UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Reconhecimento Automático de Voz

por

RICARDO FELIPE CUSTÓDIO

Exame de Qualificação

EQ - 22 CPGCC-UFRGS Março de 1998

Prof. Dante Augusto Coute Barone
Orientador

Prof. Waldir Leite Roque
Co-orientador

Porto Alegre, agosto de 1998

CIP - CATALOGAÇÃO NA PUBLICAÇÃO

Custódio, Ricardo Felipe
Reconhecimento Automático de Voz: exame de qualificação/ por Ricardo Felipe Custódio - Porto Alegre: CPGCC da UFRGS, 1998.
70f.: il. (EQ-22 CPGCC - UFRGS 1998).

Trabalho orientado pelo prof. Dr. Dante Augusto Couto Barone e co-orientado pelo prof. Dr. Waldir Leite Roque
1. Reconhecimento automático de voz. 2. Processamento de Sinal. 3. Dimensão Fractal. 4. Expoentes de Lyapunov. 5. Caos determinístico. 6. Séries temporais I. Barone, Dante Augusto Couto. II. Título. III Série.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitora: Profa. Wrana Panizzi
Pró-Reitor de Pós-Graduação: Prof. José Carlos Hennemann Ferraz
Diretor do Instituto de Informática: Prof. Philippe Olivier Alexandre Navaux
Coordenadora do CPCGG: Profa. Carla Maria Dal Sasso Freitas

Bibliotecária-Chefe do Instituto de Informática: Beatriz Regina Bastos Haro

Sumário

Lista de Abreviaturas

ACC	Algoritmos de Contagem de Caixas
ARPA	Advanced Research Projects
CEM	Método do Expoente Crítico (*Critical Method Exponent*)
CMN	*Cepstral Mean Normalization*
CMU	Carnegie Mellow University
CPGCC	Curso de Pós-Graduação em Ciência da Computação
CSLU	Center for Spoken Language Understanding
DARPA	Defense Advanced Research Projects Agency
DFDT	Dimensão Fractal Dependente do Tempo
DSP	Processador Digital de Sinais (*Digital Signal Processor*)
DTW	Alinhamento Temporal Dinâmico (*Dynamic Time Warping*)
EEG	Eletroencefalograma
EKS	Entropia de Kolmogorov-Sinai
EM	Máxima Estimativa (*Estimate Maximize*)
FFT	Transformada Rápida de Fourier (*Fast Fourier Transform*)
HMM	Modelos Escondidos de Markov (*Hidden Markov Models*)
IBM	Industry Business Machine
LDA	Análise Linear de Discriminante (*Linear Discriminant Analysis*)
LPC	Codificação Linear Preditiva (*Linear Predictive Coding*)
MIC	Método da Integral de Correlação
MIT	Massachusetts Institute Technology
MMSE	*Minimum Mean Square Error*
MS	Método Singular
MVP	Método dos Vizinhos Próximos
NIST	National Institute of Standards and Technology
NSF	National Science Foundation
OGI	Oregon Graduated Institute
PCA	Análise do Componente Principal (*Principal Component Analysis)*
PLP	Predição Linear Perceptiva
RAV	Reconhecimento Automático de Voz
RPS	*root power sum*
RSR	Relação Sinal Ruído
SAST	Sistema de Análise de Séries Temporais
SRI	Stanford Research Institute
STFT	*Short Time Fourier Transform*
SUR	Speech Understanding Research
TI	Texas Instruments
TIMIT	Texas Instruments & Massachusetts Institute Technology
TMM	*Tied-Mixture Model*
VQ	Quantização Vetorial (*Vector Quantization*)

Lista de Figuras

Lista de Tabelas

Proposta

Abrangência: Reconhecimento Automático de Voz
Profundidade: Caracterização de Padrões Sonoros com Dimensão Fractal e Expoentes de Lyapunov

CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Aluno: Ricardo Felipe Custódio

Proposta de Exame de Qualificação

Descrição da Área de Abrangência: Reconhecimento Automático de Voz é o processo de converter um sinal acústico, capturado por um microfone ou um telefone, para um conjunto de palavras. O produto deste processo, ou seja, a palavra reconhecida, pode ser utilizado para enviar comandos e/ou dados a um computador. Em geral, um sistema de reconhecimento de voz pode ser qualificado por sua efetividade em: a) Separar palavras: consiste no grau de isolamento entre palavras necessário para o seu reconhecimento; b) Dependência do Locutor: consiste no grau para o qual o sistema é restrito a um locutor; c) Técnicas de Comparação: consiste no método utilizado para comparar uma palavra a um vocabulário de palavras conhecidas; e d) Tamanho do vocabulário: o número de palavras que pode ser reconhecido.
O reconhecimento de Voz é um problema difícil, principalmente devido às muitas fontes de variabilidade associadas com o sinal. Em primeiro lugar, a realização acústica de um fonema, a menor unidade de som do qual a palavra é composta, é altamente dependente do contexto do qual ela aparece. Segundo, estas variabilidades podem resultar de trocas no ambiente ou na posição e características do captador. Terceiro, podem existir variabilidades para um mesmo locutor dependendo de seu estado físico e emocional, taxa de produção sonora e qualidade da voz. Por último, diferenças sociolinguísticas, dialeto e tamanho e forma do trato vocal podem contribuir para esta variabilidade. Desta forma, será feito um estudo sobre as técnicas utilizadas para modelar estas variabilidades que buscam melhorar a eficiência dos sistemas de reconhecimento de voz.

Descrição do Tema de Profundidade: Ao modelar-se as fontes das variabilidades, procuram-se formas de caracterizar os diversos padrões sonoros de um sinal de voz. Recentemente, com o avanço da teoria de sistemas dinâmicos não-lineares (Caos), constatou-se que poder-se-ia utilizar a dimensão fractal e espectro Lyapunov na caracterização de séries temporais. Esta caracterização consiste na reconstrução do atrator possivelmente associado aos diversos padrões sonoros. Desta forma, estudar-se-á métodos e algoritmos que sejam adequados às particularidades de um sinal sonoro, tais como tamanho da palavra de quantização e pequeno número de amostras.

Data: 01/08/1997

Resumo

Realizou-se um estudo sobre o reconhecimento automático de voz - RAV como requisito de exame de qualificação em abrangência e a caracterização de padrões sonoros com dimensão fractal e expoentes de Lyapunov como requisito de exame em profundidade para doutoramento, no Curso de Pós-Graduação em Ciência da Computação (CPGCC) da Universidade Federal do Rio Grande do SUL (UFRGS). O tema de abrangência é uma revisão das principais tecnologias envolvidas, métodos empregados e desafios futuros no aprimoramento dos sistemas de RAV. O tema de profundidade é um estudo das propriedades dinâmicas envolvidas no processo de produção da voz pelo aparelho fonador humano. A geometria fractal tem revolucionado a aplicação de conceitos geométricos não euclidianos nas ciências naturais [MAN 82, FED 88] e estruturas auto-similares, antes consideradas extremamente complexas para serem caracterizadas, tem sido alvo do conceito de dimensão fractal. Há evidências que o sinal de voz é caótico e portanto pode ser modelado por um sistema de equação diferenciais. Em particular, é dado ênfase ao estudo de métodos que permitam estimar algumas propriedades invariantes de sistemas caóticos e que podem ser utilizados na caracterização de padrões sonoros da voz. São apresentados alguns algoritmos eficientes que possibilitam estimar a dimensão fractal e os expoentes de Lyapunov.

Palavras-Chave: Reconhecimento automático de voz, processamento de sinal, dimensão fractal, expoentes de Lyapunov, caos determinístico, séries temporais

Title: "Automatic Speech Recognition"

Abstract

A study was become fulfilled on the automatic recognition of voice as requisite of qualifier exam. The works has two parts: the first one is an overview about automatic recognition of voice with emphasis in variability associated with the speech; the second one is about the characterization of sonorous standards with fractal dimension and Lyapunov exponents. The overview is an employed walk through of the main involved technologies, methods and future challenges in the improvement of the automatic speech recognition systems. The depth subject is a study of the involved dynamic properties in the process of production of the voice for the human. This subject has revolutionized the application of not Euclidean geometric concepts in natural sciences [MAN 82, FED 88] and structures auto-similar, before considered extremely complex to be characterized, have been target of the concept of fractal dimension. It has evidences that the voice sign is chaotic and therefore can be shaped by a distinguishing system of equation. In particular, emphasis is given to the study of methods that allow estimating

some invariant properties of chaotic systems and that they can be used in the characterization of sonorous standards of the voice. Some efficient algorithms are presented that they make possible to estimate the fractal dimension and the Lyapunov exponents.

Keywords: Speech recognition, signal processing, fractal dimension, Lyapunov exponents, deterministic chaos, time series

1 Introdução

O presente trabalho foi desenvolvido para atender aos requisitos de exame de qualificação em abrangência e profundidade para doutoramento no Curso de Pós-Graduação em Ciência da Computação (CPGCC) da Universidade Federal do Rio Grande do SUL (UFRGS).

Para atender ao **tema de abrangência**, realizou-se um estudo sobre o reconhecimento automático de voz - RAV, cujo principal objetivo é melhorar a interface do homem com a máquina, área que tem sido explorada por engenheiros e cientistas nas últimas cinco décadas. A habilidade de conversar livremente com uma máquina representa um dos últimos desafios para o entendimento do processo de produção e percepção envolvido na comunicação falada humana. Em um futuro próximo, redes interativas proverão acesso a informação e serviços que afetarão fundamentalmente como as pessoas trabalham, divertem-se ou conduzem seus afazeres diários. Hoje, tais redes são limitadas às pessoas que podem ler e tenham acesso a computadores, uma pequena parcela da população mesmo nos países mais desenvolvidos.

O resultado deste estudo consiste em uma revisão das principais tecnologias envolvidas, métodos empregados e desafios futuros no aprimoramento dos sistemas de RAV.

É útil entender que o RAV está dentro do campo processamento de voz. Pode-se distinguir três principais áreas neste campo: codificação, reconhecimento e síntese. A Fig. 1.1 esquematiza o relacionamento entre estas áreas.

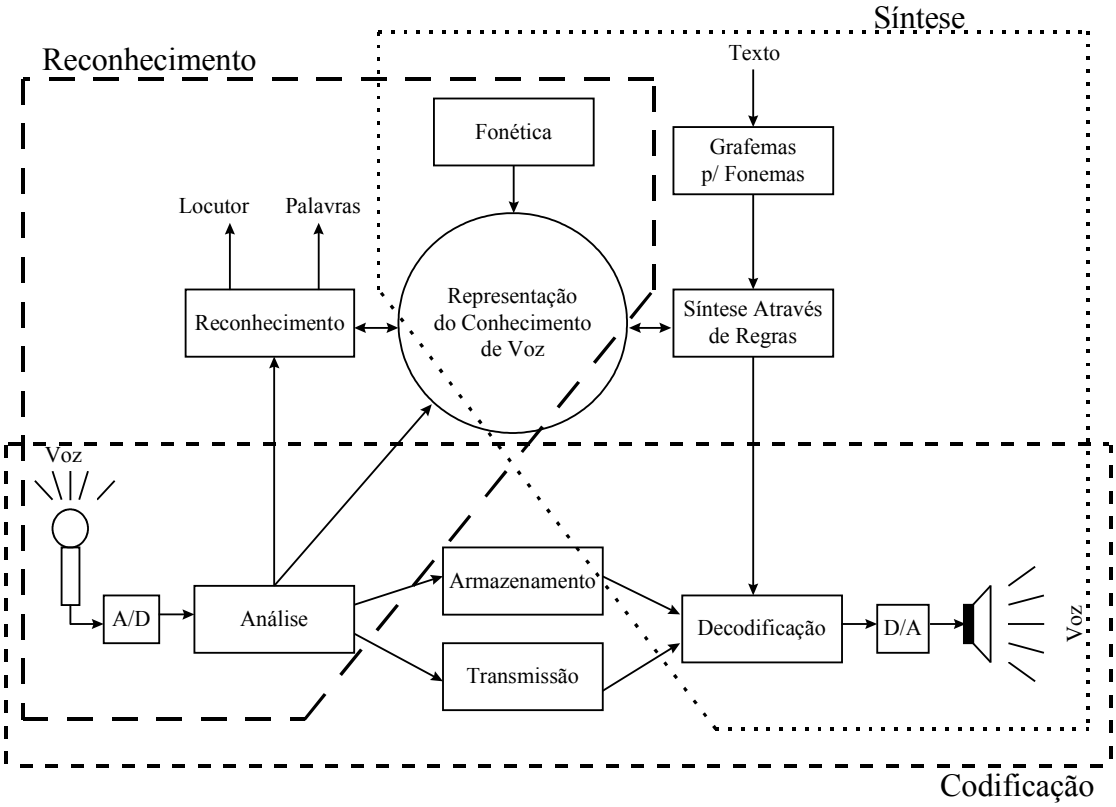


FIGURA 1.1 - Diferentes áreas no RAV

Pode-se observar, na Fig 1.1, que existe uma interseção entre estas três áreas do processamento de sinais de voz. De forma sucinta, a codificação é responsável pela representação, o mais compacta possível, do sinal de voz. O reconhecimento trata da conversão desta representação para comandos e/ou palavras. E a síntese é responsável pela geração da voz.

O RAV envolve muitas tecnologias e aplicações diferentes, tais como:

a) Conversão de um sinal acústico para um conjunto de palavras;
b) Entendimento do conteúdo lingüístico;
c) Reconhecimento do locutor:
    - identificar um locutor específico de um população conhecida;
    - verificar a identidade reclamada por um locutor;
d) Identificar a língua que está sendo falada;
e) Efetuar ligações telefônicas através de comandos falados, tal com: “ligue para minha casa”;
f) Entrada de dados simples, tal como entrar o número do cartão de crédito;
g) Preparação de documentos estruturados, tal como o resultado de um exame de radiologia;
h) Localização física de um locutor;
i) Regionalismos;
j) Tamanho do vocabulário;
k) Estado emocional;
l) Tradução automática entre diferentes línguas;
m) Vícios de fonética;
n) Problemas de articulação;
o) Compressão e Limpeza;
p) Separação de locutores;
q) Separação da voz de um locutor com música de fundo;
r) Número de pessoas falando;
s) Criptografia;
t) Pontuação.

Além dos exemplos de tecnologias e aplicações acima, na medida em que os sistemas evoluem e o tempo de implementação de uma nova aplicação diminui, conhecido como Desenvolvimento Rápido de Aplicações, mais tecnologias podem ser desenvolvidas e aplicadas a sistemas reais e novas aplicações surgem.

O tema de abrangência está desenvolvido ao longo da parte 1 deste estudo, sendo que o capítulo 2 trata das tecnologias envolvidas no processo de RAV, o capítulo 3 discute a robustez nos sistemas de RAV e o capítulo 4 discute os modelos escondidos de Markov, a técnica atualmente dominante em RAV.

Como **tema de profundidade** é realizado um estudo das propriedades dinâmicas envolvidas no processo de produção de voz pelo aparelho fonador humano. Há evidências que o sinal de voz é caótico e portanto pode ser modelado por um sistema de equação diferenciais. Em particular, é dada ênfase ao estudo de métodos que permitam estimar algumas propriedades invariantes de sistemas caóticos e que podem ser utilizados na caracterização de padrões sonoros. As propriedades invariantes mais utilizadas são: a dimensão fractal; a entropia de Kolmogorov Sinai; e o espectro de Lyapunov.

A geometria fractal tem revolucionado a aplicação de conceitos geométricos não-euclidianos nas ciências naturais [MAN 82, FED 88, ABA 96]. Estruturas auto-similares, antes consideradas extremamente complexas para serem caracterizadas, tem sido alvo do conceito de dimensão fractal. Séries temporais oriundas da digitalização de sinais como os do eletroencefalograma, voz, ruído térmico, música e outros, tem sido caracterizados e modelados por este novo conceito. Do ponto de vista qualitativo o sinal pode ser: complexo, mas periódico; composição de sinais periódicos ou quase-periódicos; caos determinístico; e processo estocástico.

Os sinais periódicos ou quase-periódicos podem ser analisados através de métodos tradicionais de análise de sinais experimentais: espectro de freqüência, através da transformada de Fourier e função de autocorrelação, entre outras ferramentas. Historicamente, o sinal de voz é estudado em segmentos. Cada segmento é considerado estacionário e periódico. Assim, pode-se utilizar a transformada de Fourier na análise desses sinais. Tratando-se de caos determinístico pode-se descrever a dinâmica por meio de um sistema de equações diferenciais. E sinais estocásticos devem ser tratados através de distribuições de probabilidade, pois estão associados a um número muito grande de variáveis. A distinção entre processos estocásticos e determinísticos vem sendo feita, com algum sucesso, por meio de técnicas ligadas aos conceitos de dimensão fractal e entropia de Kolmogorov-Sinai.

Neste trabalho é feito um estudo de algumas técnicas que permitem caracterizar as séries temporais que tem algum grau de caos determinístico. Procura-se discutir a utilização das propriedades fractais em particular na caracterização de sinais de voz. São apresentados alguns algoritmos que possibilitam estimar a dimensão fractal. Além disso, é apresentado o método do expoente crítico utilizado na estimativa da dimensão fractal dependente do tempo.

Esta temática está desenvolvida na parte II do trabalho, organizada na forma dos seguintes capítulos:

Capítulo 5: Revisão das principais técnicas utilizadas na reconstrução da dinâmica de séries temporais experimentais com comportamento caótico;
Capítulo 6: Descrição do conceito de dimensão fractal e apresentação de alguns dos métodos que permitem estimar esta dimensão a partir de sinais de voz;
Capítulo 7: Apresentação do expoente de Lyapunov como uma das medidas invariantes que podem ser utilizadas na caracterização de sinais de voz;
Capítulo 8: Discussão da entropia de Kolmogorov, outra medida invariante, utilizada na verificação do caos em sinais de voz;
Capítulo 9: Estudo de caso. Aplicação do método do expoente críticos a sinais de voz

No último capítulo, o das conclusões, procura justificar o interesse nesta área emergente, como uma alternativa para a caracterização de sinais de voz e desta forma uma alternativa interessante ao processo de RAV.

Finalmente, como incentivo à busca de referências sobre este tão instigante assunto, apresenta-se uma lista de bibliografias. O objetivo da lista não foi exaurir o que há na literatura, mas apresentar algumas fontes que poderão enriquecer e talvez elucidar dúvidas que poderão surgir na leitura deste trabalho.

Parte I: Exame de Qualificação em Abrangência

2 Reconhecimento Automático de Voz - RAV

2.1 Introdução

O RAV é o processo de converter um sinal acústico, capturado por um microfone ou um telefone, para um conjunto de palavras. O produto deste processo, ou seja, a palavra reconhecida, pode ser utilizado para enviar comandos e/ou dados a um computador. Em geral, um sistema de reconhecimento de voz pode ser qualificado por sua efetividade em: a) Separar palavras: consiste no grau de distanciamento entre palavras necessário para o seu reconhecimento; b) Dependência do Locutor: consiste no grau para o qual o sistema é restrito a um locutor; c) Técnicas de Comparação: consiste no método utilizado para comparar uma palavra a um vocabulário de palavras conhecidas; e d) Tamanho do vocabulário: respectivo número de palavras que pode ser reconhecido.

O reconhecimento de voz é um problema difícil, principalmente devido às muitas fontes de variabilidade associadas com o sinal. Em primeiro lugar, a realização acústica de um fonema, a menor unidade de som do qual a palavra é composta, é altamente dependente do contexto no qual ela aparece. Em segundo lugar, estas variabilidades podem resultar de trocas no ambiente ou na posição e características do microfone. Em terceiro lugar, podem existir variabilidades para um mesmo locutor dependendo de seu estado físico e emocional, taxa de produção sonora e qualidade da voz. Por último, diferenças sociolinguísticas, dialeto e tamanho e forma do trato vocal podem contribuir para esta variabilidade.

Existem muitas frentes de pesquisa na área de RAV. Uma classificação útil dessas frentes é: produção, percepção e caracterização acústica-fonética do sinal de voz; processamento digital de sinais e métodos de análise para reconhecimento de voz; técnicas de reconhecimento de padrões; projeto e implementação de sistemas de conhecimento de voz; e teoria e implementação de modelos escondidos de Markov.

As tecnologias empregadas no RAV podem ser classificadas em três grandes estratégias ou grupos: fonética-acústica; reconhecimento de padrões; e inteligência artificial. A Tab. 2-1 descreve sucintamente cada um destes grandes grupos.

A Tab. 2.2 mostra um resumo da tecnologia existente e projeções futuras em RAV. A tecnologia básica que está sendo utilizada e provavelmente permanecerá são os modelos escondidos de Markov (HMM). Neste trabalho dar-se-á ênfase a esta tecnologia. Cabe ressaltar que redes neurais tem sido utilizado no reconhecimento isolado de palavras e no auxílio à resolução de pequenas tarefas no reconhecimento contínuo com HMM. Veja [BEM 96, FRI 96, KOS 92, LIP 89] para maiores informações sobre a utilização de redes neurais no reconhecimento automático de voz.

Na seção 2.2 é feito um breve histórico dos principais acontecimentos que marcaram época em RAV. Na seção 2.3 discute-se em detalhes o que é um sistema de RAV e seus principais componentes, com a apresentação das tecnologias envolvidas. Uma vez descrito o que é um sistema de RAV, apresenta-se, na seção 2.4, os principais parâmetros utilizados na caracterização desses sistemas. Na pesquisa de novas tecnologias, no desenvolvimento de novas aplicações e no uso propriamente dito dos sistemas de RAV são necessárias bases de dados com amostras de voz. Essas bases são conhecidas como CORPORA. Alguns dos mais importantes corporas existentes são apresentados na seção 2.5. A seção 2.6 apresenta como podem ser modeladas as diferentes fontes de variabilidade presentes no sinal de voz. Na seção 2.7 discute

sucintamente o estado da arte em termos de tecnologias e sistemas existentes. Finalmente na seção 2.8 apresenta-se expectativas futuras nesta área.

TABELA 2.1 - Estratégias de reconhecimento automático de voz

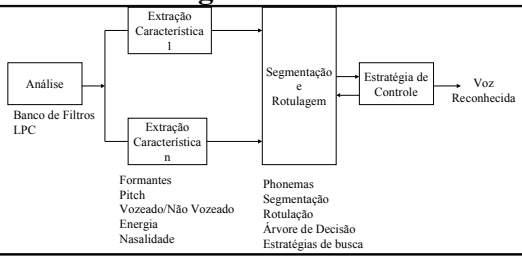
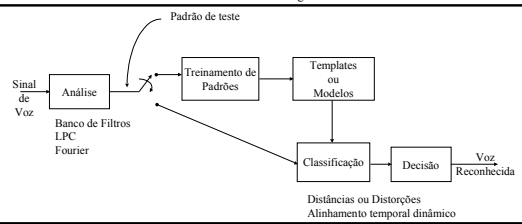
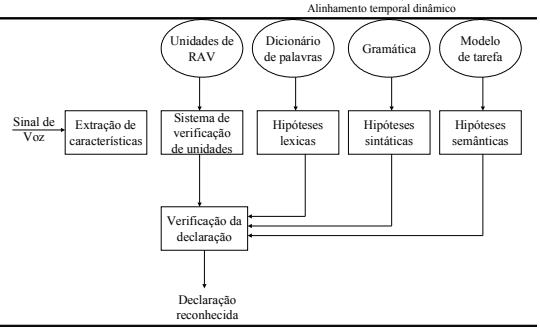
Tecnologia	Descrição	Diagrama básico
Fonética-acústica	Postula que existem unidades fonéticas finitas e distintas e que podem ser caracterizadas através do seu sinal ou espectro no tempo	
Reconhecimento de padrões	Os padrões de voz são utilizados diretamente sem determinação explícita ou segmentação como é na fonética acústica	
Inteligência artificial	Aplica técnicas de IA a idéias e conceitos das abordagens fonética-acústica e reconhecimento de padrões	

TABELA 2.2 - Projeções para reconhecimento de voz

Ano		1900 a 1998	1998 a 2003	Além de 2003
Capacidade de Reconhecimento	Palavras conectadas	Voz Contínua; Modelos de palavra inteira; Gramáticas de estado finito; tarefas restritas	Voz contínua; Reconhecimento de subpalavras; Modelos de linguagem representativo do inglês natural; Semântica de tarefa específica	Voz contínua; gramática, sintaxe e semântica de linguagem natural; Aprendizagem e adaptação neural
Tamanho do vocabulário	10 a 30	100-10.000	20.000 a 100.000	Irrestrito

Tecnologia básica	Templates; Redes Neurais	HMM	HMM	HMM
Requisitos de Processamento	2-4 DSPs (25 Mips[1]/Chip)	4-10 DSPs (50 Mips/Chip)	5-50 DSPs (200 Mips/Chip)	20-60 DSPs(1.000 Mips/Chip)
Aplicações	Discagem através de voz; Saldo bancário	Transações bancárias; Controle de robôs;	Máquinas de ditado; Acesso a base de dados;	Interações em linguagem natural; Tradução telefônica;

2.2 Histórico

A Tab. 2.3 apresenta os principais acontecimentos que marcaram o desenvolvimento da área de RAV. Dentre todos estes acontecimentos, dois foram decisivos no surgimento de sistemas com aplicabilidade real: a) O projeto da agência norte-americana ARPA na década de 1970 [KLA 77] e b) o uso dos modelos escondidos de Markov [MAR 96].

TABELA 2.3 - Histórico

Década	**Acontecimento**
1870	• Alexander Graham Bell queria construir uma máquina com o objetivo de fazer a voz visível para os deficientes auditivos. Seu trabalho culminou com o desenvolvimento do telefone;
1950	• Foi construído o primeiro sistema de reconhecimento automático de dígitos nos laboratórios AT&T Bell, Estados Unidos. O sistema consistia na comparação de segmentos de voz com padrões previamente armazenados. Era necessário um ajuste à voz do locutor para que o sistema funcionasse. Conseguiu-se uma taxa de reconhecimento de 99%. Nesta mesma época, os cientistas acordaram que o reconhecimento era uma tarefa muito complexa e estabeleceram metas menos audaciosas. Surgiu assim o conceito de sistema dependente de locutor, sistema com pausa forçada entre palavras e sistema com pequeno vocabulário (menos de 50 palavras);

[1] Milhões de instruções por segundo

1960	• Incorporação de técnicas de normalização temporal com o objetivo de minimizar a variabilidade advinda das diferentes taxas de produção sonora por um mesmo locutor; • Em 1968, o computador HAL-9000, do filme 2001: Uma Odisséia no Espaço introduziu o conceito de reconhecimento automático de voz ao público em geral;

1970	• Surgiram os primeiros sistemas comerciais de RAV. O primeiro sistema foi o VIP 100 da empresa norte-americana Threshold Technology, Inc. Este sistema conseguia reconhecer um pequeno número de palavras isoladas;
	• A agência ARPA[2] lançou o projeto SUR[3] com o objetivo de estimular as pesquisas em reconhecimento automático do voz visando grandes vocabulários, voz contínua e independente de locutor. O projeto iniciou em 1971 e terminou em 1976, com as seguintes características:
	  • Um vocabulário de 1000 palavras ou mais;
	  • Voz conectada;
	  • A voz de vários locutores;
	  • Uma aplicação realista;
	  • Pequeno tempo de reconhecimento;
	  • Taxa de erro menor que 10%.
	• Apesar de várias instituições, tais como a IBM, CMU[4] e MIT[5], terem participado deste projeto, somente o sistema desenvolvido pela CMU reuniu todos os requisitos do SUR. O sistema, denominado *Harpy*, foi capaz de reconhecer 1.011 palavras com uma taxa de erro de 5%. Uma descrição detalhada deste sistema por ser encontrada em [KLA 77].

1980	• Uso maciço dos modelos escondidos de Markov, como ferramenta para o reconhecimento automático de voz; • Grande melhoria da imunidade a ruído dos sistemas de RAV; • Desenvolvimento de modelos de linguagem, em particular o modelo *N-gram* da IBM; • Em 1986, a empresa norte-americana Speech Systems, Inc. lançou o primeiro sistema comercial capaz de reconhecer 20.000 palavras, voz contínua e independente do locutor; • No final da década, a empresa norte-americana Dragon Systems, Inc. lançou um sistema que se adaptava ao locutor, com vocabulário de 30.000 palavras e reconhecimento de palavras isoladas.

[2] Advanced Research Projects
[3] Speech Understanding Research
[4] Carnegie Mellow University
[5] Massachusetts Institute Technology

1990	• Surgiram circuitos integrados com a capacidade de reconhecer voz; • Consolidação dos primeiros corporas (TIMIT, CSLU e outros); • Em 1993 surgem os primeiros sistemas para computador pessoal. Dentre eles o Personal Dictation da IBM para o sistema operacional OS/2 e PlainTalk da Apple para o Macintosh. • Em 1994 a empresa norte-americana Philips Dictation Systems, Inc. colocou no mercado o primeiro sistema baseado em computador pessoal capaz de reconhecer voz contínua com um grande vocabulário; • Em 1996, o OS/2 Warp 4 é o primeiro sistema operacional a embutir comandos de voz, para navegação no sistema; • Em 1996, a IBM apresenta o Med Speak/Radiology, o primeiro produto para reconhecimento da voz contínua em tempo real; • Em junho de 1997 a empresa norte-americana Dragon Systems, Inc. lança o primeiro programa de uso geral para reconhecimento de voz contínua; • Em agosto de 1997 a IBM lança o ViaVoice; • Em setembro de 1997 Bill Gates, da Microsoft, declara que o reconhecimento de voz é a chave para o avanço tecnológico dos próximos anos.

Até o presente não se tem notícia de uma sistema de RAV para o português falado no brasil, para voz contínua conversacional.

2.3 Sistemas de RAV

A tarefa de um sistema de RAV pode ser realizada em diversos níveis como por exemplo fonemas, palavras, frases e declarações. Cada nível pode prover restrições temporais que podem resolver erros ou incertezas de níveis mais baixos. Assim é mais conveniente deixar as decisões discretas para os níveis mais elevados e nos níveis inferiores trabalhar com modelos probabilísticos.

Os principais componentes de um típico sistema de RAV, descritos a seguir, são mostrados na Fig. 2.1.

As três ferramentas mais utilizadas para a modelagem acústicas são: templates, redes neurais e modelos escondidos de Markov. As redes neurais são estruturas matemáticas não lineares que podem armazenar, através de treinamento, o modelo acústico utilizado durante o processo de reconhecimento. Embora existam alguns trabalhos recentes nesta área[6], os sistemas existentes de reconhecimento de voz utilizam basicamente modelos escondidos de Markov.

[6] ver [BEM 96]

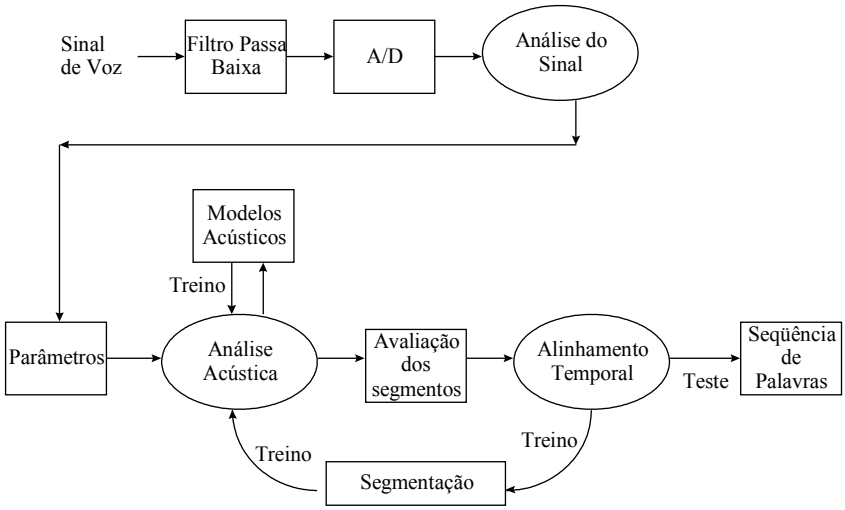


FIGURA 2.1 - Estrutura de um típico sistema de RAV

2.3.1 Sinal de Voz

A voz é produzida pelo aparelho fonador humano, transmitida através de variações de pressão do ar e recebida pelo sistema auditivo humano onde é processado o seu entendimento. Do ponto de vista de engenharia, pode-se capturar a voz em vários locais nesse caminho, conforme ilustra a Fig. 2.2. Pode-se, por exemplo, amostrar o movimento das articulações (língua, lábios, etc.). Pode-se amostrar a variação de pressão de ar através de um microfone. Pode-se realizar a leitura labial. Pode-se medir a atividade elétrica cerebral. No entanto, a forma mais utilizada é a amostragem da variação de pressão do ar, por ser a mais simples, tecnologicamente falando. O sinal de voz capturado desta forma, não é o mais indicado para a realização do seu reconhecimento. Assim, procura-se realizar uma série de transformações que visam modificar esta representação para uma outra representação o mais próxima possível daquela que o cérebro humano utiliza para efetivamente realizar o reconhecimento. A representação de segmentos de voz através da análise espectral e em particular dos coeficientes Mel Cepstrais é uma das formas mais utilizadas. A utilização deste procedimento tem mostrado, na prática, excelentes resultados [COL 95, BEM 96, JOH 97, MAR 96].

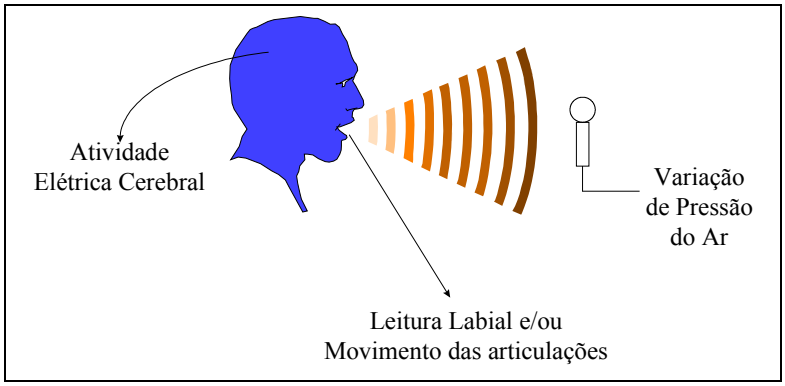


FIGURA 2.2 - Locais onde pode ser amostrado o sinal de voz

A Fig. 2.3 mostra a forma de onda da palavra “five” em inglês.

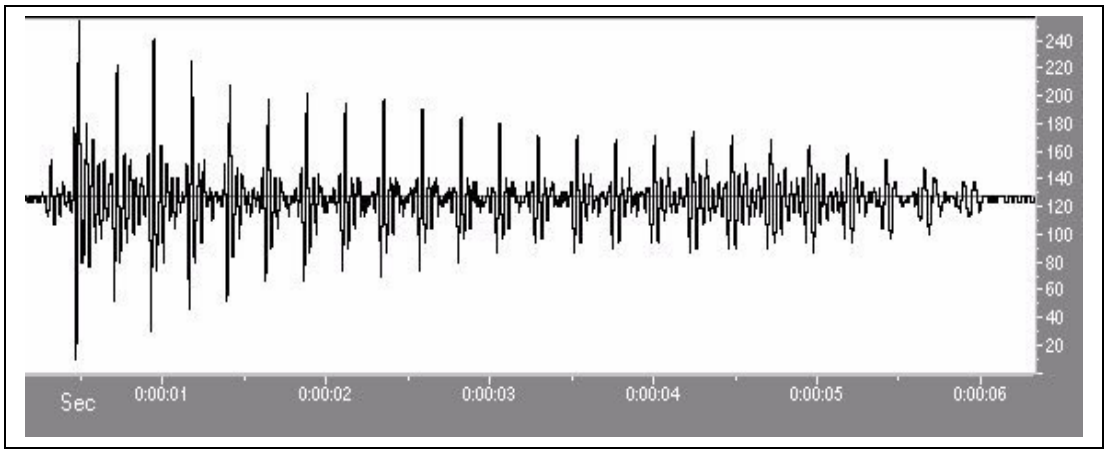


FIGURA 2.3 - Forma de onda da declaração “five” em inglês

2.3.2 Filtragem Passa Baixa

Antes do sinal de voz ser codificado pelo conversor analógico digital, é necessário filtra-lo de forma a não conter freqüências maiores que a taxa de Nyquist, metade da freqüência de amostragem, de forma a evitar a sobreposição de freqüências durante a análise espectral. O sinal de voz codificado desta forma pode ser reconstruído através de um conversor digital analógico. A Fig. 2.4 ilustra este processo: a voz que sairá do alto-falante será mesma que entra no microfone caso a freqüência de amostragem seja pelo menos o dobro da freqüência de corte do filtro passa baixa[7]. Maiores informações deste procedimento podem ser encontradas em [RAB 78].

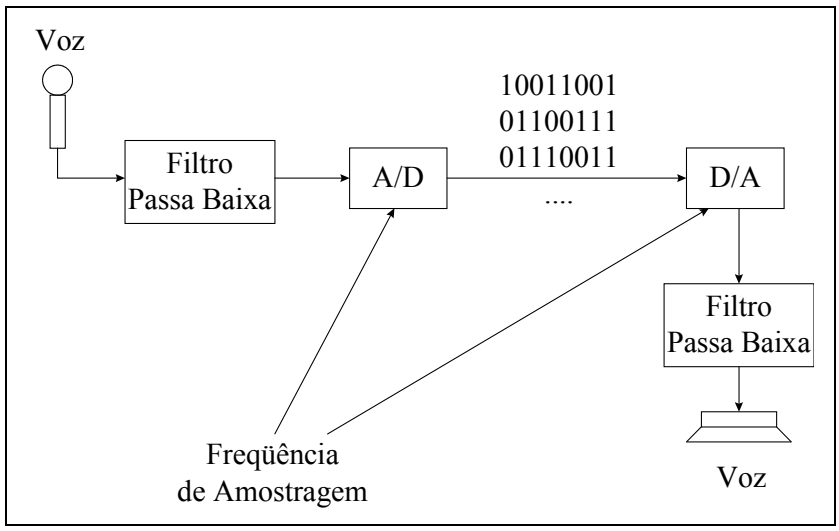


FIGURA 2.4 - Necessidade do filtro passa baixa

[7] Deve-se observar que o sinal reconstruído terá a mesma banda passante do filtro passa baixa

2.3.3 Conversão Analógico para Digital

É o responsável pela amostragem, quantização e codificação do sinal de voz analógico advindo do microfone. As principais especificações de um conversor analógico para digital são o número de bits de quantização ou tamanho da palavra de quantização e a taxa de amostragem. Para sinais de voz utiliza-se, normalmente 16 bits de quantização e taxa de amostragem entre 6kHz e 20kHz [COL 95]. Também é comum realizar-se uma codificação não linear, uma vez que a maioria das amostras de um típico sinal de voz tem pequena amplitude. Assim, é possível aproveitar melhor o espaço de memória disponível para codificação. A codificação não linear mais utilizada é a conhecida lei μ [RAB 78]. Muitas bases de dados de voz, disponíveis para a pesquisa e desenvolvimento de sistemas de processamento digital de voz, são codificados por esta lei.

A lei μ consiste em realizar-se a seguinte transformação:

$$y(n) = X_{max} \frac{\log\left[1 + \mu \frac{|x(n)|}{X_{max}}\right]}{\log[1+\mu]} sign[x(n)], \tag{2.1}$$

onde x(n) são as amostras do sinal de voz, $X_{máx}$ é o maior valor de todas as amostras, μ é a quantidade de compressão e sign(t) é a função sinal, isto é:

$$sign(t) = \begin{cases} 1 & se \quad t \geq 0 \\ -1 & se \quad t < 0 \end{cases} \tag{2.2}$$

A Fig. 2.5 mostra uma família de curvas de y(n) versos x(n) para diferentes valores de μ. Se μ=0, a equação (2.1) reduz-se a:

$$y(n) = x(n),$$

isto é, os níveis de quantização estarão uniformemente espaçados. Um estudo da relação sinal ruído para um quantizador lei-μ e de outros quantizadores não lineares pode ser encontrado em [RAB 78].

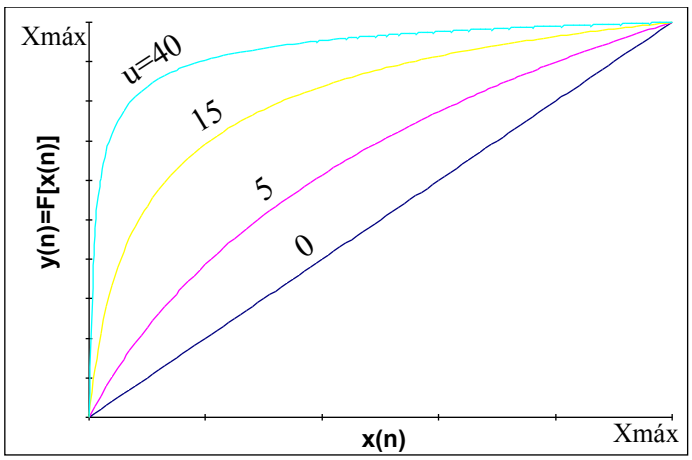


FIGURA 2.5 - Relação entre a entrada e saída para uma característica lei μ

2.3.4 Análise do Sinal

As técnicas de análise do sinal de voz normalmente extraem características úteis e as comprimem por um fator de até 10, sem perda de qualquer informação importante. Esta compressão é possível uma vez que existe uma quantidade razoável de informação redundante nos parâmetros que representam as características.

A representação, o mais compacta possível, objetiva preservar a informação necessária para determinar a identidade fonética de uma porção da voz de forma, tanto quando possível, insensível a fatores como diferentes locutores, efeitos introduzidos em canais de comunicação, estado emocional do locutor e outras variabilidades.

As representações atualmente utilizadas concentram-se primariamente sobre as propriedades do sinal de voz atribuídas à forma do trato vocal. As representações são sensíveis a sons vocálicos ou não, mas tentam ignorar efeitos devido a variações em freqüência. Essas representações quase sempre derivam do espectro de potência de tempo curto, isto é, o espectro de potência de seguimentos de 10 a 30 mSeg de sinais de voz. Este tamanho do segmento foi escolhido porque o sinal de voz pode ser considerado estacionário neste período e assim pode-se utilizar a análise espectral.

Resumindo, as técnicas mais populares são:

a) Análise espectral, realizada através da transformada de Fourier, de segmentos de voz. As freqüências são distribuídas usando a escala Mel, Bark ou outras, a qual realizam uma transformação linear nas baixas freqüências e logarítmica nas altas freqüências. Estas escalas foram inspiradas nas características fisiológicas do sistema auditivo humano. Para realizar a transformada de Fourier utiliza-se os algoritmos rápidos conhecidos como FFT (Fast Fourier Transform), os quais são extremamente eficientes com complexidade computacional na ordem de $n\log(n)$. A Fig. 2.6 mostra o espectro de potência da declaração "five" em inglês.

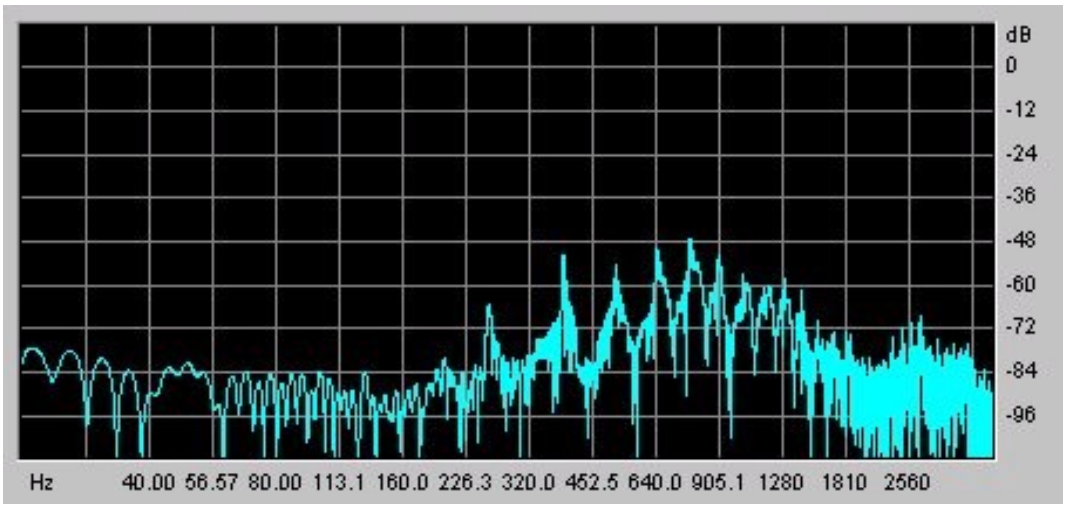


FIGURA 2.6 - Espectro de potência da declaração "five"

Uma ferramenta gráfica muito utilizada para a visualização do espectro de freqüência em função do tempo é o espectrograma. A Fig. 2.7 mostra o espectrograma da declaração "five" em inglês. O eixo das abcissas é o tempo em segundos e o das

ordenadas é a freqüência em Hz. Os tons mais escuros são onde o sinal tem maior energia e os mais claros menor energia.

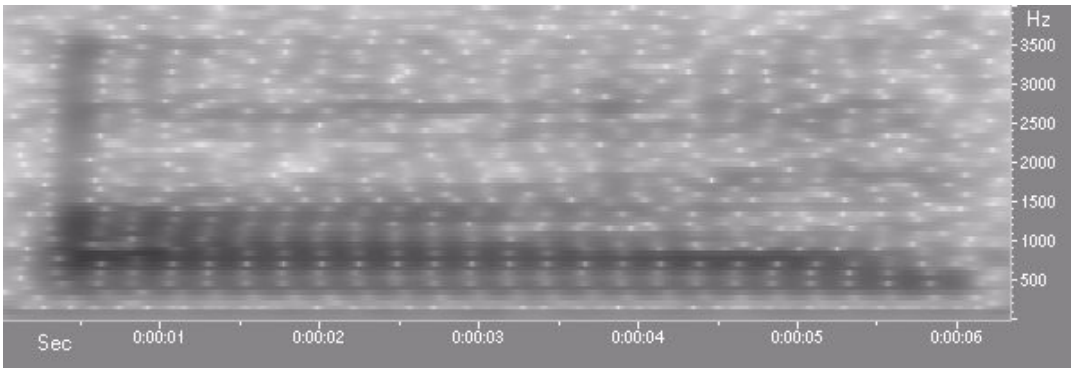


FIGURA 2.7 - Espectrograma da declaração "*five*" em inglês.

a) Predição Linear Perceptual (PLP), a qual é também fisiologicamente motivada, porém, seus coeficientes, não podem ser interpretados visualmente;

b) Codificação Linear Preditiva (LPC - Linear Predictive Coding). Esta técnica é historicamente uma das mais importantes técnicas de análise. O objetivo é estimar os coeficientes de um filtro unicamente por pólos, representando a caixa acústica do trato vocal humano. Assim, pode-se estimar a próxima amostra com base na soma ponderada de amostras passadas, ou seja:

$$\hat{s}_n = \sum_{i=1}^{p} a_i s_{n-i} \,, \tag{2.3}$$

onde s_i são as amostras do sinal de voz no tempo i e a_i são os coeficientes do filtro, cuja função transferencia é dada por:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-1}} \tag{2.4}$$

Estes coeficientes podem ser visualizados como as freqüências ressonantes desta caixa acústica realizando assim a conformação do sinal.

c) Análise cepstral, que consiste em calcular-se a transformada inversa de Fourier do espectro de potência logarítmico do sinal de voz.

Pode-se verificar que as quatro técnicas descritas acima são baseadas na representação espectral do sinal de voz. Assim, muitas pesquisas tem demonstrado que faz pouca diferença em qual técnica utilizar nos sistemas de RAV. De fato, cada uma destas técnicas tem seus defensores.

O espectro de potência é quase sempre representado na escala logarítmica. Quando o ganho aplicado ao sinal varia, a forma do espectro de potência é preservada. Filtragem linear mais complicada causada, por exemplo, pela acústica ambiental ou por variações entre linhas telefônicas, as quais aparecem como efeito modulador sobre a

forma de onda e como efeito multiplicativo no espectro de potência linear, tornam-se simples constantes aditivas no espectro de potência logarítmico.

As vantagens da escala logarítmica do espectro de potência são:

a) O sinal de excitação quase-periódico e o filtro variante no tempo que representa o trato vocal podem ser facilmente separados, uma vez que seus efeitos são aditivos no domínio do espectro de potência logarítmico;
b) Distribuições estatísticas do espectro de potência logarítmico tem propriedades convenientes para o reconhecimento.

Antes de computar o espectro de potência logarítmico, a forma de onda passa por um processo de pré ênfase através de um simples filtro passa alta com um único zero na função de transferência, dando um ganho de 6dB por oitava de forma a fazer o espectro ligeiramente plano. A equação (2.5) apresenta esse filtro.

$$s_{pre}(t) = s(t) - \alpha s(t-1) \,, \tag{2.5}$$

onde s(t) é a amostra de um sinal de voz no tempo t e α é tipicamente 0,95.

Resumindo, o espectro de potência logarítmico é assim computado:

a) Toma-se um segmento pré enfatizado, na ordem de 25 ms. de voz;
b) Aplica-se uma função janela a fim de evitar os efeitos das bordas;
c) Aplica-se a transformada de Fourier.

As funções janela mais utilizadas são mostradas na Tab. 2.4.

TABELA 2.4 - Funções janelas utilizadas na análise espectral

Janela	**Função**	**Equação**
Bartlett	$w_j = 1 - \left\| \frac{j - \frac{1}{2}N}{\frac{1}{2}N} \right\|$	(2.6)
Hamming	$w_j = 0{,}54 - 0{,}46\cos\left(\frac{2\pi j}{N}\right)$	(2.7)
Hanning	$w_j = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi j}{N}\right)\right]$	(2.8)
Welch	$w_j = 1 - \left(\frac{j - \frac{1}{2}N}{\frac{1}{2}N}\right)^2$	(2.9)

Onde N é o número de amostras do segmento.

Uma vez obtido o espectro de freqüências logarítmico, procede-se à determinação, por exemplo, dos coeficientes cepstrais. A transformada coseno, uma variação da transformada de Fourier, pode ser usada para converter eficientemente o conjunto de energias logarítmicas para um conjunto de coeficientes cepstrais. A Tab. 2.5 descreve o significado desses coeficientes.

TABELA 2.5 - Significado dos coeficientes Cepstrais

Coeficiente	**Significado**
C_0	Descreve a forma do espectro logarítmico independentemente de seu nível geral;
C_1	Mede o balanço entre as metades superior e inferior do espectro;
C_2 a C_n	São relacionados com estruturas mais finas do espectro.

O trato vocal pode ser considerado como um tubo acústico sem perdas e seu efeito no sinal da excitação é uma série de ressonâncias. Para a voz em circunstâncias acústicas favoráveis, esta é uma boa aproximação e é conhecida como codificação linear preditiva (LPC) ou modelagem autoregressiva no sentido em que os parâmetros do filtro a só pólos são ajustados ao espectro do sinal de voz, embora o espectro não necessite ser calculado explicitamente. A LPC tem problemas com certos sinais degradados e não é conveniente para produzir coeficientes cepstrais na escala Mel. A predição linear perceptiva (PLP) combina a LPC e um banco de filtros pelo ajuste de um modelo a só pólos a um conjunto de energias produzidas por um banco de filtros perceptualmente motivados e após isso é feito o cálculo dos coeficientes cepstrais.

Muitos sistemas aumentam a informação no espectro de potência de curto prazo com informação sobre sua taxa de variação no tempo. A forma mais simples de obter esta informação dinâmica seria tomar a diferença entre sucessivos segmentos. Todavia, isso pode ser muito sensível a variações entre segmentos consecutivos. Consequentemente, tendências lineares são estimadas sobre seqüências de tipicamente 5 ou 7 segmentos. Alguns sistemas vão além e estimam aceleração de características tão bem quanto taxas de variações. Estas características dinâmicas de segunda ordem necessitam de seqüências mais longas de segmentos para uma estimativa confiável.

Uma vez que os coeficientes cepstrais são grandemente descorrelacionados, um método computacionalmente eficiente para obter razoável estimativa da probabilidade consiste em calcular a distância euclidiana do vetor modelo de referência após serem os coeficientes adequadamente ponderados. Vários esquemas de ponderação tem sido usados. Um esquema empírico que trabalha bem deriva dos pesos dos primeiros 16 coeficientes da metade positiva de um ciclo da forma de onda do seno. Tem havido bons resultados a ponderação dos coeficientes cepstrais PLP pelo seu índice, conhecido como RPS[8], dando a C_0 um peso zero, etc. Maiores informações podem ser obtidas em [COL 95].

Enquanto que os coeficientes cepstrais são altamente descorrelacionados, uma técnica denominada análise do componente principal (PCA - Principal Component Analysis), pode prover uma transformação que pode remover completamente a dependência linear entre os conjuntos de variáveis. Este método pode ser usado para descorrelacionar não somente conjuntos de energia, mas também combinações de conjuntos de parâmetros tais como características dinâmicas e estáticas e parâmetros PLP e não PLP. Uma dupla aplicação da PCA com uma operação de ponderação, conhecida como análise discriminante linear (LDA), pode levar em conta a informação

[8] root power sum weighting

discriminativa necessária para distinguir entre os sons de voz para gerar um conjunto de parâmetros, algumas vezes chamados coeficientes IMELDA, adequadamente ponderados pelo cálculo de distância euclidiana. Bom desempenho tem sido reportado com um muito reduzido conjunto de coeficientes IMELDA, e existe evidencia que incorporando sinais degradados na análise pode melhorar a robustez.

A grande maioria dos sistemas comerciais de RAV usa as representações acima citadas. Com o intuito de conseguir novas tipos de representações, tem-se procurado utilizar transformadas *wavelets* e rede neurais artificiais para prover operações não lineares sobre a representação espectral logarítmica.

Na segunda parte desta monografia (exame de qualificação em profundidade) será introduzido um novo conjunto de medidas invariantes que poderão ser utilizadas para a análise do sinal de voz. Estas medidas baseiam-se no fato de que a voz é produzida por um sistema dinâmico não linear e desta forma pode ser caótica. De fato, existem alguns trabalhos publicados que comprovam a natureza caótica dos sinais de voz [NAK 93, LEV 94, SAB 95 e SAB 96].

2.3.5 Vetores de Parâmetros

O resultado da análise do sinal de voz é um conjunto de segmentos com tamanho entre 10 e 30 ms., tipicamente obtidos a intervalos de 10 ms., representados na forma de um conjunto de 10 ou 20 parâmetros.

2.3.6 Modelos Acústicos

Os modelos acústicos são utilizados para analisar o conteúdo acústico dos segmentos de voz. Existem muitos modelos acústicos. Os modelos mais populares são: *templates* e estados.

Os *templates* são simplesmente o registro de uma amostra da voz, por exemplo, uma palavra. Assim, uma palavra desconhecida pode ser reconhecida simplesmente comparando-a a um conjunto de *templates* conhecidos determinando aquele que mais se parece. Os *templates* tem muitos pontos fracos:

a) Não conseguem modelar variabilidades acústicas;
b) São limitados a modelos de palavra inteira, uma vez que na prática é muito difícil registrar ou segmentar uma amostra menor que uma palavra.

Os estados são uma representação mais flexível. O paradigma de estados mais utilizado é os modelo escondido de Markov (HMM - ***H**idden **M**arkov **M**odels*) [COL 95, BEM 96] . Um HMM é um modelo estocástico duplo na qual a geração do fonema e a sua realização acústica são ambas representadas estatisticamente como processos Markov. Nesta abordagem, cada palavra é representada por um conjunto de estados, onde cada estado representa o som mais esperado de ser escutado naquela parte da palavra. Isso é feito através da distribuição de probabilidades sobre o espaço acústico. As distribuições de probabilidade podem ser modeladas: parametricamente, assumindo que elas tem um forma simples tal como uma gaussiana; ou não parametricamente, representando estas distribuições diretamente. Isso pode ser feito através de um histograma ou através de redes neurais. Durante a fase de treinamento, os modelos acústicos são incrementalmente modificados com o intuito de melhorar o

desempenho geral do sistema de RAV. Durante a fase de teste os modelos não são alterados.

2.3.7 Análise Acústica e Avaliação dos Segmentos

A análise acústica é realizada pela aplicação do modelo acústico a cada segmento, produzindo um vetor de avaliação. Para o caso do modelo acústico *templates*, a avaliação consiste em determinar a distância euclidiana entre o segmento desconhecido e todos os outros do conjunto de segmentos conhecidos. Para os modelos acústicos baseados em estados, a avaliação consiste na determinação da probabilidade de emissão, isto é, a probabilidade do estado gerar o corrente segmento.

2.3.8 Alinhamento Temporal

As avaliações dos segmentos são convertidas para uma seqüência de palavras respeitando as restrições impostas pela trajetória dos sons no modelo acústico, obtendo-se assim um melhor caminho. Isto é necessário para a normalização temporal da duração das unidades de som da voz, uma vez que estas peças de sons são produzidas a taxas diferentes, inclusive para um mesmo locutor. O processo de determinação deste melhor caminho chama-se alinhamento temporal. As restrições podem ser nas próprias palavras e entre elas. Nas palavras, as restrições são dadas pela seqüência de possíveis segmentos no caso dos modelos *templates*, ou pela seqüência de estados no caso do modelo baseado em estados. Entre as palavras, as restrições são dadas por uma gramática, indicando quais palavras podem seguir uma determinada palavra.

O alinhamento temporal pode ser realizado eficientemente através da programação dinâmica, um algoritmo geral que usa restrições locais. Este algoritmo tem duas principais variantes: Alinhamento temporal dinâmico (DTW - Dynamic Time Warping) e o algoritmo Viterbi.

O DTW é um eficiente método para encontrar o alinhamento não linear ótimo. O algoritmo consiste em calcular os segmentos otimizados localmente do caminho de alinhamento global durante uma varredura do vetor de avaliação. Seja $D(x,y)$ a distância euclidiana entre o segmento x e y do *template* de referência. Seja também $C(x,y)$ a avaliação acumulada ao longo do caminho ótimo que leva a $D(x,y)$, ou seja:

$$C(x,y) = MIN\left(C(x-1,y), C(x-1,y-1), C(x,y-1)\right) + D(x,y) \qquad (2.10)$$

Percorre-se então todos os segmentos, iniciando em $C(0,0)$ e terminando em $C(X,Y)$. O alinhamento completo pode ser recuperado mantendo a trilha percorrida anteriormente. Assim, procede-se o cálculo do alinhamento ótimo em cada segmento de referência, e aquele com a avaliação acumulada menor será o vencedor que indicará a palavra desconhecida.

2.3.9 Seqüência de Palavras

O resultado final do alinhamento temporal é a seqüência desejada de palavras. Na prática, costuma-se gerar muitas seqüências, ou seja, aquelas de maior avaliação. Com isso pode-se fazer com que o sistema de RAV realize sua tarefa em dois passos. O primeiro passo, utilizando um modelo mais simples, determina um conjunto contendo as N melhores possíveis seqüências. Após isso avalia-se somente as declarações deste conjunto em modelos mais complexos. Este é conhecido como busca dos melhores N (*N-best search*).

2.4 Caracterização de sistemas de RAV

Um sistema de RAV pode reconhecer uma palavra isolada o que exige que o locutor realize uma pausa entre palavras ou reconhecer voz contínua que é muito mais difícil de ser feito.

Alguns sistemas necessitam de uma etapa de treinamento, ou seja, o usuário deve prover amostras de sua voz antes que possa utilizar o sistema. Tais sistemas são conhecidos como sistema dependente de locutor. O sistemas que não necessitam deste treinamento são ditos sistemas independente de locutor.

O reconhecimento é mais difícil se o vocabulário for grande ou conter muitas palavras que tenham sons similares.

Quando a voz é produzida por uma seqüência de palavras, modelos de linguagem ou gramáticas artificiais são usadas para restringir a combinação de palavras. O modelo de linguagem mais simples pode ser especificado como uma máquina de estado finito, onde as palavras possíveis, após cada palavra, são dadas explicitamente. Modelos de linguagens mais gerais, aproximando a linguagem natural, são especificados em termos de uma gramática sensível a contexto.

A perplexidade fornece a média geométrica do número de palavras que podem seguir uma palavra depois que o modelo de linguagem tem sido aplicado. É uma forma de medir a dificuldade do reconhecimento, envolvendo o tamanho do vocabulário e o modelo de linguagem.

A Tab. 2.6 apresenta um resumo dos parâmetros mais utilizados para caracterizar a capacidade de um sistema de RAV.

TABELA 2.6 - Parâmetros típicos usados para caracterizar a capacidade de um sistema de reconhecimento automático de voz

Parâmetros	**Faixa**
Modo de falar	Palavras isoladas ou contínua
Estilo de falar	De voz lida a voz expontânea
Engajamento	Dependente ou independente de locutor
Vocabulário	Pequeno (< 20 palavras) a grande (> 20.000 palavras)
Modelo de linguagem	Estado finito a sensível a contexto
Perplexidade	Pequena (<10) a grande (>100)
RSR	Alta(>30dB) e baixa(<10dB)
Transdutor	Microfone especial a telefone

O desempenho dos sistemas de RAV pode ser avaliado através da taxa de erro de palavras E, definido como:

$$E = \frac{S + I + R}{N} 100 , \tag{2.11}$$

onde N é o número total de palavras no conjunto de teste e ***S***, ***I*** e ***R*** são o número total de substituições, inserções e remoções, respectivamente.

Por substituição entende-se a troca de uma palavra falada por uma outra qualquer, normalmente similar. Por exemplo, o sistema escuta "6" quando na realidade o locutor disse "3".

Por inserção entende-se a adição de uma palavra. Por exemplo, o sistema escuta "12394" quando o locutor disse "1234".

Por remoção entende-se a perda de uma palavra. Por exemplo, o sistema escuta "156" quando o locutor disse "1256".

Além destes erros, outros podem ser incluídos nesta avaliação:

a) Falsa aceitação: Reconhecimento de uma palavra que não está no vocabulário como uma palavra que está no vocabulário. Esse erro também é conhecido como palavra fora do vocabulário;
a) Quebra: Uma palavra polissílaba é reconhecida como duas ou mais palavras. Por exemplo, "Maringá" é reconhecido como "mar e andar".
a) União: Duas ou mais palavras são reconhecidas como um simples palavra. Por exemplo, "a parte do" reconhecida como "aparelho".

2.5 Corpora

É uma base de dados de referência que contém amostras de voz que refletem o locutor, ambiente, vocabulário, perplexidade e outras informações necessárias no desenvolvimento de sistemas de RAV.

Corporas são utilizados também para o estudo e desenvolvimento de gramáticas de línguas faladas.

O número de corporas de linguagem falada está crescendo rapidamente, pelo menos para o Inglês falado nos Estados Unidos. A razão é que a ARPA e o NIST as usam para realizar a avaliação anual de sistemas de RAV com grande vocabulário e sistemas de entendimento de linguagem falada.

Um dos primeiros corporas construídos foi o da empresa norte-americana *Texas Instruments, Inc*. denominado *TI Digits corpus*, completado em 1984. Esta base de dados contém amostras de dígitos conectados falados por 326 mulheres, homens, meninas e meninos e tornou-se ferramenta padrão no desenvolvimento e teste de sistema de RAV de dígitos conectados para o inglês americano.

O corpora TIMIT, contratado pelo DARPA, foi desenvolvido para conter todos os padrões sonoros e seqüências fonéticas do inglês americano. Consistiu num esforço entre a Texas Instruments (TI), Massachusetts Institute of Technology (MIT) e Stanford Research Institute International (SRI). Ele contém amostras de voz de 630 pessoas representando todos os dialetos de todas as regiões dos Estados Unidos. O corpora TIMIT tem sido usado na pesquisa de representações acústicas da voz, no desenvolvimento de técnicas de modelagem estatística e na avaliação de sistemas de RAV baseado em fonemas.

O corpora *DARPA Resource Management*, completado em 1988, contém 2.800 sentenças faladas de uma centena de locutores nativos do inglês americano que consultaram uma base de dados da marinha sobre recursos militares.

O corpora King Corpus é usado para avaliação de sistemas de reconhecimento de voz e locutor. Este corpora foi coletado pela empresa norte-americana ITT em 1987 através de um contrato de pesquisa com o governo dos Estados Unidos da América. Existe uma versão disponível através do LDC[9] denominada King-92. Esta contém 51 locutores homem em dois canais distintos: um através de um telefone e o outro através de um microfone de alta qualidade. Para cada locutor e canal existem dez arquivos.

O corpora GlobalPhone, desenvolvido na Universidade de Karlsrure na Alemanha, contém amostras de voz das línguas: árabe; chinês (mandarim), croata, alemão, japonês, coreano, português, russo, espanhol, sueco, tamil e turco. Para cada língua, 100 locutores leram 20 minutos de artigos econômicos e políticos de jornais locais. A voz foi registrada em qualidade de escritório usando um microfone próximo a boca do locutor [SCH 98, SCH 97].

O corpora PDBCNS[10] está sendo desenvolvido pela empresa norte-americana ARCON, Inc. a pedido do Laboratório Lincoln do MIT. Consiste de uma base de dados de voz ruidosa. O objetivo é usar esta base na avaliação de métodos de enriquecimento de voz a serem usados pela justiça.

O corpora ATC[11] consiste de aproximadamente setenta horas de conversar registradas entre controladores de vôo e aeronaves nos três maiores aeroportos dos Estados Unidos.

O corpora do OGI, desenvolvido no CSLU, contém uma grande quantidade de sentenças, palavras, dígitos e pequenos textos em vários contextos e em vários ambientes. Além das amostras de voz, o CSLU tem uma série de ferramentas de manipulação deste corpora, assim como a descrição sucinta das bases. O corpora constitui de uma série de base de dados, alguns já completados e outros em desenvolvimento. O OGI trabalha com a política de ceder sem custos para instituições de pesquisas e universidades todo o seu material.

Existem muitos outros corporas. A Tab. 2.7 lista alguns deles, considerados os mais importantes.

TABELA 2.7 – Lista de Corporas de Voz

Corpora	Informações e Contato
ACCOR	Prof. W. Hardcastle (sphard@queen-margaret-college.main.ac.uk); Prof. A. Marchal (phonetic@fraix11.bitnet)
ALBAYZIN	Prof. Climent Nadeu, Departament of Speech Signal Theory and Communications, Univeristat Politecnica de Catalunya, Espanha (nadeu@tsc.upc.es)
ANDOLS[12]	Bruce Millar, Computer Sciences Laboratory, Research School of Information Sciences and Engineering, Australian National University, Camberra, ACT 0200, Australia (bruce@cslab.anu.edu.au)
ARS	Mr. G. Babini, Via G. Beis Romoli 274, I-101488, Torino, Italia

[9] Informações adicionais deste corpora podem ser obtidos em http://www.ldc.upenn.edu/ldc/noframe.html. Obs: Utilize a palavra King no campo de pesquisa.

[10] Pilot Data Base for Corpus of Noisy Speech

[11] Air Traffic Control. Esta base também pode em http://www.ldc.upenn.edu/ldc/noframe.html

[12] Australian National Database of Spoken Language

ATR, ETL & JEIDA	K. Kataoka, AI and Fuzzy Promotion Center, Japan Information Processing Development Center (JIPDEC), 3-5-8 Shibakoen, Minatoku, Tokyo 105, Japan, TEL. +81 3 3432 9390, FAX. +81 3 3431 4324
BREF	bref@limsi.fr
CAR & Waxholm	Bjorn Granstrom (bjorn@speech.kth.se)
CNSC[13]	Prof. Jialu Zhang, Academia Sinica, Institute of Acoustic, 17 Shongguanjun St, Beijing PO Box 2712, 100080 Beijing, Peoples Republic of China
CSLU	Veja http:// www.cse.ogi.edu/CSLU/corpora.html
ELRA[14]	Sarah Houston (100126.1262@compuserve.com)
ELSNET[15]	elsnet@let.ruu.nl
ERBA	Stefan Rieck, Lehrstuhl Informatik 5 (Pattern Recognition), University of Erlangen-Nurnberg, Martensstr.3 , 8520 Erlangen, Germany. (rieck@informatik.uni-erlangen.de)
EuroCocosda	A Fourcin (adrian@phonetics.ucl.ac.uk)
EUROM1	Base de dados em várias línguas. Contato geral: A. Fourcin (UCL) (adrian@phonetics.ucl.ac.uk). Contatos para línguas individuais: Alemão - D. Gibbon (Un.Bielefeld) (gibbon@asl.uni-bielefeld.de) Dinamarquês – B. Lindberg (IES) (bli@stc.auc.dk) Francês - J.F. Serignat (ICP) (serignat@icp.grenet.fr) Italiano - G. Castagneri (CSELT) (castagneri@cselt.stet.it) Norueguês - T. Svendsen (SINTEF-DELAB) (torbjorn@telesun.tele.unit.no) Holandês - J. Hendriks or L. Boves (PTT Research) (boves@lett.kun.nl) Sueco - G. Hult (Televerket) or B. Granstrom (KTH) (bjorn@speech.kth.se) Inglês - UK: A. Fourcin (UCL) (adrian@phonetics.ucl.ac.uk) Espanhol - A. Moreno (UPC) (amoreno@tsc.upc.es) Português - I. Trancoso (INESC) (imt@inesc.pt)
GRONINGEN	(els@spex.nl) (CDs disponíveis via ELSNET)
LCS[16]	Laboratório de Comunicações e Sistemas da Universidade de São Paulo. Professor Dr. Euvaldo F.Cabral Jr. (euvaldo@lcs.poli.usp.br ou http://www.lcs.poli.usp.br)
LDC[17]	Elizabeth Hodas (ehodas@unagi.cis.upenn.edu) ou http://www.cis.upenn.edu/ldc. Informação sobre o LCD e suas atividades podem ser obtidas via ftp anônimo (ftp.cis.upenn.edu) no diretório pub/ldc.

[13] Chinese National Speech Corpus
[14] European Language Resources Association
[15] European Network in Language and Speech
[16] Laboratório de Comunicações e Sistemas da Universidade de São Paulo. (http://www.lcs.poli.usp.br)
[17] Linguistic Data Consortium. Maiores informações podem ser obtidas em http://www.ldc.upenn.edu/ldc/noframe.html

LLSEC[18]	É uma base de dados de voz amostrada com múltiplos microfones em diferentes cenários para refletir realisticamente uma grande variedade de condições de voz degradadas.
LRE ONOMASTICA	M. Jack, CCIR, University of Edinburgh, (mervyn.jack@ed.ac.uk)
NSC[19]	Steve Crowdy, Longman UK, Burnt Mill, Harlow, CM20 2JE, UK
PAROLE	Mr. T. Schneider, Sietec Systemtechnik Gmbh, Nonnendammallee 101, D-13629 Berlin, Alemanha
PHONDAT2	B. Eisen, University of Munich, Alemanha
POINTER	Mr. Corentin Roulin , BJL Consult, Boulevard du Souverain 207/12, B-1160 Bruxelles, Bélgica.
POLYGLOT	Antonio Cantatore, Syntax Sistemi Software, Via G. Fanelli 206/16, I- 70125 Bari, Italy
RELATOR	A. Zampolli, Istituto di Linguistica Computazionale, CNR, Pisa, Italia, (giulia@icnucevm.cnuce.cnr.it) ou http://www.XX.relator.research.ec.org
REVOX	Prof. Dr. Dante Barone (barone@inf.ufrgs.br)
ROARS	Pierre Alinat, Thomson-CSF/Sintra-ASM, 525 Route des Dolines, Parc de Sophia Antipolis, BP 138, F-06561 Valbonne, França
SCRIBE	Mike Tomlinson, Speech Research Unit, DRA, Malvern, Worc WR14 3PS, Inglaterra
SPEECHDAT	Mr. Harald Hoege, Siemens AG, Otto Hahn Ring 6, D-81739 Munich, Alemanha
SPELL	Jean-Paul Lefevre, Agora Conseil, 185, Hameau de Chateau, F-38360 Sassenage, França
SUNDIAL	Jeremy Peckham, Vocalis Ltd., Chaston House, Mill Court, Great Shelford, Cambs CB2 5LD UK, (jeremy@vocalis.demon.co.uk)
SUNSTAR	Joachin Irion, EG Electrocom Gmbh, Max-Stromeyerstr. 160, D-7750 Konstanz, Alemanha
VERBMOBIL	B. Eisen, University of Munich, Alemanha

2.5.1 Corporas disponíveis do CSLU

Abaixo temos uma descrição dos principais corporas disponibilizados pelo CSLU até a presente data[20].

a) ISOLET

É uma base de dados das letras do alfabeto inglês falado isoladamente. A base de dados consiste em 7800 letras faladas, 2 produções de cada letra por 150 locutores. Contém

[18] Lincoln Laboratory Speech Enhancement Corpus. Em http://www.ll.mit.edu/SST/corpora.html pode ser obtida a base de dados completa

[19] Normal Speech Corpus

[20] Este foi detalhado pois tivemos acesso a todas essas base de dados, através de fitas enviadas pelo CSLU para o Instituto de Informática da UFRGS.

aproximadamente 1,5 horas de voz. As gravações foram feitas em ambiente silencioso (laboratório) com um microfone especial (*noise-canceling microphone*).

b) *Spelled and Spoken Words*

Consiste de vozes de 4000 locutores que realizaram uma chamada telefônica. Eles soletraram seu primeiro e último nome, e falaram seu nome completo, qual cidade que eles cresceram, a cidade de onde falaram, e responderam a duas perguntas de sim/não. A fim coletar exemplos suficientes de cada letra falada, aproximadamente 1000 locutores recitaram também o alfabeto inglês com pausas entre as letras. Cada chamada foi transcrita por duas pessoas e todas as diferenças foram resolvidas. Além disso, parte deste corpora foi foneticamente nomeado.

c) *Multi-Language Telephone Speech*

Consiste de vozes de telefone de 11 línguas: Inglês, Farsi, francês, alemão, Hindi, japonês, coreano, mandarim, espanhol, Tamil, e vietnamita. O corpus contém declarações fixas do vocabulário (por exemplo dias da semana) assim como voz continua fluente. A versão atual inclui declarações de aproximadamente 2000 locutores, com aproximadamente 23 horas de voz. Transcrições ortográficas não alinhadas temporalmente e transcrições fonéticas alinhadas temporalmente estão disponíveis para algumas das declarações.

d) Histórias

Consiste na declaração extemporâneas de vozes em inglês produzidas por 692 locutores diferentes (aproximadamente 10 horas). As histórias foram coletadas durante o levantamento de dados, que produziu 692 atendimentos completos por locutores falando em inglês. No fim do protocolo, foi pedido aos locutores que falassem sobre qualquer tópico durante um minuto. Estes monólogos são referidos como histórias. Cada história foi transcrita ao nível de palavra não alinhada temporalmente. 322 histórias foram transcritas ao nível de palavra com marcadores de alinhamento temporal, e 210 histórias foram etiquetadas foneticamente.

e) *Cellular Words and Phrases*

Consiste de gravações de voz faladas através do telefone celular. A versão atual inclui declarações de 344 locutores diferentes e de uma transcrição de cada declaração. Há um total de 15.109 arquivos, compreendendo aproximadamente 7,6 horas de voz.

f) *Apple Words and Phrases*

Este corpora foi financiado pela empresa norte-americana *Apple Computer Inc.* Consiste de aproximadamente 21 horas de gravação de vozes ao telefone. Um mil chamadas foram coletados através de um sistema análogo e duas mil foram coletadas através de um sistema digital. Cada locutor repetiu uma lista de frases do tipo do comando e do controle, como por exemplo “ajuda”.

g) *30K Names*

Consiste na coleção de declarações do primeiro e último nomes. A versão 1,0 do corpus de 30.000 nomes contém 15.000 arquivos. Cada arquivo tem uma transcrição ortográfica e aproximadamente 7.000 têm uma transcrição fonética.

h) *30K Numbers*

É uma coleção de números ordinais e cardinais espontâneos, de vários dígitos contínuos e de dígitos isolados. A versão 1,0 do corpora contém 30.000 números em 15.000 arquivos. Cada arquivo tem uma transcrição ortográfica e aproximadamente 7.000 têm uma transcrição fonética.

i) *Portland Cellular*

Consiste de vozes gravadas em telefone celular de aproximadamente 425 locutores da área metropolitana da cidade de Portland, estado de Oregon nos Estados Unidos. O corpus contém vocabulário fixo e declarações contínuas. Há aproximadamente 8000 declarações, cada uma com uma transcrição ortográfica. Além disso, aproximadamente 200 declarações contínuas com duração de 30 segundos cada foram foneticamente transcritas.

j) *Yes/No Words*

É uma coletânea de declarações expontâneas de "*yes/no*". A versão 1,0 contém perto de 20.000 declarações. Esta versão não contém qualquer palavra além de "*yes/no*". A próxima versão provavelmente conterá palavras como "yep" e "nope".

k) *SR4X*

É uma coleção de 36 locutores repetindo 11 palavras 6 vezes em quatro canais diferentes.

l) *Alphadigit*

Consiste de declarações gravadas de telefone de 3000 locutores. Cada locutor repetiu várias declarações alfanuméricas. As declarações foram definidas tal que todo par de letras e números fosse representado e que cada letra ou dígito aparecesse na posição inicial e final. Cada arquivo contém uma transcrição ortográfica.

m) *22 Language*

Consiste de voz gravada de telefone de 22 línguas diferentes: Árabe oriental, Cantones, Tcheco, Farsi, Francês, Alemão, Hindi, Hungariano, Japonês, Koreano, Malay, Mandarin, Italiano, Polonês, Português, Russo, Espanhol, Sueco, Swahili, Tamil, Vietnamita, e Inglês. O corpus contém declarações com vocabulários fixo (por exemplo dias da semana) e voz contínua fluente. Existem pelo menos 300 locutores diferentes para cada língua. Cada declaração é verificada por um locutor nativo para determinar se o locutor seguiu as instruções corretamente na respostas as indagações. Uma **parte** das declarações de cada língua foram foneticamente e ortograficamente transcritas.

n) *National Cellular*

Consiste de declarações em telefone celular de 676 locutores de quatro cidades americanas. Cada locutor declara vários dígitos e números assim como voz fluente. Cada arquivo tem uma transcrição ortográfica.

2.6 Fontes de Variabilidades

Os sistemas de reconhecimento automático de voz modelam as fontes de variabilidade em muitas formas. A Tab. 2.8 descreve como pode ser modelada estas variabilidades, de forma a levá-las em consideração nos sistemas de RAV.

TABELA 2.8 - Modelagem das Variabilidades da Voz

Variabilidade ao Nível da:	**Como pode ser modelada**
Representação do sinal	Enfatizando as características do sinal de voz que sejam independente do locutor e não enfatizando aquelas dependentes;
Fonética acústica	a) Usando técnicas estatísticas aplicada a grandes quantidades de dados; a) Adaptação dos modelos acústicos independente do locutor para aqueles do corrente locutor durante seu uso; a) Através do treino de modelos fonéticos em diferentes contextos.
Palavra	Através das várias possíveis pronúncias da palavra em representações conhecidas como redes de pronunciação.

2.7 Estado da Arte

A última década tem mostrado um significante progresso na tecnologia dos sistemas de RAV. Segundo [COL 95] a taxa de erro continua a cair por um fator de 2 a cada dois anos. Os fatores que tem contribuído para este avanço são:

a) Uso do HMM uma vez que a partir dos dados de treinamento, os parâmetros do modelo podem ser obtidos automaticamente para um dado desempenho ótimo;
a) Desenvolvimento de corporas que permitem quantificar as características importantes do sinal de voz;
a) Definição de padrões de avaliação de desempenho. A apenas uma década, pesquisadores treinavam e testavam seus sistemas usando dados coletados localmente. Isso tornava difícil comparar o desempenho dos sistemas e na maioria das vezes este desempenho era fortemente degradado quando eram apresentados dados que não pertenciam ao conjunto de treinamento.
a) Avanços na tecnologia do computador também influenciaram neste progresso, uma vez que eles estão cada vez mais rápidos, baratos e com grande quantidade de memória disponível. Isso permitiu avaliar algoritmos cada vez mais complexos, que exigiam grande quantidade de memória e poder de processamento. Além disso, o tempo de teste de uma nova idéia e sua efetiva implementação e coleta dos resultados foi grandemente reduzido.

Desde 1992, os pesquisadores tem direcionado seus esforços para grandes vocabulários (20.000 ou mais palavras), alta perplexidade (PP >= 200), independente de

locutor e reconhecimento de voz contínua. O melhor sistema em 1994 alcançava uma taxa de erro de 7,2% para converter voz de noticiário de negócios americanos para palavras.

Mesmo embora muito progresso tenha sido alcançado, as máquinas estão muito longe de reconhecer voz conversacional. Taxas de reconhecimento de voz ao redor de 50% é o que tem sido obtido em conversações telefônicas. Levará muitos anos antes que ilimitado vocabulário com voz contínua e independente de locutor possa ser realizado pelos sistemas de RAV.

Não se tem notícia de um sistema de RAV contínua para o português falado no Brasil até o momento.

2.8 Futuro

Em 1992, a NSF (Fundação Nacional de Ciência dos Estados Unidos) promoveu um *workshop* para identificar os desafios chaves na pesquisa da tecnologia de linguagem humana e a infra-estrutura necessária para suportar o trabalho. Pesquisas nas áreas: Robustez, Portabilidade, Adaptação, Modelos de Linguagem, Palavras fora do vocabulário, Voz espontânea, Prosódia e Modelagem Dinâmica foram os principais desafios levantados para os próximos anos. A Tab. 2.9 descreve cada um desses desafios.

Além desses desafios, o Brasil ainda necessita passar a barreira do desenvolvimento de um Corpora com amostras do português falado no Brasil com seu regionalismos, sotaques e diferentes vocabulários. Além de aspectos tais como ruído de meios de comunicação e sistema de telefonia celular. Só assim poderá os centros de pesquisa no Brasil desenvolverem modelos adequados ao português. Após o desenvolvimento desta estrutura básica é que os pesquisadores poderão e terão condições de abordar adequadamente os desafios relacionados anteriormente.

TABELA 2.9 - Desafios na tecnologia de RAV

Área	**Desafio**
Robustez	Em um sistema robusto, o desempenho degrada pouco quando as condições tornam-se diferentes daquela em que o sistema foi treinado. Diferenças nas características do sinal e ambiente acústico deverão receber especial atenção.
Hardware	Desenvolvimento de hardware especializado para a amostragem de sinais de voz, tais como microfones com compensação de ruído, pré processamento imitando o sistema auditivo humano, etc.
Portabilidade	Refere-se a habilidade de rapidamente projetar e implementar novas aplicações com RAV. Com o objetivo de alcançar picos de eficiência os sistemas devem ser treinados para exemplos específicos ao qual se destina a aplicação e isso requer tempo e muito dinheiro.
Adaptação	Como podem os sistemas se adaptarem continuamente a troca de condições (novos microfones, diferentes locutores, etc.) e melhorar através de seu uso?
Modelo de Linguagem	Sistemas atuais utilizam modelos estatísticos de linguagem para ajudar a reduzir o espaço de busca e resolver ambigüidades

	acústicas. Quando o vocabulário cresce talvez seja necessário incorporar restrições sintáticas e semânticas que não podem ser capturadas por modelos puramente estatísticos.
Palavras fora do vocabulário	Os sistemas de RAV são projetados para uso de um particular conjunto de palavras, mas os usuários do sistema podem não saber exatamente quais palavras o sistema pode reconhecer. Isso leva a um certo percentual de palavras fora do vocabulário em condições naturais. Os sistemas devem possuir algum método de detectar estas palavras fora do vocabulário.
Voz espontânea	Os sistemas que são empregados para uso real devem lidar com uma variedade de fenômenos, tais como falsos inícios, hesitações, construções gramaticalmente incorretas e outros comportamentos comuns não encontrados em voz lida.
Prosódia	Refere-se a estrutura acústica que estende-se por vários segmentos ou palavras. Acento, entonação e ritmo trazem uma importante informação para RAV e as intenções do usuário (ex.: sarcasmo, raiva). Os sistemas atuais não capturam a estrutura da prosódia. Como integrar informação da prosódia dentro da arquitetura de reconhecimento é uma questão crítica que ainda não tem sido respondida.
Modelagem dinâmica	O sistemas assumem uma seqüência de segmentos de entrada os quais são tratados como se eles fossem independentes. Mas é sabido que aspectos perceptuais de palavras e fonemas requerem a integração de características que refletem os movimentos das articulações do aparelho fonador, os quais são de natureza dinâmica. Como modelar a dinâmica e incorporar esta informação dentro de sistemas de RAV é um problema não solucionado.

2.9 Conclusão

Este capítulo apresentou as tecnologias envolvidas no processo de reconhecimento automático de voz. Inicialmente apresentou-se um breve histórico dos principais acontecimentos que marcaram o desenvolvimento da tecnologia do RAV. Após isso definiu-se o que é um sistema de RAV e seus principais componentes, com a apresentação das tecnologias envolvidas. Uma vez sabido o que é um sistema RAV, apresentou-se os principais parâmetros utilizados na caracterização desses sistemas. Também foram apresentados os principais Corporas existentes contendo amostras de sinais de voz. Após isso discutiu-se como pode ser modelada as diferentes fontes de variabilidade presentes no sinal de voz. Finalmente apresentou-se o estado da arte em termos de tecnologias e sistema existentes e a expectativa futura nesta área.

3 Robustez nos sistemas de RAV

3.1 Introdução

Robustez no reconhecimento de voz refere-se a necessidade de manter bom reconhecimento mesmo quando a qualidade da voz de entrada é degradada, ou quando as características acústicas, articulatórias ou fonéticas da voz de treinamento e ambientes de testes diferem. Obstáculos ao reconhecimento robusto incluem degradações acústicas produzidas por ruído aditivo, os efeitos da filtragem linear, não linearidades introduzidas pelo microfone ou meio de transmissão. A Fig. 3.1 apresenta uma representação esquemática de algumas fontes de variabilidade que podem degradar a precisão no reconhecimento de voz e alguns dos possíveis procedimentos de compensação.

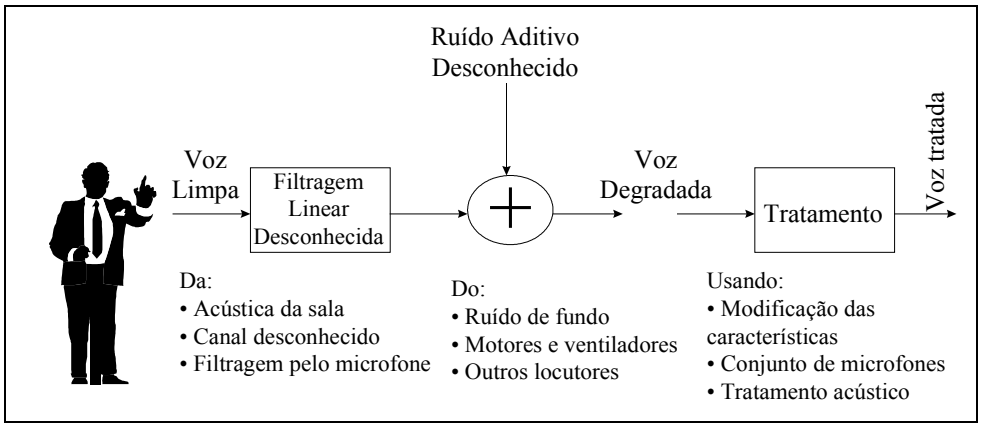


FIGURA 3.1 - Representação esquemática de algumas fontes de variabilidade

Os sistemas de RAV tem-se tornado muito mais robustos em recentes anos com respeito a variabilidade acústica e a diferenças entre locutores. Na medida em que estão sendo disponibilizados na forma de aplicações reais, a necessidade de uma maior robustez na tecnologia de reconhecimento está se tornado mais aparente. Em geral, o desempenho dos sistemas de RAV caem na medida em que:

a) A voz é transmitida através de linhas telefônicas;
b) A relação sinal ruído é extremamente baixa, particularmente quando o ruído de fundo consiste de vozes de outros locutores;
c) A língua nativa do locutor não é aquela em que o sistema foi treinado.

Progresso significativo tem sido obtido na última década na adaptação dinâmica dos sistemas de RAV a novos locutores, com técnicas que modificam ou deformam a representação fonética do sistema para refletir as características acústicas de locutores individuais.

A seção 3.2 apresenta algumas dessas técnicas para a adaptação dinâmica dos parâmetros internos que podem melhorar a robustez dos sistemas de RAV em relação ao ambiente e diferentes locutores. A seção 3.3 discute a questão do uso simultâneo de vários microfones com o objetivo de melhorar a direcionalidade e a relação sinal ruído.

A seção 3.4 enfatiza a importância do uso de técnicas de processamento digital de sinais, fisiologicamente motivados. Finalmente na seção 3.5 apresenta-se os desafios futuros para a melhoria da robustez nos sistemas de RAV.

3.2 Adaptação Dinâmica dos Parâmetros

A adaptação das características que são entradas de um sistema de RAV ou das representações internas das possíveis declarações, é a mais direta abordagem para a adaptação ambiental e de locutor. Estas adaptações podem ser realizadas por meio de:

a) Uso de procedimentos para a estimativa ótima de novos parâmetros;
b) Desenvolvimento de procedimentos de compensação baseados em comparações empíricas da voz de treinamento e ambiente de teste;
c) Uso de filtragem passa alta nos parâmetros.

3.2.1 Estimativa ótima dos parâmetros

A maior parte das técnicas de melhoria da robustez são baseadas no modelo estatístico que caracteriza as diferenças entre a voz usada para treinamento e aquela usada no teste do sistema. Os parâmetros destes modelos são estimados e as características da voz de entrada ou as representações internamente armazenadas no sistema são então modificadas. Modelos estruturais típicos para a adaptação da variabilidade acústica assumem que a voz é adicionada de ruído com um espectro de potência desconhecido ou por uma combinação de ruído adaptativo e filtragem linear. Muitos trabalhos recentes consistem na reimplementação de técnicas que removem o ruído aditivo com o propósito de enriquecer a voz. Tais abordagens foram capazes de substancialmente reduzir as taxas de erros dos sistemas de RAV. As abordagens são similares para o caso de adaptação ao locutor.

A solução dos problemas de estimação requer aproximações analíticas e numéricas ou o uso de técnicas de estimação iterativas tal como o algoritmo de máxima estimativa (EM - Estimate Maximize). Estas abordagens tem tido sucesso em aplicações onde assume-se que os modelos são razoavelmente válidos, mas eles são limitados em alguns casos pela complexidade computacional. Uma outra abordagem é usar o conhecimento do ruído de fundo dos exemplos para transformar a média e a variância dos modelos fonéticos que tem sido desenvolvido para limpar a voz. A técnica conhecida como combinação de modelos paralelos estende esta abordagem, provendo um modelo analítico da degradação que contabiliza o ruído aditivo. Estes métodos trabalham razoavelmente bem, mas são computacionalmente custosos no presente.

3.2.2 Comparação de características empíricas

A comparação de características empíricas derivadas da voz de alta qualidade com características da voz que é simultaneamente registrada sob condições degradadas podem ser usadas, ao invés de um modelo estrutural, para compensar as diferenças entre as condições de treinamento e de testes. Nesses algoritmos, os efeitos combinados das variabilidades ambiental e do locutor são tipicamente caracterizados como perturbações

aditivas nas características. Muitos dos algoritmos robustos, empiricamente baseados, com algum sucesso aplicam vetores de correção aditiva às características derivadas da forma de onda da voz de entrada ou aplicam vetores de correção aditiva de representações internas dessas características no sistema de RAV.
Esta abordagem geral pode ser estendida para casos quando o ambiente de teste é desconhecido a priori, através do uso de vetores de correção em paralelo para um determinado número de diferentes condições de teste, e através da subsequente aplicação do conjunto de vetores de correção, ou modelos acústicos. Nos casos onde a condição de teste não é uma daquelas usadas para treinar os vetores, a precisão no reconhecimento pode ser melhorada pela interpolação dos vetores de correção ou por características estatísticas representantes das condições do melhor candidato. Os procedimentos da compensação empiricamente obtidas são extremamente simples, e eles são completamente eficazes nos casos quando as condições de teste são razoavelmente similares àquelas condições usadas para desenvolver os vetores de correção.

3.2.3 Filtragem passa alta cepstral

Consiste na filtragem passa alta dos coeficientes cepstrais, o qual provém um notável maior robustez a um custo computacional insignificante. No bem conhecido método RASTA, um filtro passa alta ou passa banda é aplicado ao espectro de potência logarítmico da voz tais como os coeficientes cepstrais. O método da normalização média cepstral (CMN - Cepstral Mean Normalization), a filtragem passa alta é acompanhada pela subtração da média do coeficientes cepstrais de curto tempo do coeficientes cepstrais de entrada. Esses algoritmos compensam diretamente os efeitos da filtragem linear desconhecida porque eles forcam os valores médios dos coeficientes cepstrais sejam zero nos domínios de teste e de treinamento, e assim se igualam. Um extensão do algoritmo RASTA conhecido como J-RASTA melhora a robustez para voz com baixa RSR. Em uma avaliação usando 13 dígitos isolados sobre linhas telefônicas, foi mostrado que o método J-RASTA reduziu as taxas de erro em 55% comparado com o RASTA quando o ruído e os efeitos da filtragem estão presentes. A filtragem cepstral passa alta é tão barata e efetiva que é correntemente enriquecida de alguma forma em praticamente todos os sistemas que são requeridos eficácia no reconhecimento robusto.

3.3 Uso de múltiplos microfones

Maiores melhoramentos na precisão do reconhecimento pode ser alcançados mediante o uso de múltiplos microfones pois produzem sensibilidade diretiva direcional na direção da fonte sonora. De fato, resultados de recentes experimentos pilotos em ambientes de escritório confirmam que o uso de múltiplos microfones em combinação com um algoritmo de pós processamento pode reduzir as taxas de erro de reconhecimento em até 61%.

Conjunto de processadores que fazem uso do erro quadrado médio mínimo (MMSE - Minimum Mean Square Error) baseado em técnicas de filtragem adaptativa clássicas podem trabalhar bem quando a degradação do sinal é dominada por ruído aditivo independente, mas ele não trabalha bem em ambientes reverberantes quanto a distorção é uma versão atrasada do sinal de voz.

Uma terceira abordagem para processamento em conjunto de microfones é o uso de algoritmos de correlação cruzada os quais tem a habilidade de reforçar as componentes de um som chegando de um determinado lugar. Esses algoritmos são apelativos, uma vez que são inspirados no processamento feito pelo sistema auditivo humano, mas eles tem demonstrado somente uma modesta superioridade.

3.4 Uso de Processamento de Sinal Fisiologicamente Motivado

Tem sido desenvolvido esquemas de processamento de sinal para RAV que imitam vários aspectos da fisiologia do sistema perceptivo e auditivo humano. Recentes avaliações indicam que esses modelos podem de fato prover melhor reconhecimento do que a tradicional representação cepstral quando a qualidade da voz de entrada degrada, ou quando as condições de treino e teste diferem.

Apesar disso, modelos baseados no sistema auditivo humano ainda não são capazes de demonstrar melhor reconhecimento do que o mais eficaz algoritmo dinâmico de adaptação, e técnicas de adaptação convencionais são de longe bem mais eficazes computacionalmente.

É possível que o sucesso de modelos baseados no sistema auditivo humano tem sido limitado pelo fato que os classificadores baseados em HMM não se adequam bem as propriedades estatísticas produzidas por esses modelos. Outros pesquisadores sugerem que ainda não foi identificado as características dos modelos que proverão superior eficácia.

3.5 Futuro

O reconhecimento robusto de voz tem-se tornado somente há muito pouco tempo uma área vital de pesquisas. Até o momento, os maiores sucessos em adaptação ambiental tem sido limitado ao ruído aditivo quase estacionário e/ou filtragem linear, ou quando há disponibilidade de dados para treinamento sobre o ambiente. Algoritmos de adaptação ao locutor tem tido sucesso em prover melhor reconhecimento para locutores com a mesma língua nativa. Já para locutores com língua nativa diferente, o reconhecimento preciso permanece substancialmente ruim. A Tab. 3.1 resume o que pode ser feito para que os sistema de RAV se adaptam, ou seja, sejam mais robustos aos diferentes possíveis agentes causadores das variabilidades de um sinal de voz.

TABELA 3.1 - O que pode ser feito para que os sistemas de RAV se adaptem aos diferentes tipos de variabilidades em que o sinal de voz esta sujeito

Tipo de Desafio	**O que pode ser feito:**
Voz sobre linha telefônicas	O reconhecimento de voz telefônica é difícil porque o canal telefônico tem sua própria RSR e resposta em freqüência. A voz sobre linhas telefônicas está sujeita a interferências transientes e distorções não lineares.
Ambientes com baixa RSR	Mesmo com as técnicas do estado da arte de compensação, o reconhecimento preciso degrada quando a RSR do canal cai abaixo de 15dB, mesmo embora humanos podem obter excelente

	reconhecimento a este nível de RSR.
Interferência entre canais	Interferência por outros locutores é um desafio muito maior que a interferência por ruídos de banda larga. Até o momento não se tem técnicas que permitam resolver este problema.
Rápida adaptação a locutores não nativos	Os sistemas de RAV devem ser capazes de trabalhar também com locutores não nativos de forma a garantir seu sucesso comercial.
Corpora de vozes com degradações realísticas	O progresso rápido no reconhecimento robusto será dependente da formulação, coleta, transcrição e disseminação de corporas com voz que contenham exemplos realísticos das degradações encontradas em ambientes práticos. A seleção de tarefas apropriadas e domínios para fontes de bases de dados compartilhada é melhor realizada através da cooperação de pesquisadores, desenvolvedores de aplicações e usuários finais. O conteúdo dessas bases de dados deveriam ser realísticas o suficientes para serem úteis na solução de verdadeiros problemas.

3.6 Conclusão

Neste capítulo discutiu-se o problema da robustez dos sistemas de RAV, ou seja, como evitar que as diversas variabilidades presentes no sinal de voz diminuam a qualidade do reconhecimento. Apresenta-se também algumas técnicas para a adaptação dinâmica dos parâmetros internos que podem melhorar a robustez dos sistemas de RAV em relação ao ambiente e diferentes locutores. Discutiu-se a questão do uso simultâneo de vários microfones com o objetivo de melhorar a direcionalidade e a relação sinal ruído. Enfatizou-se a importância do uso de técnicas de processamento digital de sinais, fisiologicamente motivadas. Finalmente apresentou-se os desafios futuros para a melhoria da robustez nos sistemas de RAV.

4 Modelos Escondidos de Markov - HMM

4.1 Introdução

Este capítulo apresenta os modelos escondidos de Markov que consiste da técnica mais flexível e de maior êxito usada em RAV. A seção 4.2 apresenta os conceitos básicos. A seção 4.3 descreve os três principais algoritmos utilizados no treinamento e teste dos modelos escondidos de Markov. Finalmente, a seção 4.4 discute: como aplicar o modelo escondido de Markov ao espaço acústico; os algoritmos de treinamento; e algumas de suas deficiências.

4.2 Conceitos Básicos

Um modelo escondido de Markov consiste de um conjunto de estados conectados por transições. Ele inicia em um estado inicial. Em cada passo de tempo discreto, uma transição é feita para um novo estado e um símbolo de saída é produzido neste novo estado. A palavra ***escondida*** é derivada do fato de que a seqüência de estados visitada no tempo é escondida, sendo somente visível a seqüência de símbolos produzidos em cada estado.

Em RAV, os estados representam os modelos acústicos, indicando a seqüência dos sons que deverão ser produzidas nos mais diversos segmentos de voz. Já as transições provém restrições temporais, indicando quais estados podem seguir cada outro na seqüência. A Fig. 5.1 ilustra como estados e transições em um HMM podem ser estruturados hierarquicamente com o objetivo de representar fonemas, palavras e sentenças.

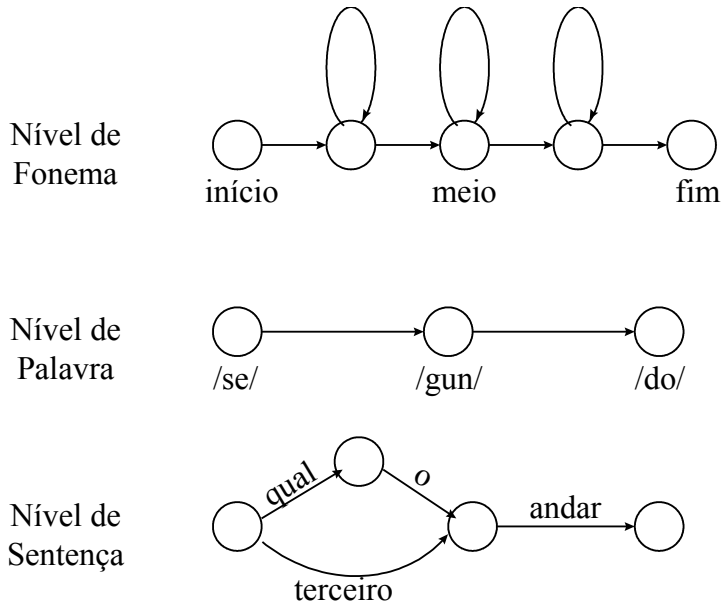


FIGURA 5.1 - Um HMM hierarquicamente estruturado

Um HMM consiste dos seguintes elementos:

$\{s\}$ - conjunto de estados

$\left\{a_{ij}\right\}$ - conjunto de probabilidades de transição, onde a_{ij} é a probabilidade de ocorrer a transição do estado i para o estado j

$\left\{b_i(u)\right\}$ - conjunto de probabilidades de emissão, onde b_i é a distribuição de probabilidade sobre o espaço acústico descrevendo a probabilidade de emissão de todo possível som u no estado i.

As probabilidades a e b devem satisfazer as seguintes propriedades:

$$a_{ij} \geq 0,\ b_i(u) \geq 0,\ \forall i, j, u \tag{4.1}$$

$$\sum_{j=1}^{n} a_{ij} = 1,\ \forall i \tag{4.2}$$

$$\sum_{u=1}^{M} b_i(u) = 1,\ \forall i \tag{4.3}$$

onde n é o número total de estados e M o número total de sons.

Pode-se verificar que a e b dependem somente do estado atual, independente da história passada da seqüência de estados, o que implica que está-se trabalhando com modelos HMM de primeira ordem. Com isso o número de parâmetros que deve ser treinado é pequeno e consequentemente os algoritmos de treinamento e de testes tornam-se muito eficientes.

4.3 Algoritmos

Os três principais algoritmos de treinamento e teste de modelos HMM são:

a) O algoritmo progressivo, muito usado no reconhecimento de palavras isoladas;
a) O algoritmo Viterbi, usado no reconhecimento de voz contínua;
a) O algoritmo *forward-backward*, útil para o treinamento de um HMM.

4.3.1 O algoritmo progressivo

Para que seja possível realizar o reconhecimento isolado de palavras, deve-se avaliar a probabilidade que um dado modelo HMM de palavra produza uma dada seqüência de forma a poder-se comparar as medidas de cada modelo de palavra e escolher aquela com a maior avaliação. Em outras palavras, dado um modelo HMM ***M***, consistindo de $\{s\}$, $\left\{a_{ij}\right\}$ e $\left\{b_i(u)\right\}$, deve-se calcular a probabilidade que seja gerado a

seqüência de saída $y_1^T = (y_1, y_2, y_3, \ldots, y_T)$. Uma vez que todo estado i pode gerar cada saída u com probabilidade $b_i(u)$ então toda seqüência de estado de comprimento T contribui para a probabilidade total. Uma forma de calcular esta probabilidade seria listar todas as possíveis seqüências de estados de comprimento T e acumular suas probabilidades de gerar y_1^T. Isso não é prático, pois trata-se de um algoritmo com complexidade exponencial. Uma solução muito mais eficiente é o conhecido algoritmo progressivo, com complexidade linear, o qual é uma instancia da classe de algoritmos conhecidos como programação dinâmica. Primeiro define-se $\alpha_j(t)$ como a probabilidade de gerar uma seqüência parcial y_1^t para o estado j no tempo t. $\alpha_j(t=0)$ é inicializado para 1,0 no estado inicial, e 0,0 em todos os outros estados. Se já foi calculado $\alpha_i(t-1)$ para todo i na seqüência de tempo anterior t-1, então $\alpha_j(t)$ pode ser avaliado recursivamente em termos da probabilidade incremental de entrar no estado j de cada estado i enquanto gerando o símbolo de saída y_t, ou seja:

$$\alpha_j(t) = \sum_{i=1}^{n} \alpha_i(t-1) a_{ij} b_j(y_t), \tag{4.4}$$

onde n é o número total de estados.

Se F é o estado final, então por indução vê-se que $\alpha_F(T)$ é a probabilidade que o modelo HMM gere a seqüência completa de saída y_1^T.

4.3.2 O algoritmo Viterbi

O número de modelos HMM, um para cada possível declaração, no reconhecimento contínuo de voz seria impraticável se fosse utilizado o algoritmo progressivo. Desta forma deve-se inferir a seqüência real de estados que geram a dada seqüência de observação. Desta seqüência de estados pode-se facilmente recuperar a seqüência da palavra. O estado real é escondido, por definição, e não pode ser unicamente identificado assim como qualquer caminho poderia ter produzido esta seqüência de saída, com alguma pequena probabilidade. A melhor forma de fazer isso seria encontrar a seqüência de estados que fosse mais provável de ter gerado a seqüência observada.

Um algoritmo com complexidade linear para fazer isso é o algoritmo Viterbi, o qual é também baseado na programação dinâmica. É um algoritmo muito similar ao algoritmo progressivo, porém ao invés de avaliar-se a soma em cada estado, é avaliado o máximo, ou seja:

$$v_j(t) = \underset{i=1}{\overset{n}{MAX}} \left[v_i(t-1) a_{ij} b_j(y_t) \right], \tag{4.5}$$

onde n é o número total de estados.

É fácil verificar que a partir do registro do caminho realizado pode-se facilmente recuperar a seqüência de palavras.

4.3.3 O algoritmo *Forward-Backward*

É um algoritmo cujo objetivo é treinar o modelo HMM. Consiste em otimizar a e b em relação a probabilidade do modelo de gerar todas as seqüências de saídas do conjunto de treinamento. Esta otimização é difícil de se obter e não se conhece um algoritmo direto para isso. Assim, utilizam-se iterações, ou seja, inicia-se a e b com algum valor e iterativamente, após avaliação da probabilidade, altera-se seus valores até que a probabilidade obtida seja considerada razoável ou até que um determinado número finito de iterações seja realizado. O algoritmo *Forward-Backward*, também conhecido como algoritmo *Baum-Welch*, é baseado neste técnica.

Na exposição do algoritmo progressivo foi definido $\alpha_j(t)$ como a probabilidade de gerar uma seqüência parcial y_1^t para o estado j no tempo t. Define-se agora sua imagem, $\beta_j(t)$, como a probabilidade de gerar o remanescente da seqüência y_{t+1}^T, iniciando do estado j no tempo t. Chamaremos de termo progressivo $\alpha_j(t)$ e termo regressivo $\beta_j(t)$. Assim como $\alpha_j(t)$, $\beta_j(t)$ pode ser avaliado recursivamente, mas desta vez na direção contraria, ou seja:

$$\beta_j(t) = \sum_{k=n}^{1} \beta_k(t+1) a_{ik} b_k\left(y_{t+1}\right), \tag{4.6}$$

onde n é o número total de estados no modelo HMM.

A iteração é inicializada no tempo T, fazendo $\beta_k(T) = 1$ para o estado final e 0,0 para os outros estados. Define-se também $\gamma_{ij}(t)$ como a probabilidade de ocorrer uma transição do estado i para o estado j no tempo t para uma dada seqüência de saída, ou seja:

$$\gamma_{ij}(t) = P\left(i_t \to j \middle| y_1^T\right) = \frac{P\left(i_t \to j, y_1^T\right)}{P\left(y_1^T\right)} = \frac{\alpha_i(t) a_{ij} b_j\left(y_{t+1}\right) \beta_j(t+1)}{\sum_{k=n}^{1} \alpha_k(T)}, \tag{4.7}$$

onde n é o número total de estados no modelo HMM.

Seja agora $N\left(i \to j\right)$ o número esperado de vezes que a transição do estado i para o estado j, para todo tempo t, ou seja:

$$N\left(i \to j\right) = \sum_{t=1}^{T} \gamma_{ij}(t), \tag{4.8}$$

e desta forma o número esperado de vezes que o estado i seja visitado será dado por:

$$N(i) = \sum_{j=1}^{n} \sum_{t=1}^{T} \gamma_{ij}(t), \tag{4.9}$$

onde n é o número total de estados do modelo.

O número de vezes que o estado i produz o símbolo u será dado então por:

$$N(i,u) = \sum_{t=1,(y_t=u)}^{T} \sum_{j=1}^{n} \gamma_{ij}(t) \tag{4.10}$$

Finalmente, uma primeira aproximação para a e b, produzindo $\overline{a}$ e $\overline{b}$ será dada por:

$$\overline{a_{ij}} = P(i \to j) = \frac{N(i \to j)}{N(i)} = \frac{\sum_{t=1}^{T} \gamma_{ij}(t)}{\sum_{j=1}^{n} \sum_{t=1}^{T} \gamma_{ij}(t)} \tag{4.11}$$

$$\overline{b_i(u)} = P(i,u) = \frac{N(i,u)}{N(i)} = \frac{\sum_{t=1,(y_t=u)}^{T} \sum_{j=1}^{n} \gamma_{ij}(t)}{\sum_{t=1}^{T} \sum_{j=1}^{n} \gamma_{ij}(t)} \tag{4.12}$$

Pode ser provado que fazendo $a = \overline{a}$ e $b = \overline{b}$, e repetindo-se o procedimento, a probabilidade $P(y_1^T)$ convergirá para máximos locais, pressionando os parâmetros do modelo HMM em direção à sua otimização.

4.4 Aplicação ao espaço acústico

De forma a aplicar o modelo HMM a sinais de voz são necessárias modificações no modelo básico acima descrito. As três variações mais utilizadas e populares são:

a) Modelo de densidade discreto

Neste modelo o espaço acústico é dividido em um certo número de regiões através de um procedimento de aglomeração conhecido como quantização vetorial (VQ - Vector Quantization). É comum dividir-se em 256, 512 ou 1024 subespaços. O centróide de cada aglomerado é representado por um elemento denominado ***codeword***. O conjunto dos todos os ***codeword*** definem o ***codebook***. O problema deste modelo são os erros de quantização quando o tamanho do ***codeword*** é pequeno e a diminuição da quantidade de informação acústica para treinamento do modelo quando o tamanho do ***codeword*** é grande.

b) Modelo de densidade contínuo.

Os erros de quantização podem ser eliminados usando o modelo de densidade contínua, ao invés de ***codeword***. Uma forma de fazer-se isso é assumir-se que o espaço acústico tenha uma certa forma paramétrica. Em geral utiliza-se uma mistura de

gaussianas para realizar esta parametrização. O problema desta abordagem consiste no fato de os dados não serem compartilhados entre os diversos estados, tal que se o número total de estado for muito grande, será exigido um número muito grande de gaussianas para que o modelo seja treinado.

c) Modelo híbrido.

Também conhecido como modelo de mistura amarrada (TMM - Tied-Mixture Model), consiste na utilização de gaussianas para a representação dos aglomerados, evitando-se desta forma os erros de quantização.

Os três modelos acima são grandemente utilizados, embora o modelo contínuo apresente melhores resultados para um vocabulário grande de palavras.

Nos sistema de RAV utiliza-se normalmente vários conjuntos de parâmetros que representam o sinal de voz. Os mais utilizados são baseados no espectro de potência do sinal janelado, ou seja, os coeficientes cepstrais, os delta cepstrais, a energia e a variação da energia. Mesmo sendo possível concatena-los em um único vetor, é preferível trata-los independentemente, tal que o conjunto seja mais coerente e possa ser modelado com um número mínimo de parâmetros. Assim é necessário modificar os modelos supracitados para que contemplem múltiplos conjuntos de parâmetros, se ou seja:

$$b_j(u) = \prod_{i=1}^{N} b_j(u_i), \tag{4.13}$$

onde u_i são os vetores de parâmetros de N independes conjuntos.

Apesar de estar o HMM revolucionando a tecnologia de RAV, estes modelos tem pontos fracos. O primeiro ponto fraco reside no fato de assumir-se que todas as probabilidades dependem exclusivamente do estado atual e isso não vale para sinais de voz. Com isso, por exemplo, esses modelos não conseguem representar adequadamente as coarticulações[21]. Além disso, a duração é modelada de forma não precisa pela distribuição com decaimento exponencial ao invés de uma distribuição mais precisa como a Poisson e outras. O segundo ponto fraco reside no fato de assumir-se que não há correlação entre segmentos adjacentes. O terceiro ponto fraco são os modelos de densidade de probabilidade utilizados que não conseguem representar fielmente a verdadeira densidade do espaço acústico. Finalmente a técnica que busca os parâmetros, baseada na busca iterativa, além de ser ineficiente, não consegue uma boa discriminação dos modelos acústicos. Para amenizar este último ponto fraco, poder-se-ia realizar o treinamento através da máxima informação mútua, porém esta é muito mais complexa e difícil de implementar.

[21] **Coarticulação**: Entende-se aqui coarticulação como a influência sobre um determinado fonema, dos fonemas anteriores e posteriores de uma declaração sonora.

4.5 Conclusão

Neste capítulo foi apresentado os modelos escondidos de Markov que consistem da mais flexível e de maior êxito técnica usada no RAV. Apresentou-se os conceitos básicos e descreveu-se os principais algoritmos utilizados no treinamento e teste dos modelos. Finalmente, discutiu-se como aplicar o modelo escondido de Markov ao espaço acústico, os algoritmos de treinamento e algumas de suas deficiências.

5 Reconstrução da Dinâmica de Séries Temporais

5.1 Introdução

Pode-se classificar os sinais em: sinais estacionários, quase-estacionários e transientes. Um sinal é estacionário se suas propriedades estatísticas são invariantes no tempo. A ferramenta adequada para estudar esse tipo de sinal é a transformada de Fourier, uma vez que as funções harmônicas (senos e cosenos) tem suporte global, ou seja, é possível decompor o sinal canonicamente em uma combinação linear de senos e cosenos. Sinais não estacionários (quase-periódicos e transientes) não podem ser analisados pela transformada de Fourier. Neste caso, a ferramenta mais indicada é a transformada *Wavelet*. Utiliza-se *wavelets* do tipo tempo-frequência para os sinais quase-periódicos e wavelets do tipo tempo-escala, para os sinais tendo uma estrutura fractal ou caóticos [DAU 92].

Os algoritmos utilizados para a caracterização de séries temporais obtidas de sinais caóticos, são usualmente testados em mapas e fluxos com dinâmica conhecida. No caso de mapas, simula-se o efeito de uma série temporal utilizando uma única variável, a partir da qual o atrator[22] é reconstruído. Contudo, o número de pontos na

série experimental é normalmente reduzido, tipicamente alguns milhares e muitas vezes insuficiente. Além disso, não se tem controle sobre o nível de ruído presente e nem sempre é possível manter a série somente com amostras do estado estacionário do sistema físico associado. Os procedimentos mais usados neste tipo de série são: o cálculo de dimensões de atratores; a determinação da entropia de Kolmogorov; o espectro de expoentes de Lyapunov; a reconstrução da dinâmica; e a redução de ruído [BRO 86].

Métodos clássicos tais como a análise espectral, ondaletas (*wavelets*) e a função de autocorrelação são normalmente utilizados para analisar as séries temporais regulares, mas podem ser úteis na caracterização de séries com comportamento caótico. Desta forma, revisa-se a seguir estes métodos, adaptados ao caso particular de séries temporais.

A seção 5.2 formaliza o que vem a ser transformada de Fourier. A seção 5.3 apresenta a transformada Gabor. A seção 5.4 apresenta a transforma *Wavelet*. A seção 5.5 apresenta a função de autocorrelação e como esta pode ser utilizada para avaliar o grau de semelhança de um sinal à medida que o tempo evolui. A seção 5.6 discute as séries caóticas e estocásticas. Finalmente, a seção 5.7 apresenta o procedimento de Takens para a reconstrução do atrator.

5.2 Transformada de Fourier

[22] **Atrator**: Conjunto de pontos no espaço de fase visitado pela solução das equações que representam a dinâmica de um sistema, após os transientes. Um atrator pode ter dimensão inteira (**atrator regular**) ou uma dimensão fracionária (**atrator estranho**).

Seja uma série temporal resultado de uma série de medidas realizadas a intervalos de tempos regulares Δt dada por $\{x_j\} = x(t_j)$, onde $t_j = j\Delta t$. Essa série pode ser representada pela superposição de componentes periódicas, ou seja, pode ser decomposta em uma soma de funções sinusoidais, as quais são bem localizadas na freqüência. Desta forma, a determinação do peso de cada uma dessas componentes é chamada análise espectral.

A transformada de Fourier de uma série temporal é definida por uma outra série $\{\hat{x}_k\}$ tal que

$$\{\hat{x}_k\} = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} x_j e^{\left[i\frac{2\pi jk}{N}\right]} \tag{5.1}$$

A transformada inversa reconstrói o sinal original $\{x_j\}$, isto é,

$$\{x_j\} = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \hat{x}_k e^{\left[-i\frac{2\pi kj}{N}\right]} \tag{5.2}$$

O espectro de potências $P(w)$ é definido como o módulo quadrado de $\{\hat{x}_k\}$, ou seja,

$$P(w) = \left|\hat{x}_k\right|^2 \tag{5.3}$$

O cálculo da transformada de Fourier pode ser feito com rapidez através do uso de algoritmos conhecidos como FFT. Estes algoritmos são de complexidade $O(n\log(n))$. Uma descrição detalhada destes algoritmos podem ser encontradas em [DUH 90].

5.3 Transforma Gabor

A transformada de Fourier fornece a representação do conteúdo de freqüência, mas não dá nenhuma informação sobre a localização no tempo; por exemplo, um variação brusca, bem localizada na tempo, altera todos os elementos na freqüência.

Um alternativa para resolver o problema é multiplicar-se o sinal por uma função janela, delimitando-o no tempo, e aplicar-se a transformada de Fourier ao sinal janelado. Analisa-se o espectro de Fourier pela posição da janela no tempo. Este tipo de análise é conhecido por STFT (Short Time Fourier Transform). No caso particular de a janela ser uma janela Gaussiana, tem-se a transformada Gabor [GAB 46].

5.4 Transformada Wavelet

Define-se uma família de funções:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right), \qquad a > 0,\ b \in R, \tag{5.4}$$

onde Ψ é uma função fixa, denominada ***wavelet* mãe**, a qual é bem localizada no tempo e na freqüência. Define-se a transformada *wavelet* $Tf(a,b)$ como:

$$Tf(a,b) = \left\langle f, \Psi_{a,b} \right\rangle = \frac{1}{\sqrt{a}} \int f(t)\overline{\Psi\left(\frac{t-b}{a}\right)}dt \tag{5.5}$$

onde f(t) é a função da qual se deseja fazer a decomposição.

A existência de uma transformada inversa depende da escolha de Ψ. Mais precisamente se Ψ é tal que:

$$C_\Psi = \int_{-\infty}^{+\infty} \frac{\left|\hat{\Psi}(\omega)\right|^2}{|\omega|} < +\infty, \tag{5.6}$$

então f(t) pode ser reconstruído através de:

$$f(t) = C_\Psi^{-1} \int_0^{+\infty} \frac{da}{a^2} \int_{-\infty}^{+\infty} Tf(a,b)\Psi_{a,b}(t)db \tag{5.7}$$

Existem muitos trabalhos na literatura que tratam de *Wavelets*. Porém o primeiro livro em português sobre o assunto só saiu em julho de 1997, no 21º. Colóquio Brasileiro de Matemática [GOM 97].

Existem muitas técnicas que permitem fazer a decomposição atômica de sinais. O trabalho descrito em [CUS 97], propõe que se utilize o paradigma de algoritmos genéticos para esta decomposição. Isto facilita a utilização de átomos que tenham algum significado físico de interesse.

5.5 Função de autocorrelação de um sinal

Seja ϕ_m a média do produto dos valores do sinal $x(t)$ nos instantes t e $t + m\Delta t$,

$$\phi_m = \frac{1}{N}\sum_{j=1}^{N} x_j x_{j+m} \tag{5.8}$$

Essa função indica por quanto tempo o valor do sinal no instante t depende de seus valores prévios ou ainda o grau de semelhança existente no sinal à medida que o tempo passa. Para x_j periódico com período N, tem-se $\phi_m = \phi_{m+N}$.

Pode-se mostrar que:

$$P(w) = \left|\hat{x}_k\right|^2 = \sum_{m=1}^{N} \phi_m \cos\left(\frac{2\pi mk}{N}\right) \tag{5.9}$$

ou seja, o espectro de potências é proporcional à transformada de Fourier da função de autocorrelação.

As séries temporais oriundas de medidas de um sistema físico real possuem diferentes espectros de potências. Um pico numa determinada freqüência indica que o sinal é periódico de período T correspondente a esta freqüência.

Em geral não se conhece esse período T, e desta forma não se pode escolher o tempo total de medida como um múltiplo de T e a resolução dos picos de $P(w)$ fica comprometida. Ao invés de um pico na freqüência $w = \frac{2\pi}{T}$, tem-se um pico mais largo seguido de picos secundários conhecidos como *side-lobes*. O espectro de potências de um sinal periódico de período T é portanto composto de um pico na freqüência $w = \frac{2\pi}{T}$, *side-lobes* e picos menores nos harmônicos, também seguidos de *side-lobes*.

Se o sinal é não periódico, o resultado da superposição de r sinais periódicos de períodos incomensuráveis[23], tem-se que o espectro de potências será formado por picos nas freqüências

$$f_n = \sum_{j=1}^{r} \alpha_j \omega_j \text{ ,} \tag{5.10}$$

onde α_j são inteiros arbitrários e ω_j são as freqüências associadas aos períodos de cada sinal independente.

Um sinal com espectro de potência contínuo define um sinal caótico ou estocástico. Quando o sinal é composto por muitos sinais periódicos com um número grande de freqüências independentes, obtém-se, na prática, um espectro de potências que parece contínuo.

A Tab. 5.1 descreve o comportamento da função de autocorrelação de alguns sinais.

TABELA 5.1 - Comportamento da função de autocorrelação por tipo de sinal

Sinal	**Comportamento de ϕ_m**
periódico ou quase-periódico	• permanece diferente de zero quando o tempo tende ao infinito • é igualmente periódica
multi-periódico	se confunde com aquela de um sinal caótico
Caótico	Caótico
Estocástico	Estocástico

[23] **Período Incomensurável**: A relação entre as freqüências é um número irracional.

5.6 Séries Caóticas e Estocásticas

Um modelo capaz de descrever a dinâmica que deu origem a uma série temporal só pode ser obtido caso a série tenha algum grau de caos determinístico. Uma série estocástica, em geral associada à presença de ruído, só pode ser descrita por métodos probabilísticos. Um sinal completamente aleatório é conhecido como ruído branco. O espectro de potências do ruído branco forma um patamar onde as amplitudes são independentes das freqüências. Um caso particular muito importante e bastante estudado é o conhecido processo $1/f$, ou seja, sinais com espectro de potências inversamente proporcional à freqüência. Séries temporais oriundas da amostragem de um processo físico real, normalmente, tem um espectro com esse comportamento para muitas décadas de freqüência. Uma lista parcial de fenômenos naturais que exibem este comportamento, tirada de [WOR 96], inclui:

a) Série temporal geográfica tais como a variação da temperatura e pluviosidade, medidas de corrente oceânicas, variação do nível de inundação do rio Nilo, oscilação do eixo terrestre, freqüência de variação da rotação da Terra, e variações nas manchas solares;
b) Série temporal econômica tal como a Média Industrial *Dow Jones*;
c) Série temporal fisiológica tal como a taxa instantânea de batida do coração em pacientes cardíacos, e variações do eletroencefalograma (EEG) sob condições de estímulos;
d) Série temporal biológica tal como a tensão elétrica nos nervos e membranas sintéticas;
e) Flutuações eletromagnéticas tais como na radiação ruidosa galáctica, na intensidade de fontes de luz, e no fluxo de corrente em supercondutores;
f) Ruído em dispositivos eletrônicos como o transistor bipolar e de efeito de campo, válvulas e diodos;
g) Resistência de flutuação em filmes metálicos, filmes semicondutores e contatos, e termocélulas;
h) Variação na freqüência em relógios com osciladores de cristal de quartzo, atômicos, e em ressonadores com cavidade supercondutora;
i) Fenômenos induzidos pelo homem incluindo variações no fluxo de tráfico e variação em amplitude e freqüência nas músicas do oeste americano, africana, asiática e indiana, moderna e tradicional;
j) Padrões de erros de transmissão em canais de comunicação;
k) Variações na textura nos terrenos naturais, paisagens e formação de nuvens.

5.7 Reconstrução do Atrator

Para que se possa analisar as propriedades de um possível atrator associado a uma série temporal com comportamento caótico determinístico é necessário em primeiro lugar reconstrui-lo num espaço de fases de dimensão conveniente. Um dos primeiros trabalhos de como realizar esta reconstrução foi proposto por [PAC 80]. Esse método tinha problemas, pois o cálculo das derivada amplificava os erros experimentais e o algoritmo tornava-se pouco prático, principalmente se o número de variáveis independentes envolvido fosse grande. Esses problemas foram resolvidos através do

procedimento proposto por Takens [TAK 81] que permite reconstruir certas propriedades topológicas do atrator a partir da série temporal $\{x_j\}$ por meio do vetor

$$\xi_i = \{x(t_i), x(t_i + p), \ldots, x(t_i + (m-1)p)\}$$

onde m é a dimensão de imersão e p é o passo de reconstrução. O método de Takens é também chamado método dos atrasos temporais. Embora o atrator reconstruído não seja idêntico ao original, as propriedades topológicas são preservadas. A dimensão m do espaço de fases reconstruído não precisa ser necessariamente idêntica à dimensão D do espaço de fases real dos vetores x_i que representam a dinâmica da série. O número total de vetores obtidos através desta reconstrução é dado por:

$$S = \frac{1}{2}\big(N - mp\big)\big(N - (m-1)\big) \tag{5.11}$$

Para $m, p << N$, então

$$S \approx \frac{1}{2} N^2 \tag{5.12}$$

Assim, para uma série temporal de 10.000 pontos, tem-se aproximadamente $5x10^7$ vetores.

Em geral é necessário reconstruir-se o atrator em espaços de fases com dimensão suficientemente elevada ($m > 2D_0 + 1$, onde D_0 é a dimensão de Hausdorff [FED 88] do atrator). Na maioria dos casos práticos, mesmo para $m < 2D_0 + 1$, obtém-se bons resultados.

A análise tradicional de sinais baseada na análise do espectro de potências ou da função de autocorrelação não permite, em geral, a distinção entre uma dinâmica caótica determinística e comportamento estocástico. Vários procedimentos têm sido desenvolvidos com essa finalidade, inclusive nos casos em que não se sabe, ou não é possível, modelar ou descrever a dinâmica em termos de equações diferenciais ou mapas.

Através da técnica de reconstrução de Takens, Grassberger-Procaccia propôs um método numérico muito eficiente para a estimativa da dimensão de correlação D_2. Uma dimensão D_2 muito elevada normalmente indica que a série é estocástica. Outra técnica muito utilizada é a estimativa da entropia de Kolmogorov-Sinai. Há também o método proposto em [KAP 92], o qual verifica a existência de uma dinâmica determinística em uma série temporal partindo do vetor de Takens e através dele reconstruindo-se não somente o atrator, mas o vetor associado à direção do fluxo em espaços de fases de dimensões crescentes. O espaço de fases é dividido em caixas e procede-se a análise estatística do valor médio do vetor associado à direção do fluxo em cada caixa. Para um sinal estocástico essa média é próxima de zero. Isso não ocorre em sinais determinísticos, onde se espera uma significativa correlação entre os vetores numa mesma caixa. Outro método que também utiliza a reconstrução de Takens é o método proposto por [WOL 85], a partir do qual pode-se determinar o maior expoente de Lyapunov positivo, cuja presença indica ser a série caótica.

5.7.1 A Escolha do Passo

Takens demonstrou que para um número infinito de pontos e na ausência de ruído, a escolha do passo de reconstrução p é na grande maioria dos casos arbitrária. Entretanto, as séries temporais experimentais são finitas, usualmente contaminadas com ruído externo e obtidas com o uso de filtros. Nessa situação a reconstrução depende, e muito, da escolha correta do passo. Se o passo p for muito pequeno, $x(t)$ e $x(t+p)$ terão praticamente o mesmo valor. Como conseqüência, o atrator com dimensão maior que dois reconstruído em uma dimensão de imersão igual a dois (m = 2), por exemplo, fica comprimido em torno da diagonal $y = x$, já que $\xi_1 \approx \xi_2$, ou seja, esse atrator apresentará uma dependência linear entre ξ_1 e ξ_2, que não ocorre nas componentes reais x e y. Por outro lado, como a trajetória real está restrita a um volume finito do espaço de fases, o passo p não pode ser muito grande, sob pena dos vetores reconstruídos serem completamente descorrelacionados.

Inúmeros critérios tem sido propostos na literatura para orientar a escolha do passo correto. Um dos mais simples e talvez por isso muito difundido, sugere que se use um passo p da ordem do tempo de autocorrelação τ do sinal, que é definido por

$$\phi_\tau = \frac{1}{2}\phi_m, \tag{5.13}$$

para $m = 0$, onde ϕ é a função de autocorrelação definida pela equação (5.8). O passo p é definido de modo que $p \approx \tau$. Tal critério garante que x_i e x_{i+p} sejam linearmente independentes, mas não completamente descorrelacionados.

Um segundo critério, proposto por [FRA 86], escolhe-se o passo $p \approx \tau'$, onde τ' é o intervalo de tempo que minimiza a informação mútua contida em vetores vizinhos ao longo de uma trajetória de evolução. O método é de mais difícil implementação numérica, consumindo um tempo significativamente maior de computação e inadequado para séries temporais com poucos pontos.

Um terceiro método, conhecido como método do fator de preenchimento, é de fácil implementação e usa usa argumentos geométricos para completar e tornar mais precisa a informação obtida pelo método do tempo de autocorrelação. O procedimento baseia-se no fato que um atrator reconstruído com um passo pequeno demais colapsa sobre a hiper-diagonal do espaço de fases reconstruído e ocupa, portanto, um volume mínimo desse espaço, enquanto que um passo muito grande gera atratores que ocupam hiper-volumes muito grandes. Assim calcula-se a função fator de preenchimento, que estima a parcela do espaço de fases reconstruído ocupada pelo atrator. O valor do passo é obtido como o intervalo de tempo que seja, ao mesmo tempo, menor que o pseudo-período e maximize o fator de preenchimento.

O chamado **diagrama de primeiro retorno**, que permite avaliar-se a sensibilidade do processo de reconstrução em relação ao passo consiste em fazer o gráfico de $x(t_i)\,\mathrm{X}\,x(t_{i+p})$, reconstruindo-se para tanto vetores bidimensionais com passo p. Uma simples inspeção visual desse gráfico nos fornece informações sobre os valores de p para os quais $x(t_i)$ e $x(t_{i+p})$ estão ainda fortemente correlacionados, situação em que o atrator reconstruído fica comprimido próximo à diagonal, e nos diz quão sensível é a reconstrução do atrator relativamente à escolha do passo, quão homogêneo é o

atrator e qual o número típico de vetores reconstruídos necessário para caracterizá-lo completamente.

A título de exemplo, a Fig. 5.1 mostra o diagrama de primeiro retorno para a equação logística:

$$x_{n+1} = 4x_n\left(1 - x_n\right) \tag{5.14}$$

Pode-se verificar na Fig. 5.1 que para p=1, os pontos já não estão mais correlacionados.

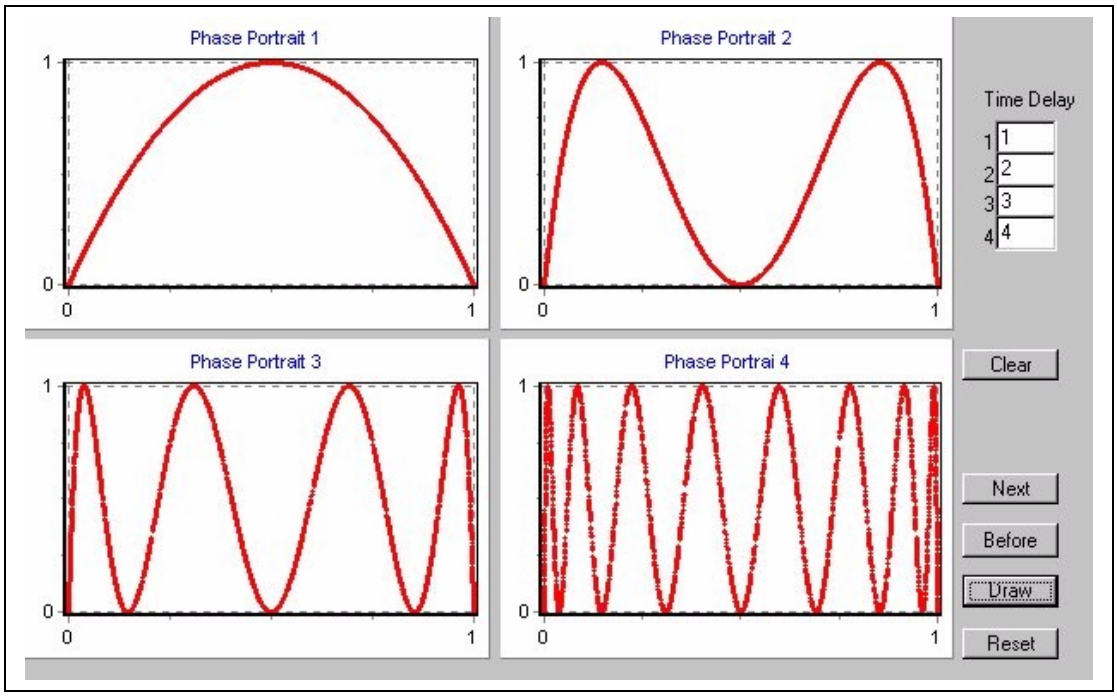


FIGURA 5.1 - Diagrama de primeiro retorno para a equação logística

5.7.2 A Escolha da Dimensão de Imersão

Foi visto, através da equação (5.11), que o número total de vetores S de reconstrução de Takens diminui na medida em que m e p crescem. Contudo se o número de pontos for grande, conforme a equação (5.12), o número de vetores independe de m e p. Isso faz com que a maioria dos algoritmos exija uma eficiência computacional muito elevada para dimensões de imersão maiores que 2 ou 3 [FER 94].

Em alguns casos, como nos algoritmos para cálculo da dimensão de correlação ou da entropia de Kolmogorov, deve-se reconstruir o atrator sucessivamente em espaços de imersão de dimensões crescentes. A comparação dos resultados obtidos em cada uma destas dimensões permitirá a analise dos valores finais das grandezas que se deseja obter. Em geral, as características dos dados experimentais devem ser tais que permitam a reconstrução com estatística suficiente até espaços de imersão com dimensão da ordem de $m > 2D_0 + 1$. Infelizmente, tanto informações sobre o valor de D_0, como sobre o número de pontos na série temporal necessários a uma estatística suficiente só são possíveis depois de um razoável esforço computacional. Outro agravante é que na maioria dos casos práticos dispõe-se de uma série temporal de tamanho fixo, ou seja,

não se pode obter mais amostras para a série. Assim, é necessário o uso de algoritmos que possibilitem esta reconstrução a partir dos dados disponíveis.

Para sistemas com baixa dimensionalidade, de uma ou duas dimensões, a inspeção visual continua sendo o método mais conveniente para a escolha da dimensão de imersão. No entanto, existem propostas de se formalizar métodos para estimar a dimensão de imersão mínima necessária à reconstrução de um atrator. Uma dessas propostas é conhecida pelos nomes de processo de Kahunen-Loeve, análise de fatores, análise da componente principal ou ainda decomposição por valor singular, a qual consiste em realizar uma transformação de coordenadas, passando das coordenadas originais usadas para reconstruir o atrator a um sistema no qual as variáveis de estado tenham a mínima correlação. Para isso diagonaliza-se a chamada matriz de covariância ou matriz de correlação. Essa transformação de variáveis leva o atrator para uma representação onde as variáveis de estado são estatisticamente independentes.

A dimensão de imersão do atrator reconstruído é estimada pela simples contagem dos autovalores não nulos da matriz de correlação. Além disso, deve ser estipulado um limiar inferior abaixo do qual o autovalor será considerado nulo.

Há também o método sugerido em [KEN 92] que parece ser bastante promissor. O método consiste em contar o número de falsos vizinhos para cada um dos seus pontos. Quando o número de falsos vizinhos cai a zero o atrator terá sido suficientemente "desdobrado" e é possível, desta forma, identificar a menor dimensão de imersão capaz de representá-lo adequadamente. A identificação dos falsos vizinhos pode ser feita através de dois testes. Primeiro, seja $R_m^2(n)$ o quadrado da distância entre um dado ponto ξ_n e o seu ponto mais próximo ξ_{n^*} num atrator reconstruído com dimensão m. Se ξ_{n^*} é um falso vizinho, $R_m^2(n)$ provavelmente aumentará muito quando se passa da dimensão m para a dimensão m+1. Desta forma ξ_{n^*} é um falso vizinho se:

$$\sqrt{\left[\frac{R_{m+1}^2(n) - R_m^2(n)}{R_m^2(n)}\right]} > L_c \tag{5.15}$$

onde L_c é uma distância crítica. Esse critério estabelece uma condição necessária mas não suficiente para se identificar os falsos vizinhos. Segundo, devido ao número finito de pontos com os quais se lida na prática, pode-se ter situações onde o vizinho mais próximo está a uma distância da ordem do tamanho do atrator. Pode-se mostrar que se $R_m(n) \approx L_A$ (L_A é um tamanho típico do atrator , então $R_{m+1}(n) \approx 2L_A$ para um falso vizinho. Assim estabelece-se a condição adicional:

$$\frac{R_{m+1}(n)}{L_A} \approx A_c, \tag{5.16}$$

onde A_c é um limite crítico. Um vizinho é considerado falso se obedecer estas duas condições.

Apesar dos vários métodos disponíveis para a determinação da dimensão de imersão apropriada à reconstrução, a baixa sensibilidade dos resultados finais com relação a pequenas variações em m é o indicador mais seguro de que a reconstrução faz sentido e é robusta.

5.7.3 Cuidados Durante a Reconstrução

Os seguintes cuidados devem ser tomados quando se procede ao estudo de uma série temporal experimental:

a) A série temporal deve corresponder à amostragem de um estado estacionário do sistema físico real;
b) Uma quantidade razoável de amostras, normalmente maior que 10.000, de forma que o atrator seja visitado várias vezes, ou seja, que haja pontos suficientes para representar estatisticamente o atrator;
c) Que a série seja amostrada a uma freqüência tal que leve em consideração o teorema da amostragem [GAB 46].

O número total de amostras deve ser maior se o atrator for não-homogêneo. Na prática só se tem informações sobre a homogeneidade do atrator após se despender um razoável esforço computacional para se proceder às reconstruções e calcular dimensões ou probabilidades. A inspeção visual do atrator feita através do diagrama de primeiro retorno é o único recurso disponível para se ter alguma idéia de antecipada do grau de uniformidade do possível atrator. O método é limitado uma vez que se analisam apenas reconstruções bidimensionais. A projeção bidimensional de uma atrator tridimensional, por exemplo, pode ser percorrido de forma bastante homogênea, sem que o atrator o seja.

Em relação ao número de pontos na série experimental, já foram relatados, na literatura, resultados bastante bons obtidos a partir de cerca de 1.000 pontos. Contudo é mais freqüente operar-se com cerca de 10.000 pontos na série temporal.

Em geral é possível mostrar-se que existe um limite máximo para a dimensão do atrator calculada a partir de um número finito de pontos. Eckmann e Ruelle [ECK 92] demonstraram que não se deve estimar dimensões maiores que:

$$d_{max} = \frac{2\log N}{\log\left(1/\rho\right)} \tag{5.17}$$

onde

N é o número de pontos da série temporal

$\rho = \frac{\varepsilon}{D}$

D é o diâmetro do atrator

ε é a distância mínima entre os vetores reconstruídos pela técnica de Takens

A Tab. 5.2 mostra quantos pontos são necessários para se reconstruir atratores com dimensão até 10, considerando $\rho = 0{,}1$. Nesta tabela também é mostrado o número aproximado de vetores de Takens e a quantidade de memória necessária para armazenar este vetor, considerando a utilização da representação em ponto flutuante padrão IEEE *double*, ou seja, 8 bytes.

TABELA 5.2 - Pontos, vetores e memória para reconstrução de Takens

Dimensão	Número de Pontos	Número de Vetores	Memória
4	100	5.000	40 Kbytes
5	320	50.000	400 Kbytes
6	1000	500.000	4 Mbytes
7	3200	5.000.000	40 Mbytes
8	10.000	50.000.000	400 Mbytes
9	32.000	500.000.000	4 Gbytes
10	100.000	5.000.000.000	40 Gbytes

As restrições são ainda mais drásticas no caso do cálculo de expoentes de Lyapunov, onde torna-se necessário cerca do dobro de pontos em relação à estimativa da dimensão de correlação.

A freqüência de amostragem deve ser suficientemente alta para que se registre toda a estrutura fina do sinal a ser analisado. Em geral deve-se ter pelo menos 10 pontos num período de correlação ou pseudo-período do sinal. O intervalo de tempo entre duas medidas consecutivas deve ser escolhido de tal forma que a distância média entre pontos sucessivos da trajetória seja maior que a distância média entre pontos vizinhos sobre o atrator. De forma resumida tem-se que a freqüência de amostragem dependerá:

a) Do número total de pontos;
b) Da possível dimensão do atrator;
c) Das características métricas do atrator.

5.8 Redução de Ruído

A utilização da reconstrução de Takens no cálculo da dimensão fractal e do expoente de Lyapunov pressupõe que a série temporal não esteja contaminada por ruído. Contudo, as séries temporais oriundas de medidas de sistemas reais, normalmente apresentam algum grau de contaminação de ruído. Basicamente, existem duas classes de ruído, do qual se tem interesse, nesse tipo de séries:

a) Ruído localizado em freqüência;
b) Ruído com comportamento caótico.

Para “limpar” a série contaminada com ruído localizado em freqüência basta filtrá-la na região do espectro onde localiza-se o ruído[DAV 97].

Já as séries contaminadas com ruído que apresentam comportamento caótico podem ser limpas de duas formas diferentes, as quais são descritas a seguir:

a) Caso se conheça de antemão a dinâmica do sinal sem ruído, a limpeza por ser feita da seguinte forma:

Ruído = Sinal Medido em (t+1) - Dinâmica Aplicada ao Sinal Medido em t;

b) Caso a dinâmica não seja conhecida, campo de intensa pesquisa atualmente, utilizam-se procedimentos conhecidos como sombreamento. O sombreamento consiste em obter-se uma dinâmica para o sinal através da expansão em série em alguma base, do conjunto de mapas acoplados obtidos a partir da reconstrução de Takens. Essas aproximações podem ser feitas em locais específicos da série, como também de forma global. As expansões mais utilizadas são: séries de Taylor, razões entre polinômios funções de base radiais e redes neurais. Não significa, contudo, que esta dinâmica seja uma representação real do sistema. Busca-se apenas uma representação com plausibilidade física para este possível comportamento. Uma vez obtida a dinâmica, pode-se reduzir a quantidade de ruído e em muitos casos ainda, aumentar o número de amostras da série.

5.9 Conclusão

Neste capítulo apresentou-se as principais ferramentas utilizadas na análise de séries temporais. Em primeiro lugar classificou-se as séries temporais em séries periódicas, quase-periódicas, caóticas e estocásticas, assim como as ferramentas adequadas ao estudo de cada um desses tipos de séries. O sinal de voz pode ser visto como uma série temporal caótica. Assim apresentou-se formalmente como pode-se proceder a reconstrução do possível atrator associado.

6 Dimensão Fractal

6.1 Introdução

A geometria fractal tem revolucionado a caracterização de estruturas auto-similares e fenômenos da natureza. Os fractais podem ser utilizados como um novo método para caracterizar estruturas aparentemente complexas e irregulares na natureza.

Quando se fala em dimensão logo vem a mente o conceito de dimensão euclidiana. Todavia, existem na natureza estruturas geométricas complexas que são melhor caracterizadas por uma dimensão não inteira, ou fracionária, ou ainda, fractais. Existem conjuntos ainda mais complexos que os fractais. Esses são conhecidos como multifractais. Um fractal pode ser caracterizado por uma única dimensão. Já um conjunto multifractal só pode ser caracterizado por um conjunto infinito de dimensões.

Fractais tem sido muito utilizado no modelamento de estruturas auto-similares tais como montanhas, nuvens, rochas e a auto-afinidade encontradas em ruído térmico, eletroencefalograma humano (EEG), música e recentemente em sons vocálicos. Os sinais biológicos são conhecidos por ter uma componente auto-afim a qual pode ser verificada através da dimensão fractal associada à propriedade auto-afim. Além disso, os atratores reconstruídos oriundos de sinais biológicos podem ser considerados ter propriedades caóticas. A existência de um expoente de Lyapunov positivo indica a presença de caos e as propriedades auto-similares do atrator caótico reconstruído também podem ser caracterizadas através da dimensão fractal relacionada a auto similaridade.

Pichover e Khorosani [PIC 86] foram os primeiros a afirmarem que existe uma certa auto-afinidade nos sons vocálicos e encontraram a dimensão fractal da forma de onda do sinal na ordem de $D = 1{,}66$.

Os sons vocálicos também são apontados como tendo caos em sua dinâmica, e assim um outro método de determinar as propriedades fractais destes sons é caracterizar as propriedade fractais do atrator caótico reconstruído da forma de onda do sinal de voz. A existência de caos pode ser verificada através do maior expoente de Lyapunov aplicado ao atrator reconstruído a partir do método de reconstrução de Takens. Um dos primeiros trabalhos publicados com uma estimativa deste maior expoente de Lyapunov foi o de [SAB 96], onde encontrou-se que todas as vogais da língua japonesa tem expoente de Lyapunov positivo. Além disso, também foi mostrado que todas as vogais da língua japonesa tem propriedades multi-fractais, indicando que o atrator reconstruído tem uma estrutura fractal não uniforme e complexa.

Normalmente um segmento de sinal de voz contém vários fonemas, vogais e outros sons. Na prática é muito difícil separar-se estas unidades de forma a poder caracteriza-las por uma única dimensão fractal ou multifractal global. Assim, assume-se que as propriedades fractais variam com o tempo. Com o objetivo de caracterizar efetivamente estes sons de voz, [SAB 96] propôs a dimensão fractal dependente do tempo (TDDS - *Time Dependent Fractal Dimension*) e as dimensões multifractais dependente do tempo (TDMFD - *Time Dependent Multifractal Dimension*), as quais mostram a variação da dimensão fractal ao longo do tempo.

As dimensões fractais são, por definição, insensíveis às escalas no tempo e amplitude, sendo portanto indicada como um bom parâmetro para o reconhecimento de voz. De fato, o trabalho de [SAB 96] mostrou que estas técnicas podem efetivamente

serem utilizadas para melhorar a qualidade de reconhecimento de voz dos sistemas existentes.

A seção 6.2 define formalmente o que vem a ser dimensão fractal. A seção 6.3 apresenta os algoritmos mais conhecidos para a estimativa desta dimensão. A seção 6.4 procede-se à generalização destes métodos com o objetivo de poder-se verificar a homogeneidade do possível atrator associado. E finalmente na seção 6.5 apresenta-se o algoritmo de Grassberger-Procaccia para essa generalização.

6.2 Definição de Dimensão Fractal

Seja um conjunto de pontos A. Define-se dimensão de Hausdorff-Besicovitch ou dimensão fractal como:

$$D_0 = \lim_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{\ln\left(1/\varepsilon\right)} \tag{6.1}$$

onde $N(\varepsilon)$ é o número mínimo de caixas de lado ε necessário para cobrir todo o conjunto de pontos A. Na prática, devido à limitação dos recursos computacionais, tal como a utilização de representação em ponto flutuante e um número finito de pontos, procura-se determinar o número de caixas com pelo menos um ponto do conjunto A. A dimensão calculada desta forma denomina-se **capacidade** e é uma boa aproximação para a dimensão fractal.

6.3 Cálculo da Dimensão Fractal

Existem muitos algoritmos diferentes para o cálculo da dimensão de atratores associados a séries temporais. Quase todos eles têm como ponto de partida a reconstrução proposta por Takens. Os mais conhecidos são:

a) Algoritmos de Contagem de Caixas;
b) Método dos Vizinhos Próximos;
c) Método da Integral de Correlação;
d) Método Singular;
e) Método do Expoente Crítico.

6.3.1 Algoritmos de Contagem de Caixas - ACC

Algoritmos de contagem de caixas são bastante populares e são adequados a cálculos com mapas e fluxos conhecidos, na medida em que necessitam de muitos pontos para conseguirem estimar as propriedades topológicas do possível atrator associado à série temporal. A Tab. 6.1 mostra os passos do algoritmo.

TABELA 6.1 - Passos do Algoritmo de Contagem de Caixas

Passo 1	Divide-se o espaço de fases ocupado pelo conjunto de pontos A em caixas de tamanho ε
Passo 2	$N(\varepsilon)$ vai ser o número de caixas com pelo menos um ponto do conjunto A
Passo 3	Repetir os passo 1 e 2 para diversos valores de ε
Passo 4	Determinação de D_0 como a inclinação do gráfico $\log N(\varepsilon) \, x \log\left(\frac{1}{\varepsilon}\right)$

Observa-se que a medida que ε diminui $N(\varepsilon)$ aumenta significativamente, aumentando a quantidade de memória e de desempenho computacional necessária para executar o algoritmo. Esse algoritmo torna-se completamente ineficiente para dimensões maiores que 2[FER 94].

6.3.2 Método dos Vizinhos Próximos - MVP

Um método muito estudado é o dos vizinhos próximos, o qual, aparentemente é mais adequado ao estudo de atratores com dimensão alta.

Este método consiste em cobrir-se o conjunto de pontos com caixas de modo que cada caixa tenha o mesmo número de pontos. A dimensão é obtida pelo estudo da variação do raio dos aglomerados ao longo do conjunto de pontos A.

6.3.3 Método da Integral de Correlação - MIC

O método da Integral de Correlação é o mais popular dos métodos por ser de fácil implementação. Também é conhecido com **algoritmo de Grassberger-Procaccia** ou dimensão de correlação.

Seja a Integral de Correlação dada por:

$$C(\varepsilon) = \frac{1}{N^2} \lim_{N \to +\infty} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} H\left(\varepsilon - \left| \vec{x}_i - \vec{x}_j \right| \right) \tag{6.2}$$

onde

$$H(x) = \begin{cases} 1 & se \quad x \geq 0 \\ 0 & se \quad x < 0 \end{cases} \tag{6.3}$$

A dimensão é dada por:

$$D_2 \sim \lim_{\varepsilon \to 0} \frac{\ln C(\varepsilon)}{\ln \varepsilon} \tag{6.4}$$

O método funciona muito bem para séries temporais longas. Contudo, na prática há limitações experimentais fortes para a obtenção de séries longas restringindo a utilização do algoritmo de Grassberger-Procaccia no caso de conjuntos de alta

dimensão. Na grande maioria das aplicações o algoritmo de Grassberger-Procaccia não permite estimar com segurança dimensões de correlação maiores que 4 ou 5, as quais correspondem a dimensões de imersão da ordem de 9 a 11.

6.3.4 Método Singular - MS

O Método Singular aparentemente é mais vantajoso no caso de séries temporais contaminados por um nível alto de ruído [FER 94].

6.3.5 Método do Expoente Crítico - MEC

Este método aproveita-se da característica de auto-afinidade que algumas séries temporais possuem, estimando a dimensão fractal a partir do momento associado ao espectro de potência do sinal dado por:

$$I_\alpha = \int_1^U duP(u)u^\alpha , \qquad (-\infty < \alpha < +\infty) \tag{6.5}$$

onde U é o limite superior da integração e u é a freqüência normalizada cuja freqüência de corte inferior corresponde a 1. Aqui α toma valores reais e geralmente varia sobre (-∞,+∞). Dentro do intervalo de integração, $P(u)$ é assumido como seguindo a lei da potência:

$$P(u) \approx u^{-\beta} \tag{6.6}$$

Isso é conseqüência da alto-afinidade da série temporal em consideração. Substituindo-se a equação (6.6) na equação (6.5) tem-se a seguinte equação para o momento:

$$I_\alpha \approx \int_1^U duP(u)u^{\alpha-\beta} = \int_1^U duu^{X-1} = \frac{1}{X}\left(U^X - 1\right) = \frac{2}{X}e^{\frac{\nu X}{2}} sinh\left(\frac{\nu X}{2}\right), \tag{6.7}$$

onde X e ν são definidos como

$$X = \alpha - \beta + 1 \tag{6.8}$$

e

$$\nu = \log U \tag{6.9}$$

Tomando-se a derivada terceira do logaritmo do momento I_α, tem-se:

$$\frac{d^3}{d\alpha^3}\log I_\alpha = -2\frac{1}{X^3} + \frac{1}{4}\nu^3 \cos ech^3\left(\frac{\nu X}{2}\right)\cosh\left(\frac{\nu X}{2}\right) = 0 \qquad (X = 0) \tag{6.10}$$

onde α_C é o valor crítico que satisfaz a equação acima. Da relação acima, o expoente β na equação (6.6) é dado como:

$$\beta = \alpha_C + 1 = 2H + 1 \tag{6.11}$$

onde H é o expoente de Hurst [FED 88].

Finalmente, a dimensão fractal D pode ser dada como:

$$D = 2 - H = 2 - \frac{\alpha_C}{2} \tag{6.12}$$

Levando-se em conta aspectos práticos, [NAK 96] propôs a seguinte fórmula para o cálculo da equação (26), dada a densidade espectral de potência P(u):

$$\frac{d^3}{d\alpha^3}\log I_\alpha = \frac{I_\alpha''' I_\alpha^2 - 3I_\alpha'' I_\alpha' I_\alpha + 2\left(I_\alpha'\right)^3}{I_\alpha^3}, \tag{6.13}$$

onde a n-ésima derivada de I_α, $I_\alpha^{(n)}$ pode ser avaliada da seguinte equação:

$$I_\alpha^{(n)} = \frac{d^n}{d\alpha^n}\int_1^U duu^\alpha P(u) = \int_1^U du(\log u)^n u^\alpha P(u) \tag{6.14}$$

A Fig. 6.1 mostra o comportamento típico da derivada terceira do momento espectral de uma série temporal auto-afim. Nesta figura o eixo das abcissas é o alfa e o das ordenadas é a derivada primeira do momento.

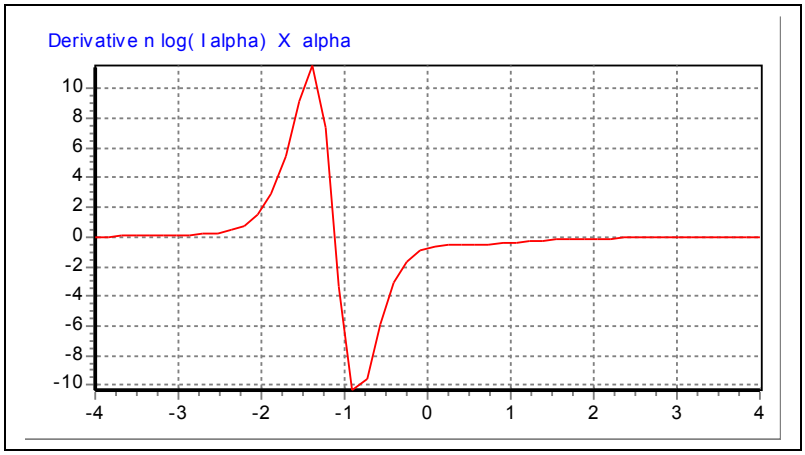


FIGURA 6.1 - Derivada Terceira do Momento Espectral

Após vários testes práticos e com o objetivo de melhorar o desempenho do algoritmo que determina α_C, pode-se utilizar as seguintes fórmulas alternativas:

a) Cálculo direto do $\log I_\alpha$. Neste caso o valor de α_C pode ser determinado pela interseção das duas retas que aproximam, por exemplo, o primeiro terço e o terceiro terço da curva α_C X $\log I_\alpha$. A Fig. 6.2 mostra o comportamento típico do momento espectral de uma série temporal auto-afim. Nesta figura o eixo das abcissas é o alfa e o das ordenadas é a derivada primeira do momento.

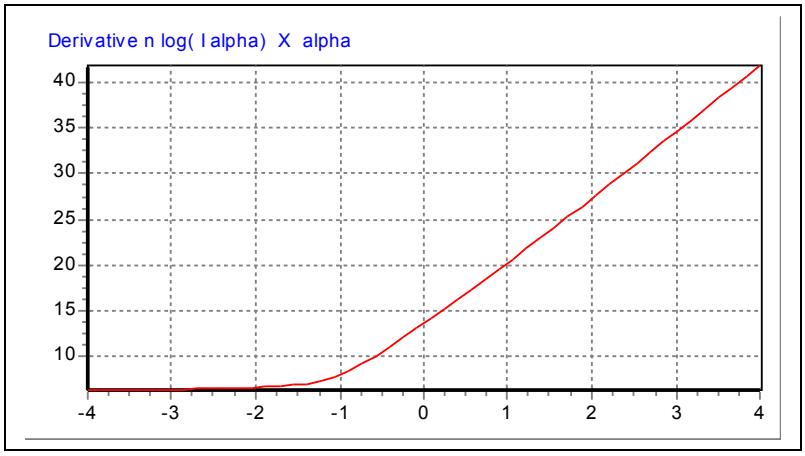


FIGURA 6.2 - Momento Espectral

b) Cálculo da derivada primeira do $\log I_\alpha$, ou seja,

$$\frac{d}{d\alpha}\log I_\alpha = \frac{I'_\alpha}{I_\alpha} \tag{6.15}$$

Neste caso o valor de α_C pode ser determinado como o ponto de inflexão da curva. A Fig. 6.3 mostra o comportamento típico da derivada primeira do momento espectral de uma série temporal auto-afim. Nesta figura o eixo das abcissas é a variável alfa e o das ordenadas é a derivada primeira do momento. Para esta figura $\alpha_C \approx 1$.

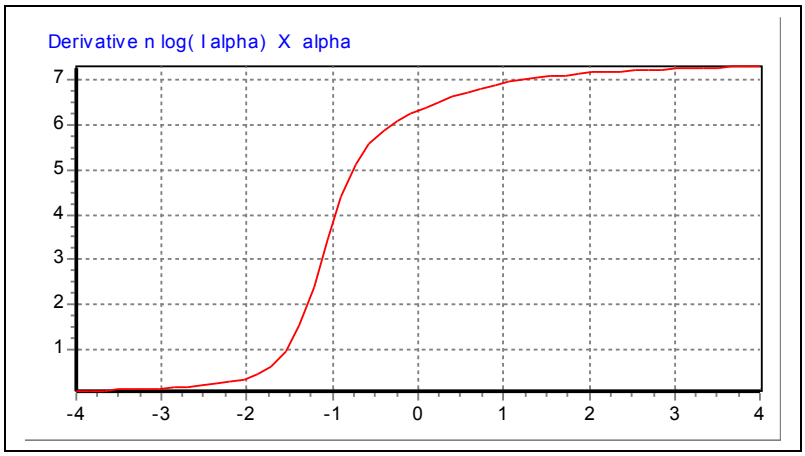


FIGURA 6.3 - Derivada Primeira do Momento Espectral

c) Cálculo da derivada segunda do $\log I_\alpha$, ou seja,

$$\frac{d^2 \log I_\alpha}{d\alpha^2} = \frac{I''_\alpha I_\alpha - \left(I'_\alpha\right)^2}{{I_\alpha}^2} \tag{6.16}$$

Neste caso o valor de α_C pode ser determinado como o ponto de máximo desta curva. A Fig. 6.4 mostra o comportamento típico da derivada segunda do momento espectral de uma série temporal auto-afim. Nesta figura o eixo das abcissas é o alfa e o das ordenadas é a derivada segunda do momento.

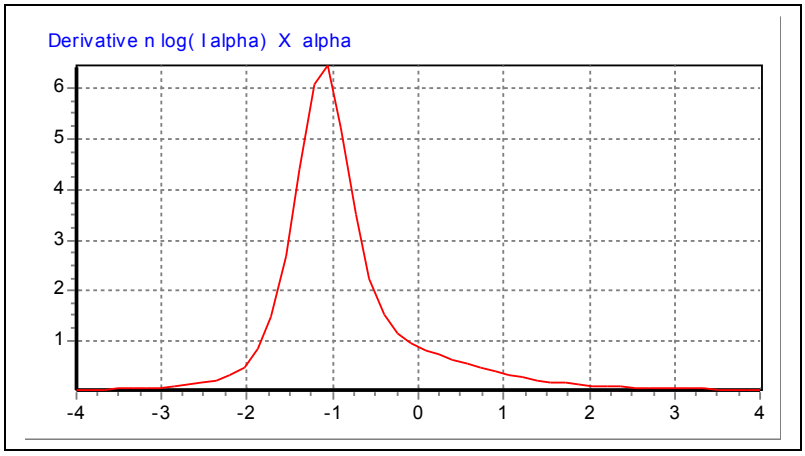


FIGURA 6.4 - Derivada Segunda do Momento Espectral

6.4 Dimensões Generalizadas ou Multifractais

Define-se dimensão generalizada como:

$$D_q = \frac{1}{q-1} \lim_{\varepsilon \to 0} \left(\frac{\ln \sum_{i=1}^{N(\varepsilon)} {p_i}^q}{\ln \varepsilon} \right) \tag{6.17}$$

onde

$q \in \Re$, $q \neq 1$

$$p_i = p_i(\varepsilon) = \lim_{N \to +\infty} \frac{N_i}{N} \tag{6.18}$$

N_i é o número de pontos na caixa i
N é o número total de pontos

A Tab. 6.2 mostra algumas das dimensões mais utilizadas.

Tabela 6.2 - Dimensões Fractais mais Utilizadas

Dimensão	**Algoritmo**	**Equação**
Dimensão Fractal ou Capacidade	$D_0 = \lim_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{\ln\left(1/\varepsilon\right)}$	(6.19)
Dimensão de Informação	$D_1 = \lim_{\varepsilon \to 0} \left(\frac{\sum_{i=1}^{N(\varepsilon)} p_i \ln p_i}{\ln \varepsilon} \right)$	(6.20)
Dimensão de Correlação	$D_2 = \lim_{\varepsilon \to 0} \left(\frac{\ln \sum_{i=1}^{N(\varepsilon)} p_i^{\,2}}{\ln \varepsilon} \right)$	(6.21)

6.5 Algoritmo de Grassberger-Procaccia para Dimensões Generalizadas

Da mesma forma como foi feito para as Dimensões Generalizadas, pode-se obter uma equação equivalente para a estimativa da integral de correlação utilizada na proposta de Grassberger-Procaccia. Assim, o espectro de dimensões da série temporal pode ser obtido através de:

$$C_q(e) = \lim_{N \to +\infty} \left[\frac{1}{N} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \left\{ \frac{1}{N} \sum_{j=1}^{N} H\left(e - \left|\vec{x}_i - \vec{x}_j\right|\right) \right\}^{q-1} \right]^{\frac{1}{q-1}} \tag{6.22}$$

Desta forma tem-se:

$$D_q \sim \lim_{\varepsilon \to 0} \frac{\ln C_q(\varepsilon)}{\ln \varepsilon} \tag{6.23}$$

6.6 Conclusão

Este capítulo apresentou os métodos mais utilizados na estimativa da dimensão fractal. Procedeu-se também com a generalização dos procedimentos de forma a contemplar e desta forma poder verificar a homogeneidade do possível atrator associado.

7 Espectro de Expoentes de Lyapunov

7.1 Introdução

Normalmente, na prática, não são conhecidas as matrizes Jacobianas ou derivadas associadas à dinâmica do sistema físico donde a série temporal foi obtida. Isso dificulta o cálculo do espectro de expoentes de Lyapunov. Assim faz-se a reconstrução de Takens que dá acesso ao atrator e à sua medida invariante. Os diversos métodos propostos para a estimativa de expoentes de Lyapunov diferem exatamente na maneira de contornar este problema.

Uma vez reconstruído o atrator, define-se uma trajetória fiducial ou de referência a partir da seqüência de vetores reconstruídos. Analisa-se o que ocorre com pontos na vizinhança dessa trajetória, buscando-se informações a respeito da taxa de divergência dos pontos próximos e, portanto, dos expoentes de Lyapunov.

Existem vários métodos para o cálculo do espectro de expoentes de Lyapunov [BRI 90, BRO 91]. Eles diferem quanto ao número de expoentes calculados, e, principalmente, quanto ao modo de se aproximar a dinâmica em torno da trajetória fiducial. Em quase todos eles, a eficiência na obtenção dos expoentes de Lyapunov associados a uma série temporal depende da quantidade de pontos disponíveis e da sua qualidade. Em particular, os expoentes de Lyapunov negativos são de difícil estimativa pois associam-se a direções onde há contração, e nessas o atrator reconstruído não contém informação com resolução suficiente para uma estimativa adequada e confiável.

A seção 7.2 apresenta o método de Wolf. Este constitui um dos métodos mais utilizados devido a existência de um código robusto e confiável. A seção 7.3 apresenta o método de Eckmann e Ruelle. A seção 7.4 apresenta a conjectura de Kaplan-York que pode ser utilizada para a estimativa da dimensão fractal a partir do espectro de Lyapunov.

7.2 Método de Wolf

O método proposto por Wolf permite a estimativa dos expoentes de Lyapunov não negativos de uma série experimental. Num primeiro momento calcula-se o maior expoente de Lyapunov positivo λ_1 e depois, o segundo maior expoente λ_2 (se positivo), e assim sucessivamente. A separação entre dois pontos próximos define um eixo principal e a reortonormalização é substituída pela procura de um novo ponto, próximo à trajetória fiducial, que preserve ao máximo a orientação desse eixo.

A implementação do algoritmo de Wolf não apresenta maiores dificuldades. Nos apêndices do artigo original de Wolf são fornecidos códigos FORTRAN para tal fim. Diversos testes mostram que o método é robusto com relação a escolha dos parâmetros envolvidos.

7.3 Método de Eckmann e Ruelle

O método sugerido por Eckmann e Ruelle [ECK 85] permite determinar todo o espectro de expoentes de Lyapunov. Nele estimam-se as equações variacionais, isto é, as equações linearizadas do comportamento dinâmico do sistema, e portanto são obtidas as matrizes Jacobianas ao longo da trajetória fiducial. A partir daí obtém-se, pela definição, o espectro de Lyapunov.

7.4 Conjetura de Kaplan-York

Através desta conjetura é possível estimar-se a dimensão fractal através do espectro de Lyapunov. Esta dimensão é conhecida como Dimensão de Kaplan-York ou Dimensão de Lyapunov. Assim:

$$D_{xy} = j + \frac{\sum_{i=1}^{j} \lambda_i}{\left|\lambda_{j+1}\right|} \tag{7.1}$$

onde $\lambda_1 > \lambda_2 > \cdots > \lambda_d$ são os expoentes de Lyapunov ordenados de forma decrescente e j é o maior inteiro tal que $\sum_{i=1}^{j} \lambda_i > 0$.

A Tab. 7.1 mostra a dimensão obtida através desta conjetura. Através da observação da Tab. 7.1 pode-se constatar os bons resultados obtidos.

TABELA 7.1 - Comparação entre a Dimensão de Hausdorff e Kaplan-York

Sistema	**D_0**	**D_{ky}**
Mapa de Hénon (a=1,4; b=0,3)	1,261 ± 0,003	1,264 ± 0,002
Mapa de Hénon (a=1,2; b=0,3)	1,202 ± 0,003	1,200 ± 0,003
Mapa de Kaplan-York (α=0,2)	1,4316 ± 0,0016	1,4306766 (analítico)
Mapa de Zaslavskii (Γ=3,0; ε=0,3; ν=4/3x102)	1,380 ± 0,007	1,387 ± 0,001

7.5 Conclusão

Neste capítulo foi discutido o que é expoente de Lyapunov e sua importância na caracterização do nível de caoticidade na séries temporais. Apresentou-se o método de Wolf e o método de Eckmann e Ruelle muito utilizados na estimativa desses expoentes. E, finalmente, apresentou-se a conjectura de Kaplan-York que pode ser utilizada para a estimativa da dimensão fractal a partir do espectro de Lyapunov.

8 Entropia de Kolmogorov

A entropia de Kolmogorov é outra importante característica que descreve o grau de caoticidade de uma série temporal. A entropia fornece a taxa média de informação perdida sobre a posição do ponto de fase no atrator. É sabido que:

- Se $K = 0$ tem-se uma série periódica ou quase-periódica;
- Se K é infinito tem-se uma série aleatória;
- Se $0 < K < \infty$ tem-se uma série caótica determinística .

A entropia generalizada pode ser obtida da mesma forma que a dimensão generalizada [RAI 96], ou seja:

$$K_q = -\lim_{\varepsilon \to 0} \lim_{\Delta t \to 0} \lim_{N \to +\infty} \frac{1}{N \Delta t} \frac{1}{q-1} \ln \sum_{i=1}^{N} p_i^q \tag{8.1}$$

A abordagem mais adequada para se estimar a entropia de Kolmogorov-Sinai foi proposta por Grassberger e Procaccia, ou seja:

$$K_2 \approx \lim_{m \to \infty} \lim_{\varepsilon \to 0} K_2^m(\varepsilon), \tag{8.2}$$

onde

$$K_2^m(\varepsilon) = \frac{1}{k \Delta t} \ln \frac{C^m(\varepsilon)}{C^{m+k}(\varepsilon)} \tag{8.3}$$

onde k é um número inteiro suficientemente pequeno. Na prática não se pode satisfazer os limites de integração das fórmulas acima. Busca-se então a saturação do valor de K_2^m para m incrementando. Esta será uma boa aproximação para K_2.

Outra forma de se obter K é somar os expoentes de Lyapunov positivos, isto é:

$$K = \sum_{\substack{i=1 \\ \lambda_i > 0}}^{n} \lambda_i \tag{8.4}$$

A Tab. 8.1, tirada de [FER 94] resume as características de séries temporais regulares (periódicas e quase-periódicas), caóticas e estocásticas.

TABELA 8.1 – Resumo das características de séries temporais regulares, caóticas e estocásticas

Tipo de Série	**Expoente de Lyapunov**	**Entropia de Kolmogorov-Sinai**	**Dimensão da dinâmica assintótica** $(t \to \infty)$
Regular	Não há expoentes positivos	$K = 0$	$D < m$, $D \in N$
Caótica	Existe pelo menos um expoente positivo	$0 < K \leq L$, onde L é a soma de todos os expoentes positivos de Lyapunov	$D < m$, $D \in \Re$ Obs: em sistemas contínuos: $m \geq 3$
Estocástico	-	$K \to \infty$	$D = m$

9 Aplicação a Sinais de Voz

9.1 Introdução

Como estudo de caso, procurou-se aplicar as técnicas descritas neste trabalho para analisar-se as propriedades fractais de algumas séries temporais. Utilizou-se arquivos de voz do corpora SR4X[24] desenvolvido pelo OGI. O objetivo seria, demonstrar que é possível caracterizar estas séries a partir da avaliação do seu comportamento caótico. É sabido que sons de voz tem um espectro de freqüência do tipo $1/f$. Desta forma o método de melhor desempenho para estimar-se a dimensão fractal é o método do expoente crítico. Foram medidas: a dimensão fractal global da série e a dimensão fractal dependente do tempo. A sessão 9.2 estudou-se a dinâmica de duas séries temporais. A primeira foi resultado da amostragem da forma de onda da declaração em inglês da palavra "computer". A segunda série foi obtidas através da amostragem da palavra **"**abracadabra". A sessão 9.3 apresenta a estimativa da dimensão fractal dependente do tempo destas duas séries temporais. E a sessão 9.4 mostra os resultados obtidos para os sons vocálicos da língua japonesa.

9.2 Resultados Obtidos

Escolheu-se, a título de exemplo, para a avaliação da dimensão fractal, o som da palavra inglesa "computer". A Fig. 9.1 mostra a forma de onda no tempo desta declaração. O eixo das abscissas é o tempo e o das ordenadas é a amplitude.

A Fig. 9.2 mostra o espectro de freqüência da declaração "computer". O eixo das abscissas é a freqüência em escala logarítmica. E o eixo das ordenadas é a amplitude em decibeis. Pode-se observar que praticamente toda a energia está concentrada entre 100Hz e 3000 Hz.

A Fig. 9.3 mostra o momento espectral desta declaração. O eixo das abscissas é a variável alfa e o eixo das ordenadas é o momento espectral. Pode-se observar que existem duas regiões lineares no gráfico. Se traçarmos duas retas nestas duas regiões, teremos que a interseção das duas retas será no ponto alfa crítico igual a -0,8.

[24] Ver capítulo 2, sessão 5 para uma descrição desta base

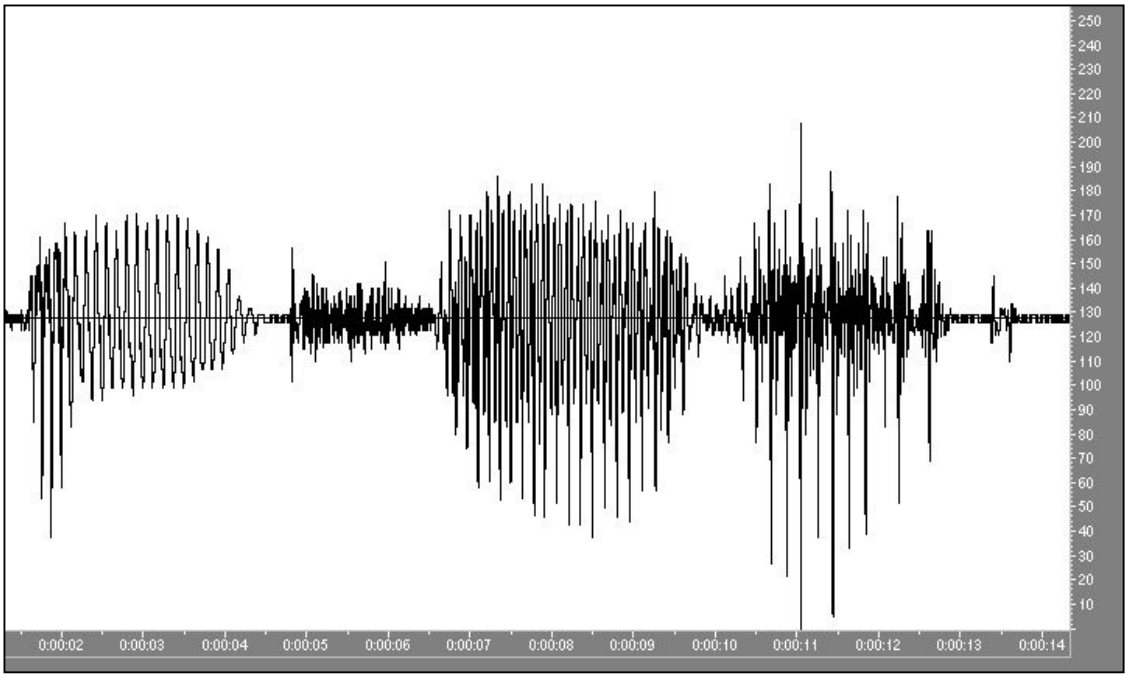


FIGURA 9.1 – Forma de onda da palavra “computer”

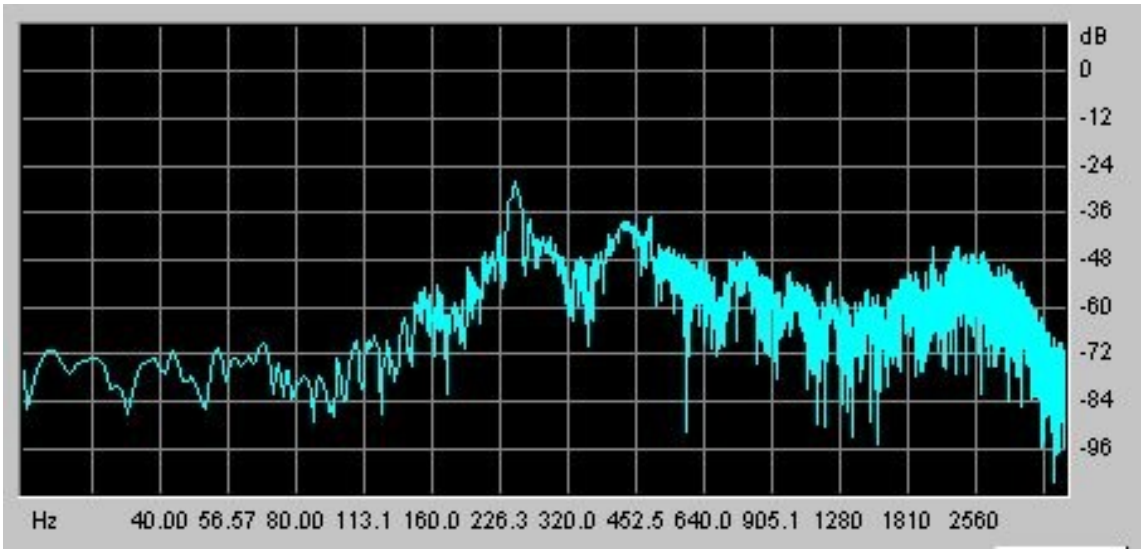


FIGURA 9.2 – Espectro de freqüência da declaração “computer”

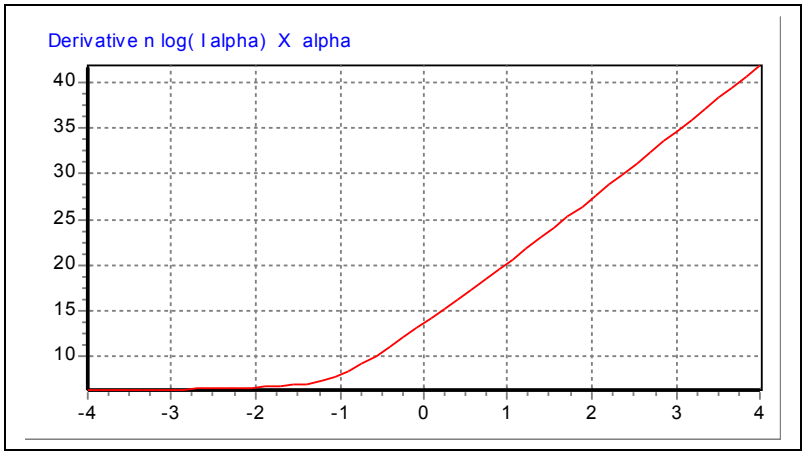


FIGURA 9.3 – Momento espectral da declaração “computer”

A Fig. 9.4 mostra a derivada primeira do momento espectral desta declaração. O eixo das abscissas é a variável alfa e o eixo das ordenadas é a derivada primeira do momento espectral. Pode-se observar que o ponto de inflexão da curva será no ponto alfa crítico igual a –1,0.

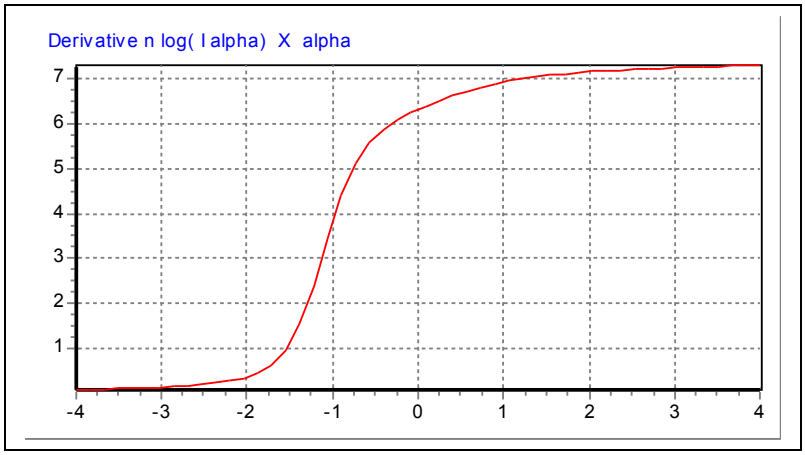


FIGURA 9.4 – Derivada primeira do momento espectral da declaração “computer”

A Fig. 9.5 mostra a derivada segunda do momento espectral desta declaração. O eixo das abscissas é a variável alfa e o eixo das ordenadas é a derivada segunda do momento espectral. Pode-se observar que o ponto de máximo da curva será no ponto alfa crítico igual a –1,0.

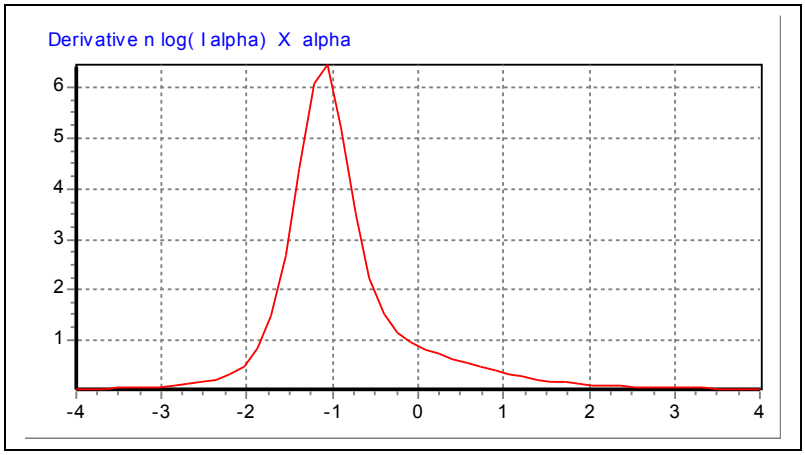


FIGURA 9.5 – Derivada primeira do momento espectral da declaração “computer”

A Fig. 9.6 mostra a derivada terceira do momento espectral desta declaração. O eixo das abscissas é a variável alfa e o eixo das ordenadas é a derivada terceira do momento espectral. Pode-se observar que o ponto onde a curva corta o eixo das abscissas será aproximadamente para alfa crítico igual a –1,0.

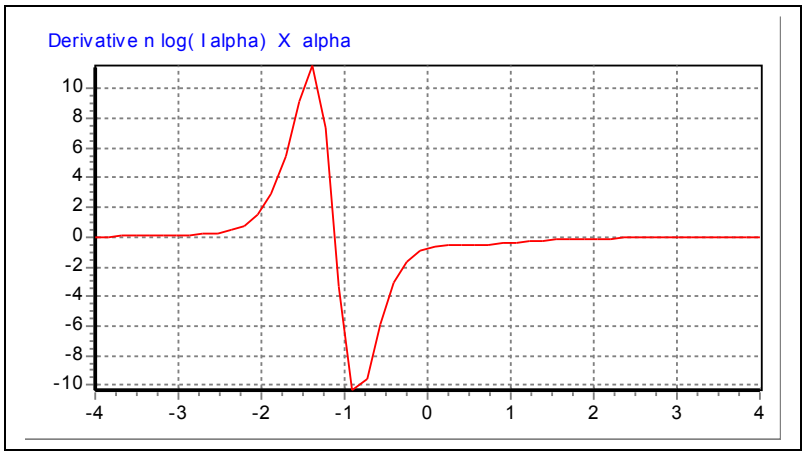


FIGURA 9.6 – Derivada primeira do momento espectral da declaração “computer”

Verificou-se que os resultados obtidos são muito próximos aos apresentados na literatura [NAK 93].

Tanto o momento, quando todas as derivadas apresentadas podem ser utilizadas para se estimar o alfa crítico e por conseguinte a dimensão fractal. Todavia, diferente do proposto em [NAK 93], a obtenção a partir do momento é muito mais eficiente computacionalmente do que pelas derivadas.

9.3 Dimensão Fractal Dependente do Tempo

Para ilustrar uma aplicação de caracterização dinâmica de uma série temporal, aplicou-se o método acima à mesma série, porém, utilizou-se a mesma abordagem daquela adotada na transformada de Gabor, ou seja, o sinal foi janelado e em cada janela estimou-se a dimensão fractal. Foi feita uma sobreposição de 12,5% entre as janelas. Utilizou-se a janela de Hanning na estimativa do espectro de potência. Cada janela teve 1024 amostras. O valor médio das dimensões fractais obtidas está apresentado na Tab. 9.1:

TABELA 9.1 – Dimensão Fractal das palavras “computer” e “abracadabra”

	Palavra	"Computer"	Palavra	"Abracadabra"
Locutor	**D**	**Desvio Padrão**	**D**	**Desvio Padrão**
1030	2,6319	0,1111	2,5869	0,0819
1063	2,6852	0,1236	2,6939	0,0848
1111	2,6304	0,0673	2,5784	0,1183
1159	2,6066	0,1073	2,6854	0,0647
1227	2,6920	0,0623	2,6607	0,0550
1234	2,6991	0,0535	2,6709	0,0483
1305	2,5977	0,1251	2,5774	0,1129
1309	2,7169	0,0991	2,7000	0,0637
1348	2,7294	0,0666	2,6853	0,0508
1381	2,7708	0,0803	2,6673	0,0793

O valor médio da dimensão fractal obtida foi de aproximadamente 2,66 com um desvio padrão médio de 0,08. Isso dá um bom indicativo da presença de caos em sinais de voz. O valor baixo da dimensão média evidencia que através de um pequeno número

de equações diferencias ordinárias pode-se representar de maneira adequada o comportamento o sistema fonador humano.

O gráfico da Fig. 9.7 mostra a variação temporal da dimensão fractal no tempo da declaração “computer” obtida pelo método CEM.

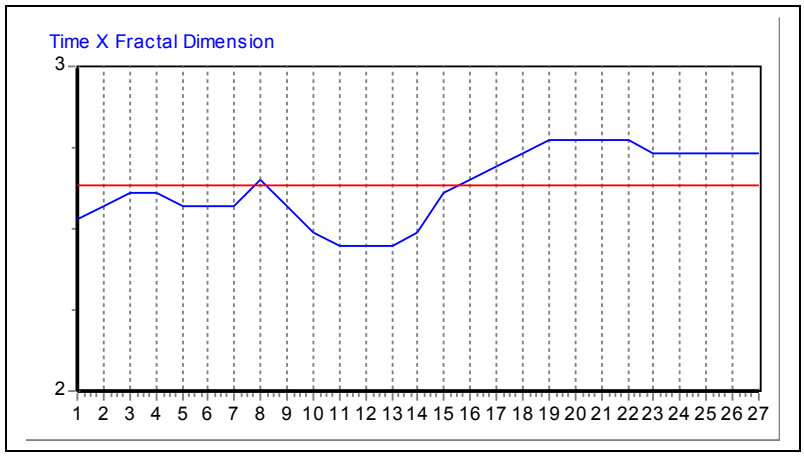


FIGURA 9.7 – Dimensão Fractal Dependente do Tempo

Pode-se observar nesta figura uma pequena variação com o tempo da dimensão fractal. Alguns trabalhos recentes tem proposto que estas variações estão associadas ao grau de liberdade dos sons produzidos pelo aparelho fonador humano. Quanto menor a dimensão fractal menor será o grau de liberdade do sistema fonador, ou seja, menor a quantidade de variáveis necessárias para representar o aparelho fonador naquele instante de tempo.

9.4 Avaliando sons vocálicos

A Tab. 9.2 mostra: os cinco primeiros expoentes de Lyapunov para as vogais da língua japonesa; a dimensão de Lyapunov; a dimensão e o alfa crítico obtidos pelo método do expoente crítico. Estes valores forem retirados do trabalho de [NAK 93]. Todos os maiores expoentes de Lyapunov das cinco vogais analisadas são positivos indicando a presença de caos determinístico. Observa-se também que o valor da dimensão fractal obtida pelo CEM coincide aproximadamente com os valores obtidos para os sons da língua inglesa. Além disso, os valores da dimensão obtida através dos expoentes de Lyapunov confirmam a baixa dimensionalidade do possível atrator associado ao sistema fonador. O autor deste trabalho propõe que se utilize a dimensão como um parâmetro adicional na caracterização dos padrões sonoros. Assim, além dos parâmetros tradicionais baseados no espectro de potência do sinal, poder-se-ia utilizar o a DFDT no processo de reconhecimento automático de voz.

TABELA 9.2 – Espectro de Lyapunov e Dimensão Fractal de sons da língua japonesa

Vogal	$\overline{\lambda_1}$	$\overline{\lambda_2}$	$\overline{\lambda_3}$	$\overline{\lambda_4}$	$\overline{\lambda_5}$	D_L	α_c	D
'a'	0,3	0,7	-0,1	-0,4	-1,2	3,99	-0,9	2,45
'i'	0,6	0,2	-0,2	-0,8	-2,2	3,64	-0,68	2,34
'u'	0,5	0,0	-0,4	-1,0	-2,3	3,16	1,65	1,17
'e'	0,6	0,0	-0,2	-0,6	-1,7	3,65	1,07	2,53
'o'	1,0	0,2	-0,2	-1,2	-3,1	3,84	0,69	1,65

9.5 Conclusão

Neste capítulo aplicou-se as técnicas descritas neste trabalho para analisar-se as propriedades fractais de algumas séries temporais. Utilizou-se arquivos de voz do corpora SR4X[25] desenvolvido pelo OGI. O objetivo foi demonstrar que é possível caracterizar estas séries a partir da avaliação do seu comportamento caótico. A dimensão fractal foi estimada através do método do expoente crítico uma vez que os sons de voz apresentam a propriedade auto-afim. Foram medidas: a dimensão fractal global da série e a dimensão fractal dependente do tempo. O valor médio da dimensão fractal obtido foi de 2,67 e desvio padrão de 0,08. Também foi apresentado os resultados obtidos por [NAK 93] para os sons vocálicos da língua japonesa. Os valores obtidos foi muito próximos dos obtidos para as declarações em inglês acima citadas.

[25] Ver capítulo 2, sessão 5 para uma descrição desta base

10 Conclusões

Este trabalho está dividido em duas partes: tema de abrangência e tema em profundidade. Na primeira parte realizou-se um estudo sobre a tecnologia de reconhecimento automático de voz, com particular ênfase no quesito das variabilidades associadas ao sinal de voz. Na segunda parte estudou-se como estimar algumas das propriedades dinâmicas envolvidas no processo de produção de voz pelo aparelho fonador humano. Em particular, deu-se ênfase ao estudo de métodos para a determinação de algumas propriedades invariantes de forma a poder-se caracterizar os padrões sonoros da voz humana. As propriedades invariantes estudadas foram: a dimensão fractal; a entropia de Kolmogorov Sinai; e o espectro de Lyapunov.

No capítulo 2 tem-se um breve histórico dos acontecimentos mais importantes que marcaram a evolução da tecnologia de RAV. Foi discutido em detalhes o que é e como pode ser caracterizado um sistema de RAV e seus principais componentes, com a apresentação das tecnologias envolvidas. Além disso discutiu-se a necessidade de bases de dados de voz padronizadas, denominadas de Corporas, para o treinamento e avaliação dos sistemas e algoritmos de RAV. Também foram listados os principais Corporas existentes. Infelizmente pouco se tem feito em termos do desenvolvimento de um corpora para o português falado no Brasil. Ainda neste capítulo discutiu-se como podem ser modeladas as diferentes fontes de variabilidade presentes no sinal de voz e o estado da arte em termos de tecnologias e sistemas existentes. Finalmente apresentou-se as expectativas futuras nesta área.

O capítulo 3 dedicou-se ao estudo da robustez no reconhecimento de voz a qual refere-se a necessidade de manter bom reconhecimento mesmo quando a qualidade da voz de entrada é degradada, ou quando as características acústicas, articulatórias ou fonéticas da voz de treinamento e ambientes de testes diferem. Foram apresentadas algumas técnicas para a adaptação dinâmica dos parâmetros internos que podem melhorar a robustez dos sistemas de RAV em relação ao ambiente e diferentes locutores. Discutiu-se também o uso simultâneo de vários microfones com o objetivo de melhorar a direcionalidade e a relação sinal ruído e a importância do uso de técnicas de processamento digital de sinais, fisiologicamente motivadas. Finalmente apresentou-se os desafios futuros para a melhoria da robustez nos sistemas de RAV.

O capítulo 4 apresentou os modelos escondidos de Markov que consiste da técnica mais flexível e de maior êxito usada em RAV. Foram apresentados os conceitos básicos e os principais algoritmos utilizados no treinamento e teste destes modelos. Finalmente, discutiu-se como aplicar o modelo escondido de Markov ao espaço acústico e algumas de suas deficiências.

O capítulo 5 preocupou-se com a reconstrução da dinâmica de séries temporais, e em particular das séries oriundas da digitalização de sinais de voz. Justificou-se porque deve-se estudar os sinais de voz sobre esta nova ótica e como isso poderia ajudar no seu reconhecimento. As séries temporais foram classificadas em três tipos: as regulares, estudas através da análise de Fourier; as estocásticas, estudadas através de ferramentas estatísticas; e as caóticas. Para estudar estas últimas apresentou-se algumas metodologias e técnicas, entre elas a dimensão fractal, a entropia de Kolmogorov-Sinai e o espectro de expoentes de Lyapunov. Os capítulos 6, 7 e 8 apresentam estas técnicas em detalhes.

O capítulo 9 apresenta alguns resultados da aplicação das técnicas de reconstrução da dinâmica a sinais de voz. Utilizou-se arquivos de voz do corpora SR4X desenvolvido pelo OGI. O objetivo foi demonstrar que é possível caracterizar estas

séries a partir da avaliação do seu comportamento caótico. A dimensão fractal foi estimada através do método do expoente crítico uma vez que os sons de voz apresentam a propriedade auto-afim. Foram medidas: a dimensão fractal global da série e a dimensão fractal dependente do tempo. O valor médio da dimensão fractal obtido foi de 2,67 e desvio padrão de 0,08. Os resultados obtidos foram muito próximos daqueles obtidos por [NAK 93] para os sons vocálicos da língua japonesa.

Bibliografia

[ABA 96] ABARBANEL, Henry D. I**. Analysis of Observed Chaotic Data**. San Diego: Springer, 1996. 272p.

[BEM 96] BENGIO, Yoshua. **Neural Networks for Speech and Sequence Recognition**. UK: Ed. International Thonson Computer Press, 1996.

[BRI 90] BRIGGS, Keith. An Improved Method for Estimating Lyapunov Exponents of Chaotic Time Series. **Physics Letters A**, [S.l.], v. 151, n. 1-2, p. 27-32, 1990.

[BRO 86] BROOMHEAD, D. S.; KING, G. P. Extracting Qualitative Dynamics from Experimental Data. **Physica D**, Amsterdam, v. 20, p. 217-236, 1986.

[BRO 91] BROWN, Reggie et al. Computing the Lyapunov Spectrum of a Dynamical System from an Observed Time Series. **Physical Review A**, [S.l.], v. 43, n. 6, p. 2787-2806, 1991.

[COL 95] COLE, Ronald A. et al. **Survey of the State of the Art in Human Language Technology**. USA: Oregon Graduated Institute, 1995. 590p.

[CUS 90] CUSTÓDIO, Ricardo F. **Codificação Paramétrica de Sinais de Voz com Excitação Multi-Pulso**. Florianópolis: UFSC, 1990. Dissertação de Mestrado.

[CUS 97] CUSTÓDIO, Ricardo F. **Caracterização de uma Série Temporal Através da Dimensão Fractal**. Porto Alegre: CPGCC da UFRGS, 1997. (TI-703).

[CUS 97] CUSTÓDIO, Ricardo F.; BARONE, Dante A. C. Atomic Decomposition with Genetic Algorithm. In: Nolta, 1997, Hawaii. **Proceedings…**[S.l.:s.n.], 1997.

[DAU 92] DAUBECHIES, I. **Ten Lectures on Wavelets**. Philadelphia, PA: SIAM, 1992.

[DAV 97] DAVIES, M. E. Reconstructing Attractors from Filtered Time Series. **Physica D**, Amsterdam, v. 101, p. 195-206, 1997.

[DRA 94] DRAZIN, P. G. **Nonlinear Systems.** Melborne: Cambridge, 1994. 317p.

[DUH 90] DUHAMED, P.; VETTERLI, M. Fast Fourier Transform: A Tutorial Review and A State of the Art. **Signal Processing**, [S.l.], v. 19, p. 259-299, Apr. 1990.

[ECK 85] ECKMANN, J. P.; RUELLE, D. Ergodic Theory of Chaos and Strange Attractors. **Rev. Mod. Phys**., [S.l.], v. 57, p. 617, 1985.

[ECK 86] ECKMANN, J. P. et al., **Phys. Rev**. **A**, [S.l.], v. 34, p. 4971, 1986.

[ECK 92] ECKMANN, J. P. e RUELLE, D. Fundamental Limitations for Estimating Dimensions and Lyapunov Exponents in Dynamical Systems. **Physica D**, Amsterdam, v. 56, p. 185-187, 1992.

[EDM 96] EDMONDS, Andrew Nicola. **Time Series Prediction Using Supervised Learning and Tools from Chaos Theory**. UK: University of Luton, 1996. Ph.D. Thesis.

[ELT 87] ELTON, John. An Ergodic Theorem for Iterated Maps, **Ergodic Theory and Dynamical Systems,** [S.l.], v. 7, p. 481-488, 1987.

[FED 88] FEDER, Jens. **Fractals**. New York: Plenum, 1988. 283p.

[FER 94] FERRARA, Nelson F.; PRADO, Carmem P. C. **Caos:** Uma Introdução. São Paulo: E. Blücher, 1994. 402p.

[FRA 86] FRASER, A. M.; SWINNEY, H. L. Independent Coordinates for Strange Attractors. **Phys. Rev**., [S.l.], v. 33A, p. 1134-1140, 1986.

[FRI 96] FRITSCH, Jürgen. **Modular Neural Networks for Speech Recognition**. USA: Carnegie Mellow University, 1996. Ph.D. Thesis.

[GAB 46] GABOR, D. Theory of Communication. **Journal of the IEEE**, New York, p. 429-457, 1946.

[GLE 87] GLEICK, J. **Chaos:** Making a New Science, New York: P. Books, 1987.

[GOM 97] GOMES, Jonas et al. **Wavelets:** Teoria, Software e Aplicações. Rio de Janeiro: IMPA/CNPq, 1997. 216p. Trabalho apresentado no Colóquio Brasileiro de Matemática, 21., 1997, Rio de Janeiro.

[GRA 83] GRASSBERGER, P.; PROCACCIA, I. Estimation of Kolmogorov Entropy from a Chaotic Signal. **Phys. Rev. A**, [S.l.], v. 28, p. 2591, 1983.

[HUT 81] HUTCHINSON, John. Fractals and Self-Similarity, **Indiana University Mathematics Journal**, [S.l.], v. 30, n. 5, p. 713-747, 1981.

[JAC 93] JACQUIN, Arnaud. Fractal Image Coding: A Review. **Proceedings of the IEEE**, New York, v. 81, n. 10, p. 1451-1465, 1993.

[JOH 97] JOHNSON, Keit; MULLENNIX, John W. **Talker Variability in Speech Recognition**. San Diego: Academic Press, 1997.

[KAP 92] KAPLAN, D. T., GLASS, L. **Physical Reviews Letters**, [S.l.], v. 68, p. 427, 1992.

[KEN 92] KENNEL M. et al., **Phys. Rev**. **A**, [S.l.], v. 45, p. 3403, l992.

[KLA 77] KLATT, Dennis. Review of the ARPA Understanding Project. **Journal of the Acoustic Society of America**, [S.l.], v. 62, p. 1345-66, 1977.

[KOS 92] KOSKO, Bart. **Neural Networks for Signal Processing**. New Jersey: Prentice Hall, 1992.

[LEV 94] LEVY-VEHEL, Jacques et al. Fractal Modeling of Speech Signals, Fractals. **An Interdisciplinary Journal On The Complex Geometry of Nature**, [S.l.], v. 2, n. 3, p. 379-386, 1994.

[LIP 89] LIPPMANN, Richard P. **Review of Neural Networks for Speech Recognition**. USA: MIT, p. 374-92, 1989.

[LU 93] LU, Jian. **Signal Recovery and Noise Reduction with Wavelets**. Hanover: Dartmouth College, 1993. Ph.D. Thesis.

[LUN 92] LUNDHEIM, Lars. **Fractal Signal Modeling for Source Coding**, Norway: The Norwegian Institute of Technology, 1992. Ph.D. Thesis.

[MAN 82] MANDELBROT, B. B. **The Fractal Geometry of Nature**. San Francisco: Freeman, 1982.

[MAR 96] MARKOWITZ, Judith A. **Using Speech Recognition**. New Jersey: Prentice Hall, 1996.

[MAZ 92] MAZEL, David. **Fractal Modeling of Time Series Data**. Atlanta, USA: Georgia Institute of Technology, 1992. Ph.D. Thesis.

[MOE 97] MOECKEL, Richard; MURRAY, Brad. Measuring the Distance Between Time Series. **Physica D**, Amsterdam, v. 102, p. 187-194, 1997.

[MOO 92] MOON, F. C. **Chaotic and Fractal Dynamics**. New York: John Wiley & Sons, 1992

[NAK 93] NAKAGAWA, Masahiro. A Critical Exponent Method to Evaluate Fractal Dimension of Self-Affine Data. **Journal of the Physical Society of Japan**, [S.l.], v. 62, n. 12, p. 4233-4239, Dec. 1993.

[OSB 90] OSBORNE, A. R.; PROVENZALE, A. Finite Correlation Dimension for Stochastic Systems with Power-Law Spectra. **Physica D**, Amsterdam, v. 35, p. 357-381, 1989.

[PAC 80] PACKARD, N. H. et al. Geometry from a Time Series. **Physical Reviews Letters**. [S.l.], v. 45, p. 712-16, 1980.

[PEN 84] PENTLAND, Alex. Fractal Based Description of Natural Scenes, **IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),** New York, v. 6, n. 6, p. 661-674, 1984.

[PIC 86] PICKOVER, C. A.; KHOROSANI, A**.** Fractal Characterization of Speech Waveform Graphs. **Computer Graphics**, p. 10-51, 1986.

[POT 97] POTAPOV, Alexei. Distortions of Reconstruction for Chaotic Attractors. **Physica D**, Amsterdam, v. 101, p. 207-226, 1997.

[PRE 94] PRESS, Willian H. et al. **Numerical Recipes in C:** The Art of Scientific Computing. Melborne: Cambridge University, 1994. 994p.

[RAB 78] RABINER, Lawrence R.; SHAFER, Ronald W. **Digital Processing of Speech Signals**. Englewood Cliffs, NJ: Prentice Hall, 1978.

[RAB 93] RABINER, Laurence; JUANG, Biing H. **Fundamentals of Speech Recognition**. Englewood Cliffs, NJ: Prentice Hall, 1993.

[RAI 96] RAIDL, Ales. Estimating the Fractal Dimension, K2-entropy and Predictability of the Atmosphere, **Czechoslovak Journal of Physics**, [S.l.] v. 46, p. 293-328, 1996.

[ROS 93] ROSENTEIN, Michel T. et al. A Practical Method for Calculating Largest Lyapunov Exponents from Small Data Sets. **Physica D**, Amsterdam, v. 65, p. 117-134, 1993.

[SAB 95] SABANAL, Salvador; NAGAGAWA, Masahiro. A Study of Time-Dependent Fractal Dimensions of Vocal Sounds. **Journal of the Physics Society of Japan**, [S.l.], v. 64, n. 9, p. 3226-3238, 1995.

[SAB 96] SABANAL, Salvador; NAGAGAWA, Masahiro. The Fractal Properties of Vocal Sounds and Their Application in the Speech Recognition Model. **Chaos, Solution & Fractals**, [S.l.], v. 7, n. 11, p. 1825-1843, 1996.

[SAT 87] SATO, Shinichi et al. Practical Methods of Measuring the Generalized Dimension and the Largest Lyapunov Exponent in High Dimensional Chaotic Systems. **Prog. Ther. Phys**. [S.l.], v. 77, n. 1, 1987.

[SCH 97] SCHULTZ, Tanja; WESTPHAL, M**.** The GlobalPhone Project: Multilingual LVCSR[26] with Janus-3. **Proc. SQEL**, p. 20-27, 1997.

[SCH 98] SCHULTZ, Tanja.; WAIBEL, Alex**.** Multilingual and Crosslingual Speech Recognition. **DARPA Broadcast News Workshop**, 1998.

[SMI 84] SMITH, Avy R. Plants, Fractals, and Formal Languages, **SIGGRAPH `84 Computer Graphics Conference Proceedings**, v. 18, n. 3, p.

[26] Large Vocabulary Continuous Speech Recognition

1-10, 1984.

[STO 91] STOOD, R.; PARISI, J. Calculation of Lyapunov Exponents Avoiding Spurious Elements. **Physica D**, Amsterdam**,** v. 50, p. 89-94, 1991.

[STR 94] STROGATZ, S. H. **Nonlinear Dynamics and Chaos With Applications to Physics, Biology, Chemistry, and Engineering**. Massachusetts: Addison Wesley Publishing Company. 1994.

[TAK 81] TAKENS, F. **Detecting Strange Attractors in Turbulence**. Dynamical Systems and Turbulence. New York: Springer Verlag, 1981, 898p. (Lecture Notes in Math).

[THE 92] THEILER, J. et al. Testing for Nonlinearity in Time Series: The Method of Surrogate Data. **Physica D**, Amsterdam, v. 58, p. 77-94, 1992.

[TIM 98] TIMOSZCZUK, Antonio P**. Reconhecimento Automático do Locutor com Redes Neurais Artificiais do Tipo Radial Basis Function (RBF) e Minimal Temporal Information (MTI)**. São Paulo: USP, 1998. Dissertação de Mestrado.

[VIN 93] VINES, Greg. **Signal Modeling With Iterated Function Systems**. USA: Georgia Institute of Technology, 1993. Ph.D. Thesis.

[WOL 85] WOLF, Alan et al. Determining Lyapunov Exponents from a Time Series. **Physica D**, Amsterdam, v. 16, p. 285-317, 1985.

[WOR 96] WORNELL, Gregory W. **Signal Processing with Fractals:** A Wavelet-Based Approach. New Jersey: Prentice Hall, 1996. 177p.

[ZEN 91] ZENG, X. et al. Estimating the Lyapunov-Exponent Spectrum from Short Time Series of Low Precision. **Physical Reviews Letters**, [S.l.], v. 66, n. 25, p. 3229-3232, 1991.