UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA
Departamento de Ciências Exatas
Colegiado de Ciência da Computação

Um Modelo Intervalar para Reconhecimento de Fala por Computadores

Aspectos Teóricos e Estado da Arte

Diogo Pereira Silva de Novais

Vitória da Conquista – BA
Março de 2012

Um Modelo Intervalar para Reconhecimento de Fala por Computadores

Aspectos Teóricos e Estado da Arte

Diogo Pereira Silva de Novais

Orientador: Prof. Dr. Roque Mendes Prado Trindade

Trabalho de conclusão de curso apresentado ao Colegiado do Curso de Ciência da Computação do Departamento de Ciências Exatas – DCE-UESB como pré-requisito para obtenção do título de bacharel em Ciência da Computação

Áreas de concentração: Processamento de Sinais Digitais, Reconhecimento de Fala, Matemática Intervalar.

Vitória da Conquista – BA
Março de 2012

UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA
Departamento de Ciências Exatas
Colegiado de Ciência da Computação

DECLARAÇÃO DE APROVAÇÃO

Título: UM MODELO INTERVALAR PARA RECONHECIMENTO DE FALA POR COMPUTADORES: ASPECTOS TEÓRICOS E ESTADO DA ARTE

Autor: Diogo Pereira Silva de Novais

Aprovada como parte das exigências para obtenção do Título de Bacharel em Ciência da Computação, pela Banca Examinadora:

_____
Prof. Dr. Roque Mendes Prado Trindade – DCE/UESB
Orientador

_____
Prof. Dr. Benedito Melo Acióly – DCE/UESB
Parecerista

_____
Prof.ª Dr ª Vera Pacheco – DEL/UESB
Parecerista

_____
Prof.ª Esp. Crijina Chagas Flores – FAINOR
Parecerista

Data de realização: 26 de março de 2012.

Agradecimentos

Aos meus pais, Raimundo e Marlene, que participam de minha formação, orientado, apoiando e que, assim como todo o resto na vida, viveram a graduação lado a lado comigo.

Aos meus irmãos, Weldon, Valéria e Leandro que, junto aos meus pais, formam a base que me permite cada passo na vida. Que além disso, compartilharam conversas que renderam um aprendizado interessante, sobre outras ciências e sobre a ciência como um todo.

A Cátia Dias (e sua família, que hoje tenho como minha), mulher com quem tenho a honra de compartilhar a vida e, como parte disso, este trabalho. Com ela pude discutir, revisar e reorganizar o projeto por diversas vezes.

Aos meus colegas de turma (e Jorge, veterano com quem cursei algumas disciplinas), pessoas a quem atribuo maior parte da minha formação. Alunos exemplares que mostraram que os corredores, bibliotecas e laboratórios possibilitam a convivência com pessoas com as quais se aprende só por estar por perto.

Aos professores do curso de Ciência da Computação e demais cursos que mediaram a busca do conhecimento e forneceram fundamentos e fontes para desenvolver o conhecimento em diversas áreas.

A Celina, secretária do colegiado de extrema eficiência. Além disso, amiga, conselheira, ouviu muitos desabafos e encorajou muitas ações, uma espécie de “mãe” dos alunos do curso.

Aos amigos da AEESP, companheiros de ônibus por 5 anos, dividimos muita dificuldades de quem sai de Poções diariamente para realizar o sonho da graduação. Dividimos também boas feijoadas.

Ao meu Orientador e amigo (Roque Trindade), que me apresentou a ciência, possibilitou minha inserção neste mundo; que mais do que isso, encoraja, apoia e carrega com sua disposição e exemplo vários alunos do curso para carreiras de sucesso.

Resumo

O reconhecimento de fala tem sido foco de pesquisas nas últimas décadas, contribuindo para o desenvolvimento de novas interfaces de interação com computadores digitais, seja como meio para entrada de dados, suporte para melhor entendimento da fala e produção de fala sintética, ou como ferramenta para linguistas. No campo de processamento de sinais, a Análise Intervalar tem apresentado resultados importantes no tratamento de imprecisões. Este trabalho discute o processo de reconhecimento de fala, apresentando modelos matemáticos utilizados em tal atividade para, por fim, apresentar a teoria de Análise Intervalar, junto aos modelos intervalares já existentes análogos a modelos usados no reconhecimento de fala, concluindo com a proposta de um Modelo Intervalar para Reconhecimento Computacional de Fala.

Abstract

The speech recognition have been the focusing of researches in the last decades, contributing to the development of new interacting interfaces for digital computers, either as a mean of data entry, base for better understanding of speech and production of synthetic speech, or as a tool for linguists. In the area of signal processing, the Interval Analysis has shown important results in the imprecision treatment. This work discuss the speech recognition process, presenting the mathematics models used in this task to ultimately, presents the Interval Analysis Theory, together with the interval models already existing analog to the models used in the speech recognition, concluding with a propose of a Interval Model for Computer Speech Recognition.

Sumário

Sumário.....i

Lista de Figuras.....iii

Lista de tabelas.....iv

Lista de Abreviações.....v

1.INTRODUÇÃO.....1

1.1. Justificativa.....2

1.2. Objetivo.....3

1.3. Metodologia.....3

1.4 Organização do Trabalho.....4

2.FONÉTICA.....5

2.1. Fonética Articulatória.....5

2.2. Fonética Acústica.....11

3.USO DE MODELOS MATEMÁTICOS QUE LIDAM COM IMPRECISÃO PARA RECONHECIMENTO DE VOZ.....16

3.1 Conjuntos Fuzzy.....16

3.2. Hidden Markov Models.....18

3.3. Redes Neurais.....20

3.4 Análise Intervalar.....23

4.O PROCESSO DE RECONHECIMENTO DE FALA COMPUTACIONAL.....26

4.1 Aquisição do Sinal de Voz.....28

4.2 Pré-Processamento.....31

4.3 Extração de Descritores.....33

4.4 Treinamento e Classificação.....35

4.5 Reconhecimento.....36
5.O MODELO INTERVALAR.....37
5.1 Sinais Digitais Intervalares.....37
5.2 Normalização Intervalar.....38
5.3 Transformada Discreta de Fourier Intervalar.....39
5.4 Hmms Intervalares.....39
5.5 Redes Mapas Auto-Organizáveis Intervalares.....40
5.6 Apresentação do Modelo.....41
6.CONSIDERAÇÕES FINAIS.....44
7.REFERÊNCIAS.....46

Lista de Figuras

Figura 2.1: Representação Esquemática da Área Vocálica. 7

Figura 2.2: Esquema das partes da língua, alvéolos e úvula.....11

Figura 2.3: Gráficos da palavra "pato" gerados pelo Praat (programa de análise de fala)..…13

Figura 3.1: Uma unidade de uma Rede Neural.....21

Figura 4.1: Diagrama do ASR proposto por Bresolin (2008).....27

Figura 5.1: Modelo Intervalar Proposto.....42

Lista de tabelas

Tabela 2.1: Classificação de Vogais.....8

Lista de Abreviações

ASR Reconhecimento Automático de Fala (do inglês – *Automatic Speech Recognition*)

DFT Transformada Discreta de Fourier (do inglês – *Discrete Fourier Transform*)

FFT Transformada Rápida de Fourier (do inglês – *Fast Fourier Transform*)

HMM *Hidden Markov Model*

SOM Mapa Auto-organizável (do inglês – *Self Organizing Maps*)

IA Inteligência Artificial

IPA Alfabeto Fonético Internacional (do inglês – *International Phonetics Alphabet*)

LPC *Linear Predictive Coding*

MFCC *Mel-frequency Cepstral Coefficients*

PLN Processamento de Linguagem Natural

STFT Transformada de Fourier de Tempo Curto (do inglês – *Short-Time Fourier Transform*)

1. INTRODUÇÃO

Com o desenvolvimento dos dispositivos portáteis, aparelhos menores são criados a cada dia e as interfaces convencionais dos periféricos de entrada para computadores não se adéquam ou não atendem eficientemente as demandas destes novos aparelhos. Um teclado padrão, por exemplo, com mais de 100 teclas não pode ser inserido em um celular. Para isso, novos teclados, monitores sensíveis ao toque e teclados virtuais foram criados.

Estes novos dispositivos ainda impõem grandes limitações na relação $tamanho \times usabilidade$. A exemplo, um *mp3 player shuffle* de pouco mais de 11cm² contém uma quantidade muito limitada de comandos disponíveis. Parte disso pode ser atribuído à dificuldade de inserção de novos botões, por falta de espaço no aparelho.

Outro problema das interfaces de comunicação usuais (teclado, mouse, monitores sensíveis a toque, etc.) é a dificuldade, muitas vezes impossibilidade, de uso das mesmas por portadores de deficiências motoras.

O reconhecimento de fala por computadores trouxe uma nova possibilidade de interface de comunicação com o usuário, que resolve tanto os problemas de adaptação a dispositivos pequenos, quanto o uso por portadores de deficiências motoras. Apesar disso, o reconhecimento de fala por computadores é uma tecnologia recente e pouco desenvolvida, que enfrenta problemas como alto custo computacional e imprecisão nos resultados.

Desde Moore (1979), vários estudos vem sendo desenvolvidos em torno do uso de intervalos em ambientes computacionais, seja em computação científica para redução dos erros de aproximação, ou em circuitos e processamentos de sinais para tratar a imprecisão dos sistemas como em (TRINDADE, 2009) e (LORDELO E FERREIRA, 2005).

A fonética acústica já possui vários estudos que possibilitam o entendimento físico de fala, tanto no que diz respeito a sua formação no locutor, quanto sua propagação, recepção e entendimento pelo ouvinte.

A fala, segundo Fry (1979), assim como os outros sons, se propaga através de vibrações em partículas e pode ser estudada através da análise acústica. Além disso,

conforme Pickett (1998), por mais complexos que possam ser inicialmente, os sinais gerados podem ser convertidos e analisados como ondas senoidais, que simplificam a análise de sinais fazendo uso apenas das informações que se desejam dos sinais.

A literatura atual apresenta uma quantidade significativa de trabalhos que discutem, desenvolvem e implementam soluções intervalares para tratamento de imprecisão em diversas áreas. Da mesma forma, uma vasta gama de trabalhos em torno de reconhecimento de fala podem ser encontrados. No entanto, pouca ou nenhuma associação ainda foi feita entre a análise intervalar e as imprecisões no processo de reconhecimento computacional de fala.

Este trabalho propõe fundamentar a criação de um modelo de reconhecimento de fala com o uso de análise intervalar, apresentando os conceitos fundamentais que envolvem as áreas chaves de pesquisa nesta área (processamento de sinais, tratamento de imprecisões e fonética acústica), para, por fim, sugerir um modelo intervalar para reconhecimento computacional de fala.

1.1. JUSTIFICATIVA

Com os avanços da Inteligência Artificial (IA), o reconhecimento de fala por computadores passou a ser utilizado em diversos setores. Entre as principais aplicações estão o uso de computadores por comandos de voz, fornecendo acessibilidade, e a automação de atendentes de *call-centers.*

As tecnologias que fornecem estes serviços ainda necessitam de melhorias em diversos aspectos, por se tratarem de tecnologias recentes e em fase de desenvolvimento.

Uma das atividades do reconhecimento de fala por computadores é o processamento de sinais digitais de áudio. Parte dos problemas relacionados ao reconhecimento de fala está relacionado a problemas como o ruído e as variações do sinal gerado pelas mesmas palavras ditas por pessoas diferentes e, até mesmo, uma única palavra dita por uma pessoa e diferentes ocasiões.

O desenvolvimento de um modelo intervalar para o reconhecimento de fala pode oferecer melhorias significativas no reconhecimento de fala, uma vez que já possibilitou

melhorias em outras aplicações, como é apresentado no decorrer do trabalho. Apesar disso, a literatura existente não relaciona ainda os modelos que possuem versão intervalar com o reconhecimento de fala.

Entre outros fatores, este trabalho apresenta à comunidade de científica de maneira compilada, uma discussão mais completa sobre o reconhecimento de fala, os modelos matemáticos utilizados, suas versões intervalares (quando existentes) e fundamenta a proposta de aplicação de análise intervalar no reconhecimento de fala.

1.2. OBJETIVO

1.2.1.Objetivo Geral

Discutir aspectos teóricos relacionados ao reconhecimento de fala e à análise intervalar, de modo a fundamentar uma proposta de modelo intervalar para reconhecimento computacional de fala.

1.2.2 Objetivos Específicos

- Apresentar aspectos teóricos em torno do reconhecimento de fala;
- Apresentar modelos de tratamento de imprecisão neste processo;
- Elencar trabalhos que desenvolvam versões intervalares de modelos matemáticos usados no reconhecimento de fala;
- Propor um modelo baseado na teoria existente conforme modelos discutidos.

1.3. METODOLOGIA

O trabalho tem uma abordagem essencialmente teórica, apresentando conceitos básicos relacionados ao estudo da fala humana, tendo como base as discussões apresentadas em (KENT e READ, 2001), (LADEFOGED, 1996), (MASSINI-CAGLIARI e CAGLIARI, 2001), e (PICKETT, 1998). Outra base de referências são os trabalhos de análise intervalar, tendo como principais fontes (MOORE, 1979) e (TRINDADE, 2009).

Dois trabalhos na mesma área deste trabalho (processamento digital de sinais de fala)

tem contribuição essencial na fundamentação da discussão apresentada no decorrer do texto: (BRESOLIN, 2008) e (RABBINER e JUANG, 1993).

1.4 ORGANIZAÇÃO DO TRABALHO

O capítulo 2 apresenta conceitos básicos da fonética (área da linguística que se ocupa do estudo dos sons da fala), de modo a permitir um maior entendimento da formação da fala humana e das maneiras de representação propostas pelo linguistas.

No capítulo 3 são apresentados aspectos matemáticos sobre o tratamento de imprecisão, com enfoque no reconhecimento de fala.

O capítulo 4 apresenta com maior aprofundamento o processo de reconhecimento de fala e a imprecisão inerente ao processo.

No capítulo 5, são apresentados análogos intervalares de modelos matemáticos existentes encontrados na literatura disponível, para então ser proposto um modelo intervalar para o reconhecimento computacional de fala.

2. FONÉTICA

A fonética e a fonologia são duas ciências que tratam do mesmo objeto: a fala. Cada uma delas, porém, possui seu foco de estudo bem definido. A fonética se ocupa em descrever os sons da fala, enquanto a fonologia trata da interpretação dos resultados obtidos pela análise fonética dos sons da fala (MUSSALIM e BENTES, 2001).

A fonética trata de como os sons são produzidos, propagados e percebidos, enquanto a fonologia trata do que eles representam em determinada língua (MUSSALIM e BENTES, 2001).

A fonética pode ser vista sobre três diferentes pontos de vista: articulatório, descrevendo como o som é produzido no locutor; acústico, descrevendo como o som se propaga até chegar ao ouvinte e; auditivo, descrevendo como ele é percebido pelo ouvinte (MUSSALIM e BENTES, 2001).

Neste trabalho, serão considerados os aspectos articulatórios e acústicos da fala. O que não significa que os aspectos auditivos sejam menos importantes, mas o papel do ouvinte será realizado por um computador digital, portanto, o uso de analogias e a criação de modelos similares entre o computador e a audição humana aumentariam demasiadamente o escopo do trabalho.

2.1. FONÉTICA ARTICULATÓRIA

Segundo Bresolin (2008), “A fonética articulatória descreve a maneira natural com que os serem humanos articulam o trato vocal para a produção dos sons básicos[...]”. O processo da fala envolve diversos órgãos do corpo humano, chegando a atingir mais da metade do mesmo (MUSSALIM e BENTES, 2001). Bresolin (2008), destaca em seu trabalho três sistemas de maior importância na formação da fala: O sistema respiratório, o fonatório e o articulador.

Na maior parte do tempo, a fala é formada a partir de modificações na corrente de ar gerada pelo sistema respiratório. Os sons são formados através de vibrações no ar causadas pelo ar dos pulmões que passa pela glote. Alguns sons raros, os

cliques de algumas línguas africanas por exemplo, são gerados pela corrente de ar gerada por movimentos da laringe enquanto a glote está fechada, não fazendo uso da corrente de ar da respiração.

O sistema fonatório é formado pela laringe, que contém as cordas vocais. Conforme em (MASSINI-CAGLIARI e CAGLIARI, 2001), “a passagem que se forma entre as cordais vocais é chamada de glote”. A glote pode assumir diferentes configurações durante a fala.

Essas variações articulatórias da glote e da laringe modificam acusticamente o ar liberado pelos pulmões. Esse processo é conhecido por processo de fonação (MASSINI-CAGLIARI e CAGLIARI, 2001).

Tendo passado pelo sistema fonatório, o ar ainda pode sofrer obstruções que o modificam acusticamente ao passar pelas cavidades supraglotais, formadas pelos órgãos do sistema articulatório (faringe, língua, nariz, dentes e lábios), finalizando a formação do som da fala no locutor.

A existência ou não de obstrução na passagem de ar pelas cavidades supraglotais na formação de um som permite sua classificação em segmento consonantal ou vocálico, sendo respectivamente com obstrução e sem obstrução (BRESOLIN, 2008). Alguns segmentos não podem ser definidos como vogais ou consoantes, já que a passagem do ar não é bem definida. Exemplos dos mesmos são as semivogais e as aproximantes.

Por conta da diferença articulatória, critérios distintos são usados na subclassificação de vogais e consoantes. A Associação Fonética Internacional propôs uma classificação dos segmentos de acordo com suas características articulatórias, conhecida como Alfabeto Fonético Internacional (IPA), no qual vogais, consoantes e segmentos que não se enquadram como nenhum dos dois são classificados por diferentes critérios.

O IPA mostra não apenas os fonemas identificáveis nas línguas conhecidas, mas também as combinações articulatórias possíveis de produção ainda não

encontradas e as impossíveis de realização devido a fisiologia do aparelho fonador humano.

2.1.1 – Classificação De Vogais

Conforme supracitado, vogais são caracterizadas pela ausência de obstrução do ar nas cavidades supraglotais. Isso acontece porque durante a pronúncia das vogais a ponta da língua está sempre abaixada e sua superfície de forma convexa, limitando à área em que o som poderá passar sem obstrução das cavidades supraglotais (MUSSALIM e BENTES, 2001). Estes limites são mostrados pelo trapézio na figura 2.1.

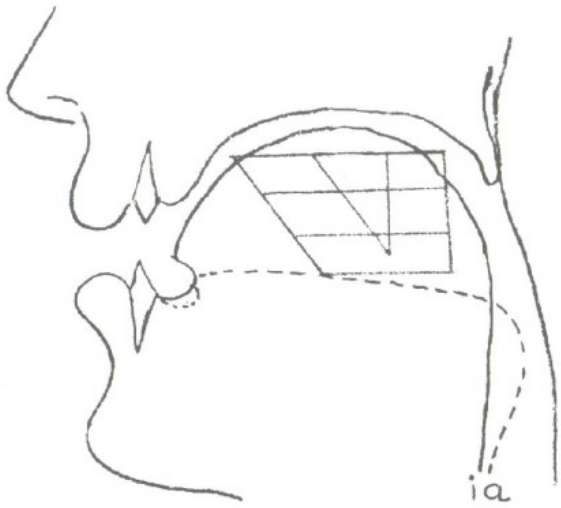

Figura 2.1: Representação Esquemática da Área Vocálica. Adaptada de (BRESOLIN, 2008).

As vogais são classificadas de acordo com as movimentações da língua no sentido horizontal e vertical, sendo quatro diferentes níveis para um e três para o outro, respectivamente. Além disso disso, ainda podem ser classificadas de duas diferentes formas de acordo com o arredondamento ou não dos lábios. A tabela 2.1 mostra a classificação das vogais conforme esses critérios e sua representação gráfica no IPA.

As vogais ainda podem sofrer um outro efeito que é a nasalização, caracterizada

pela passagem de ar pela cavidade nasal. Ela acontece no Português, por exemplo, quando uma vogal é escrita antes de n ou m, ou com o uso do acento “~”, como em “banho” e “pão”.

	Regiões Articulatórias					
	Anterior		Central		Posterior	
Altura: Fechada	i	y	ɨ	ʉ	ɯ	u
Meio-fechada	e	ø	ə	ɵ	ɤ	o
Meio-aberta	ɛ	œ	ɐ		ʌ	ɔ
Aberta	a	ɶ			ɑ	ɒ
	não arredondada	arredondada	não arredondada	arredondada	não arredondada	arredondada
	Labialização					

Tabela 2.1: Classificação de Vogais. Adaptado de (MASSINI-CAGLIARI e CAGLIARI, 2001).

2.1.2 – Classificação De Consoantes

Consoantes podem ser classificadas através do modo e o lugar de articulação, da vibração ou não das cordas vocais, e pelo mecanismo aerodinâmico envolvido (MUSSALIM e BENTES, 2001).

Pelo modo de articulação, os sons das consoantes podem ser classificados como (MUSSALIM e BENTES, 2001):

- Oclusivas: O som é produzido por um bloqueio na corrente de ar. Ex: da<u>d</u>o, <u>p</u>ato;

- Nasais: O som é produzido com o bloqueio do ar na cavidade oral e o rebaixamento do palatino permite a passagem de ar pelas narinas. Ex: so<u>nh</u>o, da<u>m</u>a;

- Fricativos: O som é produzido com o estreitamento de alguma parte do aparelho fonador, sofrendo fricção. Ex: <u>f</u>aca, <u>s</u>apato;

- Africados: O som é produzido inicialmente pelo bloqueio da passagem de ar dentro da cavidade oral, sofrendo posteriormente uma obstrução que provoca

fricção. É uma combinação de sons oclusivos e fricativos que ocorre no mesmo local de articulação. Ex: Tiago, ou diagrama, pronunciado como no dialeto carioca e o baiano.

- Laterais: A cavidade oral anterior bloqueia a passagem central do ar, permitindo apenas uma passagem lateral. Ex: labirinto, calha;

- Vibrantes ou vibrantes múltiplos: Caracterizados por batidas rápidas da língua no véu palatino;

- Vibrante simples ou tepe: Uma batida rápida da ponta da língua nos alvéolos dos incisos superiores, provocando uma rápida obstrução do ar. Ex: bravo;

- Retroflexo: O som é produzido pelo curvamento da ponta da língua para cima e para trás, como na pronúncia do “r” nos dialetos caipiras;

- Aproximantes: São sons formados acima da área das vogais, mas a passagem de ar é maior que a pressão que causa a fricção.

No que diz repeito aos locais de articulação os sons podem ser classificados em onze categorias diferentes (MUSSALIM e BENTES, 2001):

- Labial ou bilabial: É produzido pela aproximação do lábio superior no lábio inferior. Ex: pá, Maria.

- Labiodental: É produzido pela aproximação dos lábios inferiores nos dentes incisivos superiores. Ex: farinha, fé;

- Dental: É produzido com a ponta da língua junto à parte posterior dos dentes incisivos superiores ou entre os dentes incisivos superiores e inferiores. No português, só ocorre entre a língua e a parte posterior dos dentes, como em “sapato”. A outra forma é comum no Inglês em palavras como Theory.

- Alveolar: É produzido pela aproximação da ponta da língua nos alvéolos. Ex: nata;

- Palatal: É produzido com a aproximação da parte média da língua com o palato duro;

- Palatoalveolar: É produzido na região imediatamente posterior à região onde o som alveolar é produzido;

- Alveopalatal: É produzido na região imediatamente anterior a que os sons palatais são produzidos;

- Velar: É produzido quando a parte posterior da língua se aproxima do palato mole. Ex: ca<u>v</u>alo;

- Uvular: É produzido pela parte posterior da língua pressionando o fundo da cavidade oral (palato mole e úvula);

- Faringal: É produzida pela constrição da ponta da língua com a faringe;

- Glotal: É produzido pela articulação das cordas vocais. Ex: esca<u>rr</u>ar.

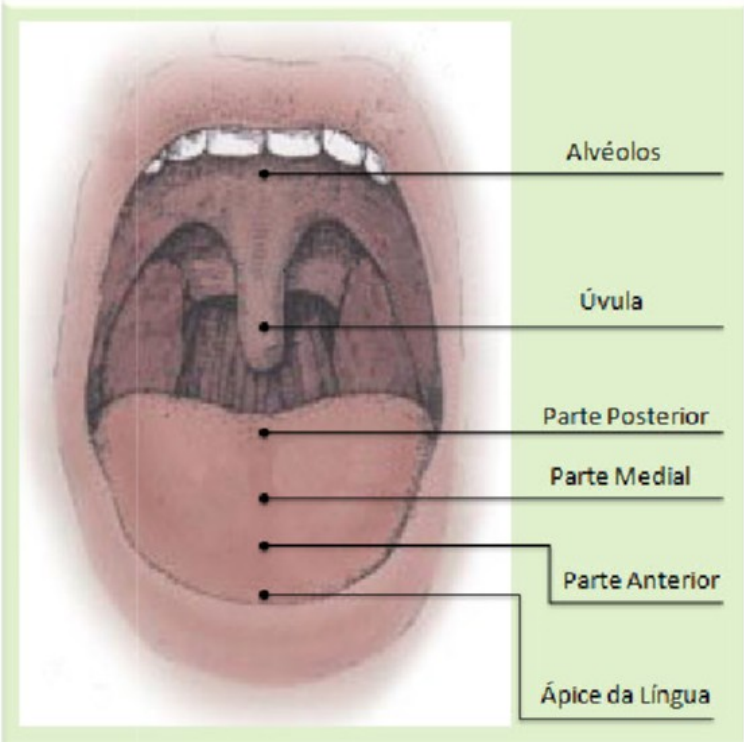

Figura 2.2: Esquema das partes da língua, alvéolos e úvula. Adaptada de (BRESOLIN, 2008).

A figura 2.2 mostra as partes da língua, alvéolos e úvula, órgãos utilizados na articulação da fala.

Quanto à vibração das cordas vocais, os sons podem ser classificados como surdos ou sonoros, sendo respectivamente sem vibração e com vibração das cordas vocais.

2.2. FONÉTICA ACÚSTICA

Segundo Fry (1979) os sons da fala, assim como os demais sons da natureza, se propagam através da vibração em partículas e podem ser estudados através de análise acústica. O produto final da fala é um sinal acústico, o qual representa a mensagem linguística do emissor (KENT e READ, 2001), que ao chegar no receptor é interpretada e entendida completando sua transmissão.

A fonética acústica é definida por Pickett (1998, pp. 5) como “a ciência da linguagem que trata do código sonoro da fala” e, ainda, conforme Pickett (1998), tem

como foco principal os padrões sonoros da fala e não sua função na linguagem.

Para a análise acústica, os sons podem ser encontrados em três diferentes formas (KENT e READ, 2001):

- Ondas acústicas: o som que se propaga no meio através da vibração em partículas. É dessa forma que o som chega ao aparelho auditivo humano, é convertido em impulsos elétricos e enviado ao cérebro;

- Sinal analógico armazenado: o som pode ser captado e transformado em sinais analógicos, por um microfone, por exemplo, que transforma a vibração no ar em sinais elétricos que podem ser armazenados em fitas magnéticas, preservando as características analógicas do sinal;

- Sinal digital armazenado: o som analógico, pode ser convertido em séries de números, passando neste caso a uma forma digital ou discreta. Computadores trabalham apenas com dados digitais e sons neste formato podem ser analisados com o apoio de computadores digitais.

A fonética acústica se ocupa basicamente de três linhas de pesquisa: a estrutura física dos sons da fala, a fala sintética e o reconhecimento automático da fala (MUSSALIM e BENTES, 2001), sendo o terceiro, utilizado no escopo deste trabalho.

Para o entendimento do reconhecimento automático da fala, é importante entender como os sinais da fala são armazenados e processados por computadores e, posteriormente, entender que características dos sons representados nesses sinais podem permitir sua distinção dos demais sons da fala humana.

2.2.1. Processamento Dos Sinais Digitais Da Fala

Para serem processados em computadores, após a conversão de analógicos para digitais, os sinais da fala são convertidos em uma série de números, chamados amostra, que representam a amplitude do sinal naquele intervalo de tempo. Cada amostra representa um espaço de tempo, desta forma, o número de amostras por segundo com o qual o sinal foi representado é informação importante para

mensuração da qualidade do áudio digital (LADEFOGED, 1996).

Esse sinal digital representa o som em forma de onda sinusoidal, e pode ser visto graficamente, de forma que sua amplitude é apresentada no eixo vertical, em função do tempo (eixo horizontal).

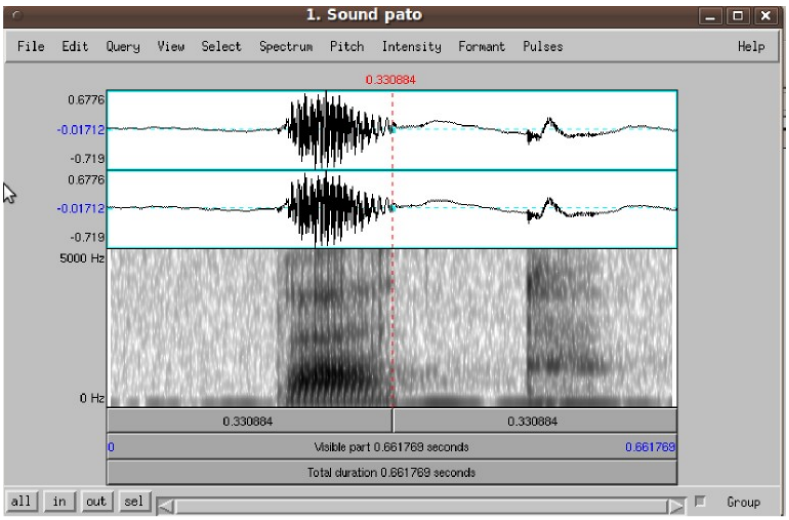

Figura 2.3: Gráficos da palavra "pato" gerados pelo Praat (programa de análise de fala). Acima gráfico em forma de onda e abaixo espectrograma.

A produção fala humana é contínua mas, apesar disso, pode ser dividida e analisada em partes menores como: frases, palavras, sílabas, etc. Para análise acústica, muitas vezes se deseja analisar o som por segmentos (consonantais e vocálicos), essa segmentação para análise pode ser realizada através do uso de “janelas”.

Janelas podem ser vistas como máscaras com buracos inseridas em torno da onda do sinal. No momento da análise, apenas as informações dentro da janela (dos buracos) serão consideradas e as demais descartadas. Na análise de Fourier, por exemplo, todas os pontos fora da janela, são comumente consideradas como zero (LADEFOGED, 1996).

Os sinais em forma de onda estão no domínio do tempo e muitas das informações que permitem distinguir um som de outro encontram-se no domínio da

frequência. Para análise do sinal no domínio da frequência, pode ser usada a Transformada de Fourier que, conforme Trindade (2009), converte um sinal no domínio do tempo para o domínio da frequência.

Uma questão importante a ser observada é que os sons da fala são formas de onda complexas formadas pela combinação de ondas simples em diferentes frequências (LADEFOGED, 1996) e, relações entre estas frequências são de extrema importância para a identificação de um som em detrimento de outros.

Ao analisar sinais no domínio da frequência pela transformada de Fourier, a informação sobre o tempo se perde, ou seja: dois sinais complexos compostos pelas mesmas ondas simples, mas em tempos diferentes, tem a mesma representação no domínio da frequência pela Transformada de Fourier (BRESOLIN 2008). Sendo importante, portanto, o janelamento dos sinais antes da Análise de Fourier para que o sinal de um som não interfira indevidamente na identificação dos demais.

Outra forma de análise dos sinais da fala é o espectrograma, um gráfico que possibilita a visão de três dimensões do sinal da fala, o tempo, a frequência e a intensidade, respectivamente no eixo horizontal, no eixo vertical e pela variação do tom de cinza (do mais claro para o mais escuro) (KENT e READ, 2001). Na figura 2.3 pode ser observado o gráfico de um som em forma de onda e o espectrograma de um mesmo som.

2.2.2. Pistas Acústicas No Reconhecimento De Fala

O processo de percepção e reconhecimento de fala por humanos envolve diversos aspectos físicos, sociais, culturais e de contexto. Ao escutar um som, normalmente o ouvinte tenta associá-lo a uma palavra conhecida que faça parte do contexto da conversa, por conta disso podemos, por exemplo, entender diferentes pronúncias da mesma palavra.

Em se tratando de computadores, a inserção do contexto pode ser atividade complexa. Além disso, o sinal acústico dos sons da fala traz consigo “pistas” que servem para identificação de características articulatórias do som, que podem

permitir sua classificação.

A classificação dos sons exclusivamente por suas características acústicas pode resultar em um sistema de difícil aplicação para processamento de linguagem natural, já que o contexto é desprezado e duas diferentes pronúncias da mesma palavra terão diferentes representações no sistema. A classificação preliminar dos sons acusticamente, no entanto, pode fornecer parâmetros mais sólidos que melhorem a eficiência do processamento de linguagem natural, aliados a métodos que considerem o contexto da fala.

Existem várias pistas acústicas que podem ser usadas tanto para identificação de vogais quanto de consoantes e as mesmas refletem características da articulação do som.

Na identificação de vogais pode-se destacar o uso da primeira e segunda frequências formantes, conhecidas na literatura por F_1 e F_2, apesar de outras pistas poderem ser usadas a depender do idioma, a duração da vogal no inglês, por exemplo. Para consoantes, algumas pistas podem separar sons surdos de vozeados, além de permitir inferências sobre o local e a maneira de articulação dos sons (FRY, 1979).

3. USO DE MODELOS MATEMÁTICOS QUE LIDAM COM IMPRECISÃO PARA RECONHECIMENTO DE VOZ

Uma das dificuldades da análise acústica da fala são as imprecisões que envolvem o processo. Conforme Ladefoged (1996), grande parte desta dificuldade é, muitas vezes, a impossibilidade de análise do som original. Pois com o uso de aparelhos, sejam digitais ou analógicos, o que se estuda é o som captado ou gravado e não o som propriamente emitido.

No processo de reconhecimento de fala, mesmo quando o som é gravado em cabines acústicas, traz consigo uma série de imprecisões tanto por parte dos circuitos envolvidos (microfone, placa de áudio, etc.) quanto pelo comportamento da fala humana. Além disso a conversão de áudio analógico em digital envolve dois processos de discretização, a quantização, tornando o sinal discreto no tempo, e a amostragem, tornando o sinal discreto em amplitude (STRANNEBY, 2001).

Assim como qualquer outra substituição de modelo ou processo infinito por finito, a transformação de sinais analógicos em digitais provoca erros (CLAUDIO e MARINS, 1994), que refletem diretamente na precisão do reconhecimento de fala.

Vários modelos matemáticos lidam com imprecisão e incerteza no domínio dos problemas. Neste capítulo são apresentados alguns destes modelos e suas aplicações no reconhecimento de fala.

3.1 CONJUNTOS FUZZY

A teoria tradicional dos conjuntos se adéqua muito bem a problemas nos quais é sempre possível definir quais elementos pertencem a determinado conjunto de acordo com determinadas características. O problema é que nem sempre é possível se ter uma visão exata e completa da realidade e, muitas vezes, a forma de percepção ou detecção da realidade é o uso de circuitos ou sensores que por sua própria natureza atuam sobre determinada margem de certeza. Além disso, em algumas situações fica difícil definir limites entre estados conceituais como entre o morno e o quente, ou o bom e o ruim.

Para tratar essas incertezas inerentes a certos ambientes, foi criada a teoria dos conjuntos *fuzzy*, que pode ser vista como alternativa à teoria tradicional dos conjuntos para lidar com problemas onde é impossível (ou muito difícil) definir a pertinência de um elemento a um conjunto com total certeza. Neste trabalho, os conjuntos tradicionais serão chamados de conjuntos "*crisp*", como por Klir e Yuan (1995), para que possam ser distinguidos dos conjuntos *fuzzy*.

Um conjunto *crisp*, segundo Klir e Yuan (1995), pode ser definido da seguinte forma:

$$A = \{x | P(x)\}, \tag{1}$$

onde $P(x)$ diz se x pertence ou não ao conjunto A. Além disso, os conjuntos podem ser definidos por uma função característica X_A conforme abaixo:

$$X_A(x) = \begin{cases} 1 \; se \; x \in A \\ 0 \; se \; x \notin A \end{cases} \tag{2}$$

A função X_A mapeia elementos de X em $\{0,1\}$ e pode ser expressa formalmente por (KLIR e YUAN, 1995):

$$X_A : X \rightarrow \{0,1\} \tag{3}$$

Por outro lado, dado um conjunto X qualquer, um conjunto *fuzzy* arbitrário A do conjunto X pode ser definido pela função (KLIR e YUAN 1995):

$$A : X \rightarrow [0,1] \tag{4}$$

, onde o valor retornado pela função indica o grau de pertinência do elemento ao conjunto *fuzzy* A.

Conforme pode-se observar comparando as definições 3 e 4, a principal diferença entre os conjuntos *fuzzy* e os conjuntos *crisp* é o contradomínio de suas funções características. Enquanto nos conjuntos *crisp*, um elemento pode apenas pertencer ou não a um conjunto (2 possibilidades), em um conjunto *fuzzy*, um

elemento pode pertencer a um conjunto com qualquer grau de pertinência entre 0 e 1. E é essa característica dos conjuntos *fuzzy* que permite sua aplicação no tratamento de incertezas, onde o grau o de pertinência de um elemento a um conjunto pode ser considerado o grau de certeza de que ele atende determinada propriedade.

Um exemplo de aplicação da teoria *fuzzy* no tratamento de imprecisões é o trabalho de MILLS (1996), no qual foi implementado um reconhecedor de dígitos isolados, usando uma versão *fuzzy* de algoritmos tradicionalmente usados no reconhecimento de voz.

Uma palavra dita pela mesma pessoa sofre, entre outras, variações no volume e na velocidade de pronúncia. Para minimizar estas variações, Mills (1996) fez uso de uma versão *fuzzy* do algoritmo de Programação Dinâmica Simétrica, para aproximar dois sons em forma de onda em relação ao tempo. Este algorítimo possui complexidade $O(n^2)$ e para evitar custos adicionais de processamento com análise de Fourier ou LPC, os sons não foram analisados no domínio da frequência, o que pode ter reduzido a precisão da análise, mas reduziu seu custo (MILLS, 1996).

O sistema *fuzzy* garantiu durantes os testes no mínimo a mesma precisão da versão *crisp* do sistema (MILLS ,1996).

3.2. HIDDEN MARKOV MODELS

Um *Hidden Markov Model* (HMM) é um modelo estocástico de máquina de estados que possui basicamente dois tipos de estados, os observáveis e os ocultos. Normalmente os estados ocultos representam alguma característica física do problema (CHING e NG, 2006). As transições dos estados observáveis são pré-definidas e as transições entre os estados ocultos são definidas de acordo com uma matriz de probabilidade gerada por um algorítimo específico baseado em eventos aleatórios.

Um HMM pode ser caracterizado pelos seguintes elementos (RABINER, 1989):

- O número N de estados ocultos. O conjunto de estados ocultos é denotado:

$$S=\{s_{1,}s_{2,}s_{3,}\ldots,s_n\} \quad ;$$

- O número M de símbolos observáveis distintos por estado oculto. O conjunto de símbolos individuais é denotado:

$$V=\{v_{1,}v_{2,}v_{3,}\ldots,v_M\} \quad ;$$

- A distribuição de probabilidades de transição dos estados $[A]_{ij}=\{a_{ij}\}$, onde

$$a_{ij}=P(Q_{t+1}=s_j|Q_t=s_i),1\leqslant j\leqslant N \quad ;$$

- A distribuição de probabilidades dos símbolos de observação no estado oculto j , $[B]_{jk}=\{b_j(v_k)\}$, onde:

$$b_j(v_k)=P(Q_t=v_k|Q_t=s_j),1\leqslant j\leqslant N,1\leqslant k\leqslant M \quad ;$$

- A distribuição de estado inicial $\prod=\{\pi_i\}$, onde:

$$\pi_i=P(Q_1=s_i),1\leqslant i\leqslant N \quad .$$

HMMs possuem três questões básicas (CHING e NG, 2006):

1. Dada uma sequência de observação $O=\{O_1O_{2\ldots}O_T\}$ e um HMM, como computar de maneira eficiente a probabilidade da sequência de observação?

2. Dada uma sequência de observação $O=\{O_1O_2\ldots O_T\}$ e um HMM, como escolher a sequência de estados correspondente $Q=\{Q_1Q_2\ldots Q_T\}$ que é ótima em determinado sentido?

3. Dada a sequência de observação $O=\{O_1O_2\ldots O_T\}$ como escolher os parâmetros do modelo em um HMM?

Para resolver o problema 1, normalmente é usado um algoritmo de programação dinâmica conhecido como *forward-backward*. O segundo problema consiste na

descoberta da parte oculta do modelo, onde normalmente é usado critério de otimalidade para resolver o problema tão bem quanto for possível. Uma técnica de programação dinâmica conhecida como algoritmo de Viterbi é comumente usado para resolver este problema. Para o terceiro problema, é usado um algoritmo conhecido por *Expectation-Maximization*.

Mais detalhes sobre a fundamentação matemática dos HMMs podem ser encontrados em (RABINER, 1989) e em (CHING e NG, 2006).

Uma característica importante dos HMMs para seu uso no reconhecimento de fala é a capacidade de lidar com eventos dinâmicos em função do tempo. Desta forma, eles podem ser implementados para realizar a identificação de padrões na fala tanto através de análise temporal quanto através de análise acústica ou as duas em conjunto (JUANG e RABINER, 1991).

Trabalhos em torno do reconhecimento de palavras isoladas com HMMs chegaram a mostrar eficiência superior a 95% para alguns vocabulários em torno de 1000 palavras, para diferentes interlocutores (JUANG e RABINER, 1991).

3.3. REDES NEURAIS

Vários algorítimos e modelos matemáticos para resolução de problemas por computadores são inspirados em processos e agentes da natureza. Diversas estruturas abstratas, a exemplo de pilhas, filas, árvores, são inspiradas em entidades do mundo concreto.

Assim como estes modelos, as redes neurais foram desenvolvidas tendo como inspiração uma entidade da natureza, neste caso, a organização e comunicação dos neurônios do cérebro humano. O termo “inspiração” tem sido usado pelo fato dos algoritmos de IA, segundo Russel e Norvig (1995), privilegiarem comportamento estritamente racional em detrimento de imitações do comportamento humano.

Uma rede neural é composta por um conjunto de nós ou unidades que são interconectados por ligações, as quais possuem, cada uma, um peso a ela associado. Cada unidade tem ligações de entrada de outras unidades e ligações de

saída para outras unidades, o valor da saída depende do nível de ativação que varia de acordo com os valores de entrada e os pesos das ligações. A ideia central é que cada unidade possa trabalhar dependendo apenas de suas entradas, sem a necessidade de uma visão global da rede (RUSSEL e NORVIG, 1995).

Esta estruturação independente e inerentemente paralela dos nós de uma rede neural, como menciona Braga *et al.* (2000), criam a possibilidade de obtenção de desempenho superior aos modelos convencionais, já que tal estrutura pode facilitar o uso de computação paralela ou distribuída.

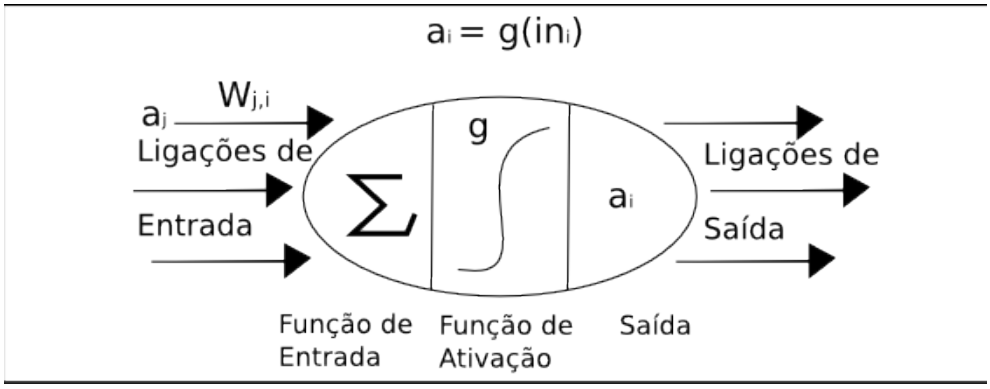

Figura 3.1: Uma unidade de uma Rede Neural. Adaptado de (RUSSEL e NORVIG, 1995)

A figura 3.1 mostra uma unidade de rede neural, onde $W_{j,i}$ é o peso da ligação para a unidade i , a_j , valor de entrada da ligação, $g(\text{in}_i)$, o valor de entrada da unidade aplicado à função de ativação e a_i , o valor de saída da unidade i .

Uma rede neural, assim que criada, pode ter seus pesos ajustados aleatoriamente ou de maneira que otimize sua convergência , e seu aprendizado é dado através do ajuste dos pesos das ligações por um algoritmo de treinamento. De acordo com a forma de aprendizado, as redes podem ser classificadas em supervisionadas (ou com professor) e não supervisionas (ou sem professor), sendo que na primeira, os pesos são reajustados com o fornecimento de entradas e resposta esperadas e na segunda, necessita-se apenas das entradas, que serão

automaticamente agrupadas.

Redes neurais são comumente aplicadas para o reconhecimento de padrões em ambientes de tempo estático ou em um segmento localizado do tempo (TEBELSKIS, 1995). Por conta disso, redes neurais dificilmente podem ser aplicadas para encontrar padrões através de análise temporal de sinais da fala (TEBELSKIS, 1995).

Devido às características supracitadas das redes neurais, elas geralmente são usadas em conjunto com outros modelos, realizando apenas a análise acústica dos sinais. O modelo usado por Tebelskis (1995) faz uso de uma combinação de redes neurais com HMM, no qual o HMM é usado para análise temporal dos sinais e redes neurais, na análise acústica.

Conforme os resultados do trabalho de Tebelskis (1995), redes neurais lidam bem com ruído nos dados, além de suportarem paralelismo com mais facilidade, sendo uma área promissora nos estudos de reconhecimento de fala.

3.3.1 – Redes Mapas Auto-Organizáveis

As redes mapas auto-organizáveis ou SOM (do inglês *Self-organizing Maps*), são um tipo de neural não supervisionada, criadas por Kohonen (1988), com aplicações em diversas áreas como biologia (MAHONY *et al.*, 2005), e, inclusive no reconhecimento de fala pelo próprio Kohonen (1988).

Redes SOM são capazes de agrupar topologicamente conjuntos de dados inicialmente dispersos através de um conjunto de características dos padrões de entrada (Braga *et al.*, 2000) ou, como menciona Haykin (2000), reduzem um padrão de sinal incidente de dimensões arbitrárias em um mapa discreto, comumente de uma ou duas dimensões, de forma topologicamente ordenada.

Em outras palavras, dado um conjunto de dados, uma rede SOM os agrupa de acordo com suas características semelhantes, construindo um mapa nos quais vetores semelhantes estão topologicamente próximos. Após a fase de treinamento, os agrupamentos podem ser rotulados, de modo que assim que um padrão é apresentado, a rede procura a unidade mais semelhante ao mesmo, que poderá ser

classificado pelo rótulo dos agrupamentos.

Usando uma rede SOM, Kohonen (1988) construiu um datilógrafo digital comandado por voz, no qual palavras são reconhecidas isoladamente através da classificação dos fonemas que as formam. O “datilógrafo” de Kohonen atingiu uma precisão que variava de 92% a 97%, dependendo da complexidade do texto.

3.4 ANÁLISE INTERVALAR

Sabe-se que entre dois números reais existem infinitos números. Uma das dificuldades advindas do uso de computadores digitais para resolução de problemas é a representação deste espaço contínuo dos reais em um ambiente discreto.

Tradicionalmente, as linguagens de programação fazem uso da aritmética de ponto flutuante para representação de um subconjunto dos reais em computadores. Números em ponto flutuante são gerados a partir de arredondamentos ou truncamentos. Desta forma, o que se obtém como resposta de um problema é uma solução que se aproxima do número real que representa a solução do sistema.

A análise intervalar propõe como solução a estes problemas a visão de intervalos de números reais como um novo tipo numérico composto por dois números reais, um representando o limite inferior e o outro o limite superior do intervalo (MOORE, 1979).

Um intervalo é definido por Moore (1979) como um conjunto limitado fechado de números reais tal que:

$$[a,b]=\{x : a \leq x \leq b\} \qquad (1)$$

Uma notação muito utilizada na literatura é $[\underline{X}, \bar{X}]$ onde $\underline{X}$ e $\bar{X}$ são respectivamente o limite inferior e superior do intervalo.

A principal vantagem do uso de intervalos é a possibilidade de representação do número junto à informação que representa sua imprecisão, enquanto na representação em ponto flutuante, após o arredondamento ou truncamento se perde

a informação sobre a imprecisão do processo.

Outra questão importante é a existência dos intervalos degenerados, que são os intervalos que possuem o limite superior igual ao limite inferior. Através deste intervalos, podemos representar os números reais, fazendo deles um caso particular dos intervalos.

Importante lembrar que, se tratando de computadores, não é possível se trabalhar com intervalos reais, já que os limites terão que ser representados por números limitados, provavelmente com aritmética de ponto flutuante (MOORE, 1979).

A diferença no uso dos intervalos está na consideração da imprecisão durante todo processo. Para tal, Moore (1979) definiu uma aritmética específica para os números intervalares, com as operações usuais de soma, multiplicação, oposto e inverso, através das quais também pode se definir a subtração e a divisão. As operações podem ser definidas segundo Trindade (2009) como:

$$X * Y = \{ x * y : x \in X, y \in Y \}, \tag{3}$$

onde $*$ representa qualquer uma das operações usuais de soma, subtração, multiplicação ou divisão.

Trindade (2009) apresenta em seu trabalho uma série de recursos para o processamento de sinais digitais usando análise intervalar, entre eles estão a transformada-*Z* e a transformada de Fourier intervalares. Outro recurso importante é uma métrica estritamente intervalar para a distância entre dois intervalos, que pode representar a distância entre dois números intervalares, mantendo a informação de imprecisão durante os cálculos.

Analogamente ao trabalho de Mills (1996), no qual é usada lógica *fuzzy* na representação das incertezas do processo de reconhecimento de fala, a análise intervalar pode ser vista como uma possível alternativa para ampliação da eficiência dos reconhecedores de voz através do melhor tratamento das imprecisões inerentes ao processo.

4. O PROCESSO DE RECONHECIMENTO DE FALA COMPUTACIONAL

Em alguns trabalhos (RABINER e JUANG (1993), por exemplo), o processo de reconhecimento de fala é definido de maneira a englobar todas as atividades desde a captação do som, passando pelo processamento, reconhecimento das estruturas fonéticas, léxicas e sintáticas, até o entendimento semântico da sentença dita.

Por outro lado, a partir do momento em que as estruturas fonéticas foram identificadas e, possivelmente, realizado o reconhecimento léxico, o processo de reconhecimento se torna independente do meio através do qual a mensagem foi captada (fala, entrada de texto, reconhecimento óptico de carácter etc). A partir deste ponto, o processo acaba se confundindo com a área de Processamento de Linguagem Natural (PLN). Há ainda trabalhos , como em (JURAFSKY e MARTIN, 2000), que incluem o reconhecimento de fala como parte do PLN.

Trabalhos voltados à área de processamento de sinais como: (BRESOLIN, 2008), (TEBELSKIS, 1995), (MILLS, 1996), tratam o reconhecimento de fala apenas como as etapas que vão da captação do sinal ao nível léxico. Por se tratar de estudo relacionado à mesma área destas pesquisas, será utilizado neste trabalho um modelo de reconhecimento de fala similar aos citados, baseado no modelo utilizado por BRESOLIN (2008), no qual o reconhecimento de fala é dividido em seis fases:

a) **Aquisição do sinal de voz**: as vibrações no ar provocadas pela fala do emissor são captados por algum aparelho e convertidas para uma representação digital do sinal emitido;

b) **Pré-processamento**: preparação do sinal para otimização da análise através de filtros, normalizadores, além de janelamento do sinal, quando necessário, para posterior análise;

c) **Extração de descritores**: uso de ferramentas de processamento digital de sinal, aliadas a técnicas de fonética acústica para extração dos elementos determinantes para a classificação dos sons emitidos;

d) **Treinamento**: treinamento de uma máquina inteligente, que através dos

dados fornecidos seja capaz de classificar os sons emitidos na fase de reconhecimento;

e) **Classificação**: após treinada, ao receber um novo som, a máquina deve classificá-lo, de acordo com suas características, como pertencente a um dos padrões encontrados durante o treinamento, ou a nenhum deles;

f) **Reconhecimento**: após a classificação do som ou de suas partes, o sistema deve reconhecer a sentença recebida através de decisões lógicas.

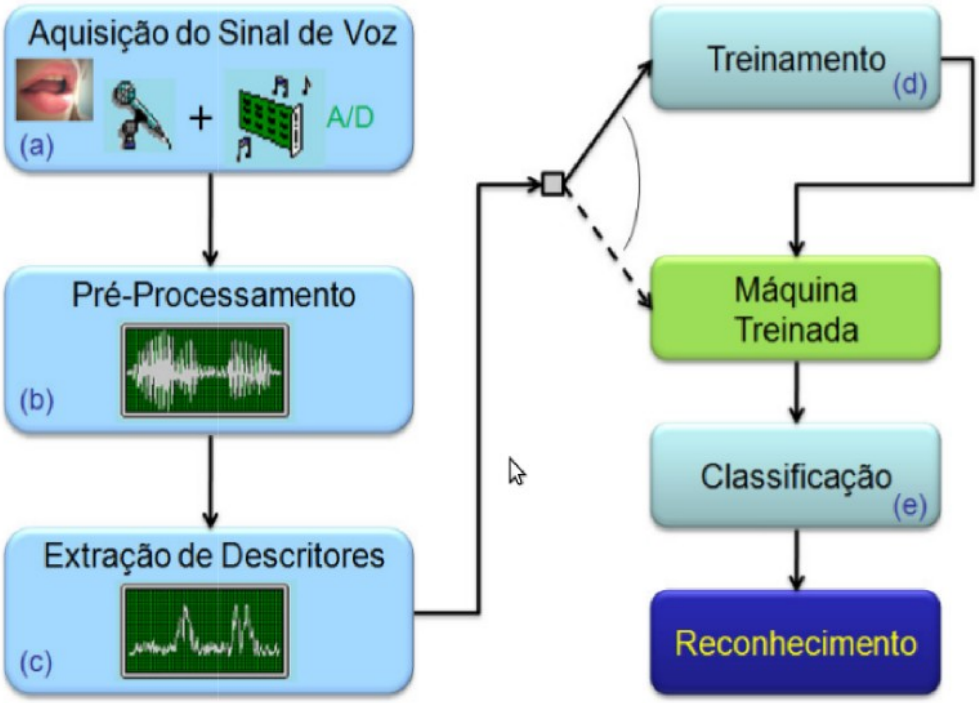

Figura 4.1: Diagrama do ASR proposto por Bresolin (2008). Fonte: (BRESOLIN, 2008)

Em seu trabalho, Bresolin (2008) concentra-se nas etapas (*c), (d)* e *(e)* da figura 4.1, pois nas fases anteriores, a literatura disponível apresenta propostas com poucas variações e os modelos se encontram mais sólidos. No entanto, ao se propor um modelo intervalar, as fases anteriores ganham maior importância na discussão, pois na fase de aquisição *(a)* é que surgem as primeiras imprecisões do processo e uma vez iniciado seu tratamento, é coerente que se trate a imprecisão em todos

passos posteriores.

Na etapa de reconhecimento, porém, não faz sentido se falar em intervalos, uma vez que neste momento o padrão de sinal recebido já foi classificado e as decisões a serem tomadas estão mais ligadas ao contexto semântico que à estrutura do sinal.

Nos tópicos que seguem, as etapas do reconhecimento de fala serão discutidas mais aprofundadamente, ao tempo que relacionadas com a proposta de análise intervalar de sinais digitais como possível alternativa aos modelos convencionais.

4.1 AQUISIÇÃO DO SINAL DE VOZ

Todo o processo de reconhecimento de fala, tem início na aquisição do sinal de voz. É nele que as vibrações no ar provocadas pela fala do emissor da mensagem são captadas e armazenadas digitalmente para que sejam realizados o processamento e análise necessários.

Poucos trabalhos discutem esta etapa do reconhecimento, uma vez que os Sistemas de Reconhecimento Automático de Fala (ASR do inglês *Automatic Speech Recognition*) devem funcionar em ambientes heterogêneos em relação à captação do sinal, pois essa etapa é realizada em nível de hardware e impor restrições deste tipo limitaria consideravelmente o número de ambientes no qual o ASR poderia ser utilizado.

Para discussão de uma proposta intervalar, no entanto, o entendimento desta fase permite dar coerência à imprecisão representada nos primeiros sinais digitais intervalares. Faria pouco sentido o início do uso de intervalos de maneira aleatória, simplesmente supondo a imprecisão e substituindo números de ponto flutuante por intervalos.

Em computadores, a aquisição do sinal passa normalmente por dois processos: a captação do sinal por **microfone** e a conversão do sinal analógico para digital pela **placa de captura de áudio**. Um maior entendimento do funcionamento destes equipamentos, pode fornecer parâmetros para definição de intervalos que representem a imprecisão desde o início do trabalho do reconhecedor.

4.1.1 Microfones

Microfones são dispositivos eletroacústicos que atuam na conversão de energia acústica em elétrica. De maneira geral, possuem um diafragma ou outra superfície qualquer flexível que se movimenta em resposta às ondas acústicas recebidas (BALLOU, 1987). A saída fornecida pelo microfone é um sinal elétrico que equivale em forma e amplitude à onda acústica.

Existem diversos tipos de microfones com diferentes precisões e qualidades, eles podem variar quanto à forma de captação das variações de pressão no ar (microfone de pressão ou de velocidade) quanto a direção de captação do som (omnidirecional, bidirecional ou unidirecional), quanto ao material utilizado na produção (carbono, cristal, cerâmica etc) (BALLOU, 1987) entre outros.

De acordo com a combinação dessas características acima citadas, cada tipo de microfone apresentará diferenças em:

a) **relação sinal/ruído:** além de captar o sinal que se deseja, o microfone também pode registrar, misturado ao sinal, ruídos do ambiente, ou até mesmo resultantes do meio de transmissão. Esta relação mostra quanto ruído é captado juntamente ao sinal desejado.

b) **faixa de frequência sensível:** a depender do tipo, os microfones podem ser sensíveis a maiores faixas de frequência, representando assim com maior fidelidade a onda recebida. Importante lembrar que maior sensibilidade pode fazer também com que não se alcance a relação sinal/ruído desejada em ambientes ruidosos.

A partir das duas características supracitadas, pode se ter noção da imprecisão adicionada pelos microfones na aquisição do sinal de voz pelo sistema ASR. Ao se pensar em um modelo genérico, deve-se levar em consideração que intervalos muito pequenos podem não representar satisfatoriamente a imprecisão nos piores casos, e que o uso de intervalos muito grandes pode subutilizar o potencial de microfones melhores, além de gerarem resultados possivelmente inconclusivos.

4.1.2 Placas De Captura De Áudio

Após ser captado pelo microfone, o sinal de voz é recebido pela placa de captura de áudio, que realizará a conversão do sinal analógico em digital (A/D).

Um sinal analógico é contínuo em relação ao tempo e à amplitude. Uma vez que computadores digitais são capazes de lidar apenas com dados discretos, o sinal de voz captado pelo microfone deve ser discretizado em tempo e amplitude. Conforme Stranneby (2001), uma sinal é considerado digital quando discreto em tempo e amplitude.

A conversão de um sinal continuo para discreto em relação ao tempo é chamado quantização e consiste (como o próprio nome sugere) na extração de amostras do sinal em intervalos de tempo com tamanho habitualmente constante. Este período de tempo T entre a extração das amostras é conhecido como período de amostragem e está relacionado com a frequência de amostragem da seguinte forma (STRANNEBY, 2001):

$$f_s = \frac{1}{T}$$

Em placas de captura de áudio, a frequência de captura suportada, normalmente expressa em Hertz (Hz), representa a quantidade de amostras por segundo que o dispositivo é capaz de capturar. Quanto maior este valor, menos do sinal original é perdido.

Outra conversão que o áudio passa na placa de captura é a quantização, que transforma o sinal contínuo em discreto em relação à amplitude, através da extração dos pontos do sinal que pertencem a um conjunto discreto de amplitudes (STRANNEBY, 2001).

A discretização em amplitude está diretamente relacionada a capacidade de representação numérica do sistema computacional utilizado, e representa perdas similares às decorrentes dos cálculos realizados em aritmética de ponto flutuante. É difícil de mensurá-la considerando isoladamente a qualidade das placas de captura

utilizadas, mas é importante saber que este processo de discretização, assim como qualquer outro adicionará imprecisão no processo.

A relação sinal/ruído utilizada nos microfones também pode ser utilizada em placas de captura para complementar a análise de sua qualidade.

É importante lembrar que sistemas ASR, de modo geral, devem lidar com a heterogeneidade dos equipamentos usados na fase de aquisição do som da fala. A discussão apresentada neste tópico objetiva apresentar de maneira mais detalhada as principais origens da imprecisão no reconhecimento de fala e apresentar ideias que levem a parâmetros para mensuração da imprecisão a ser considerada no início do processo de reconhecimento de fala.

4.2 PRÉ-PROCESSAMENTO

Uma vez que o sinal é captado e armazenado digitalmente, o próximo passo do reconhecedor é preparar o sinal para que possa ter suas principais características extraídas de maneira mais eficiente pelos descritores. Nesta etapa, o sinal pode ser otimizado para análise, através do uso de filtros, normalização e separação de *background* (separação do sinal de voz dos ruídos de fundos) como fez Bresolin (2008) e, posteriormente janelado, caso a análise seja baseada em unidades menores que a palavra.

Filtros normalmente são usados no domínio da frequência para eliminar aquelas que representam ruídos amplamente conhecidos, como em (BRESOLIN, 2008) que removeu frequências acima 9*kHz,* que podem ser gerados pela corrente elétrica. A normalização apenas traz todo o sinal para um intervalo que facilite sua análise (geralmente $[0,1]$). Já a separação de *background,* pode ser feita no domínio do tempo, usando a energia do sinal (BRESOLIN, 2008).

4.2.1 Janelamento Do Sinal De Voz

O reconhecimento de fala por unidades menores que a palavra parte do princípio que ela é resultado da conexão sequencial de fonemas. Além disso, a segmentação de fonemas dentro do sinal assume que as propriedades do sinal de fala alteram-se

instantaneamente na transição de um fonema para outro. O que nem sempre é verdade, pois a pronúncia de um fonema pode provocar alterações no fonema seguinte (efeito conhecido como coarticulação) (ALMPANIDIS e KOTROPOULOS, 2008). Por exemplo, a pronuncia de uma consoante fricativa [s], como em “sabão”, pode dar características fricativas à vogal [a] imediatamente posterior.

Apesar dos efeitos da coarticulação, a ideia do janelamento do sinal para identificação de fonemas é definir regiões nas quais estão presentes de maneira mais determinante as características de cada fonema, ou como explicado em (ALMPANIDIS e KOTROPOULOS, 2008), fornecer ponteiros para aproximadamente o início e fim de um fonema.

Uma das propostas para o janelamento é a consideração de estacionaridade do sinal de 10 a 30ms (RABBINER 1975, apud BRESOLIN, 2008), e então o janelamento simétrico do sinal, que apesar de não separar com grande precisão os fonemas, possui baixa complexidade de implementação e baixo custo em desempenho computacional.

Uma implementação deste tipo de janelamento pode ser vista em (BRELOSIN, 2008), no qual foi utilizado janelamento de 30ms que superposição de 33,33%.

Uma vez que a audição humana possui uma capacidade de reconhecimento muito maior que os sistemas ASR, é plausível que se busque implementações computacionais que usem características similares ao modelo humano. Motivado por este aspecto, juntamente ao fato de que as células do ouvido humano possuem resposta assimétrica a impulsos, Rozman e Kodek (ROZMAN e KODEC, 2007) propõem o janelamento assimétrico do sinal de voz.

Em (ROZMAN e KODEC, 2007) pode-se observar que apesar de não existir estudos que mostrem que a segmentação ótima de fonemas leve a um nível ótimo de reconhecimento, a segmentação assimétrica de fonemas oferece maior resolução para análise espectral, melhorando o desempenho de sistema ASR na presença de ruídos, até mesmo quando usada após o treinamento do classificador.

Além de (ROZMAN e KODEC, 2007), outros trabalhos como (ALMPANIDIS e KOTROPOULOS, 2008) e (GOH e RAVEENDRAN, 2009), discutem o janelamento assimétrico de sinais de voz, respectivamente com o uso de Critério de Informação Bayesiano com Distribuição Gama no domínio da frequência e, Análise de Autocorrelação em dados gerados pela Transformada Wavelet Diádica.

4.3 EXTRAÇÃO DE DESCRITORES

No capítulo 2 são discutidas as pistas presentes no sinal acústico que possibilitam a identificação de fonemas através de análise acústica. Estas características do sinal que permitem tal identificação são conhecidos como os descritores da fala (BRESOLIN, 2008).

Este trabalho discutirá apenas dois extratores de descritores conhecidos, a Transformada de Fourier e a Transformada Wavelet. Outros modelos largamente usados como a LPC (Linear Predictive Coding), banco de filtros, MFCC (Mel Frequency Cepstral Coefficients) podem ser encontrados em (BRESOLIN, 2008), (RABBINER e JUANG, 1993) e (MILLS, 1996).

4.3.1 Transformada Discreta De Fourier (DFT – *Discrete Fourier Transform*)

Como já foi dito, no contexto de processamento de sinais, a Transformada de Fourier leva um sinal do domínio do tempo para o domínio da frequência. Sendo originalmente contínua, possui uma versão discreta em relação ao tempo, outra discreta em relação ao tempo e à frequência.

Como será apresentada no próximo capítulo a versão intervalar da DFT, apresentaremos a definição da DFT.

Dada uma sequência $X[n]$ de tamanho finito N, a sua DFT é definida por:

$$X(k)=\sum_{n=0}^{N-1} X[n]e^{\frac{-2\pi i}{N}}, 0 \leqslant k \leqslant N-1$$

Existe também uma versão da DFT que representa melhor variações em função do tempo, utilizando em sua definição um espécie de janela do sinal. Essa versão é

conhecida como Transformada de Fourier Dependente do Tempo ou STFT (Transformada de Fourier de Tempo Curto – do inglês *Short-Time Fourier Transform*) (RABBINER e SCHAFER, 1978).

Os coeficientes da DFT podem ser usados isoladamente para análise acústica, por exemplo para localização das frequências formantes citadas no capítulo 2.

Além disso, a DFT atua como parte de outros descritores como bancos de filtros, na produção do espectrograma (RABBINER e SCHAFER, 1978) e como parte da MFCC (BRESOLIN, 2008).

A computação da DFT possui complexidade $O(n^2)$ e, para resolver este problema, foi desenvolvido um algoritmo conhecido como Transformada Rápida de Fourier (FFT – do inglês *Fast Fourier Transform*), que possui complexidade $O(n \log n)$ e é geralmente usado em implementações computacionais (BRESOLIN, 2008).

4.3.2 Transformada Wavelet

A Transformada de Fourier apresenta uma grande limitação ao se tratar de sinais não estacionários. A análise de Fourier não trata as variações no tempo, ou seja a junção de sinais com frequência diferentes, independente da ordem de junção dos sinais apresenta o mesmo resultado com a Transformada de Fourier (BRESOLIN, 2008).

As Transformadas Wavelet permitem a análise de sinais em relação ao tempo e à frequência ao mesmo tempo, sendo mais indicada neste caso para sinais não estacionários como a fala, como exemplifica BRESOLIN (2008).

Apesar do poder de análise oferecido, a Transformada Wavelet Packet (um dos tipos de Transformadas Wavelet) possui custo computacional de ordem $O(n \log n)$, assim como a FFT. Esse motivo somado às vantagens já citadas, levaram BRESOLIN (2008) a fazer uso da Transformada Wavelet Packet como descritor em seu trabalho.

4.4 TREINAMENTO E CLASSIFICAÇÃO

Estas duas etapas (treinamento e classificação) apesar de similares, atuam em momentos diferentes do processo de reconhecimento de fala. Como já citado, o objetivo da fase de treinamento é preparar uma máquina que seja posteriormente capaz de classificar sons através dos dados fornecidos. Já a classificação acontece após o período de treinamento, quando o reconhecedor já está em pleno funcionamento e deve classificar sons recebidos.

Dois importantes modelos matemáticos utilizados no reconhecimento de fala são discutidos no capítulo 3 deste trabalho, as Redes Neurais e os HMM. Uma vez que suas definições e aplicações em sistemas ASR já foram discutidas, apresentaremos aqui apenas aspectos referentes ao treinamento e classificação.

Em HMMs, a classificação é realizada através de decisões probabilísticas. Desta forma, na fase de treinamento é fornecida uma sequência de exemplos, chamada sequência de observação e o HMM encontra um modelo probabilístico que represente esta sequência. Após definidos estes estados, na fase de classificação, para uma sequência de observação dada, o HMM deve fornecer o modelo que possui maior probabilidade de produzir a sequência (RABBINER, 1989).

Já em redes neurais, a forma de treinamento dependerá do tipo de rede. Na rede não supervisionada, são fornecidos exemplos, que a rede agrupará em vizinhanças de acordo com suas semelhanças. Nas supervisionadas, exemplos são fornecidos em par com a resposta esperada e a rede ajusta seus pesos de acordo com o exemplo dado (BRAGA et. al, 2000). Após treinada, ao ser fornecida uma sequência de dados, a rede vai determinar a que grupo ela pertence.

4.4.1 Independência De Locutor

Devido às diferenças entre o sinal de voz produzido por diferentes pessoas, vários estudos são realizados em torno do reconhecimento de fala independente de locutor, ou para um grupo de locutores. O maior problema é que, muitas vezes, um reconhecedor treinado com a voz de uma pessoa não consegue reconhecer a voz

de outros locutores e reconhecedores genéricos apresentam taxas de reconhecimento menores que os dependentes de locutor (TEBELSKIS, 1995).

Tebelkis (1995), apresenta duas técnicas que aprimoram o reconhecimento para mais de um locutor. A primeira, que funciona para um grupo limitado de locutores, treina redes neurais ou HMMs para cada locutor e treina um reconhecedor de locutor. A outra técnica, conhecida por normalização, consiste no treinamento de uma rede neural para mapear outras vozes para representação equivalente em uma voz tida como padrão.

Na primeira técnica, o maior gargalo é o reconhecedor de locutor, que além de adicionar tempo no processamento, se falhar, comprometerá todo o reconhecimento. Já na segunda técnica, o modelo pode funcionar bem com padrões de voz similares à voz escolhida, mas pode apresentar grandes dificuldades com pessoas que apresentem padrão de voz muito distante do comum.

4.5 RECONHECIMENTO

Uma vez identificados os sinais fonéticos da fala, através de decisões lógicas e uso de modelos simbólicos (como máquinas de estados), a sentença dita pode ser convertida em texto, para posteriormente passar pela análise sintática, semântica e pragmática, para que possa ser interpretada pelo computador a mensagem de voz recebida.

Esta etapa envolve importantes áreas do PLN, como análise para desambiguação, recuperação de informação e recuperação de contexto através da análise pragmática.

Após este momento (para fins de análise computacional) o fato da linguagem ser falada, difere pouco da escrita para, exceto possivelmente pela coloquialidade do texto falado. Tal discussão excede o escopo do presente trabalho, mas pode ser encontrada em (JURAFSKY e MARTIN, 2000).

5. O MODELO INTERVALAR

O capítulo anterior discutiu o processo de reconhecimento de fala e apresentou modelos matemáticos que constituem ferramentas para este processo. Foram ainda apresentadas características dos sistemas ASR que inserem imprecisões nos mesmos.

Em momento anterior (item 3.4) foi apresenta a proposta de Moore de análise de intervalos como tipo numérico, representando incertezas de processos computacionais.

Neste capítulo, são apresentados fundamentos de análise intervalar de sinais digitais, sugerindo aplicações dos mesmos em reconhecimento de fala, enquanto são apresentados trabalhos que discutem versões intervalares de modelos matemáticos apresentados no capítulo anterior para, por fim, ser apresentado um esboço de modelo intervalar para reconhecimento de fala.

5.1 SINAIS DIGITAIS INTERVALARES

Como já foi visto anteriormente, as imprecisões dos sistemas ASR tem início no processo de aquisição do sinal que, se representado por sequências de números de ponto flutuante, não manterá informações sobre a incerteza do processo. Para o modelo intervalar, sugere-se então, que o sinal de voz capturado seja armazenado como um sinal digital intervalar.

Para maior compreensão do que vem a ser um sinal digital intervalar, é importante definir primeiro o que é um sinal digital, para posteriormente definir os sinais digitais intervalares como um caso particular. Pode-se definir um sinal intervalar como:

> “[...] um sinal que carrega em si uma quantidade de incerteza e é representado por um intervalo limitado nos seus extremos pelo mínimo e o máximo que o sinal pode assumir, sendo o diâmetro do intervalo a quantidade de incerteza que o sinal carrega”. (TRINDADE, Roque Mendes Prado. *Uma fundamentação matemática para processamento digital de sinais intervalares*. Tese (Doutorado em Engenharia Elétrica) - UFRN. Natal. 2009. p. 75).

Um sinal digital intervalar pode ser visto como: “[...] um sinal representado por uma sequência de intervalos digitais $X[n]$, sendo o diâmetro de cada termo da sequência a quantidade de incerteza que o sinal carrega” (TRINDADE, 2009).

Desta forma, para o modelo intervalar, os sinais de voz podem ser armazenados como sinais digitais intervalares formados por sequências de intervalos de diâmetro n , com centro em p_i . Onde n representa a imprecisão do sistema de captação do sinal (microfones, placa de som, computador etc.) e p_i , o valor medido pelo sistema no $i-ésimo$ ponto do sinal.

5.2 NORMALIZAÇÃO INTERVALAR

Bresolin (2008), realiza a normalização do sinal de voz dividindo todos os pontos do sinal pelo valor máximo de sua amplitude pela equação:

$$xnorm_i = \frac{x_i}{max|x_i|}$$

Para versão intervalar, tomando intervalos do tipo $[\underline{X}, \overline{X}]$, o sinal pode ser normalizado pela equação:

$$Xnorm_i = \frac{X_i}{max|\overline{X}_i|}$$

Outra opção, supondo o uso de hardware e software de sons convencionais que retornam os sinais captados com números de ponto flutuante, é normalizar o sinal digital captado e, após a normalização, converter o sinal para um sinal digital intervalar, onde cada termo da sequência que forma o sinal é um intervalo de diâmetro $\frac{n}{max|\overline{X}_i|}$.

Desta segunda forma, pode-se reduzir o custo computacional da operação, uma vez que operações em ponto flutuante custam menos que operações em intervalos.

5.3 TRANSFORMADA DISCRETA DE FOURIER INTERVALAR

A Transformada de Fourier, já apresentada anteriormente, possui grande importância no processamento de sinais, podendo ser usada na construção de filtros de frequência, na extração de coeficientes MFCC, entre outras formas no pré-processamento e na extração de descritores do sinal da fala.

Trindade (2009) definiu e discutiu algumas propriedades da Transformada-*Z* discreta intervalar e, para tal, definiu uma versão intervalar da DFT provando que, analogamente às definições existentes para os reais, no conjunto dos intervalos, a Transformada de Fourier é um caso particular da Transformada-*Z* (com raio de convergência igual ao intervalo degenerado $[1,1]$).

A Transformada Discreta de Fourier Intervalar é definida em (TRINDADE, 2009) como:

$$X(\boldsymbol{e}^{jw}) = \sum_{k=-\infty}^{k=\infty} X[k]\boldsymbol{e}^{-jwk} \quad ,$$

onde $\boldsymbol{e}$ representa o intervalo degenerado $[e,e]$.

Apesar de ter sido definida para um sinal discreto em amplitude, a DFT intervalar proposta por Trindade produz um sinal contínuo no domínio da frequência. Para que possa se produzir um sinal discreto no domínio da frequência, a partir de amostras de um sinal de um sinal contínuo no tempo, podemos alterar os limites do somatório para de $k=1$ para $k=n$, onde n é o número de amostras do sinal.

5.4 HMMS INTERVALARES

Parte significativa da essência dos HMMs está concentrada no cálculo da matriz de distribuição de probabilidades da transição dos estados $[A]_{ij}$ e da matriz de distribuição de probabilidades dos símbolos de observação $[B]_{ij}$. Desta forma, partindo da definição de probabilidades intervalares, surge a proposta de HMMs Intervalares.

Santos *et al.* (SANTOS *et al.*, 2006) define um HMM intervalar como um HMM que faz uso de probabilidades intervalares.

A ideia proposta por Campos *et al.* (CAMPOS *et al.*, 2002) para definição de probabilidades intervalares é baseada nas limitações de representação numérica em computadores digitais. Após associada a um evento A, uma probabilidade $P(A)=p$ é associada a um intervalo que a contenha, de modo que dada uma probabilidade não possível de representação em um sistema qualquer de ponto flutuante, ela deve ser representada pelo menor intervalo de máquina que a contenha.

A partir das probabilidades intervalares, em (SANTOS *et al.*, 2006) são definidas versões intervalares dos algoritmos *Forward, Backward, Viterbi* e *Baum Welch*, usados para soluções das três questões básicas dos HMMs citadas no capítulo 3.

Um exemplo de aplicação de HMMs intervalares pode ser visto em (DIMURO *et al.*, 2008) que os utiliza no reconhecimento de traços de personalidade em interações sociais.

5.5 REDES MAPAS AUTO-ORGANIZÁVEIS INTERVALARES

Sendo um tipo de rede neural com aplicações conhecidas em reconhecimento de fala, como no datilógrafo fonético em (KOHONEN, 1988), os SOM podem ser uma boa alternativa para a implementação de um reconhecedor de fala intervalar. SOM intervalares já foram utilizados com êxito para identificação e controle de sistemas não lineares (LIU *et al.*, 2008).

Liu *et al.* (LIU *et al.*, 2008), construiu SOM intervalares através do uso de entradas intervalares e pesos intervalares nas conexões entre os nós da rede. Para tal, foi definida a seguinte regra de aprendizado para atualização de um peso:

$$\overline{w_i}(t+1)=\overline{w_i}(t)+\Delta\overline{w_i}(t),$$

$$\underline{w_i}(t+1)=\underline{w_i}(t)+\Delta\underline{w_i}(t),$$

Onde:

$$\Delta \overline{w_i}(t) = \alpha_1(t) h_{ci}(t)[x - mid(w_i(t)) - rad(w_i(t))],$$

$$\Delta \underline{w_i}(t) = \alpha_2(t) h_{ci}(t)[x - mid(w_i(t)) + rad(w_i(t))].$$

Onde: $t = 0, 1, 2, \ldots n$ é a coordenada discreta do tempo, com $\alpha(t)$ iniciando uma sequência adequada monotonicamente decrescente de coeficientes de ganho intervalares, $0 < \alpha(t) < 1$ e h_{ci} é uma função de vizinhança denotando as coordenadas dos nós c e i pelos vetores r_i e r_c, respectivamente e h_{ci} uma função kernel adequada. As funções mid e rad fornecem, respectivamente, o centro do intervalo e o raio do intervalo.

Um questão interessante em relação ao SOM intervalar proposto por Liu *et al.* (2008) é que ele faz uso de uma distância entre intervalos representada por um número real, de modo que a distância não preserva informação de imprecisão. Liu *et al.* calcula a distância entre dois intervalos X e Y da seguinte forma:

$$d(X, Y) = |mid(X) - mid(Y)| + |rad(X) - rad(Y)|$$

Trindade (2009) define a distância essencialmente intervalar entre dois intervalos X e Y como:

$$d(X, Y) = [inf\{d_e(x, y) x \in X \, e \, y \in Y\}; \sup\{d_e(x, y) x \in X \, e \, y \in Y\}] \quad .$$

Onde, $d_e(x, y)$ é a distância euclidiana entre os dois pontos pertencentes aos intervalos.

Uma possível alteração no SOM intervalar de Liu *et al.* (2008), seria o uso da distância essencialmente intervalar de Trindade (2009), mas deve-se garantir preliminarmente que a substituição da distância preserve a propriedade de convergência do modelo. As provas de convergência do SOM intervalar podem ser encontradas em (LIU *et al*., 2008).

5.6 APRESENTAÇÃO DO MODELO

Uma vez apresentados as características gerais do processo de reconhecimento de fala por computadores e os modelos matemáticos utilizados neste processo que

possuem versão intervalar, segue, neste tópico a apresentação de uma proposta de modelo intervalar para reconhecimento computacional de fala.

Desta forma, serão apresentados a seguir os métodos de análise intervalar que podem ser aplicados em cada fase do processo de reconhecimento de fala. A figura 5.1 relaciona os modelos matemáticos com cada fase do reconhecimento.

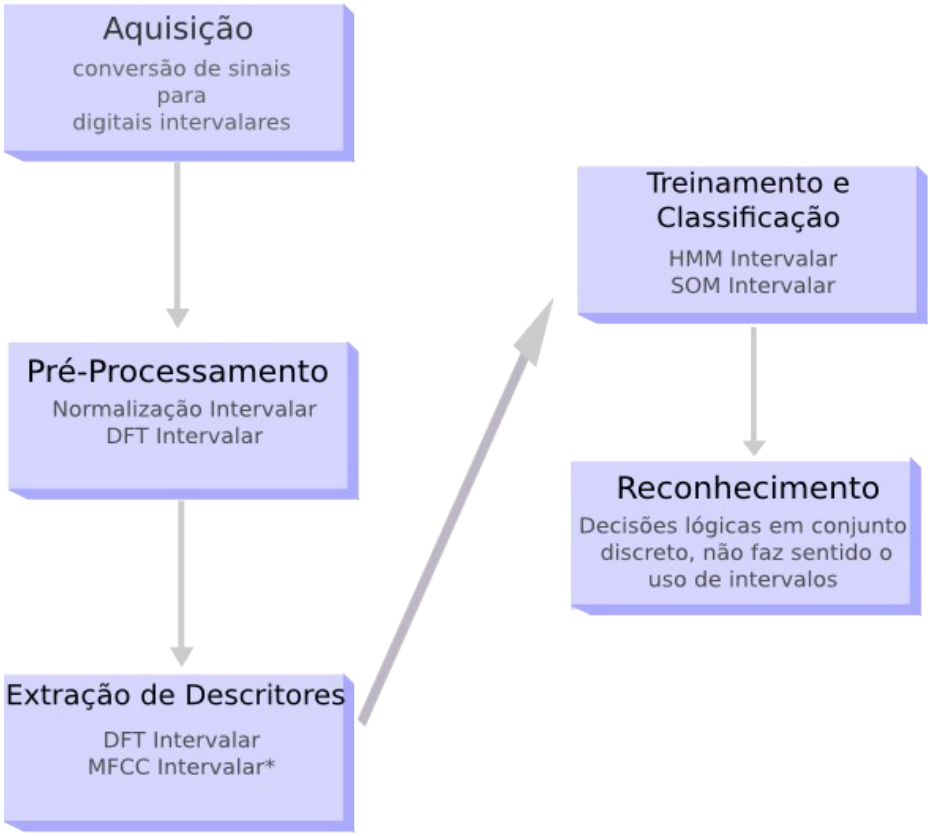

Figura 5.1: Modelo Intervalar Proposto

** A implementação de um MFCC intervalar envolveria, entre outras coisas o desenvolvimento de uma transformada cosseno intervalar.*

De acordo com a relação sinal/ruído e a taxa de amostragem presente nos sistemas de **aquisição** da fala normalmente utilizados em computadores domésticos tradicionais, combinados com fundamentação empírica, pode ser definido um nível médio de imprecisão, que servirá de parâmetro para conversão dos sinais digitais de ponto flutuante em sinais digitais intervalares.

Na fase de **pré-processamento,** podem ser usados a **normalização intervalar** (apresentada no tópico 5.2), e a conversão do sinal para análise no domínio da frequência com a **DFT intervalar**.

Após o pré-processamento, a **extração de descritores** pode ser feita utilizando pistas acústicas através de análises no domínio da frequência com a **DFT**. Apesar de não terem sido encontrados trabalhos que discutam uma versão intervalar da MFCC, a **DFT intervalar** é um primeiro passo para implementação da mesma.

Durante as fases de **treinamento e classificação** podem ser usados **HMMs intervalares** e **SOM intervalares,** respectivamente para análise temporal da fala e no domínio da frequência.

Após a classificação pelos **reconhecedores**, o conjunto de resultados é discreto, não justificando o uso de intervalos na fase final do reconhecimento. Através do reconhecimento hierárquico, como em (BRESOLIN, 2008), pode-se identificar as características de cada fonema e localizá-lo no IPA. A partir da identificação dos fonemas no IPA, pode-se realizar a escrita fonética do som recebido, e daí serem aplicadas técnicas de PLN para determinar a sentença dita.

6. CONSIDERAÇÕES FINAIS

Este trabalho apresentou os conceitos importantes em torno do reconhecimento computacional de fala e sobre a análise intervalar, relacionando-os de modo a propor um modelo intervalar para este processo.

Apesar de entrar em detalhes matemáticos do modelo, apenas uma discussão inicial sobre o assunto é apresentada e muitos detalhes sobre os modelos, principalmente sobre sua implementação computacional ainda devem ser estudados e definidos.

Não foram encontrados análogos intervalares para alguns métodos computacionais utilizados no processamento digital da fala, por exemplo Transformadas Wavelets e filtros de pré-ênfase. O desenvolvimento de versões intervalares destes métodos pode aprimorar o modelo proposto.

Entre esses métodos, que carecem de desenvolvimento de análogos intervalares podemos destacar:

- Transformadas Wavelets;
- Transformada Cosseno;
- Com a Transformada Cosseno, a MFCC;
- Truncagem.

Algumas dificuldades foram encontradas no decorrer do trabalho, principalmente em relação aos trabalhos em torno de implementações intervalares em processamento de sinais. Poucos trabalhos tem foco em implementações computacionais e mesmo os que discutem tal problemática normalmente não apresentam detalhes suficientes para que se possa dar continuidade ao que foi pesquisado.

Além disso, alguns trabalhos que possivelmente contribuiriam de maneira

significante com a pesquisa tem difícil acesso, ou o restringem ao pagamento de importâncias consideráveis para uma pesquisa não financiada. De modo que modelos importantes de métodos intervalares para processamento de sinais podem ter sidos desenvolvidos mas não se tornaram amplamente conhecidos por conta de tais limitações.

Por fim, este trabalho apenas apresenta uma sugestão de modelo teórico para processamento da fala como sinal digital intervalar, necessitando ainda ser refinado e analisado, principalmente quanto ao custo computacional antes de ser implementado. Apesar disso, a discussão apresentada no trabalho fornece os primeiros parâmetros para o processamento da fala como sinal digital intervalar, servindo de referência para próximas pesquisas que venham a ser realizadas.

7. REFERÊNCIAS

ALMPANIDIS, George; KOTROPOULOS, Constantine. Phonemic segmentation using the generalized gamma distribution and small sample bayesian information criterion. **Speech communication 50**. p.33-55. 2008.

BALLOU, Glen. Microphones. In: BALLOU, Glen (ed). **Handbook for sound engineers:** The new audio encyclopedia. Indianapolis: Howard W. Sans & Company, 1987.

BRAGA, Antônio de Pádua; LUDERMIR, Teresa Bernarda; CARVALHO, André C. P. de Leon Ferreira. **Redes Neurais Artificiais:** teoria e aplicações. Editora LTC: Rio de Janeiro, 2000.

BRESOLIN, Adriano de Andrade. **Reconhecimento de voz através de unidades menores do que a palavra, utilizando *Wavelet Packet* e SVM, em uma nova Estrutura Hierárquica de Decisão.** Tese (Doutorado em Engenharia Elétrica) - UFRN. Natal. 2008.

CAMPOS, M. A., DIMURO, G. P., COSTA, A. C. R., ARAÚJO, J. F. F., DIAS, A. M.. Probabilidade Intervalar e Cadeias de Markov Intervalares no Maple . **Tendências em Matemática Aplicada e Computacional, 3.** Nº 2. p. 53-62. 2002.

CLAUDIO, D. e MARINS, J. **Cálculo numérico computacional: teoria e prática**. 2ª Edição. Editora Atlas. São Paulo. 1994.

CHING, Wai-Ki e NG, Michael K.. **Markov Chains: models, algorithms and applications**. Springer. New York. 2006.

DIMURO, G. P., COSTA, A. C. R., GONÇALVES, L. V. HÜBNER, A.. Interval-Valued Hidden Markov Models for Recognizing Personality Traits in Social Exchanges in Open Multiagent Systems . **TEMA: Tendências de Matemática Aplicada e Computacional, 9.** Nº 1. p. 83-93. 2008.

FRY, D. B. **The physics of speech**. Cambridge. New York. 1979.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 4ª Edição. Editora Atlas. São Paulo, 2002.

GOH, Y. H.; RAVEENDRAN, P. Phoneme segmentation of speech signal. **International Conference for Technical Postgraduates (TECHPOS)**. p.1-3. 2009.

HAYKIN, Simon. **Redes neurais:** Princípios e prática. Tradução de Paulo Martins Engel. 2. ed. Porto Alegre. Bookman, 2001.

JURAFSKY, Daniel e MARTIN, James H. **Speech and Language Processing: an introduction to natural language processing, computational linguistics, and speech recognition.** New Jersey. Prentice Hall. 2000.

JUANG, B. H. e RABINER, L. R.. Hidden Markov models for speech recognition. **Technometrics**, v. 33, n. 3, pp. 251-272. 1991.

KENT, Ray. D. e READ, Charles. **The acoustic analysis of speech**. Singular Thomson Learning. 2ª Edição. Madinson. 2001.

LADEFOGED, Peter. **Elements of acoustic phonetics**. 2ª Edição. The University of Chicago Press. Chicago. 1996.

LIU, L., XIAO, J., e YU, L. Interval Self-Organizing Maps for Nonlinear System Identification an Control. **Proceedings of the 5th international symposium on Neural Networks: Advances in Neural Networks.** Spring-Verlag. Berlin. 2008.

LORDELO, Alfredo D. S.; FERREIRA, Paulo A. V.. Análise intervalar e projeto de controladores robustos via programação alvo. **Sba Controle & Automação**, Campinas,v. 16, n. 2, Junho de 2005.

KOHONEN, Teuvo. The 'Neural' phonetic typewriter". **Journal Computer**, v. 21, n. 3, pp.11-22 . IEEE Computer Society. California. 1988.

KLIR, George. J. and YUAN, Bo. **Fuzzy sets and fuzzy logic: theory and applications**. Prentice Hall. New Jersey. 1995.

MAHONY, Shaum, HENDRIX, David, SMITH, Terry J. e GOLDEN, Aaron. Self-Organizing Maps of Weight Matrix for Motif Discovery in Biological Sciences. **Artificial Intelligence Review**. Springer. 2005.

MASSINI-CAGLIARI, Gladis; CAGLIARI, Luiz Carlos. **Fonética**. In: MUSSALIM, Fernanda; BENTES, Anna Christina. **Introdução à Lingüistica: Domínios e Fronteiras**. São Paulo: Cortez, 2001.

MILLS, Patrick M. **Fuzzy speech recognition**. Tese (Mestrado em Engenharia Elétrica) – University of South Carolina. South Carolina. 1996.

MOORE, R. M. **Methods and applications of interval analysis**. Siam. Philadelphia. 1979.

PICKETT, J. M. **The acoustics of speech communication: fundamentals, speech perception theory, and technology**. Allyn and Bacon. Boston. 1998.

RABBINER, Lawrence R. e SCHAFER, Ronald W.. **Digital Processing of Speech Signals**. Prentice Hall. New Jersey. 1978.

RABINER L. R.. A tutorial on Hidden Markov Models and selected applications on speech recognition. **Proceedings of the IEEE**. v. 77. p. 257-286. 1989.

RABINER, L. e JUANG, B.. **Fundamentals of Speech Recognition**. Prentice

Hall. 1993.

ROZMAN, Robert; KODEC, Dusan M. Using asymmetric window in speech recognition. **Speech communication 49.** p.268-276. 2007.

RUSSEL, Stuart J. e NORVIG Peter. **Artificial intelligence: A modern approach**. Prentice Hall. New Jersey. 1995.

SANTOS, A. V., DIMURO, G. P., BARBOSA, L. V., COSTA, A. C. R. REISER, R. H. S., CAMPOS, M.A. Probabilidades Intervalares em Modelos Ocultos de Markov. **TEMA :Tendencias de Matemática Aplicada e Computacional**. **7,** Nº 2. p.361-370. 2006.

STRANNEBY, Dag. **Digital Signal Processing: DSP and Applications**. Oxford. Oxford. 2001.

TEBELSKIS, Joe. **Speech recognition using neural networks**. Tese (Doutorado em Ciência da Computação) – Canegie Mellon University. Pittsburgh. 1995.

TRINDADE, Roque Mendes Prado. **Uma fundamentação matemática para processamento digital de sinais intervalares**. Tese (Doutorado em Engenharia Elétrica) - UFRN. Natal. 2009.