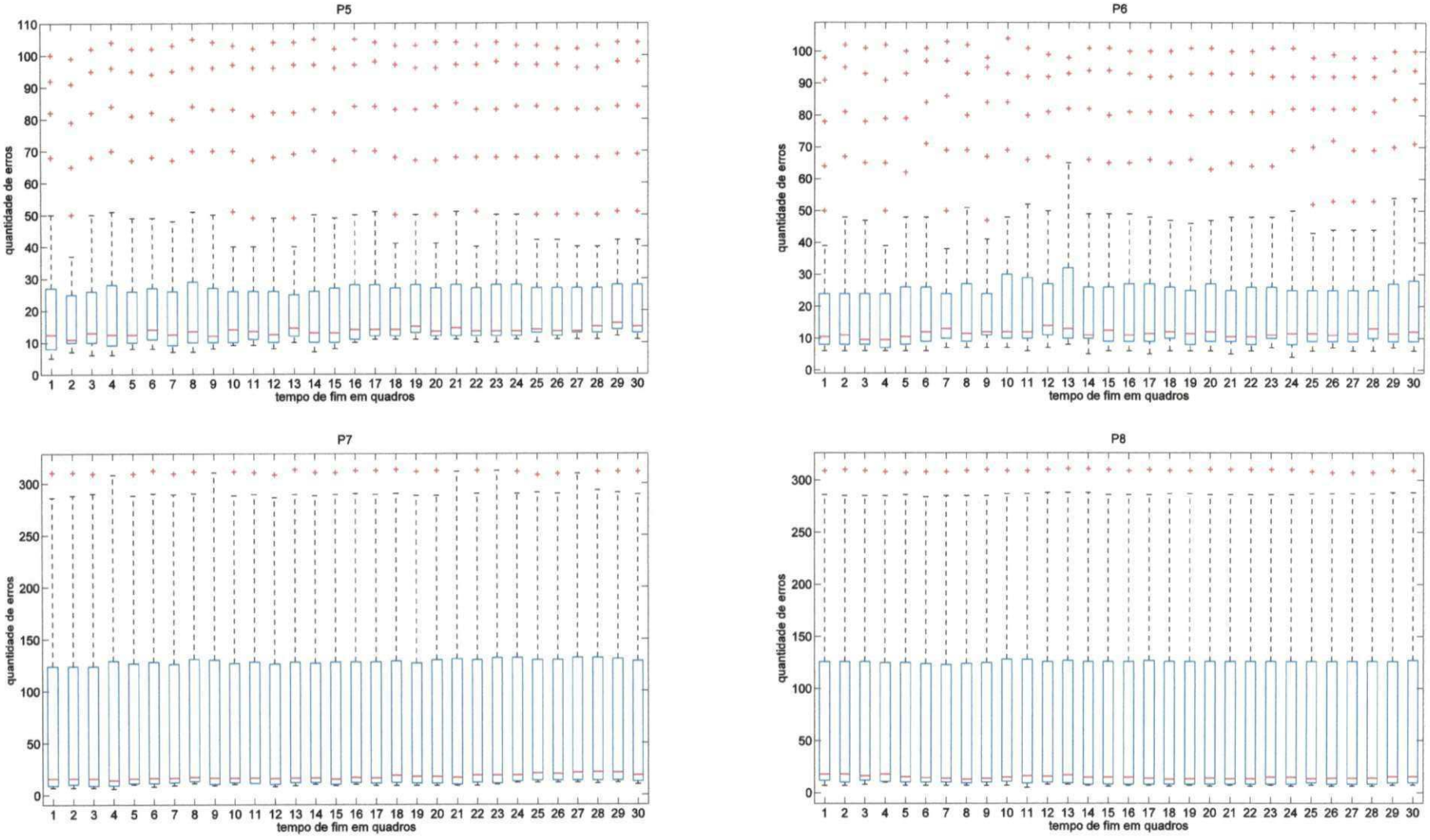


Figura 4.5: Comparação dos processamentos com extração de características cepstral sob a influência da variação de TF na métrica QE.



Então, pode-se concluir que o limiar de tempo de início possui mais influência no reconhecimento que o limiar de tempo de fim, ou seja, detectar precisamente o início de uma palavra é mais crítico para o reconhecimento do que detectar com exatidão o fim da palavra.

Além disso, o limiar TI não deve assumir valores maiores que 20 quadros. Em se tratando do limiar TF, não foram observadas restrições na faixa de 1 a 30 quadros. Este resultado é muito importante para uma futura implementação em *hardware*, uma vez que, é possível economizar memória para o armazenamento das amostras de voz sem que haja impactos significativos no reconhecimento das palavras.

Os melhores pares de limiares de tempo, que tornam mínima a quantidade de erros de uma dada configuração de sistema relacionada a um determinado processamento, são apresentados na Tabela 4.3.

Tabela 4.3: Limiares de tempo que minimizam QE.

Limiares de tempo (TI, TF)	QE Mínima	Processamento
(1,13)	46	P1
(3,2)	46	P1
(2,1)	42	P2
(2,21)	42	P2
(6,1)	45	P3
(8,2)	45	P3
(4,13)	50	P4
(12,1)	5	P5
(6,24)	4	P6
(6,4)	6	P7
(1,11)	5	P8

Conforme informações da Tabela 4.3, as configurações dos processamentos, que apresentam QE mínima, possuem limiares de tempo de início compreendidos na faixa de 1-12 quadros e limiares de tempo de fim na faixa de 1-24. Essas faixas de valores são compatíveis com o comportamento esperado (TF sem restrições de faixa e TI menor que 20 quadros).

Ainda de acordo com as informações Tabela 4.3, a configuração que apresenta o menor valor de QE (quatro erros) está compreendida no processamento P6. Portanto, essa configuração foi gerada com limiares de energia iguais a 0,001002 (LI) e 0,007760 (LF),

limiares de tempo iguais a 6 (TI) e 24 (TF), extração de características cepstrais e uso da regra de decisão I (vide Tabela 4.1). Na Tabela 4.4, é apresentada a matriz de confusão para a configuração de menor erro.

Tabela 4.4: Matriz de confusão para a configuração de menor erro.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*
go	108	2	0	0	0	0	0
help	0	110	0	0	0	0	0
no	0	1	109	0	0	0	0
repeat	0	0	0	110	0	0	0
stop	0	0	0	0	109	1	0
start	0	0	0	0	0	110	0
yes	0	0	0	0	0	0	110

A matriz de confusão apresenta a distribuição das palavras reconhecidas de forma correta e errada. Para a base de dados utilizada, a matriz de confusão ideal deveria apresentar na sua diagonal principal somente o valor 110[17], porém, há a ocorrência de falsos reconhecimentos para as palavras *go*, *no* e *stop*. Contudo, estes erros não são significativos e a matriz de confusão para a configuração de menor erro aproxima-se da matriz ideal.

É importante destacar, que essa configuração de menor erro foi gerada com os limiares de energia mais indicados para a detecção dos trechos de atividade de voz, conforme exposto anteriormente.

Além disso, conforme literatura da área (RABINER e JUANG, 1993; CAEIROS, MIYATAKE e MEANA, 2009), a extração de características cepstrais mostrou-se mais adequada para o reconhecimento de palavras do que a extração de características LPC. Esse fato é evidenciado na Tabela 4.3, visto que, a quantidade mínima de erros para os processamentos com extração LPC foi 7 vezes maior que o maior erro mínimo dos processamentos com extração cepstral. As demais figuras e tabelas desta seção também evidenciam esse fato.

[17] Corresponde ao número total de elocuções de teste de uma dada palavra da base de dados utilizada.

4.2 Análise das Regras de Decisão

Na Seção 4.1 foi apresentada uma análise dos processamentos da Tabela 4.1. O objetivo desta análise foi identificar a configuração de sistema que produz a quantidade de erros mínima, dado a utilização da base de dados descrita na Seção 3.1. Lembrando que, essa base de dados possui 1.232 elocuções de treinamento e 770 elocuções de teste, conforme apresentado na Seção 3.1. Ainda conforme a Seção 4.1, a configuração destacada na Tabela 4.5 apresentou uma resposta de 4 erros e 766 acertos.

Tabela 4.5: Configuração com QE mínima.

Limiares Detector de Voz				Extração de Características	Regra de Decisão
LI	LF	TI	TF	Coeficientes cepstrais	RD-I
0,001002	0,007760	6	24		

Nesta seção, faz-se uma análise da aplicação das regras de decisão II, III e IV, descritas na Seção 3.8, uma vez que, a análise inicial, apresentada na Seção 4.1, somente levou em consideração a aplicação de RD-I. Para a análise dessas regras, os limiares do detector de voz e a extração de características não foram alterados e assumiram os valores apresentados na Tabela 4.5.

O objetivo desta análise é identificar uma regra e um limiar de decisão (LD), que minimize a métrica QE, mas que não cause impactos significativos na quantidade de acertos[18] (QA). A introdução de um limiar nas regras de decisão pode alterar o resultado obtido com a configuração apresentada na Tabela 4.5, uma vez que a resposta "desconhecido" é inserida ao sistema. Contudo, a introdução de um limiar de decisão no processo de reconhecimento de palavras visa evitar a ocorrência de falsos reconhecimentos (erros), tanto no universo de palavras cadastradas quanto no de palavras não cadastradas no sistema.

Para a análise das regras de decisão II, III e IV foi utilizada uma sequência de três mil limiares de decisão, na qual a diferença entre limiares consecutivos é constante e igual a 0,00001, sendo o primeiro e o último limiares da sequência iguais a 0,00001 e 0,03000, respectivamente.

---

[18] Número de sentenças de teste que foram associadas corretamente aos seus respectivos padrões de referência.

Nas Figuras 4.6, 4,7 e 4.8, são apresentados os gráficos das funções de resposta do sistema quando da variação do limiar de decisão de RD-II.

Figura 4.6: Função Acerto quando da variação do limiar de decisão de RD-II.

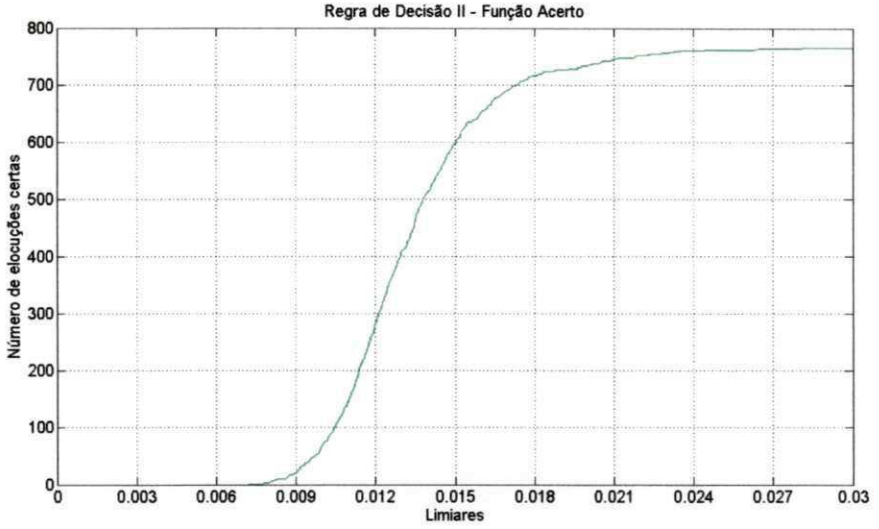
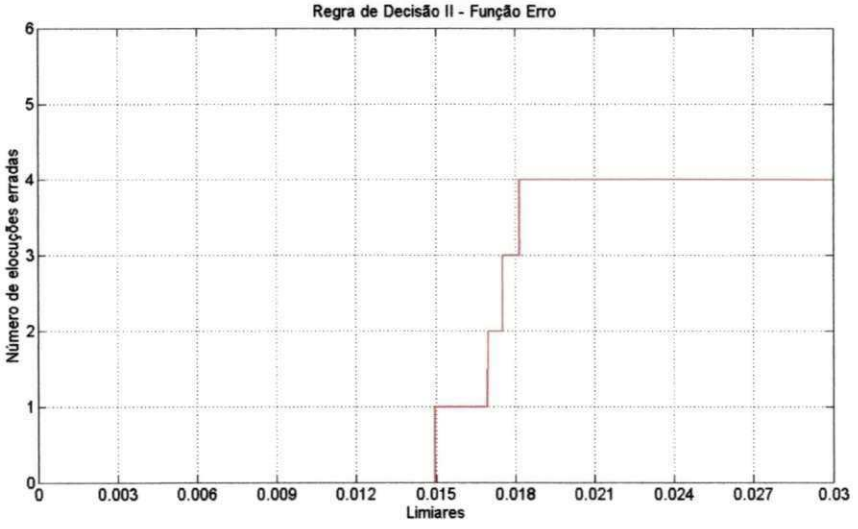


Figura 4.7: Função Erro quando da variação do limiar de decisão de RD-II.



As funções Acerto (Figura 4.6) e Erro (Figura 4.7) apresentam um comportamento ascendente à medida que o limiar de decisão aumenta. Na faixa de limiares de 0,00001 a aproximadamente 0,01500 a função Erro é constante e igual a zero, enquanto que, a função Acerto varia de zero a aproximadamente 600 elocuções. No trecho entre os limiares 0,015 e 0,018 a função Erro assume três valores (1, 2 e 3), ao passo que, neste trecho, a função Acerto

varia de aproximadamente 600 a 718 elocuções. A partir do limiar 0,018 a função Erro é igual a quatro e a função Acerto assume valores na faixa de 719 a 770 elocuções. A função Desconhecido, por sua vez, apresenta um comportamento descendente quando o limiar de decisão aumenta, conforme está apresentado na Figura 4.8. Na Figura 4.9, está apresentada uma comparação das funções de resposta do sistema para a regra de decisão II.

Figura 4.8: Função Desconhecido quando da variação do limiar de decisão de RD-II.

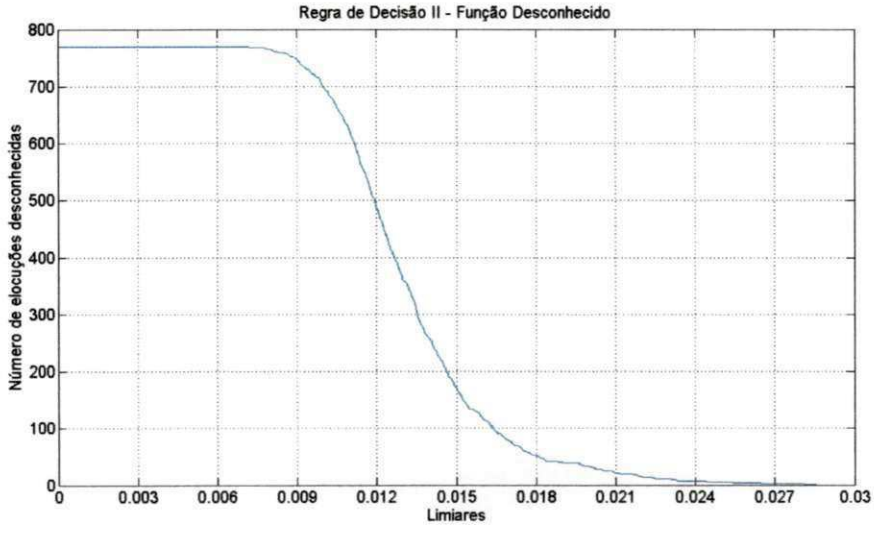
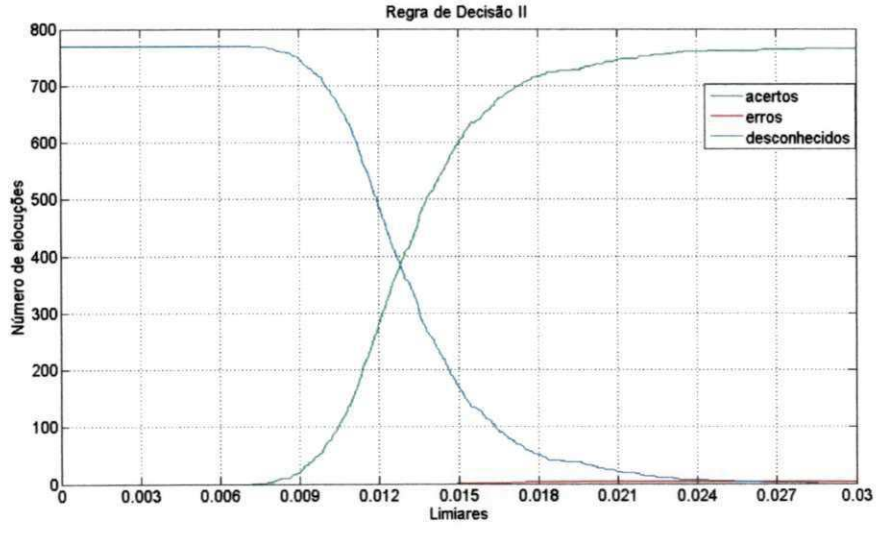


Figura 4.9: Comparação das funções de resposta do SRF quando da variação do limiar de decisão de RD-II.



Os gráficos das funções de resposta do sistema para RD-III são apresentados nas Figuras 4.10, 4.11 e 4.12.

Figura 4.10: Função Acerto quando da variação do limiar de decisão de RD-III.

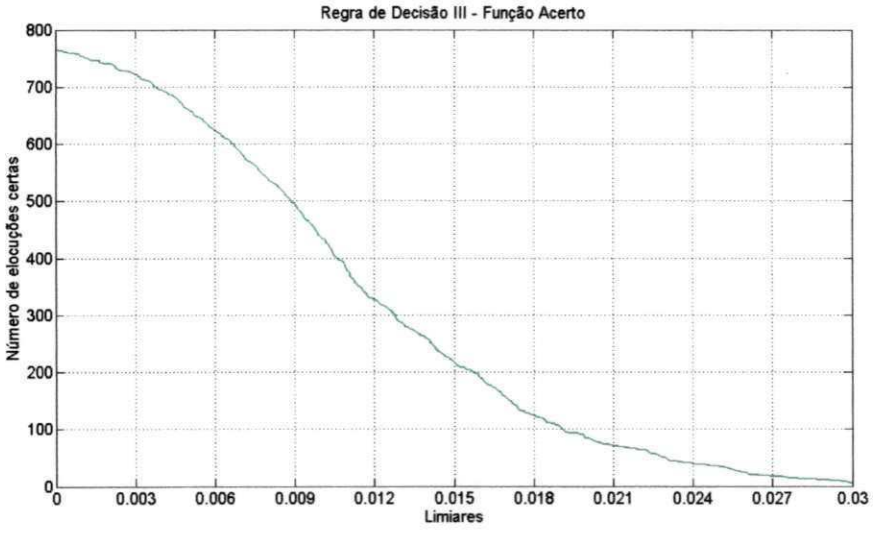
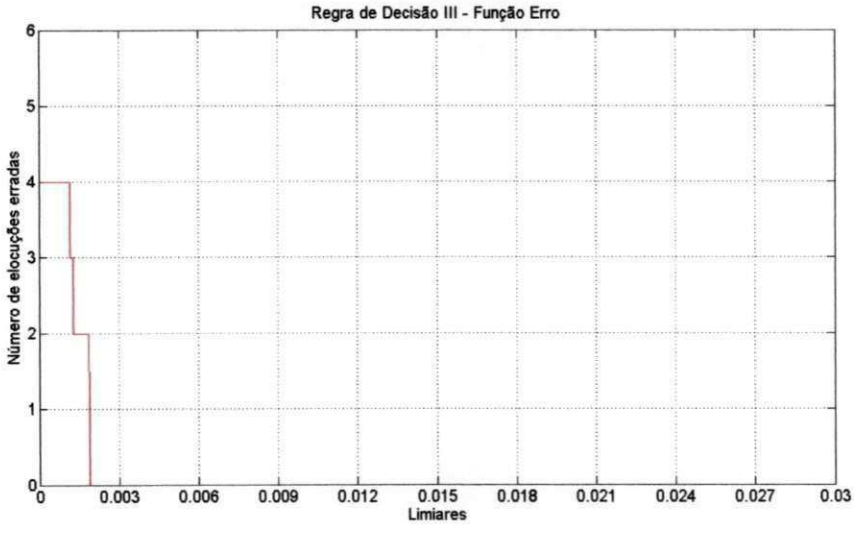


Figura 4.11: Função Erro quando da variação do limiar de decisão de RD-III.



As funções Acerto (Figura 4.10) e Erro (Figura 4.11) da regra de decisão III, diferentemente de RD-II, apresentam um comportamento descendente à medida que o limiar de decisão aumenta. Na faixa de limiares de 0,00001 a 0,00188 a função Erro assume quatro valores (4, 3, 2 e 1), enquanto que, a função Acerto varia de 766 a 742 elocuções. Ainda de acordo com os gráficos das funções Erro e Acerto (Figura 4.11 e Figura 4.10), a partir do

limiar 0,00189 a função Erro permanece com valor zero, ao passo que, a função Acerto varia na faixa de 742 a 6 elocuções. O comportamento da função Desconhecido (Figura 4.12) é ascendente quando o limiar de decisão de RD-III aumenta, uma vez que a tolerância para diferenciar as palavras aumenta (ver Seção 3.8). A comparação das funções de resposta do sistema para RD-III está apresentada na Figura 4.13.

Figura 4.12: Função Desconhecido quando da variação do limiar de decisão de RD-III.

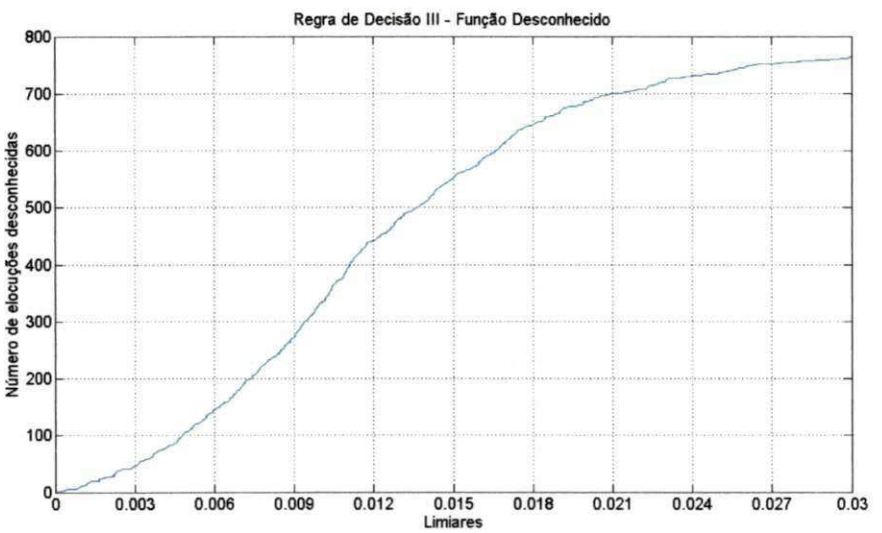
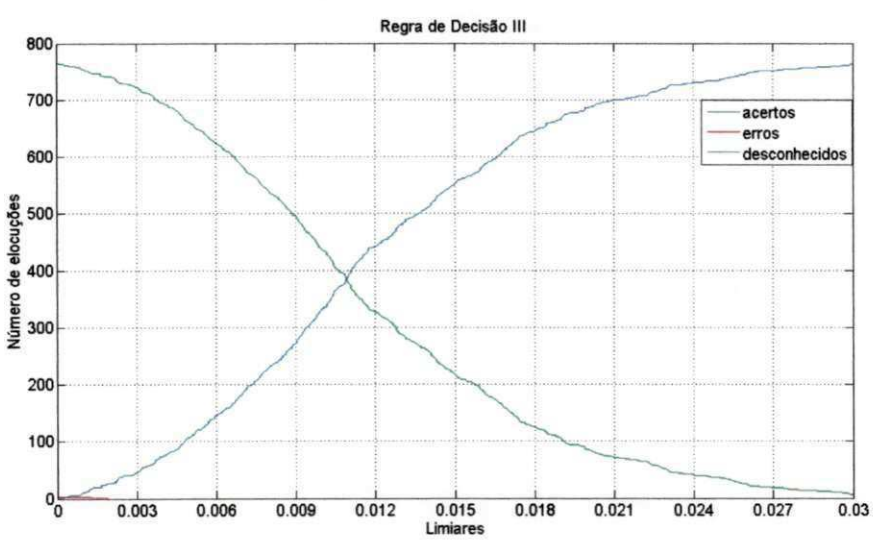


Figura 4.13: Comparação das funções de resposta do SRF quando da variação do limiar de decisão de RD-III.



Nas Figuras 4.14, 4.15 e 4.16, são apresentados os gráficos das funções de resposta do sistema quando da aplicação de RD-IV.

Figura 4.14: Função Acerto quando da variação do limiar de decisão de RD-IV.

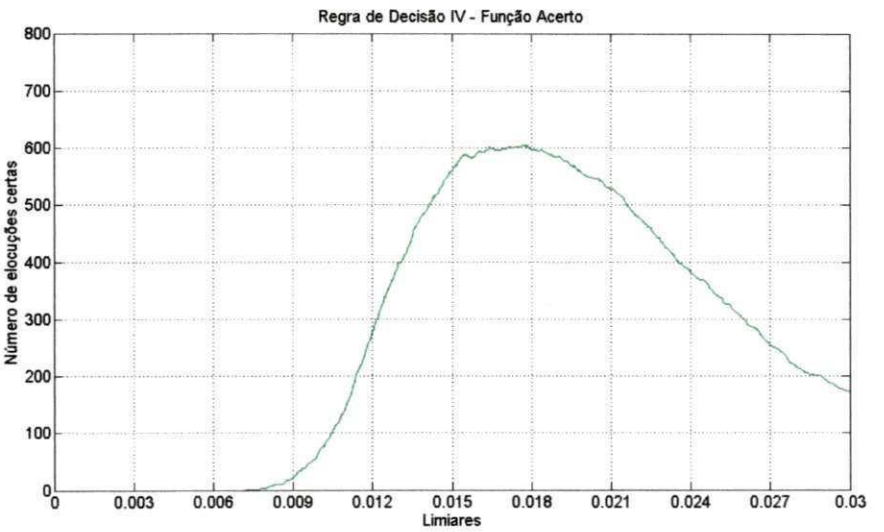
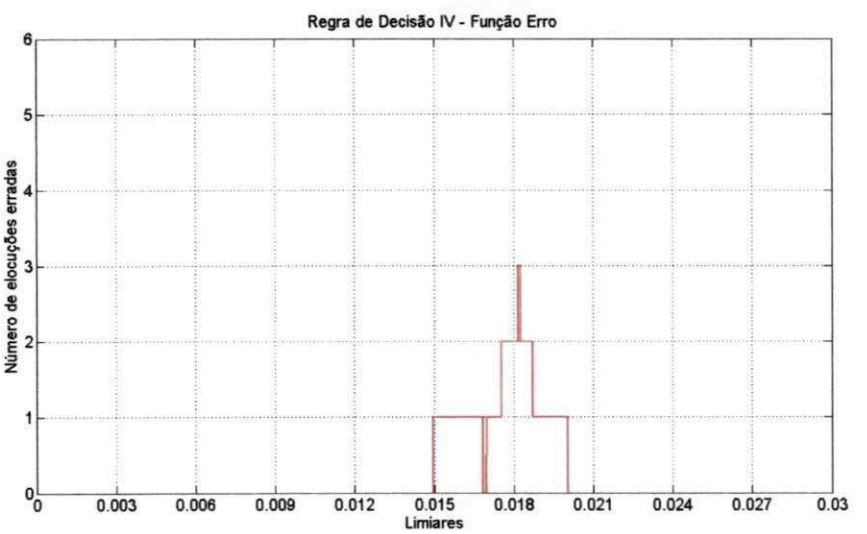


Figura 4.15: Função Erro quando da variação do limiar de decisão de RD-IV.



Para o universo de limiares de decisão analisados o valor máximo da função Erro (Figura 4.15) foi de 3 elocuções. Esse valor da função Erro ocorre para a faixa de limiares de 0,01817 a 0,01824, sendo que, a variação da função Acerto (Figura 4.14) nesse trecho é de 597 a 594 elocuções. Contudo, o valor máximo da função Acerto, 606 elocuções, foi obtido com um limiar de 0,01777. Para este limiar, a função Erro assumiu o valor 2 e a função Desconhecido (Figura 4.16) o valor 162. Na Figura 4.17 tem-se a comparação das funções de resposta do sistema para a regra de decisão IV.

Figura 4.16: Função Desconhecido quando da variação do limiar de decisão de RD-IV.

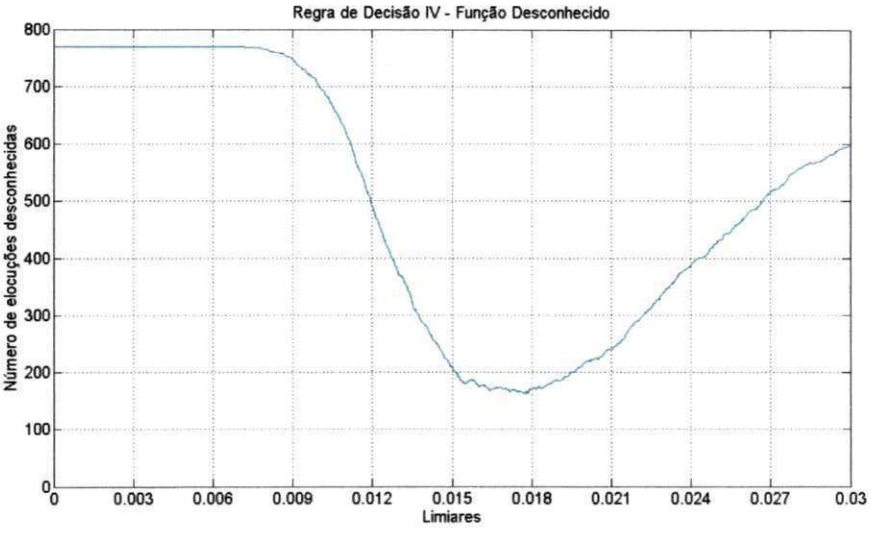
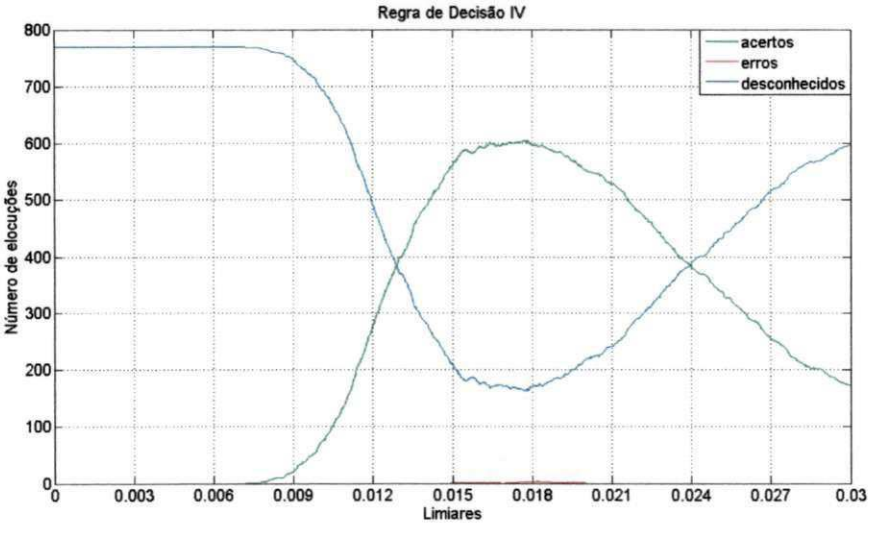


Figura 4.17: Comparação das funções de resposta do SRF quando da variação do limiar de decisão de RD-IV.



Nas Figuras 4.6, 4.7 e 4.8 foram apresentados de forma geral os gráficos das funções de resposta do sistema quando da aplicação de RD-II, RD-III e RD-IV, respectivamente. No entanto, uma análise comparativa entre essas regras se faz necessária, uma vez que, é desejável identificar qual regra de decisão apresenta o melhor desempenho, considerando-se as métricas QE, QA e QD (Quantidade de Desconhecidos)[19]. Na Tabela 4.6, é apresentada uma comparação entre essas três regras. As matrizes de confusão para as configurações de sistema apresentadas na Tabela 4.6 estão apresentadas no Apêndice B.

[19] Número de sentenças de teste que não foram associadas a nenhum dos padrões de referência armazenados.

Tabela 4.6: Comparação as regras de decisão.

QE	RD-II			RD-III			RD-IV		
	QA	QD	LD	QA	QD	LD	QA	QD	LD
0	596	174	0,01493	742	28	0,00189	599	171	0,01693
1	690	79	0,01693	742	27	0,00188	602	167	0,01712
2	707	61	0,01751	749	19	0,00129	606	162	0,01777
3	718	49	0,01812	751	16	0,00117	597	170	0,01817
4	766	0	0,02855	766	0	0,00001	-	-	-

Na coluna QA da Tabela 4.6, está apresentada a quantidade de acertos máxima, conseguida com a aplicação de uma dada regra e um dado limiar de decisão, para uma faixa em que a quantidade de erros é fixa em um dado valor. Por exemplo, na faixa em que a função Erro assume o valor zero, a quantidade de acertos máxima obtida é de 569 elocuções, enquanto que, a quantidade de desconhecidos é de 174 elocuções. Essas métricas foram obtidas com a aplicação de RD-II e um limiar de decisão de 0,01493.

Conforme informações da Tabela 4.6, a regra de decisão III apresenta um desempenho melhor, uma vez que, esta regra gerou uma quantidade de acertos maior ou igual às demais regras, considerando-se uma faixa fixa de erros. Ainda de acordo com a Tabela 4.6, a configuração de sistema que apresentou como resposta um reconhecimento de 742 elocuções e uma quantidade de 28 elocuções desconhecidas, com nenhum erro, foi conseguida com a aplicação de RD-III e com um limiar de decisão de 0,00189. Portanto, as taxas de acertos, de erros e de desconhecidos obtidas foram de 96,36%, 0% e 3,64%, respectivamente.

4.3 Comparação com Sphinx3

Nas seções anteriores deste capítulo, foram apresentadas as análises de diversas configurações de sistema com o objetivo de identificar a configuração que proporciona a melhor resposta do sistema, tendo como parâmetros de avaliação as métricas QA, QE e QD.

Essa seção, por sua vez, tem como objetivo apresentar os resultados obtidos com a ferramenta Sphinx3, utilizando-se a base de dados descrita na Seção 3.1. O uso da mesma base de dados utilizada nas fases de treinamento e teste do SRF desenvolvido possibilita uma comparação entre esses dois sistemas.

Na Tabela 4.7, é apresentada a configuração utilizada na ferramenta Sphinx3. Com essa configuração, obteve-se uma taxa de acerto de 100%, ou seja, as 770 elocuções de teste, da base de dados descrita na Seção 3.1, foram reconhecidas corretamente com o uso da ferramenta. Como descrito na Seção 4.2, a configuração mais indicada para o sistema desenvolvido apresenta uma taxa de acerto de 96,36% e uma taxa de desconhecidos de 3,64%. Logo, a partir do critério da taxa de acerto, o Sphinx3 apresenta melhor desempenho que o sistema desenvolvido. Contudo, é importante destacar que, o Sphinx3 utiliza de técnicas mais complexas para o reconhecimento de voz, tais como, HMM contínuo, modelo linguístico e coeficientes mel cepstrais. O sistema desenvolvido utiliza técnicas mais simples, tais como, quantização vetorial, comparação por distância euclidiana e coeficientes cepstrais obtidos a partir dos coeficientes LPC. A opção por técnicas mais simples é justificada pelos requisitos de baixo consumo e de simplificação das operações visando uma implementação futura em *hardware*.

Tabela 4.7: Configuração utilizada no Sphinx3.

Fator de pré-ênfase	0,9375 (15/16)
Tamanho da janela	20ms
Extração de características	mel cepstrais
Número de coeficientes	12
Número de pontos da FFT (*Fast Fourier Transform*)	512
Frequência de amostragem	11.025 Hz
Frequências de filtragem	130 Hz (baixa) e 5400 Hz (alta)
Número de filtros	36

4.4 Considerações para uma Implementação em *Hardware*

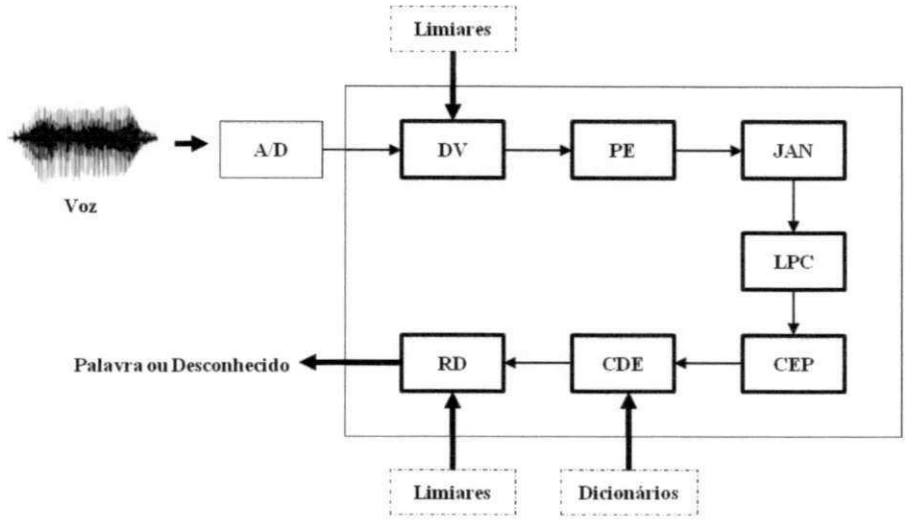
Nas seções anteriores, foram apresentadas análises das configurações de sistema simuladas neste trabalho. O objetivo dessas análises foi indicar a configuração de sistema mais adequada para o SRF desenvolvido.

Essa seção aborda, em linhas gerais, considerações para uma implementação do sistema proposto em *hardware*. Para tanto, várias adaptações se fazem necessárias, como por exemplo, utilização de ponto fixo nas operações com números fracionários, paralelismo, *pipeline*, controle da largura de *bits*, simplificação das funções matemáticas, substituição de operações de multiplicação e divisão por deslocamentos, dentre outras (CIPRIANO, 2001;

DIAS, 2011). É importante destacar, que em um modelo de referência em *software*, estas abordagens não são comumente exploradas, visto que, o objetivo é a modelagem do sistema.

Na Figura 4.18, está apresentado o diagrama em blocos da arquitetura em *hardware* do sistema de reconhecimento de palavras isoladas proposto. Esta arquitetura é baseada na modelagem arquitetural de (FECHINE et al., 2010) e é composta por 7 blocos: DV (Detector de Voz), PE (Pré-ênfase), JAN (Janelamento e Segmentação), LPC, CEP (Cepstrais), CDE (Comparação por Distância Euclidiana) e RD (Regra de Decisão).

Figura 4.18: Diagrama em blocos da Arquitetura do SRF.



Adaptado de (FECHINE et al., 2010).

A comunicação entre blocos nesta arquitetura deverá ser feita pelo protocolo AMBA[20] e todos os blocos funcionais devem ser implementados com máquinas de estados finitas, objetivando a redução do consumo energético (FECHINE et al., 2010). Com o uso de máquinas de estados finitos é possível otimizar o número de transições das portas lógicas, principalmente as transições desnecessárias resultantes de atrasos de propagação do sinal. Uma menor quantidade de transições possibilita uma redução no consumo de energia (MEIXEDO, 2008).

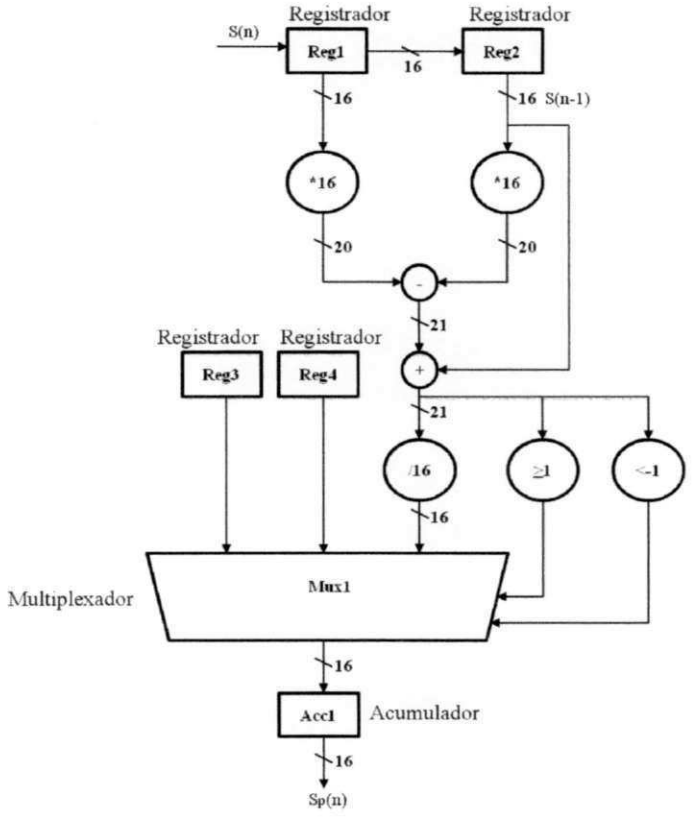
O sinal de voz amostrado (a uma frequência de 11.025 Hz), quantizado (16 bits) e normalizado (compreendido na faixa de [-1:1].), constitui-se uma das entradas do bloco DV,

[20] *AMBA ® AXI Protocol Specification*, disponível em www.arm.com.

juntamente com os limiares de energia e de tempo apresentados na Tabela 4.5. Este bloco necessita armazenar uma determinada quantidade de amostras de voz para a tomada de decisão sobre o início e o fim das palavras, conforme algoritmo apresentado na Seção 3.2. Para a decisão de início, 660 amostras[21] são necessárias, enquanto que para a decisão de fim são necessárias 2640 amostras[22]. Logo, o DV precisará de uma memória RAM R/W de 4Kx16 bits para endereçar 3300 amostras.

O próximo bloco PE aplica o filtro de pré-ênfase nas amostras do sinal S(n). O fator de pré-ênfase de 15/16 possibilitou uma simplificação da arquitetura. Na Figura 4.19, está apresentada a arquitetura simplificada para a realização da pré-ênfase.

Figura 4.19: Arquitetura da pré-ênfase.



Esta arquitetura foi idealizada pela equipe Brazil-IP CG (FECHINE et al., 2010). Nesta arquitetura, os registradores Reg1 e Reg2 são inicializados com o valor zero e os

[21] Corresponde a 6 segmentos de 110 amostras.
[22] Corresponde a 24 segmentos de 110 amostras.

registradores Reg3 e Reg4 com os valores -0,99999 e 0,99999 (notação em ponto fixo), respectivamente. No primeiro pulso de relógio, Reg1 é atualizado com o valor da primeira amostra de voz S(1) e Reg2 continua com o valor zero. No pulso seguinte, o cálculo da pré-ênfase é realizado. O cálculo da pré-ênfase nesta arquitetura sofreu uma reformulação, uma vez que, a Equação 2.8 foi multiplicada por 16, resultando na Equação 4.1.

$$s_p(n) = s(n) - s(n-1) + \frac{s(n-1)}{16} \therefore 16s_p(n) = 16s(n) - 16s(n-1) + s(n-1). \qquad (4.1)$$
$$\frac{16s_p(n)}{16} = 16s_p(n).$$

As operações de cálculo da pré-ênfase ocorrem em paralelo e no próximo pulso de relógio são realizadas comparações com o objetivo de identificar se o cálculo da pré-ênfase resultou em um valor fora do intervalo [-1:1[. Em caso afirmativo, é feita a saturação para um dos valores armazenados em Reg3 ou Reg4. Na saída da operação de multiplexação, a amostra pré-enfatizada Sp(1), é armazenada no acumulador Acc1.

Na etapa seguinte, um novo dado é lido, S(2). Este dado é armazenado em Reg1 e o valor anterior de Reg1, a amostra de voz S(1), é transferido para Reg2. Nos dois pulsos subsequentes, acontece o cálculo da pré-ênfase e a multiplexação, resultando na amostra pré-enfatizada Sp(2) que é armazenada no acumulador. Este procedimento será repetido até que todas as amostras do sinal de voz sejam processadas.

As operações de multiplicação e divisão por 16 da Equação 4.1, são realizadas fazendo-se deslocamentos de quatro *bits*. Nas operações de multiplicação, o deslocamento é feito para a esquerda, enquanto que, na divisão, para a direita.

Após a pré-ênfase, as amostras Sp(n) são enviadas ao bloco JAN, que realiza as operações de janelamento (função Hamming) e divisão em quadros de 220 amostras com superposição de 50%. Este bloco possui uma memória ROM de 2^7x16 bits que armazena os valores da função Janela de Hamming (Equação 2.10). São necessários apenas 7 bits para o endereçamento dos valores da função devido às propriedades de simetria da janela (J(n), com *n* variando de 0 a 109 é simétrico a J(n) com *n* variando de 110 a 220).

No próximo passo, as amostras janeladas seguem de forma serial para o bloco LPC, no qual serão calculados os coeficientes LPC pelo método da autocorrelação (Equações 2.22 e 2.23). É necessário armazenar 220 amostras de voz para o cálculo de 12 coeficientes, que

serão enviados de forma serial para o bloco CEP que realiza o cálculo dos coeficientes cepstrais pela Equação 2.29, descrita na Seção 2.3.3.2.

O bloco CDE recebe 12 coeficientes cepstrais de forma serial e calcula a distância deste vetor em relação aos 7 dicionários. A distância empregada neste bloco é a distância euclidiana (vide Equação 2.32). Desta forma, são enviadas 7 distâncias (serial) para o bloco RD. Para armazenar os 7 dicionários, cada um com dimensão 12 e 64 níveis, torna-se necessário o uso de uma memória RAM R/W de 8Kx11 bits para o endereçamento de 5376 valores[23].

No bloco RD é implementado o algoritmo de decisão RD-III (Seção 3.8) e efetuado o cálculo da média de cada uma das 7 distâncias enviadas, armazenando a menor e a segunda menor média, como também o índice do dicionário que caracteriza a menor média. A decisão é tomada a partir da comparação da diferença entre a segunda menor e a menor média das distâncias. Caso a diferença seja maior que o limiar de decisão, a saída do bloco é o índice da palavra associado à menor distância, caso contrário a saída é "desconhecido".

4.5 Discussão

Neste capítulo, foram analisadas diversas configurações do sistema. A configuração que apresentou a resposta mais adequada, com taxa de acerto de 96,36%, está apresentada na Tabela 4.8.

Tabela 4.8: Configuração do SRF desenvolvido.

Frequência de Amostragem	11025 Hz
Fator de Pré-ênfase	0,9375
Tamanho do Segmento para Detecção de Voz	10 ms (110 amostras)
Limiares do Detector de Voz	LI=0,001002; LF= 0,007760; TI= 6 e TF=24
Segmentação e Janelamento	Segmentos de 20 ms (220 amostras) e janela de Hamming com superposição de 50%
Extração de Características	Coeficientes cepstrais
Número de Coeficientes	12
Comparação	Distância Euclidiana
Regra de Decisão	RD-III com limiar de decisão de 0,00189

[23] Os 5376 valores correspondem a 7 dicionários x 64 níveis x dimensão 12.

Além disso, também foi apresentado um resultado com 100% de reconhecimento com a ferramenta Sphinx3, utilizando-se a mesma base de dados (descrita na Seção 3.1). Destaca-se, porém, a complexidade da modelagem do Sphinx3 em comparação ao modelo proposto.

Por fim, foram feitas considerações para a implementação do sistema de reconhecimento de palavras isoladas em *hardware*.

5 Considerações Finais e Sugestões para Trabalhos Futuros

Por ser a voz um meio de comunicação prático e usual entre humanos, é de interesse que as facilidades dessa comunicação sejam estendidas a comunicação homem-máquina, uma vez que esse tipo de comunicação tem se tornado cada vez mais habitual (RABINER e SCHAFER, 1978; VIDAL, 2006).

Contudo, apesar dos avanços teóricos e técnicos no estudo das interfaces de voz na interação com as máquinas, ainda não é possível atender às expectativas dos usuários em um reconhecimento com a mesma complexidade e taxa de acerto do ser humano (GOMES, 2007). Além disso, quando as máquinas em questão são dispositivos embarcados com dependência de bateria, que apresentam o requisito de baixo consumo energético, uma dificuldade adicional é inserida no projeto dessas interfaces de voz.

O objeto de estudo deste trabalho está inserido na área de Reconhecimento de Fala e consistiu no estudo das configurações mais adequadas para se construir um sistema de reconhecimento automático de palavras isoladas, independente de locutor, a ser utilizado como modelo de referência para implementações em *hardware* que apresentem o requisito de baixo consumo energético.

Este capítulo, apresenta as conclusões obtidas no estudo das configurações do SRF desenvolvido, destacando as contribuições mais relevantes, além de indicar sugestões para trabalhos futuros.

5.1 Contribuições

A partir dos resultados obtidos, pode-se destacar as contribuições apresentadas a seguir.

- Construção de base de dados, composta de um vocabulário de 7 palavras (*go*, *help*, *no*, *repeat*, *start*, *stop* e *yes*), gerada a partir de 11 locutores, que possui 1.232 elocuções de treinamento e 770 elocuções de teste (Seção 3.1).

- Descrição de um método para auxiliar a determinação dos limiares do detector de voz para uma dada base de dados (Seção 3.2).
- Análise dos parâmetros do detector de voz, de modo que se verificou que o bom desempenho do sistema está associado aos ajustes dos limiares de energia e de tempo (Seção 4.1). A eficiência na detecção da voz proporciona também uma redução no consumo energético, uma vez que os períodos de inatividade da voz, tais como, silêncio e ruído de fundo, deixam de ser processados.
- Análise de regras de decisão, de forma que se constatou que o bom desempenho do sistema está associado à determinação de um bom algoritmo de decisão, como também, de ajustes dos limiares de decisão desses algoritmos (Seção 4.2).
- Indicação da configuração de sistema que apresentou a resposta mais adequada, dentre as configurações analisadas, tendo como critério a análise das métricas de desempenho QE, QA e QD (Capítulo 4), como também, a simplificação das operações e a redução do consumo, tendo em vista uma implementação futura do sistema em *hardware*.
- Comparação do desempenho do sistema desenvolvido com o software de reconhecimento Sphinx3 (Seção 4.3).

Diante do exposto, pode-se concluir que as simplificações adotadas com vistas à implementação em *hardware*, com restrições de consumo energético, aliadas à análise criteriosa para a escolha de limiares para detecção de voz (eliminação de intervalos de silêncio) foram eficientes. Essa abordagem não impactou em uma redução significativa do desempenho do sistema em comparação àqueles que utilizam técnicas mais complexas, a exemplo de coeficientes mel cepstrais para a composição do vetor de características e GMM para o classificador.

5.2 Sugestões para trabalhos futuros

Como sugestões para trabalhos futuros, tem-se a implementação em linguagem de descrição de *hardware* do sistema apresentado. Essa implementação deverá ser levada a efeito com o auxílio de uma metodologia de verificação funcional para garantir que o *hardware* prototipado reflita o modelo de referência em *software*.

Recomenda-se investigar também a aplicação de outras técnicas para a extração de características do SRF, tais como, coeficientes *wavelets* e mel cepstrais; para o cálculo da distorção na construção do dicionário do quantizador, tal como, a distância de Manhattan; e para a construção do classificador, a exemplo de, HMM e redes neurais, levando-se em consideração os requisitos de baixo consumo energético. Também é indicada uma análise quantitativa do consumo com a implementação das técnicas propostas.

Por fim, sugere-se a ampliação da base de dados de palavras isoladas, de modo a torná-la mais representativa visando o uso do sistema por um universo significativo de usuários e de aplicações. Recomenda-se que essa ampliação adicione ao conjunto de teste elocuções de locutores que não pertençam ao conjunto de treinamento.

6 Referências Bibliográficas

AMUDHA, V.; VENKATARAMANI, B.; VINOTHKUMAR, R.; RAVISHANKAR, S. SOC Implementation of HMM Based Speaker Independent Isolated Digit Recognition System. *IEEE conference VLSI design*, p.848-853, 2007.

AMUDHA, V.; VENKATARAMANI, B.; MANIKANDAN, J. FPGA Implementation of Isolated Digit Recognition System Using Modified Back Propagation Algorithm. *International Conference on Eletronic Design*, 2008.

BENZEGHIBA, M. F.; BOULARD, H. User-customized password speaker verification based on HMM/ANN and GMM models. *Seventh International Conference on Spoken Language Processing*, p. 1325-1328, 2002.

BERGERON, J. *Writing Testbenches: Functional Verification of HDL Models*. 2 ed. Springer, 2003, 512 p.

BOGERT, P.; HEALY, M. J. R; TUKEY, J. W. The frequency analysis of times series for echos: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. *Proceedings of the Symposium on Time Series Analysis*, 1963.

BOURLARD, H.; MORGAN, N. *Connectionist Speech Recognition: A Hybrid Approach*. Springer, 1994, 312 p.

CAEIROS, A. M.; MIYATAKE, M. N; MEANA, H. P. Isolate Speech Recognition Based on Time-Frequency Analysis Methods. *In:* CORROCHANO, E. B.; EKLUNDH, J. O. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer, 2009. cap. 6, p. 297-304.

CAMPBELL, J. P. Speaker Recognition: A Tutorial. *Proceedings of the IEEE*, v. 85, no. 9, p. 1437-1462, set. 1997.

CARDOSO, D. P. *Identificação de locutor usando modelos de mistura de gaussianas*. 88 f. Dissertação – Universidade de São Paulo, São Paulo. 2009.

CARVALHO, L. M. P. *Sistema de Verificação do Orador, Baseado em Modelos de Markov, Compactado num Objecto COM para Windows*. 134 f. Dissertação – Universidade do MINHO, Guimarães, Portugal. 2007.

CHEN, F.; MASI, C. Effect of noise on automatic speech recognition system error rate. *IEA 2000/HFES 2000 Congress,* San Diego, USA, p. 606-609, 2000.

CHEN, F. *Designing Human Interface In Speech Technology*. Springer, 2005. 382 p.

CIPRIANO, J. L. G. *Desenvolvimento de Arquitetura para Sistemas de Reconhecimento Automático de Voz Baseados em Modelos Ocultos de Markov*. 125 f. Tese (Doutorado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul, Porto Alegre. 2001.

CIPRIANO, J. L. G.; NUNES, R. P.; BARONE, D. A. C. Implementation of Voice Processing Algorithms in FPGA Hardware. *Proceedings of XIX Workshop Iberchip*, 2003

CUNHA, C. F. *Gramática do Português Contemporâneo*. Lexikon, 2007. 424 p.

DIAS, D. *Desenvolvimento de um IP Core de Pré-processamento Digital de Sinais de Voz para Aplicação em Sistema Embutidos.* 108 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Campina Grande. 2006.

DIAS, D. *Reconhecimento de Fala Contínua para o Português Brasileiro em Sistemas Embarcados.* 178 f. Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Campina Grande. 2011.

FECHINE, J. M.; AGUIAR-NETO, B. G. Modelagem de identidade vocal utilizando modelos de markov escondidos. *XVI Congresso Nacional de Matemática Aplicada e Computacional – CNMAC*, 1993.

FECHINE, J. M. *Verificação de Locutor Utilizando Modelos de Markov Escondidos (HMMs) de Densidades Discretas*. 147 f. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal da Paraíba, Campina Grande. 1994.

FECHINE, J. M. *Reconhecimento automático de identidade vocal utilizando modelagem híbrida: Paramétrica e Estatística*. 212 f. Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Campina Grande. 2000.

FECHINE, J. M.; PAIXÃO, L.; JÚNIOR, A.; MELO, F.; ESPÍNOLA, S. SPVR: An IP core for Real-Time Speaker Verification. *IP-SOC Conference*, 2010.

FELLBAUM, K. *Sprachsignalverarbeitung and Sprachübertragung*. Springer-verlag, 1984.

FURUI, S. Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, v. 29, no. 2, p. 254-272, abr. 1981.

GOMES, R. J. R. *Teste de Interfaces de Voz*. 114 f. Dissertação – Faculdade de Engenharia da Universidade do Porto, Porto. 2007.

GUIMARÃES, I. *A Ciência e a Arte da Voz Humana*, ESSA – Escola Superior de Saúde do Alcoitão, 2007.

HAYES, M. H. *Processamento Digital de Sinais*. Bookman, 1999. 462 p.

HERBIG, T.; GERL, F.; MINKER, W. *Self-Learning Speaker Identification: A System for Enhanced Speech Recognition*. Springer, 2011. 172 p.

JIANG, X.; MA, M. Y.; CHEN, C. W. *Mobile Multimedia Processing: Fundamentals, Methods, and Applications*, 1. ed. Springer, 2010. 287 p.

KIM, S-N; HWANG, I-C; KIM, Y-W; KIM, S-W A VLSI Chip for Isolated Speech Recognition System. *International Conference on Consumer Electronics*, p. 118-119, 1996.

LAMAS, R. M. L. S. *Avaliação de Codificadores de Voz em Ambiente VoIP*. 90 f. Dissertação – Universidade Federal do Rio de Janeiro. 2005.

LEE, C.; HYUN, D.; CHOI, E.; GO, J.; LEE, C. Optimizing Feature Extraction for Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, v. 11, n. 1, p. 80-87, jan. 2003.

LEE, K.; HAUPTMANN, A. G.; RUDNICKY, A. The Spoken Word, *Byte*, p. 225-232, 1990.

LI, Q.; ZHENG, J.; TSAI, A.; ZHOU, Q. Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition. *IEEE Transactions on Speech and Audio Processing*, v. 10, no. 3, 2002.

LINDE, Y.; BUZO, A.; GRAY, R. M. An algorithm for vector quantizer design. *IEEE Transaction on Communications*, v. 28, n. 1, p. 84-95, 1980.

MAIA, B. M. F. *Descaracterização Perceptiva da Assinatura Vocal*. 109 f. Dissertação – Faculdade de Engenharia da Universidade do Porto, Porto. 2010.

MALKIN, R. G. *Machine Listening for Context-aware Computing*. 201 f. Tese – Carnegie Mellon University, Pittsburgh. 2006.

MARPLE, S. L. *Digital Spectral Analysis with Applications*. Prentice Hall, 1987. 492 p.

MEIXEDO, J. M. R. *Metodologias de projecto de baixo consumo para implementações em FPGA*. 75 f. Dissertação (Mestrado em Engenharia Eletrotécnica e de Computadores) – Faculdade de Engenharia da Universidade do Porto, Porto. 2008.

MILNER, B. Speech feature extraction and reconstruction. *In:* TAN, Z-H; LINDBERG, B. Automatic Speech Recognition on Mobile Devices and over Communication Networks. Springer, 2008. cap. 6, p. 107-130.

MÜLLER, C. *Speaker Classification I: Fundamentals, Features, and Methods.* Springer, 2007. 355 p.

NAKAMURA, K.; ZHU, Q.; MARUOKA, S.; HORIYAMA, T.; KIMURA, S.; WATANABE, K. Speech Recognition Chip for Monosyllables, *Asia and South Pacific Design Automation Conference*, p. 396-399, 2001.

NEVES, C.; VEIGA, A.; SÁ, L.; PERDIGÃO, F. Efficient Noise-Robust Speech Recognition Front-End Based on the ETSI Standard. *9th International Conference on Signal Processing*, Beijing, p. 609-612, 2008.

O'SHAUGHNESSY, D. *Speech Communications: Human and Machine.* 2. ed. Wiley-IEEE Press, 2000. 548 p.

O'SHAUGHNESSY, D. Interacting with computers by voice: automatic speech recognition and synthesis, *Proceedings of the IEEE*, v. 91, no. 9, p. 1272-1305, 2003.

OLIVEIRA, H. F. A. *BVM: Reformulação da metodologia de verificação funcional VeriSC*. 139 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Campina Grande. 2010.

OLIVEIRA, M. P. B. *Verificação Automática do Locutor, Dependente de Texto, Utilizando Sistemas Híbridos MLP/HMM*. 103 f. Dissertação (Mestrado em Engenharia Elétrica) – Instituto Militar de Engenharia, Rio de Janeiro. 2001.

PAPAMICHALIS, P. E. *Practical Approaches to Speech Coding*. Prentice-Hall, 1987. 400 p.

PETRY, A.; ZANUZ, A.; BARONE, D. A. C. Utilização de Técnicas de Processamento Digital de Sinais para a Identificação Automática de Pessoas pela Voz. *SSI99*, 1999.

PETRY, A.; ZANUZ, A.; BARONE, D. A. C. Reconhecimento Automático de Pessoas Pela Voz Através de Técnicas de Processamento Digital de Sinais. *SEMAC 2000*, 2000.

PHADKE, S.; LIMAYE, R.; VERMA, S.; SUBRAMANIAN, K. On Design and Implementation of an Embedded Automatic Speech Recognition System. *17th International Conference on VLSI Design*, p. 127-132, 2004.

RABINER, L. R.; SCHAFER, R. W. *Digital Processing of Speech Signals.* Prentice Hall, 1978. 512 p.

RABINER, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, v. 77, n. 2, p.257-286, 1989.

RABINER, L. R.; JUANG, B. H. *Fundamentals of Speech Recognition.* Prentice Hall, 1993. 496 p.

RABINER, L. R.; SCHAFER, R. W. *Theory and Applications of Digital Speech Processing.* Prentice Hall, 2010. 1056 p.

REZENDE, J. A. M. *Efeitos da Segmentação em Sistemas Híbridos ANN+HMM*. 125 f. Dissertação (Mestrado em Engenharia Elétrica) – Instituto Nacional de Telecomunicações, Santa Rita do Sapucaí. 2005.

SHERWANI, J.; PALIJO, S.; MIRZA, S.; AHMED, T.; ALI, N.; ROSENFELD, R. Speech vs. Touch-tone: Telephony Interfaces for Information Access by Low Literate Users, *ICTD*, 2008.

SHIRALI-SHAHREZA, S.; SAMETI, H.; SHIRALI-SHAHREZA, M. Parental Control Based on Speaker Class Verification, *IEEE Transactions on Consumer Electronics*, v. 54, n. 3, ago. 2008.

SILVA, A. G. *Reconhecimento de Voz para Palavras Isoladas*. 60 f. Monografia – Universidade Federal de Pernambuco, Recife. 2009.

SILVA, K. R. G. *Uma Metodologia de Verificação Funcional para Circuitos Digitais.* 121 f. Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Campina Grande,

Campina Grande. 2007.

SINGH, A.; GARG, D. *Soft Computing*. Allied Publishers, 2005. 610 p.

SOONG, F. K.; JUANG, B.-H. A Vector Quantization Approach to Speaker Recognition. *AT&T technical Journal*, New Jersey, v. 66, no. 2, p. 16-26, 1987.

VAREJÃO, R. *Implementação em tempo real de um sistema de reconhecimento de dígitos conectados.* 79 f. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Estadual de Campinas, Campinas. 2001.

VIDAL. E., Computer-assisted translation using speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, v.14, p. 941-951, maio 2006.

VIEIRA, M. N. *Módulo Frontal para um Sistema de Reconhecimento Automático de Voz.* Dissertação (Mestrado em Engenharia Elétrica) – Universidade Estadual de Campinas, Campinas. 1989.

VOJTKO, J.; KOROSI, J.; ROZINAJ, G. Comparison of automatic speech recognizer SPHINX 3.6 and SPHINX 4.0 for creating systems in Slovak language. *15th International Conference on Systems, Signals and Image Processing*, p. 37-539, 2008.

WALKER, W.; LAMERE, P.; KWOK, P.; RAJ, B.; SINGH, R.; GOUVEA, E.; WOLF, P.; WOELFEL, J. Sphinx-4: A Flexible Open Source Framework for Speech Recognition. *Sun Microsystems Laboratories*, 2004.

YING, G.S; MITCHELL, C. D.; JAMIESON, L. H. Endpoint detection of isolated utterances based on a modified Teager energy measurement. *IEEE International Conference on Acoustic, Speech and Signal Processing*, v.2, p.732-735, 1993.

YUANYUAN, S.; JIA, L.; RUNSHENG, L. Single-chip Speech Recognition System Based on 8051 Microcontroller Core. *IEEE Transactions on Consumer Electronics*, v. 47, p. 149-153, 2001.

ZHOU, H.; HAN, H. Design and implementation of speech recognition system based on field programmable gate array. *Modern Applied Science*, v. 3, no. 8, p. 106-111, ago. 2009.

Apêndice A Resultados do Processamento dos dados Coletados

Neste trabalho, foram simuladas 7.200 configurações do SRF agrupadas em 8 processamentos, cada processamento com 900 configurações. Essas configurações foram obtidas com variações dos limiares do módulo de detecção de voz, como também, da alternância entre o uso de coeficientes LPC e cepstrais. Não houve alterações nos demais módulos do SRF para a realização das simulações, esses módulos foram utilizados conforme descrito no Capítulo 3 – uma ressalva deve ser feita com relação ao módulo de decisão, visto que, foi utilizada a regra de decisão I em todas as configurações.

Nas Tabelas A.1, A.2, A.3, A.4, A.5, A.6, A.7 e A.8, é apresentado o resultado de cada um dos 8 processamentos no formato de uma matriz de ordem 30, tal que, cada elemento da matriz corresponde a métrica QE de uma dada configuração. As linhas e colunas destas matrizes representam os limiares TI e TF, respectivamente, do módulo de detecção de voz. Nos 8 processamentos, esses limiares de tempo variam de 1 a 30 quadros[24].

Os resultados apresentados nas Tabelas A.1, A.2, A.3 e A.4 foram gerados com coeficientes LPC, nas demais tabelas desse apêndice utilizou-se os coeficientes cepstrais. Quanto aos limiares de energia do DV, os resultados das Tabelas A.1, A.5, A.4 e A.8 foram gerados com limiares iguais, sendo que, para os resultados das duas primeiras tabelas utilizou-se os limiares LI e LF iguais a 0,001002, enquanto que, para as duas últimas tabelas esses limiares assumiram o valor 0,007760. Para as Tabelas A.2 e A.6 empregou-se limiares de 0,001002 e 0,007760 para LI e LF, respectivamente. Nas Tabelas A.3 e A.7, LI e LF assumiram, respectivamente os valores 0,007760 e 0,001002.

---

[24] Cada quadro de entrada do módulo DV corresponde a um trecho de voz de aproximadamente 10ms.

Tabela A.1: Resultados do Processamento 1[25].

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	73	65	70	64	58	52	58	54	56	58	51	50	46	54	54	57	61	63	51	69	56	70	66	63	70	60	63	50	57	60
2	59	69	76	77	74	76	82	70	72	69	74	55	56	55	65	71	70	59	57	56	83	50	62	75	57	70	67	66	62	53
3	62	46	63	64	55	60	57	72	67	71	75	60	72	68	68	71	74	72	78	69	64	73	76	82	70	81	75	70	70	74
4	65	58	53	61	59	59	66	67	72	69	78	67	67	76	72	75	78	69	73	76	69	65	85	68	75	86	87	93	74	83
5	61	64	61	69	74	75	69	65	64	72	67	69	70	78	71	72	70	70	71	71	74	67	82	81	65	79	84	82	76	80
6	53	71	66	60	70	73	70	76	73	74	83	79	82	80	81	83	77	72	82	73	78	66	81	86	79	83	92	81	94	95
7	57	60	52	59	62	62	73	64	69	62	68	63	70	69	66	64	64	63	67	65	65	64	72	65	63	64	74	71	70	66
8	55	61	61	57	58	65	69	61	60	63	63	60	61	60	60	62	65	57	65	67	71	66	62	70	65	64	74	69	68	72
9	59	61	65	59	64	63	65	67	74	63	67	71	65	64	62	63	64	67	67	69	72	68	67	76	75	72	69	77	66	74
10	60	58	59	58	56	52	61	62	69	62	67	67	62	69	55	58	58	67	61	72	70	65	65	72	69	68	67	70	64	71
11	64	59	59	57	66	57	58	60	58	72	63	61	62	62	67	62	60	64	66	72	69	64	65	70	72	74	68	70	64	70
12	62	63	59	59	63	58	62	56	57	72	59	69	63	64	67	68	65	67	69	73	71	66	69	76	74	74	70	71	66	72
13	68	60	50	55	67	55	56	58	59	74	64	67	62	59	65	64	55	69	66	73	71	66	67	68	76	69	64	71	73	75
14	63	63	61	66	62	64	64	56	61	72	65	69	63	67	68	67	57	71	69	73	72	68	70	75	75	72	66	71	74	78
15	59	59	53	55	56	53	55	59	62	61	59	67	63	63	66	60	57	68	64	64	64	57	60	64	65	65	57	65	59	63
16	66	55	53	58	54	56	61	59	58	70	70	69	64	59	61	59	62	69	65	69	68	58	59	65	65	67	59	66	58	63
17	70	57	60	64	61	60	61	61	60	66	75	70	65	63	67	59	60	71	64	70	71	60	61	67	66	67	63	70	61	67
18	67	68	54	61	61	67	69	69	72	66	69	69	70	69	68	63	67	63	72	74	71	62	63	69	67	69	65	71	63	69
19	68	70	64	65	61	65	69	72	72	67	72	66	73	72	69	67	65	64	74	75	73	64	64	70	69	71	68	72	65	70
20	69	65	60	71	66	66	66	63	73	69	73	70	78	73	68	68	66	65	75	76	74	67	69	73	77	74	68	74	66	74
21	66	69	62	80	71	70	71	68	79	77	81	71	82	77	75	72	73	72	82	81	80	73	73	74	80	76	69	77	68	76
22	70	71	71	80	77	77	79	77	81	74	83	78	85	80	80	74	77	74	83	81	81	75	78	82	84	82	73	80	77	81
23	77	78	72	82	78	82	86	80	86	79	84	79	83	80	86	80	79	77	86	85	83	78	81	87	84	84	78	87	81	84
24	89	84	88	86	87	94	89	94	89	94	91	85	92	89	93	85	87	84	92	94	91	85	86	91	93	91	83	92	88	92
25	100	94	90	97	96	96	97	99	98	96	98	93	99	98	100	93	93	92	101	103	96	92	94	99	99	99	93	97	95	96
26	104	101	95	107	105	102	105	108	113	108	107	102	111	105	105	104	107	102	110	114	113	104	108	111	111	110	102	104	100	104
27	126	120	118	130	122	123	124	126	131	121	126	122	125	124	125	125	124	126	134	133	131	125	127	133	133	129	127	127	122	125
28	140	135	132	143	136	139	145	146	145	136	142	138	143	143	141	140	140	139	148	147	146	140	142	146	147	147	141	143	141	143
29	154	145	145	153	143	152	154	155	161	152	158	147	153	152	152	152	154	153	159	160	160	151	153	157	156	156	151	157	152	156
30	167	151	161	166	158	162	164	162	166	160	160	156	161	156	161	155	163	159	164	165	165	157	162	156	163	165	161	160	159	162

[25] Neste processamento utilizou-se coeficientes LPC e limiares de energia de início e fim iguais a 0,001002.

Tabela A.2: Resultados do Processamento 2^{26}.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	55	52	58	55	52	49	57	58	56	53	55	57	57	61	60	57	56	64	60	66	62	62	62	59	59	57	60	69	69	58
2	42	54	54	57	56	60	61	49	57	58	51	54	53	55	57	60	65	64	64	59	42	55	58	54	51	49	56	57	55	59
3	50	58	58	55	52	64	59	53	63	60	67	51	52	51	59	64	55	56	51	55	51	48	53	49	54	55	55	55	55	58
4	50	66	59	57	52	60	59	53	55	56	61	53	57	60	51	58	57	48	52	55	52	51	55	57	59	56	61	70	64	72
5	58	54	54	50	52	55	52	53	52	49	51	46	53	51	44	52	50	59	48	50	55	50	48	55	54	53	55	68	64	56
6	54	54	58	46	58	48	56	53	52	52	56	54	48	57	56	49	57	58	54	56	58	58	52	59	62	59	57	62	63	64
7	50	56	58	49	57	61	54	59	57	53	64	56	55	66	56	68	59	61	62	61	64	60	57	58	63	63	59	64	61	67
8	53	56	53	45	55	57	53	47	52	51	60	56	55	52	58	58	49	52	52	55	47	54	55	53	48	54	57	55	59	55
9	55	63	56	56	53	57	54	60	56	58	53	59	53	59	55	56	56	57	58	58	57	56	54	56	59	59	59	60	58	53
10	52	56	59	50	62	63	55	55	56	54	55	53	57	56	63	55	57	53	56	60	57	55	50	55	59	55	54	50	55	55
11	54	48	50	57	51	54	55	58	48	54	54	58	57	52	60	59	59	48	60	54	52	52	55	64	51	58	48	50	56	57
12	47	53	53	58	52	54	47	56	58	61	57	56	56	54	53	60	54	47	49	57	62	61	59	59	51	60	63	61	59	57
13	54	51	53	60	55	54	61	59	61	54	57	60	58	58	58	59	54	57	60	54	55	50	58	56	53	59	55	54	58	61
14	64	59	59	56	58	58	60	53	56	58	55	57	62	57	59	59	53	52	46	59	58	55	57	57	58	65	54	59	61	60
15	57	58	56	51	64	56	57	56	59	57	58	52	60	58	57	58	58	60	59	52	61	58	55	55	53	53	57	53	58	58
16	59	57	62	60	65	57	61	65	60	64	56	64	68	65	61	59	60	57	59	64	62	51	56	54	63	63	59	67	66	65
17	69	66	63	68	63	67	60	67	68	69	69	71	72	65	64	59	59	52	56	56	60	60	65	62	62	64	66	62	61	66
18	63	69	63	69	73	67	68	66	62	73	69	72	74	67	64	73	60	66	63	63	64	58	63	66	66	65	64	60	62	66
19	65	60	55	65	62	62	63	62	71	71	62	70	63	68	62	70	66	60	66	63	64	66	64	65	67	63	64	65	63	65
20	64	67	64	66	68	70	73	69	69	72	67	65	72	70	71	70	61	61	62	62	64	55	67	61	62	63	62	63	61	63
21	69	69	66	67	67	70	75	70	76	67	67	75	68	77	73	71	70	68	68	70	69	71	71	73	75	70	73	73	73	74
22	70	71	79	74	70	78	72	79	80	73	77	77	79	80	75	82	76	82	81	80	76	75	82	73	74	74	72	72	71	73
23	82	84	81	76	86	84	74	85	83	80	83	84	78	72	76	76	75	76	74	73	75	72	73	80	81	84	81	81	80	80
24	84	85	86	82	91	85	87	92	90	100	94	99	91	85	79	84	85	80	79	79	82	80	83	84	81	89	87	87	87	87
25	94	91	97	89	94	96	96	98	102	95	100	102	97	92	91	94	90	94	94	92	93	89	87	88	91	95	94	96	94	97
26	98	100	101	101	100	106	109	114	113	113	106	107	103	101	100	98	100	101	98	96	101	98	98	99	98	96	100	98	99	100
27	125	120	118	115	120	123	129	129	130	136	128	126	125	122	116	119	117	122	118	117	119	118	116	116	122	122	124	128	127	125
28	136	137	133	140	138	133	137	140	144	144	136	137	135	133	133	130	133	135	134	131	133	133	133	128	127	131	134	135	133	137
29	147	152	144	150	149	156	155	153	153	150	144	143	147	146	144	142	144	144	143	144	145	143	146	147	145	149	146	149	151	150
30	158	155	155	156	166	163	164	162	162	163	153	151	152	156	151	150	152	152	155	151	155	156	152	156	155	158	158	157	158	156

[26] Neste processamento utilizou-se coeficientes LPC e limiares de energia de início e fim iguais a 0,001002 e 0,007760, respectivamente.

Tabela A.3: Resultados do Processamento 3^{27}.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	56	57	56	57	55	62	62	62	75	67	73	76	68	71	76	69	68	70	75	78	68	73	76	79	75	71	75	75	66	76
2	51	53	57	61	57	61	60	60	57	66	71	73	62	69	66	64	64	61	67	75	65	64	70	68	68	69	69	67	73	73
3	53	53	53	54	52	63	59	60	59	62	64	66	68	72	64	66	62	63	64	60	67	62	69	71	71	69	67	68	73	72
4	52	49	55	54	54	59	61	60	62	69	64	68	70	75	66	61	56	65	64	61	68	65	73	67	71	70	70	70	71	73
5	53	49	50	52	48	59	61	62	64	66	63	68	68	74	66	68	66	65	64	60	63	60	63	66	70	72	70	70	71	73
6	45	49	52	51	57	59	63	64	63	67	65	71	73	76	71	75	58	68	67	61	63	64	74	73	71	77	71	71	75	73
7	49	50	52	59	60	55	62	65	62	68	69	72	77	81	71	75	63	65	71	66	69	63	75	68	69	76	73	73	76	73
8	50	45	53	54	61	61	59	69	63	67	71	70	74	80	68	74	67	67	69	66	67	69	78	72	75	76	74	73	75	73
9	53	49	54	61	64	57	63	64	69	74	67	73	80	83	71	79	67	70	71	68	70	69	78	74	78	79	76	75	78	77
10	54	48	53	55	63	60	64	68	71	76	67	73	80	85	72	80	68	70	71	68	70	69	78	75	79	80	78	76	80	78
11	56	52	58	56	59	58	68	67	70	72	68	75	81	85	78	80	76	73	73	68	73	69	77	75	78	77	77	75	82	79
12	54	47	63	60	63	66	67	72	67	75	72	75	77	87	76	79	73	72	75	68	75	70	79	76	79	79	78	77	82	80
13	55	47	57	54	63	59	63	71	69	75	74	75	76	84	77	80	72	72	75	67	75	71	79	77	80	79	77	77	83	79
14	55	53	56	53	63	57	67	70	67	72	65	74	74	88	76	79	70	72	76	68	77	70	80	78	80	78	79	78	82	80
15	66	51	58	59	61	64	67	69	70	69	69	68	71	78	69	71	67	73	71	66	73	69	77	74	80	76	77	75	82	78
16	73	57	64	71	70	68	79	77	77	78	77	76	79	89	76	83	78	79	79	73	79	76	84	79	84	84	83	84	89	82
17	75	60	66	73	71	72	80	77	83	84	77	79	83	92	77	81	76	81	80	79	81	80	88	82	87	85	85	83	89	85
18	92	80	84	90	86	86	94	93	96	99	97	96	97	104	97	98	93	98	97	92	98	96	102	98	102	101	100	98	105	101
19	107	92	97	106	101	97	109	112	114	112	110	110	115	120	110	114	110	114	109	108	113	110	116	116	119	116	116	113	120	114
20	123	112	122	118	121	116	130	128	131	131	131	129	132	139	130	134	128	132	132	128	136	129	138	132	140	138	137	135	140	136
21	148	137	141	141	144	138	152	157	150	155	157	156	158	169	154	159	157	160	159	154	159	154	161	154	161	157	159	159	166	161
22	157	154	150	157	157	154	165	172	165	171	171	172	170	178	166	171	167	169	173	169	172	166	173	172	176	173	173	171	177	172
23	176	168	166	173	178	172	180	187	181	183	182	182	186	192	182	185	186	183	182	179	183	181	196	188	194	191	186	185	190	188
24	190	181	180	184	189	188	198	199	197	202	204	204	203	213	203	211	210	212	210	211	205	203	212	207	214	213	210	208	209	207
25	221	206	210	213	217	212	226	227	222	227	225	232	235	241	236	236	231	235	234	233	232	228	236	231	234	235	233	231	237	236
26	244	237	242	241	247	240	247	254	253	256	253	257	263	267	261	261	258	261	263	257	261	255	261	258	262	265	264	261	263	262
27	278	266	272	278	282	272	287	289	283	288	286	291	296	301	292	293	289	299	296	295	292	289	298	293	298	298	298	295	299	297
28	309	294	299	304	306	304	313	314	312	316	313	316	320	325	317	321	319	319	321	318	322	318	323	316	323	322	321	319	325	326
29	326	318	320	326	325	322	323	329	326	332	328	329	332	340	328	328	325	326	331	329	330	329	335	334	330	332	333	338	342	340
30	344	336	339	339	341	342	346	356	350	359	349	345	346	356	348	352	348	346	351	350	348	349	356	354	350	357	354	355	363	363

[27] Neste processamento utilizou-se coeficientes LPC e limiares de energia de início e fim iguais a 0,007760 e 0,001002, respectivamente.

Tabela A.4: Resultados do Processamento 4[28].

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	66	59	63	63	63	62	63	54	57	58	53	51	54	60	57	56	55	56	55	57	58	56	54	57	60	62	62	62	58	60
2	66	68	73	63	64	59	58	63	58	62	56	55	57	58	55	57	54	59	59	59	58	60	57	58	55	57	57	57	53	57
3	73	73	72	71	66	70	65	64	64	60	60	55	52	55	61	59	57	60	60	61	59	60	61	65	66	64	64	64	64	63
4	68	75	68	73	62	71	64	64	67	65	64	60	50	59	60	57	63	61	59	62	62	58	61	65	67	69	69	69	69	68
5	69	72	67	76	77	70	67	63	65	64	63	62	55	62	62	59	61	58	59	58	59	61	62	65	69	64	64	64	63	63
6	74	74	75	74	74	74	71	72	65	66	64	63	61	64	61	58	60	64	62	65	62	62	61	67	64	65	65	65	66	66
7	70	75	76	82	72	72	74	67	70	66	67	66	61	64	64	62	64	68	68	66	65	64	65	69	67	69	69	69	69	69
8	73	68	75	81	76	77	73	71	70	75	68	68	59	67	64	62	63	69	71	71	69	69	66	71	70	70	70	70	71	71
9	79	71	74	76	77	75	74	68	71	69	69	65	61	61	70	67	69	66	66	65	65	67	64	70	70	69	69	69	71	71
10	77	75	79	79	78	79	73	71	73	69	74	66	62	64	64	65	66	69	68	66	68	63	65	69	70	73	73	73	72	72
11	75	69	78	80	71	72	70	69	71	74	67	72	63	66	66	62	64	69	67	66	69	66	65	68	70	68	68	68	71	71
12	79	72	81	84	66	72	72	76	73	71	62	65	64	63	65	61	62	63	68	67	66	67	66	69	69	74	74	74	73	73
13	72	67	80	75	73	73	75	71	73	66	63	65	67	65	66	65	66	66	73	66	68	63	62	66	67	68	68	68	70	71
14	71	70	71	65	63	69	70	68	68	64	64	64	67	69	63	63	67	67	61	64	61	69	69	72	70	71	71	71	75	75
15	77	68	71	68	71	74	72	75	77	67	70	67	62	63	62	58	67	67	65	66	65	63	64	65	67	68	68	68	68	68
16	84	68	83	80	83	85	84	81	82	85	79	75	64	67	65	68	71	71	73	72	70	68	67	69	73	72	72	72	73	73
17	83	77	76	81	85	79	85	77	79	80	82	77	72	74	73	68	72	76	75	75	74	74	74	71	72	72	72	72	73	73
18	95	93	93	96	98	95	94	92	93	95	102	93	93	94	90	86	90	88	91	92	90	87	87	91	93	95	95	95	96	96
19	114	107	107	106	112	107	108	106	114	110	111	109	106	105	104	100	106	104	105	106	106	102	102	104	106	105	105	105	106	106
20	135	128	119	125	129	122	127	132	132	129	125	126	125	126	121	114	123	122	121	121	124	119	120	124	126	126	126	126	128	128
21	160	153	154	150	149	147	146	149	149	144	145	142	141	148	145	141	148	145	144	143	143	143	145	147	147	146	146	146	150	150
22	167	163	167	163	166	161	166	166	166	162	158	156	157	161	158	157	158	156	157	157	153	158	159	163	162	158	158	158	164	164
23	180	177	183	175	185	176	179	180	177	178	177	176	179	178	175	177	180	179	179	179	179	185	185	188	187	184	184	184	188	188
24	195	191	198	194	194	194	197	190	192	192	193	190	193	195	195	197	192	191	191	189	189	193	193	194	195	193	193	193	195	195
25	216	214	219	217	216	218	219	217	217	218	218	216	219	224	223	220	213	216	212	215	212	212	212	213	214	210	210	210	214	214
26	244	240	247	242	244	245	242	239	243	243	243	243	243	250	250	250	245	242	244	243	241	246	246	249	250	243	243	243	248	248
27	279	275	281	278	281	276	279	275	281	280	277	278	275	282	281	280	277	273	273	275	274	280	281	282	279	277	277	277	284	284
28	300	299	301	299	303	301	305	299	301	303	302	306	300	307	307	305	310	309	305	308	308	311	311	312	312	307	307	307	313	313
29	327	322	326	329	326	325	327	326	326	328	329	334	330	333	329	330	324	328	328	331	329	329	329	332	332	332	332	332	329	329
30	346	345	343	343	348	346	348	347	346	348	344	350	346	348	347	346	348	350	348	348	349	350	349	351	353	353	353	353	350	350

[28] Neste processamento utilizou-se coeficientes LPC e limiares de energia de início e fim iguais a 0,007760.

Tabela A.5: Resultados do Processamento 5[29].

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	14	11	14	13	10	19	18	17	21	22	20	21	20	21	15	18	14	16	16	17	15	19	19	18	19	17	13	18	20	18
2	12	12	12	15	15	16	14	14	11	15	14	11	12	13	19	20	13	12	20	17	20	23	22	26	21	21	22	22	20	24
3	9	11	13	7	13	11	13	17	12	15	13	13	13	13	13	14	14	14	12	12	16	12	13	16	15	15	13	16	16	14
4	15	10	10	13	13	14	11	11	10	13	15	10	16	13	14	14	13	14	15	12	13	13	12	12	11	13	17	12	14	17
5	9	17	13	10	15	14	11	8	10	12	12	9	13	15	13	11	16	14	12	13	13	10	10	12	10	13	12	14	12	12
6	9	13	12	8	10	15	9	13	9	10	14	11	15	12	11	11	11	11	12	11	14	12	10	12	14	11	16	16	16	11
7	7	11	13	6	11	10	9	10	8	13	10	9	10	8	9	10	13	12	11	12	14	11	12	13	13	12	13	13	12	12
8	8	10	10	9	11	8	9	10	9	11	13	11	14	9	9	11	13	13	14	12	13	10	11	12	13	12	12	13	12	13
9	8	10	9	9	9	10	10	9	10	9	12	12	14	9	9	11	12	13	13	12	13	11	11	12	13	11	12	13	15	14
10	7	11	9	9	8	11	9	7	8	9	11	10	14	9	8	11	12	12	13	12	12	12	12	13	13	12	13	13	16	15
11	7	9	11	10	9	10	9	10	10	9	10	9	15	10	9	11	12	12	13	12	12	12	12	13	13	12	13	14	15	14
12	5	7	8	10	10	11	9	8	8	9	10	8	12	10	10	11	13	12	13	12	12	12	12	13	13	12	13	13	15	14
13	7	7	8	9	8	10	9	9	11	9	11	11	11	9	9	11	12	12	13	12	12	12	13	11	13	13	13	14	15	13
14	9	9	8	9	10	8	7	8	12	10	11	13	13	7	10	12	12	12	14	12	12	11	13	11	13	12	12	14	15	13
15	7	9	6	11	10	11	9	9	10	10	9	9	10	10	10	13	11	12	12	12	11	11	12	11	11	11	11	11	13	13
16	7	11	12	11	9	12	12	11	11	10	10	10	11	11	11	13	12	13	13	13	12	12	13	12	12	12	12	12	13	13
17	10	10	11	12	10	13	14	11	11	10	10	10	10	12	11	13	13	13	13	13	12	13	13	12	12	12	12	12	13	13
18	14	9	13	12	11	9	10	13	12	13	11	12	11	13	12	14	14	14	15	14	14	14	14	13	13	13	13	13	14	14
19	13	9	14	14	12	13	12	14	14	15	13	13	13	14	13	15	15	14	15	15	15	14	15	14	14	14	14	14	15	15
20	17	11	16	16	14	16	17	19	17	18	15	17	16	17	15	17	19	19	18	17	18	17	16	17	16	16	16	16	17	17
21	17	15	17	17	16	17	16	20	15	17	15	16	16	15	15	18	19	17	18	17	19	17	17	18	17	17	17	17	18	18
22	21	21	24	23	23	23	19	22	21	21	21	22	21	21	21	23	23	23	22	22	24	22	22	22	22	22	22	22	23	23
23	27	25	26	28	26	27	26	29	27	26	26	26	25	26	27	28	28	27	28	27	28	27	28	28	27	27	27	27	28	28
24	32	32	35	35	31	30	32	37	34	34	33	34	34	33	32	35	35	34	35	35	36	35	35	35	34	34	34	34	35	35
25	40	37	40	42	40	37	37	42	40	40	40	41	40	41	39	42	41	41	41	41	41	40	40	41	42	42	40	40	42	42
26	50	50	50	51	49	49	48	51	50	51	49	49	49	50	49	50	51	50	50	50	51	51	50	50	50	50	50	50	51	51
27	68	65	68	70	67	68	67	70	70	70	67	68	69	70	67	70	70	68	67	67	68	68	68	68	68	68	68	68	69	69
28	82	79	82	84	81	82	80	84	83	83	81	82	82	83	82	84	84	83	83	84	85	83	83	84	84	83	83	83	84	84
29	92	91	95	96	95	94	95	96	96	97	96	96	97	97	96	97	98	97	96	96	97	97	98	97	97	97	96	96	98	98
30	100	99	102	104	102	102	103	105	104	103	102	104	104	105	102	105	104	103	103	104	104	103	104	103	103	102	102	103	104	104

[29] Neste processamento utilizou-se coeficientes cepstrais e limiares de energia de início e fim iguais a 0,001002.

Tabela A.6: Resultados do Processamento 6[30].

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	8	8	8	7	9	11	14	8	11	11	8	12	9	15	14	9	9	10	11	12	9	13	12	11	11	9	9	17	11	11
2	8	10	6	6	8	13	13	11	12	10	9	12	11	10	13	8	12	14	14	11	8	12	11	7	9	9	13	11	11	8
3	11	11	13	7	13	12	10	12	12	10	12	14	10	10	10	14	7	11	7	12	10	10	10	11	11	10	9	15	12	12
4	10	12	8	12	10	12	13	11	12	8	11	12	13	9	9	13	11	11	11	12	12	6	10	9	9	10	13	15	15	12
5	11	9	11	10	12	12	12	11	12	12	7	13	12	10	9	10	9	9	8	8	5	9	10	11	10	9	8	12	14	10
6	8	12	14	10	13	9	11	10	14	9	14	10	10	7	10	8	8	9	11	8	7	7	8	4	13	13	12	12	9	10
7	11	11	9	7	9	11	10	10	9	11	6	8	11	9	8	6	6	6	6	6	7	6	7	7	8	7	6	8	9	8
8	10	8	9	8	9	11	7	8	9	7	8	11	12	6	6	10	9	6	6	6	10	6	11	7	8	7	7	7	8	6
9	7	7	9	7	8	9	8	9	9	11	10	8	12	9	9	8	6	12	11	8	10	10	8	8	11	8	11	6	12	9
10	8	8	6	7	8	9	15	10	12	10	12	11	13	7	9	11	6	12	12	9	6	7	10	7	12	10	11	11	8	8
11	8	10	7	9	9	9	7	9	8	10	8	11	9	5	6	10	9	6	7	11	11	10	11	7	6	11	11	11	7	13
12	7	8	8	8	7	9	9	9	11	8	7	7	9	11	14	12	5	9	8	10	10	7	8	12	8	11	10	9	11	8
13	7	7	8	6	10	8	9	7	11	8	11	9	10	10	8	9	10	10	7	10	7	11	9	10	10	9	10	11	7	10
14	6	7	7	6	6	6	8	12	7	10	10	9	8	11	10	7	8	10	7	7	7	7	10	7	7	7	10	8	10	6
15	6	7	7	8	7	11	10	13	11	10	10	10	10	11	11	11	12	12	12	9	9	10	10	11	14	9	7	8	9	9
16	9	7	9	7	8	9	13	9	9	13	11	16	13	11	8	9	11	12	8	11	10	9	11	13	8	8	7	10	8	8
17	8	6	9	8	9	7	10	10	13	17	10	16	14	14	12	10	10	11	9	13	9	8	8	9	9	9	7	8	8	9
18	10	9	8	7	8	13	11	9	10	12	13	15	14	12	13	11	13	11	11	11	13	9	9	12	13	13	12	16	11	16
19	13	12	10	10	11	15	14	14	12	13	14	14	14	10	10	11	12	11	12	13	14	14	15	12	10	12	11	10	10	12
20	14	13	12	13	12	12	16	12	14	17	13	17	16	16	16	14	14	17	14	15	17	16	17	15	14	15	15	14	15	16
21	14	12	13	13	12	13	16	17	14	16	19	17	13	15	15	15	15	14	13	16	13	15	14	15	15	15	14	15	17	16
22	19	19	21	18	21	18	19	23	24	21	21	24	22	20	22	24	25	24	22	24	24	23	24	22	21	21	20	21	20	22
23	24	24	24	24	26	26	24	27	24	30	29	27	32	26	26	27	27	26	25	27	25	26	26	25	25	25	25	25	27	28
24	33	32	31	30	34	31	30	31	36	32	31	36	33	31	32	33	33	32	31	31	31	31	32	32	32	34	34	33	33	34
25	39	37	37	39	38	37	38	36	41	40	39	40	41	39	39	39	40	43	42	41	42	37	39	41	43	44	44	44	46	46
26	50	48	47	50	48	48	50	51	47	48	52	50	48	49	49	49	48	47	46	47	48	48	48	50	52	53	53	53	54	54
27	64	67	65	65	62	71	69	69	67	69	66	67	65	66	65	65	66	65	66	63	65	64	64	69	70	72	69	69	70	71
28	78	81	78	79	79	84	86	80	84	84	80	81	82	82	80	81	81	81	80	81	81	81	81	82	82	82	82	81	85	85
29	91	95	93	91	93	97	97	93	95	93	92	92	93	94	94	93	92	92	93	93	93	93	92	92	92	92	92	92	94	94
30	98	102	101	102	100	101	103	102	98	104	101	99	98	101	101	100	100	100	101	101	100	100	101	101	98	99	98	98	100	100

[30] Neste processamento utilizou-se coeficientes cepstrais e limiares de energia de início e fim iguais a 0,001002 e 0,007760, respectivamente.

Tabela A.7: Resultados do Processamento 7[31].

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	7	11	10	11	12	11	13	13	9	12	12	14	12	10	11	12	10	12	12	15	9	14	18	12	15	14	13	11	12	10
2	10	11	10	7	10	13	12	12	11	13	11	10	12	13	11	13	11	14	12	11	10	13	12	13	13	12	12	11	13	13
3	8	11	7	10	12	12	12	11	12	12	11	8	11	12	12	14	10	13	10	12	11	11	10	13	13	14	15	12	13	13
4	9	10	7	11	11	13	11	12	13	12	11	10	11	13	10	14	9	12	9	10	11	11	10	14	12	14	13	14	13	13
5	7	8	9	9	10	11	13	12	10	11	11	10	9	10	10	11	9	9	9	9	10	11	12	14	12	14	13	14	12	13
6	8	7	8	6	10	11	11	12	10	11	12	10	9	10	10	13	10	11	10	10	12	9	12	12	14	13	14	13	13	13
7	8	7	9	9	11	10	11	13	11	12	11	12	10	11	11	13	11	12	11	9	9	11	12	12	12	13	13	14	15	14
8	8	10	9	8	10	12	10	11	9	10	11	8	12	11	10	10	11	14	10	10	10	13	11	12	14	14	14	14	14	13
9	9	8	9	8	12	8	10	13	11	12	12	9	11	12	9	11	12	12	10	11	10	12	11	12	14	14	14	13	14	13
10	10	10	8	7	13	10	13	14	10	12	11	9	12	13	10	12	12	12	10	11	10	13	11	13	14	15	15	14	14	13
11	11	7	9	7	11	11	9	13	10	10	12	9	12	13	11	12	11	11	11	10	10	13	10	13	14	14	15	14	15	12
12	15	10	13	9	11	11	10	11	10	12	11	9	13	12	10	12	11	12	11	10	10	12	10	13	15	15	15	14	15	12
13	13	9	9	10	10	10	13	13	12	10	11	10	13	12	10	12	11	13	11	11	10	12	10	13	15	15	15	14	15	12
14	14	9	9	10	13	12	14	14	10	11	13	9	12	10	9	10	11	13	12	11	11	14	10	13	15	14	15	14	15	12
15	14	13	12	9	12	12	13	14	13	13	13	11	13	13	12	13	13	16	15	13	14	15	13	15	17	17	18	18	18	15
16	17	19	19	17	19	20	19	21	20	20	20	18	20	20	19	20	20	22	21	21	20	23	20	23	25	24	25	26	25	23
17	21	22	23	20	23	24	25	26	24	24	22	22	24	24	23	24	24	27	25	24	24	26	24	24	27	26	26	27	27	25
18	36	39	39	37	39	39	39	41	40	40	38	38	40	40	39	40	40	42	41	40	41	43	41	41	43	42	44	45	44	42
19	52	54	55	54	55	56	55	57	56	56	54	54	56	56	55	56	56	58	56	56	56	59	56	58	59	58	59	61	60	58
20	74	76	75	75	77	77	76	78	77	77	75	75	77	77	76	77	77	79	77	77	77	79	77	78	80	79	80	81	80	78
21	97	98	96	100	101	100	99	101	100	100	98	98	100	100	99	100	100	102	100	101	100	102	100	101	103	102	103	104	103	101
22	109	111	109	113	113	113	112	114	114	113	111	112	112	113	112	114	113	113	113	114	115	115	116	117	115	115	116	117	116	114
23	124	124	124	129	127	128	126	131	130	127	128	126	128	127	128	128	128	129	127	130	131	130	132	132	130	130	132	132	131	129
24	141	144	141	144	144	144	144	147	146	145	145	145	146	146	145	146	147	147	146	148	144	146	147	148	148	147	149	149	150	147
25	170	169	167	170	172	173	171	172	173	174	170	170	172	171	172	175	176	176	175	176	175	177	177	177	175	175	177	178	174	174
26	199	201	198	199	203	201	201	201	201	201	201	201	203	202	201	202	201	202	201	202	204	203	201	203	204	204	205	207	206	203
27	232	234	233	237	238	238	238	239	240	239	236	235	239	237	237	237	237	238	238	239	240	238	236	236	237	240	240	240	239	238
28	260	260	261	263	265	264	264	264	265	265	264	263	263	265	266	267	265	265	265	264	263	264	263	265	266	264	268	268	265	264
29	286	288	290	287	288	290	289	290	289	288	289	286	289	288	289	290	289	290	288	288	288	290	288	290	291	290	291	293	291	289
30	310	310	309	308	309	312	309	311	310	311	310	308	313	310	310	312	312	313	311	312	311	312	312	311	308	309	309	311	311	311

[31] Neste processamento utilizou-se coeficientes cepstrais e limiares de energia de início e fim iguais a 0,007760 e 0,001002, respectivamente.

Tabela A.8: Resultados do Processamento 8[32].

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	11	12	11	10	10	10	10	10	11	7	5	10	9	9	7	7	8	10	10	8	10	8	9	8	10	11	11	11	10	9
2	11	11	12	10	12	10	11	10	10	10	9	9	9	7	8	9	7	7	8	8	8	7	8	8	8	8	8	8	10	9
3	13	11	11	10	9	9	11	8	8	11	11	10	13	8	7	8	8	8	7	6	7	7	7	7	9	9	9	9	9	9
4	7	7	12	11	8	11	11	11	12	13	12	12	9	8	6	7	8	9	8	8	7	7	7	7	9	10	10	10	9	8
5	12	11	11	13	9	10	11	8	9	11	9	8	8	7	9	11	8	8	8	8	8	9	11	11	10	10	10	10	11	11
6	8	9	10	11	10	11	10	10	9	10	9	10	9	7	7	8	7	8	8	8	8	9	9	9	10	10	10	10	10	10
7	9	10	8	11	10	9	8	8	8	12	10	11	9	9	10	10	8	7	8	11	9	9	10	10	10	11	11	11	11	11
8	8	10	9	11	7	10	9	7	7	8	9	9	8	8	8	7	8	8	8	7	7	7	8	7	10	8	8	8	8	8
9	9	10	12	11	7	7	7	8	7	9	8	9	9	8	8	8	8	6	8	8	8	6	7	6	10	8	8	8	9	9
10	13	9	13	11	10	10	8	9	8	11	9	10	9	9	8	8	8	7	7	7	7	8	7	7	7	7	7	7	8	8
11	13	9	11	10	10	10	8	9	10	13	10	11	8	9	8	8	7	7	8	8	8	8	7	8	8	9	9	9	7	7
12	13	14	15	14	10	10	8	10	11	12	10	11	9	8	7	7	7	7	7	7	8	8	8	8	7	6	6	6	10	10
13	15	17	16	17	13	11	8	8	11	10	9	10	8	8	7	7	8	8	8	8	7	8	8	8	9	6	6	6	7	7
14	14	9	17	16	9	9	10	8	10	11	10	10	10	8	9	8	7	7	8	8	7	8	8	8	7	7	7	7	7	7
15	16	14	12	14	12	11	12	10	11	11	13	12	14	11	11	11	10	8	10	10	10	9	10	11	9	8	8	8	11	11
16	20	19	15	19	18	18	16	15	16	17	20	20	20	19	19	19	18	16	17	17	17	18	19	19	17	17	17	17	20	20
17	24	24	24	23	20	20	20	21	20	25	24	25	25	25	25	25	23	20	22	22	23	23	24	24	22	22	22	22	25	25
18	38	38	39	38	37	37	35	36	36	40	41	40	40	40	40	40	39	37	37	37	37	38	39	38	36	36	36	36	39	39
19	55	54	57	54	55	55	54	55	54	57	57	56	56	56	55	55	55	53	54	54	54	54	55	55	53	53	53	53	56	56
20	76	77	77	76	75	76	74	78	77	79	76	76	77	75	75	77	75	73	74	74	74	74	75	75	73	73	73	73	76	76
21	99	100	98	99	97	97	96	99	98	102	100	100	98	99	98	98	99	96	97	97	97	97	97	97	96	96	96	96	97	98
22	111	111	111	109	109	108	108	108	109	114	112	113	111	111	111	111	111	110	110	110	110	110	110	110	109	108	108	108	111	112
23	126	126	126	125	125	124	123	124	125	128	128	126	127	126	126	126	127	126	126	126	126	126	126	126	126	126	126	126	126	127
24	141	142	141	141	140	139	142	142	143	146	145	145	145	145	146	146	144	142	142	144	144	144	144	144	143	143	143	143	144	144
25	171	170	169	167	168	167	169	170	171	173	171	171	173	171	172	172	172	169	169	169	169	168	168	168	168	168	168	168	169	169
26	201	199	199	199	199	199	198	200	201	200	202	200	201	201	201	201	201	199	199	199	199	199	199	199	199	199	199	199	201	202
27	234	234	234	235	233	233	234	234	235	236	236	235	235	238	237	237	237	234	233	234	233	233	233	233	233	233	233	233	235	235
28	263	262	262	261	260	260	260	261	261	264	262	262	262	263	263	263	263	261	261	261	261	261	261	261	261	261	261	261	262	262
29	286	285	285	285	286	284	285	285	285	287	287	288	288	288	286	286	286	287	287	286	286	286	286	286	287	287	287	287	288	288
30	309	310	309	308	307	308	308	309	310	309	309	310	311	311	310	309	310	309	309	310	310	310	310	310	308	307	307	307	309	309

[32] Neste processamento utilizou-se coeficientes cepstrais e limiares de energia de início e fim iguais a 0,007760.

Apêndice B Matrizes de Confusão

Neste apêndice, são apresentadas as matrizes de confusão para cada uma das configurações apresentadas na Tabela 4.6.

Tabela B.1: Matriz de confusão para a configuração RD-II e LD de 0,01493.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	71	0	0	0	0	0	0	39
help	0	84	0	0	0	0	0	26
no	0	0	92	0	0	0	0	18
repeat	0	0	0	101	0	0	0	9
stop	0	0	0	0	94	0	0	16
start	0	0	0	0	0	69	0	41
yes	0	0	0	0	0	0	85	25

Tabela B.2: Matriz de confusão para a configuração RD-II e LD de 0,01693.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	89	0	0	0	0	0	0	21
help	0	101	0	0	0	0	0	9
no	0	0	98	0	0	0	0	12
repeat	0	0	0	109	0	0	0	1
stop	0	0	0	0	104	1	0	5
start	0	0	0	0	0	89	0	21
yes	0	0	0	0	0	0	100	10

Tabela B.3: Matriz de confusão para a configuração RD-II e LD de 0,01751.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	93	0	0	0	0	0	0	17
help	0	104	0	0	0	0	0	6
no	0	1	99	0	0	0	0	10
repeat	0	0	0	110	0	0	0	0
stop	0	0	0	0	104	1	0	5
start	0	0	0	0	0	94	0	16
yes	0	0	0	0	0	0	103	7

Tabela B.4: Matriz de confusão para a configuração RD-II e LD de 0,01812.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	97	1	0	0	0	0	0	12
help	0	107	0	0	0	0	0	3
no	0	1	100	0	0	0	0	9
repeat	0	0	0	110	0	0	0	0
stop	0	0	0	0	104	1	0	5
start	0	0	0	0	0	95	0	15
yes	0	0	0	0	0	0	105	5

Tabela B.5: Matriz de confusão para a configuração RD-II e LD de 0,02855.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	108	2	0	0	0	0	0	0
help	0	110	0	0	0	0	0	0
no	0	1	109	0	0	0	0	0
repeat	0	0	0	110	0	0	0	0
stop	0	0	0	0	109	1	0	0
start	0	0	0	0	0	110	0	0
yes	0	0	0	0	0	0	110	0

Tabela B.6: Matriz de confusão para a configuração RD-III e LD de 0,00189.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	94	0	0	0	0	0	0	16
help	0	110	0	0	0	0	0	0
no	0	0	105	0	0	0	0	5
repeat	0	0	0	110	0	0	0	0
stop	0	0	0	0	106	0	0	4
start	0	0	0	0	0	108	0	2
yes	0	0	0	0	0	0	109	1

Tabela B.7: Matriz de confusão para a configuração RD-III e LD de 0,00188.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	94	1	0	0	0	0	0	15
help	0	110	0	0	0	0	0	0
no	0	0	105	0	0	0	0	5
repeat	0	0	0	110	0	0	0	0
stop	0	0	0	0	106	0	0	4
start	0	0	0	0	0	108	0	2
yes	0	0	0	0	0	0	109	1

Tabela B.8: Matriz de confusão para a configuração RD-III e LD de 0,00129.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	97	1	0	0	0	0	0	12
help	0	110	0	0	0	0	0	0
no	0	0	107	0	0	0	0	3
repeat	0	0	0	110	0	0	0	0
stop	0	0	0	0	107	1	0	2
start	0	0	0	0	0	109	0	1
yes	0	0	0	0	0	0	109	1

Tabela B.9: Matriz de confusão para a configuração RD-III e LD de 0,00117.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	97	1	0	0	0	0	0	12
help	0	110	0	0	0	0	0	0
no	0	1	108	0	0	0	0	1
repeat	0	0	0	110	0	0	0	0
stop	0	0	0	0	107	1	0	2
start	0	0	0	0	0	110	0	0
yes	0	0	0	0	0	0	109	1

Tabela B.10: Matriz de confusão para a configuração RD-III e LD de 0,00001.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	108	2	0	0	0	0	0	0
help	0	110	0	0	0	0	0	0
no	0	1	109	0	0	0	0	0
repeat	0	0	0	110	0	0	0	0
stop	0	0	0	0	109	1	0	0
start	0	0	0	0	0	110	0	0
yes	0	0	0	0	0	0	110	0

Tabela B.11: Matriz de confusão para a configuração RD-IV e LD de 0,01693.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	67	0	0	0	0	0	0	43
help	0	99	0	0	0	0	0	11
no	0	0	69	0	0	0	0	41
repeat	0	0	0	102	0	0	0	8
stop	0	0	0	0	77	0	0	33
start	0	0	0	0	0	85	0	25
yes	0	0	0	0	0	0	100	10

Tabela B.12: Matriz de confusão para a configuração RD-IV e LD de 0,01712.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	68	0	0	0	0	0	0	42
help	0	100	0	0	0	0	0	10
no	0	1	68	0	0	0	0	41
repeat	0	0	0	103	0	0	0	7
stop	0	0	0	0	75	0	0	35
start	0	0	0	0	0	87	0	23
yes	0	0	0	0	0	0	101	9

Tabela B.13: Matriz de confusão para a configuração RD-IV e LD de 0,01777.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	70	1	0	0	0	0	0	39
help	0	103	0	0	0	0	0	7
no	0	1	65	0	0	0	0	44
repeat	0	0	0	102	0	0	0	8
stop	0	0	0	0	73	0	0	37
start	0	0	0	0	0	89	0	21
yes	0	0	0	0	0	0	104	6

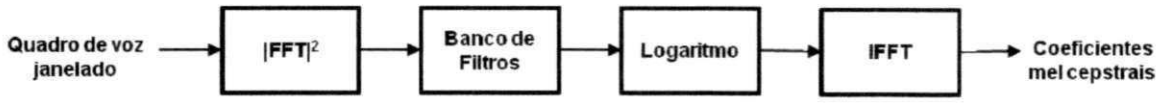
Tabela B.14: Matriz de confusão para a configuração RD-IV e LD de 0,01817.

	go	*help*	*no*	*repeat*	*stop*	*start*	*yes*	Desconhecido
go	69	2	0	0	0	0	0	39
help	0	103	0	0	0	0	0	7
no	0	1	64	0	0	0	0	45
repeat	0	0	0	101	0	0	0	9
stop	0	0	0	0	68	0	0	42
start	0	0	0	0	0	90	0	20
yes	0	0	0	0	0	0	102	8

Anexo A Coeficientes Mel Cepstrais

Os coeficientes mel cepstrais podem ser obtidos conforme ilustração apresentada na Figura A.1 (O'SHAUGHNESSY, 2000).

Figura A.1: Diagrama em blocos do cálculo dos coeficientes mel cepstrais.



Inicialmente, calcula-se o espectro da energia por meio da transforma de Fourier no quadro em análise. Após o cálculo da FFT, aplica-se ao sinal um banco de filtros triangulares na escala Mel[33], geralmente utiliza-se um conjunto de 20 filtros (O'SHAUGHNESSY, 2000). O objetivo de tais filtros são uma tentativa de aproximar a resposta do ouvido humano aos sinais sonoros (PETRY, ZANUZ e BARONE, 1999).

Depois da filtragem do sinal, calcula-se o logaritmo da energia na saída de cada um dos filtros e, por fim, calcula-se a transformada inversa de Fourier sobre estes valores, obtendo-se os coeficientes mel cepstrais (DIAS, 2000)[34]. O processo de obtenção desses coeficientes é definido matematicamente pela Equação A.1 (PETRY, ZANUZ e BARONE, 1999).

$$c(n) = \sum_{k=1}^{K} \log |S(k)| \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], 0 \leq n \leq P. \qquad \text{(A.1)}$$

Em que:

$c(n)$ – O n-ésimo coeficiente mel cepstral;

P – Número de coeficientes mel cepstrais extraídos;

---

[33] Mel é a unidade de medida de frequências ou picos percebidos de um tom (PETRY, ZANUZ e BARONE, 1999).

[34] DIAS, R. S. F. *Normalização De Locutor Em Sistema De Reconhecimento De Fala.* 113 f. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Estadual de Campinas, Campinas. 2000.

K – Número de filtros digitais;

$S(k)$ – Sinal de saída do banco de filtros digitais;

A diferença entre o cálculo dos coeficientes cepstrais, obtidos pelo método da FFT, e dos mel cepstrais reside na aplicação do banco de filtros digitais ao espectro real do sinal, antes da função logarítmica (PETRY, ZANUZ e BARONE, 1999).