UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

Reconhecimento Automático de Locutor Utilizando Medidas de Invariantes Dinâmicas Não-Lineares

por

ADRIANO PETRY

Tese submetida à avaliação,
como requisito parcial para a obtenção do
grau de Doutor em Ciência da Computação

Prof. Dr. Dante Augusto Couto Barone
Orientador

Porto Alegre, agosto de 2002.

CIP - CATALOGAÇÃO NA PUBLICAÇÃO

Petry, Adriano

Reconhecimento automático de locutor utilizando medidas de invariantes dinâmicas não-lineares / por Adriano Petry. - Porto Alegre : PPGC da UFRGS, 2002.

155 p. : il.

Tese (doutorado) - Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR - RS, 2002. Orientador: Barone, Dante A. C.

1.Reconhecimento de Locutor. 2.Reconhecimento de Voz 3.Processamento Digital de Sinais. 4.Sistemas Dinâmicos. 5.Teoria do Caos 6.Séries Temporais. I.Barone, Dante. II.Título.

Agradecimentos

Agradeço a todas as pessoas que contribuíram, de forma direta ou indireta, para que eu pudesse desenvolver o trabalho que culmina com esta tese de doutorado. Dentre essas pessoas, não posso deixar de nomear algumas que foram especialmente importantes. Primeiramente meus pais, Claudio Enio Petry e Elisabeth Prüfer Petry, que foram de fundamental importância desde o início de minha vida estudantil, através de grande incentivo e apoio incondicional. Além disso, minha estabilidade emocional e autoconfiança é em muito fruto de um ambiente familiar favorável, muito amor e carinho. Agradeço também ao meu saudoso avô materno, Othmar Arnulf Prüfer, pelo incentivo aos estudos, apoio e amizade. Ao meu orientador, professor Dante Barone, agradeço pela oportunidade de ingresso no meio acadêmico e conseqüente realização profissional. Sua amizade e inspiradora motivação me ofereceram um grande desenvolvimento pessoal e uma visão interessante sobre vários aspectos da vida. Agradeço também à minha noiva e futura esposa, Cristiane Righi Franchi, pela compreensão, amor, carinho e atenção dedicados a mim. A tranqüilidade e segurança emocional que nosso relacionamento me proporcionou foram indispensáveis para que eu pudesse avaliar claramente a importância das coisas na minha vida, e "colocar as pedras" antes de tudo. Gostaria também de mencionar meu grande amigo e cunhado, Tibério Silva Caetano, que me auxiliou em vários momentos de dúvidas técnicas, além de revisar praticamente todos meus trabalhos e apontar importantes aspectos a serem melhorados. Agradeço à Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), que me proporcionou auxílio financeiro ao longo do curso de doutorado. Também ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que me apoiou financeiramente durante o período anterior ao início do curso de doutorado. Agradeço por fim a Universidade Federal do Rio Grande do Sul (UFRGS), especialmente à Escola de Engenharia e ao Instituto de Informática, que me deram a oportunidade de acesso ao ensino superior gratuito de qualidade indiscutível.

Sumário

Lista de Abreviaturas

A/D	Analógico/Digital
ANC	Cancelamento Adaptativo de Ruído (*Adaptive Noise Canceling*)
bps	Bits por Segundo (*Bits per second*)
CEM	Método do Expoente Crítico (*Critical Exponent Method*)
CMS	Subtração da Média Cepstral (*Cepstral Mean Subtraction*)
CPU	Unidade Central de Processamento (*Central Processing Unit*)
DET	Detecção de Erro Balanceado (*Detection Error Tradeoff*)
DFT	Transformada de Fourier Discreta (*Discreet Fourier Transform*)
DSP	Processamento Digital de Sinais (*Digital Signal Processing*)
DTW	Alinhamento Temporal Dinâmico (*Dynamic Time Warping*)
ECG	Eletrocardiograma
EEG	Eletroencefalograma
EER	Taxa de Erro Igual (*Equal Error Rate*)
EM	Algoritmo *Expectation-Maximization*
FA	Falsa Aceitação
FFT	Transformada Rápida de Fourier (*Fast Fourier Transform)*
FR	Falsa Rejeição
GMM	Modelos de Mistura de Gaussianas (*Gaussian Mixture Models*)
HMM	Modelos Ocultos de Markov (*Hidden Markov Models)*
IDFT	Transformada de Fourier Discreta Inversa (*Inverse Discreet Fourier Transform*)
ICT	Transformada Inversa do Cosseno (*Inverse Cossine Transform*)
LPC	Coeficientes de Predição Linear (*Linear Prediction Coefficients*)
MIPS	Milhões de Instruções por Segundo
ML	Máxima Verossimilhança (*Maximum likelihood*)
MPSV	Volume Mínimo do Espaço de Fases (*Minimum Phase Space Volume*)
MFCC	Coeficientes *mel*-cepstrais (*Mel-frequency cepstral coefficients*)
PCM	Codificação por Modulação de Pulso (*Pulse Code Modulation*)
PPGC	Programa de Pós-Graduação em Computação
QV	Quantização Vetorial
RAF	Reconhecimento Automático de Fala
RAL	Reconhecimento Automático de Locutor
RASTA	Metodologia Espectral Relativa (*Relative Spectral Metodology*)
RNA	Redes Neurais Artificiais
SNR	Relação Sinal Ruído (*Signal to Noise Ratio*)
SVD	Decomposição por Valor Singular (*Singular Value Decomposition*)
TDFD	Dimensão Fractal Dependente do Tempo (*Time-dependent Fractal Dimension*)
TDMFD	Dimensão Multi-fractal Dependente do Tempo (*Time-dependent Multifractal Dimension*)
UFRGS	Universidade Federal do Rio Grande do Sul
UML	Linguagem de Modelamento Unificada (*Unified Modeling Language*)
UBM	Modelo de Impostores Universal (*Universal Background Model*)

VLSI	Circuitos integrados de larga escala (*Very Large Scale Integrated Circuits*)

Lista de Símbolos

$\rightarrow$	tende a
∞	infinito
<<	muito menor que
>>	muito maior que
$\approx$	aproximadamente igual a
π	3,14159..., valor de pi
e	2,71828..., base dos logaritmos neperianos
log	logaritmo base 10
ln	logaritmo natural
dx	diferencial de x
A^T	Matriz transposta de A
A^{-1}	Matriz inversa de A
det(A)	Determinante da matriz A
sin	seno
cos	cosseno
sinh	seno hiperbólico
$\lvert x \rvert$	módulo de x
Δt	variação de t
$n!$	fatorial de n

Lista de Figuras

Lista de Tabelas

Resumo

As técnicas utilizadas em sistemas de reconhecimento automático de locutor (RAL) objetivam identificar uma pessoa através de sua voz, utilizando recursos computacionais. Isso é feito a partir de um modelamento para o processo de produção da voz. A modelagem detalhada desse processo deve levar em consideração a variação temporal da forma do trato vocal, as ressonâncias associadas à sua fisiologia, perdas devidas ao atrito viscoso nas paredes internas do trato vocal, suavidade dessas paredes internas, radiação do som nos lábios, acoplamento nasal, flexibilidade associada à vibração das cordas vocais, etc. Alguns desses fatores são modelados por um sistema que combina uma fonte de excitação periódica e outra de ruído branco, aplicadas a um filtro digital variante no tempo. Entretanto, outros fatores são desconsiderados nesse modelamento, pela simples dificuldade ou até impossibilidade de descrevê-los em termos de combinações de sinais, filtros digitais, ou equações diferenciais. Por outro lado, a Teoria dos Sistemas Dinâmicos Não-Lineares ou Teoria do Caos oferece técnicas para a análise de sinais onde não se sabe, ou não é conhecido, o modelo detalhado do mecanismo de produção desses sinais. A análise através dessa teoria procura avaliar a dinâmica do sinal e, assumindo-se que tais amostras provêm de um sistema dinâmico não-linear, medidas qualitativas podem ser obtidas desse sistema. Essas medidas não fornecem informações precisas quanto ao modelamento do processo de produção do sinal avaliado, isto é, o modelo analítico é ainda inacessível. Entretanto, pode-se aferir a respeito de suas características qualitativas, como o número de graus de liberdade que esse modelo apresenta, a taxa de perda de informação do modelo relativamente às condições iniciais, ou a estabilidade local de sua trajetória no espaço de fases.

O problema analisado ao longo deste trabalho trata da busca de novos métodos para extrair informações úteis a respeito do locutor que produziu um determinado sinal de voz. Com isso, espera-se conceber sistemas que realizem a tarefa de reconhecer um pessoa automaticamente através de sua voz de forma mais exata, segura e robusta, contribuindo para o surgimento de sistemas de RAL com aplicação prática. Para isso, este trabalho propõe a utilização de novas ferramentas, baseadas na Teoria dos Sistemas Dinâmicos Não-Lineares, para melhorar a caracterização de uma pessoa através de sua voz. Assim, o mecanismo de produção do sinal de voz é analisado sob outro ponto de vista, como sendo o produto de um sistema dinâmico que evolui em um espaço de fases apropriado. Primeiramente, a possibilidade de utilização dessas técnicas em sinais de voz é verificada. A seguir, demonstra-se como as técnicas para estimação de invariantes dinâmicas não-lineares podem ser adaptadas para que possam ser utilizadas em sistemas de RAL. Por fim, adaptações e automatizações algorítmicas para extração de invariantes dinâmicas são sugeridas para o tratamento de sinais de voz. A comprovação da eficácia dessa metodologia se deu pela realização de testes comparativos de exatidão que, de forma estatisticamente significativa, mostraram o benefício advindo das modificações sugeridas. A melhora obtida com o acréscimo de invariantes dinâmicas da forma proposta no sistema de RAL utilizado nos testes resultou na diminuição da taxa de erro igual (EER) em 17,65%, acarretando um intrínseco aumento de processamento. Para sinais de voz contaminados com ruído, o benefício atingido com o sistema proposto foi verificado para relações sinal ruído (SNRs) maiores que aproximadamente 5 dB.

O avanço científico potencial advindo dos resultados alcançados com este trabalho não se limita às invariantes dinâmicas utilizadas, e nem mesmo à caracterização de locutores. A comprovação da possibilidade de utilização de técnicas da Teoria do Caos em sinais de voz permitirá expandir os conceitos utilizados em qualquer sistema que processe digitalmente sinais de voz. O avanço das técnicas de Sistemas Dinâmicos Não-Lineares, como a concepção de invariantes dinâmicas mais representativas e robustas, implicará também no avanço dos sistemas que utilizarem esse novo conceito para tratamento de sinais vocais.

Palavras-Chave: Reconhecimento de Locutor, Reconhecimento de Voz, Processamento de Sinais, Sistemas Dinâmicos, Teoria do Caos, Séries Temporais.

TITLE: "AUTOMATIC SPEAKER RECOGNITION USING NONLINEAR DYNAMICAL FEATURES"

Abstract

The techniques used to perform automatic speaker recognition aim to identify a person through his/her speech, using computational resources. These techniques are done considering a speech production model. The detailed model for that process should take in account the temporal variation in the vocal tract shape, resonances associated to vocal tract physiology, losses due to heat conduction and viscous friction in the vocal tract walls, softness of these walls, sound radiation in the lips, nasal coupling, flexibility associated to the vocal cords vibration, etc. Some of these physical factors are modeled by a system, which combines a periodic source of excitation and a source of white noise, applied to a time-variant digital filter. However, other factors are not considered in this modeling, because of the intrinsic difficulty or even impossibility of description in terms of signal combinations, digital filters, or differential equations. On the other hand, the Nonlinear Dynamical Systems Theory or Chaos Theory offers techniques to signal analysis whose detailed production mechanism is not known. The analysis using this theory try to evaluate the signal dynamics and, assuming that the samples come from a nonlinear dynamical system, qualitative measures can be estimated. These measurements do not provide precise information about the real model of signal production since the analytical model is not available yet. However, it is possible to estimate the model's qualitative information, such as the number of degrees of freedom, the rate of loss of information related to initial conditions, or the local stability of its trajectory in the phase space.

The problem analyzed in this work is based on the search for new methods to extract useful information about the speaker who produced a speech signal. By solving this problem, new systems that recognize a person through his/her speech in a more exact, secure and robust way can be developed, contributing to the advance of automatic speaker recognition systems with practical application. To accomplish these tasks, this work proposes new tools, based on Nonlinear Dynamical System Theory, to improve the characterization of a speaker using his/her speech. The speech production mechanism is analyzed under a different point of view, as the result of a dynamical system that evolves in an appropriated state space. First, the possibility of using these techniques in speech signals is checked. After, it is demonstrated how the techniques used to estimate nonlinear dynamical invariants can be adapted to be used in speaker recognition systems. At the end, algorithmic adaptations and automatizations to extract dynamical invariants are suggested for speech signals. The effectiveness of this methodology was proved through statistically significative tests of comparative performance, which shown improvement in accuracy using the suggested modifications. The improvement obtained with the addition of dynamical invariants in the proposed way resulted in a decrease of equal error rate (EER) in 17.65%, with an intrinsic increase of processing. For speech signal corrupted by noise, an improvement was obtained with the proposed system for signal to noise ratios (SNRs) higher than 5 dB.

The potential scientific advance obtained from the achieved results from this work is not limited to dynamical invariants used, neither to the characterization of speakers. The possibility of using Chaos Theory in speech signals will allow concepts expansion in any system that digitally processes speech signals. The advance of Nonlinear Dynamical System techniques, like the development of more representative

and robust dynamical invariants, will result in the advance of systems that use this new concept to treat speech signals.

Keywords: Speaker Recognition, Speech Recognition, Signal Processing, Dynamical Systems, Chaos Theory, Time Series.

1 Introdução

O desenvolvimento tecnológico atual na área de Informática permite a construção de sistemas extremamente complexos. Mais especificamente, a área de Processamento Digital de Sinais (DSP) requer recursos que apresentem alto desempenho, pois os algoritmos utilizados podem ter complexidades elevadas. Algumas das técnicas de DSP foram aprofundadas dentro de temas específicos, a fim de tratarem tipos particulares de sinais, como os sinais de voz, de forma a obter resultados melhores. Desse modo, outros campos de atuação surgem para profissionais especializados no tratamento de sinais, como o campo de Processamento Digital da Fala. Essa área compreende algumas sub-áreas bem definidas [RAB 78]: a codificação do sinal vocal, técnicas para melhorar a qualidade do sinal, síntese de fala, reconhecimento automático de fala e reconhecimento automático de locutor.

Várias técnicas foram desenvolvidas a fim de codificar o sinal vocal [MAK 85], objetivando seu armazenamento ou transmissão de forma segura, eficiente e rápida. Assim, foram criados os chamados *vocoders* ou *voice coders*. Um dos principais objetivos dos *vocoders* é a redução da largura de banda necessária à transmissão do sinal de voz. A tecnologia de codificação da forma de onda do sinal vocal tem sido utilizada por décadas e suas técnicas estão bem desenvolvidas. Com o desenvolvimento da tecnologia de circuitos integrados em grande escala (VLSI), a implementação de algoritmos muito complexos está sendo possível, de forma a requerer taxas de transmissão menores que as necessárias anteriormente.

Muito esforço também tem sido feito no sentido de melhorar a qualidade do sinal [RAB 78]. Essa área abrange a remoção de reverberação (ou eco) do sinal de voz e a eliminação de ruído ou a reconstrução de sinais degradados. Tal área divide-se em quatro classes distintas, cada uma com suas vantagens e limitações. A primeira classe concentra-se no domínio espectral em tempos curtos. As técnicas utilizadas nessa classe visam atenuar o ruído através de uma subtração espectral, estimando as componentes do ruído nos intervalos de tempo entre palavras. A segunda classe é baseada no modelamento da voz utilizando métodos iterativos. Esses sistemas estimam os parâmetros que caracterizam o sinal de voz analisado, e reconstroem o sinal livre de ruído, baseando-se na filtragem de Wiener. A terceira classe é baseada em algoritmos que realizam o cancelamento adaptativo de ruído (ANC). A última classe utiliza informações sobre a periodicidade que os sons vozeados apresentam. São utilizadas técnicas baseadas em rastreamento da freqüência fundamental do sinal de voz para sua reconstrução livre de ruído.

Os sistemas que utilizam tecnologia de síntese de fala [KLA 87] objetivam responder aos comandos solicitados através de sons compreensíveis pelo homem. Assim, a síntese de fala é basicamente a conversão de um texto qualquer em um conjunto de fonemas, que podem ser entendidos. As primeiras técnicas desenvolvidas apenas reproduziam algumas palavras ou frases pré-gravadas. Com o avanço tecnológico, a concatenação de fonemas pré-gravados possibilitou a síntese de um número bem maior de palavras, sem a necessidade de armazenamento de todas elas. Os sistemas que utilizam a tecnologia de síntese de fala são usualmente empregados em aplicações remotas, onde apenas um teclado ou terminal é disponível, como no caso de aplicações telefônicas.

O reconhecimento automático de fala (RAF) [RAB 93] consiste, basicamente, na conversão de uma onda acústica para uma equivalência escrita da mensagem contida nessa onda. Muita pesquisa tem sido feita no sentido de descobrir, de forma automática, o que foi dito. Inicialmente, o desafio foi o reconhecimento de palavras isoladas, como

os números de 0 a 9, falados por apenas um locutor. As técnicas então evoluíram para sistemas que reconheciam a concatenação de palavras em uma frase. Atualmente, vários sistemas fazem o reconhecimento automático do texto falado de forma contínua, sem a necessidade de inserção de um período de silêncio entre palavras consecutivas. Com o amadurecimento da pesquisa na área, o enfoque tem sido atribuído aos sistemas para reconhecimento de fala espontânea, identificando-se a ocorrência de eventos comuns numa conversação, como tosse, interjeições, gaguejo ou correções posteriores da pronúncia de uma palavra. Assim, os sistemas para RAF vêm sendo desenvolvidos e fornecem informações específicas a respeito do texto que foi falado em uma locução.

Reconhecimento de locutor consiste em identificar uma pessoa através da análise de uma amostra de sua voz. Esta tarefa pode ser realizada de forma automática através da utilização de recursos computacionais. De acordo com o objetivo desejado, o reconhecimento automático de locutor (RAL) [FUR 97] [ROS 76] é dividido em: a) Identificação: visa determinar qual, dentre os locutores conhecidos, pronunciou a amostra de voz que está sendo avaliada; b) Verificação: deseja-se avaliar uma amostra de voz e aceitá-la (ou não) como pronunciada por um locutor específico. Várias técnicas têm sido utilizadas com sucesso para a identificação e/ou verificação automática de pessoas através da fala. Algumas apresentam melhores resultados quando o texto da amostra de voz avaliada é o mesmo utilizado para o treinamento. Outras enfocam o reconhecimento de uma amostra de voz independentemente do que foi dito. Todas utilizam algoritmos para extração da informação útil da onda acústica, e posterior classificação ou comparação de tais parâmetros.

Este trabalho concentra-se na área de RAL, e propõe novas técnicas para a extração de diferentes tipos de informação útil para caracterizar uma pessoa através de sua voz. Para isso, o processo de produção do sinal de voz é avaliado sob um novo ponto de vista, diferente do modelamento matemático linear existente: é visto como um sistema dinâmico não-linear, associado a atratores com comportamento caótico. Assim, verificando-se primeiramente a validade desse tipo de análise, técnicas são aplicadas que possibilitam a estimação desse novo conjunto de informações. O conjunto de técnicas e ferramentas utilizadas nesse tipo de análise é chamada de Teoria do Caos. O passo final a ser dado na direção de disponibilizar as particularizações dessas técnicas em sinais de voz é através de experimentos práticos que comprovam a eficiência desse novo conjunto de informações, no sentido de caracterizar de forma mais completa a voz de um locutor.

Por vezes a literatura aponta o termo “caos” como algo ruim ou indesejável. Normalmente, as pessoas referem-se a algo desorganizado ou confuso como caótico. Cientificamente, caos implica a existência de imprevisibilidade no sentido da dinâmica que se apresenta, sem ser contudo algo ruim ou indesejável, muitas vezes ao contrário. O conjunto de ferramentas matemáticas, numéricas e geométricas apropriadas para a análise de problemas não-lineares para o qual não existem soluções gerais explícitas fazem parte da Teoria do Caos. Por causa de sua generalidade, a Teoria do Caos ou Teoria dos Sistemas Dinâmicos Não-Lineares pode ser utilizada para analisar uma grande variedade de problemas. Nas últimas décadas, a análise de sistemas dinâmicos não-lineares evoluiu consideravelmente [BAN 99] [KOH 2000] [STA 98] devido a diferentes aspectos. Um dos principais fatores para essa evolução foi o desenvolvimento de computadores poderosos, com grande capacidade de processamento, que permitiram a implementação prática de técnicas para a análise de sinais provenientes de sistemas dinâmicos não-lineares. Além disso, pesquisadores têm encontrado cada vez mais sistemas físicos cujas observações podem ser analisadas através da Teoria do Caos, aumentando o grau de entendimento de sua complexidade. O número de trabalhos

científicos que analisam fenômenos naturais através da Teoria do Caos vem crescendo muito, envolvendo áreas do conhecimento distintas como medicina, física, computação, meteorologia, geologia, engenharias, etc. Atualmente, diversos sistemas são analisados através dessa teoria, apresentando resultados promissores. Como exemplo, pode-se citar sistemas de escoamento de água, previsão do tempo, análise de sinais biológicos como eletroencefalograma (EEG), eletrocardiograma (ECG) e sons pulmonares, análise mercadológica, reações químicas, evolução de populações de animais e plantas, entre outros.

O problema analisado ao longo deste trabalho trata da busca de novos métodos para extrair informações úteis a respeito do locutor que produziu um determinado sinal de voz. Com isso, espera-se conceber sistemas que realizem a tarefa de reconhecer um pessoa automaticamente através de sua voz de forma mais exata, segura e robusta, contribuindo para o surgimento de sistemas de RAL com aplicação prática. Para isto, o mecanismo de produção do sinal de voz é analisado sob outro ponto de vista, como sendo o produto de um sistema dinâmico que evolui em um espaço de fases apropriado. A modelagem detalhada do processo de produção dos sons vocais deve levar em consideração diversos fatores como a variação temporal da forma do trato vocal, as ressonâncias associadas à sua fisiologia, perdas devidas ao atrito viscoso nas paredes internas do trato vocal, suavidade dessas paredes internas, radiação do som nos lábios, acoplamento nasal, flexibilidade associada à vibração das cordas vocais, etc [KUM 96] [DEL 87]. Alguns desses fatores são matematicamente modelados por um sistema que combina uma fonte de excitação periódica e outra de ruído branco, aplicadas a um filtro digital variante no tempo [RAB 78] [DEL 87]. Entretanto, outros fatores são desconsiderados nesse modelamento, pela simples dificuldade ou até impossibilidade de descrevê-los em termos de combinações de sinais, filtros digitais, ou equações diferenciais. Assim, a Teoria dos Sistemas Dinâmicos Não-Lineares ou Teoria do Caos oferece técnicas para a análise de sinais onde não se sabe, ou não é conhecido, o modelo detalhado do mecanismo de produção desses sinais. A dinâmica do sinal é avaliada e medidas qualitativas podem ser obtidas desse sistema. Essas medidas não fornecem informações específicas quanto ao modelamento do processo de produção do sinal avaliado, isto é, o modelo matemático é ainda inacessível. Entretanto, pode-se aferir a respeito de suas características qualitativas, como o número de graus de liberdade que esse modelo apresenta, a taxa de perda de informação do modelo relativo às condições iniciais, ou a estabilidade local de sua trajetória no espaço de fases. Assim, espera-se que as estimativas de invariantes dinâmicas como dimensão e expoentes de Lyapunov agreguem informação até então não considerada nos sistemas de RAL atuais, elevando as taxas de reconhecimento desses sistemas.

1.1 Aplicações do RAL

A possibilidade de utilização de técnicas para RAL em sistemas reais, de aplicação imediata, é uma das grandes motivações dos trabalhos desenvolvidos na área. Outras técnicas também poderiam ser utilizadas para a confirmação automática da identidade de uma pessoa através de medidas biométricas, sem intervenção humana. Alguns exemplos são o reconhecimento da íris, reconhecimento de face, análise da impressão digital, verificação das características geométricas da mão, avaliação da forma do caminhar, análise da assinatura, etc. Por outro lado, a autenticação biométrica através da voz possui várias vantagens intrínsecas, quando comparadas às outras técnicas biométricas, como:

- Facilmente adquirida, sem a necessidade de hardware específico. Apenas um microfone de custo reduzido e um conversor A/D são capazes de adquirir a voz do locutor com precisão.
- Não intrusiva, isto é, não existe a necessidade de contato físico do usuário com o sistema de captação da voz.
- Natural e de fácil aceitação. A maioria das pessoas se comunica através da voz, que é um método extremamente difundido e utilizado, excetuando-se as pessoas com necessidades específicas.
- Não requer treinamento, ou seja, os usuários podem ser autenticados sem que necessitem aprender a lidar com o sistema.
- Pode ser facilmente utilizada em redes telefônicas, permitindo a autenticação remota, sem a necessidade de qualquer equipamento extra, uma vez que o próprio aparelho telefônico já possui o transdutor para a captação do sinal de voz.

Pode-se identificar muitas aplicações práticas para os sistemas de RAL. De uma forma geral, os sistemas para controle de acesso, ou validação de identidade através de senhas ou cartões poderiam ser substituídos por (ou agregados a) sistemas de RAL, estabelecendo-se uma confiabilidade superior. Isso se deve ao fato de que cartões magnéticos ou senhas podem ser extraviados, e a maioria dos sistemas atuais não possui qualquer meio para identificar biometricamente o usuário. A identificação realizada se limita apenas à leitura de um código magnético, associado (na melhor das hipóteses) à verificação de uma seqüência de dígitos (senha).

Dentre as aplicações para os sistemas de RAL, pode-se destacar:

- **Aplicações bancárias**: neste caso, pode-se desejar que uma senha possa ser substituída ou agregada a uma verificação da identidade do cliente pela voz. Tal aperfeiçoamento poderia também ser estendido aos caixas 24 horas. O nível de confiabilidade em tais sistemas deve ser elevado.
- **Controle de acesso a áreas restritas**: localidades onde o acesso deve ser restrito a um número limitado de pessoas, como instalações militares, laboratórios industriais ou presídios, poderiam fazer uso da tecnologia de RAL para o controle do acesso de forma segura, prática e confiável.
- **Controle de acesso em softwares**: alguns locais da memória de computadores podem conter informações confidenciais. Tais informações armazenadas magneticamente podem ter seu acesso restrito a apenas algumas pessoas. A autenticação da identidade dessas pessoas pode ser feita através de técnicas de RAL.
- **Aeroportos**: até mesmo os cartões de embarque poderiam ser substituídos por amostras de voz do passageiro, o que poderia evitar fraudes no transporte de pessoas.
- **Ponto eletrônico**: a tarefa de registrar a presença de um funcionário em uma empresa é normalmente realizada através dos chamados cartões-ponto, ou cartões magnéticos. Tais métodos dão margem a pequenas fraudes, ou seja, não há um controle eficiente de quem está registrando a presença. Isso poderia ser evitado através da introdução de sistemas para RAL, que apresentam maior confiabilidade e segurança.
- **Assinatura eletrônica**: qualquer comercialização à distância (*e-commerce*) poderia ser autenticada através de uma assinatura eletrônica por voz. Até mesmo as transações que envolvem cartões de crédito poderiam ser efetuadas de forma segura, com a confirmação da identidade vocal do usuário.

- **Elevadores**: em prédios residenciais, pode-se imaginar a aplicação das técnicas de RAL nos elevadores, de forma a permitir o acesso dos moradores do edifício, e restringir o acesso às demais pessoas [PET 99a] [PET 99b] [PET 99c]. Assim, um morador poderá entrar no elevador, dizer seu nome e a verificação da identidade é automaticamente realizada, fazendo com que o elevador conduza-o ao andar onde mora.
- **Hotéis**: quando uma pessoa é hospedada em um hotel, ela recebe uma chave (às vezes magnética). Tal chave deve ser guardada, ou deixada na recepção. Pode-se pensar em substituir a utilização de chaves por senhas vocais. Dessa forma, não existirá mais a necessidade de cuidar de uma chave, ou o perigo de uma pessoa desconhecida ter acesso ao quarto de um hóspede.
- **Acesso a redes de computadores**: em substituição aos mecanismos de proteção de redes de computadores (*firewalls*), que normalmente utilizam senhas, pode-se pensar numa verificação biométrica de identidade, como o RAL. Isso tornaria o sistema mais seguro e mais prático, uma vez que não seria necessária memorização da seqüência de dígitos.
- **Aplicações militares**: pode-se imaginar muitas aplicações dos sistemas para RAL enfocando atividades militares. Desde a troca de informações entre agentes militares até mesmo a ativação de alguma arma perigosa podem ser feitas de forma bem mais segura através da utilização de sistemas para RAL.

1.2 Porque o RAL ainda é um problema em aberto ?

Existem muitas dificuldades práticas que devem ser superadas antes da utilização dos sistemas de RAL em aplicações onde o grau de confiabilidade deva ser elevado. Alguns desses fatores vêm sendo foco de intensas pesquisas nos últimos anos, e suas influências são cada vez mais reduzidas. Entretanto, a inserção das técnicas de RAL em sistemas práticos ainda é relativamente pequena, se considerarmos as aplicações em potencial existentes. Isso em muito se deve à dificuldade em sobrepor alguns dos fatores que mais atrapalham o correto reconhecimento de locutores. Quando a influência desses fatores estiver substancialmente controlada, restarão poucas barreiras técnicas para a utilização em larga escala do RAL no dia-a-dia das pessoas.

Em um sistema de RAL, deve-se prever a possibilidade de que pessoas tentem enganar o sistema, através da imitação da voz de outras. Outro fator a ser avaliado é a possibilidade do usuário utilizar gravações obtidas de locutores cadastrados para acessar o sistema de RAL. As modificações na voz decorrentes de patologias no aparelho vocal ou mudanças no estado emocional dos locutores podem acarretar uma falsa rejeição por parte do sistema. Também o ambiente de gravação deve ser cuidadosamente escolhido e suas principais características devem ser mantidas o máximo possível constantes ao longo das etapas de treinamento e reconhecimento. Outro potencial problema encontrado é relativo à eficiência computacional requerida, uma vez que técnicas computacionalmente pesadas podem ser utilizadas. A utilização de diferentes tipos de transdutores e durações de fala excessivamente curtas também podem prejudicar o reconhecimento.

1.2.1 Imitadores

Muitas pessoas conseguem imitar a voz de outras. Isso ocorre com mais freqüência com personalidades famosas (políticos, artistas, intelectuais, etc). Quando

ouvidas por pessoas comuns, a voz do imitador parece ser muito semelhante à do imitado. De fato, os imitadores conseguem modificar a configuração de seu trato vocal e produzir sons com algumas características que imitam a produção natural do som de outras pessoas.

Uma maior robustez dos sistemas de RAL contra imitadores é adquirida quando várias informações são avaliadas simultaneamente. Os imitadores conseguem, por exemplo, reproduzir a freqüência fundamental (*pitch*) do imitado com maior facilidade. Entretanto, outras informações, como as freqüências das formantes, não podem ser facilmente imitadas por estarem associadas fortemente com a anatomia do trato vocal. Assim, quando são utilizadas várias informações simultaneamente, ou quando as características analisadas forem de difícil imitação, o sistema para RAL será bem mais imune aos imitadores.

1.2.2 Gravações

Pode-se imaginar que a pré-gravação da voz de uma pessoa e sua reprodução para o sistema de RAL possa acarretar a aceitação incorreta de impostores. Isso é especialmente verdade com a utilização de recursos de gravação e reprodução sofisticados. A figura 1.1 ilustra a utilização de pré-gravações e posterior reprodução.

Os sistemas para RAL pouco podem fazer contra esse tipo de erro. Os métodos mais eficientes que se pode conceber seriam uma vigilância permanente do dispositivo de aquisição de voz, ou o sistema requerer uma palavra pseudo-aleatória específica. Uma vigilância permanente, a fim de garantir que aparelhos de reprodução sonora não sejam utilizados, normalmente não é algo desejado em um sistema.

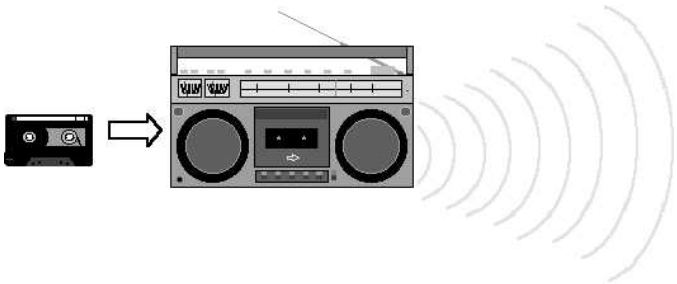

FIGURA 1.1 - Utilização de gravações em sistemas de RAL

O aspecto de o sistema requerer uma palavra pseudo-aleatória específica abrange a identificação ou verificação dependente de texto. Neste caso, palavras diferentes da requerida apresentariam um valor superior de distorção, impedindo uma falsa aceitação. O fato de a palavra ser pseudo-aleatória dificulta o sucesso de uma pré-gravação. O problema decorrente dessa escolha pseudo-aleatória de palavras está na aquisição de amostras de voz para treinamento do sistema. Os locutores cadastrados devem neste caso fornecer amostras de voz com diferentes textos, de forma a expandir as opções de aleatoriedade possíveis.

1.2.3 Patologias do aparelho vocal

Mudanças no aparelho vocal humano devidas a patologias, como resfriados ou rouquidão, são freqüentes. Essas modificações alteram bastante a voz produzida. Isso

pode prejudicar muito a exatidão do sistema de RAL utilizado, através da rejeição incorreta de um locutor cadastrado.

A melhor forma de contornar esse problema seria a extração apropriada de informações relativas à anatomia do trato vocal. Desta forma, a produção de sons distorcidos diminuiriam o erro por parte do sistema. Entretanto, essa extração de informações pode não ser bem sucedida, pois pode haver uma mudança fisiológica no trato vocal, como inchaço, presença de material orgânico indesejado, ou reposicionamento de estruturas constituintes do aparelho vocal. Nesses casos, haverá uma iminente degradação do desempenho do sistema de RAL, podendo haver um rejeição indesejada de um locutor cadastrado.

1.2.4 Mudanças no estado emocional

Podem ser percebidas as variações na voz das pessoas, quando estão submetidas a um estado emocional intenso. É fácil perceber a modificação na voz de uma pessoa deprimida, estressada ou com medo, por exemplo. As mudanças que ocorrem na voz de pessoas submetidas a fortes estados emocionais podem ocasionar um reconhecimento incorreto de identidade.

Da mesma forma que as patologias do aparelho vocal, devem ser utilizadas informações que pouco variem, quando a voz efetivamente produzida apresenta distorções. Assim, informações que estão relacionadas com a fisiologia ou estrutura do trato vocal são mais eficientes nesses casos. Entretanto, a utilização das informações atualmente conhecidas pouco contribui para amenizar tais problemas.

1.2.5 Ambiente de gravação

A questão do ambiente onde as amostras de voz são capturadas e transmitidas pode modificar significativamente ao desempenho do sistema de RAL. Aspectos relativos à conversão analógico-digital (A/D), como a taxa de amostragem e a resolução (número de bits/amostra) adequadas, o tipo de transdutor (microfone) utilizado, o canal de transmissão (linha telefônica, rádio, fio de cobre, etc) e o isolamento acústico devem ser analisados com cuidado. O ruído associado deve ser especialmente avaliado. Uma baixa relação sinal ruído (SNR) prejudica consideravelmente o desempenho de qualquer sistema de RAL.

A escolha da taxa de amostragem e resolução adequadas garante a preservação da informação contida no sinal, durante o processo de digitalização. O canal de transmissão utilizado deverá também preservar da melhor forma possível o espectro do sinal transmitido. O isolamento acústico no local da aquisição da voz e o ruído associado são especialmente importantes. Esses fatores devem ser especificados com cautela. O ruído deverá ser, na medida do possível, eliminado do sinal adquirido através de técnicas de DSP eficientes.

1.2.6 Eficiência computacional

Os algoritmos que vêm sendo desenvolvidos para processar os sinais de voz requerem hardware de alto desempenho para que possam fornecer resultados em tempo hábil. De acordo com a técnica utilizada, esse desempenho requerido pode ser maior ou menor. De qualquer forma, qualquer sistema de RAL deve ser capaz de fornecer uma

resposta em frações de segundo, evitando com isso a espera excessiva do usuário e possível rejeição a sua utilização.

De uma maneira geral, esse problema tende a ser gradualmente amenizado com o avanço da microeletrônica. A figura 1.2, adaptada de [ARC 2002], ilustra o acentuado crescimento do poder computacional do hardware existente, para os microprocessadores da família Intel. Por outro lado, as técnicas também têm adquirido uma alta sofisticação, o que acaba contrabalançando o problema. Assim, a tendência é que se possam construir sistemas cada vez mais elaborados, com exatidões crescentes, mantendo-se um tempo de resposta adequado.

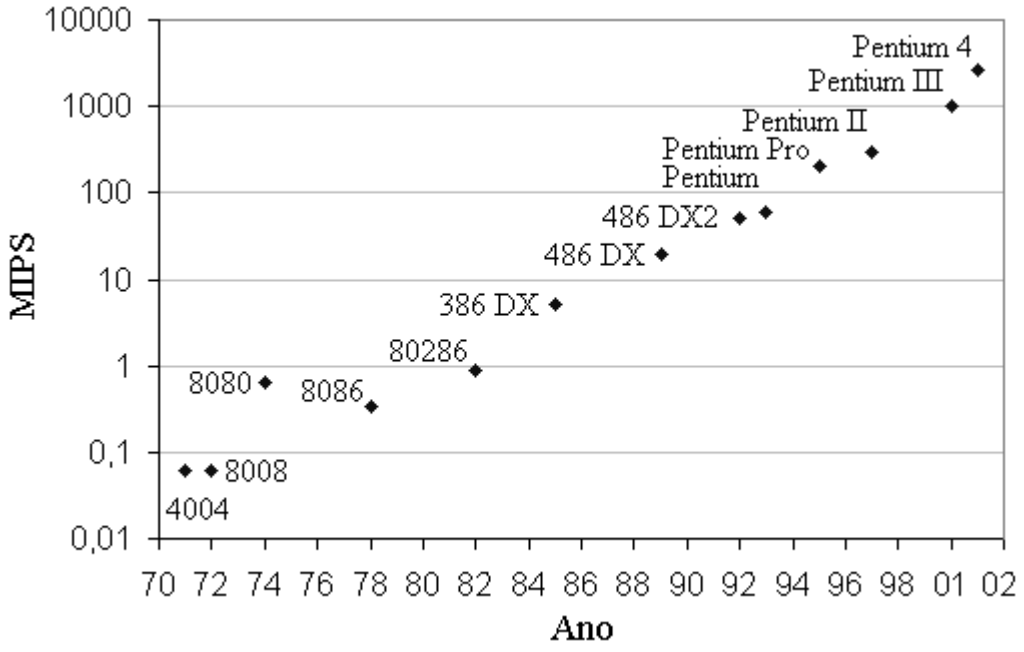

FIGURA 1.2 - Avanço do poder de processamento dos microprocessadores Intel

1.2.7 Transdutores

Existem tipos diferentes de transdutores onde é possível realizar a aquisição da voz. Microfones construídos com transdutores à base de carbono diferem bastante dos construídos com eletreto, por exemplo. É possível observar que a variabilidade de microfones causa degradação significativa no desempenho de sistemas de RAL. Outro grande problema que advém da utilização de microfones refere-se a sua localização durante a fala. De acordo com o posicionamento do microfone, vozes de um mesmo locutor são gravadas com características distintas.

Existem evidências que os efeitos de transdutores são não-lineares [QUA 98] [QUA 2000], e assim difíceis de serem removidos. Tais efeitos podem permanecer nas informações extraídas do sinal de voz, e o modelo para as características acústicas de um locutor poderá estar associado com essas distorções. O que se procura fazer para reduzir esse efeito é uma normalização no grau de similaridade obtido através dos parâmetros de um locutor, utilizando amostras de locutores em tipos diferentes de transdutores [REY 97].

1.2.8 Duração da fala

Outro fator importante relativo à confiabilidade de resultados em sistemas de RAL diz respeito ao tempo de fala disponível tanto para treinamento dos padrões vocais de um locutor, quanto para o reconhecimento das amostras de voz. Quanto maior for a duração das amostras de voz de treinamento, mais similar será o padrão vocal gerado às características do locutor, e da mesma forma, quanto maior for a duração da amostra de voz a ser reconhecida, maior será a probabilidade que o resultado da análise corresponda ao resultado correto.

Pouco se pode fazer se não houver amostras de voz suficientes fornecidas ao sistema. A degradação do desempenho dos sistemas de RAL é grande quando pouca informação é fornecida. O melhor que se pode esperar é que os dados inseridos no sistema sejam aproveitados ao máximo, viabilizando um reconhecimento correto com menores quantidades de amostras de voz.

1.3 Outras abordagens

Uma vez que vários problemas intrínsecos na área de RAL ainda podem ser revistos e amenizados, um vasto campo de pesquisa pode ainda ser explorado. Assim, várias abordagens podem ser utilizadas para diminuir a influência de um ou mais fatores que degradam o desempenho dos sistemas de RAL. Algumas dessas abordagens que vêm sendo experimentadas são:

- **Identificação do tipo de transdutor** utilizado na aquisição do sinal de voz e posterior tratamento direcionado do sinal [REY 97] [REY 2000]. Neste caso, a identificação do tipo de transdutor utilizado é realizada antes do reconhecimento do sinal de voz. A seguir, uma normalização nos valores de similaridade é aplicada, buscando maior independência do tipo de microfone.
- **Análise de ocorrências específicas** na produção da voz como a coarticulação entre fonemas [ADA 98], ou transições entre consoantes e vogais [YEG 98]. Aqui, eventos específicos são utilizados para distinguir locutores.
- **Extração de novos tipos de informação** que possam melhor caracterizar os locutores [PET 2002a] [PET 2001] [PET 2002b] [PET 2002c] [GOP 99] [ZIL 98] . Neste caso, busca-se a extração de parâmetros que possuam a capacidade de melhorar os sistemas de RAL, agregando informação não considerada anteriormente.
- **Remoção da influência do canal de transmissão** nas informações extraídas. Algumas técnicas objetivam reduzir a influência do canal de transmissão, como a subtração dos parâmetros cepstrais médios (CMS) [ROS 94b] [WES 97] ou filtragem RASTA [HER 94].
- **Utilização de classificadores diferentes** [HE 97] [YUA 99]. Neste caso, procura-se encontrar ou desenvolver classificadores que apresentem elevada exatidão no RAL, ou combinar diferentes classificadores.
- **Construção de modelos universais** de padrões de voz [REY 2000]. Assim, o padrão de cada locutor que é cadastrado no sistema é derivado desse padrão universal, chamado Modelo de Impostores Universal (UBM). Quando um sinal for avaliado, ele é comparado com o modelo do locutor e com o UBM, e os valores de similaridade são subtraídos, para que o resultado seja o mais independente possível de fatores externos, como ruído, que degradaria ambas similaridades, mas não a diferença entre elas.

Este trabalho trata especificamente da abordagem de extração de novos tipos de informação. A aplicação da Teoria do Caos em sistemas de RAL, explorada neste trabalho, vem com o objetivo de analisar a produção do sinal de voz como produto de um sistema dinâmico não-linear e com isso extrair informações qualitativas desse sinal, de forma que esses novos tipos de informação possam ajudar na diferenciabilidade de locutores.

Uma das motivações deste trabalho fundamenta-se principalmente no fato de que atratores reconstruídos a partir de sinais biológicos podem conter propriedades caóticas devidas ao caos inerente na dinâmica do mecanismo de geração desses sinais. Essa análise já foi feita em alguns sinais biológicos como EEGs, ECGs e sons pulmonares, que apresentaram características caóticas e estão sendo estudadas sob o ponto de vista dinâmico não-linear [KOH 2000] [OLI 99] [ABA 93] [STA 98]. Especificamente para sons vocais, uma caracterização do ponto de vista dinâmico não-linear poderá oferecer informações adicionais a respeito do locutor que produziu os sinais, contribuindo para o melhoramento dos sistemas de RAL existentes. Além disso, acreditamos ser possível particularizar a implementação de algoritmos computacionais de análise de séries temporais através das técnicas da Teoria do Caos especificamente para os sinais de voz. Com isso, espera-se que a utilização dessas ferramentas possa ser otimizada para a análise de sons vocais.

Com a inclusão de informações obtidas através da análise dinâmica não-linear da voz de locutores, espera-se uma redução de alguns problemas encontrados em sistemas de RAL, como maior robustez contra imitadores, aumento do reconhecimento mesmo quanto existem patologias no aparelho vocal ou quando o estado emocional do locutor está alterado, maior robustez em ambientes de gravação distintos, maior independência de tipos diferentes de transdutores e menor necessidade de fornecimento de grandes amostras de voz pelos locutores. Por outro lado, a eficiência computacional do sistema deve ser reduzida através da inclusão de algoritmos para análise do sinal através da Teoria do Caos. Alguns desses efeitos são explorados extensamente ao longo deste trabalho.

1.4 Pesquisa na área

A área de RAL têm sido foco de intensa pesquisa e estudo atualmente. Muitos trabalhos podem ser encontrados que tratam de métodos de extração de informações ou classificação de tais parâmetros em sistemas de RAL. Vários congressos fazem referência, nos tópicos de interesse, especificamente à área de RAL. Da mesma forma, revistas altamente renomadas fazem clara referência à área de RAL, sendo este um assunto de grande interesse.

Na área de Teoria dos Sistemas Dinâmicos Não-Lineares, muitos trabalhos buscam identificar aplicações práticas para o conjunto de técnicas disponíveis para a análise de sinais caóticos. Várias revistas e congressos abordam a análise de séries temporais, incluindo sinais biológicos, através da Teoria do Caos. Alguns dos principais veículos de divulgação de importantes trabalhos em ambas áreas de pesquisa são mostrados nas tabelas 1.1 e 1.2.

TABELA 1.1 - Alguns dos principais jornais nas áreas de RAL e Teoria do Caos

Veículo	Acesso através da Internet
IEEE Transactions on Speech and Audio Processing	http://www.ieee.org/organizations/society/sp/tsa.html
Speech Communication	http://www.elsevier.nl/locate/specom
The Journal of the Acoustical Society of America	http://asa.aip.org/jasa.html
Computer, Speech & Language	http://www.academicpress.com/www/journal/0/@/0/la.htm
Chaos Solitons, & Fractals	http://www.elsevier.nl/locate/chaos
Physica D	http://www.elsevier.nl/locate/physd
Journal of the Physical Society of Japan	http://wwwsoc.nacsis.ac.jp/jps/jpsj/
Physical Review Letters	http://prl.aps.org/prlinfo.html
Journal of Nonlinear Science	http://link.springer-ny.com/link/service/journals/00332/index.htm
Stochastics and Dynamics Journal	http://ejournals.wspc.com.sg/sd/sd.html

TABELA 1.2 - Alguns dos principais congressos e *workshops* nas áreas de RAL e Teoria do Caos

Veículo	Acesso através da Internet
IEEE International Conference on Acoustics, Speech and Signal Processing	http://www.icassp2002.com/
European Conference on Speech Communication and Technology	http://eurospeech2001.org/
International Conference on Spoken Language Processing	http://cslr.colorado.edu/icslp2002/
International Conference On Signal Processing Applications and Technology	http://www.icspat.com/
Audio and Video-based Person Authentication	http://www.hh.se/avbpa/
2001: A Speaker Odyssey – The Speaker Recognition Workshop	http://www.odyssey.westhost.com/
Dynamics Day Europe 2002	http://www.iwr.uni-heidelberg.de/dd02/
International conference on New Directions in Dynamical Systems	http://ndds.math.h.kyoto-u.ac.jp/
Fractal 2002 Complexity and Fractals in Nature	http://www.kingston.ac.uk/fractal/
International Conference on Chaos and Nonlinear Dynamics	http://www.chaos.umd.edu/DDays2002/
SIAM Conference on Applications of Dynamical Systems	http://www.siam.org/meetings/ds01/index.htm
Nonlinear Dynamics of Electronic Systems	http://www.ndes2001.tudelft.nl/

Os trabalhos que podem ser encontrados através dos veículos mostrados nas tabelas 1.1 e 1.2 constituem grande parte da informação atualizada disponível tanto na área de RAL quanto na área de Teoria do Caos (com interesse especial em possíveis aplicações).

1.5 Conteúdo da tese de doutorado

Este trabalho está dividido em 5 capítulos. O primeiro capítulo abrange basicamente uma introdução, a identificação clara do problema a ser resolvido, motivações e aspectos relativos ao posicionamento da proposta de tese no contexto das áreas cobertas. A abordagem adotada para a concretização dos resultados é também mencionada.

O segundo capítulo trata das técnicas conhecidas para o RAL. Este capítulo mostra o estado-da-arte dessa tecnologia e referencia os principais trabalhos encontrados na literatura. Há um aprofundamento descritivo das principais técnicas, como a aquisição do sinal de voz, extração e seleção da informação útil, e modelamento.

O terceiro capítulo desenvolve primeiramente os principais conceitos por trás da Teoria do Caos. Um enfoque em séries temporais é utilizado, uma vez que a aplicação desses conceitos se dará em sinais de voz. A seguir, a aplicação dessas técnicas é mostrada através da descrição dos principais algoritmos existentes para a reconstrução do atrator associado a séries temporais. Por fim, esse capítulo descreve os métodos para a extração de invariantes dinâmicas não-lineares, que serão amplamente empregados nos capítulos seguintes.

O quarto capítulo descreve em detalhes as propostas deste trabalho, fundamentando-se nas questões técnicas assimiladas nos capítulos anteriores. Inicialmente é verificada a possibilidade de utilização das técnicas disponibilizadas pela Teoria do Caos em sinais de voz. A seguir, são apresentados os métodos desenvolvidos para a correta aplicação dessas técnicas em séries temporais provenientes de sinais de voz. É mostrado como utilizar as informações já conhecidas para o RAL, agregando-se os novos tipos de informações aos sistemas existentes. É verificada a real eficácia dos métodos propostos através de experimentos práticos. Neste capítulo também é feita uma análise de significância estatística, objetivando validar os resultados.

Ao final, o quinto capítulo enfoca as principais conclusões do trabalho desenvolvido. Um resumo dos resultados alcançados é fornecido, assim como são evidenciadas as principais contribuições originais do candidato. Por fim, trabalhos futuros são apontados, e os resultados são discutidos. Em anexo a esta tese, as principais publicações do candidato são reproduzidas.

1.6 Resumo

Este capítulo procurou de forma geral situar o leitor na área de análise explorada neste trabalho, descrever os principais problemas que serão tratados e fornecer motivações para suas soluções. Inicialmente, as ramificações da área de processamento da fala foram mencionadas e foi identificada a sub-área de atuação enfocada. A seguir, algumas motivações foram apresentadas para a escolha dessa sub-área de atuação, assim como algumas das principais dificuldades encontradas. As propostas para o melhoramento técnico do estado-da-arte desses sistemas foram mencionadas e o esforço

acadêmico existente foi descrito através da citação dos principais veículos de divulgação existentes que tratam de áreas afins.

2 Reconhecimento Automático de Locutor

2.1 Introdução

O objetivo principal desse capítulo é descrever a metodologia atualmente empregada em sistemas de RAL. Primeiramente são descritas as formas com que os sistemas de RAL podem atuar. A seguir, é mostrada uma visão geral da arquitetura desses sistemas. O estudo das principais etapas dessa arquitetura é aprofundado, abrangendo os métodos para aquisição do sinal de voz, extração da informação útil, geração dos padrões vocais e medida de distorção.

2.2 Tipos de reconhecimento de locutor

As técnicas utilizadas para o RAL variam de acordo com a natureza do problema que se deseja solucionar. O problema do RAL divide-se primeiramente em **Identificação** e **Verificação**, conforme o tipo de reconhecimento que se deseja. Os sistemas de identificação contêm um grupo conhecido e limitado de locutores cadastrados. A amostra avaliada pode necessariamente pertencer a um desses locutores (**Grupo Fechado**), ou pode vir de um locutor desconhecido (**Grupo Aberto**). Tanto os sistemas de identificação quanto de verificação também podem ser **Dependentes do Texto** ou **Independentes do Texto**. Essa divisão refere-se à necessidade (ou não) das amostras de voz utilizadas para treinamento do sistema conterem o mesmo texto das amostras de reconhecimento.

De acordo com o objetivo desejado, o reconhecimento automático de locutor pode ser dividido em [FUR 97] [ROS 76]:

- **Identificação**: visa determinar qual, dentre os locutores conhecidos, pronunciou a amostra de voz que será avaliada. Dada uma amostra de voz desconhecida, o sistema deverá compará-la com todos os padrões de referência de cada locutor cadastrado no sistema, e identificar o mais semelhante. Desta forma, o processamento necessário para identificar uma amostra de voz desconhecida será maior, quanto maior for o número de locutores cadastrados no sistema. Analogamente, a probabilidade de ocorrência de erros de classificação aumenta quando a população de locutores cadastrados aumentar.
- **Verificação**: deseja-se avaliar uma amostra de voz e aceitá-la (ou não) como pronunciada por um locutor específico. Assim são fornecidos ao sistema uma amostra de voz e uma identificação correspondente ao suposto falante. O sistema deverá então determinar se a amostra de voz fornecida é (ou não) suficientemente similar aos padrões de referência do locutor apontado. Neste caso, apenas uma comparação é realizada, independente do número de locutores cadastrados. Desta forma, mesmo em grandes populações de locutores que utilizam o sistema, o processamento necessário para realizar uma verificação praticamente não muda. A probabilidade de ocorrência de erros de verificação mantém-se a mesma para cada locutor, mesmo quando há um aumento na população de locutores cadastrados.

As técnicas de identificação de locutor utilizam, na etapa de treinamento, padrões obtidos a partir de um conjunto finito de locutores. Na avaliação da amostra a ser reconhecida, podemos identificar um divisão entre:

- **Grupo Fechado**: quando a amostra de voz sob avaliação pertence, necessariamente, a um dos locutores cadastrados no sistema. Assim, deve-se apenas identificar qual dos padrões conhecidos apresenta a maior semelhança com a amostra avaliada.
- **Grupo Aberto**: há possibilidade da amostra sob avaliação não pertencer a qualquer um dos locutores conhecidos. Assim, mesmo identificando qual dos padrões conhecidos apresenta a maior semelhança com a amostra avaliada, deve-se verificar se esse valor de similaridade é suficiente para reconhecer o locutor. O grau de dificuldade associado a esse tipo de identificação é maior uma vez que há a possibilidade da amostra avaliada não pertencer a qualquer um dos locutores conhecidos.

Quanto à necessidade do texto da amostra de voz sob avaliação ser o mesmo das amostras de treinamento, tanto os sistemas de identificação, quanto de verificação podem ser:

- **Dependentes de Texto**: nos sistemas dependentes de texto, o usuário é induzido a pronunciar uma palavra específica. Tal palavra foi previamente utilizada para o treinamento do sistema e geração dos padrões de referência.
- **Independentes de Texto**: o RAL independente de texto visa identificar ou verificar a identidade de um locutor utilizando qualquer amostra de voz produzida pelo mesmo. Nesses sistemas, a amostra de voz fornecida é avaliada e deseja-se identificar nela algumas singularidades que a associe diretamente aos padrões de referência, previamente armazenados. Há um grau de dificuldade superior para implementação de técnicas para RAL independentes de texto, quando comparadas às dependentes de texto. Também há possibilidade dos sistemas independentes de texto não chegarem a resultados conclusivos, caso a mostra de voz fornecida não contenha muita informação útil, impossibilitando um reconhecimento adequado.

2.3 Visão geral de sistemas de RAL

Um sistema para RAL é composto basicamente pelas etapas de **treinamento** e **reconhecimento**. A etapa de treinamento normalmente é realizada antes de se tentar reconhecer qualquer amostra de voz. Essa etapa abrange a aquisição do sinal vocal dos locutores que serão cadastrados no sistema, extração de informações úteis dessas amostras, geração dos padrões que serão utilizados como referência na etapa de reconhecimento e identificação dos limiares de similaridade associados aos padrões gerados, para o caso de verificação de locutor. No caso de um sistema de identificação de locutor, a geração dos limiares não é necessária. A etapa de reconhecimento abrange a aquisição do sinal de voz que será avaliado, extração de informações úteis e comparação dessas informações com os padrões gerados na etapa anterior. Se o sistema for de verificação de locutor, a medida de distorção é realizada apenas com um padrão, e o limiar indicará se a identidade foi ou não aceita. Para a identificação de locutor, a medida de distorção é realizada com cada um dos padrões de voz gerados no treinamento. Nesse caso, o padrão que apresentar menor distorção é escolhido. A figura 2.1 ilustra as etapas de treinamento e reconhecimento para um sistema genérico de RAL.

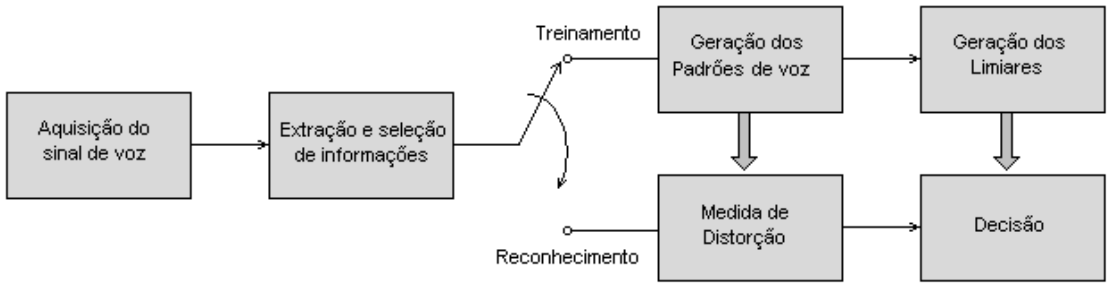

FIGURA 2.1 - Etapas de treinamento e reconhecimento de um sistema de RAL

2.4 Aquisição do sinal de voz

A aquisição do sinal de voz consiste na conversão das ondas sonoras em sinais elétricos através de um transdutor (microfone), filtragem desse sinal elétrico e conversão analógico-digital (A/D) desse sinal. Esse processo é ilustrado na figura 2.2.

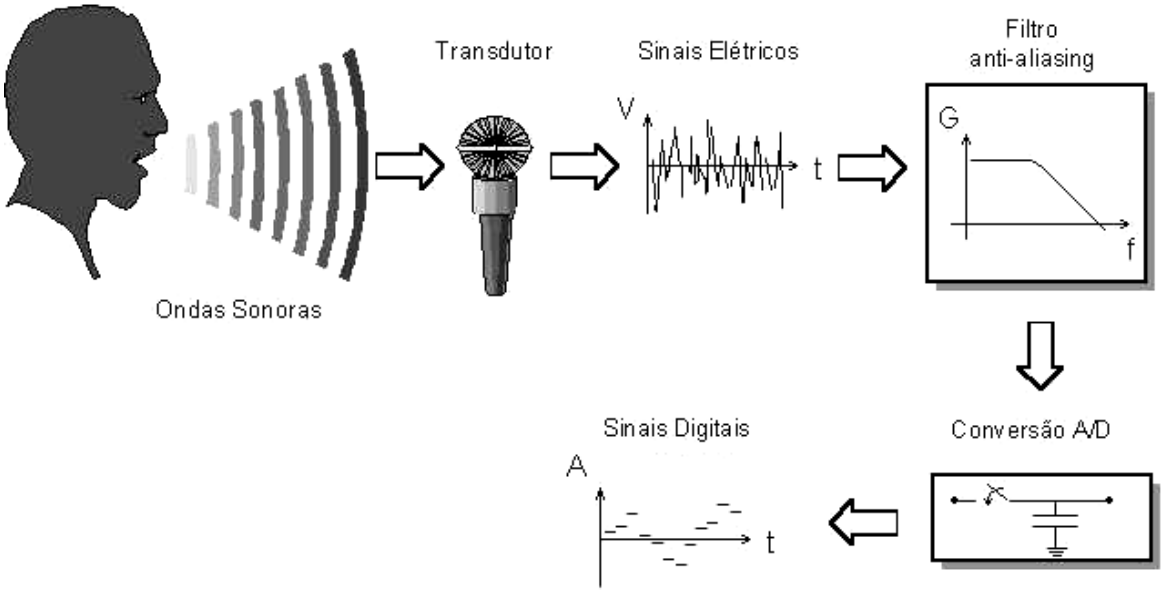

FIGURA 2.2 - Processo de aquisição do sinal de voz

A conversão das ondas acústicas em sinais elétricos é realizada através de um transdutor. As especificações técnicas dos transdutores encontrados no mercado variam de acordo com o modelo e a aplicação específica.

O sinal de voz não é limitado em freqüência. Ele possui componentes de freqüência em uma faixa bastante larga, mesmo que com amplitudes muito baixas. A filtragem do sinal analógico deve ser realizada através de um filtro passa-baixas, antes da conversão A/D [RAB 78]. Tal filtro, chamado filtro *anti-aliasing*, objetiva suprimir as componentes de freqüência do sinal analógico superiores à metade da freqüência de amostragem. Esse critério é conhecido como critério de Nyquist [PRO 96].

A conversão analógico-digital do sinal elétrico filtrado é realizada através de um amostrador. O sinal amostrado é representado de forma digital, e pode-se então proceder à aplicação de técnicas de processamento digital de sinais. A freqüência de amostragem

utilizada deve ser tal que preserve a maior parte da informação contida no sinal de voz. O sinal de voz geralmente possui componentes de freqüência superiores a 5KHz com amplitudes baixas. Assim, freqüências de amostragem superiores a 11KHz preservam grande parte da informação contida no sinal. Outro fator a ser avaliado na conversão analógico-digital é o número de níveis que o sinal amostrado poderá assumir. Quanto maior for esse número, mais parecido o sinal digital será do sinal analógico original. Esses níveis são normalmente representados por uma cadeia de bits. Para sinais de voz, uma resolução de 16 bits (equivalente a 65536 níveis) por amostra é apropriada.

2.5 Extração de informação útil

A avaliação direta do sinal digital, que normalmente possui grande quantidade de dados, além de requerer tempo e processamento consideráveis, provavelmente não trará resultados significativos. Certamente muitos dos pontos avaliados apresentarão informação redundante ou não conterão qualquer informação que possa ser utilizada. O classificador utilizado para avaliar todos os pontos fornecidos dificilmente distinguirá entre amostras de locutores diferentes. Desta forma, uma redução no volume de dados a serem avaliados deve ser realizada. É necessário eliminar informações redundantes ou que não se relacionem diretamente com o locutor. Fornecendo-se apenas um conjunto pequeno mas consistente de elementos para um classificador, viabiliza-se uma classificação robusta e confiável.

O desempenho de todo um sistema para RAL está diretamente relacionado à qualidade da informação que é fornecida. Qualquer classificador, por melhor que seja, apresentará resultados pouco significativos se os parâmetros utilizados para treinamento ou reconhecimento não contiverem informações relevantes. Assim, fica evidente a importância da etapa de extração e seleção da informação para a etapa de classificação.

Algumas características são altamente desejáveis nos padrões que derivam de um sinal vocal, quando se objetiva a aplicação de técnicas para RAL. Dentre outros fatores, a seleção dos parâmetros a serem utilizados deve levar em consideração [ATA 76]:

- **Representatividade**: é interessante que parâmetros utilizados em sistemas de RAL representem, de forma específica, o locutor que os produziu, e variem significativamente para outros locutores. Assim, deseja-se que a variação intra-locutor seja pequena, mas que haja uma grande variação inter-locutor.
- **Fácil medição**: um parâmetro facilmente mensurável é desejado, de forma a inserir baixa probabilidade de erro de medidas e menor complexidade de aquisição.
- **Estabilidade**: parâmetros que apresentam resultados absurdos para algumas situações têm seu uso limitado. É interessante que haja uma estabilidade e que os valores extraídos dos sinais sejam permanentemente coerentes.
- **Ocorrência natural e freqüente**: é interessante que, a partir de qualquer amostra de voz, seja possível avaliar parâmetros representativos. Assim, a ocorrência dos parâmetros avaliados deve ser freqüente.
- **Pouca variação com ambiente**: a independência das condições de aquisição do sinal é um fator desejável em qualquer sistema que utilize amostras de voz. Desta forma, modificações nas condições de aquisição da voz pouco interfeririam no desempenho geral do sistema.
- **Não suscetível à mímica**: através da imitação da voz de pessoas é possível reproduzir de forma razoavelmente semelhante algumas características de

> outro locutor. Entretanto, muitos parâmetros permanecem invariantes, mesmo que o som produzido, para nossos ouvidos, seja muito parecido. A utilização desses parâmetros que pouco variam, mesmo quando uma imitação é realizada, permite a construção de sistemas imunes à mímica.

A construção de métodos para obtenção de parâmetros representativos normalmente utiliza uma modelagem matemática para a produção do sinal de voz. Dessa forma, normalmente o sinal é pré-enfatizado, e a análise em tempos curtos, mostrada adiante, é realizada. Parâmetros que se relacionam com a avaliação do sinal ao longo de alguns segundos são também úteis nos sistemas de RAL.

2.5.1 Modelo da produção da fala

Com o estudo do mecanismo de produção da fala é possível identificar parâmetros, associados ao locutor, que possam ser utilizados nos sistemas para RAL. Os sons vocais podem ser classificados diferentemente, de acordo com o modo de excitação utilizado para sua produção, o local de geração e a configuração do trato vocal. Por exemplo, os sons *vozeados* são produzidos quando o ar é forçado através das cordas vocais tensionadas de forma a vibrarem, produzindo pulsos de ar quasi-periódicos, que excitam o trato vocal. A freqüência de oscilação desses pulsos é chamada freqüência fundamental, ou *pitch*, e está diretamente relacionada à tensão imposta nas cordas vocais. Já os sons *fricativos* são gerados através da formação de uma contração em algum ponto do trato vocal, onde a passagem de ar é forçada de forma a gerar turbulência. Essa turbulência é modelada através de uma fonte de ruído branco que excita o trato vocal. Outro exemplo são os sons *plosivos*, que são gerados através da obstrução completa do fluxo de ar através do trato vocal, e posterior liberação abrupta. Esse tipo de excitação produz um som com bastante intensidade, uma vez que é estabelecida uma rápida variação de pressão nas paredes do trato vocal, devido à obstrução e rápida liberação. Muitos sons produzidos pelo aparelho vocal são, na verdade, uma combinação de tipos distintos.

O processo de produção da voz é modelado analiticamente como mostra a figura 2.3 [RAB 78] [DEL 87]. Há um chaveamento entre as fontes de excitação periódica e ruído branco, que produzem os sons vozeados e fricativos, respectivamente. Os ganhos para vozeados e fricativos correspondem à amplitude do sinal gerado. Esse modelo foi desenvolvido através do estudo da produção do sinal de voz, e representa uma simplificação desse processo.

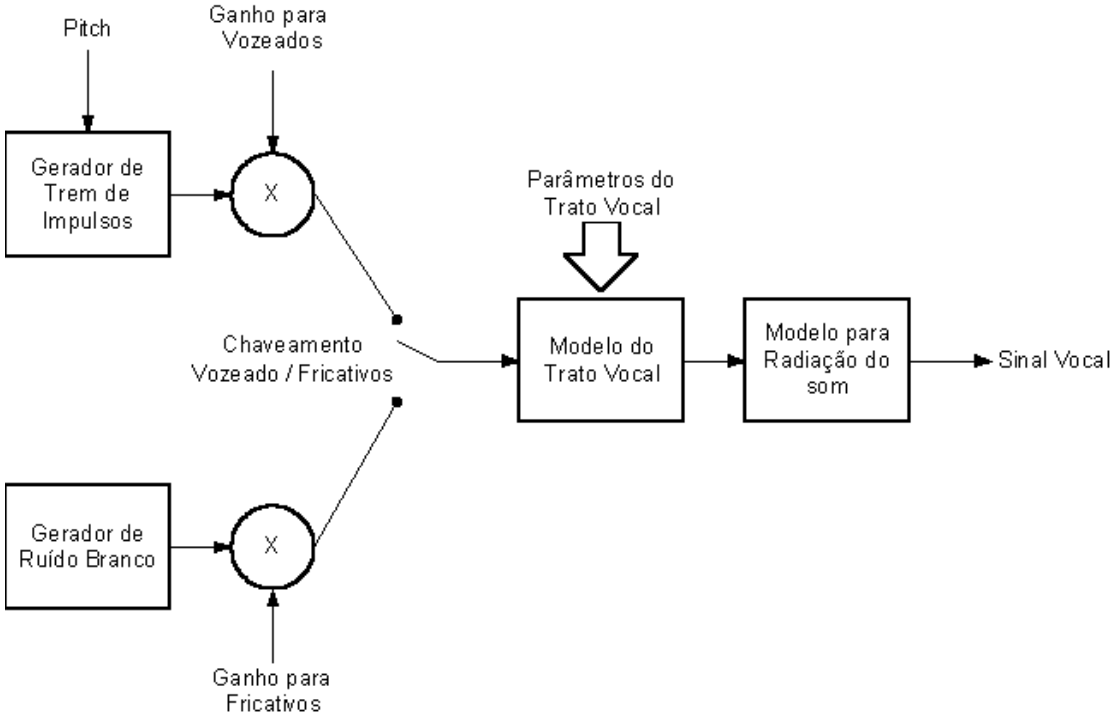

FIGURA 2.3 - Modelo matemático simplificado para a produção de sinais vocais

2.5.2 Pré-ênfase do sinal de voz

Ao sinal de voz digitalizado normalmente é aplicado um filtro de pré-ênfase objetivando nivelar seu espectro [PIC 93] [PET 99d] [PET 99e] [PET 2000a]. Os sinais sonoros apresentam uma atenuação espectral de aproximadamente 6dB/oitava, decorrentes de características fisiológicas do trato vocal. Podemos eliminar essa atenuação espectral, através da aplicação de filtro FIR de primeira ordem, de acordo com a função de transferência mostrada na equação 2.1.

$$H(z) = 1 - az^{-1} \tag{2.1}$$

onde ***a*** é o coeficiente de pré-ênfase, escolhido tipicamente entre 0,4 e 1.

A função de transferência mostrada na equação 2.1 pode ser aplicada ao sinal amostrado através da operação mostrada na equação 2.2.

$$y(n) = x(n) - ax(n-1) \tag{2.2}$$

onde *x(n)* é o sinal de voz amostrado e *y(n)* é o sinal pré-enfatizado.

A figura 2.4 a seguir mostra o espectro de freqüências antes e após a aplicação do filtro de pré-ênfase (a = 0,95) em um sinal de voz. Como pode ser observado, houve um nivelamento no espectro do sinal.

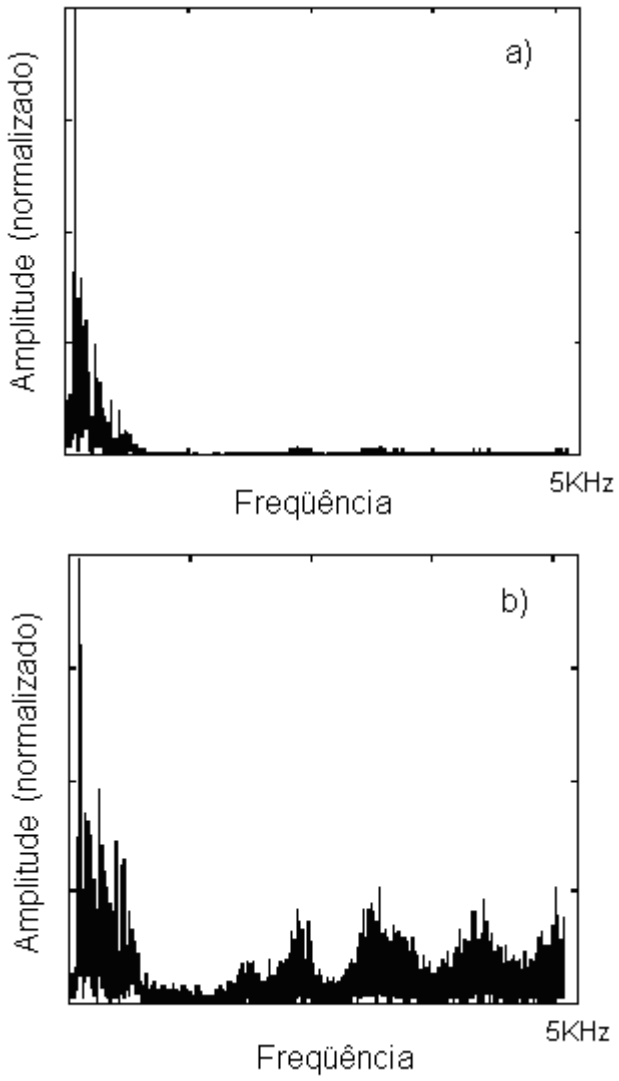

FIGURA 2.4 - Espectro de freqüências para um sinal de voz a) sem pré-ênfase e b) com pré-ênfase

2.5.3 Análise para tempos curtos

Pode-se considerar que os parâmetros do trato vocal variam lentamente. Assim, para intervalos de tempo de aproximadamente 30ms assume-se que o modelo do trato vocal não varia [DEL 87] [PIC 93]. Desta forma, o sistema é considerado aproximadamente estacionário quando analisado localmente nesses intervalos. A forma de limitar a faixa temporal sob análise em um sinal se dá através da aplicação de uma janela. Muitos tipos diferentes de janelas podem ser utilizados. A tabela 2.1 mostra os tipos mais conhecidos de janelas e suas respectivas equações matemáticas, onde *N* é o número de pontos da janela e *n* o índice avaliado. A figura 2.5 ilustra o formato que essas janelas assumem.

TABELA 2.1 - Equação matemática para tipos mais conhecidos de janelas

Janela	Equação Matemática
Retangular	$\begin{cases} 1 & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$
Bartlett	$\begin{cases} 1 - \left\| \dfrac{n - \frac{1}{2}N}{\frac{1}{2}N} \right\| & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$
Blackman	$\begin{cases} 0{,}42 - 0{,}5\cos\left(\dfrac{2\pi n}{N-1}\right) + 0{,}08\cos\left(\dfrac{4\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$
Hamming	$\begin{cases} 0{,}54 - 0{,}46\cos\left(\dfrac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$
Hanning	$\begin{cases} 0{,}5 - 0{,}5\cos\left(\dfrac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$
Welch	$\begin{cases} 1 - \left(\dfrac{n - \frac{1}{2}N}{\frac{1}{2}N} \right)^2 & 0 \le n \le N-1 \\ 0 & n > N-1 \end{cases}$

Normalmente, a janela de *Hamming* é utilizada nos sistemas de RAL, por apresentar características espectrais interessantes [DEL 87]. As sucessivas janelas usualmente possuem uma região de sobreposição, para que a variação dos parâmetros entre janelas adjacentes seja mais gradual, e para que a avaliação dos elementos localizados nos extremos de alguma janela não seja prejudicada.

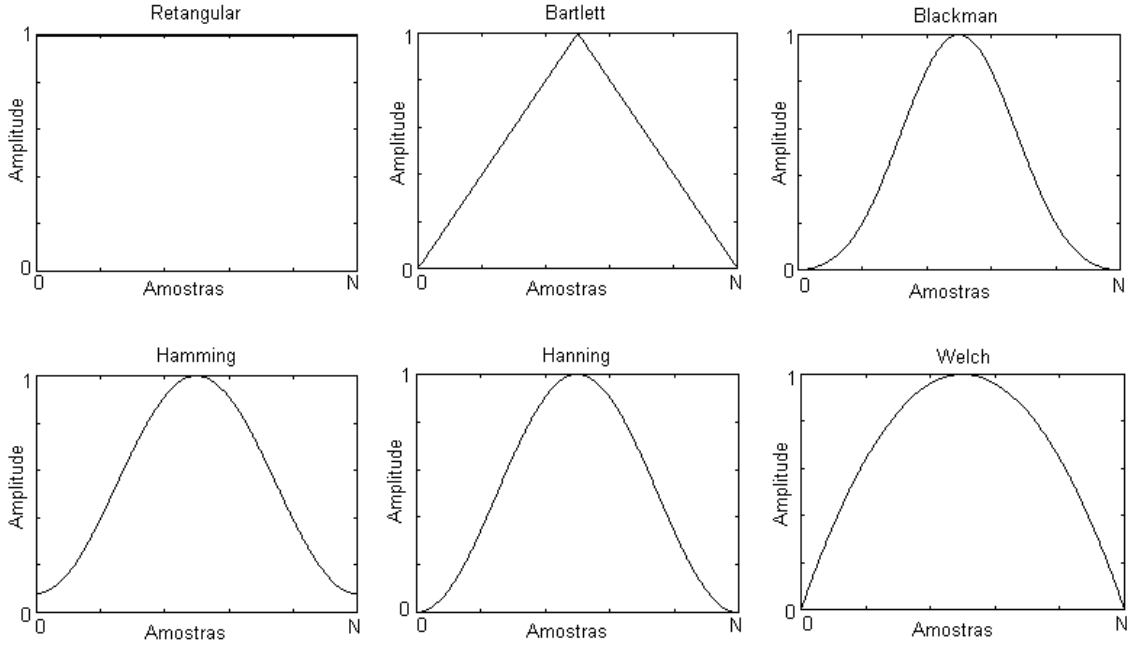

FIGURA 2.5 - Formato dos tipos mais conhecidos de janelas

2.5.4 Extração de parâmetros

As amostras de voz selecionadas através da aplicação de uma janela são utilizadas para a extração de parâmetros representativos, que são posteriormente utilizados. Tais parâmetros representam o sinal de voz por um período de tempo chamado de quadro (em inglês, "*frame*"). Alguns desses parâmetros são mais representativos do locutor que pronunciou a amostra de voz sob análise. Outros carregam informação a respeito do texto falado. Alguns dos principais parâmetros que são utilizados são:

- **Banco de filtros**

Basicamente é a aplicação de filtros passa-banda deslocados em freqüência [PIC 93] [RAB 78] [RAB 93] [DEL 87]. Os parâmetros utilizados são a energia na saída de cada filtro;

- **Energia**

A amplitude do sinal de voz em um quadro pode ser facilmente medida, como mostra a equação 2.3 [PIC 93] [DEL 87]. A variação de tal medida ao longo do tempo pode fornecer informações importantes a respeito da geração da amostra de voz sob análise.

$$E(n) = \frac{1}{N}\sum_{m=0}^{N-1}\left[w(m)y\left(n - \frac{N}{2} + m \right)\right]^2 \qquad (2.3)$$

onde *E(n)* é a energia do sinal, *w(m)* é a janela de *N* pontos aplicada ao sinal pré-enfatizado *y(n)*, sendo *n* sendo o índice de amostragem (tempo discreto) do centro da janela.

A medida de energia do sinal vocal pode também ser utilizada para a determinação dos limites das palavras. Neste caso, é estabelecido um valor limiar abaixo do qual o sinal é classificado como ruído de fundo (*ground noise*), e acima é classificado como voz.

- **Taxa de Cruzamento por Zero**

É definida como o número de vezes que o sinal cruzou o nível médio, por vezes chamado de nível zero. A taxa de cruzamento por zero pode ser utilizada para determinação automática dos limites de palavras. Sendo *y(n)* a *n*-ésima amostra do sinal de voz pré-enfatizado, formalmente pode-se definir a taxa de cruzamento por zero (ZCR) [DEL 87] [RAB 78] como mostra a equação 2.4.

$$ZCR = \frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{2}\left|sign(y(n)) - sign(y(n-1))\right| \tag{2.4}$$

onde N é o número de amostras do quadro de voz avaliado e a função *sign(y(n))= 1* se *y(n)* for positivo, e *sign(y(n))= -1* se *y(n)* for negativo.

- **Coeficientes Cepstrais**

Tendo-se em mente o modelo de produção do sinal de voz, pode-se definir que o sinal de voz é o resultado da operação de convolução (⊗) do sinal de excitação *u(t)* com a resposta impulsiva do trato vocal *h(t)*. A operação de convolução transforma-se numa multiplicação, quando se troca o domínio tempo pelo domínio freqüência [PRO 96]. Aplicando-se a função logarítmica a dois sinais multiplicados, é possível transformar tal multiplicação em sobreposição (soma) desses sinais. Como a maior parte da energia espectral do sinal de excitação *u(t)* e resposta impulsiva do trato vocal *h(t)* ocupam bandas espectrais diferentes, é possível utilizar as informações de apenas um deles. Sabe-se que o trato vocal varia mais lentamente que a excitação. Logo, aplicando-se a transformada inversa e utilizando os primeiros coeficientes, estarão sendo avaliadas as informações a respeito da configuração do trato vocal, para o quadro analisado. A figura 2.6 ilustra o método utilizado para desconvoluir os sinais.

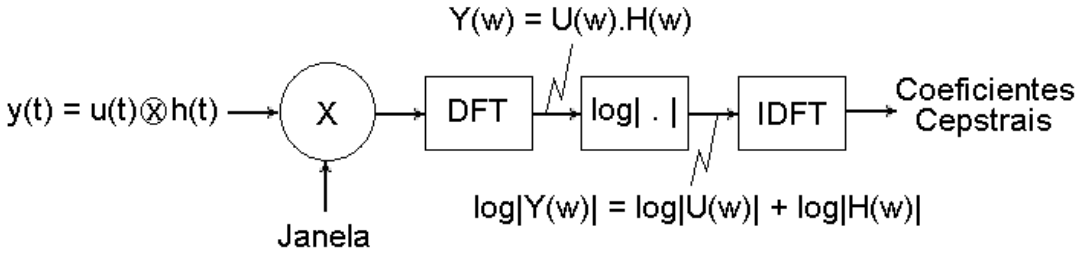

FIGURA 2.6 - Processo de obtenção do coeficientes cepstrais

Assim, os coeficientes cepstrais *c(n)* [DEL 87] [RAB 78] são definidos como a transformada de Fourier inversa (IDFT) do logaritmo do espectro de um sinal, como mostra a equação 2.5.

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1}\log\left|Y(k)\right|e^{j2\pi kn/N} \tag{2.5}$$

onde *N* é o número de pontos da janela utilizada, *Y(k)* é a transformada de Fourier do sinal pré-enfatizado (e janelado) *y(n)*.

- **Coeficientes *mel*-cepstrais**

Mel é a unidade de freqüências ou picos percebidos de um som. Tal unidade não corresponde linearmente à freqüência física, bem como o ouvido humano também não o faz. A escala *mel*, mostrada na equação 2.6, reflete a forma com que seres humanos percebem as freqüências e foi obtida através de experiências subjetivas com voluntários.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.6}$$

A diferença básica entre a obtenção dos coeficientes cepstrais e dos *mel*-cepstrais é a aplicação de um banco de filtros digitais triangulares tipo passa-banda espaçados segundo a escala *mel*, como pode ser visto na figura 2.7. Muitas vezes, a escala *mel* é aproximada a uma escala linear até freqüência de 1KHz, e logarítmica acima de 1KHz. A tabela 2.2, obtida de [PIC 93] fornece a freqüência central e largura de banda dos primeiros 20 filtros triangulares que podem ser aplicados ao espectro do sinal, fornecendo os coeficientes *mel*-cepstrais [DEL 87].

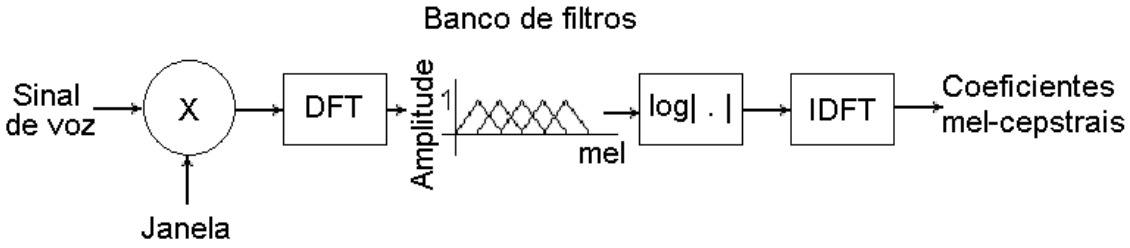

FIGURA 2.7 - Processo de obtenção dos coeficientes *mel*-cepstrais

Os coeficientes *mel*-cepstrais têm sido utilizados extensivamente em sistemas de RAF. Alguns sistemas de RAL utilizam os coeficientes *mel*-cepstrais, e seus resultados são similares aos obtidos com coeficientes cepstrais. Muitas vezes, a etapa de aplicação da IDFT do processo de obtenção dos coeficientes *mel*-cepstrais é substituída pela transformada inversa do cosseno (ICT), a fim de concentrar a informação nos coeficientes de mais baixa ordem. Esse procedimento será adotado neste trabalho.

- **Coeficientes de Predição Linear**

Dada uma amostra de voz *y(n)* no instante *n*, é possível aproximá-la como uma combinação linear das *p* amostras passadas, como mostra a equação 2.7.

$$y(n) \approx a_1 y(n-1) + a_2 y(n-2) + ... + a_p y(n-p) \tag{2.7}$$

onde a_1, a_2, ..., a_p são os coeficientes de predição linear (LPC) [RAB 78] [RAB 93] [DEL 87], assumidos como constantes para o quadro sob análise.

O método de cálculo dos LPC inicia-se com o cálculo dos coeficientes de autocorrelação de cada quadro, como mostra a equação 2.8.

$$r(m) = \sum_{n=0}^{N-1-m} y(n) y(n+m) \tag{2.8}$$

onde $m=0,1,...,p$, p é a ordem da análise LPC, $r(m)$ são os coeficientes de autocorrelação, N o número de amostras em uma janela, e $y(n)$ é o sinal pré-enfatizado e janelado.

TABELA 2.2 - Freqüência central e largura de banda de filtros triangulares espaçados segundo escala *mel*

Índice do Filtro	Freqüência Central (Hz)	Largura de Banda (Hz)
1	100	100
2	200	100
3	300	100
4	400	100
5	500	100
6	600	100
7	700	100
8	800	100
9	900	100
10	1000	124
11	1149	160
12	1320	184
13	1516	211
14	1741	242
15	2000	278
16	2297	320
17	2639	367
18	3031	422
19	3482	484
20	4000	556

O método formal para conversão dos coeficientes de autocorrelação em LPC é conhecido como método de Durbin [RAB 93] [RAB 78] [DEL 87] [PIC 93], mostrados nas equações 2.9 a 2.13.

$$E^{(0)} = r(0) \tag{2.9}$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(i-j)}{E^{(i-1)}} \qquad 1 \le i \le p \tag{2.10}$$

$$\alpha_i^{(i)} = k_i \tag{2.11}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \tag{2.12}$$

$$E^{(i)} = (1 - k_i^{\,2}) E^{(i-1)} \tag{2.13}$$

As equações 2.9 a 2.13 são resolvidas recursivamente para *i=1,2,...p* e os LPC (a_m) são obtidos como mostra a equação 2.14.

$$a_m = \alpha_m^{(p)} \qquad 1 \le m \le p \tag{2.14}$$

- **Freqüência Fundamental**

A freqüência fundamental, por vezes chamada de *pitch*, representa a principal freqüência com que as cordas vocais vibram. Seu valor está diretamente associado com a morfologia da fonte de excitação do trato vocal, e é um parâmetro bastante representativo de um locutor. Mulheres ou crianças, cujas cordas vocais tem comprimento pequeno, tendem a apresentar um valor para a freqüência de *pitch* maior que os homens.

Algumas técnicas para detecção do *pitch* a partir de um quadro vozeado de um sinal de voz são disponíveis. Os coeficientes de predição linear (LPC) podem ser utilizados no algoritmo denominado Rastreamento Simples do Filtro Inverso (SIFT) [DEL 87] [RAB 78]. Algumas propriedades dos coeficientes cepstrais também podem ser utilizadas para estimar o valor da freqüência fundamental [RAB 78]. Nesse caso, os coeficientes cepstrais de mais alta ordem são utilizados, pois contêm a informação a respeito da excitação do trato vocal. Na prática, procura-se um pico na vizinhança do período de *pitch* esperado (entre 3 e 20ms – equivalente a 333Hz e 50Hz). Se a amplitude do pico estiver acima de um limite pré-estabelecido, sua posição é uma boa aproximação para o período de *pitch*. A figura 2.8 mostra os coeficientes cepstrais obtidos a partir de um quadro vozeado de um sinal de voz. Claramente, o período de *pitch* é salientado nos coeficientes de mais alta ordem. Como a freqüência de amostragem utilizada foi 11025Hz, o período de *pitch* mostrado equivale a aproximadamente 9,3ms, equivalente à freqüência de *pitch* de 107Hz.

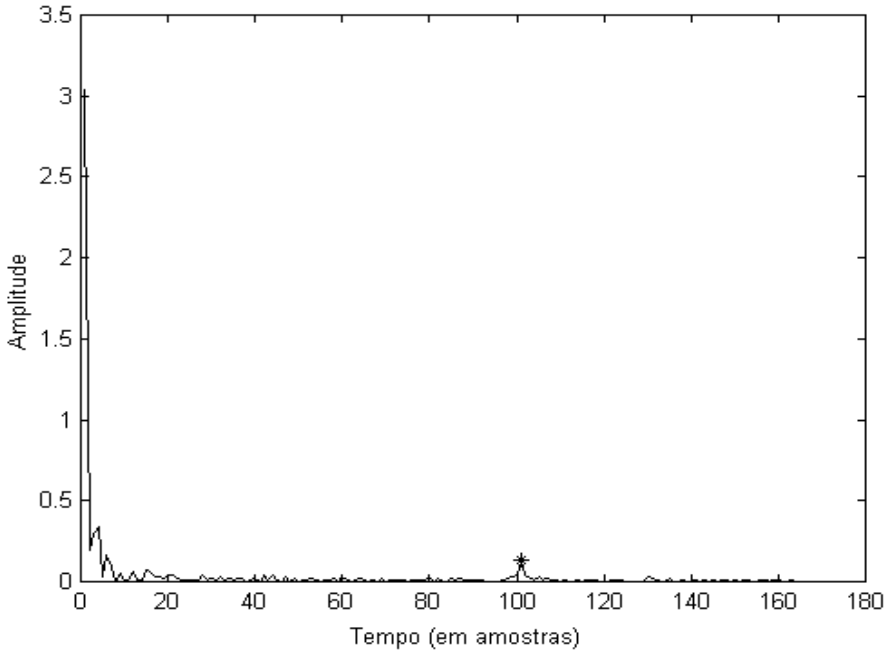

FIGURA 2.8 - Coeficientes cepstrais de um quadro vozeado de um sinal de voz e identificação do *pitch*

- **Freqüência das Formantes**

Basicamente, o caminho vocal humano é composto por uma concatenação de tubos, com variadas áreas seccionais. A função de transferência de energia, a partir da fonte de excitação até a saída, pode ser descrita em termos de freqüências naturais ou ressonâncias desse conjunto de tubos. Tais ressonâncias são chamadas formantes e são as freqüências onde há maior transmissão de energia. Tipicamente, as três primeiras ressonâncias são as de maior importância.

Os métodos disponíveis para cálculo das freqüências das formantes basicamente obtém uma estimativa do espectro "suavizado" do quadro vozeado sob análise, e identificam os picos desse espectro. As localizações de tais picos representam as freqüências das formantes. Para isso, pode-se utilizar a transformada de Fourier dos primeiros coeficientes cepstrais [RAB 78], ou o inverso da transformada de Fourier dos primeiros coeficientes de predição linear (LPC) [DEL 87]. A figura 2.9 mostra em a) o espectro normalizado de um quadro vozeado de um sinal de voz, b) o espectro suavizado e normalizado obtido através da utilização de 10 LPC e c) o espectro suavizado e normalizado obtido através da utilização de 10 coeficientes cepstrais. Para as reconstruções mostradas, os valores de freqüências de formantes obtidos foram aproximadamente 330Hz, 2.4KHz e 3.5KHz.

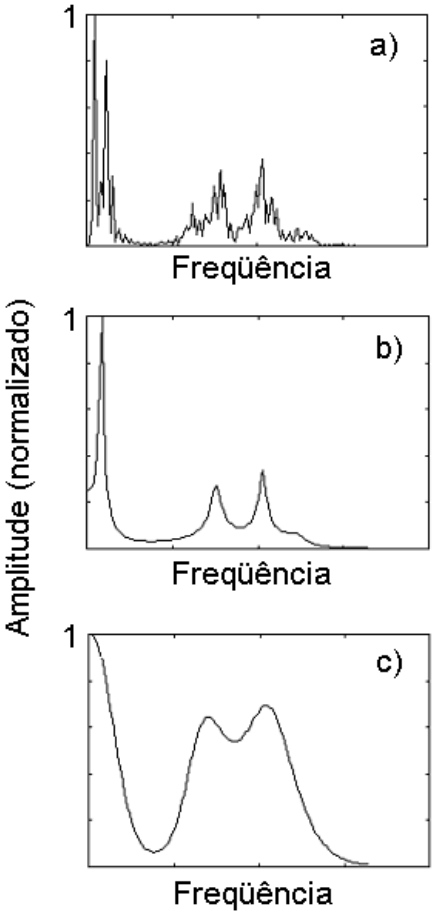

FIGURA 2.9 - Para um quadro vozeado: a) espectro normalizado b) espectro suavizado através da utilização de 10 LPC c) espectro suavizado através de utilização de 10 coeficientes cepstrais

2.5.5 Parâmetros diferenciais

A utilização de derivadas temporais aplicadas aos parâmetros extraídos geralmente são associadas aos próprios parâmetros [DEL 87], a fim de caracterizar sua variação ao longo dos quadros. Derivadas de primeira e segunda ordens são normalmente utilizadas. A equação 2.15 mostra o cálculo dos parâmetros diferenciais de primeira ordem. Os parâmetros diferenciais de segunda ordem podem ser obtidos através da aplicação da equação 2.15 nos parâmetros diferenciais de primeira ordem.

$$\frac{\partial}{\partial n} C_i(n) = \mu\left[C_i(n+K) - C_i(n-K)\right] \tag{2.15}$$

onde $C_i(n)$ é o parâmetro de ordem i do n-ésimo quadro, μ é uma constante de contribuição e K é a variação temporal considerada.

2.5.6 Subtração da média

A extração de parâmetros representativos de um locutor é realizada avaliando-se um quadro do sinal, ou seja, em intervalos de tempo suficientemente curtos, no qual o

sistema é considerado estacionário. É possível também avaliar ao longo de vários quadros do sinal o valor médio dos parâmetros extraídos. Esses parâmetros médios são usualmente subtraídos dos parâmetros de cada quadro [ROS 94b] [WES 97] [MAR 77], eliminando com isso distorções que tendem a elevar, na média, seus valores. Esse tipo de distorção é normalmente produto do canal de transmissão do sinal de voz, e essa técnica é por vezes chamada subtração da média cepstral (CMS). Dessa forma, os parâmetros utilizados nas etapas seguintes podem ser "normalizados", de forma a fornecer informações mais confiáveis. A equação 2.16 descreve a aplicação dessa técnica.

$$C_i^{novo}(n) = C_i(n) - \frac{1}{N}\sum_{j=1}^{N} C_i(j) \tag{2.16}$$

onde $C_i(n)$ é o parâmetro de ordem i do n-ésimo quadro e N o número total de quadros da amostra de voz avaliada.

2.6 Modelamento

Em um sistema de RAL, após a extração de parâmetros do sinal de voz, eles devem ser comparados com os padrões previamente armazenados, de forma a quantificar a semelhança entre o trecho de voz a ser reconhecido e os locutores cadastrados no sistema. Uma comparação adequada resulta no melhor aproveitamento da informação contida nesses parâmetros. Os sistemas de RAL podem utilizar técnicas distintas. Vários modelos estatísticos diferentes foram utilizados em sistemas de RAL e são descritos na literatura, dentre os quais destacam-se [VUU 99]:

- Quantização Vetorial [FIN 97] [MAT 94] [PET 99a] [PET 99b] [PET 2000a] [SOO 87] [MOR 2001];
- Redes Neurais Artificiais [ADA 97] [FAR 94] [LIP 87] [CAS 97] [YEG 2001];
- Medidas de distância de Bhattacharyya e Divergência [CAM 97] [KAI 67] [PET 2000b] [PET 2001] [VER 99];
- Modelo de Mistura de Gaussianas (GMM) [REY 95] [REY 92] [VUU 99] [MIY 2001];
- Modelos Ocultos de Markov (HMM) [MAT 94] [ROS 90].

Ao longo deste trabalho, a medida de distância de Bhattacharyya será extensivamente utilizada para avaliação dos parâmetros dinâmicos não-lineares. As principais motivações para essa escolha abrangem o forte embasamento estatístico que essa análise utiliza, simplicidade de compreensão do funcionamento geral do sistema, ausência de parâmetros de sintonia e rapidez de execução dos algoritmos envolvidos. Os parâmetros de sintonia referem-se às modificações que podem ser feitas no sistema a fim de alcançar maior exatidão. Como exemplos desses parâmetros podemos mencionar: o número de vetores utilizados, técnica para sua divisão e medida de proximidade entre vetores, em quantização vetorial; a arquitetura adotada e função de ativação, em redes neurais; o número de Gaussianas que representam um locutor, na modelagem por mistura de Gaussianas; as possibilidades de transição entre estados, nos modelos ocultos de Markov, entre outros. As vantagens mencionadas não prejudicam a exatidão do sistema, uma vez que as taxas de acerto obtidas com essa técnica são comparáveis às de outras técnicas.

O modelamento através de Gaussianas, e posterior análise através da distância de Bhattacharyya, compõem um método relativamente fácil de ser compreendido e que não possui ajustes internos, como configurações complexas ou valores empiricamente estabelecidos. Essa característica beneficia em muito a análise que será conduzida nesse trabalho, uma vez que o interesse concentra-se na extração de parâmetros representativos, que agreguem novos tipos de informações. Assim, a influência dos fatores associados ao modelamento utilizado é reduzida, permitindo que as mudanças em taxas de exatidão do sistema reflitam exclusivamente as modificações nos parâmetros que são utilizados para caracterizar os locutores.

O grau de exatidão absoluto do sistema não apresenta informações conclusivas a respeito da utilidade do acréscimo de parâmetros dinâmicos não-lineares. O principal objeto de estudo é, na verdade, o desempenho do sistema relativamente a uma configuração que utiliza os principais parâmetros explorados na literatura. Assim, é possível verificar a variação da confiabilidade do sistema, robustez a ruído, tempo de execução e outras informações, quando são acrescidos parâmetros dinâmicos não-lineares. Uma análise de significância estatística nos resultados poderá indicar uma genuína melhora na caracterização de um locutor, que deverá se refletir também em outros métodos de modelagem.

2.6.1 Distância de Bhattacharyya

Em estatística, o grau de proximidade entre duas funções densidade de probabilidade diferentes está relacionado com a noção de medida de distorção. Uma forma de calcular essa medida é através da distância de Bhattacharyya [KAI 67].

Considere duas funções densidade de probabilidade $p_1(x)$ e $p_2(x)$, obtidas a partir de duas classes diferentes de parâmetros. A distância de Bhattacharyya é definida por [BHA 43]

$$B = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)\, p_2(x)}\, dx \tag{2.17}$$

Casos especiais dessa medida de distorção genérica podem ser calculados explicitamente para vários tipos de funções densidade de probabilidade. Um caso importante refere-se a distribuições Gaussianas multivariadas. Considerando $p_i(x)$ como funções densidade de probabilidade Gaussianas, é possível demonstrar [KAI 67] que a equação 2.17 pode ser escrita como mostra a equação 2.18,

$$B = \frac{1}{8}(m_1 - m_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (m_1 - m_2) + \frac{1}{2} \ln \left(\frac{\det(\Sigma_1 + \Sigma_2)/2}{\sqrt{\det(\Sigma_1)}\sqrt{\det(\Sigma_2)}} \right), \tag{2.18}$$

onde m_i é o vetor média, e Σ_i é a matriz covariância obtidos a partir dos vetores da classe i.

A distância de Bhattacharyya pode ser aplicada em uma grande variedade de distribuições de probabilidade conhecidas, de acordo com o melhor modelamento estabelecido às densidades de probabilidade. A utilização de funções densidade de probabilidade Gaussianas para os parâmetros não é arbitrária, uma vez que é suficiente que a densidade seja essencialmente unimodal e aproximadamente Gaussiana no centro de sua excursão. Essas propriedades são freqüentemente respeitadas em sistemas físicos. Através da inspeção de histogramas obtidos a partir de parâmetros reais avaliados de

sinais de voz, é possível verificar que suas distribuições podem ser modeladas como funções densidade de probabilidade Gaussianas.

Apenas como ilustração, a figura 2.10 mostra quatro histogramas obtidos a partir da mesma dimensão dos coeficientes cepstrais, extraídos de dois locutores diferentes que falam a mesma palavra. Primeiramente, é possível verificar uma forte semelhança com a distribuição Gaussiana. Além disso, os histogramas de um mesmo locutor apresentam uma proximidade maior que os do outro locutor, indicando o provável sucesso de um classificador baseado na distância de Bhattacharyya.

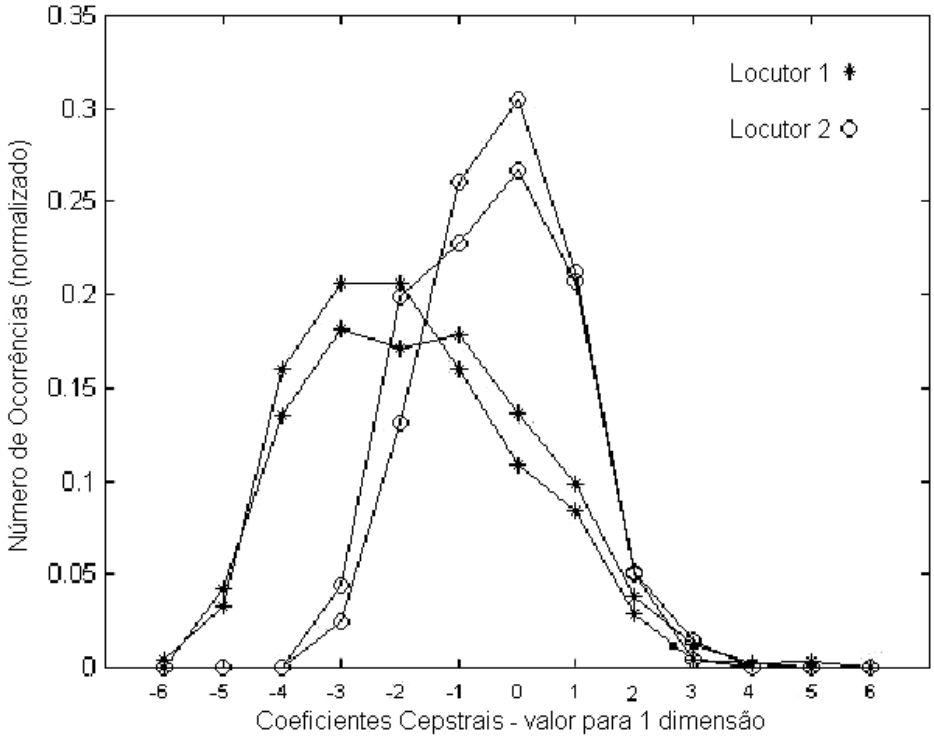

FIGURA 2.10 - Histogramas para uma dimensão dos coeficientes cepstrais extraídos a partir de amostras de voz de locutores diferentes

Um sistema de reconhecimento de locutor baseado na distância de Bhattacharyya funciona através de comparações entre densidades de probabilidades, tanto na etapa de treinamento quanto no reconhecimento. Assim, os parâmetros obtidos a partir de alguns sinais de voz de cada pessoa cadastrada no sistema são utilizados para compor os padrões vocais, nesse caso representados pela densidade Gaussiana, que é completamente descrita por seu vetor média e matriz covariância. Na etapa de reconhecimento, os parâmetros obtidos a partir do sinal de voz a ser reconhecido constituirão a densidade Gaussiana que será utilizada na comparação, através da equação 2.18.

2.7 Resumo

Este capítulo abordou aspectos técnicos relacionados aos sistemas de RAL conhecidos. Inicialmente, a perturbação no ar gerada pela produção do som vocal é convertida para o domínio digital, utilizando técnicas para conversão A/D. A seguir, as informações que caracterizam o locutor devem ser extraídas das amostras digitais de voz. O sinal é primeiramente pré-enfatizado, e após é analisado em janelas de tempo de

curta duração onde os parâmetros podem ser extraídos levando-se em conta o modelo matemático de produção da voz. Para os sistemas de RAL que realizam a verificação da identidade, ou os sistemas de identificação em grupo aberto, a geração de um limiar associado a cada locutor cadastrado deve ser feita. Ao final, um classificador é utilizado para avaliar a similaridade entre a amostra de voz desconhecida e os padrões dos locutores cadastrados. Avaliando-se a medida que o classificador apresenta é possível aferir a respeito da identidade do locutor que produziu o sinal de voz analisado.

3 Teoria do Caos e medidas de invariantes dinâmicas

3.1 Introdução

Este capítulo pretende fornecer uma fundamentação teórica para a aplicação e adaptação das técnicas disponibilizadas pela Teoria do Caos em sinais de voz. Os conceitos desenvolvidos darão maior ênfase para o tratamento de sinais discretos oriundos da observação de sistemas físicos, como é o caso do sinal de voz. Primeiramente uma breve introdução é apresentada, de forma a situar o foco do estudo utilizado. A seguir, aspectos relativos às séries temporais são mostrados. Após, é realizada uma análise comparativa entre sistemas lineares e sistemas dinâmicos não-lineares, enfocando sua utilização em séries temporais. Em seguida o método utilizado para reconstrução da dinâmica do sistema a partir de amostras unidimensionais é mostrado. Essa reconstrução é necessária para a estimativa de informações qualitativas do sistema. Assim, as principais técnicas empregadas para extração de algumas invariantes dinâmicas são apresentadas, e serão amplamente empregadas no capítulo seguinte.

3.2 Teoria do Caos

Muitos fenômenos naturais apresentam um comportamento complexo e dinâmico, o que dificulta sua análise. O escoamento não laminar da água, por exemplo, apresentaria grande dificuldade para determinação exata da posição de cada partícula ao longo do tempo, apesar de que elas seguem leis físicas relativamente simples e conhecidas. A influência de fatores aparentemente desprezíveis é capaz de causar grandes mudanças na evolução da posição e velocidade das partículas. Esse padrão de aparente imprevisibilidade é um sinal claro de sistemas caóticos. O conjunto de ferramentas analíticas, numéricas e geométricas apropriadas para a análise de problemas não-lineares para o qual não existem soluções gerais explícitas constitui a Teoria do Caos. Por causa de sua generalidade, a Teoria do Caos ou Teoria dos Sistemas Dinâmicos Não-Lineares pode ser utilizada para analisar uma grande variedade de problemas, mesmo quando eles são não-caóticos.

A Teoria do Caos permite estudar a complexidade de sistemas, servindo como um dos paradigmas mais poderosos e gerais conhecidos. Os sistemas caóticos necessariamente apresentam não-linearidades de forma a serem sensíveis às condições iniciais. Isto é, mesmo conhecendo-se todas as leis que determinam a evolução do sistema, qualquer modificação nas condições que o sistema se encontra no início da análise, por menor que seja, implicará em um comportamento completamente distinto do comportamento analiticamente esperado. O caos que em geral se estuda considera os sistemas de um ponto de vista determinístico, ou seja, a análise realizada não assume que o comportamento do sistema seja aleatório. Ao contrário, o sistema apresenta um comportamento que, mesmo sem que se possa determinar de forma exata e precisa a sua evolução, o objeto de estudo está na determinação de seu comportamento, através da avaliação de características qualitativas dessa evolução.

Caos pode ocorrer em dois tipos diferentes de sistemas:

- **sistemas conservativos**: onde a energia do sistema é constante, ou seja, não há perdas;

- **sistemas dissipativos**: onde poderá existir uma variação na energia do sistema devido a fatores dissipativos;

O sistemas enfocados nesse trabalho são os dissipativos, que são facilmente encontrados na natureza do dia-a-dia. A produção dos sinais vocais também tem caráter intrinsecamente dissipativo por estar sujeita a interações diversas com o meio.

3.3 Caos em série temporais

Dados podem ser obtidos e mostrados de formas diferentes. Se os dados estiverem ordenados como uma seqüência no tempo, eles são chamados de série temporal. Esse tipo de medida discreta será fortemente utilizado ao longo deste trabalho, uma vez que as amostras capturadas a partir da digitalização de sinais de voz constituem por si só séries temporais. As séries temporais podem ser analisadas de várias formas. Quando se possui uma série temporal e uma análise dinâmica não-linear é aplicada, normalmente deseja-se determinar que tipo de sistema dinâmico a produziu. Os dados obtidos a partir de uma série temporal são normalmente digitalizados, através de uma amostragem com taxa fixa. A equação 3.1 mostra matematicamente como é feita a amostragem de um sinal analógico $x_a(t)$, mantendo-se constante o período T de amostragem.

$$x(t_n) = x_a(t_n T) \qquad -\infty < t_n < \infty \tag{3.1}$$

para t_n assumindo apenas valores inteiros.

O valor do período T de amostragem deverá ser tal que respeite o critério de Nyquist (ou teorema da amostragem) [PRO 96] [RAB 78], ou seja, a freqüência de amostragem ($1/T$) deverá ser maior que o dobro da maior componente em freqüência existente no sinal analógico $x_a(t)$. Os dados digitais obtidos a partir da série temporal analógica podem ser então processados por computadores.

As tarefas que são impostas aos analistas a partir de sistemas lineares e não-lineares são muito parecidas, mas os métodos de análise são substancialmente diferentes. As principais similaridades e diferenças podem ser vistas na tabela 3.1, adaptada de [ABA 93].

TABELA 3.1 - Comparação entre processamento digital de sinais lineares e não-lineares

Objetivo	Processamento de sinais lineares	Processamento de sinais dinâmicos não-lineares
Determinação do Sinal	Redução de Ruído Separar ruído do sinal, através da análise espectral das características do sinal	Redução de Ruído Utilização da dinâmica ou distribuições invariantes, e probabilidade de transições Markovianas
Determinação do Espaço de Atuação	Transformada de Fourier	Reconstrução do Espaço de Fases
Classificação do Sinal	Identificação dos picos espectrais – freqüências de ressonância do sistema	Medidas de Invariantes orbitais, como expoentes de Lyapunov, dimensões e entropia

3.4 Classificação dos atratores

O termo “atrator” deriva da observação de um sistema no espaço de fases. Esse sistema tenderá a se desenvolver assumindo o estado representado por seu atrator. Os sistemas dinâmicos são atraídos para os atratores. O termo “espaço de fases” é utilizado para referenciar a região *p*-dimensional onde o sistema sob análise evolui. Projetando-se esse comportamento do sistema para uma única dimensão, tem-se a série temporal, obtida a partir de dados experimentais, como ilustra a figura 3.1. Desta forma, o que será mostrado ao longo desse capítulo é como, a partir da projeção unidimensional da trajetória do sistema, pode-se estimar a evolução *p*-dimensinal que gerou a série temporal dada. Para avaliar as propriedades de um possível atrator associado a uma série temporal de dados experimentais com comportamento caótico determinístico, é necessário primeiramente reconstruí-lo num espaço de fases de dimensão apropriada.

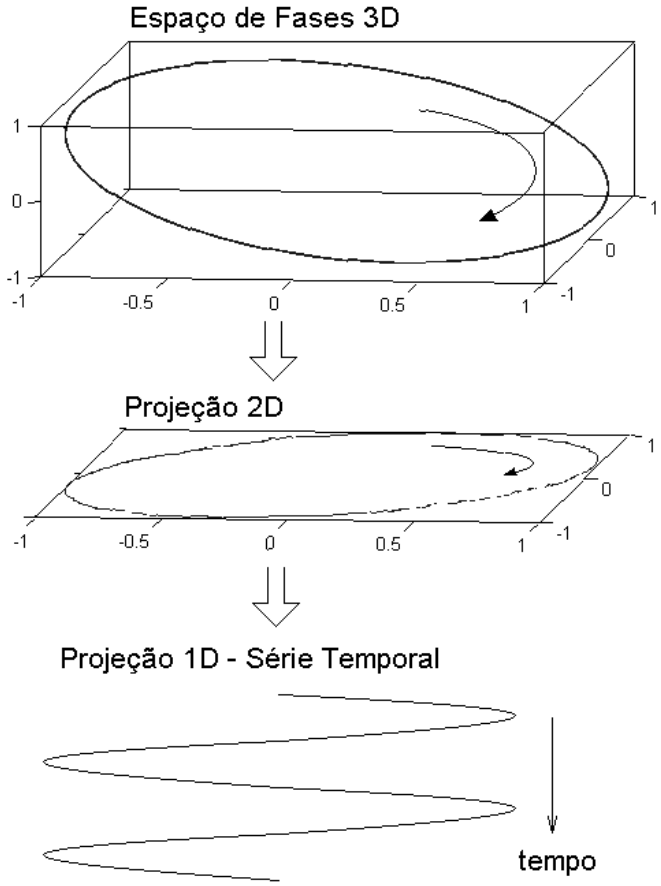

FIGURA 3.1 - Projeção unidimensional de um sistema dinâmico que evolui em um espaço de fases tri-dimensional

Um marco importante foi alcançado quando, a partir de uma série temporal experimental, algoritmos possibilitaram reconstruir a trajetória no espaço de fases do sistema. A validação dessa reconstrução foi provada pelo teorema de Takens [TAK 81], também chamado “método dos atrasos temporais”. O atrator reconstruído através desse método não é idêntico ao original, mas pode-se demonstrar que as propriedades topológicas são preservadas. Vários algoritmos foram então introduzidos para a estimação de medidas de invariantes dinâmicas não-lineares, como dimensão fractal,

dimensão de correlação e expoentes de Lyapunov, para um atrator reconstruído através desse método. Essas invariantes dinâmicas não-lineares serão estudadas mais profundamente adiante.

Os atratores podem apresentar formas e complexidades diversas. Os quatro tipos diferentes de atratores que são observados em sistemas dinâmicos são [ÇAM 93]:

- **Atrator de ponto fixo**: são atratores que atingem um ponto de equilíbrio. Esse tipo de atrator não é sensível às condições iniciais e não leva o sistema a um estado caótico.
- **Atrator de ciclos-limites**: atratores que possuem um ciclo de atuação, onde o sistema deverá permanecer. Mesmo aplicando-se uma energia externa que force o sistema a se afastar do ciclo de atividade original, ele tenderá a voltar.
- **Atrator toroidal**: atratores encontrados em sistemas que possuem muitos graus de liberdade. Atratores toroidais conservam energia, e sua imagem no espaço de fases é semelhante a um “novelo de lã”.
- **Atrator estranho**: são atratores altamente irregulares, possuindo propriedades geométricas complicadas. São encontrados associados a sistemas dinâmicos dissipativos não periódicos.

3.5 Cuidados anteriores à reconstrução do atrator

Antes de passarmos à reconstrução propriamente dita do atrator associado a uma série temporal, alguns cuidados devem ser tomados, de forma a garantir que a trajetória reconstruída corresponda fielmente à dinâmica do sistema. O primeiro dos cuidados a ser tomado diz respeito à freqüência de amostragem adequada do sinal analógico proveniente do sistema dinâmico. A questão da degradação do sinal amostrado pela inserção de ruído, e formas de contornar esse problema, também são apontados. Outro fator importante abrange a questão da estacionariedade do sistema analisado. Por fim, os métodos disponíveis para reconstrução de atratores e para a estimativa de invariantes dinâmicas têm necessidade de uma quantidade mínima de amostras disponíveis para oferecerem estimativas que reflitam de forma apropriada as características do sistema.

3.5.1 Freqüência de amostragem

A análise de sinais realizada pelos computadores digitais requer que o sinal físico analógico seja digitalizado. Independentemente da natureza do sinal que se está analisando, ele deverá ser amostrado a uma freqüência maior ou igual ao dobro da máxima componente em freqüência existente nesse sinal. Essa condição, conhecida como teorema da amostragem ou critério de Nyquist [RAB 78] [PRO 96], é aplicada também na análise de sinais provenientes de sistemas dinâmicos não-lineares. Para o caso específico dos sinais vocais, o estudo dos mecanismos de aquisição são abordados no item 2.4 deste trabalho.

3.5.2 Eliminação do ruído

Pode-se descrever o processo de aquisição das amostras constituintes de uma série temporal como tendo algumas etapas:

- produção do sinal pela fonte, que pode apresentar distúrbios em sua propagação;

- propagação do sinal através de um canal de comunicação, potencialmente contaminado pelo ambiente;
- medição do sinal no receptor, o qual pode apresentar perturbações na recepção e agir como um filtro aplicado ao sinal transmitido.

No caso de sinais de banda curta, inseridos em um ambiente que apresenta componentes de freqüência espalhadas pelo espectro (banda larga), a diferença é clara e a análise de Fourier é a indicada para separar o sinal de interesse. Analogamente, se o sinal e a contaminação estão localizados em bandas espectrais diferentes, as técnicas de Fourier ainda são certamente indicadas. Já no caso de sinais obtidos de fontes caóticas, tanto o sinal de interesse, quanto as interferências podem se apresentar como banda larga, e a análise de Fourier já não ajudará muito na separação do sinal de interesse [ABA 93].

Para remover interferência ou ruído estocástico de um sinal caótico observado, que tipicamente não podem ser separados no domínio freqüência, existem três abordagens diferentes [ABA 93]:

- **conhecendo-se a dinâmica**: se as equações que descrevem a evolução do sinal são conhecidas, pode-se determinar uma função para avaliar o grau de contaminação do sinal observado. Se apenas as propriedades dinâmicas são desejadas, as observações contaminadas são descartadas, e tais propriedades são extraídas a partir das equações conhecidas;
- **tendo-se observado algum sinal sem ruído**: quando algum sinal "limpo" (sem ruído) obtido do sistema caótico sob análise estiver disponível. Nesse caso, qualquer questão geral ou estatística sobre o sistema será melhor respondida utilizando-se esse sinal sem contaminação, ao invés de se tentar extrair informações menos úteis de um sinal ruidoso. O método apropriado aqui consiste em uma redução de ruído probabilística do sinal contaminado, baseada em ajustes, de forma semelhante à técnica padrão de máxima probabilidade *a posteriori*. Maiores detalhes de implementação podem ser encontrados em [ABA 93];
- **não sabendo qualquer informação**: quando a dinâmica do sistema não é conhecida, e não se tem acesso a qualquer sinal livre de ruído. Esse caso é bastante comum em sistemas que utilizam séries temporais de dados experimentais. Neste caso, duas estratégias podem ser adotadas:

1- Construção de mapas polinomiais locais: utiliza-se informações da vizinhança dos pontos, e então se ajusta cada ponto de acordo com o mapa construído com um conjunto de pontos [ABA 93] [KOS 91]. Isso é feito em dois passos distintos. O primeiro considera o movimento de um conjunto de pontos na vizinhança de cada ponto do atrator, de forma a computar uma aproximação linear para sua dinâmica. No segundo passo, é avaliado quão bem uma trajetória individual obedece tal aproximação.

2- Utilização de filtros lineares: apenas filtros do tipo FIR ou de médias são permitidos, de forma a não alterar a estrutura do espaço de fases da dinâmica [ABA 93] [SCH 91].

3.5.3 Estacionariedade

As amostras, que posteriormente serão utilizadas para a reconstrução da trajetória do sistema dinâmico no espaço de fases, deverão ser provenientes de um estado estacionário desse sistema.

Para o caso de sinais de voz, sabe-se que a forma do trato vocal varia lentamente. Assim, para janelas temporais de aproximadamente 30-40 ms, os

parâmetros do trato vocal podem ser considerados aproximadamente estacionários [RAB 78] [DEL 87].

3.5.4 Número de amostras

O número de amostras disponível deverá ser tal que o atrator seja visitado várias vezes, de modo a representá-lo de forma estatisticamente confiável. Na prática, bons resultados já são alcançados em séries experimentais compostas por aproximadamente 1000 pontos [FER 94]. Um limite superior típico para o número de pontos tratável pela maior parte dos algoritmos disponíveis para avaliação das propriedades dinâmicas é por volta de 20000, isto a um custo computacional já relativamente elevado [FER 94].

3.6 O Atrator: sua trajetória no espaço de fases

A partir de uma série temporal de dados experimentais discretos $x(t_i)$ para $i=1,2,...,N$, como mostra a equação 3.2.

$$x(t) = \{x(t_1), x(t_2),..., x(t_N)\} \tag{3.2}$$

são reconstruídos vetores *m*-dimensionais através da equação 3.3.

$$\vec{\xi}_i = \{x(t_i), x(t_i + p), x(t_i + 2p),..., x(t_i + (m-1)p)\} \tag{3.3}$$

onde *p* é o chamado passo de reconstrução (*time delay*) e *m* é a chamada dimensão de imersão (*embedding dimension*).

Os vetores ξ_i reconstruídos representam a trajetória do sinal temporal $x(t_i)$ no espaço de fases *m*-dimensional. Esse método de reconstrução é chamado "método dos atrasos temporais", e sua validação já foi provada [TAK 81]. Takens demonstrou que a trajetória do sinal temporal no espaço de fases não é idêntica à trajetória real que gerou a série temporal, mas as características topológicas do atrator reconstruído são preservadas. Dessa forma, a medida de invariantes dinâmicas, a partir da reconstrução proposta, produz resultados verdadeiros.

A figura 3.2 a) mostra o conhecido atrator de Lorenz, em um espaço de fases bi-dimensional. Esse atrator foi obtido utilizando-se as três equações diferenciais de Lorenz, mostradas na equação 3.4, com parâmetros r=45.92, b=4.0 e σ=16.0, que fornecem um atrator caótico.

$$\begin{aligned} \frac{dx(t)}{dt} &= \sigma(y(t) - x(t)), \\ \frac{dy(t)}{dt} &= -x(t)z(t) + rx(t) - y(t), \\ \frac{dz(t)}{dt} &= x(t)y(t) - bz(t). \end{aligned} \tag{3.4}$$

Se utilizarmos apenas o sinal $x(t)$ desse atrator e reconstruirmos sua trajetória em um espaço de fases bi-dimensional através do método dos atrasos temporais, obtemos a trajetória da figura 3.2 b). Apesar de não serem idênticos, ambos atratores da figura 3.2 são topologicamente equivalentes.

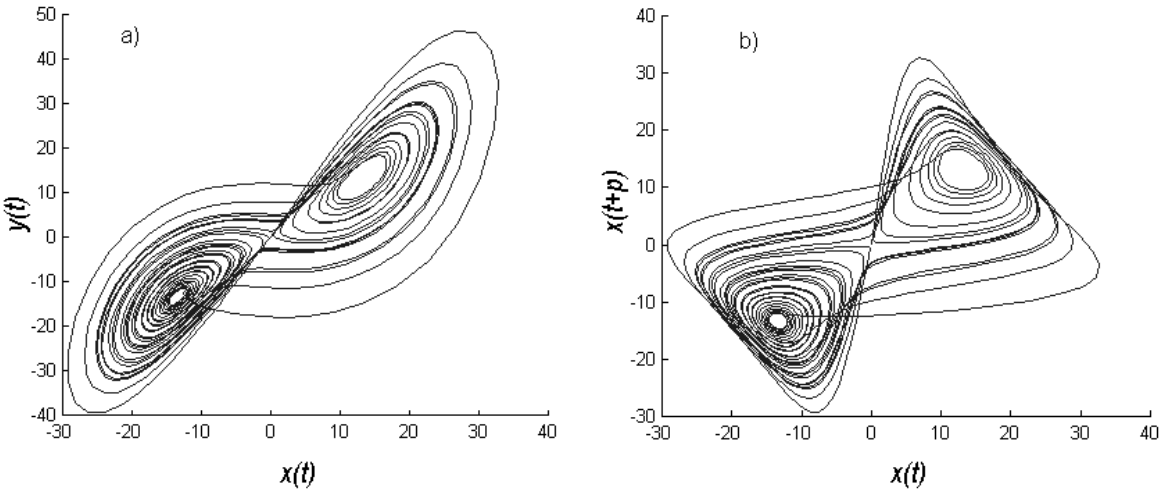

FIGURA 3.2 - Atrator de Lorenz em um espaço de fases tri-dimensional: a) original e b) reconstruído a partir da série temporal *x(t)*.

A identificações completa dos vetores ξ_i reconstruídos só se dá quando forem calculados os valores adequados para o passo *p* de reconstrução e para a dimensão de imersão *m*. Os métodos para o cálculo de tais valores são mostrados a seguir.

3.6.1 Passo de reconstrução

A escolha do passo *p* adequado para a reconstrução dos vetores ξ_i deve ser realizada com cuidado. Um valor de *p* excessivamente pequeno faz com que os valores de *x(t)* e *x(t+p)* sejam muito parecidos e, conseqüentemente, os vetores ξ_i e ξ_{i+1} também o serão. Na prática a escolha do passo *p* muito pequeno se apresenta na forma de uma trajetória alongada no espaço de fases, devido à grande correlação existente entre as componentes dos vetores ξ_i. Por outro lado, um valor do passo de reconstrução *p* excessivamente grande faz com que os vetores ξ_i e ξ_{i+1} sejam pouco correlacionados. Desta forma, a reconstrução da trajetória no espaço de fases apresenta-se dispersa. Como ilustração, a figura 3.3 apresenta a reconstrução em duas dimensões da série temporal obtida a partir das equações de Lorenz (3.4) com parâmetros r=45.92, b=4.0 e σ=16.0, utilizando um passo de reconstrução: a) excessivamente pequeno, b) adequado e c) excessivamente grande.

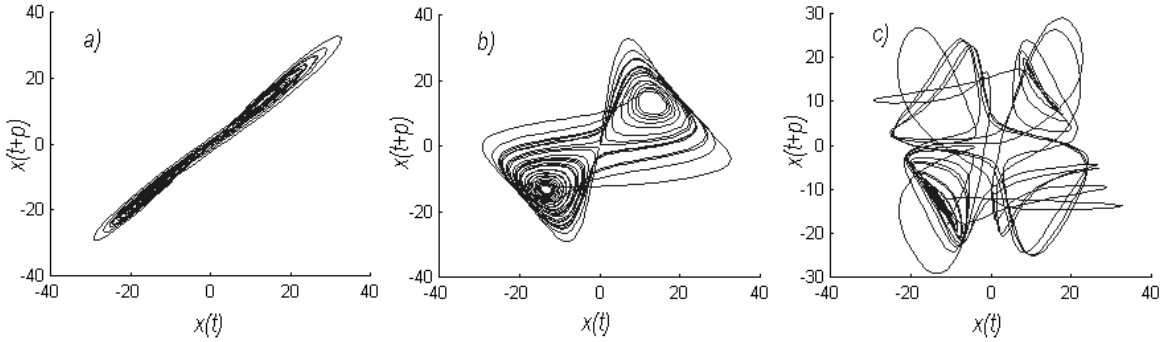

FIGURA 3.3 - Reconstrução da trajetória em duas dimensões do atrator de Lorenz, utilizando um passo de reconstrução: a) excessivamente pequeno, b) adequado e c) excessivamente grande

A visualização bidimensional da trajetória da série temporal no espaço de fases é também conhecida como diagrama de primeiro retorno. Através de uma inspeção visual, é possível identificar se os vetores estão ou não altamente correlacionados, ou seja, se o passo *p* escolhido é ou não adequado. Entretanto, análise visuais são muito sujeitas a interpretações subjetivas. Assim, métodos foram desenvolvidos para encontrar analiticamente o valor apropriado para o passo de reconstrução.

Um dos métodos bastante difundidos para a escolha do passo *p* de reconstrução sugere que se utilize um valor próximo ao tempo de correlação do sinal. Assim, calcula-se a função linear de autocorrelação [ABA 93], como mostra a equação 3.5,

$$C(\tau) = \frac{\frac{1}{N}\sum_{m=1}^{N}[x(m+\tau)-\bar{x}][x(m)-\bar{x}]}{\frac{1}{N}\sum_{m=1}^{N}[x(m)-\bar{x}]^2} \tag{3.5}$$

onde $\bar{x}$ pode ser calculado como mostra a equação 3.6, e utiliza-se o valor de tempo τ onde $C(\tau)$ cai para uma certa fração do valor inicial, por exemplo $1/e$ [ROS 94].

$$\bar{x} = \frac{1}{N}\sum_{m=1}^{N} x(m) \tag{3.6}$$

Dessa forma, a estimação do passo de reconstrução poderá ser feita através da análise da função de autocorrelação do sinal. Outros métodos de escolha analítica do passo de reconstrução podem ser encontrados em [ROS 94]. De qualquer forma, uma inspeção visual através do diagrama de primeiro retorno poderá sempre ser realizada, de forma a garantir a escolha correta do passo de reconstrução, sem contudo oferecer uma avaliação precisa.

3.6.2 Dimensão de imersão

A dimensão de imersão *m* do espaço de fases reconstruído não precisa ser a mesma do espaço de fases real. Em geral, é necessário apenas utilizar uma dimensão suficientemente elevada, como mostra a equação 3.7.

$$m \geq 2D_0 + 1 \tag{3.7}$$

onde D_0 é a dimensão de *Hausdorff* do atrator, cujo cálculo é mostrado adiante.

Na prática, dimensões inferiores ao valor definido na equação 3.7 muitas vezes são suficientes para estabelecer uma reconstrução de trajetórias no espaço de fases tal que apresentem invariantes dinâmicas adequadas.

A escolha da dimensão de imersão *m* elevada causa a diminuição do número de vetores ξ_i utilizados para a reconstrução, entretanto para sinais com um número elevado de amostras isso não será muito prejudicial. Entretanto, os algoritmos disponíveis que medem invariantes dinâmicas tornam-se significativamente mais complicados, e alguns até mesmo inviáveis, quando aplicados num espaço de fases de dimensão elevada.

Um dos métodos para estimar a dimensão de imersão mínima necessária à reconstrução do atrator é conhecido como processo de *Karhunen-Loève* ou Decomposição por Valor Singular (SVD) [KUM 96] [ABA 93]. O método SVD foi um dos primeiros propostos para reconstruir os vetores ξ_i a partir de séries temporais de dados experimentais. Uma desvantagem desse método é que ele não consegue distinguir duas séries temporais que possuem a mesma estrutura de covariância mas apresentam diferenças em estruturas de ordem superior. A eficácia desse método é alcançada através da redução do ruído. Se o nível de ruído é conhecido, pode-se descartar o subespaço da matriz de trajetórias que corresponde aos valores singulares abaixo do limiar de ruído [KUM 96].

Outro método que vem sendo utilizado com sucesso é o chamado critério dos "Falsos Vizinhos" [KEN 92]. Esse método fundamenta-se na idéia de que, em dimensões de imersão demasiadamente pequenas, nem todos os pontos que parecem estar próximos são realmente vizinhos devido à dinâmica do sistema. Alguns estarão, na verdade, bem longe e apenas parecem vizinhos porque a estrutura geométrica do atrator foi projetada em um espaço dimensionalmente inferior ao real. Dessa forma, quando a dimensão for aumentada, a distância entre "falsos vizinhos" deverá crescer bastante. Algoritmicamente, esse método inicia assumindo a dimensão de imersão *m* igual a um. São calculados então os vetores ξ_i para essa dimensão. A seguir, para cada um dos vetores ξ_i, é verificado quem é o seu vizinho mais próximo, de acordo com a distância euclidiana. A dimensão de imersão *m* é então incrementada e são verificados quantos dos vetores "vizinhos" eram, na verdade, "falsos vizinhos". A distância entre dois vetores considerados "falsos vizinhos" deverá aumentar muito quando se incrementar a dimensão *m*. A condição estabelecida na equação 3.8 mostra o método de avaliação de "vizinhança" entre dois vetores.

$$\left[\frac{D_{m+1}^2 - D_m^2}{D_m^2}\right]^{1/2} > L_c \tag{3.8}$$

onde L_c é um valor de distância crítica (normalmente escolhido um valor superior a dez), D_m^2 é o quadrado da distância entre os vetores para a dimensão *m*, e D_{m+1}^2 é o quadrado da distância entre os vetores para a dimensão *m*+1.

Em aplicações práticas, o número de pontos disponíveis, geralmente, não é muito elevado. Isso pode acarretar a escolha de dimensões inferiores às desejadas. Outro critério foi então estabelecido a fim de lidar com um conjunto limitado de dados. Seja D_A o tamanho do atrator. Se o vizinho mais próximo a um vetor qualquer está muito longe, ou seja, $D_m \approx D_A$, e ele é um falso vizinho, então a distância D_{m+1} deverá ser $D_{m+1} \approx 2.D_A$. Esse segundo critério, que é utilizado juntamente com o primeiro, pode ser escrito como mostra a equação 3.9.

$$\frac{D_{m+1}}{D_A} > A \tag{3.9}$$

onde A é uma constante (normalmente igual a 2,0), e D_A uma medida do tamanho do atrator, que pode ser calculada como mostra a equação 3.10.

$$D_A^2 = \frac{1}{N}\sum_{n=1}^{N}\left[x(n) - \bar{x}\right]^2 \tag{3.10}$$

onde N é o número de pontos da série temporal de dados experimentais discretos $x(t)$, e $\bar{x}$ é o valor médio dos pontos $x(t)$. Outras escolhas para a estimação de D_A, como o valor absoluto da variância de $x(t)$, podem ser utilizadas.

Assim, o vizinho mais próximo é considerado falso se obedecer a qualquer uma das duas condições anteriores (equação 3.8 ou 3.9). O processo é repetido, sempre se aumentando a dimensão utilizada, até que o número de "falsos vizinhos" encontrados esteja suficientemente próximo a zero. A figura 3.4 ilustra a variação do percentual de falsos vizinhos com relação à dimensão m, para a série temporal obtida a partir das equações do atrator de Lorenz (3.4). Foi utilizado um passo de reconstrução p igual a 25, obtido a partir da análise do tempo de correlação do sinal, distância crítica L_c igual a 10 e a constante A igual a 2.0 (valores normalmente utilizados).

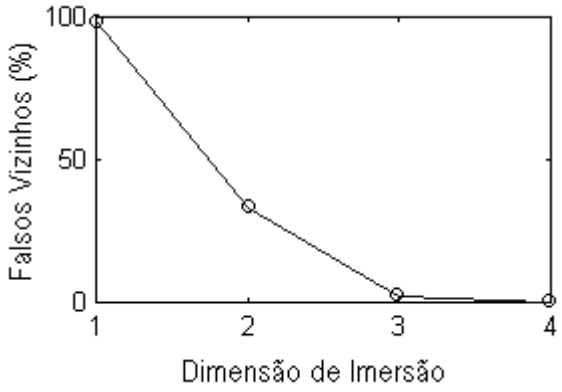

FIGURA 3.4 - Variação do percentual de "falsos vizinhos" para o atrator de Lorenz

Vários métodos são disponíveis para a determinação da dimensão de imersão m apropriada, como a inspeção visual da trajetória reconstruída, a análise SVD, e critério dos "falsos vizinhos". É possível também avaliar os valores da dimensão fractal, cujos métodos de estimação são mostrados adiante, avaliando-se o sistema em dimensões de imersão m variadas. Conforme a dimensão de imersão m for incrementada, o valor da dimensão fractal calculada tenderá a um valor adequado, indicando assim o valor apropriado da dimensão de imersão.

3.7 Invariantes Dinâmicas

A impossibilidade de predizer por que ponto do espaço de fases a trajetória de um atrator estranho passará num certo instante de tempo é uma característica intrínseca de todos os sistemas caóticos. No entanto, ainda podemos fazer previsões muito precisas, mas elas se referem às características qualitativas do comportamento do sistema, e não aos valores exatos de seu estado num determinado instante. A análise de sistemas dinâmicos não lineares, em termos das características topológicas de seus

atratores, é conhecida como “análise qualitativa”. Através dessa análise é possível estimar as invariantes dinâmicas não lineares [ABA 93], como dimensão, que fornece uma medida dos graus de liberdade de um sistema e expoentes de Lyapunov, que é uma medida da estabilidade local de uma trajetória no espaço de fases.

Esta seção apresenta técnicas que podem ser utilizadas para a estimativa de algumas informações topológicas que podem ser obtidas através da análise do comportamento do atrator associado a um sistema dinâmico não-linear.

3.7.1 Dimensão fractal

A medida da dimensão de um atrator fornece informações topológicas a respeito da estrutura que esse atrator apresenta. Para sistemas com vários graus de liberdade, a trajetória no espaço de fases apresenta-se altamente irregular. A avaliação estrutural de um atrator pode ser realizada através da subdivisão do espaço de fases em $N(\varepsilon)$ hiper-cubos de lado ε, estudando-se a distribuição de probabilidades p_i ao longo do atrator, quando ε tende a zero. A distribuição de probabilidades p_i é definida como a probabilidade de se ter um ponto da trajetória no i-ésimo hiper-cubo, como mostra a equação 3.11.

$$p_i = \lim_{N\to\infty} \frac{N_i}{N} \tag{3.11}$$

onde N é o número total de pontos da trajetória e N_i é o número de pontos contidos no i-ésimo hiper-cubo de lado ε.

Assim, são definidas as dimensões generalizadas como mostra a equação 3.12

$$D_q = \frac{1}{q-1} \lim_{\varepsilon\to 0} \frac{\log \sum_{i=1}^{N(\varepsilon)} p_i^q}{\log \varepsilon} \tag{3.12}$$

onde q é um número real que identifica as infinitas dimensões que podem ser calculadas.

Considerando-se os casos onde q=0 e q=2, é possível demonstrar-se [FER 94] [ABA 93] que esses valores fornecem respectivamente medidas das chamadas dimensão fractal (D_0) e dimensão de correlação (D_2).

A diferença entre as infinitas dimensões D_q está nas regiões de interesse analisadas. Quando q aumenta, as regiões mais densas do atrator são realçadas. Analogamente, à medida que q diminui, são destacadas regiões mais rarefeitas.

A dimensão fractal [BAD 84] [TER 83] [FER 94], ou dimensão de *Hausdorff*, é obtida analisando-se um conjunto de pontos num espaço de dimensão p. Cobrindo-se todos esses pontos com hiper-cubos de lado ε, define-se a dimensão fractal D_0 desse conjunto de pontos como mostra a equação 3.13, que pode ser facilmente obtida a partir da equação 3.12 para dimensões generalizadas fazendo-se q=0 .

$$D_0 = \lim_{\varepsilon\to 0} \frac{\log N(\varepsilon)}{\log(1/\varepsilon)} \tag{3.13}$$

onde $N(\varepsilon)$ é o número mínimo de hiper-cubos de lado ε necessário para cobrir todo o conjunto de pontos.

Aplicando-se a definição de dimensão fractal pode-se calcular, por exemplo, o valor para uma superfície quadrada, como mostra a figura 3.5, adaptada de [FER 94].

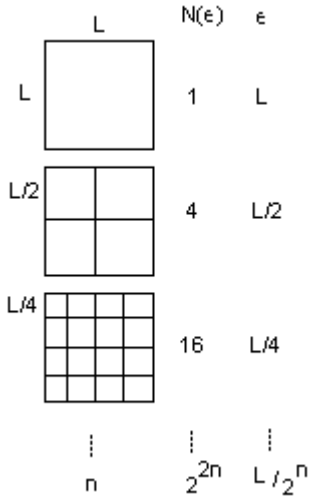

FIGURA 3.5 - Cálculo da dimensão fractal para superfícies quadradas

Neste caso, temos, a partir da equação 3.13, $D_o = \lim_{\varepsilon \to 0} \frac{\log(2^{2n})}{\log(2^n / L)} = 2$.

Quanto mais "recortado" for o elemento sob análise, maior será sua dimensão fractal. Dessa forma, aplicando-se o conceito de dimensão fractal aos atratores, é possível determinar um valor para o grau de irregularidade geométrica desse atrator. O valor da dimensão fractal é independente da freqüência com a qual uma trajetória visita as várias partes do atrator, por ser uma medida puramente geométrica.

A medida de dimensão fractal serve também para avaliar a dimensão de imersão *m*, utilizada para reconstruir o atrator. Se a dimensão de imersão *m* não for grande o suficiente de forma a desenvolver a estrutura do atrator completamente, a projeção do atrator tende a "preencher" o espaço, resultando numa estimativa para a dimensão fractal igual à dimensão de imersão [ABA 93]. Conforme a dimensão de imersão *m* for incrementada, a dimensão fractal calculada "satura" para um valor adequado.

Um dos métodos disponíveis para o cálculo da dimensão fractal dos atratores reconstruídos é chamado de algoritmo de contagem de caixas [ABA 93] [BAD 84]. Nesses algoritmos, a região do espaço de fases ocupada pelo atrator é dividida em caixas de dimensão igual à dimensão de imersão, e lado de valor ε. Contam-se, então, quantas dessas caixas possuem pelo menos um ponto do atrator. O cálculo deve ser repetido para diversos valores de ε. O valor da dimensão fractal pode ser obtido através da inclinação do gráfico $\log N(\varepsilon)$ *versus* $\log(1/\varepsilon)$, onde *N(ε)* é o número mínimo de hiper-cubos de lado ε necessário para cobrir todo o conjunto de pontos. Os programas que utilizam algoritmos de contagem de caixas exigem espaço em memória e tempo de processamento elevados, principalmente quando ε diminui. Desta forma, mesmo utilizando-se algoritmos eficientes, a idéia de contagem de caixas é em geral impraticável para dimensões maiores que dois.

Um dos métodos mais conhecidos e utilizados para o cálculo da dimensão fractal é o método do Expoente Crítico (CEM) [NAK 93] [SAB 96] [CUS 99]. Para séries temporais de dados com componentes auto-similares, a dimensão fractal D_0 pode ser avaliada como mostra a equação 3.14.

$$D_0 = 2 - H \tag{3.14}$$

onde *H* é o expoente de Hurst.

Basta então calcular o expoente de Hurst *H*, e o CEM proposto baseia-se na análise do momento I_α associado ao espectro de potências do sinal, definido de acordo com a equação 3.15

$$I_\alpha = \int_1^U duP(u)u^\alpha \qquad para \quad -\infty < \alpha < +\infty \tag{3.15}$$

onde I_α é o α-ésimo momento, *P(u)* é a densidade de energia espectral modelada pela equação 3.16, *u* é a freqüência normalizada cujo limite inferior é 1 e superior é *U*. Em [NAK 93] é proposto utilizar apenas as freqüências que aparentemente seguem o modelo de energia espectral da equação 3.16. Especificamente para sinais de voz, as componentes de baixa freqüência são desprezadas, fazendo-se $u=k/k_c$, onde *k* é a freqüência real, k_c é a freqüência abaixo da qual o modelo da equação 3.16 não é seguido.

$$P(u) \approx u^{-\beta} \tag{3.16}$$

Aplicando-se o modelo da equação 3.16 na equação 3.15, obtemos a equação 3.17.

$$I_\alpha \approx \int_1^U duu^{\alpha-\beta} = \int_1^U duu^{X-1} = \left[\frac{u^X}{X}\right]_{u=1}^{u=U} = \frac{U^X}{X} - \frac{1}{X} \tag{3.17}$$

onde foi realizada uma substituição de variáveis, fazendo $X=\alpha-\beta+1$.

A equação 3.17 pode ser re-escrita como mostra a equação 3.18.

$$I_\alpha \approx \frac{\left(e^{\ln U}\right)^X - 1}{X} = \frac{\left(e^{\ln U}\right)^X - e^0}{X} = \frac{2}{X} e^{\left(\frac{X \ln U}{2}\right)} \cdot \left(\frac{e^{\left(\frac{X \ln U}{2}\right)} - e^{\left(\frac{-X \ln U}{2}\right)}}{2} \right) \tag{3.18}$$

A partir da equação 3.18 anterior e da definição da função de seno hiperbólico, o momento pode ser escrito como mostra a equação 3.19.

$$I_\alpha \approx \frac{2}{X} e^{\left(\frac{X \ln U}{2}\right)} sinh\left(\frac{X \ln U}{2}\right) \tag{3.19}$$

Avaliando-se o gráfico do logaritmo natural do momento com relação à α, é possível determinar o expoente crítico α_c do momento. Esse valor é determinado através da projeção dos dois regimes de $ln(I_\alpha)$ para $\alpha<<0$ e para $\alpha>>0$. É possível obter esse valor através da análise da derivada segunda com relação à α do logaritmo do momento. Para efeitos práticos, essa expressão pode ser desenvolvida como mostra a equação 3.20 através da utilização direta da densidade de energia espectral *P(u)*,

$$\frac{\partial^2 \ln(I_\alpha)}{\partial \alpha^2} = \frac{I_\alpha''}{I_\alpha} - \left(\frac{I_\alpha'}{I_\alpha}\right)^2 \tag{3.20}$$

onde a *n*-ésima derivada de I_α, I_α^n, pode ser calculada [SAB 96] como mostra a equação 3.21.

$$I_{\alpha}^{n} = \frac{\partial^{n}}{\partial \alpha^{n}} \int_{1}^{U} du u^{\alpha} P(u) = \int_{1}^{U} du (\ln(u))^{n} u^{\alpha} P(u) \tag{3.21}$$

O valor crítico α_c que faz com que a derivada segunda com relação à α do logaritmo do momento seja máxima fornece o valor do expoente β como mostra a equação 3.22.

$$\beta = \alpha_c + 1 = 2H + 1 \tag{3.22}$$

De posse do valor do expoente de Hurst *H*, é possível facilmente estimar o valor da dimensão fractal D_0, através da equação 3.14 anterior.

A figura 3.6 mostra o gráfico do logaritmo do momento com relação à α, obtido a partir de 10000 amostras do sinal *x(t)* do atrator de Lorenz (equações 3.4), onde a intersecção das projeções dos regimes para $\alpha << 0$ e $\alpha >> 0$ indicam o valor do expoente crítico α_c.

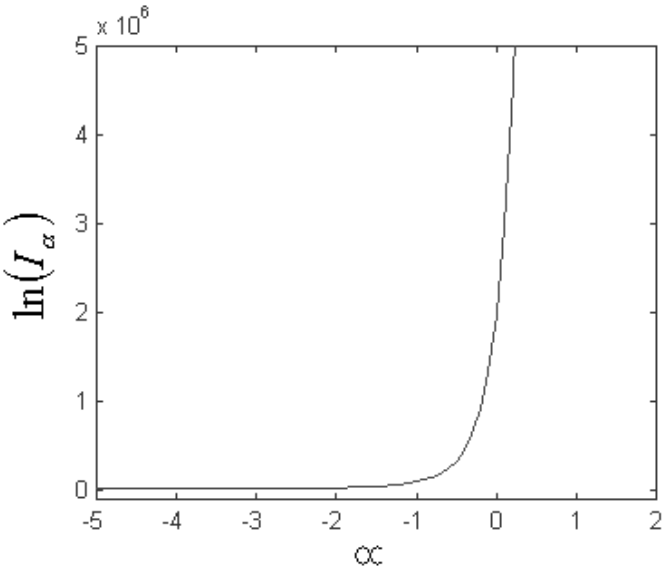

FIGURA 3.6 - Evolução do logaritmo do momento associado ao espectro de potências

A figura 3.7 mostra a variação da derivada segunda do logaritmo do momento com relação à α para o mesmo sinal da figura 3.6. Esse gráfico foi obtido através da integração (soma) mostrada na equação 3.21 utilizando o espectro do sinal. A seguir foi estimado o valor da derivada segunda do logaritmo do momento de acordo com a equação 3.20. O valor crítico α_c é aproximadamente igual a -0,6. Dessa forma, é possível avaliar os demais parâmetros: o expoente β vale 0,4 e o expoente de Hurst *H* vale –0,3. Isso fornece um valor para a dimensão fractal D_0 aproximadamente igual a 2,3.

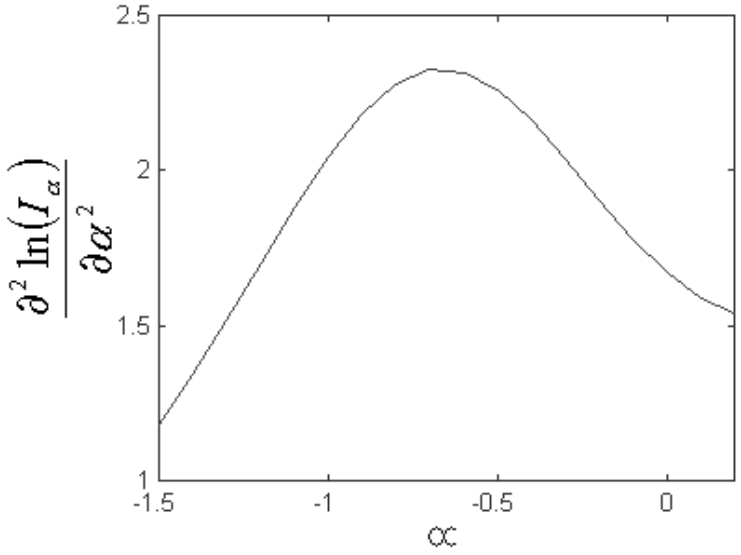

FIGURA 3.7 - Evolução da derivada segunda do logaritmo do momento associado ao espectro de potências

Dados obtidos a partir de sons vocais complicados contêm flutuações na amplitude, freqüência, forma de pronúncia, etc, que variam com o tempo. Assim, se a série temporal for dividida em intervalos curtos, é possível analisar essas flutuações e caracterizar tal série através de valores sucessivos de dimensão fractal. Essa seqüência de valores é chamada Dimensão Fractal Dependente do Tempo (TDFD) [SAB 96].

3.7.2 Dimensão de correlação

O cálculo da dimensão de correlação pode ser obtido fazendo-se q=2 na equação 3.12 para dimensões generalizadas, que fornece a expressão mostrada na equação 3.23

$$D_2 = \lim_{\varepsilon \to 0} \frac{\log \sum_{i=1}^{N(\varepsilon)} p_i^2}{\log \varepsilon} \tag{3.23}$$

O algoritmo de Grassberger-Procaccia [GRA 83a] [KUM 96] para o cálculo da dimensão de correlação aproxima $\sum_{i=1}^{N(e)} p_i^2$ como o somatório de correlação $C(\varepsilon)$, calculado conforme a equação 3.24.

$$C(\varepsilon) = \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} \Theta\left(\varepsilon - \left|\xi_i - \xi_j\right|\right) \tag{3.24}$$

onde ξ_i é o i-ésimo vetor de dimensão m (dimensão de imersão) da série temporal reconstruída através do método de Takens, N é o número de vetores ξ_i reconstruídos e $\Theta(arg)$ é a função de Heaviside, que retorna 1 para $arg \geq 0$ e retorna 0 para $arg < 0$.

O cálculo do somatório de correlação pode ser alternativamente escrito [KUM 96] como mostra a equação 3.25.

$$C(\varepsilon) = \frac{1}{(N-W)(N-W+1)} \sum_{j=1}^{N} \sum_{\substack{i=1 \\ |i-j|\geq W}}^{N} \Theta\left(\varepsilon - \left|\xi_i - \xi_j\right|\right) \tag{3.25}$$

onde *W* é empiricamente determinado, cujo valor típico é 10.

A proposta de modificação da equação 3.24 objetiva excluir os pares de pontos (ξ_i,ξ_j) que estão próximos no espaço de fases apenas porque estão temporalmente próximos.

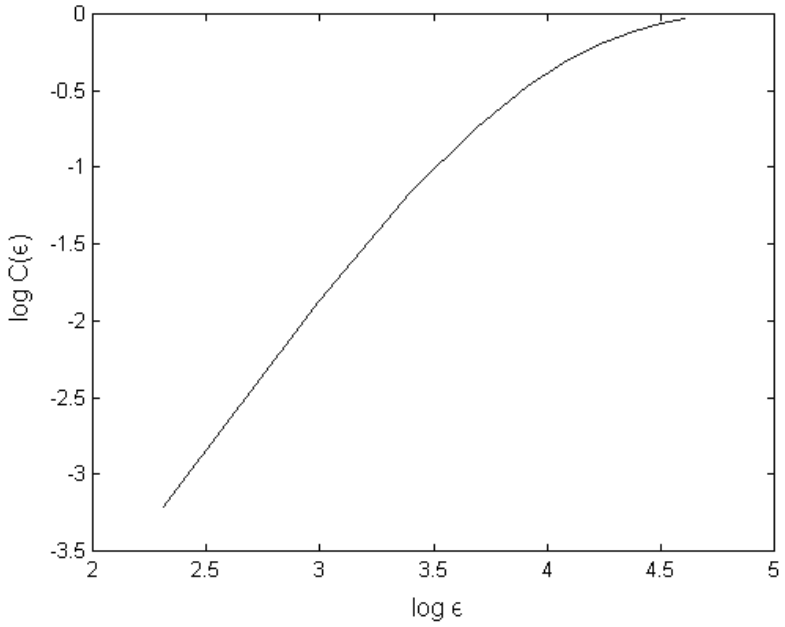

FIGURA 3.8 - Curva log C(ε) *versus* log ε para o atrator de Lorenz

Na prática, deve-se obter a curva de *log C(ε) versus log ε*, e aplicar um ajuste linear para obter a declividade dessa curva, que aproxima o valor da dimensão de correlação D_2. Para o atrator de Lorenz (equações 3.4), essa curva é mostrada na figura 3.8, e a dimensão de correlação obtida vale aproximadamente 1,37.

Na grande maioria das aplicações em sinais temporais experimentais, o algoritmo de Grassberger-Procaccia [GRA 83a] descrito não consegue estimar com segurança dimensões de correlação maiores que 5, que corresponde a dimensões de imersão da ordem de 10 [FER 94].

3.7.3 Expoentes de Lyapunov

Para sistemas dinâmicos, os expoentes de Lyapunov (λ_i) fornecem uma medida de sensibilidade do sistema às condições iniciais. Quando o atrator associado ao sistema é caótico, ao considerar-se duas trajetórias cujas condições iniciais são muito próximas, as trajetórias divergem, em média, a uma taxa exponencial caracterizada pelo maior expoente de Lyapunov. A presença de um expoente positivo é suficiente para garantir que as órbitas sejam exponencialmente sensíveis às condições iniciais ou a perturbações, diagnosticando-se o caos. O atrator de um sistema dissipativo com um ou mais expoentes de Lyapunov positivos é classificado como estranho [WOL 85]. Nenhuma informação sobre o comportamento local de um atrator é extraída dos expoentes de Lyapunov.

É possível estimar o valor da dimensão fractal (D_0) através do espectro de expoentes de Lyapunov [KAP 79]. A dimensão obtida aproxima a dimensão fractal para atratores estranhos, e é muitas vezes chamada de dimensão Kaplan-York ou dimensão de Lyapunov (D_{KY}). O método de cálculo é mostrado na equação 3.26.

$$D_{KY} = j + \frac{\sum_{i=1}^{j} \lambda_i}{\left|\lambda_{j+1}\right|} \tag{3.26}$$

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

O problema de determinar os expoentes de Lyapunov quando apenas observações escalares são disponíveis é outro desafio. Como sempre, o primeiro passo é reconstruir o espaço de fases através das técnicas mostradas anteriormente. Procede-se então a análise do que ocorre com pontos na vizinhança de uma trajetória. O maior expoente de Lyapunov é estimado como a taxa média de separação entre pontos vizinhos.

Os métodos disponíveis para a estimativa do espectro de expoentes de Lyapunov diferem em relação ao número de expoentes calculados, e a forma de se aproximar a dinâmica do sistema nas proximidades da trajetória de referência. Estes métodos, por um lado populares, quase sempre produzem resultados precisos apenas para o cálculo do maior expoente de Lyapunov (λ_1), devido a inevitáveis erros numéricos [ABA 93].

- **Método de Wolf**

Este algoritmo, proposto por Wolf [WOL 85], permite a estimativa dos expoentes de Lyapunov não negativos de uma série experimental. No início, calcula-se o maior expoente de Lyapunov positivo da série (λ_1). A seguir, é calculado o segundo maior expoente de Lyapunov (λ_2), e assim por diante.

Considera-se uma trajetória de referência no atrator reconstruído através dos elementos da série experimental sob avaliação. Desta forma, é avaliada a taxa de divergência que um par de pontos muito próximos possui.

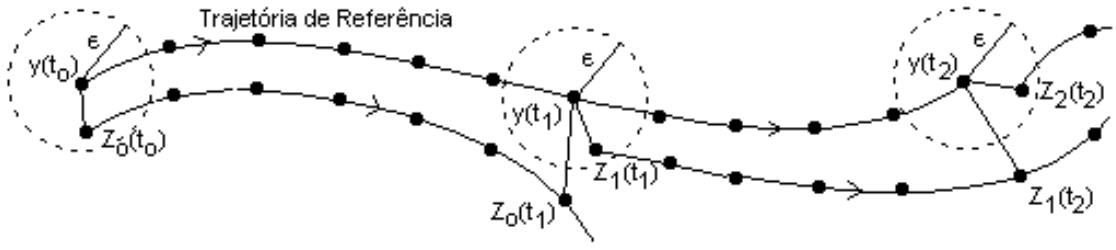

FIGURA 3.9 - Diagrama esquemático do método de Wolf para o cálculo dos expoentes de Lyapunov

A figura 3.9 ilustra o processo de obtenção dos expoentes de Lyapunov utilizando o método de Wolf. Dada a série temporal *x(t)*, o espaço de fases *m*-dimensional é primeiramente reconstruído, de forma que um ponto qualquer do atrator

pode ser escrito como $\vec{\xi}_i = \{x(t_i), x(t_i + p), x(t_i + 2p), ..., x(t_i + (m-1)p)\}$. Deve-se primeiramente localizar o ponto mais próximo ao ponto inicial $y(t_0)$ da trajetória de referência (fiducial), através da medida euclidiana de distância. Esse ponto mais próximo é escrito como $Z_0(t_0)$. A distância entre esses dois pontos é chamada $L(t_0)$ que deve ser suficientemente pequena para que $Z_0(t_0)$ esteja contido dentro de uma hiper-esfera de raio ε,, centrada em $y(t_0)$. Acompanha-se a evolução temporal da distância entre os pares de pontos até que, num próximo instante de tempo (t_1), essa distância exceda o valor ε. Essa distância superior a ε é chamada $L'(t_1)$. Quando isso ocorrer, deve-se encontrar o ponto mais próximo de $y(t_1)$, que esteja numa direção próxima a do segmento formado pelos pontos $y(t_1)$ e $Z_0(t_1)$. Esse novo ponto chamamos de $Z_1(t_1)$, e a distância entre $y(t_1)$ e $Z_1(t_1)$ vale $L(t_1)$. Esse processo de substituição é repetido quando novamente a distância entre os pontos exceder o valor ε. Isso é feito até que todos os pontos da trajetória de referência tenham sido percorridos. O maior expoente de Lyapunov positivo pode ser então obtido através da equação 3.27.

$$\lambda_1 = \frac{1}{t_M - t_0} \sum_{i=0}^{M-1} \log_2 \frac{L'(t_{i+1})}{L(t_i)} \tag{3.27}$$

onde M é o número total de substituições de pontos realizadas, t_i é o tempo da i-ésima substituição.

O cálculo do somatório dos dois maiores expoentes de Lyapunov ($\lambda_1+\lambda_2$) é similar ao método descrito para o cálculo do maior expoente, mas é de mais complicada implementação. O ponto inicial da trajetória de referência é utilizado e seus dois pontos mais próximos são escolhidos. A área $A(t_0)$ definida por esses três pontos é monitorada até que o passo de substituição seja desejável e possível. Assim, é definido um número mínimo de passos entre substituições, o número de passos que serão desfeitos quando uma substituição se mostra inadequada, e uma medida de área máxima anterior à substituição. A propagação e substituição são repetidas através da trajetória principal e os dois maiores expoentes de Lyapunov são estimados conforme a equação 3.28.

$$\lambda_1 + \lambda_2 = \frac{1}{t_M - t_0} \sum_{i=0}^{M-1} \log_2 \frac{A'(t_{i+1})}{A(t_i)} \tag{3.28}$$

onde M é o número total de substituições de pontos realizadas, t_i é o tempo da i-ésima substituição.

- **Método de Rosenstein**

Vários métodos para estimar o espectro de expoentes de Lyapunov são disponíveis. Muitos deles são baseados no método de Wolf que certamente é o mais conhecido. Nesta seção, é apresentado um outro método, desenvolvido por Rosenstein [ROS 93], que parece ser o mais adequado a séries temporais provenientes de sinais de voz.

Um dos problemas de se estimar invariantes dinâmicas a partir de um sinal de voz é a necessidade de que a série temporal represente um estado estacionário do sistema. Isso já foi abordado anteriormente e acarreta a necessidade de se trabalhar com quadros temporais de aproximadamente 30 ms do sinal de voz. Como conseqüência, o número de pontos disponível para a reconstrução do atrator e posterior estimativa de invariantes dinâmicas é tipicamente da ordem de algumas centenas. A principal vantagem do método de Rosenstein é a possibilidade de se extrair informações a partir

de séries temporais pequenas, utilizando-se a maior quantidade de informação possível. Esse método, por outro lado, permite a estimativa apenas do maior expoente de Lyapunov (λ_1).

O primeiro passo do método de Rosenstein é a reconstrução do atrator no espaço de fases apropriado. Após, deve-se localizar o vizinho mais próximo a cada ponto da trajetória, através da medida euclidiana de distância. Uma condição para a identificação dos vizinhos mais próximos deve ser satisfeita: eles devem ter uma separação temporal maior que o período médio da série temporal. Essa condição permite considerar cada par de vizinhos como condições iniciais de trajetórias diferentes. O período médio pode ser aproximado através da freqüência média do espectro de potências do sinal.

Ao considerar-se duas trajetórias cujas condições iniciais são muito próximas, as trajetórias divergem, em média, a uma taxa exponencial caracterizada pelo maior expoente de Lyapunov (λ_1), como mostrado na equação 3.29.

$$d_j(i) \approx C_j e^{\lambda_1 (i\Delta t)} \tag{3.29}$$

onde $d_j(i)$ é a distância entre o j-ésimo par de vizinhos próximos após i passos discretos (equivalente a $i.\Delta t$ segundos), Δt é o período de amostragem da série temporal e C_j é a separação inicial entre os vizinhos.

Aplicando-se o logaritmo natural aos dois lados da equação 3.29, ficamos com a equação 3.30.

$$\ln d_j(i) \approx \ln C_j + \lambda_1 (i\Delta t) \tag{3.30}$$

Dessa forma, para cada um dos pontos da trajetória reconstruída, é calculado seu vizinho mais próximo, respeitando-se o critério de separação temporal mencionado. A seguir, é monitorada a evolução da distância entre cada par de pontos. O logaritmo dessa distância deverá se apresentar com uma reta, de acordo com a equação 3.30 anterior. Para todos os pares de pontos, um conjunto de linhas aproximadamente paralelas serão obtidas. Cada linha possui uma inclinação razoavelmente proporcional a λ_1. Calcula-se a “linha média” representando a evolução do logaritmo da distância entre todos os pares de pontos. Aproximando-se a “linha média” a uma reta, pode-se estimar o maior expoente de Lyapunov como a declividade dessa reta. A figura 3.10 ilustra o gráfico dessa evolução para o atrator de Lorenz (equação 3.4), onde é possível observar sua declividade, cujo valor está bastante próximo ao expoente de Lyapunov esperado de 1,50 [ROS 93].

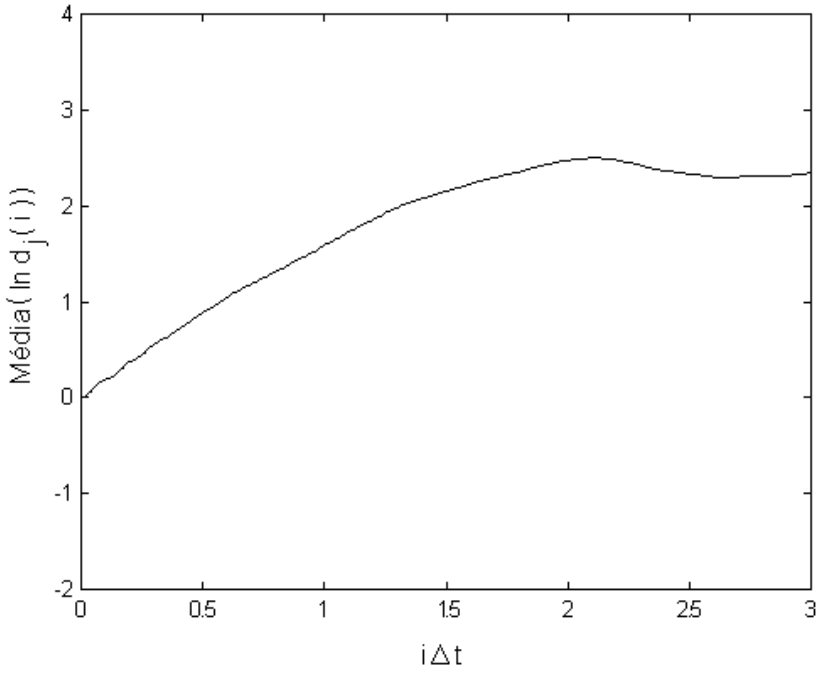

FIGURA 3.10 - Evolução da média do logaritmo da distância entre os pares de pontos para o atrator de Lorenz

3.8 Resumo

Este capítulo mostrou algumas das principais técnicas relacionadas à Teoria do Caos, dando ênfase ao tratamento de séries temporais. Primeiramente uma breve introdução à área foi apresentada, visando à ambientação do leitor relativamente aos métodos apresentados ao longo do capítulo. A seguir, foi realizada uma análise comparativa entre o tratamento apropriado aos sistemas lineares e dinâmicos não-lineares. Também foram mencionados aspectos relativos à produção da voz, a fim de explicar a possível utilização da Teoria dos Sistemas Dinâmicos Não-Lineares em sinais vocais.

O principal método de reconstrução de atratores a partir de séries temporais foi apresentado, o chamado método dos atrasos temporais, assim como técnicas utilizadas para estimação do passo de reconstrução e dimensão de imersão adequados. Alguns cuidados que devem ser observados para uma reconstrução robusta foram indicados. A partir do sistema reconstruído no espaço de fases adequado, é possível extrair algumas informações qualitativas. Algumas técnicas empregadas para estimação da dimensão fractal, dimensão de correlação e expoentes de Lyapunov foram apresentadas.

4 Proposta de tese: melhoria do RAL a partir da utilização de invariantes dinâmicas

4.1 Introdução

Os capítulos anteriores focaram aspectos técnicos relativos aos sistemas de RAL existentes e métodos utilizados para diferenciar locutores. Além disso, as técnicas empregadas em séries temporais para a estimação de invariantes dinâmicas foram mostradas, de acordo com a Teoria do Caos. A forma como esse conjunto de informações pode ser utilizado para melhorar os sistemas de RAL foi mencionada. Este capítulo pretende aprofundar-se nos métodos que serão utilizados para efetivamente aplicar as ferramentas da Teoria do Caos em sinais de voz, e obter com isso sistemas de RAL com maior exatidão.

Ao longo desse capítulo, alguns aspectos chave serão abordados relativos à validade da análise proposta. Inicialmente, a possibilidade de conexão entre a área de RAL e Teoria do Caos será explorada. A seguir, será feita uma análise detalhada do sinal de voz, sob o ponto de vista dinâmico não-linear, procurando-se verificar se este apresenta características caóticas. Será mostrado exatamente como é possível aplicar as técnicas da Teoria do Caos em sinais de voz. As considerações especiais e modificações algorítmicas que devem ser realizadas serão detalhadas. Por fim, este capítulo utilizará as considerações explicitadas para validar a utilização de informações dinâmicas não-lineares em sistemas de RAL. A metodologia utilizada para a aplicação dos algoritmos descritos é detalhada e a escolha para o sistema-base utilizado nos experimentos é justificada. Os resultados com os principais experimentos que utilizam invariantes dinâmicas em sistemas de RAL são analisados. A seguir, é mostrada a fundamentação teórica para a análise de significância estatística dos resultados obtidos nos experimentos. Essa análise estatística é aplicada nos principais experimentos realizados, de forma a verificar a consistência dos resultados. Por fim, são avaliados os aspectos relativos ao intrínseco aumento de processamento acarretado pela estimação de invariantes dinâmicas.

4.2 Trabalhos correlatos

Foram encontrados alguns trabalhos na literatura que utilizam as ferramentas disponibilizadas pela Teoria dos Sistemas Dinâmicos Não-Lineares aplicadas em sinais biológicos, especialmente em sinais de voz. Na maioria dos casos, é avaliada a possibilidade de caracterização do mecanismo de produção desses sinais através de um sistema dinâmico. Alguns trabalhos relatam experimentos que objetivam reconhecer comandos vocais utilizando invariantes dinâmicas. Não foram encontrados trabalhos que utilizassem invariantes dinâmicas não-lineares agregadas a sistemas de RAL, assim como não foram encontrados trabalhos que desenvolvessem quaisquer mecanismos para sua incorporação nos sistemas existentes ou verificassem a sua real possibilidade de utilização.

A tese de doutorado de Kumar [KUM 94] apresenta um estudo da análise dinâmica de sinais de voz em termos de observações do mecanismo de produção da fala, do sinal de voz, das limitações do modelo linear e dos avanços nas técnicas de análise e modelamento dinâmico não-linear. Na continuação de seu trabalho, Kumar *et*

al. [KUM 96] relatam a estimativa dos expoentes de Lyapunov, dimensão e entropia a partir de fonemas, e sugerem a possibilidade de utilização dessas invariantes dinâmicas para sua caracterização.

Em [POR 95], Port *et al*. propõem que a Teoria dos Sistemas Dinâmicos oferece ferramentas apropriadas para modelar muitos aspectos fonológicos da produção e percepção da fala. Uma contagem dinâmica do ritmo da fala mostra-se útil para a descrição das temporizações da língua japonesa e inglesa na repetição de frases, sendo proposto um oscilador adaptativo para modelamento dos padrões temporais da produção da fala.

Sabanal *et al*. [SAB 96] examinaram as propriedades fractais de sons vocais simples, como vogais japonesas, através da avaliação da dimensão fractal e multifractal dependente do tempo (TDFD e TDMFD). Os expoentes de Lyapunov foram também estimados a fim de comprovar a existência de caos nos atratores reconstruídos a partir de sons vocais. Foi construído um reconhecedor de dígitos japoneses utilizando uma rede neural.

Em [OLI 99], Oliveira *et al*. propuseram a utilização da dimensão fractal dependente do tempo (TDFD) para análise de sons pulmonares, visando o possível diagnóstico médico de doenças associadas ao aparelho respiratório. A tese de doutorado de Custódio [CUS 99] também explorou de forma extensiva a utilização da TDFD em sons pulmonares.

No trabalho de Sciamarella *et al*. [SCI 99], são relatados experimentos para determinação de equivalência topológica na análise da dinâmica não-linear reconstruída a partir de sons vocais.

Banbrook *et al*. [BAN 99] investigaram a geração de sons vocais como um processo dinâmico não-linear, extraindo invariantes geométricas a partir de uma base de dados composta de vogais sustentadas, e apresentou um método para síntese de voz inspirado na estimativa dos expoentes de Lyapunov.

Chan *et al*. [CHA 99] propuseram a utilização de uma técnica dinâmica não-linear (MPSV) para melhorar a qualidade de sinais sonoros e fala corrompidos por ruído aditivo.

Kohlmorgen *et al*. [KOH 2000] analisaram séries temporais extraídas a partir de sinais fisiológicos, como EEG e sons respiratórios, sugerindo que esses sinais provêm de sistemas dinâmicos com diferentes modos de operação, apresentando ferramentas para estudo de sistemas não-estacionários.

Bohez e Senevirathne investigam em [BOH 2001] a utilização da dimensão fractal visando o RAF. Neste trabalho é proposto um método para reconhecimento de fonemas e separação de palavras baseado nessa invariante dinâmica.

No trabalho [GUO 2002], Guo *et al.* avaliaram as propriedades fractais da linguagem mandarin, composta de sílabas chinesas. Os resultados experimentais mostraram que o sinal de voz falado na linguagem mandarin possui propriedades na qual a dimensão fractal apresenta bom potencial de caracterização. É também sugerido a combinação com outros métodos, como técnicas tradicionais de DSP e dimensão multifractal.

4.3 A Teoria do Caos e o RAL

A compreensão de processos naturais não se inicia com um conjunto de equações para um sistema dinâmico. Pelo contrário, o modelamento matemático é normalmente obtido após vários experimentos, que tentam reproduzir o processo sob análise de forma controlada e conhecida. Muitas vezes o modelo gerado é capaz de

fornecer os principais parâmetros que governam o processo natural sob análise. Entretanto, para muitos sistemas, vários fatores são desprezados devido a sua complexidade de representação ou pouca influência no resultado final.

Se, por um lado, um modelo completo pode ser inatingível, ainda assim é possível estimar com precisão informações acerca do sistema de comportamento complexo. A dinâmica do sinal pode ser avaliada e, assumindo-se que tais amostras provêm de um sistema caótico, medidas qualitativas, como dimensão e expoentes de Lyapunov, podem ser obtidas desse sistema. Essas medidas não fornecem informações específicas quanto ao modelamento do processo de produção do sinal avaliado, entretanto, pode-se aferir a respeito de seu comportamento, como, por exemplo, o grau de complexidade que o sistema apresenta.

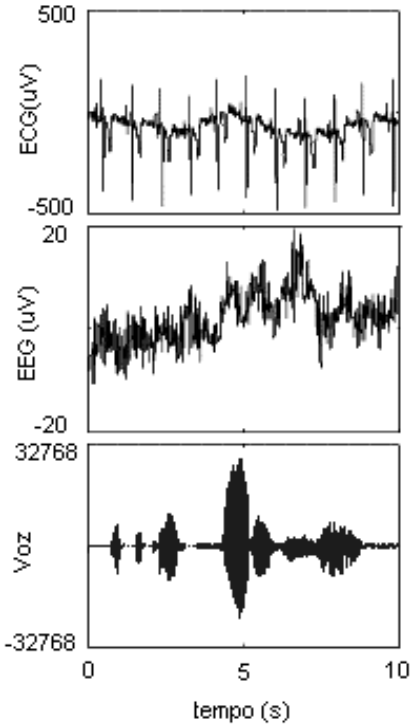

FIGURA 4.1 - Séries temporais obtidas a partir de sinais biológicos

Sinais biológicos, como os mostrados na figura 4.1, representam um grande desafio com relação a sua análise e modelagem. Um modelo detalhado para a produção da voz deveria considerar a variação temporal da configuração do trato vocal, as ressonâncias existentes no trato vocal, as perdas devidas ao atrito viscoso nas paredes do trato vocal, acoplamento da cavidade nasal, suavidade das paredes internas do trato vocal, o efeito do acoplamento subglotal (pulmões e traquéia) com a estrutura ressonante do trato vocal, e radiação do som nos lábios [KUM 96] [DEL 87]. A modelagem matemática linear, que utiliza diferentes fontes de excitação aplicadas a um filtro digital variante no tempo, é capaz de modelar com sucesso alguns dos fatores mencionados, enquanto outros são desprezados, pela dificuldade ou impossibilidade de descrevê-los em termos de combinações de sinais, filtros digitais, ou equações diferenciais. Cada locutor apresenta características fisiológicas diferentes, de forma a produzir sons diferentes. Essas diferenciações fisiológicas não se limitam apenas aos fatores matematicamente modelados. Assim, podem existir informações que caracterizem um locutor nos fatores desconsiderados no modelo matemático atual.

Essas informações, quando agregadas àquelas atualmente utilizadas, poderão conduzir o sistema que as utiliza a um grau de exatidão maior, no que diz respeito à identificação unívoca de uma pessoa através de amostras de sua voz.

A proposta de tese de doutorado apresentada neste trabalho fundamenta-se na possibilidade de utilização de técnicas da Teoria do Caos para a caracterização de pessoas através de sua voz, objetivando oferecer informações que melhorem os sistemas de RAL existentes. Defende-se primeiramente que o mecanismo de produção dos sons vocais pode ser visto como um sistema dinâmico não-linear com características caóticas. Dessa forma, estimam-se invariantes dinâmicas extraídas a partir da reconstrução do possível atrator associado ao sinal de voz, e caracteriza-se o locutor que produziu tal sinal através desses parâmetros. Nesse sentido, mecanismos foram desenvolvidos para possibilitar a análise dos sinais de voz sob o ponto de vista dinâmico não-linear e a eficiência dos parâmetros extraídos, no sentido de diferenciar locutores distintos, é analisada através de experimentos práticos.

4.4 Características caóticas no sinal de voz

As ferramentas disponibilizadas pela Teoria dos Sistemas Dinâmicos Não-Lineares são válidas quando aplicadas em séries temporais ditas caóticas. A série temporal é caótica quando é produto de um estado estacionário de um sistema dinâmico que necessariamente apresenta não-linearidades de forma a ser sensível às condições iniciais. Ou seja, uma pequena modificação nas condições na qual o sistema está inserido deverá acarretar uma significativa diferença no seu comportamento. A sensibilidade às condições iniciais está diretamente relacionada com a imprevisibilidade do comportamento do sistema.

Existem várias formas de verificação da existência de caos associado a uma série temporal. Inicialmente, a trajetória do possível atrator associado à série temporal deve ser reconstruída. A figura 4.2 ilustra a reconstrução da trajetória de atratores a partir de sinais randômicos, periódicos e conhecidamente caóticos. O sinal caótico mostrado foi obtido a partir das equações de Lorenz, mostradas na equação 3.4, com parâmetros r=45.92, b=4.0 e σ=16.0, que fornecem um sinal conhecidamente caótico [ABA 93].

Séries temporais caóticas apresentam necessariamente atratores caóticos. Sabe-se que quando um atrator associado ao sistema sob análise é caótico, ao considerar-se duas trajetórias cujas condições iniciais são muito próximas, elas divergem, em média, a uma taxa exponencial caracterizada pelo maior expoente de Lyapunov. Assim, a presença de um expoente de Lyapunov positivo é condição suficiente para garantir a existência de caos na série temporal sob análise [ABA 93] [BAN 99] [FER 94] [KUM 94] [ROS 93]. Da mesma forma, a extração da entropia de Kolmogorov é capaz de quantificar o grau de caos presente no sistema. Assim, sinais puramente randômicos apresentam entropia de Kolmogorov infinita, trajetórias regulares apresentam entropia igual a zero, e sistemas caóticos apresentam entropia finita e maior que zero [GRA 83b] [COH 85].

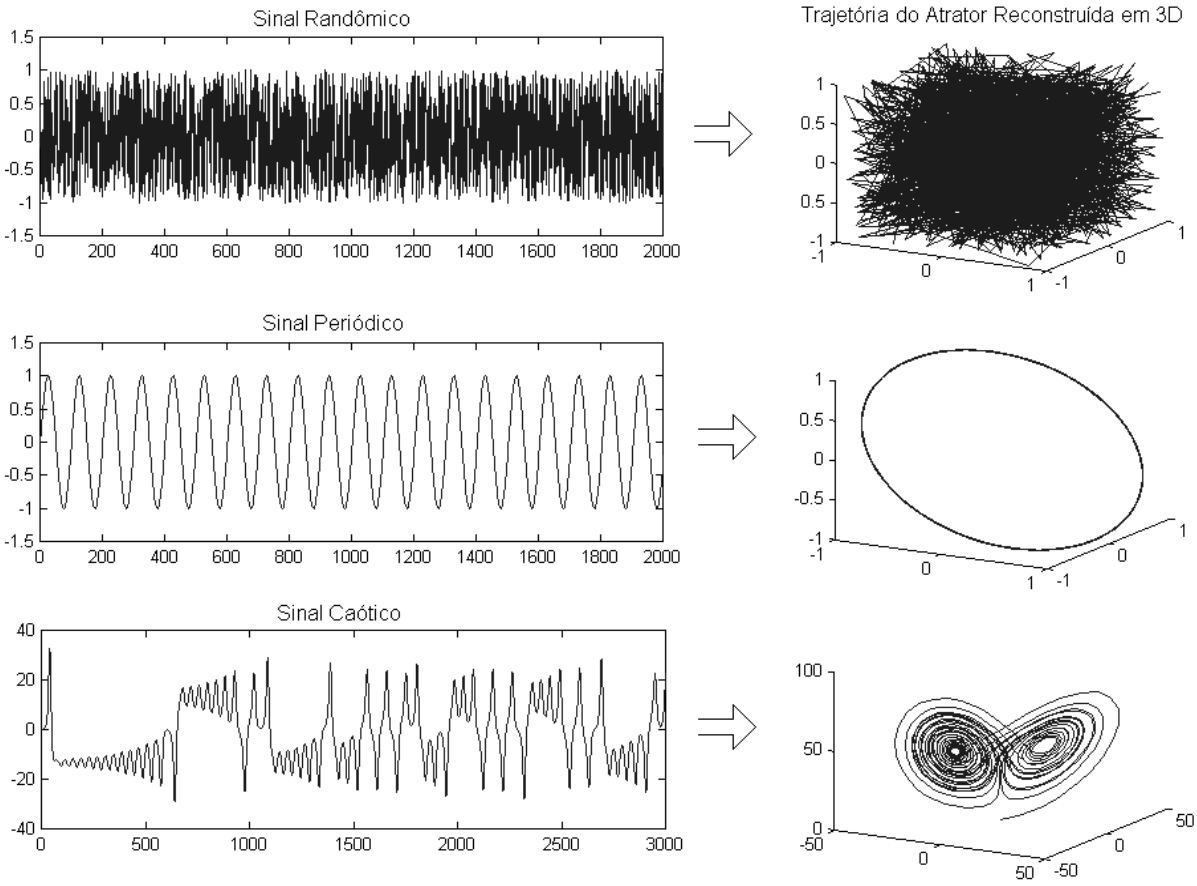

FIGURA 4.2 - Exemplo de reconstrução em 3D da trajetória do atrator para sinal randômico, periódico e caótico

A análise que deve ser utilizada para avaliar sinais de voz, em busca de características caóticas, parte do pressuposto que os dados sob análise foram obtidos a partir de um estado estacionário do sistema. Para intervalos de tempos longos (da ordem de segundos), sabe-se que o sinal de voz não é estacionário, uma vez que o aparato articulatório vocal está continuamente modificando sua posição de forma a gerar sons diferentes que compõem aquilo que está sendo dito. Por outro lado, pequenas porções de som (algumas dezenas de milisegundos) poderão ser assumidas como provenientes de um estado estacionário do sistema, pois se sabe que a variação de configuração do trato vocal é lenta [DEL 87] [RAB 78] [RAB 93]. Assim, a análise da presença de caos em sinais de voz deverá ser realizada utilizando quadros sucessivos, que selecionem apenas algumas porções para avaliação.

Para verificar a existência de sensibilidade às condições iniciais em sinais de voz, foi realizado um experimento que utiliza amostras de voz faladas por diferentes locutores. A análise realizada utilizou quadros de voz, e a partir de cada quadro foi reconstruída a trajetória do possível atrator associado, utilizando os métodos apropriados para estimação da dimensão de imersão e passo de reconstrução mostrados no terceiro capítulo. A seguir, o maior expoente de Lyapunov foi estimado para todos os quadros de voz, utilizando o método de Rosenstein [ROS 93]. A figura 4.3 ilustra o processo de estimação do maior expoente de Lyapunov para um quadro de voz. A repetição desse processo para todos os quadros do sinal avaliado fornece a variação temporal do maior expoente de Lyapunov.

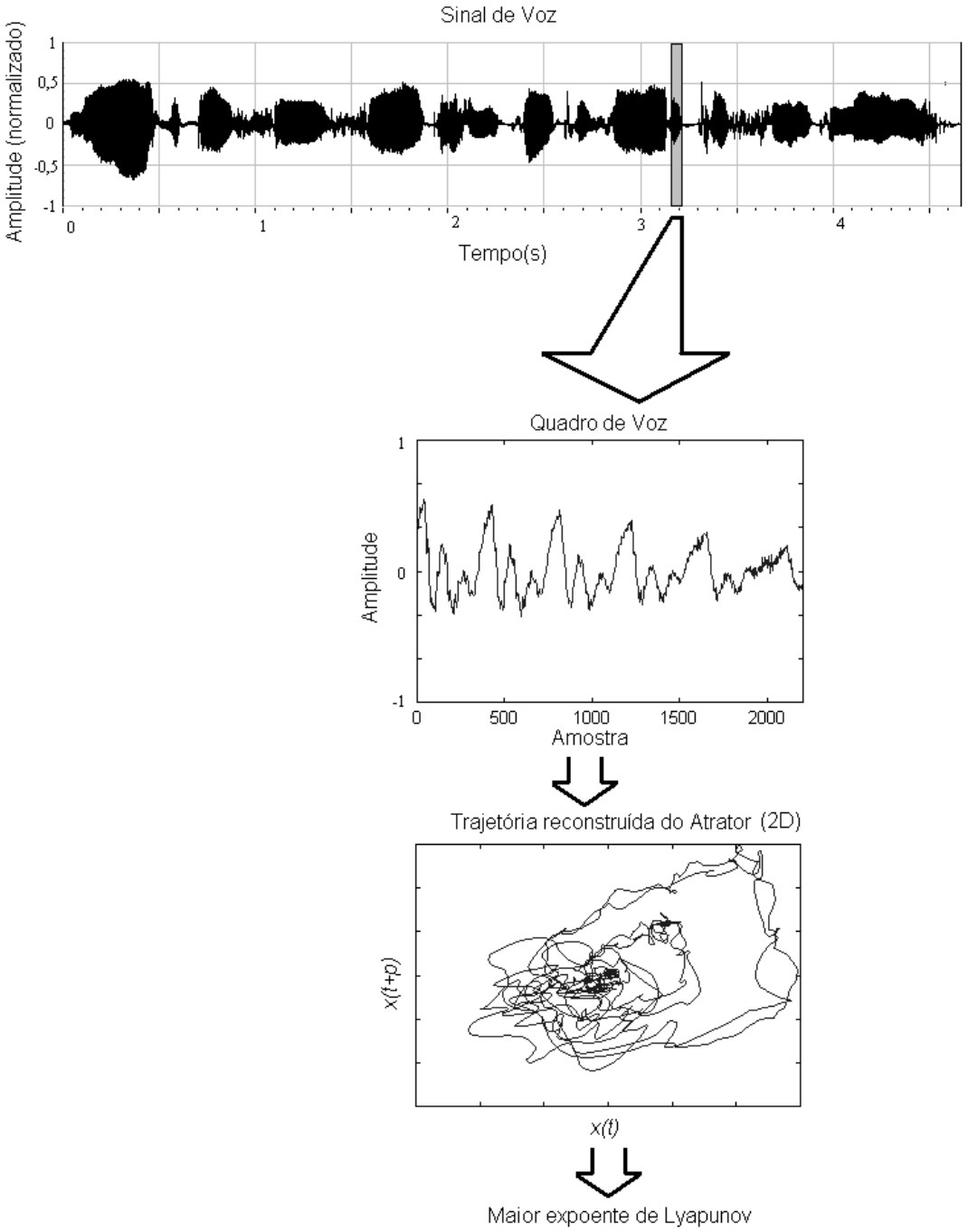

FIGURA 4.3 - Processo de estimação do maior expoente de Lyapunov a partir de um quadro de voz aproximadamente estacionária

A estimação da variação temporal do maior expoente de Lyapunov para o sinal de voz da figura 4.3 pode ser vista na figura 4.4. É possível notar que nem todos os quadros de voz possuem maior expoente de Lyapunov positivo. Isso acontece principalmente em transições entre palavras, onde a coarticulação ou um período de silêncio (que é predominante considerando apenas o quadro em questão) faz com que a respectiva estimativa para o valor do maior expoente de Lyapunov forneça valores negativos. A ausência de amostras suficientes para estimar com exatidão o expoente de Lyapunov poderá também causar um aumento da população de estimativas de

expoentes negativos, pela limitação intrínseca dos algoritmos utilizados. Isso poderá acontecer em sinais de voz que foram amostrados a taxas muito baixas.

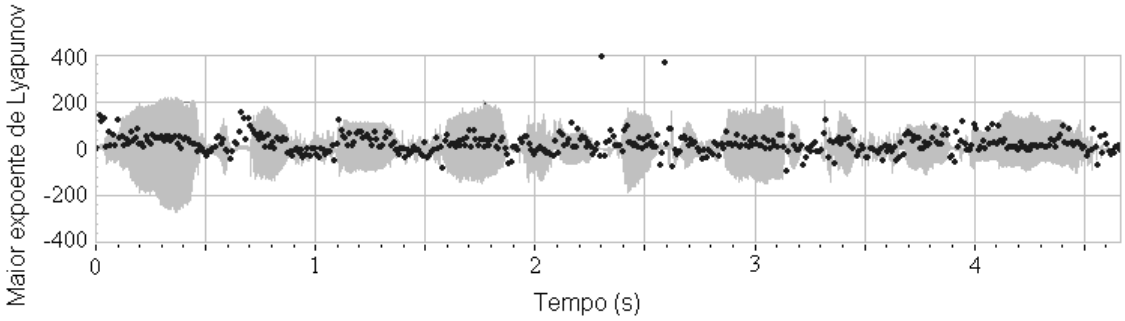

FIGURA 4.4 - Estimativa do maior expoente de Lyapunov para um sinal de voz, a partir de quadros com 30 ms de duração, extraídos a cada 10 ms

A fim de abordar a questão da presença de caos em sinais de voz, não se pode limitar a análise a apenas um arquivo de voz. Uma análise mais confiável pode ser realizada quando a estimação do maior expoente de Lyapunov é aplicada em um grande grupo de locutores diferentes. Assim, esse processo foi repetido e foram estimados 10000 valores para o maior expoente de Lyapunov, a partir de arquivos de voz obtidos de um grupo composto por 50 locutores, onde cada locutor fala uma seqüência de números distinta. O histograma para os valores do maior expoente de Lyapunov estimados são mostrados na figura 4.5. As vozes foram gravadas a uma taxa de amostragem de 44100Hz, e foram utilizados quadros de 30 ms de duração, extraídos a cada 10 ms. Através da análise da figura 4.5 fica evidenciada a presença de caos em grande parte dos atratores reconstruídos a partir de quadros oriundos de sinais de voz.

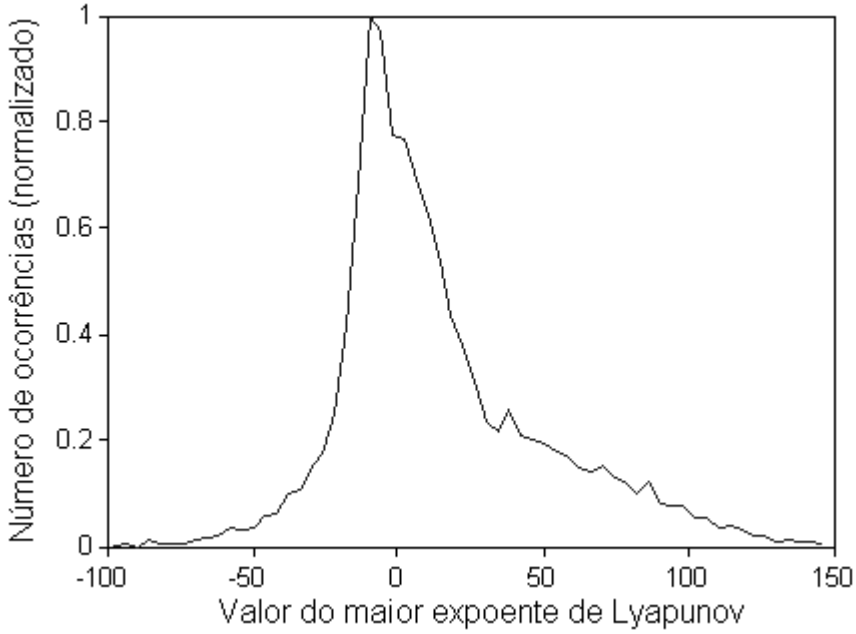

FIGURA 4.5 - Histograma dos valores para o maior expoente de Lyapunov, utilizando amostras de voz de 50 locutores distintos.

4.5 Análise dinâmica não-linear aplicada em sistemas de RAL

Muitos sistemas de RAL utilizam a combinação de parâmetros diferentes para compor um vetor de características representante de um quadro de voz. Por exemplo, em [LI 2000] Li *et al*. realizam experimentos utilizando um vetor de parâmetros composto por 39 elementos, que incluem 12 coeficientes cepstrais, 12 coeficientes delta-cepstrais, 12 coeficientes delta-delta cepstrais, 1 coeficiente de energia, 1 coeficiente de delta-energia e 1 coeficiente de delta-delta energia. A combinação de parâmetros é bastante comum, e deve ser realizada tomando-se o cuidado de não utilizar parâmetros cuja informação já esteja incluída naqueles parâmetros já utilizados. A inclusão de parâmetros sem o acréscimo de informação útil pode ser comparada ao acréscimo de ruído nos parâmetros existentes, o que certamente prejudicaria a exatidão do sistema. Assim, uma vez que as estimativas de invariantes dinâmicas não-lineares devem também ser realizadas quadro a quadro, é possível conceber sistemas de RAL que utilizem parâmetros que comprovadamente possuem eficiência na caracterização de locutores conjuntamente às invariantes dinâmicas não-lineares. Se a caracterização do sistema sob o ponto de vista dinâmico não-linear for realizada corretamente e contiver informação que realmente caracterize locutores, é provável que o desempenho do sistema de RAL melhore com o acréscimo dessas informações.

4.6 Metodologia de testes

Os testes realizados ao longo deste trabalho objetivam verificar a eficiência de sistemas de RAL quando são acrescidas informações dinâmicas não-lineares. Procurou-se adotar uma metodologia sistemática a fim de identificar a influência de alguns fatores de forma isolada. O banco de vozes utilizado, composto por 50 locutores distintos, foi coletado para avaliação de eficiência das técnicas empregadas. Foi definido um sistema-base para testes e, partir dessa configuração, alguns parâmetros foram variados e verificou-se a sua influência em testes que abrangeram a verificação de locutor.

4.6.1 Bancos de vozes

A utilização de uma base de dados apropriada é requisito fundamental para medidas comparativas de exatidão, e avaliação da influência de diversos fatores. Inicialmente, o banco de vozes deve possuir um número elevado de locutores, de ambos gêneros e idades variadas, a fim de garantir que os resultados não são atribuídos ao acaso. Além disso, cada locutor deverá fornecer várias repetições de sua locução, para que o sistema possa armazenar seu padrão vocal e ainda utilizar as gravações restantes para teste. É interessante que a taxa de amostragem utilizada na conversão A/D seja elevada, o que possibilita a verificação da influência da sua variação, através de filtragem digital. O ambiente de gravação não deverá ser muito ruidoso, uma vez que é possível inserir artificialmente ruído nos sinais e verificar sua influência. O método de coleta de voz também deverá ser idêntico para todos os locutores. Assim, controlando-se os principais fatores envolvidos no processo de criação de um banco de vozes, é possível obter o material necessário para análise consistente de resultados, e avaliação das principais variações existentes.

O trabalho desenvolvido utilizou um banco de vozes composto por 50 locutores diferentes, de forma a possibilitar a análise estatística nos resultados que garanta a confiabilidade das técnicas apresentadas. Cada locutor forneceu 20 gravações distintas

do mesmo grupo de palavras. O texto utilizado nas gravações foi escolhido de forma a garantir uma duração entre 5 e 6 segundos, sendo composto pela seqüência de palavras “nove dois seis sete um quatro zero três cinco meia oito”. As gravações foram realizadas em ambiente pouco ruidoso, a uma taxa de amostragem de 44100Hz, o que permite a repetição dos testes em taxas de amostragem inferiores. A resolução de 16 bits por amostra foi adotada. As principais características do banco de vozes utilizado para realização dos testes mostrados neste trabalho podem ser vistas na tabela 4.1.

O banco de vozes coletado foi subdividido para a avaliação de exatidão das técnicas empregadas. As 5 primeiras gravações de cada locutor foram utilizadas para a geração do padrão vocal (etapa de treinamento), e as demais foram utilizadas na etapa de testes. Nos testes de verificação de locutor realizados, as 15 gravações de teste de um determinado locutor são utilizadas para avaliar a taxa de falsa rejeição (FR), quando comparadas com o padrão vocal desse mesmo locutor. Da mesma forma, essas mesmas 15 gravações são utilizadas nos padrões vocais dos demais locutores, como gravações de impostores, permitindo a avaliação da taxa de falsa aceitação (FA). Assim, dispõe-se ao total de 750 tentativas de acesso para medição da taxa de FR, e 36750 tentativas de acesso para a taxa de FA.

TABELA 4.1 - Características do banco de vozes utilizado nos testes

Tipo de Informação	Descrição	Utilizado
Aquisição do Sinal de Voz	Microfone	Headset AC-300R mono Cyber Acoustics
	Taxa de Amostragem	44100 Hz
	Resolução	16 bits/amostra
	Modo de Gravação	Mono
	Codificação	PCM
Gravações	Locutores masculinos	28
	Locutores Femininos	22
	Duração aproximada da Locução	5-6 segundos
	Repetições por Locutor	20

4.6.2 Sistema-base para testes

O sistema-base para testes utilizou a medida de distância de Bhattacharyya para a comparação de sinais de voz, assumindo distribuições Gaussianas multivariadas. As informações-base que serão utilizadas são 16 coeficientes *mel*-cepstrais, que apresentam grande capacidade de diferenciar locutores. Além desses coeficientes, a sua derivada temporal de primeira ordem também é utilizada (coeficientes delta). A técnica de subtração do coeficiente médio (CMS) é usada, a fim de reduzir a influência do canal de transmissão. Os quadros foram obtidos através de uma janela de *Hamming* de 30 ms de duração, aplicada a cada 10 ms havendo, portanto, sobreposição de quadros sucessivos. Os quadros que apresentaram pouca energia (silêncio) são descartados. A configuração de teste inicial pode ser vista na tabela 4.2.

TABELA 4.2 - Configuração de testes do sistema-base

Tipo de Informação	Descrição	Utilizado
Aquisição do Sinal de Voz	Taxa de Amostragem	44100 Hz
	Resolução	16 bits/amostra
	Relação Sinal-Ruído (SNR)	Grande
Informações Extraídas	Tipo de Informação	Coeficientes *mel*-cepstrais
	Número de bancos de filtros	19
	Número de coeficientes	16
	Derivada de Primeira ordem	Sim
	Coeficiente de pré-ênfase	0.95
	Taxa de quadros por segundo	100 Hz
	Tamanho do quadro	30 ms
	Subtração da Média	Sim
	Remoção de quadros de silêncio	Sim
Modelos dos locutores	Tipo de modelo	Gaussiana Multivariada
	Medida de distância	Bhattacharyya

Há um compromisso entre as taxas de FA e FR, uma vez que é possível, através da variação no limiar de aceitação, reduzir uma delas acarretando o aumento da outra. Neste trabalho, a verificação do desempenho é também avaliado utilizando a chamada taxa de erro igual (EER). A EER equaliza as taxas de FA e FR, assumindo o valor onde ambas medidas são iguais. Esse parâmetro é um bom indicativo da exatidão geral do sistema de verificação de locutor [VUU 99]. A figura 4.6 mostra o gráfico de FA *versus* FR, chamado curva de detecção de erro balanceado (DET), utilizando o sistema-base para testes, onde é apontado o ponto onde ambas medidas são iguais a 2,04%.

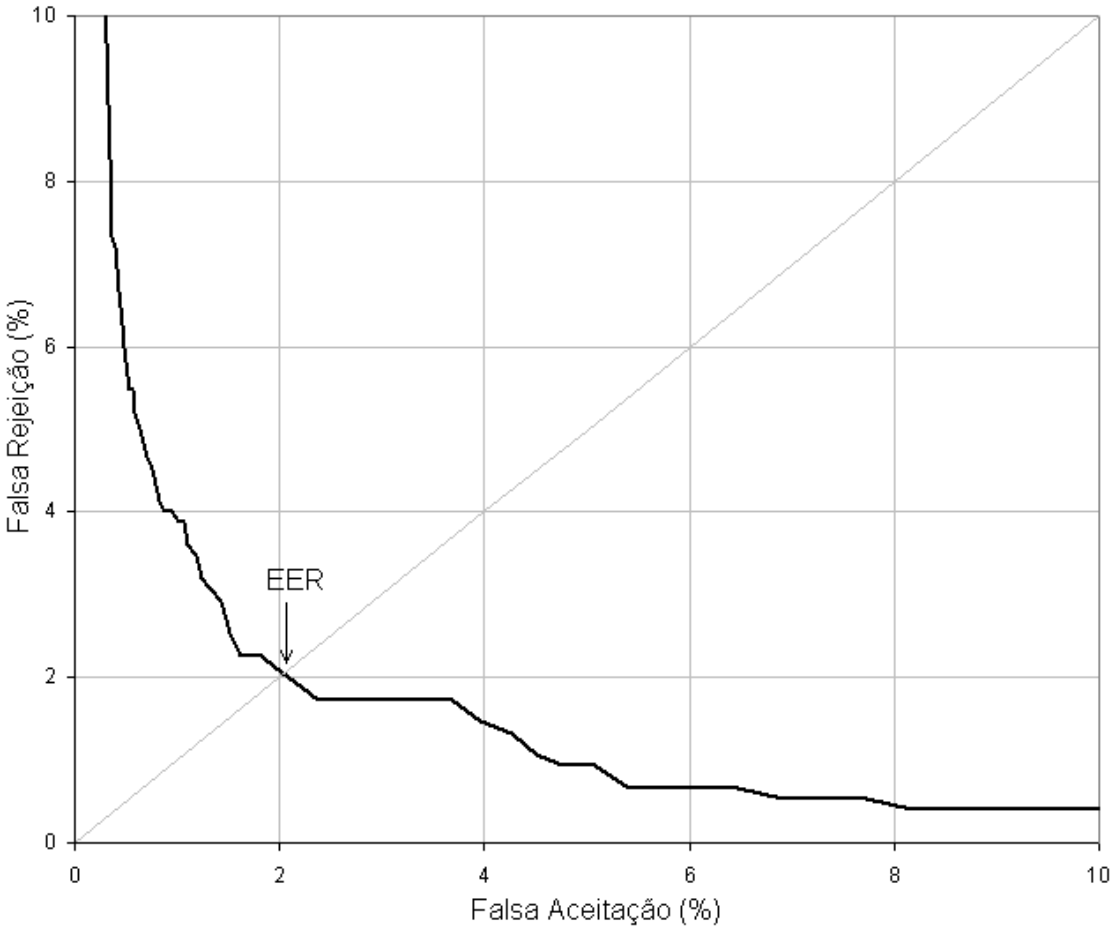

FIGURA 4.6 - Curva DET utilizando configuração de testes inicial

4.7 Adição de invariantes dinâmicas

Invariantes dinâmicas foram adicionadas ao sistema-base a fim de comparar a exatidão relativa alcançada com essa modificação. Foram considerados os testes de verificação de locutor, através da análise do EER obtido e das curvas DET. A influência de fatores como taxa de amostragem e duração do quadro, variação no número de coeficientes *mel*-cepstrais utilizados, utilização de delta coeficientes e aplicação da técnica CMS foi avaliada comparativamente ao acréscimo de invariantes dinâmicas não-lineares. As invariantes dinâmicas adicionadas ao sistema foram: dimensão fractal, dimensão de correlação e maior expoente de Lyapunov. As invariantes dinâmicas foram extraídas através dos mesmos quadros de voz utilizados para extração dos coeficientes *mel*-cepstrais.

A figura 4.7 mostra a curva DET para o sistema-base de testes e o sistema proposto (com adição das invariantes dinâmicas). Esse teste será referenciado na análise de significância estatística como T0. É possível observar uma clara redução nas taxas de erro do sistema proposto em relação ao sistema-base. O EER para o sistema proposto é 17,65% inferior ao do sistema-base, alcançando o valor de 1,68%. Além de melhorar o EER do sistema, a redução nos erros se dá ao longo de praticamente toda a curva DET.

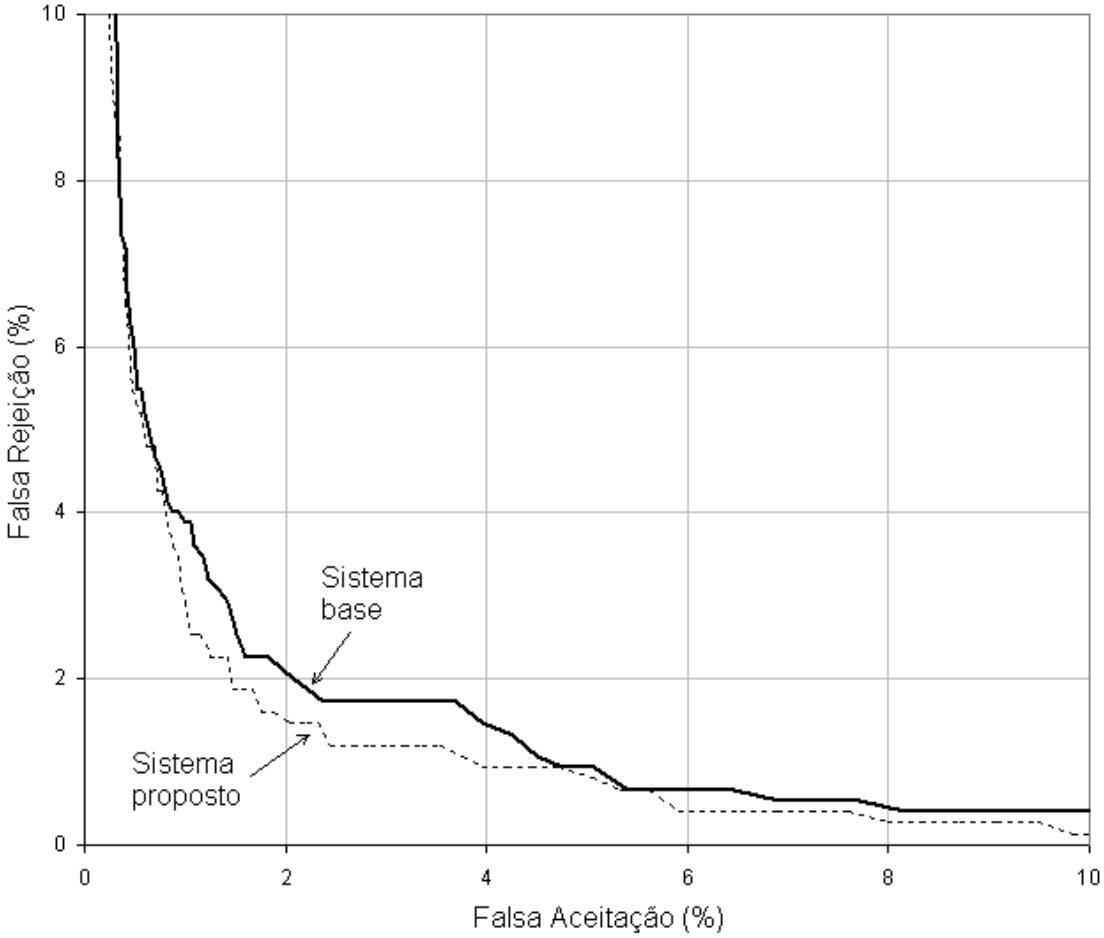

FIGURA 4.7 - Curva DET para o sistema-base e o sistema proposto

4.7.1 Taxa de amostragem e duração do quadro

A análise dinâmica não-linear utilizada em quadros sucessivos de sinais não-estacionários pode ser aplicada utilizando diferentes tamanhos de quadro, em sinais amostrados a diferentes taxas, que oferecem um número variável de amostras por quadro. A figura 4.8 mostra o número de amostras disponíveis em quadros de voz de durações diferentes, para algumas das principais taxas de amostragem. Essa informação é importante, pois fornece indicativo a respeito da quantidade de amostras requeridas para uma estimativa robusta das invariantes dinâmicas não-lineares.

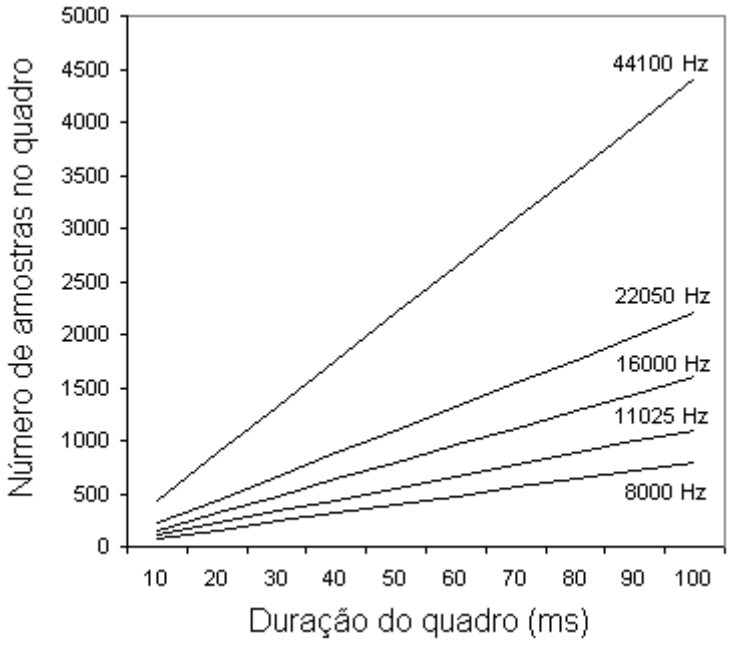

FIGURA 4.8 - Tamanho do quadro *versus* número de amostras disponíveis para algumas importantes taxas de amostragem

Inicialmente o fator de variação do número de amostras disponíveis em cada quadro de voz devido à variação na taxa de amostragem do sinal é observado. Para o sistema-base, a estimação dos coeficientes *mel*-cepstrais considera o sinal através da aplicação de bancos de filtros espaçados em freqüência. Entretanto, a utilização de taxas de amostragens maiores fornece informações mais precisas a respeito do comportamento do sinal, e conseqüentemente melhor caracterização do locutor que o produziu, devido à maior quantidade de amostras disponíveis no quadro de voz avaliado (maior faixa de freqüências contemplada). Isso pode ser visto na figura 4.9, onde pode ser observada uma redução no EER para taxas de amostragem maiores. Para o sistema proposto, essa redução no EER é mais acentuada, indicando a influência da estimação das invariantes dinâmicas. Assim, pode-se inferir que a necessidade de amostras para a estimativa das informações que caracterizam um locutor é maior para as invariantes dinâmicas que para os coeficientes *mel*-cepstrais. E na medida que uma estimativa mais robusta das invariantes dinâmicas é feita, maior é a caracterização do locutor para fins de verificação, acarretando um aumento na exatidão do sistema. É possível verificar que a utilização de uma taxa de amostragem igual a 11025 Hz já permite a estimação de invariantes dinâmicas com informações que caracterizam um locutor, aumentando a exatidão do sistema. Essa caracterização pode, entretanto, ser efetuada com maior sucesso quando taxas de amostragem maiores são utilizadas. Os testes mostrados na figura 4.9 serão referenciados na análise de significância estatística como T1, T2 e T3 para a utilização de taxas de amostragem de 11025 Hz, 22050 Hz e 44100 Hz, respectivamente.

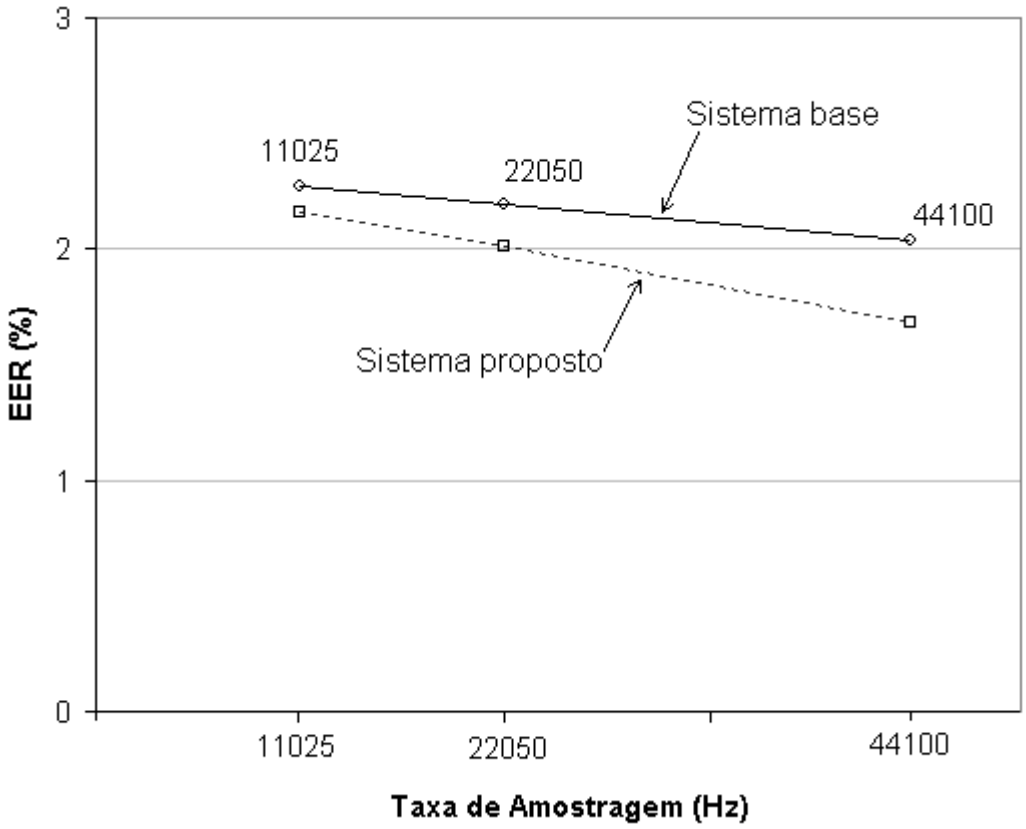

FIGURA 4.9 - Variação no EER para o sistema-base e sistema proposto, utilizando taxas de amostragem distintas

A figura 4.10 mostra a variação no EER em função do tamanho do quadro utilizado, mantendo-se a taxa de amostragem em 44100 Hz. É possível observar a necessidade das informações dinâmicas não-lineares em relação ao número de amostras disponíveis para sua estimação. Para quadros de apenas 10 ms de duração, a utilização das invariantes dinâmicas não proporcionou melhoras ao sistema. A partir de quadros de 20 ms de duração as informações dinâmicas não-lineares puderam ser estimadas de forma mais exata, acarretando a diminuição nas taxas de erro do sistema. Para quadros de 40 ms ou mais de duração a não-estacionariedade do sinal avaliado acaba prejudicando a exatidão do sistema. A análise de significância estatística referenciará os testes mostrados na figura 4.10 como T4, T5, T6 e T7 para utilização de quadros de 10 ms, 20 ms, 30 ms e 40 ms, respectivamente.

Os testes que variaram a taxa de amostragem e duração do quadro evidenciaram uma necessidade importante relativamente a uma estimativa robusta das invariantes dinâmicas não lineares: número de amostras disponíveis em cada quadro. O potencial máximo da caracterização de locutores através de invariantes dinâmicas não-lineares é alcançado para sistemas cujos quadros apresentam a maior quantidade de amostras possíveis. Entretanto, a partir de quadros constituídos de 882 amostras, a contribuição do acréscimo dessas informações já é bastante importante.

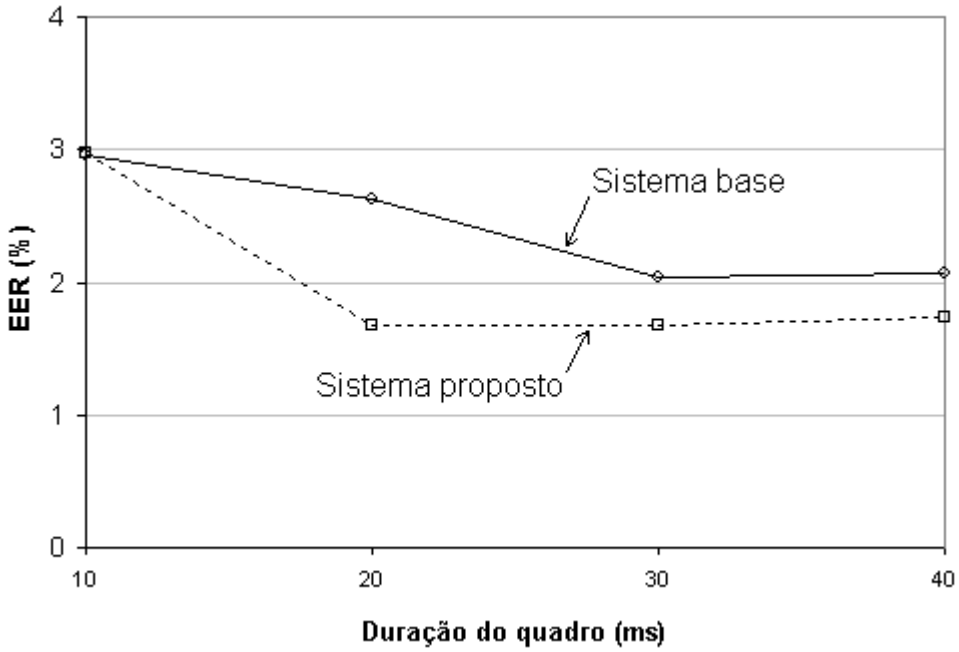

FIGURA 4.10 - Variação no EER para o sistema-base e sistema proposto, utilizando tamanhos de quadros distintos

4.7.2 Análise individual da contribuição das invariantes dinâmicas não-lineares

A figura 4.11 mostra as curvas DET obtidas com o acréscimo individualizado das invariantes dinâmicas utilizadas, e a curva resultante do acréscimo de todas invariantes dinâmicas. É possível verificar que, ao longo da maior parte da excursão das curvas, o benefício obtido com a inclusão de todas invariantes dinâmicas foi maior que o obtido individualmente a partir de cada invariante dinâmica. Com isso é possível verificar que a informação contida em cada invariante dinâmica contribui de forma individual e não redundante para a melhor caracterização do locutor.

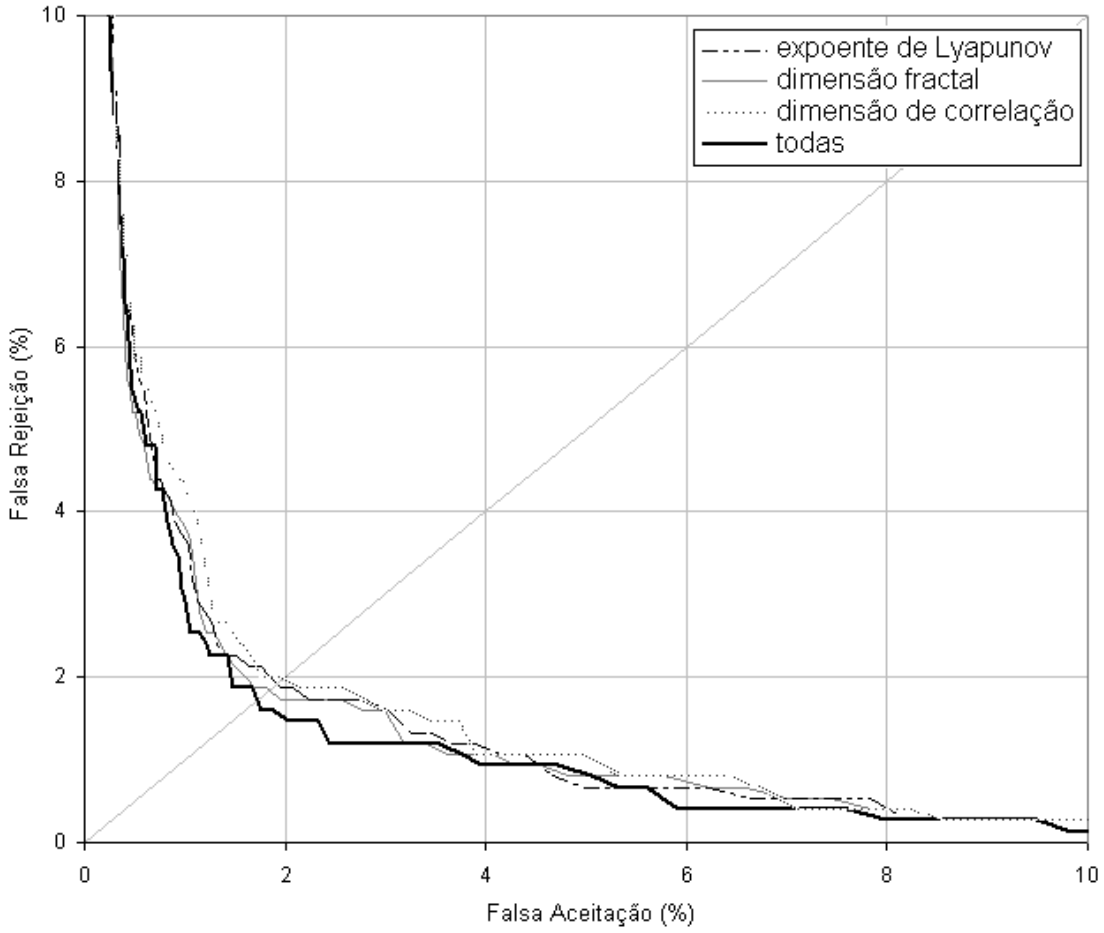

FIGURA 4.11 - Contribuição individual das invariantes dinâmicas no sistema de RAL

4.7.3 Robustez ao ruído

Um fator importante a ser avaliado é a sensibilidade das informações dinâmicas não-lineares com relação à inserção de ruído nas amostras de voz. Esse estudo deve ser realizado de forma comparativa com a sensibilidade das informações obtidas pela análise espectral tradicional. Assim, foram realizados testes que verificaram a exatidão alcançada pelo sistema de RAL para sinais de voz contaminados com ruído branco. A energia do ruído inserido foi variada, fornecendo o gráfico mostrado na figura 4.12, que mostra os resultados para diferentes relações sinal-ruído (SNR). A análise de significância estatística referenciará os testes mostrados na figura 4.12 como T8 e T9, para SNRs de 0 dB e 5 dB, respectivamente. O valor mostrado na figura 4.12 como elevado é a exatidão alcançada pelo sistema em ambiente pouco ruidoso sem o acréscimo artificial de ruído nos arquivos.

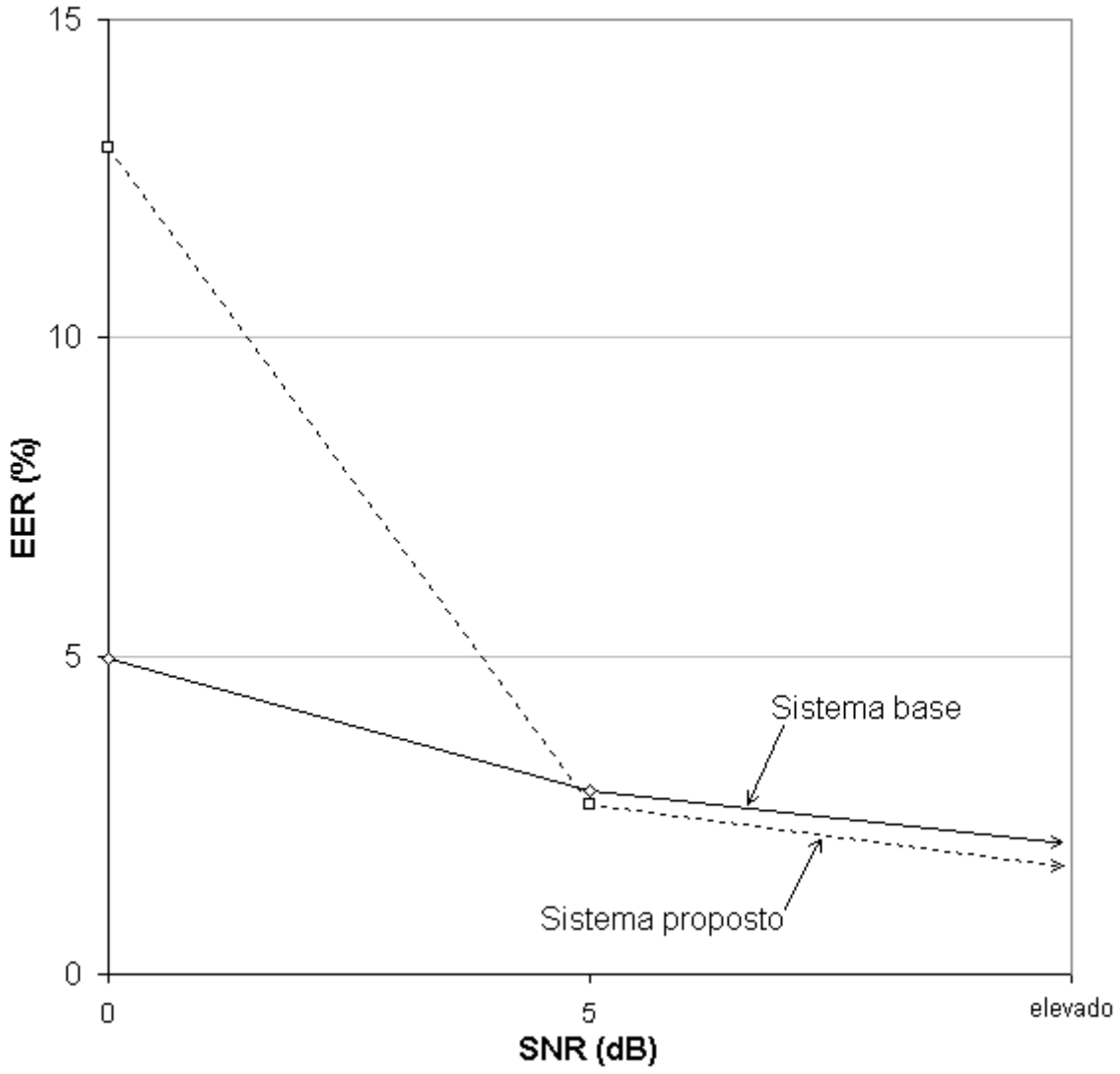

FIGURA 4.12 - Variação no EER para o sistema-base e sistema proposto, utilizando sinais contaminados com ruído branco

É possível observar a deterioração da exatidão do sistema de RAL quando a SNR diminui. Entretanto, essa deterioração é bem mais acentuada para o sistema que utiliza as invariantes dinâmicas não-lineares, fazendo com que sua exatidão seja pior que o sistema sem essas informações. Quando a SNR aumenta, as informações dinâmicas não-lineares passam a ser estimadas de forma mais exata, e acabam por melhorar a caracterização do locutor. Isso acontece para SNRs a partir de 5 dB. Mas é possível verificar uma alta sensibilidade dessas informações para SNRs baixas, o que acaba prejudicando bastante o sistema.

4.8 Significância estatística

Esta seção analisa estatisticamente a significância dos resultados obtidos. É comum encontrar na literatura algoritmos que são demonstrados como superiores a outros algoritmos, sem contudo verificar se a diferença de desempenho é significativa,

do ponto de vista estatístico. Primeiramente algumas considerações são realizadas, objetivando-se definir as variáveis envolvidas na análise e as hipóteses a serem validadas. A seguir, o teste de McNemar é apresentado para validação das hipóteses, e os resultados obtidos neste trabalho são avaliados.

4.8.1 Considerações iniciais

Suponha que os sistemas A_1 e A_2 devam ser testados e comparados no mesmo banco de vozes, composto por n arquivos de voz de diversos locutores. Seja N_{ij} o número de acertos e erros para ambos sistemas, quando utilizados no banco de vozes, como mostra a tabela 4.3.

TABELA 4.3 - Considerações para análise de significância estatística

		Sistema A_2	
		Acertos	Erros
Sistema A_1	Acertos	N_{00}	N_{01}
	Erros	N_{10}	N_{11}

A partir da tabela 4.3, é possível estabelecer as variáveis:

N_{00}: é o número de amostras de voz onde ambos sistemas realizam o reconhecimento corretamente;

N_{11}: é o número de amostras de voz onde ambos sistemas realizam o reconhecimento erradamente;

N_{01}: é o número de amostras de voz onde o sistema A_1 realiza o reconhecimento corretamente e o sistema A_2 erradamente;

N_{10}: é o número de amostras de voz onde o sistema A_1 realiza o reconhecimento erradamente e o sistema A_2 corretamente.

Em geral, N_{00} e N_{11} serão maiores que zero, de forma que os erros cometidos pelo sistema A_1 não serão independentes dos erros do sistema A_2. Isso sugere que um teste que diretamente compara os erros para os dois sistemas sobre o banco de vozes não seria apropriado [VUU 99].

Para sistemas de reconhecimento de locutor, suponha que as taxas de erro dos sistemas A_1 e A_2 são, respectivamente, p_1 e p_2. A estimativa para essas variáveis, para os testes realizados, é mostrada na equação 4.1.

$$p_1 = \frac{N_{11} + N_{10}}{n} \qquad p_2 = \frac{N_{11} + N_{01}}{n} \tag{4.1}$$

Neste caso, objetiva-se concluir, através da análise estatística, que tipo de relação existe entre as taxas de erro, ou seja, verificar se:

$$p1 = p2 \text{ ou } p1 \neq p2 \tag{4.2}$$

com um dado nível de significância $\propto$. Assim, deseja-se testar a hipótese nula H_0, mostrada na equação 4.3, contra a hipótese alternativa H_1, mostrada na equação 4.4.

$$H_0: \ p_1 = p_2 = p \tag{4.3}$$

$$H1: p1 \neq p2 \tag{4.4}$$

4.8.2 Teste de McNemar

Para determinar se a diferença no desempenho de dois sistemas é ou não estatisticamente significativo, o teste de McNemar pode ser utilizado [GIL 89]. Esse teste considera necessária a análise dos reconhecimentos onde apenas um dos sistemas cometeu um erro. Não há informação a respeito do desempenho *relativo* dos sistemas nos casos onde ambos reconhecem corretamente ou ambos erram. Assim, a hipótese nula assume que ambos sistemas avaliados possuem a mesma probabilidade de cometer erros, considerando os casos onde apenas um dos sistemas comete erro. Dessa forma, de acordo com H_0, N_{10} tenderá a seguir a distribuição binomial $\beta(k,1/2)$, onde $k=N_{10}+N_{01}$. As probabilidades podem ser diretamente calculadas como mostram as equações 4.5 e 4.6:

$$P = 2\sum_{m=N_{10}}^{k}\binom{k}{m}\left(\frac{1}{2}\right)^{k} \quad \text{quando } N_{10} > k/2, \tag{4.5}$$

$$P = 2\sum_{m=0}^{N_{10}}\binom{k}{m}\left(\frac{1}{2}\right)^{k} \quad \text{quando } N_{10} < k/2, \tag{4.6}$$

onde $\binom{k}{m}$ é o número binomial k sobre m, que pode ser expresso como mostra a equação 4.7.

$$\binom{k}{m} = \frac{k!}{m!(k-m)!}. \tag{4.7}$$

Dessa forma, H_0 será rejeitada quando P for menor que um nível de significância $\propto$. Alternativamente, se k for grande o suficiente ($k > 50$) e N_{10} não for muito próximo a k ou a 0, uma aproximação normal poderá ser utilizada, em aproximação à probabilidade binomial [GIL 89].

4.8.3 Análise dos resultados obtidos

Os testes efetuados e descritos anteriormente deverão satisfazer critérios estatísticos que validem as conclusões extraídas de suas análises. Assim, aplicando-se o teste de McNemar [GIL 89] descrito anteriormente nos resultados obtidos nos testes mostrados ao longo deste capítulo, obtemos os resultados mostrados na tabela 4.4. A comparação foi realizada entre sistema-base A_1 e o sistema proposto A_2 que adicionou ao sistema-base as invariantes dinâmicas não-lineares: dimensão fractal, dimensão de correlação e expoentes de Lyapunov. O nível de significância estatística utilizado para a análise dos resultados foi de $\propto = 0{,}02$.

TABELA 4.4 - Resultados da análise de significância estatística

Teste	EER sistema Proposto	EER sistema Base	Redução relativa percentual no EER	N_{01}	N_{10}	Estatisticamente significativo ($\propto = 0{,}02$)
T0, T3, T6	1,68 %	2,04 %	17,65 %	42	163	Sim
T1	2,16 %	2,28 %	5,26 %	69	105	Sim
T2	2,01 %	2,20 %	8,64 %	58	151	Sim
T4	2,97 %	2,97 %	0 %	143	191	Sim
T5	1,68 %	2,63 %	36,12 %	50	360	Sim
T7	1,74 %	2,07 %	15,94 %	35	175	Sim
T8	12,99 %	4,96 %	- 61,82 %	3428	353	Sim
T9	2,67 %	2,89 %	7,61 %	167	281	Sim

A análise de significância estatística realizada evidencia que os resultados obtidos através dos experimentos práticos representam o real comportamento do sistema proposto em relação ao sistema-base.

4.9 Tempo de processamento

É importante que o incremento de esforço computacional requerido para a utilização das invariantes dinâmicas não-lineares seja avaliado. A figura 4.13 mostra comparativamente a necessidade de tempo de CPU para a extração dos coeficientes *mel*-cepstrais em relação à extração das invariantes dinâmicas não-lineares. O computador utilizado nas medições possui processador Atlon, rodando a freqüência de 1,1 GHz. O tempo de processamento indicado na figura 4.13 equivale ao tempo médio utilizado para extração das informações em apenas um quadro de voz, cuja duração foi variada.

A análise da figura 4.13 evidencia claramente a alta necessidade de esforço computacional requerido para a estimação dos expoentes de Lyapunov e dimensão de correlação. Quanto maior for o número de amostras de voz existentes nos quadros avaliados, o recurso computacional necessário para a estimação dessas duas informações dinâmicas não-lineares aumenta de forma aproximadamente exponencial. Esse fator inviabiliza a utilização prática dessas informações em sistemas que devem apresentar uma resposta em poucos segundos para quadros que contenham mais que 1000 amostras. Através da figura 4.8 é possível estabelecer uma combinação entre taxa de amostragem e duração do quadro que forneça quadros com menos de 1000 amostras.

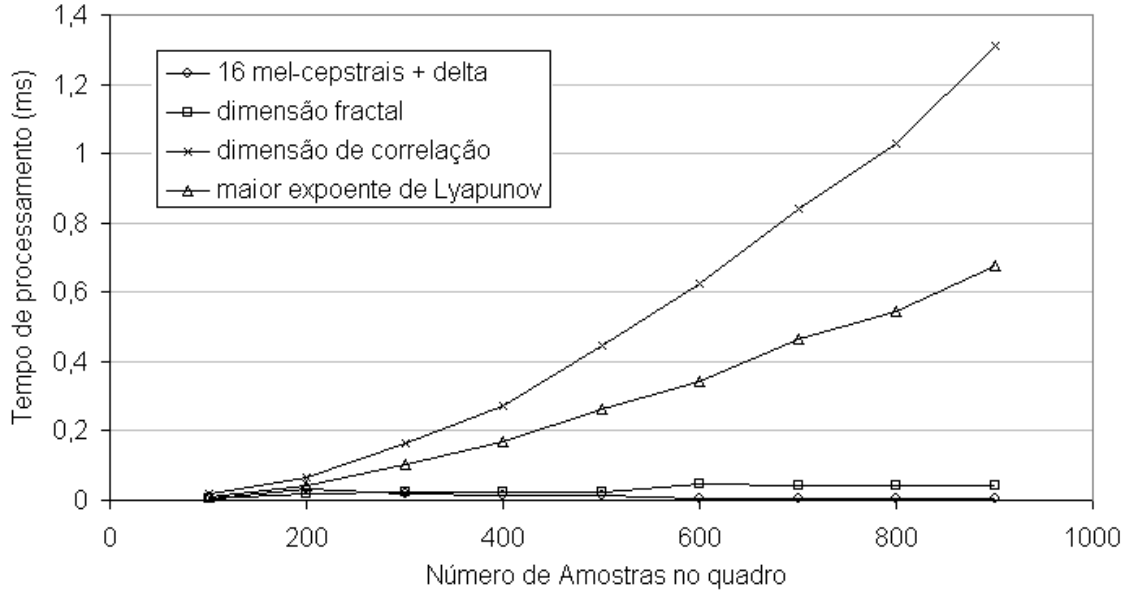

FIGURA 4.13 - Tempo de CPU requerido para estimação das informações utilizadas no sistema de RAL

Já a estimação da dimensão fractal não se apresenta como empecilho em relação ao tempo de resposta do sistema de RAL que a utilizar. Isso se deve principalmente por causa do método utilizado para sua estimação, o CEM, que consegue obter valores confiáveis avaliando o momento associado ao espectro de potências do sinal de voz. No caso de utilização de outros métodos para essa estimativa, como algoritmos de contagem de caixas, esse processamento seria inevitavelmente maior.

4.10 Resumo

Nesse capítulo a possibilidade de utilização de invariantes dinâmicas para a caracterização de locutores foi evidenciada, através da análise de sinais de voz e respectiva verificação da presença de atratores caóticos associados a quadros aproximadamente estacionários desses sinais. Assim, a conexão entre o RAL e a Teoria do Caos foi apontada, e os métodos utilizados para a aplicação de invariantes dinâmicas não-lineares nos sistemas de RAL foram mostrados. Procurou-se utilizar as invariantes dinâmicas como informações acrescidas àquelas já utilizadas e consagradas na literatura. A metodologia de testes utilizada para verificar se as invariantes dinâmicas podem realmente melhorar a caracterização de locutores foi mostrada. O sistema de RAL que utilizou esse tipo de informação apresentou redução nas taxas de erro, confirmadas pela inspeção em curvas DET e acompanhamento da variação do EER do sistema. Os principais resultados mostraram uma redução de 17,65 % no EER do sistema proposto em relação ao sistema-base. Outros testes indicaram a maior necessidade das invariantes dinâmicas relativamente ao número de amostras de voz disponíveis para sua estimação. Para sinais contaminados com ruído branco, a presença das invariantes dinâmicas não-lineares é benéfica para SNRs maiores que aproximadamente 5 dB. Quando o ruído presente é maior, a inclusão das invariantes dinâmicas não-lineares prejudica de forma intensa a exatidão geral do sistema. Os benefícios alcançados com o acréscimo de invariantes dinâmicas não-lineares foram obtidos através de acréscimo expressivo de esforço computacional, principalmente para a estimação da dimensão de correlação e expoentes de Lyapunov. A estimativa dessas

informações apresentou um aumento aproximadamente exponencial no tempo para processamento dos quadros de voz, na medida que o número de amostras contidas no quadro era aumentado. Verificou-se que a utilização de mais que 1000 amostras de voz em um quadro acabaria inviabilizando a construção de sistemas de RAL que apresentem tempos de resposta da ordem de segundos.

5 Conclusão

Ao longo deste trabalho foram enfocados aspectos relativos ao reconhecimento automático de pessoas através da análise de amostras de sua voz. Um dos principais objetivos buscados nos capítulos anteriores foi o de conectar os avanços obtidos na análise de sistemas (provenientes de séries temporais) através da Teoria do Caos com as técnicas utilizadas para a caracterização de locutores através da voz, viabilizando a construção de sistemas de RAL que apresentem maior exatidão. Inicialmente, foi realizada uma reflexão a respeito das possíveis aplicações e principais dificuldades existentes na área, apontando os esforços acadêmicos encontrados para solução dessas dificuldades. A seguir, foi fornecida uma visão geral das técnicas atualmente empregadas nos sistemas de RAL, mostrando os diferentes tipos de sistemas de RAL existentes, técnicas utilizadas para a aquisição do sinal de voz, extração e seleção da informação útil contida nesse sinal, e comparação de padrões. Também foi mostrada uma visão geral dos métodos existentes para a análise de sinais caóticos, dando especial atenção aos sinais provenientes de séries temporais obtidas a partir de medidas experimentais. Neste momento foram mostradas as ferramentas disponíveis para a correta reconstrução da trajetória do sistema no espaço de fases adequado, e posterior estimação de algumas invariantes dinâmicas não-lineares. O objetivo seguinte tratou de mostrar a conexão existente entre a área de RAL e Teoria do Caos, além de sugerir formas de aproveitar os conhecimentos de uma área na outra. Inicialmente, fez-se um estudo avaliando sinais de voz e verificando a existência de características caóticas em trechos aproximadamente estacionários desses sinais. Em um segundo momento, foram apontados métodos para a adaptação das ferramentas disponibilizadas pela Teoria do Caos nos sistemas existentes de RAL. A comprovação da eficácia dessas adaptações se deu pela realização de testes comparativos de exatidão que, de forma estatisticamente significativa, mostraram o benefício advindo das modificações sugeridas. A melhora obtida com o acréscimo de invariantes dinâmicas nos sistemas de RAL da forma proposta resultou na diminuição do EER em 17,65%, acarretando um intrínseco aumento de processamento. Para sinais de voz contaminados com ruído, o benefício atingido com o sistema proposto foi verificado para SNRs maiores que aproximadamente 5 dB. O restante deste capítulo resume o trabalho realizado e as principais conclusões descritas ao longo desta tese, enfocando as contribuições originais apresentadas e direções para trabalhos futuros.

5.1 Análise do trabalho desenvolvido e resultados

Esta tese apresentou e discutiu as idéias e o trabalho desenvolvido pelo candidato ao longo de seu doutoramento. Inicialmente foi descrito o problema abordado, que sucintamente consistia no desenvolvimento de técnicas para elevação da confiabilidade dos sistemas de RAL. Em um segundo momento, o estado-da-arte das técnicas atualmente disponíveis para o RAL foi apresentado e várias referências bibliográficas foram apontadas. Além disso, um estudo profundo a respeito das técnicas utilizadas para tratamento de sistemas dinâmicos não-lineares com comportamento caótico foi apresentado. Da mesma forma, muitas referências foram sugeridas para a melhor compreensão do tema. A seguir, foi discutida a possibilidade de utilização das ferramentas da Teoria do Caos em sinais de voz. Foi comprovada a presença de caos em quadros aproximadamente estacionários de voz, e foi mostrado exatamente como as invariantes dinâmicas podem caracterizar o locutor e contribuir para o aumento da

exatidão dos sistemas de RAL que as utilizam. Os testes realizados comprovaram, de forma estatisticamente significativa, a melhora do sistema de RAL utilizado quando invariantes dinâmicas não-lineares são adicionadas. Além disso, outras contribuições científicas obtidas ao longo do desenvolvimento dos trabalhos foram apresentadas.

O capítulo introdutório mostrou inicialmente algumas aplicações para os sistemas de RAL, evidenciando assim o imediato interesse em trabalhar com essa tecnologia e disponibilizá-la para o uso na sociedade. A seguir, foram apontados alguns dos principais problemas em aberto nessa área, salientando-se o grande esforço científico empregado na solução desses problemas. Foram também mostradas as abordagens existentes para reduzir o efeito dos fatores degradantes de exatidão dos sistemas de RAL, e foi sugerida a utilização de invariantes dinâmicas para a caracterização de locutores. Alguns dos principais veículos de divulgação científica nas áreas de RAL e Teoria do Caos foram listados e, ao final, os tópicos cobertos ao longo deste trabalho foram mencionados.

O segundo capítulo concentrou-se principalmente nas técnicas empregadas para efetuar o reconhecimento de pessoas pela voz, desde a gravação do sinal até o fornecimento de um resultado. Inicialmente, os diferentes tipos de sistemas de RAL foram discutidos. A seguir, foi mostrada uma visão geral do funcionamento desses sistemas, onde foram evidenciadas as principais etapas englobadas. Depois disso, a aquisição do sinal de voz foi discutida, envolvendo a conversão desse sinal para o domínio digital. Quando a representação digitalizada da voz está disponível, há uma redução no volume de dados através da extração da informação útil desse sinal. Isso é feito utilizando-se um modelamento matemático para o processo de produção da fala, pré-ênfase do sinal e análise de quadros de voz de curta duração (tipicamente com 30 ms). Em cada quadro de voz podem ser extraídos parâmetros distintos, entre os quais os coeficientes *mel*-cepstrais vêm se mostrando altamente eficazes para a diferenciação de locutores. A comparação ou medida de distorção entre parâmetros provenientes de sinais de voz diferentes pode ser feita através de técnicas distintas. Foi escolhida a medida de distorção de Bhattacharyya para a realização de testes, pela sua representatividade estatística, pequeno poder de processamento requerido, simplicidade de compreensão e ausência de ajustes a serem feitos, garantindo que as modificações na exatidão do sistema devem-se exclusivamente às informações utilizadas.

Ao longo do terceiro capítulo, foram introduzidos os principais conceitos e técnicas utilizadas para tratamento de sinais dinâmicos não-lineares com comportamento caótico. O tratamento de séries temporais obtidas a partir de dados experimentais foi aprofundado, uma vez que a voz é uma série temporal. Os diferentes tipos de atratores foram mostrados, e o método mais difundido para a reconstrução da trajetória de atratores associados a séries temporais foi discutido: o método dos atrasos temporais. Os cuidados que se deve ter para garantir uma reconstrução da trajetória adequada foram mencionados, assim como técnicas para a escolha de valores apropriados para o passo de reconstrução e dimensão de imersão. Assim, garante-se que o atrator reconstruído é topologicamente igual ao atrator real, dessa forma apresentando características idênticas, estimadas através de invariantes dinâmicas. As técnicas utilizadas para a estimação de algumas importantes invariantes dinâmicas foram mostradas. Também o cálculo da dimensão de correlação foi discutido, utilizando-se o algoritmo de Grassberger-Procaccia, além dos métodos de Wolf e Rosenstein para a estimação dos expoentes de Lyapunov. O método de Rosenstein mostrou-se mais interessante para a aplicação em atratores reconstruídos a partir de sinais de voz, por necessitar de uma quantidade menor de amostras para fornecer valores confiáveis.

O quarto capítulo descreve a proposta de tese em si, que associa as áreas de RAL e Teoria do Caos, propondo uma metodologia para utilização de informações dinâmicas não-lineares, conjuntamente com informações advindas da análise espectral tradicional. Inicialmente, foram apontados os trabalhos encontrados na literatura que utilizam a análise dinâmica não-linear em sinais biológicos, sem que houvesse qualquer trabalho diretamente relacionado à proposta de aplicação das técnicas em sistemas de RAL. A conexão entre as técnicas de RAL atuais e a análise dinâmica não-linear da voz está na utilização das invariantes dinâmicas estimadas como padrões que caracterizem o locutor. Assim, o atrator associado à série temporal vocal pode ser avaliado através da estimativa de suas invariantes dinâmicas. Dessa forma, as invariantes dinâmicas estimadas podem fornecer uma estimativa fisiológica do sistema produtor de voz. Com isso, seria possível aplicar tais parâmetros a um classificador, de forma a reconhecer a identidade de uma pessoa através da análise dos parâmetros obtidos a partir da avaliação dinâmica não-linear. Inicialmente foi realizada uma análise em quadros de voz de curta duração, considerados aproximadamente estacionários, a fim de verificar a existência de componentes caóticas na série temporal. Pela avaliação do maior expoente de Lyapunov foi possível verificar a presença de caos na maioria dos quadros de voz. Foram realizados vários testes de eficiência em um sistema de RAL, utilizando um banco de vozes composto de 50 locutores distintos, objetivando a posterior análise de significância estatística nos resultados. Esses testes variaram alguns parâmetros associados às informações utilizadas para caracterização dos locutores, e avaliaram comparativamente a exatidão do sistema ao acrescentar as invariantes dinâmicas dimensão fractal, dimensão de correlação e maior expoente de Lyapunov. O sistema proposto apresentou redução de 17,65% no EER, de forma estatisticamente significativa. Ao final foi discutida a questão do intrínseco aumento de necessidade computacional, que foi mais acentuado para a estimação da dimensão de correlação e expoentes de Lyapunov. Verificou-se que essa penalização no tempo de resposta do sistema de RAL inviabilizaria a utilização de invariantes dinâmicas não-lineares em sistemas que necessitem apresentar o resultado da análise vocal de forma rápida (em poucos segundos), quando quadros com mais de aproximadamente 1000 amostras forem utilizados.

O trabalho apresentado mostrou que o sinal de voz pode ser visto como proveniente de um sistema caótico determinístico, de acordo com o tratamento descrito nos experimentos realizados. Isso permite a utilização de vários conceitos e técnicas de análise de sistemas caóticos em sinais de voz, possibilitando a inserção de informações antes ignoradas por sistemas de processamento digital de sinais de voz, não se limitando apenas aos sistemas de RAL. Espera-se, assim, que o avanço da ciência na área de Teoria do Caos implique também no avanço dos sistemas que utilizam essas ferramentas para a caracterização do sinal de voz.

5.2 Contribuições originais

Durante o desenvolvimento dos trabalhos que resultaram nesta tese, algumas contribuições originais puderam ser detectadas. Essas contribuições foram detalhadas ao longo dos capítulos anteriores e são resumidamente mostradas a seguir.

5.2.1 Caracterização de locutor através de invariantes dinâmicas não-lineares

Uma das principais contribuições que este trabalho apresenta é possibilitar a caracterização de locutores utilizando medidas de invariantes dinâmicas não-lineares, de forma a comprovadamente aumentar a exatidão do sistema de RAL que utilizar esse novo tipo de informação. A proposta discutida sugere a avaliação das amostras de voz como provenientes de um sistema dinâmico não-linear com comportamento caótico que evolui em um espaço de fases. Dessa forma, foram apontadas as ferramentas apropriadas para o tratamento desse tipo de sistema, e as adaptações que devem ser feitas para que elas forneçam informações confiáveis. Basicamente, procede-se a reconstrução da dinâmica de quadros aproximadamente estacionários do sinal vocal e análise topológica do atrator associado. Assim, o sinal temporal é primeiramente mapeado ao espaço de fases de dimensão apropriada, e invariantes dinâmicas são extraídas desses atratores. Essas medidas possuem a capacidade de diferenciação entre locutores distintos, e quando são agregadas aos parâmetros atualmente utilizados nos sistemas de RAL, resultam no aumento da exatidão desses sistemas. Isso mostra que esse tipo de informação não está presente nos parâmetros atuais e ocasionam uma melhora genuína. Essa contribuição foi explorada nos trabalhos [PET 2000c] [PET 2001] [PET 2002a] [PET 2002b] [PET 2002c].

5.2.2 Extração de invariantes dinâmicas a partir de quadros estacionários de sinais de voz, extrapolando o conceito de TDFD

A estimativa de invariantes dinâmicas não-lineares em sinais de voz não pode ser aplicada diretamente em todo o sinal, obtendo assim um único valor que represente a locução. Isso não pode ser feito pela intrínseca não-estacionariedade dos sinais de voz, e conseqüente variação de suas configurações e parâmetros ao longo do tempo. Assim, aproveitando-se o conceito de quadros aproximadamente estacionários de voz, atualmente já empregados para análise espectral do sinal vocal, foi proposto o conceito de estimação de parâmetros dinâmicos não-lineares dependentes do tempo. Assim, a caracterização do sinal de voz através de informações dinâmicas não-lineares pode ser efetuada, fornecendo um conjunto de valores para as estimativas correspondentes que poderão modificar-se ao longo da locução. Essa contribuição foi explorada nos trabalhos [PET 2001] [PET 2002a] [PET 2002b] [PET 2002c].

5.2.3 Contribuição ao CEM particularmente para sinais de voz

O algoritmo utilizado para a estimação da dimensão fractal a partir de séries temporais foi o CEM, amplamente empregado na literatura. Essa técnica modela o espectro de potências do sinal sob análise como uma exponencial decrescente. Para a estimação do valor da dimensão fractal em sinais de voz, devem ser desconsideradas freqüências inferiores a uma freqüência crítica k_c, abaixo da qual o espectro de potências $P(u)$ não aproxima o modelo da equação 3.16 [NAK 93]. Fazendo-se $u=k/kc$, onde k é a freqüência real, essas baixas freqüências são corretamente desprezadas. Entretanto, a estimativa dos valores apropriados para k_c é feita heuristicamente, através da análise do espectro de potências do sinal, visualmente definindo o valor correto. Neste trabalho, sugerimos uma forma original para determinação automática de uma boa aproximação para k_c baseado em uma representação suavizada da densidade de

energia espectral. Esse método está detalhado na publicação [PET 2002a] e foi também utilizado no trabalho [PET 2001].

A figura 5.1 mostra uma típica resposta em freqüência a partir de uma quadro de voz de 30 ms de duração, obtida através do algoritmo da transformada rápida de Fourier (FFT). É importante notar que, a partir de uma determinada freqüência, a energia espectral diminui, seguindo aproximadamente a modelagem mostrada na equação 3.16. Assim, para uma representação suavizada da energia espectral, é possível realizar uma busca pelo máximo e considerá-lo a freqüência crítica k_c, acima da qual um decréscimo aproximadamente exponencial é verificado. A utilização de uma representação suavizada da resposta em freqüência, em substituição à escolha da freqüência de maior magnitude no quadro, considera toda a evolução do sinal e evita a escolha baseando-se apenas em um (possivelmente incorreto) valor. Além disso, a representação suavizada é capaz de fornecer a freqüência exata onde a magnitude pára de crescer e inicia o decréscimo, evitando picos locais e isolados. A representação suavizada do contorno da densidade de energia espectral pode ser obtida utilizando o espectro de predição linear [DEL 87] [RAB 78], obtido a partir da estimação dos coeficientes de predição linear (LPC). A figura 5.2 mostra os espectros preditivos lineares para ordens diferentes do mesmo sinal de voz da figura 5.1. Um algoritmo simples de busca pelo pico pode facilmente encontrar uma boa aproximação para k_c.

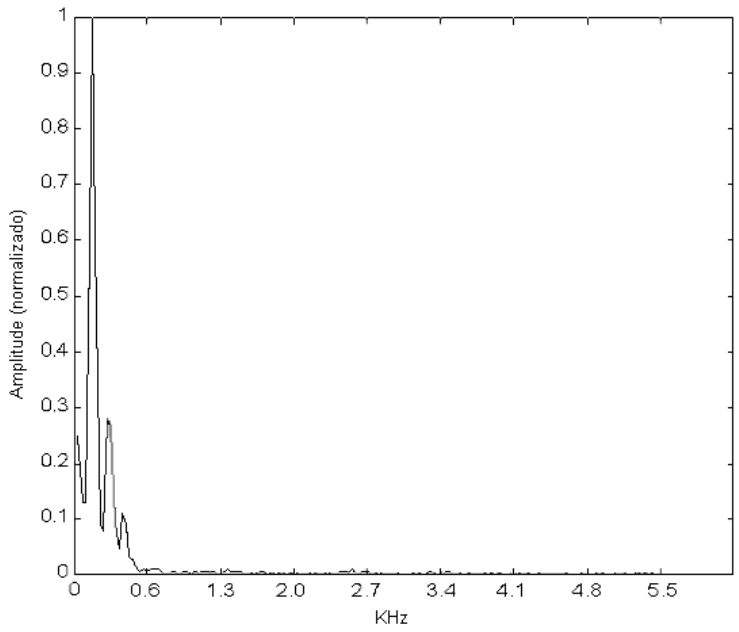

FIGURA 5.1 - Resposta em freqüência a partir de um quadro sonoro de voz de 30 ms de duração

Os valores de dimensão fractal resultantes da busca automática pela freqüência k_c são bem próximos aos resultados reportados em [SAB 96] para sinais de voz. A figura 5.3 (b) mostra os valores da TDFD a partir do sinal de voz mostrado na figura 5.3 (a), avaliando-se quadros de 30 ms de duração, extraídos a cada 10 ms.

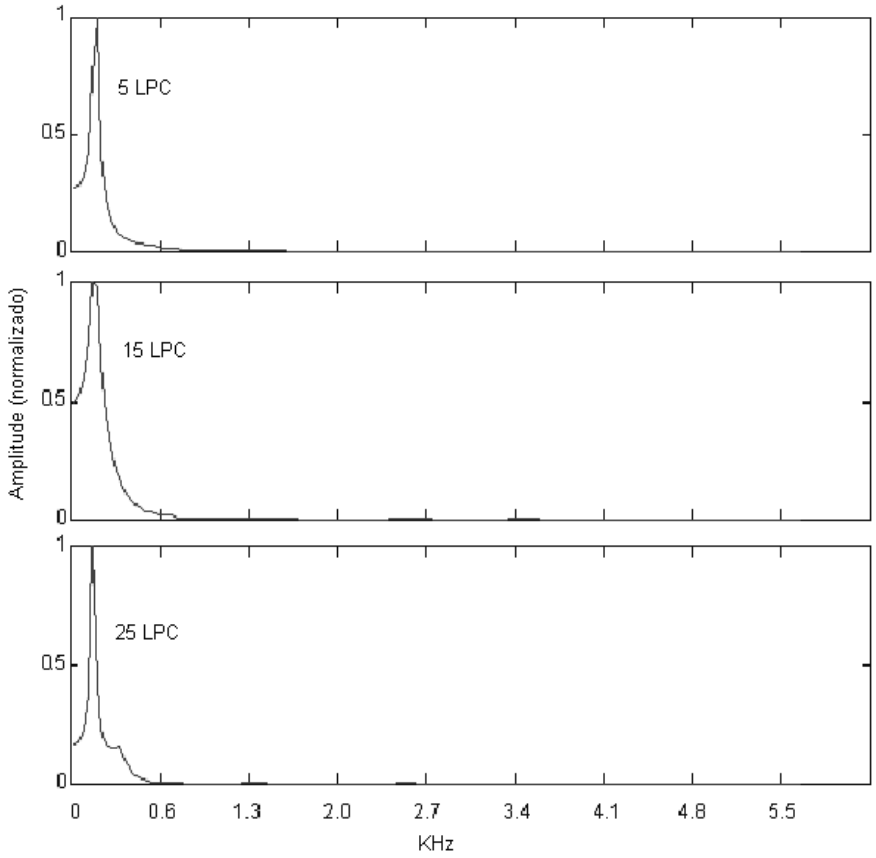

FIGURA 5.2 - Resposta em freqüência suavizada

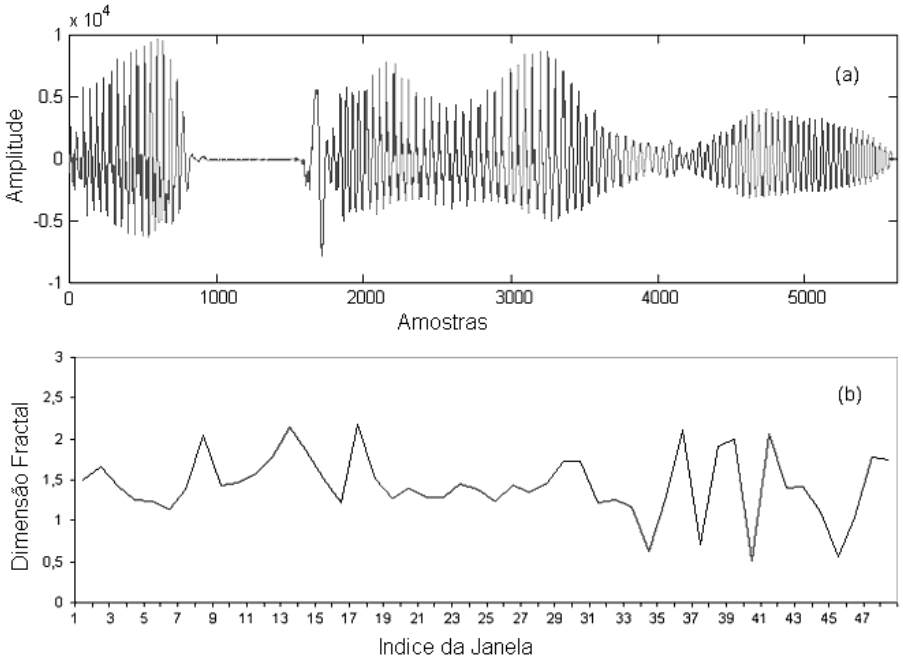

FIGURA 5.3 - TDFDs para uma sinal de voz

5.3 Trabalhos futuros

O trabalho apresentado oferece a possibilidade de exploração de diversas áreas, no sentido de continuar a avançar no desenvolvimento de técnicas da Teoria do Caos visando aplicação em sinais de voz. Inicialmente, pode-se pensar em explorar a utilização de outras invariantes dinâmicas não-lineares nos sistemas de RAL, e mesmo invariantes que venham a ser evidenciadas futuramente. A questão da redução do tempo de processamento requerido para fornecimento de estimativas robustas para invariantes dinâmicas não-lineares é, da mesma forma, uma área de atuação futura que se apresenta e visa viabilizar a utilização dessas informações em sistemas práticos. Outra área que potencialmente poderá ser estudada no futuro trata o desenvolvimento de técnicas específicas para possibilitar estimativas confiáveis de invariantes dinâmicas a partir de sinais de voz contaminados com ruído. A aplicação do trabalho desenvolvido em sistemas de RAF poderá, também, beneficiar tais sistemas, necessitando-se ajustes e desenvolvimento de técnicas específicas.

5.3.1 Utilização de outros tipos de invariantes dinâmicas não-lineares

A pesquisa que pode ser desenvolvida a partir do trabalho mostrado nesta tese é vasta. Inicialmente, pode-se pensar em utilizar outros tipos de invariantes dinâmicas para a caracterização de locutores.

Alguns exemplos de invariantes dinâmicas não exploradas ao longo desta tese são os expoentes de Lyapunov de mais alta ordem, dimensões de ordem superior, entropia ou quaisquer outras invariantes dinâmicas que venham a ser válidas. Dessa forma, a utilização das técnicas discutidas para possibilitar a utilização de estimativas de invariantes dinâmicas a partir de segmentos estacionários de voz podem ser aplicadas em outros tipos de invariantes dinâmicas, e até mesmo em invariantes que venham a ser desenvolvidas futuramente.

É provável que a caracterização de locutores através de outras invariantes dinâmicas venha a ser melhorada, possibilitando o desenvolvimento de sistemas de RAL cada vez mais robustos e confiáveis.

5.3.2 Redução do tempo de processamento requerido

A otimização do tempo de processamento requerido para a estimação robusta das invariantes dinâmicas é um fator que pode ser explorado em trabalhos futuros. Isso possibilitará a utilização de informações que atualmente não podem ser utilizadas na prática em sistemas que necessitem resultados em um curto espaço de tempo.

Os algoritmos utilizados para a caracterização de locutores através de invariantes dinâmicas, apesar de contribuírem para o correto reconhecimento, podem vir a impedir sua utilização prática devido à necessidade de tempo de resposta do sistema. Por exemplo, para um sistema que avalia um padrão vocal em casos forenses, essa necessidade de processamento não é crítica, pois a conclusão pode ser estabelecida durante um período de avaliação relativamente longo (da ordem de dias). Já em sistemas de controle de acesso a áreas restritas, onde o fluxo de pessoas que utilizam o sistema é grande, o tempo de resposta do sistema não pode ser elevado (tipicamente da ordem de segundos). Assim, a concepção de hardware que satisfaça a necessidade de processamento dinâmico não-linear em tempo hábil poderá permitir a utilização dessas técnicas mesmo em sistemas que possuam curto tempo de resposta.

5.3.3 Redução da influência do ruído na estimativa de invariantes dinâmicas não-lineares

A exploração de técnicas para trabalhar com sinais contaminados com ruídos, otimizando os resultados alcançados, poderá também beneficiar os sistemas de RAL. A particularização dessas técnicas visando a estimativa de invariantes dinâmicas não-lineares é possível e representa uma importante área de atuação.

A questão do tratamento do sinal visando eliminação de fatores degradantes, como o ruído, vem sendo foco de intensas pesquisas. Para a análise dinâmica não-linear, as necessidades e os métodos para tratamento do sinal podem ser distintos, acarretando o desenvolvimento de algoritmos e técnicas específicas e eficientes para esse tipo de análise. O benefício advindo desse desenvolvimento científico é imediato para os sistemas que utilizarem invariantes dinâmicas para a caracterização de locutores, possibilitando seu uso em ambientes ruidosos e com interferências indesejadas, tanto em nível do dispositivo de captura da voz, quanto do canal de transmissão.

5.3.4 Utilização de informações dinâmicas não-lineares em sistemas de RAF

Outra área que poderá se beneficiar dos avanços obtidos e mostrados neste trabalho é a área de reconhecimento de fala, onde se deseja identificar o texto falado, e não o falante. Nesse tipo de sistema, a introdução da análise dinâmica não-linear poderá acarretar um aumento na eficiência dos sistemas existentes, uma vez que o acréscimo de informações ainda não consideradas tenderá a melhor caracterizar o sinal avaliado.

As áreas de RAL e RAF possuem algumas semelhanças, apesar de utilizarem muitas vezes técnicas completamente distintas, uma vez que os objetivos de cada área são diferentes. Por outro lado, alguns aspectos são bastante semelhantes, e a questão da redução no volume de dados, preservando-se as principais informações do sinal, é importante em ambas áreas. Assim, é possível que a introdução da análise dinâmica não-linear nos sistemas de RAF possa trazer benefícios que poderão ser futuramente investigados.

Anexo 1 Bhattacharyya distance applied to speaker identification

Trabalho intitulado "**Bhattacharyya Distance Applied to Speaker Identification**", publicado no veículo The International Conference on Signal Processing Applications and Technology 2000, com autoria de Adriano Petry, Adriano Zanuz e Dante A. C. Barone. A formatação original do trabalho foi modificada.

BHATTACHARYYA DISTANCE APPLIED TO SPEAKER IDENTIFICATION

Adriano Petry, Adriano Zanuz, Dante Augusto Couto Barone

Universidade Federal do Rio Grande do Sul – Instituto de Informática
Av. Bento Gonçalves, 9500 bloco IV – Porto Alegre – RS – Brazil

ABSTRACT

This paper describes a new method to perform a speaker identity classification. The proposed method is based in a probabilistic distance measure – the Bhattacharyya distance. First the speech samples are pre-emphasized, then 16 LP-derived cepstral coefficients are extracted and a lifter window is applied to these feature parameters. Analyzing the feature histograms, it is possible to model them by gaussian probability densities. So the distance measure proposed by Bhattacharyya is applied in way to effectively classify an unknown speech sample. This method have been presented high accuracy, when tested with over 3s speech samples. The speech samples have been recorded from 20 different speakers.

1. INTRODUCTION

Speaker recognition is the process of identifying a person automatically using the information that his/her speech contains. Speaker recognition can be classified into speaker identification and speaker verification. In speaker identification, the unknown speech sample must be related to one of the registered speakers. The speaker verification process only accepts or not the unknown speech sample as belonging to a claimed speaker.

The overall speaker recognition process can be divided in three main phases: pre-processing, feature parameters extraction and classification. The pre-processing phase is responsible for the first transformation in the speech signal. It has the objective of eliminate the natural frequency distortions from speech signals and divide them in small pieces that can be considered stationary - the frames. The feature parameters extraction phase aims to separate the more significant information from the frames. The classification phase compare different speech signals and identify the owner of an unknown speech sample.

This work presents a new method used in classification phase, based in the Bhattacharyya distance [1][2]. The Bhattacharyya distance provides a way to measure the proximity between two distinct observations groups, modeled by probability distributions. This technique is used to perform text-dependent speaker identification.

The idea of using Bhattacharyya distance in speaker recognition is a result of an inspection about the *p*-dimensional speech vectors, generated in the feature parameters extraction. Analyzing separately the dimensions of some vectors its possible to model the histograms as gaussian probability densities. Due to this, considering each class of vectors universe as speakers representations, it has been used the Bhattacharyya distance to compare speech signals from a speakers group to validate its applicability.

The developed system work in two operational modes: training and recognition. In the training, a group of speech samples from each speaker is used to obtain their respective parameters (LP-derived cepstral coefficients). These parameters are stored to compose the registered speakers data. In the recognition, unknown speech samples are

processed in the same way to obtain the representative parameters. Then, the Bhattacharyya distance is used to evaluate the proximity between the unknown speech and each registered speaker parameters. Finally, it is selected by the minimal function the identified speaker. Figure1 shows the developed system block diagram.

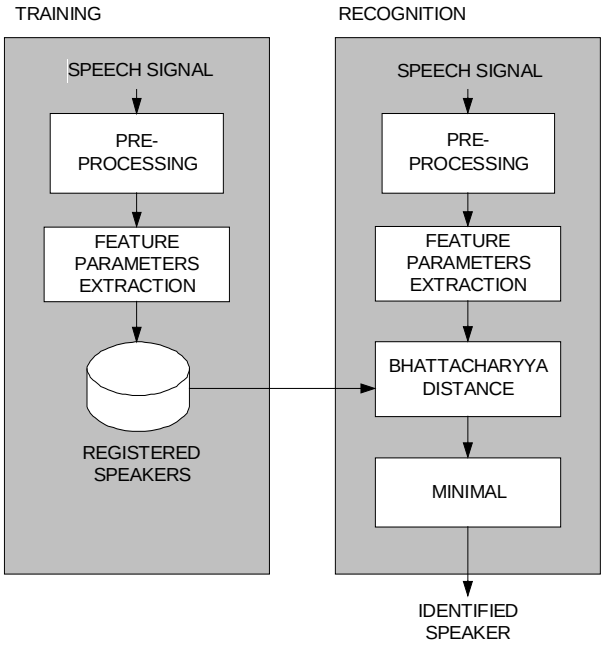

Figure 1: Developed System Block Diagram

2. PRE-PROCESSING AND FEATURE PARAMETERS EXTRACTION

After a word acquisition, the speech sample is pre-emphasized to cancel a spectral distortion introduced by the lips [3][4]. A first order FIR filter can give the desired spectral response of +6dB/octave. The FIR filter was implemented as:

$$y(n) = x(n) - 0{,}95x(n-1) \qquad 1 \le n < M$$

where M is the number of samples in the speech signal $x(n)$, and $y(n)$ is the pre-emphasized signal.

After that, a 30ms hamming window is applied every 10ms of speech signal. The feature parameters are extracted from each pre-emphasized and windowed piece of speech signal, called a speech frame. These short-term parameters can provide an effective discrimination among speakers.

One of the most commonly used short-term spectral measurements is Linear Prediction-derived (LP-derived) Cepstral Coefficients [4][5][6]. The mel-cepstral coefficients, with first and second order derivatives coefficients are also used. The results obtained using these different coefficients are similar.

Linear Prediction analysis is an important method of characterizing the spectral properties of speech in the time domain. In this analysis method, each sample of the speech signal is predicted as a linear weighted sum of the past p samples. The weights which minimize the mean-squared prediction error are called the predictor coefficients. The value of p is approximately determined by the number of poles of the vocal tract and the glottal wave transfer function, mathematically modeled.

By definition, the cepstrum (or the Cepstral Coefficients) is the inverse Fourier transform of the logarithm of the speech signal spectrum. The relationship between the cepstrum and the predictor coefficients are[6]:

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k} \quad 1 \le m \le p$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k} \qquad m > p$$

The Cepstral Coefficients obtained from the predictor coefficients are called LP-derived Cepstral Coefficients. A weighting window (named lifter window) is usually applied to the LP-derived Cepstral Coefficients. It is done to emphasize the coefficients which contain more spectral information. This window is described as:

$$w(i) = 1 + \frac{K}{2} sin\left(\frac{i\pi}{K} \right) \qquad 1 \le i \le p$$

where K is a constant usually set to 22 called lifter coefficient.

An utterance is composed by the feature parameters extracted from each one of the frames in a speech sample.

3. CLASSIFICATION USING BHATTACHARYYA DISTANCE

3.1 The Bhattacharyya Distance

In statistics, the proximity degree between two different probability densities is related with the notion of distance measure. One way to calculate this measure is throughout the Bhattacharyya distance. This distance measure is often easier to evaluate than the divergence [1].

Before discussing the speaker identity problem, it will be introduced the Bhattacharyya distance. Considering two probability densities $p_1(x)$ and $p_2(x)$, obtained from two different classes of vectors. The Bhattacharyya distance is defined by

$$B = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)\, p_2(x)}\, dx$$

Special cases of this general distance measure can be calculated explicitly to a large types of probability densities. An important case refers to the multivariate gaussian distributions. Considering $p_i(x)$ as gaussian probability densities, it is possible to show that equation 5 can be written as:

$$B = \frac{1}{8} (m_1 - m_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (m_1 - m_2) +$$

$$+ \frac{1}{2} \ln \left(\frac{\det(\Sigma_1 + \Sigma_2)/2}{\sqrt{\det(\Sigma_1)} \sqrt{\det(\Sigma_2)}} \right)$$

where m_i is the mean value and Σ_i is the covariance matrix from density i.
Since gaussian observations are common in nature, the equation 6 is specially important. It happens due to the limit central theorem.

3.2 Speaker identification

The speaker identification systems are divided into pre-processing, feature parameters extraction and classification. An important phase in a speaker identification system is the classification, where the useful information extracted from the speech

samples are used to provide a measure of similarity among different speakers. Some traditional techniques used to accomplish the classification phase are Vector Quantization [8][9][10], Hidden Markov Models [5], Neural Networks [4][5] and Dynamic Time Warping [11].

A new way to perform the classification is presented in this work. In this method, the Bhattacharyya distance is used to evaluate the proximity between an unknown speech signal and every trained speaker. The feature parameters, extracted from the trained speakers speech frames in the previous phase, are grouped in classes corresponding to each speaker. The feature parameters extracted from the unknown speech sample compose the unknown class. The unknown class and every trained speaker class are compared, two by two, through the Bhattacharyya distance measuring. Once this technique is used in a close set system, it is possible to identify the correspondent speaker by comparing the distances obtained and selecting the minimal value. (If the target is an open set system, it is necessary to establish an *a priori* threshold).

The Bhattacharyya distance can be applied to a wide variety of known probability distributions, according to the best fit [1]. The feature parameters used in this work are composed by *p*-dimensional LP-derived cepstral coefficients. Inspecting histograms obtained from these feature parameters, it is possible to see that their value distributions can be modeled as gaussians probability densities. Figure 2 shows an example of four different histograms. They have been obtained analyzing the same dimension of the LP-derived cepstral coefficients from two different speakers. It is possible to note the gaussian shape of all curves, and a significant difference for different speakers.

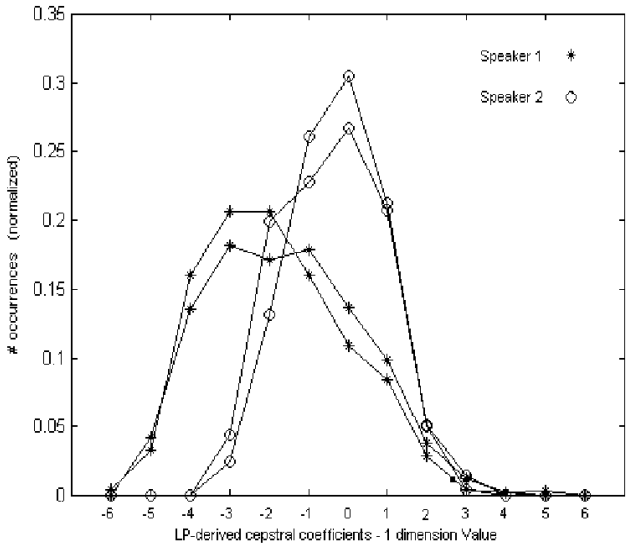

Figure 2. Histograms using different speakers feature parameters.

It is important to emphasize that the Bhattacharyya distance considers differences in both mean value and variance, from the samples provided for the two sets of vectors. Thus, the figure 2 would present high distance between the speakers because not only the mean value is different, but also because it is clear the variance difference.

4. EXPERIMENTAL RESULTS

The speech samples used for training and testing were recorded using 11025 Hz sampling rate, 16 bits per sample. A weighting window (lifter) was applied to the 16 LP-derived cepstral coefficients, using lifter coefficient equals to 22. The signal was previously pre-emphasized, and a Hamming window with the width of 20ms was

applied every 10ms. It was recorded speech samples from 20 different speakers. Each one of them spoke the word “floor”, spoken in the Portuguese language, 20 times. The first 10 samples of each speaker were used to generate the trained speakers classes. The others 10 samples were used as unknown, and the system should classify them correctly. In the tests accomplished, it was varied the medium amount of speech used as unknown, by using a single sample, two samples, and so on. The median speech time was 502 ms per sample. Figure 3 shows the system recognition rate.

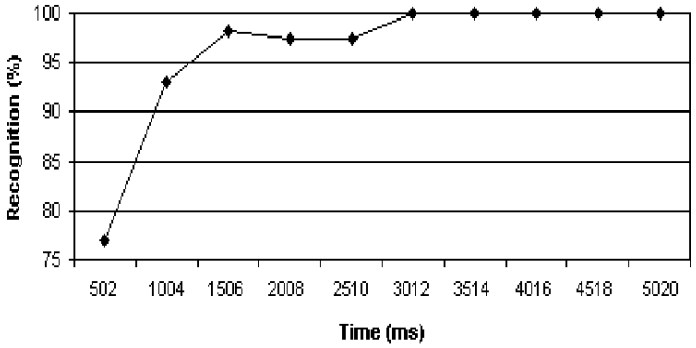

Figure 3. System performance, varying the amount of speech used as unknown.

Note that the recognition accuracy increases, if more data is provided. Using more than 3s of unknown speech, no mistakes are made.

Another interesting way of evaluate the method consistency is to measure the difference between the Bhattacharyya distances using the correct speaker and the others speakers. If the correct speaker distance is well separated from all others, the classification can be successfully accomplished. Figure 4 shows the difference between the correct speaker and the lowest incorrect distances, varying the amount of speech used as unknown.

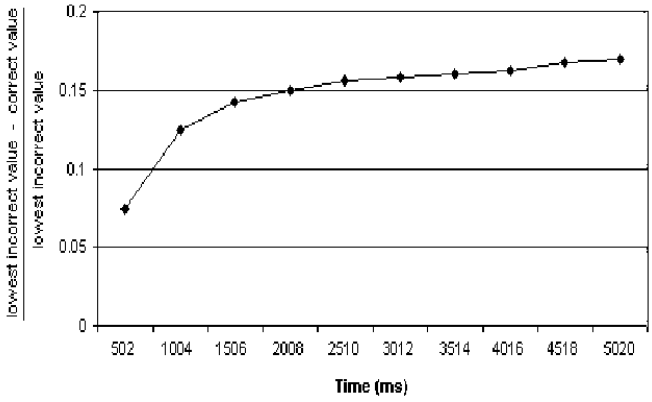

Figure 4. Diference between the correct speaker and the lowest incorrect distances, varying the amount of speech used as unknown.

It is important to salient that even after achieving a recognition rate of 100% (after 3s of speech), this difference still grows, increasing the system reliability.

5. CONCLUSIONS

It has been presented a new probabilistic method to classify unknown speech sample in a speaker recognition system. This method uses the Bhattacharyya distance applied to the feature parameters extracted from the speech.

Through the speech histogram analysis it is possible to conclude that not all feature parameters components contribute significantly to a correct classification. However, the small individual contributions are accumulated, which provides a good distinction. This is exactly the target of using Bhattacharyya distance.

The more speech is used, the greater difference is established. This is due to an intrinsic probability characteristic, where a great deal of sample data provides more precise results.

6. ACKNOWLEDGMENTS

The authors would like to thank Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for their financial support, necessary to accomplish this work.

7. REFERENCES

[1] T. Kailath, “The Divergence and Bhattacharyya Distance Measures in Signal Selection,”. *IEEE Trans. On Communication Technology*, vol. Com-15, pp. 52-60, february 1967.

[2] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bull Calcutta Math. Soc*., vol. 35, pp. 99-109, 1943.

[3] L. R. Rabiner and R. W. Schafer, “Digital Processing of Speech Signals,” *Prentice Hall*, 1978.

[4] L. Rabiner and Biing-Hwang Juang, “Fundamentals of Speech Recognition,” *Prentice Hall*, 1993.

[5] J. R. Deller, Jr. and J. G. Proakis and J. H. L. Hansen, “Discrete-time Processing of Speech Signals,” *Prentice Hall*, 1987.

[6] J. W. Picone, “Signal Modeling Techniques in Speech Recognition,” *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215-1247, september 1993.

[7] P. E. Papamichalis, “Pratical Approaches to Speech Coding,” *Prentice Hall*, 1987.

[8] A. Petry and D. A. C. Barone, “Speaker Verification Applied to an Elevator Safety System,” in Proc. of *International Conference on Signal Processing Applications and Technology - ICSPAT*, 1999.

[9] J. Makhoul, S. Roucos and H. Gish, “Vector Quantization in Speech Coding,”. *Proceedings of the IEEE*, vol. 73, no. 11, november 1985, pp. 1551-1587.

[10] R. A. Finan and A. T. Sapeluk and R. I. Damper, “VQ Score Normalisation for Text-dependent and Text-independent Speaker Recognition”. Proceedings of the *First International Conference on Audio- and Video-based Biometric Person Authentication*, 1997, pp. 211-218.

[11] H. Sakoe and S. Chiba, “Dynamic Programming Algorithm Optimization for Spoken Word Recognition,” *Proceedings of the IEEE*, 1978.

Anexo 2 Speaker identification using nonlinear dynamical features

Trabalho intitulado "**Speaker Identification Using Nonlinear Dynamical Features**", publicado no veículo Chaos, Solitons & Fractal, v. 13/2, p.221-231, February 2002, com autoria de Adriano Petry e Dante A. C. Barone. A formatação original do trabalho foi modificada.

Speaker Identification Using Nonlinear Dynamical Features

ADRIANO PETRY and DANTE AUGUSTO COUTO BARONE

Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

Abstract – This paper reports results obtained in a speaker identification system which combines commonly used feature parameters, such as LP-derived cepstral coefficients and pitch, with nonlinear dynamic features, namely, fractal dimension, entropy and largest Lyapunov exponent, extracted from every window of 30ms of speech, applied every 10ms. The corpus used in the tests is composed of 37 different speakers, and the best results are obtained when the nonlinear dynamic features are included, suggesting that the information added with these features was not present in the others so far.

1. INTRODUCTION

The digital speech processing field has been focus of intensive research lately. One of the reasons is the practical aspects involved, i.e., the possibility of developing systems with immediate real world applicability, such as code and enhance speech corrupted by noise, speech synthesis, automatic transcription of sentences and identification of a person from his/her speech. The technology used to accomplish speaker recognition can be used to verify the identity of people accessing systems, shopping, banking and also for forensic purposes [1]. Depending on the target application, a very high level of accuracy may be necessary.

The area of speaker recognition can be classified into speaker verification and identification. Speaker verification is the process of accepting or rejecting the identity claimed by a speaker, through the analysis of speech samples from that speaker. In speaker identification, the unknown speaker's speech samples are compared with all known speaker templates, and his identity is assigned to the known speaker whose template best matched. The speaker verification and identification can be accomplished using speech samples with the same text for both training and recognition (text-dependent), or a totally unconstrained vocabulary (text-independent).

A speaker identity is strongly dependent of the physiological and behavioral characteristics of the speech production system. The first step of a basic speaker recognition system is to extract from the speech samples a "good" parametric representation. These parameters must be, as much as possible, representative of a speaker, presenting low variability for that speaker's speech samples, and great difference when used with others speakers' speech samples. A mathematical speech production model have been extensively used [2][3], which combines a periodic and a stochastic source of excitation applied to a time-varying linear filter (vocal tract model). Some techniques were developed and can estimate different speaker-dependent parameters from this model. The fundamental frequency estimation of the periodic excitation (pitch) can provide information related mainly to the speaker vocal tract physiology. Others features attempt to optimally model the spectrum as an autoregressive process, providing the linear predictive coefficients (LPC). The cepstral

coefficients computed directly from the LPC are also very used to speaker recognition, and they are called LP-derived cepstral coefficients. Many important features can be extracted too, but it is important not to mix features whose information it carries is not interesting for speaker recognition task, or is not new when compared with the others features used. All these feature parameters combined can be applied with template-matching techniques to successfully distinguish speakers.

Previous papers [4][5][6][7] have worked with speech characterization and analysis using nonlinear dynamical features. Sabanal *et al*. [5] used the time-dependent fractal dimensions (TDFDs), extracted through critical exponent method (CEM), and the time-dependent multifractal dimensions (TDMFDs) to accomplish a speech recognizer. The target was to recognize Japanese digits using a neural network. Kumar *et al*. [4] estimated Lyapunov exponents, dimension and metric entropy in phonemes signals, divided into eight different types. Banbrook *et al*. [6] extracted correlation dimension, Lyapunov exponents, and short-term predictability from a corpus of sustained vowels sounds. The works mentioned used some nonlinear dynamical features to characterize a speech sound, showing the speech low dimensionality and the average exponential divergence of nearby trajectories in the reconstructed phase space.

In this work a text-dependent speaker identification is performed using a combination of commonly used feature parameters, namely pitch and LP-derived cepstral coefficients, with nonlinear dynamic invariants. While the use of standard feature parameters can perform a speaker recognition quite successfully, it may not be accurate enough for some applications. We believe a speaker recognition system performance can be improved. The characterization of speech using a nonlinear dynamic description can help on identifying people from their voices. The assumptions used to extract the standard feature parameters do not describe the nonlinear dynamic evolution of the system. It will be shown that add nonlinear dynamic qualitative information as fractal dimension, metric entropy and largest Lyapunov exponent to the standard feature parameters, is equivalent to add speaker-dependent features, not present in the standard feature parameters so far. This combination will lead a speaker recognition system to more accurate results.

The paper is organized into an introduction, a section describing the phase espace reconstruction followed by sections describing the methods used to extract nonlinear dynamic features. After that, the speaker recognition system used in this work is detailed, as well as the speech data set used to accomplish the tests. Then, experiences involving standard features and nonlinear features are showed. At the end, the performance analysis of the system is discussed.

2. PHASE ESPACE RECONSTRUCTION

In experimental applications, it is often available unidimensional measurements of a dynamical system that evolves in a multidimensional phase space. This scalar time series contains the information available from that system. In many cases, no further information is available, and an important challenge that has to be solved is the calculation of the system's real multidimensional phase space trajectory. After that, measurements that provide important knowledge about the system behavior can be done.

To evaluate the properties of an attractor associated to a time series it is first necessary to reconstruct its evolution in a proper phase space. The most used way of reconstructing the full dynamics of a system from scalar time series measurements was

proposed by Takens [8]. This method presents easy practical implementation. Given a *N*-point time series $x(t_i)$ for $i=1,2,...N$ as follows

$$x(t) = \{x(t_1), x(t_2), ..., x(t_N)\},$$

the *m*-dimensional vectors are reconstructed, according to Takens delay method [8], as

$$\vec{X}_i = \{x(t_i), x(t_i + p), x(t_i + 2p), ..., x(t_i + (m-1)p)\},$$

where *p* is called time delay and *m* is the embedding dimension. The $\vec{X}_i$ vectors represent the trajectory of the time series $x(t_i)$ in a *m*-dimensional phase space.

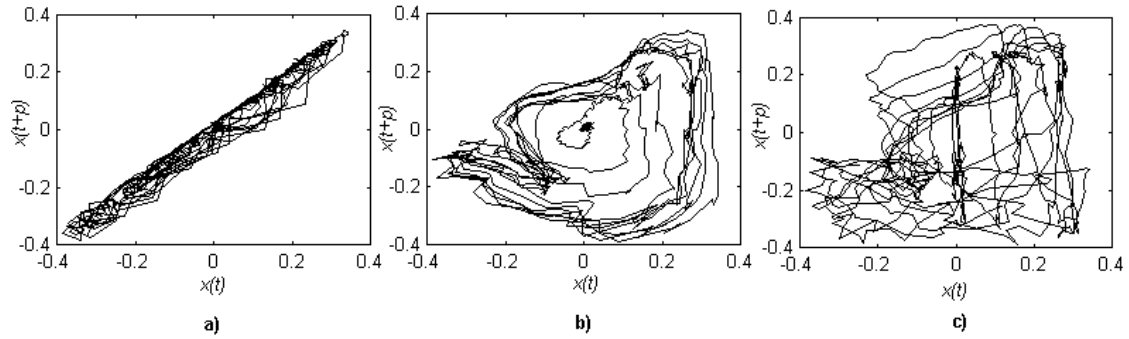

Fig. 1. 2D plot of the reconstructed phase space from 1.486 seconds of speech using time delay values (a) too small, (b) good and (c) excessive

The choice of the proper time delay (*p*) and embedding dimension (*m*) values must be made carefully. A too small value to time delay produces vectors $\vec{X}_i$ and $\vec{X}_{i+1}$ very similar, and consequently an autocorrelated attractor trajectory, probably stretched along the diagonal. When *p* value is excessive the reconstructed trajectory becomes too disperse. Figure 1 shows a 2D plot of the reconstructed phase space from 1.486 seconds of speech using time delay values (a) too small, (b) good and (c) excessive. If the attractor is unfolded into a phase space whose embedding dimension is lower than the minimum necessary, there will be vectors that remain close to one another not because of the system dynamics. On the other hand, if the chosen embedding dimension is too high, the number of vectors $\vec{X}_i$ is reduced, and it is a problem for time series composed by limited *N* numbers of points.

A criterion for an intermediate choice of time delay values is based on the analysis of autocorrelation function [9]. The autocorrelation function provides a measure of the similarity between the samples of a signal, and typically the value of *p* is set as the delay where the autocorrelation function first drops to half of the initial value. Other methods for choosing time delay can be found in [9].

An interesting method to estimate an acceptable minimum embedding dimension is called method of false neighbors [10]. Basically, for each vector of the reconstructed attractor trajectory, unfolded into a *d* embedding dimension phase space, a search for its nearest neighbor vector is made. When the embedding dimension is increased to *d+1*, it is possible to discover the percentage of neighbors that were actually “false” neighbors, and did not remain close because the *d* embedding dimension was too small. When the false neighbors percentage drops to an acceptable value, it is possible to state that the attractor was completely unfolded.

3. FRACTAL DIMENSION ESTIMATION

Nakagawa [11] estimated the fractal dimensions of self-affine data with power spectra in according with a power law based on the moment exponent. The theoretical outlines of this method, called critical exponent method (CEM), is reviewed below and a particular adaptation for speech signals is suggested.

For time series of self-affine data, the fractal dimension D_0 can be estimated as

$$D_0=2-H ,$$

where H is the Hurst exponent.

The CEM is based on analyzing the momentum I_α associated to the signal power spectrum, defined as

$$I_\alpha = \int_1^U duP(u)u^\alpha , \text{ for } -\infty < \alpha < +\infty ,$$

where U is the upper limit to the normalized frequency u, and $P(u)$ is the power spectral density, and may be assumed to follow the power law

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

Specifically to speech signals, consider k_c the lower cut frequency below which $P(u)$ does not follow the power law. By making $u=k/kc$, where k is the real frequency, these low frequencies are correctly not considered [11]. However, the estimation of proper values for k_c is made heuristically, by visualizing the speech power spectrum and "guessing" the correct value. In this work, we suggest a way to determine automatically a good approximation to k_c based on a smoothness representation of the power spectral density.

Figure 2 shows a typical frequency response from a 30ms voiced speech window, obtained through fast Fourier transform (FFT) algorithm. It is important to note that, after a determined frequency, the power spectrum decreases, following approximately the power law described previously. If it is available a smoothness representation of the power spectrum, it would be possible to search for the maximum and consider it the lower cut frequency k_c, above which an approximately exponential decrease is presented. The use of a smoothness representation of frequency response, instead of only choosing the frequency whose magnitude is maximum in the window, considers all its evolution and avoid choosing a frequency based only in one (possibly incorrect) value. Furthermore, the smoothness representation is capable of providing the exact frequency where the magnitude stops increasing and starts decreasing, avoiding small peaks. A smooth envelope representation of the power spectral density is available using the linear prediction spectrum [3][2], obtained from the estimation of the linear prediction coefficients (LPC). Figure 3 presents the linear predictive spectra for various orders for the same speech window shown in figure 2. A simple peak picking algorithm can easily find a good approximation for k_c.

After determining the lower cut off frequency, for practical effects it is possible to differentiate the logarithm of moment I_α to the 3rd order using the following equation

$$\frac{\partial^3 \ln(I_\alpha)}{\partial \alpha^3} = \frac{I_\alpha^2 I_\alpha''' - 3I_\alpha I_\alpha' I_\alpha'' + 2(I_\alpha')^3}{I_\alpha^3} ,$$

where the nth derivative of I_α, I_α^n, can be evaluated from the equation

$$I_\alpha^n = \frac{\partial^n}{\partial \alpha^n} \int_1^U duu^\alpha P(u) = \int_1^U du(\ln(u))^n u^\alpha P(u) ,$$

and solve the following equation to find the critical value α_c ,

$$\frac{\partial^3 \ln(I_\alpha)}{\partial \alpha^3} = 0 .$$

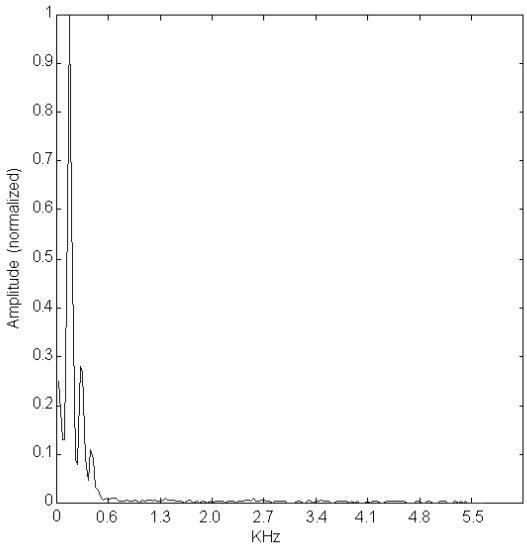

Fig. 2. Typical frequency response from a 30ms voiced speech window

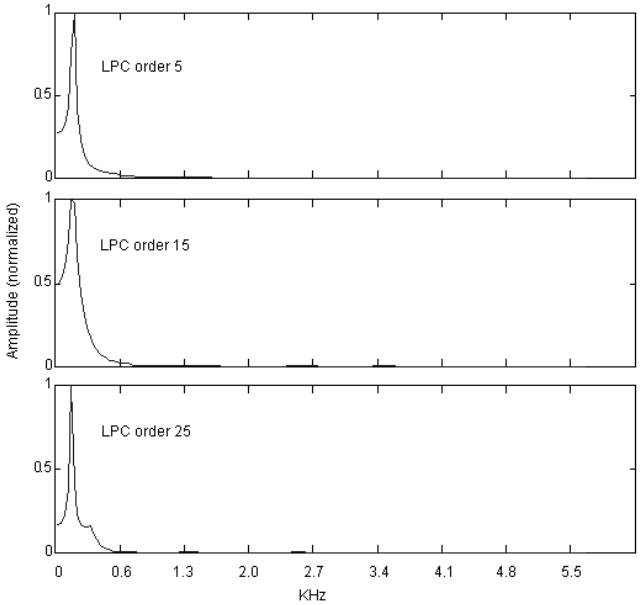

Fig. 3. Linear predictive spectra for various orders for the same speech window shown in figure 2

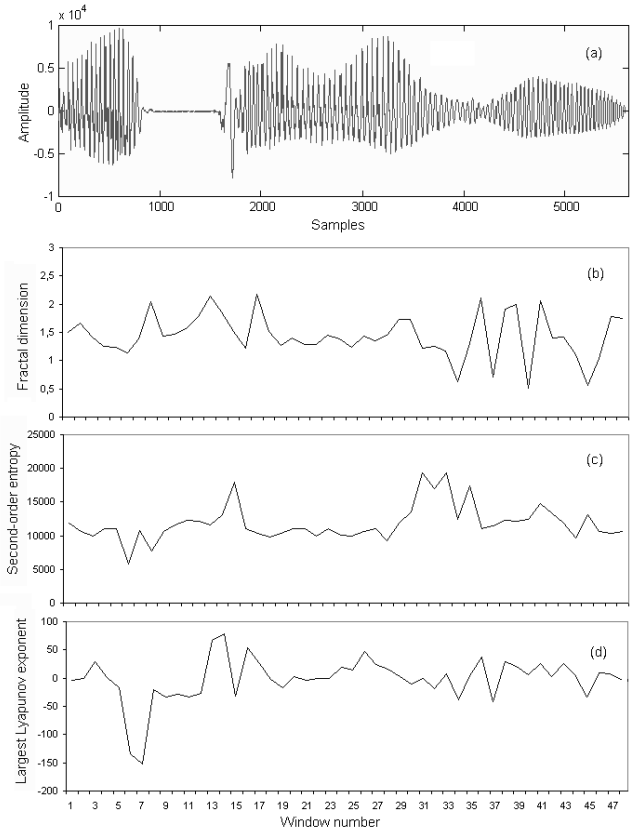

Fig. 4. Waveform of a speech signal (a), the respective time-dependent fractal dimensions (b), entropy (c) end largest Lyapunov value (d)

From the above relation, exponent β (from power law equation) is given as

$$\beta = \alpha_c + 1 = 2H + 1$$

and the fractal dimension D_0 can be calculated from

$$D_0 = 2 - H = 2 - \frac{\alpha_c}{2}.$$

The method described previously can be applied in every speech window, which provides the time-dependent fractal dimensions (TDFDs). A more detailed explanation about extracting the TDFDs using CEM can be found in [5]. The values of fractal dimension obtained with automatic search for lower cut frequency k_c are very close to the results reported in [5] for speech signals. Figure 4 (b) shows the values of fractal dimensions obtained from the speech signal in figure 4 (a), evaluating a 30ms hamming window of speech, applied every 10ms.

4. SECOND ORDER DYNAMICAL ENTROPY

Entropy is an important measure of system dynamics. The extraction of Kolmogorov or metric entropy (K) quantifies the rate of loss of information about a system state, during its evolution. A metric entropy measure positive and finite, i.e., $0<K<\infty$, characterizes the presence of chaotic behavior in the analyzed system.

The Kolmogorov entropy can be defined as follows [4][12][13]. Dividing a m-dimensional phase space into $N(\varepsilon)$ hiper-cubes with size ε^m, where observations are available at regular time intervals τ, the metric entropy can be written as

$$K = -\lim_{\tau\to 0}\lim_{\varepsilon\to 0}\lim_{b\to\infty}\frac{1}{b\tau}\sum_{i_1,\ldots,i_b} p(i_1,\ldots,i_b)\ln p(i_1,\ldots,i_b)$$

where $p(i_1,...i_b)$ is the joint probability of observing a system trajectory in hiper-cube i_1 in time τ, in hiper-cube i_2 in time 2τ, ... , and in hiper-cube i_b in time $b\tau$.

The second-order dynamical entropy (K_2) can be extracted fairly easily from experimental time series and in [13], Grassberger *et al.* proposed estimate K by the quantity K_2, once K_2 represents a lower bound for K, and for typical cases K_2 is numerically close to K. Another important advantage is that $0<K_2<\infty$ is a sufficient condition for chaos. The practical method to estimate K_2 is using the correlation sum $C_m(\varepsilon)$, through the approximation

$$K_2 \approx \frac{1}{\tau}\ln\frac{C_m(\varepsilon)}{C_{m+1}(\varepsilon)}$$

for embedding dimension m and value of ε as small as possible. The correlation sum is given by the following equation:

$$C_m(\varepsilon) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{\substack{i=1\\ i\neq j}}^{N}\Theta\left(\varepsilon - \left\|\vec{X}_i - \vec{X}_j\right\|\right)$$

where $\vec{X}_i$ is the ith vector from the attractor trajectory reconstruction, composed by N (m-dimensional) vectors, and Θ denotes the Heaviside function: $\Theta(\arg) = 1$ for arg≥0 and 0 otherwise.

In the same way it was done with fractal dimension, it is possible to extract the entropy from every window of a speech signal. Figure 4 (c) shows the values of second order entropy obtained from the same speech signal of figure 4 (a), evaluating a 30ms hamming window of speech, applied every 10ms.

5. LYAPUNOV EXPONENTS

The Lyapunov exponents quantify the sensitivity of a dynamical system to initial conditions. When an attractor is chaotic, the average exponential divergence of nearby trajectories is quantified by estimating the largest Lyapunov exponent. For time series produced by a dynamical system, the presence of a positive value for the Lyapunov exponents indicates the presence of chaos. Furthermore, in many applications it is sufficient to estimate only the largest value of the Lyapunov spectrum.

Rosenstein *et. al.* [14] proposed a method to estimate the Lyapunov exponents from time series composed by a very limited number of samples available. Good results were obtained for estimating the largest Lyapunov exponent (λ_1) of known systems using just 100-1000 samples. This characteristic is quite important when dealing with speech, once a speech signal can be considered stationary only during a small window of approximate 30ms.

The first step is the reconstruction of the attractor's trajectory in an appropriate phase space. After, the nearest neighbor of every vector of the reconstructed trajectory is found. A constraint that nearest neighbors have a temporal separation greater than the mean period of the time series must be satisfied. Doing this, it is possible to consider the pair of neighbors as belonging to different trajectories. When considering two trajectories whose initial conditions are very similar, the trajectories diverge, on average, at an exponential rate characterized by the largest Lyapunov exponent (λ_1), as follows

$$d_j(i) \approx C_j e^{\lambda_1 (i\Delta t)}$$

where $d_j(i)$ is the distance between the jth pair of nearest neighbors after i steps (equals to $i.\Delta t$ seconds where Δt is the time series sampling period) and C_j is the initial separation between the neighbors.

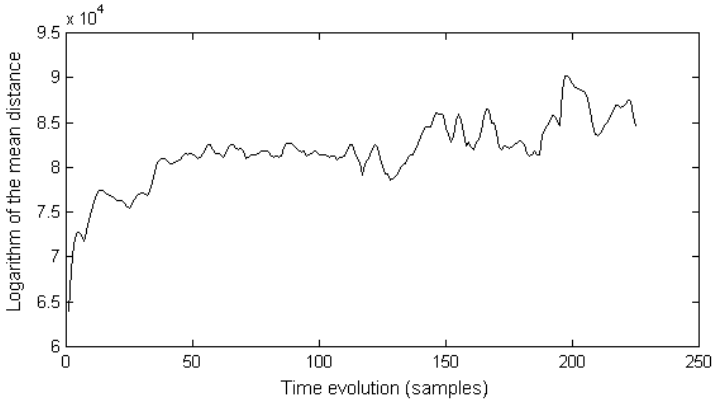

Fig. 5. Logarithm of the mean distance evolution between every pair of neighbors from the reconstructed phase space vectors of a 30ms window of speech

Applying the natural logarithm to both sides, the previous equation becomes

$$\ln d_j(i) \approx \ln C_j + \lambda_1 (i\Delta t)$$

If the logarithm of the distance evolution between every pair of neighbors is monitored, they will appear as a set of approximately parallel lines, each with a slope proportional to λ_1. The largest Lyapunov exponent is then estimated by applying least-squares method to best model the mean line. Figure 5 shows the logarithm of the mean distance evolution between every pair of neighbors from the reconstructed phase space vectors of a 30ms window of speech. It is easy to verify its positive slope, which indicates a positive value for the correspondent largest Lyapunov exponent. By repeating the same process to every window of the speech signal in figure 4 (a), the time-dependent largest Lyapunov exponents can be obtained and are showed in figure 4 (d).

6. THE SPEAKER RECOGNITION SYSTEM

A simple template-matching system, based on the Bhattacharyya distance, was used to evaluate the recognition performance that can be obtained when LP-derived cepstral coefficients and pitch are combined with nonlinear dynamic feature parameters. Basically, some speech samples of every registered speaker in the system are used to compose that speaker template. It is done by extracting the desired feature parameters from every window of all speech samples from that speaker, and calculating its mean and covariance matrix. When an unknown speech sample is presented to the system, its feature parameters are extracted the same way, the mean and covariance matrix are calculated and a similarity measure is obtained for every registered speaker template, using the Bhattacharyya distance for multivariate Gaussian distributions. The unknown speech sample is then assigned to the registered speaker whose similarity measure is maximized.

6.1 LP-derived cepstral coefficients

Linear prediction analysis is an important method of characterizing the spectral properties of speech in the time domain. In this analysis method, each sample of the speech signal is predicted as a linear weighted sum of the past *p* samples. The weights which minimize the mean-squared prediction error are called the predictor coefficients. The value of *p* is approximately determined by the number of poles of the vocal tract and the glottal wave transfer function, mathematically modeled.

An important method to estimate the linear prediction coefficients (LPC) in a window of speech starts with the calculation of the autocorrelation coefficients:

$$r(m) = \sum_{n=0}^{N-1-m} y(n)y(n+m)$$

where *m=0,1,...,p* , *p* is the LPC order, *r(m)* are the autocorrelation coefficients, *N* is the number of samples in the window, and *y(n)* represents the speech samples in that window.

To transform the autocorrelation coefficients into LPC, it is possible to use Durbin method, well detailed in [2][3]. The equations are the following:

$$E^{(0)} = r(0)$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(i-j)}{E^{(i-1)}} \qquad 1 \le i \le p$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1$$

$$E^{(i)} = (1 - k_i^{\,2}) E^{(i-1)}$$

and they are solved recursively for *i=1,2,...p* . Then, the LPC (a_m) are obtained with

$$a_m = \alpha_m^{(p)} \qquad 1 \le m \le p$$

By definition, the cepstrum (or the cepstral coefficients) is the inverse Fourier transform of the logarithm of the speech signal spectrum. The relationship between the cepstrum and the predictor coefficients are [2][3]:

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k} \quad 1 \le m \le p$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k} \qquad m > p$$

where c_m is the *m*th cepstral coefficient, a_m is the *m*th linear prediction coefficient and *p* is the predictor order.

The Cepstral Coefficients obtained from the predictor coefficients are called LP-derived cepstral coefficients.

6.2 Fundamental frequency

The fundamental frequency (also called pitch), represents the vocal chords main frequency of vibration. Its value is directly associated with the source of excitation of the vocal tract, and it is an important speaker-dependent parameter. Women and children have vocal chords with small width, consequently presenting a value for pitch usually greater than men.

There are different techniques to estimate the pitch value from a voiced speech window. Cepstrum properties can be used to calculate it with good accuracy [2]. The

higher order cepstral coefficients are considered, once it contains information concerning vocal tract excitation. In practice, the algorithm search for a peak among the values of higher order cepstral coefficients, in the neighborhood of a reasonable value for the pitch (from 3ms to 20ms, equivalent from 333Hz to 50Hz). If the amplitude of a found peak is over a threshold, its position can be considered a good approximation to pitch period. Figure 6 shows the cepstral coefficients values obtained from a voiced speech window, where clearly the desired peak is marked. The sample rate used was 11025Hz, so the pitch period is approximately 9.3ms, equivalent to fundamental frequency of 107Hz.

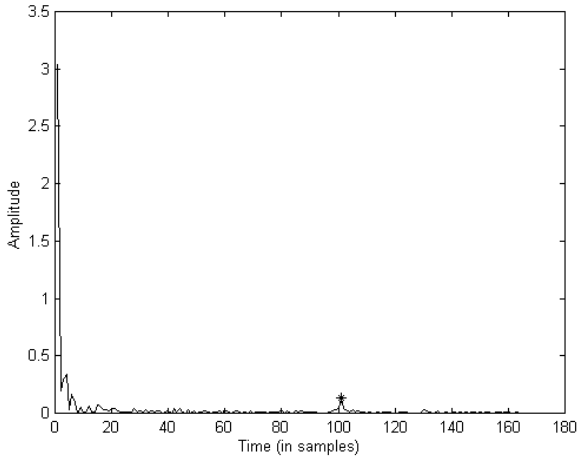

Fig. 6. Cepstrum from a voiced speech window

6.3 Bhattacharyya distance

In statistics, the proximity degree between two different probability densities is related with the notion of distance measure. One way to calculate this measure is throughout the Bhattacharyya distance. Considering two probability densities $p_1(x)$ and $p_2(x)$, obtained from two different classes of feature parameters, the Bhattacharyya distance [15] is defined by

$$B = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)\, p_2(x)}\, dx$$

Special cases of this general distance measure can be calculated explicitly to a large types of probability densities. An important case refers to the multivariate Gaussian distributions. Considering $p_i(x)$ Gaussian probability densities, it is possible to show [16] that the previous equation can be written as:

$$B = \frac{1}{8}(m_1 - m_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (m_1 - m_2) + \frac{1}{2} \ln \left(\frac{\det(\Sigma_1 + \Sigma_2)/2}{\sqrt{\det(\Sigma_1)}\sqrt{\det(\Sigma_2)}} \right)$$

where m_i is the mean value and Σ_i is the covariance matrix, obtained from the feature parameters of class i.

The Bhattacharyya distance can be applied to a wide variety of known probability distributions, according to the best fit. The assumption of Gaussian density for the parameters is not arbitrary, since it is sufficient that the density be essentially unimodal and approximately Gaussian in the center of its range. These properties are often respected in physical systems. Inspecting histograms obtained from the feature

parameters, it is possible to verify that their value distributions can be modeled as Gaussian probability densities.

6.4 The Data Set

The data set used to evaluate the speaker recognition performance is composed with speech samples from 37 different speakers, sampled at 11025Hz, with resolution of 16 bits per sample. Every speaker provided three repetitions of the vocabulary, composed by the words “first”, “second”, ... , and “tenth”, spoken in Portuguese language. The first two repetitions of the vocabulary were used to generate the speakers templates. The third repetition of the vocabulary of every speaker was used to test the system accuracy, in a total of 370 different identifications for a single test.

7. EXPERIMENTAL RESULTS

The speech samples from all 37 speakers were used to train and evaluate the speaker identification system, described previously. Different tests were accomplished, and the focus was to verify the efficiency of fractal dimension, entropy and largest Lyapunov exponent in speaker recognition task. From all speech samples, a hamming window with length of 30ms was applied every 10ms, and from every window LP-derived cepstral coefficients, pitch, fractal dimension, entropy and largest Lyapunov exponent were extracted according with the methods previously described.

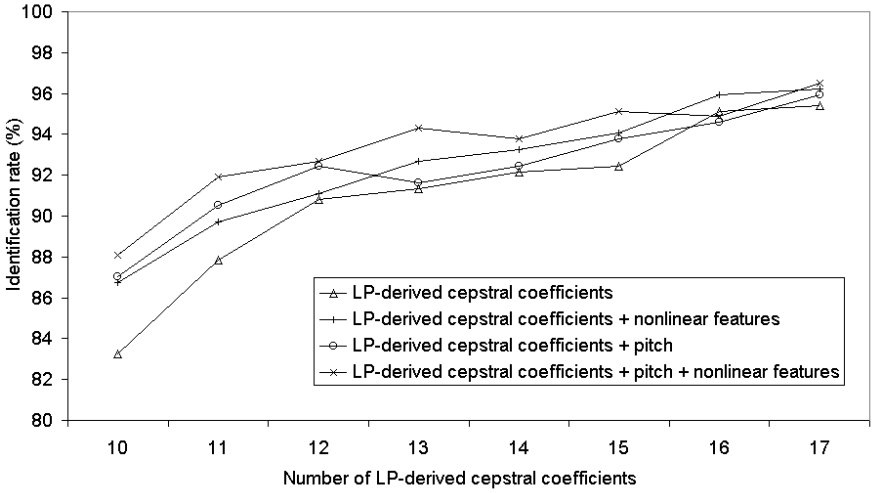

Fig. 7. System accuracy varying the number of LP-derived cepstral coefficients

When only the nonlinear features are used to fully characterize a speaker, the results obtained are poor: only 6.49% using fractal dimension, 7.57% using entropy, 9.19% using largest Lyapunov exponents, and 19.19% combining all nonlinear features. The information provided by these parameters is not enough to discard other parameters. The same idea can explain the bad performance of the speaker identification system which uses only pitch, reaching only 14.86%. Other features can achieve much higher accuracy, and LP-derived cepstral coefficients are extensively used in many speaker recognition systems, achieving 95.40% for the identification system described earlier.

Even the nonlinear feature parameters can not alone provide a speaker complete characterization, it can still help on improving the accuracy obtained with Fourier and

cepstral analysis, by providing other kind of information, not considered so far. The combination of Fourier and cepstral, with nonlinear dynamic analysis can more accurately characterize a speaker, leading the correspondent speaker recognition system to higher performance. It can be seen in figure 7, where the number of LP-derived cepstral coefficients vary. When pitch information is combined with cepstral information there is a improvement in the system's performance, indicating that it contains speaker-dependent information, which can distinguish different speakers. The same idea explains the addition of nonlinear features on helping the characterization of a speaker, even better than pitch after 13 LP-derived cepstral coefficients. The combination of cepstral analysis with pitch and nonlinear features leads the speaker identification system to even better results, achieving 96.49% of accuracy.

Figure 7 shows clearly that there is a considerable performance gain when combining nonlinear dynamic features. However, the processing time necessary to extract the nonlinear features may be much greater than to extract cepstrum and pitch. For comparison, a personal computer with an Pentium processor running at 350 MHz takes about 69.2ms to extract fractal dimension, 560.6ms for entropy and 505.2ms to extract the largest Lyapunov exponent from a window of 30ms of speech. The same computer spends only 6.9ms to extract 17 LP-derived cepstral coefficients, and 13.8ms for pitch. It is possible to verify that the processing time is increased when nonlinear features are added. It takes about 37.8 times the speech time to complete the nonlinear dynamical analysis and only 0.69 times the speech time for LP-derived cepstral coefficients and pitch extraction. The previous estimation of time was based in an average, using 289 different windows from a speech file.

8. CONCLUSIONS

This work does not intend to invalidate all research that has been developing the speaker recognition area from past years, but suggest new ideas to construct systems more robust and reliable. Extract new information that specifically distinguish different speakers is very important to continue the development of this area. In the other hand, the introduction of new techniques and new features to characterize a speaker will bring an intrinsic computational processing overhead. Particularly with nonlinear features, such processing may not allow the construction of real time systems with the hardware available today.

The standard feature parameters widely used to characterize a speaker can perform this task with relative success. However, many applications where the speaker recognition technology can potentially be introduced are still searching for more accurate systems. And the nonlinear dynamic analysis can see the speech production differently, as the result of a nonlinear dynamic process, bringing up new information to characterize it in a more complete way.

Acknowledgments - The authors would like to acknowledge Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) for its financial support.

REFERENCES

[1] Furui, S. Recent Advances in Speaker Recognition, proc. of *Audio- and Video-based Biometric Person Authentication*, AVBPA 97, Crans-Montana, Switzerland, March 1997.

[2] Rabiner, L. R. and Schafer, R. W. Digital processing of speech signals, *Prentice Hall*, 1978.
[3] Deller Jr., J. R., Proakis, J. G. and Hansen, J. H. L. Discrete-time processing of speech signals, *Prentice Hall*, 1987.
[4] Kumar, A., Mullick, S. K. Nonlinear dynamical analysis of speech, *J. Acoust. Soc. Am.* **100** (1), July 1996.
[5] Sabanal, S., Nakagawa, M. The Fractal Properties of Vocal Sounds and Their Application in the Speech Recognition Model, *Chaos, Solitons & Fractals*, Vol. 7, No. 11, pp. 1825-1843, 1996.
[6] Banbrook, M., McLaughlin, S. Mann, I. Speech Characterization and Synthesis by Nonlinear Methods, *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 1, January 1999.
[7] Chan, A. M. and Leung, H. Equalization of Speech and Audio Signals Using a Nonlinear Dynamical Approach, *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 3, May 1999.
[8] Takens, F. Detecting strange attractors in turbulence *in Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, edited by D. A. Rand and L. S. Young, Springer-Verlag, Berlin, 1981, vol. 898, pp. 366-381.
[9] Rosenstein, M. T., Collins, J. J. and De Luca, C. J. Reconstruction expansion as a geometry-based framework for choosing proper delay times, *Physica D* 73 (1994) 82-98.
[10] Kennel, M. B., Brown, R. and Abarbanel, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Physical Review A*, vol. 45, no. 6, pp. 3403-3411, March 1992.
[11] Nakagawa, M. A Critical Exponent Method to Evaluate Fractal Dimensions of Self-Affine Data, *J. of the Physical Society of Japan*, vol. 62, no. 12, pp. 4233-4239, December 1993.
[12] Cohen, A. and Procaccia, I. Computing the Kolmogorov entropy from time signals of dissipative and conservative dynamical systems, *Physical Review A*, vol. 31, no. 3, pp.1872-1882, March 1985.
[13] Grassberger, P. and Procaccia, I. Estimation of the Kolmogorov entropy from a chaotic signal, *Physical Review A*, vol. 28, no. 4, October 1983.
[14] Rosenstein, M. T., Collins, J. J., De Luca, C. J. A practical method for calculating largest Lyapunov exponents from small data sets, *Physica D* 65 (1993) 117-134.
[15] Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions, *Bull Calcutta Math. Soc.*, vol. 35, pp. 99-109, 1943.
[16] Kailath, T. The Divergence and Bhattacharyya Distance Measures in Signal Selection, *IEEE Trans. On Communication Technology*, vol. Com-15, pp. 52-60, February 1967.

Anexo 3 Fractal dimension applied to speaker identification

Trabalho intitulado “**Fractal Dimension Applied to Speaker Identification**”, publicado no veículo IEEE International Conference on Acoustics, Speech, and Signal Processing 2001, com autoria de Adriano Petry e Dante A. C. Barone. A formatação original do trabalho foi modificada.

FRACTAL DIMENSION APPLIED TO SPEAKER IDENTIFICATION

Petry, A. and Barone, D. A. C.
{adpetry,barone}@inf.ufrgs.br

Instituto de Informática, Universidade Federal do Rio Grande do Sul - Porto Alegre, Brazil

ABSTRACT

This paper reports the results obtained in a speaker identification system based in Bhattacharrya distance, which combines LP-derived cepstral coefficients, with a nonlinear dynamic feature namely fractal dimension. The nonlinear dynamic analysis starts with the phase espace reconstruction, and the fractal dimension of the correspondent attractor trajectory is estimated. This analysis is performed in every speech window, providing a measure of a time-dependent fractal dimension. The corpus used in the tests is composed by 37 different speakers, and the best results are obtained when the fractal dimension is included, suggesting that the information added with this feature was not present so far.

1. INTRODUCTION

A speaker identity is strongly dependent of the physiological and behavioral characteristics of the speech production system. The first step of a basic speaker recognition system is to extract from the speech samples a “good” parametric representation. These parameters must be, as much as possible, representative of a speaker, presenting low variability for that speaker’s speech samples, and great difference when used with others speakers’ speech samples.

Previous papers [1][2][3][4] have worked with speech characterization and analysis using nonlinear dynamical features. Sabanal *et al*. [2] used the time-dependent fractal dimensions (TDFDs), extracted through critical exponent method (CEM), and the time-dependent multifractal dimensions (TDMFDs) to accomplish a speech recognizer. The target was to recognize Japanese digits using a neural network. Kumar *et al*. [1] estimated Lyapunov exponents, dimension and metric entropy in phonemes signals, divided into eight different types. Banbrook *et al*. [3] extracted correlation dimension, Lyapunov exponents, and short-term predictability from a corpus of sustained vowels sounds. The works mentioned used some nonlinear dynamical features to characterize a speech sound, showing the speech low dimensionality and the average exponential divergence of nearby trajectories in the reconstructed phase space.

In this work a speaker identification is performed using a combination of LP-derived cepstral coefficients with a nonlinear dynamic invariant: the fractal dimension. While the use of LP-derived cepstral coefficients can perform a speaker recognition quite successfully, it may not be accurate enough for some applications. The characterization of a speaker using a nonlinear dynamic description can help on identifying people from their voices. The assumptions used to extract the standard feature parameters do not describe the nonlinear dynamic evolution of the system. It will be shown that add nonlinear dynamic qualitative information to the standard feature parameters, such as fractal dimension, is equivalent to add speaker-dependent features, not present in the

standard feature parameters so far. This combination will lead a speaker recognition system to more accurate results.

2. PHASE ESPACE RECONSTRUCTION

In experimental applications, it is often available unidimensional measurements of a dynamical system that evolves in a multidimensional phase space. This scalar time series contains the information available from that system. In many cases, no further information is available, and an important challenge that has to be solved is the calculation of the system's real multidimensional phase space trajectory. After that, measurements that provide important knowledge about the system behavior can be done.

To evaluate the properties of an attractor associated to a time series it is first necessary to reconstruct its evolution in a proper phase space. The most used way of reconstructing the full dynamics of a system from scalar time series measurements was proposed by Takens [5]. This method presents easy practical implementation. Given a *N*-point time series $x(t_i)$ for $i=1,2,...N$ as follows

$$x(t) = \{x(t_1), x(t_2),..., x(t_N)\},$$

the *m*-dimensional vectors are reconstructed, according to Takens delay method [5], as

$$\vec{X}_i = \{x(t_i), x(t_i + p), x(t_i + 2p),..., x(t_i + (m-1)p)\},$$

where *p* is called time delay and *m* is the embedding dimension. The $\vec{X}_i$ vectors represent the trajectory of the time series $x(t_i)$ in a *m*-dimensional phase space.

The choice of the proper time delay (*p*) and embedding dimension (*m*) values must be made carefully. A too small value to time delay produces vectors $\vec{X}_i$ and $\vec{X}_{i+1}$ very similar, and consequently an autocorrelated attractor trajectory, probably stretched along the diagonal. When *p* value is excessive the reconstructed trajectory becomes too disperse. If the attractor is unfolded into a phase space whose embedding dimension is lower than the minimum necessary, there will be vectors that remain close to one another not because of the system dynamics. On the other hand, if the chosen embedding dimension is too high, the number of vectors $\vec{X}_i$ is reduced, and it is a problem for time series composed by limited *N* numbers of points.

A criterion for an intermediate choice of time delay values is based on the analysis of autocorrelation function [6]. The autocorrelation function provides a measure of the similarity between the samples of a signal, and typically the value of *p* is set as the delay where the autocorrelation function first drops to half of the initial value. Other methods for choosing time delay can be found in [6].

An interesting method to estimate an acceptable minimum embedding dimension is called method of false neighbors [7]. Basically, for each vector of the reconstructed attractor trajectory, unfolded into a *d* embedding dimension phase space, a search for its nearest neighbor vector is made. When the embedding dimension is increased to *d+1*, it is possible to discover the percentage of neighbors that were actually "false" neighbors, and did not remain close because the *d* embedding dimension was too small. When the false neighbors percentage drops to an acceptable value, it is possible to state that the attractor was completely unfolded.

3. FRACTAL DIMENSION ESTIMATION

Nakagawa [8] estimated the fractal dimensions of self-affine data with power spectra in according with a power law based on the moment exponent. The theoretical outlines of

this method, called critical exponent method (CEM), is reviewed below and a particular adaptation for speech signals is suggested.

For time series of self-affine data, the fractal dimension D_0 can be estimated as

$$D_0=2-H\ ,$$

where H is the Hurst exponent.

The CEM is based on analyzing the momentum I_α associated to the signal power spectrum, defined as

$$I_\alpha = \int_1^U duP(u)u^\alpha\ ,\ \text{for}\ -\infty<\alpha<+\infty\ ,$$

where U is the upper limit to the normalized frequency u, and $P(u)$ is the power spectral density, and may be assumed to follow the power law

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

Specifically to speech signals, consider k_c the lower cut frequency below which $P(u)$ does not follow the power law. By making $u=k/kc$, where k is the real frequency, these low frequencies are correctly not considered [8]. However, the estimation of proper values for k_c is made heuristically, by visualizing the speech power spectrum and "guessing" the correct value. In this work, we suggest a way to determine automatically a good approximation to k_c based on a smoothness representation of the power spectral density.

Figure 1 shows a typical frequency response from a 30ms voiced speech window, obtained through fast Fourier transform (FFT) algorithm. It is important to note that, after a determined frequency, the power spectrum decreases, following approximately the power law described previously. If it is available a smoothness representation of the power spectrum, it would be possible to search for the maximum and consider it the lower cut frequency k_c, above which an approximately exponential decrease is presented. The use of a smoothness representation of frequency response, instead of only choosing the frequency whose magnitude is maximum in the window, considers all its evolution and avoid choosing a frequency based only in one (possibly incorrect) value. Furthermore, the smoothness representation is capable of providing the exact frequency where the magnitude stops increasing and starts decreasing, avoiding small peaks.

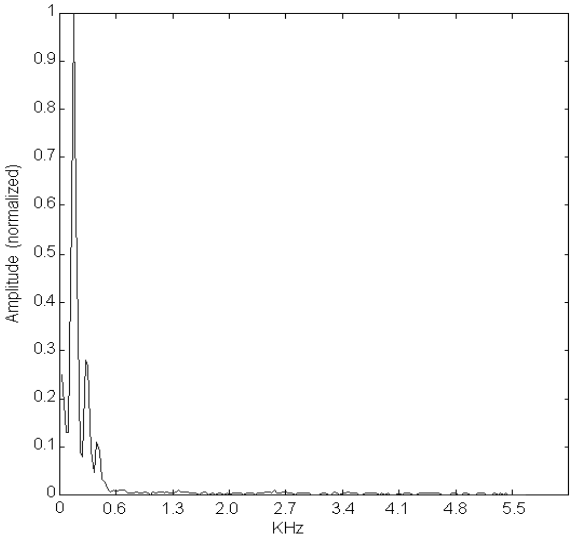

Fig. 1. Typical frequency response from a 30ms voiced speech window.

A smooth envelope representation of the power spectral density is available using the linear prediction spectrum [9][10], obtained from the estimation of the linear prediction coefficients (LPC). A simple peak picking algorithm can easily find a good approximation for k_c.

After determining the lower cut off frequency, for practical effects it is possible to differentiate the logarithm of moment I_α to the 3rd order using the following equation

$$\frac{\partial^3 \ln(I_\alpha)}{\partial \alpha^3} = \frac{I_\alpha^2 I_\alpha''' - 3 I_\alpha I_\alpha' I_\alpha'' + 2(I_\alpha')^3}{I_\alpha^3} \ ,$$

where the *n*th derivative of I_α, I_α^n, can be evaluated from the equation

$$I_\alpha^n = \frac{\partial^n}{\partial \alpha^n} \int_1^U du u^\alpha P(u) = \int_1^U du (\ln(u))^n u^\alpha P(u) \ ,$$

and solve the following equation to find the critical value α_c ,

$$\frac{\partial^3 \ln(I_\alpha)}{\partial \alpha^3} = 0 \, .$$

From the above relation, exponent β (from power law equation) is given as

$$\beta = \alpha_c + 1 = 2H + 1$$

and the fractal dimension D_0 can be calculated from

$$D_0 = 2 - H = 2 - \frac{\alpha_c}{2} \, .$$

The method described previously can be applied in every speech window, which provides the time-dependent fractal dimensions (TDFDs). The values of fractal dimension obtained with automatic search for lower cut frequency k_c are very close to the results reported in [2] for speech signals. Figure 2 (b) shows the values of fractal dimensions obtained from the speech signal in figure 2 (a), evaluating a 30ms hamming window of speech, applied every 10ms.

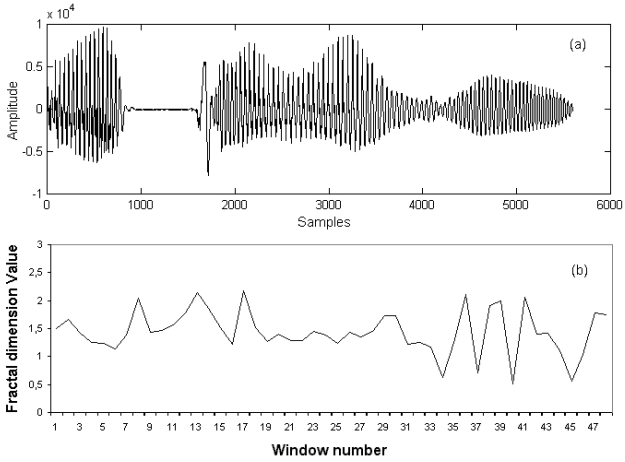

Fig. 2. Waveform of a speech signal (a), the respective time-dependent fractal dimensions (b).

4. THE SPEAKER RECOGNITION SYSTEM

A system based on the Bhattacharyya distance was used to evaluate the recognition performance that can be obtained when LP-derived cepstral coefficients is combined with fractal dimension. This system is similar to the one described in [11]. Basically, some speech samples of every registered speaker in the system are used to compose that speaker identity. It is done by extracting the desired feature parameters from every window of all speech samples from that speaker, and calculating its mean and covariance matrix. When an unknown speech sample is presented to the system, its

feature parameters are extracted the same way, the mean and covariance matrix are calculated and a similarity measure is obtained for every registered speaker, using the Bhattacharyya distance for multivariate Gaussian distributions. The unknown speech sample is then assigned to the registered speaker whose similarity measure is maximized.

4.1. LP-derived cepstral coefficients

Linear prediction (LP) analysis is an important method of characterizing the spectral properties of speech in the time domain. In this analysis method, each sample of the speech signal is predicted as a linear weighted sum of the past *p* samples. The weights which minimize the mean-squared prediction error are called the predictor coefficients. The value of *p* is approximately determined by the number of poles of the vocal tract and the glottal wave transfer function, mathematically modeled. An important method to estimate the linear prediction coefficients (LPC) is called Durbin method, well detailed in [9][10].

By definition, the cepstrum (or the cepstral coefficients) is the inverse Fourier transform of the logarithm of the speech signal spectrum. The cepstral coefficients obtained from the predictor coefficients are called LP-derived cepstral coefficients. The relationship between the cepstrum and the predictor coefficients are [9][10]:

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \quad 1 \le m \le p$$

$$c_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \qquad m > p$$

where c_m is the *m*th cepstral coefficient, a_m is the *m*th linear prediction coefficient and *p* is the predictor order.

4.2. Bhattacharyya distance

In statistics, the proximity degree between two different probability densities is related with the notion of distance measure. An estimation to the upper bound on the Bays error can be obtained using the Bhattacharyya distance. Considering two probability densities $p_1(x)$ and $p_2(x)$, obtained from two different classes of feature parameters, the Bhattacharyya distance [12] is defined by

$$B = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)\,p_2(x)}\,dx$$

Special cases of this general distance measure can be calculated explicitly to a large types of probability densities. An important case refers to the multivariate Gaussian distributions. Considering $p_i(x)$ Gaussian probability densities, it is possible to show [13] that the previous equation can be written as:

$$B = \frac{1}{8}(m_1 - m_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (m_1 - m_2) + \\ + \frac{1}{2}\ln\left(\frac{\det(\Sigma_1 + \Sigma_2)/2}{\sqrt{\det(\Sigma_1)}\sqrt{\det(\Sigma_2)}}\right)$$

where m_i is the mean value and Σ_i is the covariance matrix, obtained from the feature parameters of class *i*.

The Bhattacharyya distance can be applied to a wide variety of known probability distributions, according to the best fit. The assumption of Gaussian density for the

parameters is not arbitrary, since it is sufficient that the density be essentially unimodal and approximately Gaussian in the center of its range. These properties are often respected in physical systems. Inspecting histograms obtained from the feature parameters, it is possible to verify that their value distributions can be modeled as Gaussian probability densities.

4.3. The Data Set

The corpus used to evaluate the speaker recognition performance is composed with speech samples from 37 different speakers, sampled at 11025Hz, with resolution of 16 bits per sample. Every speaker provided three repetitions of the vocabulary, composed by the words “first”, “second”, ... , and “tenth”, spoken in Portuguese language. Every speech sample was about 618ms long, in average. The first two repetitions of the vocabulary were used to generate the speakers identity. The third repetition of the vocabulary of every speaker was used to test the system accuracy, in a total of 370 different identifications for a single test.

5. EXPERIMENTAL RESULTS

Different tests were accomplished, and the focus was to verify the efficiency of fractal dimension in speaker recognition task. From all speech samples, a hamming window with length of 30ms was applied every 10ms, and from every window LP-derived cepstral coefficients and fractal dimension were extracted according with the methods previously described.

The nonlinear feature parameters can help on improving the accuracy obtained with Fourier and cepstral analysis, by providing other kind of information, not considered so far. The combination of Fourier and cepstral, with nonlinear dynamic analysis can more accurately characterize a speaker, leading the correspondent speaker recognition system to higher performance. It can be seen in figure 3, where the number of LP-derived cepstral coefficients vary. When nonlinear dynamic information is combined with cepstral information there is a improvement in the system’s performance, indicating that it contains speaker-dependent information, which can distinguish different speakers. The combination of cepstral analysis with nonlinear features leads the speaker identification system to even better results, achieving 97.29% of accuracy, which is a good result considering the amount of speech used on trainning (about 1.2s per speaker, in average) and recognition (about 618ms per identification, in average).

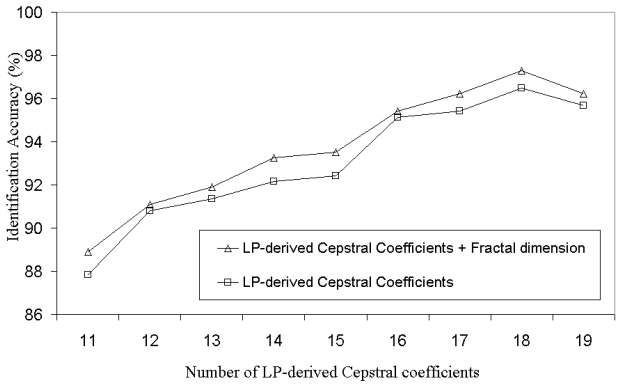

Fig. 3. System identification accuracy varying the number of LP-derived Cepstral coefficients.

Figure 3 shows clearly that there is a considerable performance gain when combining nonlinear dynamic features. However, the processing time necessary to extract the nonlinear features may be much greater than to extract cepstrum. For comparison, a personal computer with an Pentium processor running at 350 MHz takes about 69.2ms to extract fractal dimension, and only 6.9ms to extract 17 LP-derived cepstral coefficients from a window of 30ms of speech. The processing time is heavily increased when the nonlinear feature is added. It takes about 90.9% of the total processing time to complete the nonlinear dynamical analysis and only 9,07% of the total processing time for LP-derived cepstral coefficients extraction. The previous estimation of time was based in an average, using 289 different windows from a speech file.

6. CONCLUSIONS

This work suggests new ideas to construct speaker recognition systems more robust and reliable. Extract new information that specifically distinguish different speakers is very important to continue the development of this area. In the other hand, the introduction of new techniques and new features to characterize a speaker will bring an intrinsic computational processing overhead. Particularly with nonlinear features, such processing may not allow the construction of real time systems with the hardware available today.

Many applications where the speaker recognition technology can potentially be introduced are still searching for more accurate systems. The nonlinear dynamic analysis can analyze the speech production differently, as the result of a nonlinear dynamic process, bringing up new information to characterize it in a more complete way.

Acknowledgments – Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) and Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq).

7. REFERENCES

[1] Kumar, A., Mullick, S. K. Nonlinear dynamical analysis of speech, *J. Acoust. Soc. Am.* 100 (1), July 1996.

[2] Sabanal, S., Nakagawa, M. The Fractal Properties of Vocal Sounds and Their Application in the Speech Recognition Model, *Chaos, Solitons & Fractals*, Vol. 7, No. 11, pp. 1825-1843, 1996.

[3] Banbrook, M., McLaughlin, S. Mann, I. Speech Characterization and Synthesis by Nonlinear Methods, *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 1, January 1999.

[4] Chan, A. M. and Leung, H. Equalization of Speech and Audio Signals Using a Nonlinear Dynamical Approach, *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 3, May 1999.

[5] Takens, F. Detecting strange attractors in turbulence *in Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, edited by D. A. Rand and L. S. Young, Springer-Verlag, Berlin, 1981, vol. 898, pp. 366-381.

[6] Rosenstein, M. T., Collins, J. J. and De Luca, C. J. Reconstruction expansion as a geometry-based framework for choosing proper delay times, *Physica D* 73 (1994) 82-98.

[7] Kennel, M. B., Brown, R. and Abarbanel, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Physical Review A*, vol. 45, no. 6, pp. 3403-3411, March 1992.
[8] Nakagawa, M. A Critical Exponent Method to Evaluate Fractal Dimensions of Self-Affine Data, *J. of the Physical Society of Japan*, vol. 62, no. 12, December 1993.
[9] Deller Jr., J. R., Proakis, J. G. and Hansen, J. H. L. Discrete-time processing of speech signals, *Prentice Hall*, 1987.
[10] Rabiner, L. R. and Schafer, R. W. Digital processing of speech signals, *Prentice Hall*, 1978.
[11] Campbell Jr., J. P. Speaker Recognition: A Tutorial, Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, September 1997.
[12] Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions, *Bull Calcutta Math. Soc*., vol. 35, pp. 99-109, 1943.
[13] Kailath, T. The Divergence and Bhattacharyya Distance Measures in Signal Selection, *IEEE Trans. On Communication Technology*, vol. Com-15, pp. 52-60, February 1967.

Anexo 4 Text-dependent speaker verification using Lyapunov exponents

Trabalho intitulado "**Text-dependent Speaker Verification using Lyapunov Exponents**", submetido ao veículo 7th International Conference on Spoken Language Processing 2002, com autoria de Adriano Petry e Dante A. C. Barone. A formatação original do trabalho foi parcialmente modificada.

TEXT-DEPENDENT SPEAKER VERIFICATION USING LYAPUNOV EXPONENTS

Petry, A. and Barone, D. A. C.
{adpetry, barone}@inf.ufrgs.br

Instituto de Informática, Universidade Federal do Rio Grande do Sul - Porto Alegre, Brazil

Abstract

The characterization of a speech signal using nonlinear dynamical features has been focus of intense research lately. In this work, the results obtained with time-dependent largest Lyapunov exponents (TDLEs) in a text-dependent speaker verification task are reported. The baseline system used 10 cepstral coefficients and 10 delta cepstral coefficients, and it is shown how the addition of TDLEs can improve the system's accuracy. Cepstral mean subtraction (CMS) was applied to all features in the tests, as well as silence removal. The telephone speech corpus used, obtained from a subset of CSLU Speaker Recognition corpus, was composed by 91 different speakers, speaking the same sentence.

1. INTRODUCTION

Many physical phenomena present a complex behavior with fluctuations over time. Biological signals, as electroencephalograms (EEGs), electrocardiograms (ECGs), vocal sounds, and measures of arterial blood pressure, represent a great challenge related to its analysis and modeling. A detailed model of vocal tract should consider the time variation of vocal tract shape, the vocal tract resonances, losses due to heat conduction and viscous friction at the vocal tract walls, nasal cavity coupling, softness of the vocal tract walls, the effect of subglottal coupling with vocal tract resonant structure and radiation of sound at the lips [1]. A time-varying linear filter can model the effects of some of these factors, but others are very difficult to model. Some techniques have been proposed in the literature to analyze the non-linearities of dynamical systems, including those systems where it is not currently possible to merge with a mathematical model. They constitute the called Chaos theory or nonlinear dynamical systems theory.

The nonlinear dynamical systems theory can use time series to characterize the dynamical properties of the correspondent system, and extract information from these data. Thus, the speech production can be analyzed by the techniques behind this theory, and the information extracted can be applied to improve the accuracy of many speech processing systems, such as speaker recognition systems. Previous papers [2-9] have worked with speech characterization and analysis using nonlinear dynamical features.

We will explore in this paper the extraction of an important nonlinear dynamical feature, namely largest Lyapunov exponent, from every window of speech signals. This kind of analysis will provide a set of values, in this work referred to as time-dependent largest Lyapunov exponents (TDLEs). Furthermore, we intend to demonstrate that this feature can be successfully used in speech recorded at low sample rates (such as telephone speech), and it contains speaker-dependent information, which can improve the accuracy of a speaker recognition system.

2. STATE SPACE RECONSTRUCTION

In experimental applications, it is often available a set of one-dimensional measurements of a dynamical system that evolves in a multidimensional state space. This scalar time series contains the information available from that system. In many

cases, no further information is available, and an important challenge that has to be solved is the calculation of the system's real multidimensional state space trajectory. After that, measurements that provide important knowledge about the system behavior can be done.

To evaluate the properties of an attractor associated to a time series it is first necessary to reconstruct its evolution in a proper state space. The most common way of reconstructing the full dynamics of a system from scalar time series measurements was proposed by Takens [10]. This method presents an easy and practical implementation. Given a *N*-point time series $x(t_i)$ for $i=1,2,...N$ as follows

$$x(t) = \{x(t_1), x(t_2), ..., x(t_N)\}, \quad (1)$$

the *m*-dimensional vectors are reconstructed, according to Takens delay method [10], as

$$\vec{X}_i = \{x(t_i), x(t_i + p), x(t_i + 2p), ..., x(t_i + (m-1)p)\}, \quad (2)$$

where *p* is called time delay and *m* is the embedding dimension. The $\vec{X}_i$ vectors represent the trajectory of the time series $x(t_i)$ in a *m*-dimensional state space.

The choice of the proper time delay (*p*) and embedding dimension (*m*) values must be made carefully. An excessive small value assigned to time delay produces very similar vectors $\vec{X}_i$ and $\vec{X}_{i+1}$, and consequently an autocorrelated attractor trajectory, probably stretched along the diagonal. When *p* is too large, the reconstructed trajectory becomes too disperse. If the attractor is unfolded into a state space whose embedding dimension is lower than the minimum necessary, there will be vectors that remain close to one another not because of the system dynamics. On the other hand, if the chosen embedding dimension is too high, the number of vectors $\vec{X}_i$ is reduced, and it is a problem for time series composed by limited *N* numbers of points.

A criterion for an intermediate choice of time delay values is based on the analysis of the autocorrelation function. The autocorrelation function provides a measure of the similarity between the samples of a signal, and typically the value of *p* is set as the delay where the autocorrelation function first drops to half of the initial value. Other methods for choosing time delay can be found in [11].

An interesting method to estimate an acceptable minimum embedding dimension is called method of false neighbors [12]. Basically, for each vector of the reconstructed attractor trajectory, unfolded into a *d* embedding dimension phase space, a search for its nearest neighbor vector is made. When the embedding dimension is increased to *d+1*, it is possible to discover the percentage of neighbors that were actually "false" neighbors, and did not remain close because the *d* embedding dimension was too small. When the false neighbors percentage drops to an acceptable value, it is possible to state that the attractor was completely unfolded.

3. LARGEST LYAPUNOV EXPONENT ESTIMATION

Rosenstein *et. al.* [13] proposed a method to estimate the Lyapunov exponents from time series composed by few samples. Good results were obtained for the largest Lyapunov exponent (λ_1) estimation of known systems using less than 1000 samples. This characteristic is very important when dealing with speech, since a speech signal can be considered stationary only during a small window of approximate 30ms [1]. Furthermore, it allows the correct estimation of Lyapunov exponents from speech windows, using speech recorded at low sample rates, such as telephone speech.

Rosenstein's method for largest Lyapunov estimation is outlined as follows: the first step is the reconstruction of the attractor's trajectory in an appropriate state space. After, the nearest neighbor of every vector of the reconstructed trajectory is found. A constraint that nearest neighbors have temporal separations greater than the mean period

of the time series must be satisfied. Doing this, it is possible to consider the pair of neighbors as belonging to different trajectories. When considering two trajectories whose initial conditions are very similar, the trajectories diverge, on average, at an exponential rate characterized by the largest Lyapunov exponent (λ_1):

$$d_j(i) \approx C_j e^{\lambda_1 (i\Delta t)} \quad (3)$$

where $d_j(i)$ is the distance between the jth pair of nearest neighbors after i steps (equals to $i.\Delta t$ seconds, where Δt is the time series sampling period) and C_j is the initial separation between the neighbors.

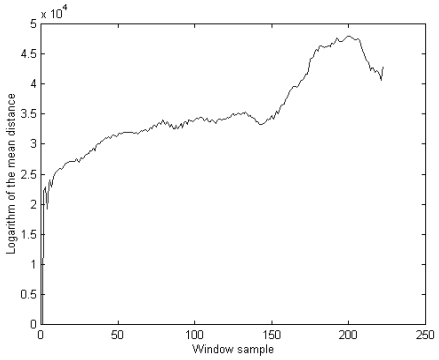

Figure 1: Logarithm of the mean distance evolution between every pair of neighbors from the reconstructed state space of a 30ms window of a speech signal

When the natural logarithm is applied to both sides, the previous equation becomes

$$\ln d_j(i) \approx \ln C_j + \lambda_1 (i\Delta t) \quad (4)$$

If the logarithm of the distance evolution between every pair of neighbors is monitored, it will appear as a set of approximately parallel lines, each with a slope proportional to λ_1. The mean line, calculated from these parallel lines, can be best modeled by applying a least-squares method, and the largest Lyapunov exponent is then estimated as the modeled line slope. Figure 1 shows an example of the logarithm of the mean distance evolution between every pair of neighbors. The m-dimensional vectors were calculated from the state space reconstruction of a 30ms hamming window, obtained from the telephone speech signal in figure 2. It is easy to verify its positive slope, which indicates a positive value for the correspondent largest Lyapunov exponent.

The speech signal showed in figure 2 was extracted from CSLU Speaker Recognition Corpus. The complete information about this corpus can be found in [14]. The speech file of figure 2 was named as “00809al1.wav” in the corpus. By repeating the same process described above in the speech signal in figure 2 for every window of length of 30ms, applied every 10 ms, the TDLEs can be obtained and are showed in figure 3.

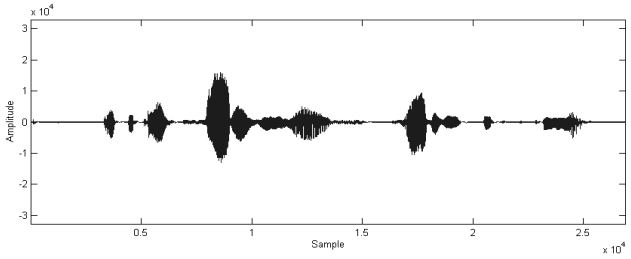

Figure 2: Telephone speech file “00809al1.wav” from CSLU Speaker recognition corpus

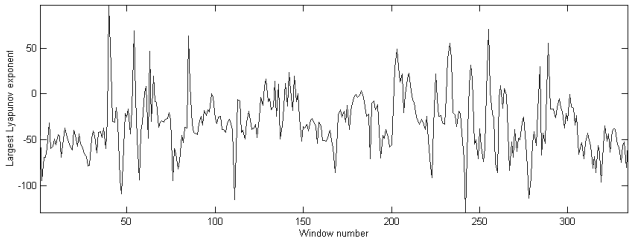

Figure 3: Time-dependent largest Lyapunov exponents from speech signal shown in figure 2

4. BASELINE SYSTEM

A system based on the Bhattacharyya distance was used to evaluate the recognition performance that can be obtained when LP-derived cepstral coefficients are combined with TDLEs. The baseline system in this work used 10 LP-derived cepstral coefficients added to 10 delta cepstral coefficients. Detailed information about cepstral coefficients estimation can be found in [1]. Cepstral mean subtraction (CMS) [15-16] was applied to the features for reduction of channel distortions, and silence frames were discarded based on energy estimation.

Basically, some speech samples of every registered speaker in the system are used to compose the speaker’s voiceprint. It is done by extracting the desired feature parameters from every window of all speech samples from that speaker, and calculating its mean vector and covariance matrix. When an unknown speech sample is presented to the system, its feature parameters are extracted the same way, the mean and covariance matrix are calculated and a distortion measure can be obtained for any registered speaker, using the Bhattacharyya distance for multivariate Gaussian distributions [17]. The functionality of this system is similar to the one described in [18].

Speaker recognition can be classified into speaker identification and speaker verification. In speaker identification task, the unknown speech sample is compared with all registered speaker’s voiceprint. In this case, the unknown speech sample is assigned to the registered speaker whose similarity measure is maximized. For speaker verification, the claimed registered speaker’s identity must be also provided. Then, the unknown speech sample is compared only with the claimed registered speaker’s voiceprint. If the distortion measure obtained is lower than a threshold, the identity is accepted; otherwise, it is rejected by the system. The system used in this work tested speaker verification in a text-dependent task, and its accuracy was compared when TDLEs were added to the baseline features.

4.1 Bhattacharyya distance

In statistics, the proximity degree between two different probability densities is related with the notion of distance measure. One way to calculate this distance between classes is throughout the Bhattacharyya distance. Considering two probability densities $p_1(x)$ and $p_2(x)$, obtained from two different classes of feature parameters, the Bhattacharyya distance [19] is defined as

$$B = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)\, p_2(x)}\, dx \qquad (5)$$

Special cases of this general distance measure can be calculated explicitly to large types of probability densities. An important case refers to the multivariate Gaussian distributions. Considering $p_i(x)$ Gaussian probability densities, it is possible to show [20] that the previous equation can be written as:

$$B = \frac{1}{8}(m_1 - m_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (m_1 - m_2) + \\ + \frac{1}{2} \ln \left(\frac{\det(\Sigma_1 + \Sigma_2)/2}{\sqrt{\det(\Sigma_1)}\sqrt{\det(\Sigma_2)}} \right) \qquad (6)$$

where m_i is the mean vector and Σ_i is the covariance matrix, obtained from the feature parameters of class i.

The Bhattacharyya distance can be applied to a wide variety of known probability distributions, according to the best fit of correspondent probability densities. The assumption of Gaussian density for the parameters is not arbitrary, since it is sufficient that the density be essentially unimodal and approximately Gaussian in the center of its range. These properties are often respected in physical systems. Inspecting the histograms obtained from feature parameters, it is possible to verify that their distributions may be modeled as Gaussian probability densities.

4.2 The speech corpus

The speech corpus used in this work is a subset of CSLU Speaker Recognition corpus. The speakers' speech samples were recorded from digital telephone lines, in various sessions during a two-year period. The sample rate used was 8000 Hz, with resolution of 16 bits per sample. Different transducers were used in the recordings, providing an unmatched condition. In this corpus, every speaker was prompted to repeat the same words and sentences as well as speak freely, allowing the tests of both text-dependent and text-independent systems. Detailed information about CSLU Speaker Recognition corpus can be found in [14].

The subset of the speech corpus used in this work was composed by speech samples from 91 different speakers. All the speakers provided 11 or 12 repetitions of speech samples with the same text: "If it doesn't matter who wins, why do we keep score ?". The speech samples whose information did not correspond to the complete expected sentence were discarded. Every repetition of the sentence above had duration of 3.2 s in mean. For training, 5 repetitions were used to generate every speaker's voiceprint and the others were used for testing the system. The testing speech samples from every speaker were used to measure false rejection (FR) rate when compared with that speaker's voiceprint, and were used to measure false acceptance (FA) rate when compared with other speakers' voiceprint. The tradeoff between these two measures is a function of the decision threshold. The plot of FA versus FR provides the ROC curves, a powerful way of analyzing the performance of a system.

5. EXPERIMENTAL RESULTS

The experimental tests accomplished used the baseline features, composed of 10 cepstral and 10 delta cepstral coefficients. The system accuracy was compared when TDLEs were added. The estimation of largest Lyapunov exponents using Rosenstein's method [13] can provide reliable results with few data. However, considering the sample rate of 8000 Hz used in the speech corpora, a window of 45 ms was used, and 360 samples from the speech window showed to be enough for Lyapunov measures, what was tested and reported in figure 4.

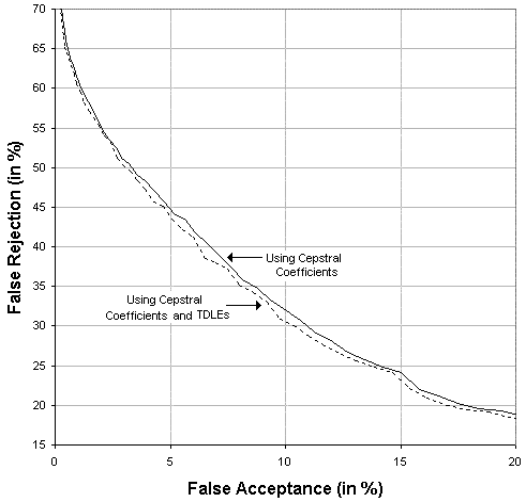

Figure 5: System performance using 45 ms windows

The addition of TDLEs led the Bhattacharyya-based, text-dependent speaker verification system to better performance. Above 45 ms of window length, the assumption of stationarity becomes unreliable.

6. CONCLUSIONS

This work suggested new ideas to construct more robust and reliable speaker recognition systems. Extraction of new information that specifically distinguishes different speakers is very important to continue the development of this area. The standard feature parameters widely used to characterize a speaker can perform this task with relative success. However, some applications where the speaker recognition technology can potentially be introduced are still waiting for more accurate systems. The nonlinear dynamics analysis can see the speech production differently, as the result of a nonlinear dynamical process, bringing up new information to characterize it in a more complete way, and perhaps pointing out a way of improving the accuracy of speaker recognition systems.

Chaos theory has been focus of intense research lately, mainly because of the great development of hardware capabilities. Computational algorithms were developed and used in real-world time series, providing good results even when the number of samples available was not large. In this work, the estimation of TDLEs was possible and accurate using telephone speech when the window length of 45ms was used. The addition of new information led the speaker recognition system under analysis to a better performance in a text-dependet speaker verification task.

Acknowledgments - The authors would like to acknowledge Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) and Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq) for financial support.

7. REFERENCES

[1] Deller Jr., J. R., Proakis, J. G. and Hansen, J. H. L. Discrete-time processing of speech signals, *Prentice Hall*, 1987.

[2] Sabanal, S.; Nakagawa, M. The Fractal Properties of Vocal Sounds and Their Application in the Speech Recognition Model, *Chaos Solitons & Fractals*, vol. 7, no. 11, pp. 1825-1843, 1996.

[3] Kumar, A.; Mullick, S. K. Nonlinear dynamical analysis of speech, *J. Acoust. Soc. Am.* 100 (1), July 1996.

[4] Banbrook, M.; McLaughlin, S.; Mann, I. Speech Characterization and Synthesis by Nonlinear Methods, *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 1, January 1999.

[5] Oliveira, L. P. L.; Roque, W. L.; Custódio, R. F. Lung Sound Analysis with Time-Dependent Fractal Dimensions. *Chaos, Solitons & Fractals*, vol. 10, no. 2, pp.1419-1423, 1999.

[6] Chan, A. M.; Leung, H. Equalization of Speech and Audio Signals Using a Nonlinear Dynamical Approach, *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 3, May 1999.

[7] Kohlmorgen, J.; Muller, K.-R., Rittweger, J.; Pawelzik, K. Identification of non-stationary dynamics in physiological recordings, *Biological Cybernetics*, 2000.

[8] Petry, A.; Barone, D. A. C. Fractal Dimension Applied to Speaker Identification, *IEEE International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, May 2001.

[9] Petry, A.; Barone, D. A. C. Speaker Identification Using Nonlinear Dynamical Features, *Chaos Solitons & Fractals*, vol 13/2, pp 221-231, September 2001.

[10] Takens, F. Detecting strange attractors in turbulence *in Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, edited by D. A. Rand and L. S. Young, Springer-Verlag, Berlin, 1981, vol. 898, pp. 366-381.

[11] Rosenstein, M. T.; Collins, J. J.; De Luca, C. J. Reconstruction expansion as a geometry-based framework for choosing proper delay times, *Physica D* 73 (1994) 82-98.

[12] Kennel, M. B.; Brown, R.; Abarbanel, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Physical Review A*, vol. 45, no. 6, pp. 3403-3411, March 1992.

[13] Rosenstein, M. T.; Collins, J. J.; De Luca, C. J. A practical method for calculating largest Lyapunov exponents from small data sets, *Physica D* 65 (1993) 117-134.

[14] Cole, R.; Noel M.; Noel, V. The CSLU Speaker Recognition Corpus, Proc. *International Conference on Spoken Language Processing*, Sydney, Australia, pp. 3167-3170, November 1998.

[15] Rosenberg, A. E.; Lee, C. H.; Soong, F. K. Cepstral Channel Normalization Techniques for HMM-based Speaker Verification, *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 1835-1838, 1994.

[16] Westphal, M. The Use Of Cepstral Means In Conversational Speech Recognition, Proc. of *Eurospeech'97*, 1997.

[17] O'Shaughnessy, D. Speech Communications Human and Machine, second edition, *IEEE Press*, 547 pp., 2000.

[18] Campbell Jr., J. P. Speaker Recognition: A Tutorial, *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462, September 1997.

[19] Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions, *Bull Calcutta Math. Soc*., vol. 35, pp. 99-109, 1943.

[20] Kailath, T. The Divergence and Bhattacharyya Distance Measures in Signal Selection, *IEEE Trans. On Communication Technology*, vol. Com-15, pp. 52-60, February 1967.

Anexo 5 Speaker recognition using time-dependent largest Lyapunov exponents

Trabalho intitulado "**Speaker Recognition Using Time-dependent Largest Lyapunov Exponents**", submetido ao veículo Computer Speech & Language journal, com autoria de Adriano Petry e Dante A. C. Barone. A formatação original do trabalho foi parcialmente modificada.

Speaker Recognition Using Time-dependent Largest Lyapunov Exponents

ADRIANO PETRY and DANTE AUGUSTO COUTO BARONE

Instituto de Informática, Universidade Federal do Rio Grande do Sul

Av. Bento Gonçalves, 9500 - Campus do Vale - Bloco IV
Bairro Agronomia - Porto Alegre - RS -Brasil
CEP 91501-970 - Brazil

Abstract – The characterization of a speech signal using nonlinear dynamical features has been focus of intense research lately. In this work, the results obtained with time-dependent largest Lyapunov exponents (TDLEs) in a text-dependent speaker verification task are reported. The baseline system used Gaussian mixture models (GMMs), obtained from the adaptation of a universal background model (UBM), for speakers' voiceprint representation. 16 cepstral and 16 delta cepstral features were used in the experiments, and it is shown how the addition of TDLEs can improve the system's accuracy. Cepstral mean subtraction was applied to all features in the tests for channel equalization, as well as silence removal. The telephone speech corpus used, obtained from a subset of CSLU Speaker Recognition corpus, was composed by 91 different speakers.

1. INTRODUCTION

Many physical phenomena present a complex behavior with fluctuations over time. Biological signals, as electroencephalograms (EEGs), electrocardiograms (ECGs), vocal sounds, and measures of arterial blood pressure, represent a great challenge related to its analysis and modeling. A detailed model of vocal tract should consider the time variation of vocal tract shape, the vocal tract resonances, losses due to heat conduction and viscous friction at the vocal tract walls, nasal cavity coupling, softness of the vocal tract walls, the effect of subglottal (lungs and trachea) coupling with vocal tract resonant structure and radiation of sound at the lips [1]. A time-varying linear filter can model the effects of some of these factors, but the remaining ones are very difficult to model. Some techniques have been proposed in the literature to analyze the non-linearities of dynamical systems, including those systems where it is not currently possible to merge with a mathematical model. A set of techniques that can perform this analysis constitutes the called Chaos theory.

The nonlinear dynamical systems theory or just Chaos theory can use time series to characterize the dynamical properties of a system and extract information from these data. Thus, the speech production can be analyzed by the techniques behind this theory, and the information extracted can be applied to improve the accuracy of many speech processing systems, such as speaker recognition systems.

Previous papers [2-9] have worked with speech characterization and analysis using nonlinear dynamical features. Sabanal and Nakagawa [2] used the time-dependent fractal dimensions (TDFDs), extracted through critical exponent method (CEM), and time-dependent multifractal dimensions (TDMFDs) to accomplish a speech recognizer. The target was to recognize Japanese digits using a neural network. Kumar and Mullick [3] estimated Lyapunov exponents, dimension and metric entropy in phonemes signals, divided into eight different types. Banbrook *et al*. [4] extracted correlation dimension,

Lyapunov exponents, and short-term predictability from a corpus of sustained vowels sounds. Oliveira *et al*. [5] proposed the use of TDFDs to evaluate the auscultation of human lung sounds, in order to diagnose diseases associated with breath apparatus. Chan and Leung [6] worked with a nonlinear dynamical technique (MPSV) for enhancing speech signals corrupted by additive noise. Kohlmorgen *et al*. [7] analyzed time series from physiological measures, and developed tools to deal with non-stationary signals. In [8] we made experiences with fractal dimension in a speaker identification system, improving its accuracy. The work [9], also from the authors, showed experiences with fractal dimension, largest Lyapunov exponents and entropy combined with cepstral coefficients in a speaker identification task.

We will explore in this paper the extraction of an important nonlinear dynamical feature, namely largest Lyapunov exponent, from every window of speech signals. This kind of analysis will provide a set of values, in this work referred to as time-dependent largest Lyapunov exponents (TDLEs). Furthermore, we will demonstrate that this feature can be successfully used in speech recorded at low sample rates (such as telephone speech), and it contains speaker-dependent information, which can improve the accuracy of a speaker recognition system.

This work is divided into an introduction, a section that describes the standard state space reconstruction techniques followed by a section focusing on the estimation of the largest Lyapunov exponent. After that, the speaker recognition system and the speech corpus used in the experiments are detailed. Then, the results are showed and final conclusions are discussed.

2. STATE SPACE RECONSTRUCTION

In experimental applications, it is often available a set of one-dimensional measurements of a dynamical system that evolves in a multidimensional state space. This scalar time series contains the information available from that system. In many cases, no further information is available, and an important challenge that has to be solved is the calculation of the system's real multidimensional state space trajectory. After that, measurements that provide important knowledge about the system behavior can be done.

To evaluate the properties of an attractor associated to a time series it is first necessary to reconstruct its evolution in a proper state space. The most common way of reconstructing the full dynamics of a system from scalar time series measurements was proposed by Takens [10]. This method presents an easy and practical implementation. Given a *N*-point time series $x(t_i)$ for $i=1,2,...N$ as follows

$$x(t) = \{x(t_1), x(t_2), ..., x(t_N)\}, \quad (1)$$

the *m*-dimensional vectors are reconstructed, according to Takens delay method [10], as

$$\vec{X}_i = \{x(t_i), x(t_i + p), x(t_i + 2p), ..., x(t_i + (m-1)p)\}, \quad (2)$$

where *p* is called time delay and *m* is the embedding dimension. The $\vec{X}_i$ vectors represent the trajectory of the time series $x(t_i)$ in a *m*-dimensional state space.

The choice of the proper time delay (*p*) and embedding dimension (*m*) values must be made carefully. An excessive small value assigned to time delay produces very similar vectors $\vec{X}_i$ and $\vec{X}_{i+1}$, and consequently an autocorrelated attractor trajectory, probably stretched along the diagonal. When *p* value is too large, the reconstructed trajectory becomes too disperse. If the attractor is unfolded into a state space whose embedding dimension is lower than the minimum necessary, there will be vectors that

remain close to one another not because of the system dynamics. On the other hand, if the chosen embedding dimension is too high, the number of vectors $\vec{X}_i$ is reduced, and it is a problem for time series composed by limited *N* numbers of points.

A criterion for an intermediate choice of time delay values is based on the analysis of the autocorrelation function. The autocorrelation function provides a measure of the similarity between the samples of a signal, and typically the value of *p* is set as the delay where the autocorrelation function first drops to half of the initial value. Other methods for choosing time delay can be found in [11].

An interesting method to estimate an acceptable minimum embedding dimension is called method of false neighbors [12]. Basically, for each vector of the reconstructed attractor trajectory, unfolded into a *d* embedding dimension phase space, a search for its nearest neighbor vector is made. When the embedding dimension is increased to *d+1*, it is possible to discover the percentage of neighbors that were actually "false" neighbors, and did not remain close because the *d* embedding dimension was too small. When the false neighbors percentage drops to an acceptable value, it is possible to state that the attractor was completely unfolded.

A 3-D plot of the reconstructed state space from a 30ms window of telephone speech is shown in figure 1. The time delay value used in the state space reconstruction was 3, based on the analysis of autocorrelation function [11]. The embedding dimension order was equal to 3, using false neighbors technique [12].

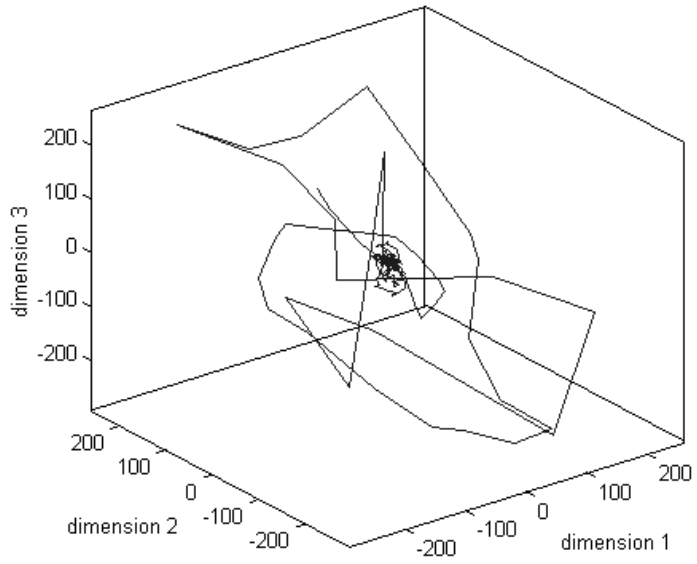

Figure 1: Reconstructed state space from a 30ms window of telephone speech

3. LARGEST LYAPUNOV EXPONENT ESTIMATION

Rosenstein *et. al.* [13] proposed a method to estimate the Lyapunov exponents from time series composed by few samples. Good results were obtained for the largest Lyapunov exponent (λ_1) estimation of known systems using less than 1000 samples. This characteristic is very important when dealing with speech, since a speech signal can be considered stationary only during a small window of approximate 30ms [1]. Furthermore, it allows the correct estimation of Lyapunov exponents from speech windows, using speech recorded at low sample rates, such as telephone speech.

Rosenstein's method for largest Lyapunov estimation is outlined as follows: the first step is the reconstruction of the attractor's trajectory in an appropriate state space. After, the nearest neighbor of every vector of the reconstructed trajectory is found. A constraint that nearest neighbors have temporal separations greater than the mean period of the time series must be satisfied. Doing this, it is possible to consider the pair of neighbors as belonging to different trajectories. When considering two trajectories whose initial conditions are very similar, the trajectories diverge, on average, at an exponential rate characterized by the largest Lyapunov exponent (λ_1):

$$d_j(i) \approx C_j e^{\lambda_1 (i\Delta t)} \qquad (3)$$

where $d_j(i)$ is the distance between the jth pair of nearest neighbors after i steps (equals to $i.\Delta t$ seconds, where Δt is the time series sampling period) and C_j is the initial separation between the neighbors.

When the logarithm is applied to both sides, the previous equation becomes

$$\log d_j(i) \approx \log C_j + \lambda_1 (i\Delta t) \quad (4)$$

If the logarithm of the distance evolution between every pair of neighbors is monitored, it will appear as a set of approximately parallel lines, each with a slope proportional to λ_1. The mean line, calculated from these parallel lines, can be best modeled by applying a least-squares method, and the largest Lyapunov exponent is then estimated as the modeled line slope. Figure 2 shows an example of the logarithm of the mean distance evolution between every pair of neighbors. The m-dimensional vectors were calculated from the state space reconstruction of a 30ms hamming window, obtained from the telephone speech signal in figure 3. It is easy to verify its positive slope, which indicates a positive value for the correspondent largest Lyapunov exponent.

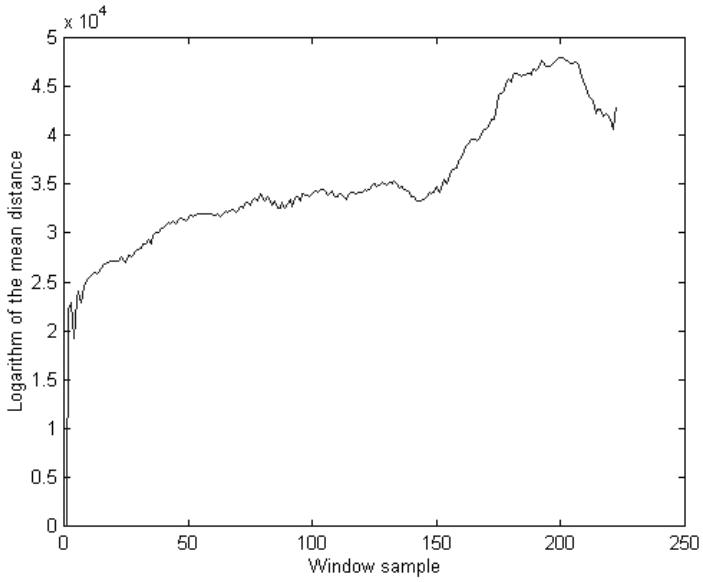

Figure 2: Logarithm of the mean distance evolution between every pair of neighbors from the reconstructed state space of a 30ms window of a speech signal

The speech signal showed in figure 3 was extracted from CSLU Speaker Recognition Corpus, available free of charge for research purposes at CSLU web site [14]. The speech samples from every speaker were recorded at 8000Hz, during a two-year period, using digital telephone lines. The complete information about this corpus can be found in [14]. The speech file of figure 3 was named as "00809al1.wav" in the

corpus. By repeating the same process described above in the speech signal in figure 3 for every window of length of 30ms, applied every 10 ms, the TDLEs can be obtained and are showed in figure 4.

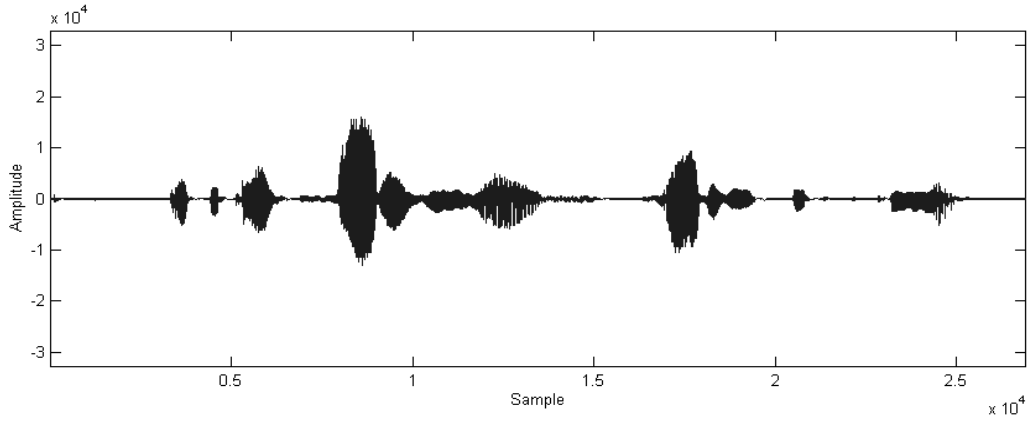

Figure 3: Telephone speech file “00809al1.wav” from CSLU Speaker recognition corpus

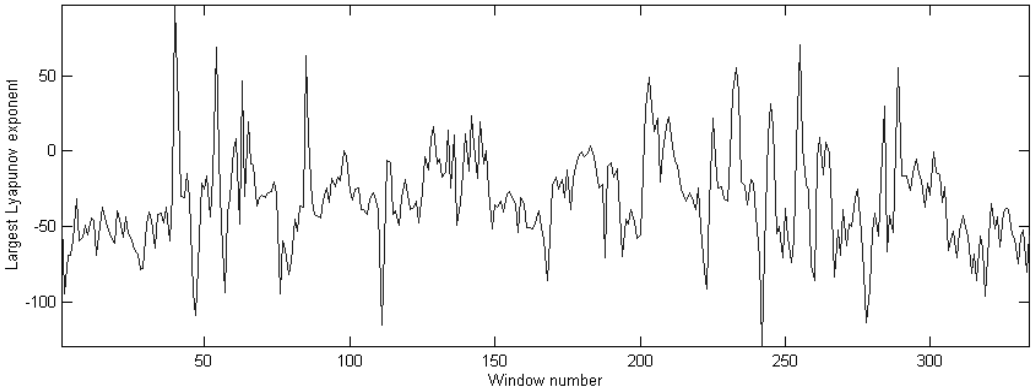

Figure 4: Time-dependent largest Lyapunov exponents from speech signal shown in figure 3

4. BASELINE SYSTEM

Speaker recognition can be classified into speaker identification and speaker verification. In speaker identification task, the unknown speech sample is compared with every registered speaker’s model. In this case, the unknown speech sample is assigned to the registered speaker whose similarity measure is maximized. For speaker verification, focused in this work, the claimed registered speaker’s identity must be also provided. Then, the unknown speech sample is compared with the claimed registered speaker’s model and if the similarity measure is greater than a threshold, the identity is accepted.

A system based on Gaussian mixture models (GMMs) as the speakers’ voiceprint representation, obtained from the adaptation of a universal background model (UBM), was used in the experiments. The system’s recognition performance that can be obtained when LP-derived cepstral coefficients are combined with TDLEs was evaluated. The baseline feature vector in this work was composed by 16 LP-derived cepstral coefficients added to 16 delta cepstral coefficients, extracted from every window of speech. Detailed information about cepstral coefficients estimation can be found in [1]. Cepstral mean subtraction (CMS) [15-16] was applied to the features for

reduction of channel distortions, and silence frames were discarded based on energy estimation.

4.1 GMM-UBM system

GMM-based speaker recognition systems have been extensively used in past years, showing good results. In order to improve speaker recognition performance under mismatched conditions, it was introduced the concept of UBM, where a GMM is then obtained from the adaptation of a general speaker-independent GMM, trained using a large number of speakers. In a GMM-UBM speaker verification approach, a log-likelihood ratio can be estimated using the scores of the sequence of feature vectors $X=\{x_1,\ldots,x_T\}$ applied to both GMMs – the claimed speaker's GMM, denoted as λ_{spk}, and UBM, denoted as λ_{UBM}:

$$\Lambda(X) = \log p(X / \lambda_{spk}) - \log p(X / \lambda_{UBM}) \quad (5)$$

where $p(X/\lambda_i)$ is the likelihood function of the sequence of feature vectors X in model i.

For Gaussian mixture speaker models, the log-likelihood of a model for X is usually computed as

$$\log p(X / \lambda) = \frac{1}{T}\sum_{t=1}^{T} \log\left(\sum_{k=1}^{N} w_k\, p_k(x_t)\right) \quad (6)$$

with mixture weights w_k and Gaussian densities $p_k(x)$. For a set of training vectors, the maximum likelihood parameters are estimated using the iterative Expectation-Maximization (EM) algoritm, well detailed in [17].

In a GMM-UBM speaker recognition system, instead of generating a GMM using the feature vectors extracted from the correspondent speaker's speech samples, every registered speaker's GMM is derived from the UBM, using Bayesian adaptation and a relevance factor. A good description of this technique can be found in [18-19].

Some advantages can be outlined in the choice of GMM-UBM systems. First an intrinsic robustness to speech degradation can be observed, since the comparison is accomplished to both claimed speaker's GMM and the UBM. So, if a corrupted speech is introduced to the system, probably both scores will be reduced, but the log-likelihood ratio tends to be more stable. The adoption of thresholds for speech acceptation or rejection is easier. Another important issue concerns the immediate possibility of updating the speaker's GMM with time, by using low relevance factors that, during a long period, can maintain the speaker's model up-to-date with speaker's speech.

4.2 The speech corpus

The speech corpus used in this work is a subset of CSLU Speaker Recognition corpus. The speakers' speech samples were recorded from digital telephone lines, in various sessions during a two-year period. The sample rate used was 8000 Hz, with resolution of 16 bits per sample. Different transducers were used in the recordings, providing an unmatched condition. Every speaker was prompted to repeat the same words and sentences as well as speak freely, allowing the tests of both text-dependent and text-independent systems. Detailed information about CSLU Speaker Recognition corpus can be found in [14]. The subset of the speech corpus used in this work was composed by speech samples from 91 different speakers. All the speakers provided 11 or 12 repetitions of speech samples with the same text: "If it doesn't matter who wins, why do we keep score ?". The speech samples whose information did not correspond to the complete expected sentence were discarded. Every repetition of the sentence above

had duration of 3.2 s in mean. For UBM training, all repetitions from 41 speakers were used. For the remaining speakers, 5 repetitions were used to adapt the UBM, and develop the correspondent GMM. The other repetitions were used for testing, providing a measure of false rejection (FR) rate when compared with the same speaker's model, and false acceptance (FA) rate when compared with other speakers' model. The tradeoff between these two measures is a function of the decision threshold. The plot of FA versus FR is called ROC curve, a powerful way of analyzing the performance of a system.

5. EXPERIMENTAL RESULTS

The experimental tests accomplished used the baseline feature vectors, composed of 16 cepstral and 16 delta cepstral coefficients. An UBM was generated with speech from 41 different speakers, and every registered speaker's GMM was derived from that UBM, using the features extracted from approximately 16 seconds of speech. The remaining speech samples were used for testing.

The system accuracy was compared when TDLEs were added. The estimation of largest Lyapunov exponents using Rosenstein's method [13] can provide reliable results with few data. However, considering the sample rate of 8000 Hz used in the speech corpora, a window of 30 ms offers only 240 samples for Lyapunov estimation, what probably will not yield the correspondent estimation to accurate values. It can be shown in the ROC curve in figure 5, where a 30 ms hamming window was used in feature extraction. Clearly, the addition of TDLEs degraded the system performance.

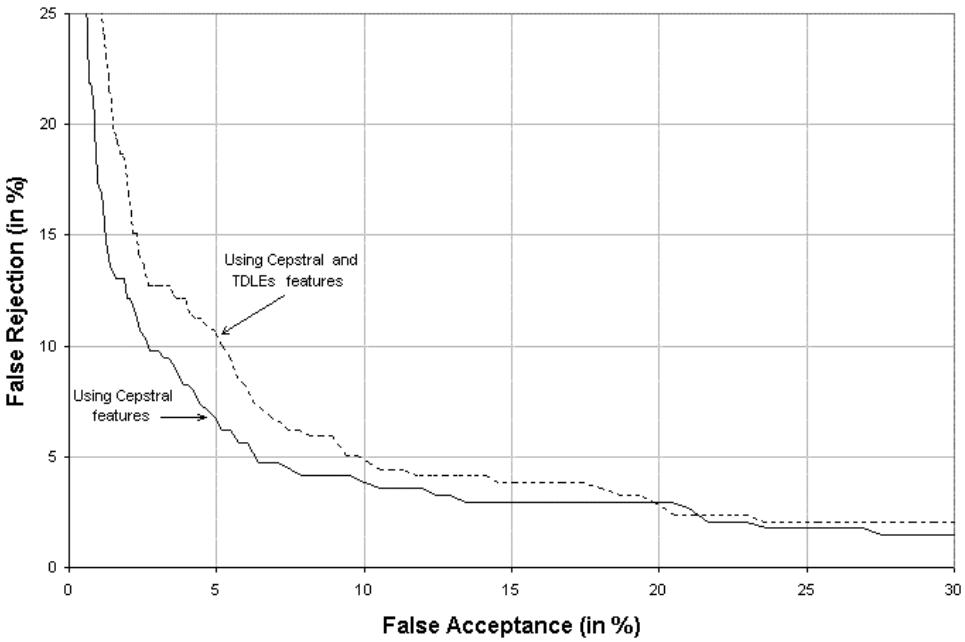

Figure 5: System performance using 30 ms windows

On the other hand, when the window length is increased to 45 ms, 360 samples from the speech window have improved the accuracy of the Lyapunov measures, what was tested and reported in figure 6. In this case, the TDLEs led the system to a comparable performance.

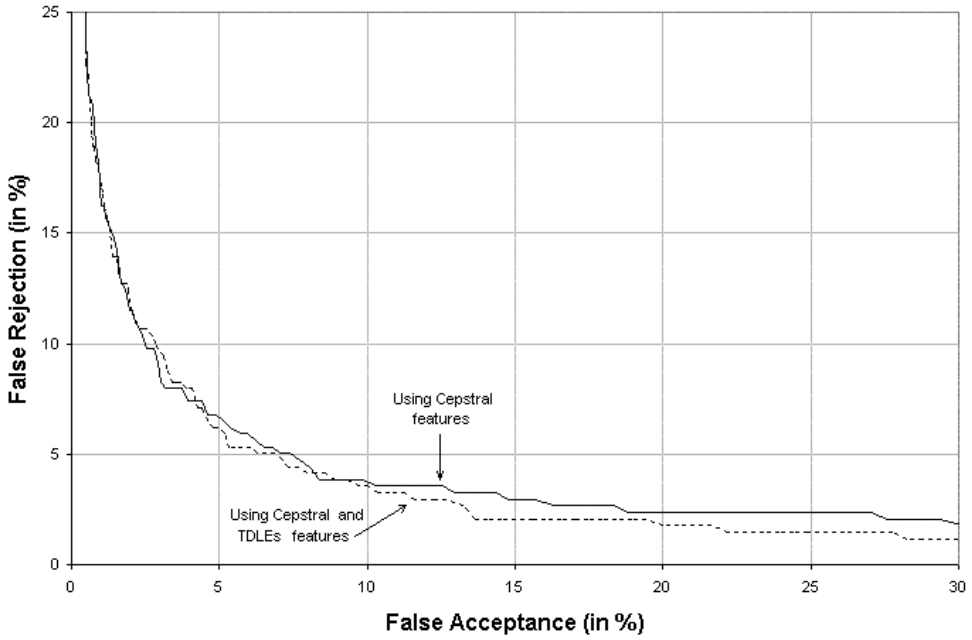

Figure 6: System performance using 45 ms windows

If the window length is again increased to 60ms, the system's accuracy using only cepstral features is practically the same of previous experiments, but with more samples for Lyapunov estimation, these values considerably improved the recognition rate, as shown in figure 7. Above 60 ms of window length, the assumption of stationarity becomes unreliable.

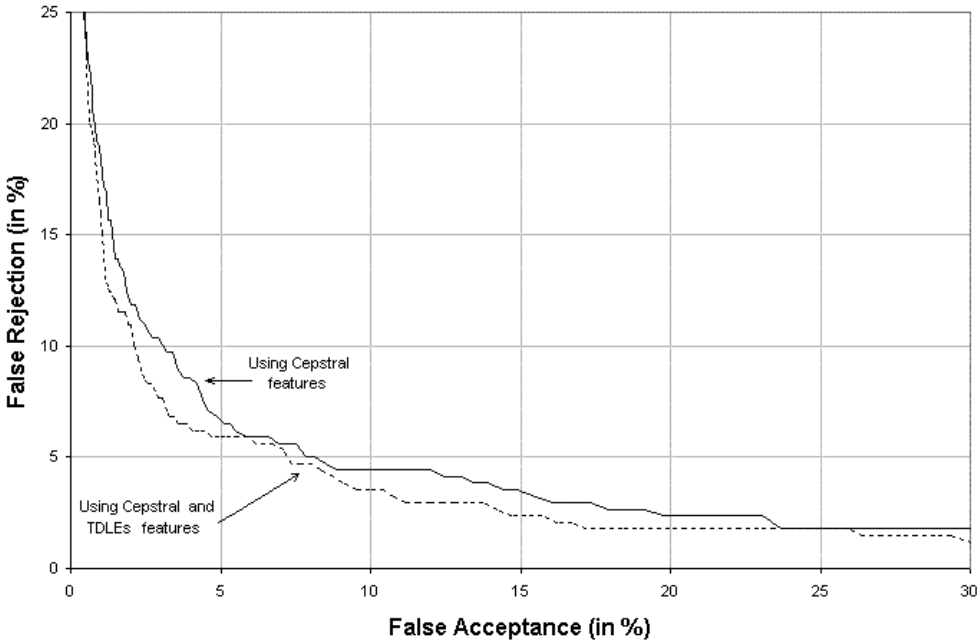

Figure 7: System performance using 60 ms windows

6. CONCLUSIONS

This work suggested new ideas to construct more robust and reliable speaker recognition systems. Extraction of new information that specifically distinguishes different speakers is very important to continue the development of this area. The standard feature parameters widely used to characterize a speaker can perform this task with relative success. However, some applications where the speaker recognition technology may be potentially introduced can be still waiting for more accurate systems. The nonlinear dynamics analysis can see the speech production differently, as the result of a nonlinear dynamical process, bringing up new information to characterize it in a more complete way, and perhaps pointing out a way of improving the accuracy of speaker recognition systems.

Chaos theory has been focus of intense research lately, mainly because of the great development of hardware capabilities. Computational algorithms were developed and used in real-world time series, providing good results even when the number of samples available was not large. In this work, the estimation of TDLEs was possible and accurate using telephone speech when the window length greater than 45ms was used. The addition of new information led the speaker recognition system under analysis to a better performance in verification task.

Acknowledgments - The authors would like to acknowledge Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) and Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq) for financial support.

REFERENCES

[1] Deller Jr., J. R., Proakis, J. G. and Hansen, J. H. L. Discrete-time processing of speech signals, *Prentice Hall*, 1987.

[2] Sabanal, S.; Nakagawa, M. The Fractal Properties of Vocal Sounds and Their Application in the Speech Recognition Model, *Chaos Solitons & Fractals*, vol. 7, no. 11, pp. 1825-1843, 1996.

[3] Kumar, A.; Mullick, S. K. Nonlinear dynamical analysis of speech, *J. Acoust. Soc. Am.* 100 (1), July 1996.

[4] Banbrook, M.; McLaughlin, S.; Mann, I. Speech Characterization and Synthesis by Nonlinear Methods, *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 1, January 1999.

[5] Oliveira, L. P. L.; Roque, W. L.; Custódio, R. F. Lung Sound Analysis with Time-Dependent Fractal Dimensions. *Chaos, Solitons & Fractals*, vol. 10, no. 2, pp.1419-1423, 1999.

[6] Chan, A. M. and Leung, H. Equalization of Speech and Audio Signals Using a Nonlinear Dynamical Approach, *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 3, May 1999.

[7] Kohlmorgen, J.; Muller, K.-R., Rittweger, J.; Pawelzik, K. Identification of non-stationary dynamics in physiological recordings, *Biological Cybernetics*, 2000.

[8] Petry, A. and Barone, D. A. C. Fractal Dimension Applied to Speaker Identification, *IEEE International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, May 2001.

[9] Petry, A. and Barone, D. A. C. Speaker Identification Using Nonlinear Dynamical Features, *Chaos Solitons & Fractals*, vol 13/2, pp 221-231, September 2001.

[10] Takens, F. Detecting strange attractors in turbulence *in Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, edited by D. A. Rand and L. S. Young, Springer-Verlag, Berlin, 1981, vol. 898, pp. 366-381.

[11] Rosenstein, M. T.; Collins, J. J.; De Luca, C. J. Reconstruction expansion as a geometry-based framework for choosing proper delay times, *Physica D* 73 (1994) 82-98.

[12] Kennel, M. B., Brown, R. and Abarbanel, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Physical Review A*, vol. 45, no. 6, pp. 3403-3411, March 1992.

[13] Rosenstein, M. T.; Collins, J. J.; De Luca, C. J. A practical method for calculating largest Lyapunov exponents from small data sets, *Physica D* 65 (1993) 117-134.

[14] Cole, R.; Noel M.; Noel, V. The CSLU Speaker Recognition Corpus, Proc. *International Conference on Spoken Language Processing*, Sydney, Australia, pp. 3167-3170, November 1998.

[15] Rosenberg, A. E.; Lee, C. H.; Soong, F. K. Cepstral Channel Normalization Techniques for HMM-based Speaker Verification, *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 1835-1838, 1994.

[16] Westphal, M. The Use Of Cepstral Means In Conversational Speech Recognition, Proc. *Eurospeech'97*, 1997.

[17] Van Vuuren, S. Speaker Verification in Time-feature Space, Ph. D. thesis, Oregon Graduate Institute of Scienceand Technology, March 1999.

[18] Reynolds, D. A. Comparison of background normalization methods for text-independent speaker verification. Proc. of the European Conference on Speech Communication and Technology, pp. 963-967, September 1997.

[19] Reynolds, D. A.; Quatieri, T. F.; Dunn, R. B. Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, vol. 10, no. 1, pp. 19-41, January 2000.

Bibliografia

[ABA 93] ABARBANEL, H. D. I. et al. The Analysis of Observed Chaotic Data in Physical Systems. **Reviews of Modern Physics**, [S.l.], v. 65, n. 4, p. 1331-1392, Oct. 1993.

[ADA 97] ADAMI, A. G**. Sistema de Reconhecimento de Locutor utilizando Redes Neurais Artificiais**. 1997. 86p. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[ADA 98] ADAMI, A. G.; DORNELES, R. V. Uma abordagem à verificação de locutor utilizando coarticulação e redes neurais artificiais. In: CONGRESSO DE ENGENHARIA E INFORMÁTICA, 1998. **Anais...** Buenos Aires: [s.n.], 1998. v. 4, p. 82-90.

[ARC 2002] ARCHIEVES BUILDERS. **Evolution of Intel Microprocessors**: 1971 to 2003. Disponível em: <www.archivebuilders.com/ whitepapers/22016v010h.html>. Acesso em: maio 2002.

[ATA 76] ATAL, B. S. Automatic Recognition of Speakers from Their Voices. **Proceedings of the IEEE**, New York, v. 64, p. 460-475, Apr. 1976.

[BHA 43] BHATTACHARYYA, A. On a measure of divergence between two statistical populations defined by their probability distributions. **Bull. Calcutta Math. Soc**., [S.l.], v. 35, p. 99-109, 1943.

[BAD 84] BADII, R.; POLITI, A. Hausdorff Dimension and Uniformity Factor of Strange Attractors. **Physical Review Letters**, [S.l.], v. 52, n. 19, p. 1661-1664, May 1984.

[BAN 99] BANBROOK, M.; McLAUGHLIN, S.; MANN, I. Speech Characterization and Synthesis by Nonlinear Methods. **IEEE Transactions on Speech and Audio Processing**, New York, v. 7, n. 1, Jan. 1999.

[BOH 2001] BOHEZ, E. L. J.; SENEVIRATHNE, T. R. Speech recognition using fractals. **Pattern Recognition**, [S.l.], v. 34, p. 2227-2243, 2001.

[ÇAM 93] ÇAMBEL, A. B. **Applied Chaos Theory**: a paradigm for complexity. [S.l.]: Academic Press, 1993. 246 p.

[CAM 97] CAMPBELL, J. P. Speaker Recognition: A Tutorial. **Proceedings of the IEEE**, New York, v. 85, n. 9, Sept. 1997.

[CAS 97] CASAGRANDE, R.; CABRAL JUNIOR, E. F. Time-Delay Neural Network Applied to Speaker Recognition. In: CONGRESSO BRASILEIRO DE REDES NEURAIS, 3., 1997. **Anais...** Florianópolis: [s.n.], 1997. p. 319-323.

[CHA 99] CHAN, A. M.; LEUNG, H. Equalization of Speech and Audio Signals Using a Nonlinear Dynamical Approach. **IEEE Transactions on Speech and Audio Processing**, [S.l.], v. 7, n. 3, May 1999.

[COH 85] COHEN, A.; PROCACCIA, I. Computing the Kolmogorov Entropy from Time Signals of Dissipative and Conservative Dynamical Systems. **Physical Review A**, [S.l.], v. 31, n. 3, p. 1872-1882, Mar. 1985.

[CUS 99] CUSTÓDIO, R. F. **Análise Não-linear no Reconhecimento de Padrões Sonoros**: estudo de caso para sons pulmonares. 1999. 119p. Tese (Doutorado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[DEL 87] DELLER JUNIOR., J. R.; PROAKIS, J. G.; HANSEN, J. H. L. **Discrete-time Processing of Speech Signals.** [S.l.]: Prentice Hall, 1987. 908 p.

[ECK 85] ECKMANN, J. P.; RUELLE, D. Ergodic Theory of Chaos. **Reviews of Modern Physics**, [S.l.], v. 57, n. 3, July 1985.

[FAR 94] FARREL, K. R.; MAMMONE, R. J. ASSALEH, K. T. Speaker recognition using neural networks and conventional classifiers. **IEEE Transactions on Speech and Audio Processing**, New York, v. 2, n. 1, Jan. 1994.

[FER 94] FERRARA, N. F.; PRADO, C. P. C. **Caos**: uma Introdução. São Paulo: E. Blücher, 1994.

[FIN 97] FINAN, R. A.; SAPELUK, A. T.; DAMPER, R. I. VQ Score Normalisation for Text-dependent and Text-independent Speaker Recognition. In: AUDIO- AND VIDEO-BASED BIOMETRICS PERSON AUTHENTICATION, AVBPA, 1997, Switzerland. **Proceedings...** Switzerland: [s.n.], 1997. p. 211-218.

[FUR 97] FURUI, S. Recent Advances in Speaker Recognition In: AUDIO- AND VIDEO-BASED BIOMETRICS PERSON AUTHENTICATION, AVBPA, 1997, Switzerland. **Proceedings...** Switzerland: [s.n.], 1997. p. 237-252.

[GIL 89] GILLICK, L.; COX, S. J. Some statistical issues in the comparison of speech recognition algoritms. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, 1989. **Proceedings...** Glascow, Scotland: [s.n.], 1989. p. 532-535.

[GOP 99] GOPALAN, K.; ANDERSON, T. R., CUPPLES, E. J. A comparison of speaker identification results using features based on cepstrum and Fourier-Bessel expansion. **IEEE Transactions on Speech and Audio Processing**, New York, v. 7, n. 3, May 1999.

[GRA 83a] GRASSBERGER, P.; PROCACCIA, I. Measuring the Strangeness of Strange Attractors. **Physica D**, [S.l.], v. 9, p. 189-208, 1983.

[GRA 83b] GRASSBERGER, P.; PROCACCIA, I. Estimation of the Kolmogorov Entropy from a Chaotic Signal. **Physical Review A**, [S.l.], v. 28, n. 4, Oct. 1983.

[GUO 2002] GUO, C. et al. A study on fractal properties of Mandarin speech. **International Journal of Non-linear Mechanics**, New York, v. 37, p. 409-417, 2002.

[HE 97] HE, J.; LIU, L.; PALM G. A new codebook training algorithm for vq-based speaker recognition. In: IEEE INTERNATIONAL

CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, Munich, 1997. **Proceedings...** Munich: [s.n.], 1997. v. 2, p. 1091-1094.

[HER 94] HERMANSKY, H.; N. MORGAN, RASTA processing of speech, **IEEE Transactions on Speech and Acoustics**, [S.l.], v. 2, n.4, p. 587-589, Oct. 1994.

[KAI 67] KAILATH, T. The Divergence and Bhattacharyya Distance Measures in Signal Selection. **IEEE Transactions on Communication Technology**, [S.l.], v. com-15, n. 1, p. 52-60, Feb. 1967.

[KAP 79] KAPLAN, J.; YORKE, J. Functional Differential Equations and Approximations of Fixed Points. **Springer Lecture Notes in Mathematics**, [S.l.], v. 730, 1979

[KEN 92] KENNEL, M. B.; BROWN, R.; ABARBANEL, H. D. I. Determining Embedding Dimension for Phase-space Reconstruction using a Geometrical Construction. **Phys. Rev A**, [S.l.], v. 45, n. 6, Mar. 1992.

[KLA 87] KLATT, D. H. Review of text-to-speech conversion for English. **Journal of the Acoustical Society of America**, [S.l.], v. 82, p. 737-793, 1987.

[KOH 2000] KOHLMORGEN, J. et al. Identification of non-stationary dynamics in physiological recordings. **Biological Cybernetics**, 2000. Disponível em:<www.researchindex.com>. Acesso em: abr. 2002.

[KOS 91] KOSTELICH, E. J.; YORKE, J. A. Noise Reduction: Finding the Simplest Dynamical System Consistent with the Data. **Physica D**, [S.l.], v. 41, p. 183-196, 1990.

[KUM 94] KUMAR, A. **Nonlinear Dynamical Analysis and Predictive Coding of Speech**. 1994. Ph.D. Thesis in Electrical Engineering, Indian Institute of Technology, Kanpur.

[KUM 96] KUMAR, A.; MULLICK, S. K. Nonlinear Dynamical Analysis of Speech. **J. Acoust. Soc. Am.**, [S.l.], v. 100, n. 1, July 1996.

[LI 2000] LI, Q.; JUANG, B.; ZHOU, Q.; LEE, C. Automatic Verbal Information Verification for User Authentication. **IEEE Transactions on Speech and Audio Processing**, [S.l.], v. 8, n. 5, Sept. 2000.

[LIP 87] LIPPMANN, R. P. An Introduction to Computing with Neural Nets. **IEEE ASSP Magazine**, [S.l.], p. 4-22, Apr. 1987.

[MAK 85] MAKHOUL, J.; ROUCOS, S.; GISH, H. Vector Quantization in Speech Coding. **Proceedings of the IEEE**, [S.l.], v. 73, n. 11, p. 1551-1588, Nov. 1985.

[MAR 77] MARKEL, J. D.; OSHIKA, B. T.; GRAY, A. H. Long-term Feature Averaging for Speaker Recognition. **IEEE Transactions Acoustic, Speech, and Signal Processing**, [S.l.], v. ASSP-25, p. 330-337, Aug. 1977.

[MAT 94] MATSUI, T.; FURUI, S. Comparision of Text-independent Speaker Recognition methods using VQ-distortion and discret/continous HMM's. **IEEE Transactions on Speech and Audio Processing**, [S.l.], v. 2, n. 3, July 1994.

[MCC 43] MCCULLOUGH, W. S.; PITTS, W. H. A Logical Calculus of Ideas Immanent in Nervous Activity. **Bull Math Biophysics**, [S.l.], v.5, p. 115-133, 1943.

[MIY 2001] MIYAJIMA, C. et al. Speaker identification using gaussian mixture models based on multi-space probability distribution. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, 2001. **Proceedings...** Salt Lake City: [s.n.], May 2001.

[MOR 2001] MORI, K.; NAKAGAWA, S. Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, 2001. **Proceedings...** Salt Lake City: [s.n.], May 2001.

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

[OLI 99] OLIVEIRA, L. P. L.; ROQUE, W. L.; CUSTÓDIO, R. F. Lung Sound Analysis with Time-Dependent Fractal Dimensions. **Chaos, Solitons & Fractals**, [S.l.], v. 10, n. 2, p.1419-1423, 1999.

[PET 99a] PETRY, A.; BARONE, D. A. C. Sistema para Controle de Elevadores por Voz. In: CONFERENCIA LATINOAMERICANA DE INFORMÁTICA, CLEI, 25., 1999, Asuncion. **Memorias...** Asuncion: Universidad Autonoma de Asuncion, 1999. v.1, p. 63-74.

[PET 99b] PETRY, A.; BARONE, D. A. C. Speaker Verification Applied to an Elevator Safety System. In: THE INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING APPLICATIONS AND TECHNOLOGY, ICSPAT, 1999, Orlando. **Proceedings...** Orlando: [s.n.], 1999.

[PET 99c] PETRY, A.; BARONE, D. A. C.; ADAMI, A. G. REVOX Voice Recognition System Applied to Industry Control. In: PROTEM-CC-PHASE III – PROJECTS INTERNATIONAL EVALUATION, 2., 1999, Rio de Janeiro. **Proceedings...** Brasília: CNPQ, 1999. p. 347-376.

[PET 99d] PETRY, A.; BARONE, D. A. C. Practical Applications of Speech Processing for Speaker Verification. In: INTERNATIONAL WORKSHOP ON NONLINEAR DYNAMICS OF ELECTRONIC SYSTEMS, NDES, 7., 1999. **Proceedings...** Dinamarca: [s.n.], 1999. p. 141-144.

[PET 99e] PETRY, A.; ZANUZ, A.; BARONE, D. A. C. Utilização de Técnicas de Processamento Digital de Sinais para a Identificação Automática de Pessoas pela Voz. In: SIMPÓSIO SOBRE SEGURANÇA EM INFORMÁTICA, SSI, 1999, São Paulo. **Anais...** São Paulo: [s.n.], 1999.

[PET 2000a] PETRY, A.; ZANUZ, A.; BARONE, D. A. C. Reconhecimento automático de pessoas pela voz através de técnicas de processamento digital de sinais. In. WORKSHOP EM INTERNET, LINUX E APLICAÇÕES, 3., 2000, São José dos Campos. **Anais...** São Paulo: [s.n.], 2000.

[PET 2000b] PETRY, A.; ZANUZ, A.; BARONE, D. A. C. Bhattacharyya Distance Applied to Speaker Identification. In: THE INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING APPLICATIONS AND TECHNOLOGY, ICSPAT, 2000, Dallas. **Proceedings...** Dallas: [s.n.], 2000.

[PET 2000c] PETRY, A. **Estudo sobre Aplicabilidade da Teoria de Sistemas Dinâmicos Não-Lineares para o Reconhecimento Automático de Locutor**. 2000. 82p. Exame de Qualificação (Doutorado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[PET 2001] PETRY, A.; BARONE, D. A. C. Fractal Dimension Applied to Speaker Identification. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, 2001. **Proceedings...** Salt Lake City: [s.n.], May 2001.

[PET 2002a] PETRY, A.; BARONE, D. A. C. Speaker Identification Using Nonlinear Dynamical Features. **Chaos, Solitons & Fractals**, [S.l.], v. 13, n. 2, p. 221-231, Feb. 2002.

[PET 2002b] PETRY, A.; BARONE, D. A. C. Text-dependent Speaker Verification using Lyapunov Exponents. Submetido para International Conference on Spoken Language Processing, ICSLP, 7., Denver, Sept. 2002.

[PET 2002c] PETRY, A.; BARONE, D. A. C. Speaker Recognition Using Time-dependent Largest Lyapunov Exponents. Submetido para Computer Speech and Language.

[PIC 93] PICONE, J. W. Signal Modeling Techniques in Speech Recognition. **Proc. of the IEEE**, [S.l.], v. 81, n. 9, Sept. 1993.

[POR 95] PORT, R.; CUMMINS, F.; GASSER, M. **A dynamic approach to rhythm in language:** toward a temporal phonology. Bloomington, Indiana: University Cognitive Science Program, 1995. (Technical Report, n. 150)

[PRO 96] PROAKIS, J. G.; MANOLAKIS, D. G. **Digital Signal Processing:** principles, algoritms, and applications. New Jersey: Prentice Hall, 1996. 968 p.

[QUA 98] QUATIERI, T. F.; REYNOLDS, D. A.; O'LEARY, G. Magnitude-only estimation of handset nonlinearity with application to speaker recognition. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, 1998. **Proceedings...** [S.l.: s.n.], 1998.

[QUA 2000] QUATIERI, T. F.; REYNOLDS, D. A.; O'LEARY, G. C. Estimation of Handset Nonlinearity with Application to Speaker Recognition. **IEEE Transactions on Speech and Audio Processing**, New York, v. 8, n. 5, Sept. 2000.

[RAB 78] RABINER, L. R.; SCHAFER, R. W. **Digital Processing of Speech Signals**. New Jersey: Prentice Hall, 1978. 512 p.

[RAB 93] RABINER, L.; JUANG, B. H. **Fundamentals of Speech Recognition.** New Jersey: Prentice Hall, 1993. 507 p.

[REY 92] REYNOLDS, D. A. **A Gaussian mixture modeling approach to text-independent speaker identification**. 1992. Ph.D. Thesis, Georgia Institute of Technology, Atlanta.

[REY 95] REYNOLDS, D. A.; ROSE, R. C. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. **IEEE Transactions on Speech and Audio Processing**, New York, v. 3, n. 1, Jan. 1995.

[REY 97] REYNOLDS, D. A. Comparison of background normalization methods for text-independent speaker verification. In: EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY, 1997. **Proceedings...** [S.l.: s.n.], 1997. p. 963-967.

[REY 2000] REYNOLDS, D. A.; QUATIERI, T. F.; DUNN, R. B. Speaker Verification Using Adapted Gaussian Mixture Models. **Digital Signal Processing**, [S.l.], v. 10, n. 1/2/3, p. 19-41, Jan./Apr./July 2000.

[ROS 76] ROSENBERG, A. E. Automatic Speaker Verification: A Review. **Proceedings of the IEEE**, [S.l.], v. 64, p. 475-487, Apr. 1976.

[ROS 90] ROSENBERG, A. E.; Lee, C.-H.; SOONG, F. K. Sub-word unit talker verification using hidden markov models. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, Albuquerque, 1990. **Proceedings...** Albuquerque: [s.n.], 1990.

[ROS 93] ROSENSTEIN, M. T.; COLLINS, J. J.; DE LUCA, C. J. A practical method for calculating largest Lyapunov exponents from small data sets. **Physica D,** [S.l.], v. 65, p. 117-134, 1993.

[ROS 94] ROSENSTEIN, M. T.; COLLINS, J. J.; DE LUCA, C. J. A. Reconstruction expansion as a geometry-based framework for choosing proper delay times. **Physica D**, [S.l.], v. 73, p. 82-98, 1994.

[ROS 94b] ROSENBERG, A. E.; LEE, C. H.; SOONG, F. K. Cepstral Channel Normalization Techniques for HMM-based Speaker Verification. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, 1994. **Proceedings...** [S.l.: s.n.], 1994. p. 1835-1838.

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

[SCH 91] SCHREIBER, T.; GRASSBERGER, P. A Simple Noise-reduction Method for Real Data. **Physics Letters A,** [S.l.], v. 160, p. 411-418, 1991.

[SCI 99] SCIAMARELLA, D.; MINDLIN, G. B. Topological Structure of Chaotic Flows from Human Speech Data. **Phys. Rev. Letters**, [S.l.], v. 82, n. 7, Feb. 1999.

[SOO 87] SOONG, F. K.; ROSENBERG, A. E.; RABINER, L. R. A Vector Quantization Approach to Speaker Recognition. **AT&T Technical Journal**, [S.l.], v. 66, p. 14-26, Mar. 1987.

[STA 98] STAM, C. J.; PIJN, J. P. M.; PRITCHARD, W. S. Reliable detection of nonlinearity in experimental time series with strong periodic components. **Physica D**, [S.l.], v. 112, p. 361-380, 1998.

[TAK 81] TAKENS, F. Detecting Strange Attractors in Turbulence. **Lect. Notes in Mathematics**, Berlin, v. 898, p. 366-381, 1981.

[TER 83] TERMONIA, Y.; ALEXANDROWICZ, Z. Fractal Dimension of Strange Attractors from Radius versus Size of Arbitrary Clusters. **Physical Review Letters**, [S.l.], v. 51, n. 14, p. 1265-1268, Oct. 1983.

[VER 99] VERGIN, R.; O´SHAUGHNESSY, D. On the use of Some Divergence Measures in Speaker Recognition. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, Phoenix, 1999. **Proceedings...** Phoenix: [S.l.], 1999.

[VIZ 98] VIZZOTTO, G. M.; PETRY, A.; BARONE, D. A. C. Utilização de um Algoritmo de Alinhamento Temporal Dinâmico para o Reconhecimento de Voz. In: **Rumos da Pesquisa**: múltiplas trajetórias. Porto Alegre: UFRGS, PROPESP, 1998. p. 293-303.

[VUU 99] VUUREN, S. V. **Speaker Verification in a Time-Feature Space**. 1999. 178p. Ph.D. Thesis, Center for Spoken Language Understanding, CSLU, Oregon Graduate Institute, Oregon.

[YEG 98] YEGNANARAYANA, B.; RAYMOND, N. J. V. Extraction of Vocal-tract system Characteristics from Speech Signals. **IEEE Transactions on Speech and Audio Processing**, New York, v. 6, n. 4, July 1998.

[YEG 2001] YEGNANARAYANA, B.; REDDY, K.; KISHORE, S. Source and System Features for Speaker Recognition using AANN Models. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, 2001. **Proceedings...** Salt Lake City: [s.n.], May 2001.

[YUA 99] YUAN, Z.; XU, B.; YU, C. Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification. **IEEE Transactions on Speech and Audio Processing**, New York, v. 7, n. 1, Jan. 1999.

[ZIL 98] ZILOVIC, M. S.; RAMACHANDRAN, R. P.; MAMMONE, R. J. Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions. **IEEE Transactions on Speech and Audio Processing**, New York, v. 6, n. 3, May 1998.

[WES 97] WESTPHAL, M. The Use Of Cepstral Means In Conversational Speech Recognition, In: EUROPEAN CONFERENCE ON SPEECH COMMUNICATION, 1997. **Proceedings...** [S.l.: s.n.], 1997.

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