Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação
Departamento de Comunicações

USO DE PARÂMETROS MULTIFRACTAIS NO RECONHECIMENTO DE LOCUTOR

Autor(a): Diana Cristina González González

Orientador: Prof. Dr. Lee Luan Ling

Co-Orientador: Prof. Dr. Fábio Violaro

Tese de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: Telecomunicações e Telemática.

Banca Examinadora

Prof. Dr. Lee Luan Ling — DECOM/FEEC/UNICAMP
Prof. Dr. Aldebaro Barreto da Rocha Klautau Junior—UFPA
Prof. Dr. Romis Ribeiro de Faissol Attux— DCA/FEEC/UNICAMP

Campinas – SP
30 Setembro 2011

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

González González, Diana Cristina
G589u Uso de parâmetros multifractais no reconhecimento de locutor / Diana Cristina González González. -- Campinas, SP: [s.n.], 2011.

Orientadores: Lee Luan Ling , Fábio Violaro.
Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Multifractal. 2. Reconhecimento automático da voz. 3. Gaussian distribution. 4. Sistema de processamento da fala. I. Ling, Lee Luan. II. Violaro, Fábio . III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Título em Inglês: Use of multifractal parameters for speaker recognition
Palavras-chave em Inglês: Multifractal, Automatic speech recognition, Gaussian distribution, Speech processing system
Área de concentração: Telecomunicações e Telemática
Titulação: Mestre em Engenharia Elétrica
Banca examinadora: Aldebaro Barreto da Rocha Klautau Junior, Romis Ribeiro de Faissol Attux
Data da defesa: 30-09-2011
Programa de Pós Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidata: Diana Cristina González González

Data da Defesa: 30 de setembro de 2011

Titulo da Tese: "Uso de Parâmetros Multifractais no Reconhecimento de Locutor"

Prof. Dr. Lee Luan Ling (Presidente): _____

Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior: _____

Prof. Dr. Romis Ribeiro de Faissol Attux: _____

Resumo

Esta dissertação apresenta a implementação de um sistema de Reconhecimento Automático de Locutor (ASR). Este sistema emprega um novo parâmetro de características de locutor baseado no modelo multifractal "VVGM" (*Variable Variance Gaussian Multiplier*). A metodologia adotada para o desenvolvimento deste sistema foi formulada em duas etapas. Inicialmente foi implementado um sistema ASR tradicional, usando como vetor de características os MFCCs (Mel-Frequency Cepstral Coefficients) e modelo de mistura gaussiana (GMM) como classificador, uma vez que é uma configuração clássica, adotada como referência na literatura. Este procedimento permite ter um conhecimento amplo sobre a produção de sinais de voz, além de um sistema de referência para comparar o desempenho do novo parâmetro VVGM. A segunda etapa foi dedicada ao estudo de processos multifractais em sinais de fala, já que eles enfatizam-se na análise das informações contidas nas partes não estacionárias do sinal avaliado. Aproveitando essa característica, sinais de fala são modelados usando o modelo VVGM. Este modelo é baseado no processo de cascata multiplicativa binomial, e usa as variâncias dos multiplicadores de cada estágio como um novo vetor de característica.

As informações obtidas pelos dois métodos são diferentes e complementares. Portanto, é interessante combinar os parâmetros clássicos com os parâmetros multifractais, a fim de melhorar o desempenho dos sistemas de reconhecimento de locutor.

Os sistemas propostos foram avaliados por meio de três bases de dados de fala com diferentes configurações, tais como taxas de amostragem, número de falantes e frases e duração do treinamento e teste. Estas diferentes configurações permitem determinar as características do sinal de fala requeridas pelo sistema. Do resultado dos experimentos foi observado que o sistema de identificação de locutor usando os parâmetros VVGM alcançou taxas de acerto significativas, o que mostra que este modelo multifractal contém informações relevantes sobre a identidade de cada locutor. Por exemplo, a segunda base de dados é composta de sinais de fala de 71 locutores (50 homens e 21 mulheres) digitalizados a 22,05 kHz com 16 bits/amostra. O

treinamento foi feito com 20 frases para cada locutor, com uma duração total de cerca de 70 s. Avaliando o sistema ASR baseado em VVGM, com locuções de teste de 3 s de comprimento, foi obtida uma taxa de reconhecimento de 91,30%. Usando estas mesmas condições, o sistema ASR baseado em MFCCs atingiu uma taxa de reconhecimento de 98,76%. No entanto, quando os dois parâmetros foram combinados, a taxa de reconhecimento aumentou para 99,43%, mostrando que a nova característica acrescenta informações importantes para o sistema de reconhecimento de locutor.

Palavras-chave: *ASR, VVGM, MFCCs, Multifractal, GMM, Cascata Multiplicativa*.

Abstract

This dissertation presents an Automatic Speaker Recognition (ASR) system, which employs a new parameter based on the "VVGM" (Variable Variance Gaussian Multiplier) multifractal model. The methodology adopted for the development of this system is formulated in two stages. Initially, a traditional ASR system was implemented, based on the use of Mel-Frequency Cepstral Coefficients (MFCCs) and the Gaussian mixture models (GMMs) as the classifier, since it is the method with the best results in the literature. This procedure allows having a broad knowledge about the production of speech signals and a reference system to compare the performance of the new VVGM parameter. The second stage was dedicated to the study of the multifractal processes for speech signals, given that with them, it is possible to analyze information contained in non-stationary parts of the evaluated signal. Taking advantage of this characteristic, speech signals are modeled using the VVGM model, which is based on the binomial multiplicative cascade process, and uses the variances of multipliers for each state as a new speech feature.

The information obtained by the two methods is different and complementary. Therefore, it is interesting to combine the classic parameters with the multifractal parameters in order to improve the performance of speaker recognition systems.

The proposed systems were evaluated using three databases with different settings, such as sampling rates, number of speakers and phrases, duration of training and testing. These different configurations allow the determination of characteristics of the speech signal required by the system. With the experiments, the speaker identification system based on the VVGM parameters achieved significant success rates, which shows that this multifractal model contains relevant information of the identity of each speaker. For example, the second database is composed of speech signals of 71 speakers (50 men and 21 women) digitized at 22.05 kHz with 16 bits/sample. The training was done with 20 phrases for each speaker, with an approximately total duration of 70 s. Evaluating the ASR system based on VVGM, with this database and using

test locutions with 3s of duration, it was obtained a recognition rate of 91.3%. Using these same conditions, the ASR system based on MFCCs reached a recognition rate of 98.76%. However, when the two parameters are combined, the recognition rate increased to 99.43%, showing that the new feature adds substantial information to the speaker recognition system.

Keywords: *ASR, VVGM, MFCCs, Multifractal, GMM, Multiplicative Cascade*.

Aos meus pais Cesar e Diana

Aos meus irmãos Cesar, Ricardo e Camilo

Luka e Sofia.

Por ser o motor de minha vida

Agradecimentos

Ao Prof. Dr Lee Luan Ling pelo apoio, orientação e motivação incondicional durante todo o trabalho.

Ao Prof. Dr Fábio Violaro pela acolhida e apoio, pela orientação e dedicação no trabalho, e pelas inúmeras discussões e idéias.

A minha família pelo carinho e motivação constante, levando-me a ser melhor cada dia.

A Gustavo, Alice, Mitchell, Cesar, Duber, Miguel, Alejandro, Jefferson M, Luisa, Felipe, Carlos, Fabio, Paul, Liz, Daniel, Andrés, Alexandre, Andrei, Juliana, pela amizade e paciência.

A meus colegas e amigos do laboratório LRPRC e da FEEC: Jeferson S, Ana, Daniel, José, Julio, Kobi, Bernardo, Natasha, Victor, Carlos.

Aos membros da Banca pelas valiosas sugestões.

A Marcela, Sandra, German, Henry, Carlos, Enrique e a "Escuela Colombiana de Ingeniería Julio Garavito" pelo apoio.

A CNPQ, pela concessão da bolsa.

A Deus, por cada minuto dado que juntos, possibilitam a realização de meus sonhos.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ASI -	Automatic Speaker Identification
ASR -	Automatic Speaker Recognition
ASV-	Automatic Speaker Verification
DCT -	Discrete Cosine Transform
DFT -	Discrete Fourier Transform
DTW -	Dynamic Time Warping
ELSDSR-	English language speech database for speaker recognition
EM -	Expectation Maximization
fBm -	fractional Brownian motion
FIR	Finite Impulse Response
GMM -	Gaussian Mixture Model
HMM -	Hidden Markov Model
i.i.d	independent and identically distributed
INRIA -	Institut National de Recherche en Informatique et en Automatique
IRCCyN -	L'Institut de Recherche en Communications et Cybernétique
LAN -	Local Area Network
LPC -	Linear Prediction Coding
ML -	Maximum Likelihood
MFCCs -	Mel-Frequency Cepstrum Coefficients
PCM -	Pulse-Code Modulation
VVGM -	Variable Variance Gaussian Multiplier
WAV -	Waveform Audio File Format
WTMM -	Wavelet Transform Modulus Maxima

1 Introdução

Nos últimos anos, houve um aumento considerável do número e variedade de produtos e serviços que incorporam a interação dos usuários por meio de tecnologia da fala, por ser a forma mais natural e flexível de comunicação humana. Algumas das aplicações de maior crescimento nessa área são sistemas biométricos que envolvem reconhecimento de locutor. Estas aplicações aproveitam o fato de que cada pessoa possui um mecanismo de produção de fala único, associado as suas características fisiológicas e aos seus hábitos linguísticos, tornando os sistemas eficientes e de baixo custo.

O acelerado desenvolvimento das telecomunicações (internet, redes celulares, entre outros) e dos dispositivos eletrônicos (computadores, reprodutores de som, entre outros) tem feito com que as aplicações baseadas em reconhecimento de locutor sejam mais atrativas para o desenvolvimento de sistemas de segurança e controle tais como:

- Autenticação de transações comerciais como método de prevenção de fraudes, através de telefone, internet.
- Operações bancárias em geral, tanto pessoalmente quanto através de algum método remoto.
- Controle de acesso para dispositivos, redes de trabalho, informação restrita.
- Auxílio a portadores de necessidades especiais.

Como mostrado nas aplicações mencionadas acima, em função desses avanços, são atribuídas responsabilidades cada vez maiores para os sistemas de reconhecimento. Esta exigência requer um estudo constante de novas tecnologias para o desenvolvimento de sistemas

mais robustos e confiáveis que satisfaçam as exigências e expectativas impostas pelo mercado, o que justifica a pesquisa nesta área.

Os sistemas de reconhecimento automático de locutor (ASR) podem ser classificados em duas categorias: identificação e verificação. Nos sistemas de identificação automática de locutor, é feito o reconhecimento de qual pessoa, pertencente a um determinado grupo de indivíduos falou, isto sem fornecer informação de sua identidade. Já nos sistemas de verificação automática de locutor, o usuário fornece sua identidade (senha específica) e o sistema decide aceitar ou recusar o usuário, dependendo da comparação com o seu padrão armazenado. Adicionalmente, os ASR podem operar de dois modos: **dependente de texto**, em que o sistema utiliza a mesma palavra ou frase tanto no treinamento quanto no teste, e **independente de texto**, em que tanto as locuções de treinamento quanto as de teste são diferentes, permitindo ao usuário falar livremente. Um sistema ASR é formado basicamente por três módulos: extração de parâmetros características do locutor; treinamento, no qual é feito o modelamento de cada locutor; reconhecimento, onde as características de um dado locutor são comparadas com os modelos armazenados previamente para efetuar a identificação/verificação.

Os sistemas tradicionais de ASR fundamentam a extração de características do sinal de fala no uso da análise espectral de curto tempo efetuada com a DFT (Campbell, J. 1997) (Reynolds, D. & Rose, R. C. 1992), focando a obtenção de aspectos estáveis e consistentes do sinal (Langi, A. & Kinsner, W. 1995). No entanto, na atualidade, novas tendências estão surgindo para o processamento de sinais caracterizados por sua natureza não estacionária, tais como os sinais de fala. Um exemplo disto é o uso da teoria fractal como método alternativo para o processamento de sinais não estacionários.

Nos sinais de fala, grande quantidade da informação se encontra concentrada nas partes não estacionárias do sinal, como é caso das transições (de vogais a consoantes, de vogal a vogal, entre outras), o que torna os métodos tradicionais pouco adequados para caracterizar estes comportamentos. Logo, intervalos de curta duração de sinais de fala podem ser considerados quase estacionários (Langit, A. Z. R., Soemintapurat, K. & Kinsners, W. 1997). A teoria multifractal é capaz de caracterizar estes tipos de mudanças rápidas chamadas de singularidades e modelar esse comportamento por meio de multi-escalas.

Existem poucas pesquisas que aplicam técnicas multifractais na área de processamento de fala até este momento. Por exemplo, os autores de (Sant'Ana, R., Coelho, R. & Alcaim, A. 2006) propõem um sistema de reconhecimento automático de fala independente de texto, o qual emprega como características estatísticas um vetor de parâmetros Hurst, obtido através da aplicação do estimador multidimensional *wavelet-based* proposto por (Veith, D. & Abry, P. 1998), e, como classificador para as tarefas de identificação e de verificação de locutor, um modelo multidimensional fBm (*fractional Brownian motion*). Em (Zhou, Y., Wang, J. & Zhang, X. 2010) é proposto um novo método de extração de características não-linear com base no método WTMM (W*avelet Transform Modulus Maxima)*, a fim de facilitar a extração de características do espectro multifractal (MSF) de sinais de fala. O principal objetivo é melhorar o desempenho de um sistema de reconhecimento de locutor utilizando as informações extraídas a partir do espectro multifractal correspondente. Em (Langi, A. & Kinsner, W. 1995), um algoritmo foi implementado a partir da "trajetória da variância da dimensão fractal" para detectar os limites externos de um enunciado e suas pausas internas, que representam ausência de fala.

1.1 Objetivos

Este trabalho tem como objetivo o desenvolvimento de um sistema de identificação automática de locutor, operando em modo independente de texto. Para o módulo de extração de parâmetros, é proposto o uso de um novo parâmetro multifractal como vetor de características do locutor, denominado VVGM (*Variable Variance Gaussian Multiplier*). Este parâmetro está baseado no modelo de cascata multiplicativa binomial, e está focado na análise da distribuição dos multiplicadores $f_{R_j}(r)$.

Adicionalmente, foi realizado em paralelo um ASR tradicional, baseado no uso dos MFCCs (*Mel-Frequency Cepstrum Coefficients*), por ser um método popular, encontrado na literatura para o processamento de fala e áudio (Kinnunen, T. & Li, Haizhou. 2010). Esta implementação foi feita com duas finalidades. A primeira é ter um sistema de referência para avaliar o desempenho do sistema ASI empregando o novo parâmetro VVGM. A segunda é integrá-lo com os parâmetros multifractais VVGM, dado que esses realizam uma abordagem

diferente do sinal, enfatizando as partes não-estacionárias e gerando informações complementares. Com isso, se consegue uma melhora de desempenho e robustez do sistema. No módulo de treinamento, os parâmetros extraídos foram modelados por uma mistura de gaussianas (GMM).

Foram desenvolvidos dois métodos para a integração dos dois parâmetros em um sistema. No primeiro método, empregam-se subsistemas de modelagem e identificação separados para cada um dos parâmetros. A decisão final de identificação é tomada ao ponderar as probabilidades a posteriori de saída de cada subsistema. No segundo método, os parâmetros são integrados em um só vetor e modelados por uma única mistura de gaussianas.

Outro objetivo deste trabalho consiste em estudar algumas características multifractais presentes em sinais de fala através de curvas multifractais como o espectro multifractal $f(\alpha)$ ou funções de escalonamento. Estas curvas (curvas de singularidade) fornecem aspectos importantes para processamento, tais como: decomposição, representação e caracterização do espectro, de forma análoga a análise de Fourier em abordagens tradicionais (Langit, A. Z. R., Soemintapurat, K. & Kinsners, W. 1997). Com isso, pretende-se abrir as portas para o uso de ferramentas multifractais em processamento de fala, como alternativa ou complemento aos métodos tradicionais.

1.2 Conteúdo da Dissertação

Este trabalho está organizado da seguinte maneira:

- No Capítulo 2, é apresentado o conceito fractal, seguido do formalismo da teoria dos processos multifractais, incluindo definições, métodos para a estimação das características multifractais e análise destas características.
- No Capítulo 3, é estudado o modelo multifractal *cascata multiplicativa*. Inicia-se com a construção de um tipo de cascata particular, chamada cascata binomial, seguido da generalização matemática. Finalmente é feita a descrição do modelo *VVGM* (*Variable Variance Gaussian Model*), o qual é proposto neste trabalho para a obtenção de parâmetros característicos de sinais de fala.

- O Capitulo 4 apresenta uma visão geral do estado da arte atual dos sistemas de reconhecimento automático de locutor (ASR), e introduz os conceitos básicos do funcionamento destes sistemas.
- No Capítulo 5, são apresentados os sistemas automáticos de identificação de locutor (ASI) independente de texto, desenvolvidos neste trabalho. São descritos os procedimentos empregados para extração de parâmetros característicos (MFCCs e VVGM), assim como o método de modelagem de cada locutor usado pelos sistemas (GMM).
- O Capítulo 6 propõe o estudo de características multifractais presentes em sinais de fala através das curvas multifractais tais como o espectro multifractal $f(\alpha)$ ou funções de escalonamento.
- No Capítulo 7, são apresentados os testes e resultados obtidos.
- Finalmente, o Capítulo 8 contém conclusões obtidas a partir das análises feitas sobre os resultados alcançados. Também são feitas sugestões para trabalhos futuros.

2 Processos Multifractais

2.1 Fractais

A noção de fractal foi divulgada pelo cientista Benoit Mandelbrot em 1975, e difundida em seu livro *Fractals: Form, Chance, and Dimension* em 1977. Este nome vem do adjetivo latino *fractus*, que significa "quebrado" ou "irregular", referindo-se a formas muito irregulares para serem descritas pela geometria tradicional. A geometria fractal é uma extensão da geometria clássica, introduzindo estruturas que não se encaixam nos padrões de Euclides e Newton (Mandelbrot 1982).

A geometria fractal é baseada em dois conceitos fundamentais: invariância na escala e dimensão fractal. Para visualizar melhor estes conceitos, será exposto um exemplo matemático formal concebido em 1904 pelo matemático sueco Helge Von Koch. Na Figura 2.1, são ilustrados quatro estágios do processo de construção da curva de Koch, também conhecida como floco de neve. Para iniciar o processo, no estágio '0' é considerado um segmento de reta unitário. No primeiro estágio este segmento unitário é dividido em três seções, e a seção do meio é trocada por um triângulo equilátero sem base. O comprimento da nova linha é de quatro seções, mas a distância entre os pontos finais é de três seções. Para o segundo estágio, cada uma das quatro seções é substituída por uma cópia do primeiro estágio, reduzida por um fator de 3. Para o desenvolvimento de mais estágios é aplicado este mesmo procedimento, tendo em consideração que, para cada nova fase, o comprimento da linha é aumentado por um fator 4/3.

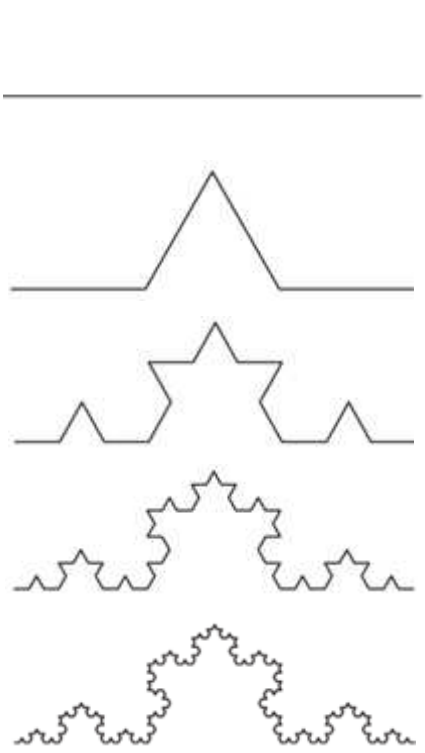


Figura 2.1 – Quatros estágios no desenvolvimento da curva de Koch.

Através da curva de Koch é possível reconhecer as propriedades básicas dos fractais. A regularidade combinatorial presente no floco de neve é essencialmente uma lei de escala. Assim, se uma pequena parte da curva é apropriadamente ampliada, a nova vista parece exatamente com alguma seção anterior. Esta propriedade é chamada *invariância na escala*, e basicamente indica que certas características de um sistema são independentes da escala de análise (García, A. P. M., Jiménez, F. J. & Ayuso, J. L 2007).

Dada a complexidade da estrutura e organização dos conjuntos fractais, não é possível estabelecer a posição dos pontos que o constituem no espaço. Em contrapartida, é definida alguma relação entre as estruturas observadas desde diferentes níveis de resolução do mesmo conjunto (Barnsley, M.F. 1993). Para esta relação é empregado o conceito *de dimensão fractal*. A dimensão fractal reflete quantitativamente a propriedade de escala fractal, ou seja, como muda sua estrutura quando se varia de estágio. No caso do floco de neve, cada seção é substituída por $N=4$ seções de comprimento $r=\frac{1}{3}$ da seção anterior (Figura 2.1). Como resultado tem-se

exatamente a mesma forma com diferente escala. O valor da dimensão é determinado relacionando esses dois números: $-\frac{\log N}{\log r} = \frac{\log 4}{\log 3} = 1.2618 \ldots$

Estes conceitos são estendidos para processos e sistemas invariantes em escala. Portanto, um processo fractal pode ser definido como aquele em que o mesmo processo elementar ocorre em diferentes escalas (Feder, J. 1988). De forma geral, os fractais podem ser divididos em duas classes: fractais determinísticos e fractais aleatórios (García, A. P. M., Jiménez, F. J. & Ayuso, J. L 2007). Uma comparação gráfica desta classificação pode ser observada na Figura 2.2, empregando um fractal tradicional, *o conjunto de Cantor*, proposto por Georg Cantor em 1883. Basicamente este fractal é gerado dividindo-se um segmento inicial em três partes e eliminando-se uns dos segmentos, sendo este processo efetuado indefinidamente sobre cada novo segmento gerado.

Fractais determinísticos são gerados através de um processo iterativo ou recorrente, regido por regras exatas de construção. Assim, um fractal pode mudar substancialmente de um estágio para outro, mas o princípio geral permanece igual, sendo equivalentes as estruturas auto-similares. Este tipo de fractal pode ser classificado em uniescalar e multiescalar.

Um fractal uniescalar pode ser gerado dividindo um objeto, definido em R^n, em N réplicas idênticas, reduzidas por um fator $r < 1$. Cada peça gerada é dividida de novo em N partes, conservando as regras de construção. Depois de infinitas iterações é obtido o fractal. Este caso é ilustrado em (2.2 a.), empregando-se $r = \frac{1}{3}$ e eliminando-se o segmento do meio.

Para um fractal multiescalar, o procedimento de construção é igual ao uniescalar, mas as peças geradas não são idênticas, pois cada divisão possui uma regra de construção própria. Assim, cada réplica será vista como uma redução do objeto original por diversos fatores $r_j < 1$, com $j = 1, \ldots, N$. Para o exemplo do conjunto de Cantor (2.2 b.), a divisão inicial é feita com os fatores de redução $r_1 = 0{,}5$, $r_2 = 0{,}25$ e $r_3 = 0{,}25$ e eliminando-se o segmento do meio.

Fractais Aleatórios são gerados através de um processo iterativo ou recorrente, que envolve aleatoriedade em cada etapa de construção. Estas condições aleatórias podem ser definidas através de diferentes técnicas estocásticas de modelagem. Embora a estrutura do fractal possa mudar para diferentes estágios, as propriedades estatísticas são as mesmas em

todas as escalas. Este tipo de fractal é associado a estruturas auto-similares estatisticamente. Para a geração de um fractal são efetuadas infinitas iterações, ou seja, infinitos passos aleatórios e, portanto é necessário usar a teoria de probabilidade (Falconer, J. K. 2003). Fractais aleatórios são amplamente usados para descrever fenômenos naturais, tais como nuvens, paisagem, ruído de fundo, entre outros (Mandelbrot 1982). Na Figura 2.2.c, é mostrada a geração de um fractal aleatório, em que cada segmento é dividido em três partes iguais, uma das quais é selecionada aleatoriamente para ser eliminada.

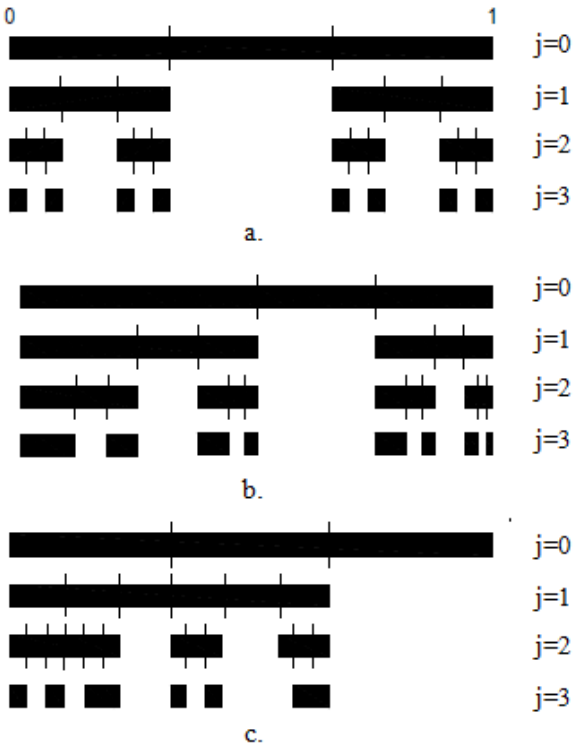


Figura 2.2 – Exemplificação dos tipos de fractais (adaptado de (de Lima, 1998)).

2.2 Processos Multifractais

Muitas séries temporais biológicas são extremamente heterogêneas, apresentando não-estacionariedade e oscilações de maneira irregular e complexa. Este comportamento pode ser descrito, pelo menos de uma forma global, por suas dimensões (Riedi, R. H., Crouse, M. S., Ribeiro, V. J. & Baraniuk, R. G. 1999). Considerando que as propriedades em escala deste tipo de sinal são causadas por uma dinâmica caótica e por processos aleatórios, existem vários comportamentos em escala diferentes (*multiple scaling*). Quando esses diversos comportamentos em escala são encontrados em diferentes instantes de tempo, tem-se o processo denominado multifractal. As propriedades de escalonamento, para processos monofractais, são caracterizadas por um único parâmetro (expoente de Hurst) durante todo o tempo do processo. No entanto, o grau de auto-similaridade deste tipo de sinais heterogêneos é variante com o tempo, sendo preciso apelar para a teoria multifractal para seu estudo.

Um processo multifractal é caracterizado por um conjunto de dimensões fractais, das quais é possível obter informações mais detalhadas, através de ferramentas como a análise multifractal. Esta análise é capaz de descrever o comportamento local de medidas, distribuições e funções de forma geométrica e estatística. Um dos critérios da análise multifractal é estimar os momentos estatísticos dos processos para avaliar suas regularidades locais (Riedi et al., 1999). Através desta análise, algumas propriedades encontradas em processos multifractais podem ser verificadas. A seguir, será apresentado o formalismo multifractal e seus métodos de análise.

2.3 Formalismo Multifractal

Sinais multifractais são geralmente caracterizados por terem um comportamento bastante irregular. Podem apresentar transições abruptas de comportamento entre um instante de tempo e o seguinte. O local destas mudanças rápidas é comumente conhecido como ponto singular.

Para caracterizar os pontos singulares presentes num sinal $f(t)$, é preciso quantificar sua regularidade. Esta medida pode ser encontrada através do expoente Lipschitz, o qual provê medidas uniformes de regularidade, tanto em intervalos de tempo quanto em pontos isolados (Stênico, J. W. e Lee, L. L. 2009). Nos processos multifractais, este expoente, também conhecido pelo nome de expoente de Hölder α_t, pode assumir uma série de valores, dependendo de t. Por conseguinte, os momentos de escala variam de maneira não-linear, gerando duas possíveis definições para os "multifractais".

A primeira definição de multifractal é vista como uma generalização do processo monofractal. Assim, diz-se que um processo $X(t)$ é monofractal se obedece à relação de escala descrita na Equação (2.1), gerando outros processos fractais, com a mesma distribuição estatística entre eles.

$$X(ct) \overset{\mathrm{d}}{=} c^H X(t) \tag{2.1}$$

Na Equação (2.1), c^H representa o fator de escalonamento, com $c > 0$ e $0 < H < 1$. Para sistemas estocásticos $\overset{\mathrm{d}}{=}$ indica igualdade em distribuição estatística entre processos. Deste modo, os sinais monofractais são considerados homogêneos no sentido que possuem as mesmas propriedades de escala, sendo caracterizados localmente por um único expoente de singularidade, o "expoente de Hurst", durante todo o tempo (Stanley, H.E. 1995) (Bund, A. & Havlin, S. 2000). Para sinais monofractais $H \equiv h_0$, o que sugere estacionariedade sob o ponto de vista de suas propriedades locais de escala (Ivanov, P. Ch. 2003).

Por outro lado, os sinais multifractais podem ser decompostos em diversos subconjuntos, que se caracterizam por diferentes expoentes de Hurst locais, assumindo o nome de expoente de Hölder α. Este expoente quantifica o comportamento local da singularidade e, portanto, refere-se ao escalonamento local da série temporal. Na Figura 2.3, pode ser observado um exemplo que apresenta a variedade de expoentes requeridos para caracterizar as propriedades de escala num sinal multifractal (Vicsek, T. 1993), mostrando que este é intrinsecamente mais complexo e heterogêneo do que o monofractal (Ivanov, P. Ch. 2003). Assim, no quadro superior são

ilustrados os expoentes locais de Hölder de um sinal multifractal no tempo, seguido da decomposição deste sinal (quadros subsequentes), com cada expoente local de Hölder indicado por uma cor diferente e cada dimensão fractal representada pela densidade de barras verticais.

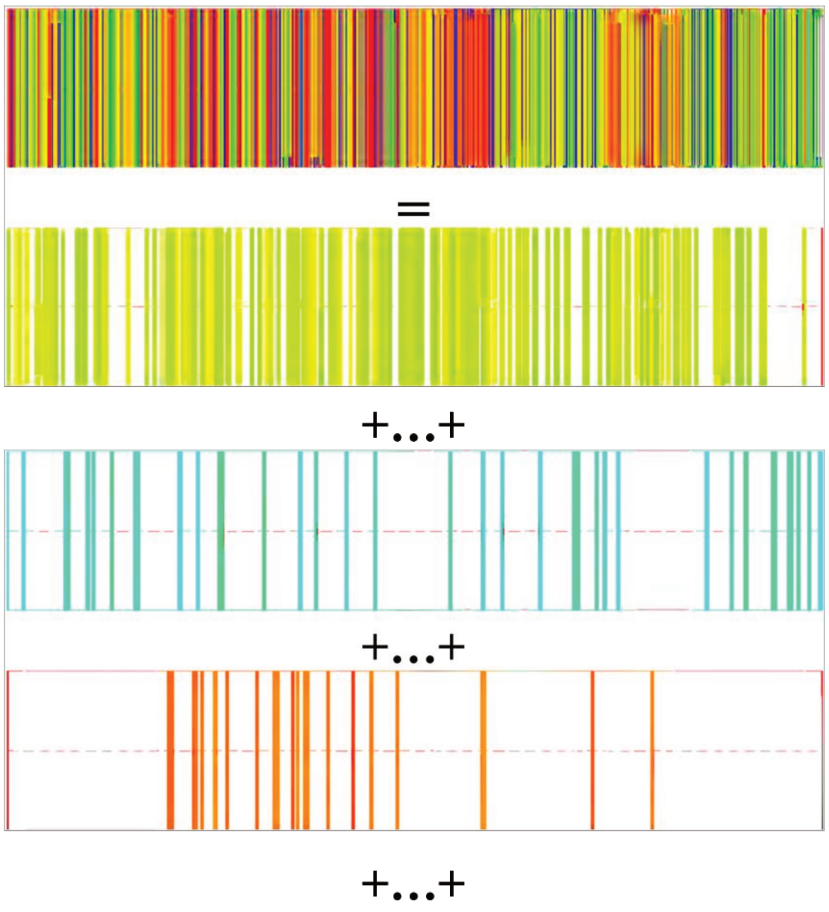


Figura 2.3 – Decomposição de expoentes locais de Hölder de um sinal multifractal. O eixo x representa o tempo e as barras verticais (eixo-y) indicam expoentes locais de Hölder (Ivanov).

Tendo por base a Equação (2.1), surge uma generalização para processos multifractais. Considerando o fator de escalonamento c^Hcomo uma nova variável aleatória $M(c)$ que não depende de t, são analisadas relações multi-escala de forma geral,

$$X(ct) \overset{d}{=} M(c)X(t) \tag{2.2}$$

onde $X(t)$ *e* $M(c)$ são dois processos estocasticamente independentes. Para os processos monofractais, $M(c) = c^H$, com H constante. Por isso, um processo monofractal possui apenas um fator de escala representado por um ponto em seu espectro de singularidades.

Analogamente, nos processos multifractais, $H(c) = log_c M(c)$, e a equação acima pode ser reescrita como:

$$X(ct) \overset{d}{=} c^{H(c)} X(t) \tag{2.3}$$

onde $H(c)$ deixa de ser uma constante (monofractais), para tornar-se uma variável aleatória dependente de c, o que permite uma melhor descrição de fenômenos irregulares. Dado o comportamento em escala descrito pela Equação (2.3) impõe algumas restrições a distribuição do processo $X(t)$, o que leva a uma primeira definição de multifractal.

Definição 2.2.1 *Um processo estocástico* $X(t)$ *é dito multifractal se satisfaz:*

$$E(|X(t)|^q) = c(q)t^{\tau(q)+1}, \forall t \in T, q \in Q \tag{2.4}$$

onde E é a função esperança, T e Q são números reais, e $\tau(q)$ e $c(q)$ são funções com domínio Q. A função $\tau(q)$ é chamada expoente de escalonamento do processo multifractal ou função de partição. Os sinais monofractais exibem um espectro $\tau(q)$ linear, considerando $\tau(q) = qH - 1$, onde H é o expoente de Hurst. Por outro lado, nos sinais multifractais, $\tau(q)$ é uma função não-linear, devido ao fato que $\tau(q) = q\alpha(q) - 1$, onde $\alpha(q) \equiv d\tau(q)/dq$ não é constante (Ivanov, P. Ch. 2003).

A segunda definição do processo multifractal é baseada no expoente de Hölder, analisando as características multi-escala locais de um processo em qualquer instante t. Assim, o comportamento errático de um processo contínuo $X(t)$, num dado tempo t, pode ser caracterizado, numa primeira aproximação, em comparação com uma função algébrica satisfazendo essa relação.

Definição 2.2.2. *Uma função ou percurso do processo $X \in C_n^{\alpha}$, se existir um polinômio P_n de grau n tal que:*

$$|X(t) - P_n(t)| \leq C|t - t_0|^{h(t_0)} \tag{2.5}$$

para valores de t suficientemente perto de t_0. A partir dessa expressão, o grau de regularidade Hölder local de X em t_0 é definido por:

$$H(t) := \sup \left\{h : X \ \in C_{t_0}^{\alpha}\right\} \tag{2.6}$$

O polinômio $P_n(t)$, nos casos mais simples, corresponde ao desenvolvimento da série de Taylor de X em t. Conhecendo o grau n deste polinômio, sabe-se que $X(t)$ é n vezes diferenciável em t_0. O expoente de Hölder determina o comportamento da função $X(t)$ na vizinhança do ponto t_0. Assim $h(t_0)$ mede o nível da singularidade neste ponto. Quanto maior for o valor do expoente de Hölder, maior será o nível de regularidade da função nesse ponto.

2.4 Espectro Multifractal

Nesta seção, é apresentada a medida μ. Esta medida é uma forma de especificar o método de distribuição, propagação ou crescimento de um objeto ou um processo fractal sobre um conjunto de regras de apoio. Esta distribuição pode ser efetuada com abordagem euclidiana,

como intervalos de retas ou quadrados, ou restrita para abordagem fractal, como o conjunto de Cantor. Tomando a curva de Koch na seção 2.1 como exemplo, é considerada como regra de apoio a divisão de cada segmento em três e a substituição deste por quatro segmentos novos de igual comprimento. Logo, ao considerar um segmento unitário para o estágio 0, a medida assume o valor de μ=4/3. Assim, a medida μ é definida formalmente como uma medida regular finita de Borel em $\mathbb{R}^n$, de modo que $0 < \mu(\mathbb{R}) < \infty$.

Para pocessos multifractais, a medida μ varia em cada intervalo de estudo. Deste modo, a análise multifractal tem como objetivo quantificar a estrutura singular das medidas e fornecer um modelo para os fenômenos em que ocorre escalonamento com uma variedade de leis de potência diferente (Falconer, J. K. 2003).

Para uma medida finita de μ em $\mathbb{R}^n$, a definição da dimensão local (expoente local de Hölder) de μ em x é dada por:

$$dim_{loc}\mu(x) = \lim_{r\to 0} \log \mu(B(x,r)) \,/\log r \tag{2.7}$$

se o limite existir. O conjunto E_α é constituído por todos os pontos x nos quais a $dim_{loc}\mu(x)$ existe, e seu valor é igual a α, gerando-se um E_α para cada $\alpha \geq 0$ (Falconer, J. K. 2003), como é definido na Equação (2.8). Para algumas medidas de μ, o conjunto E_α pode não ser vazio e ser fractal para uma gama de valores α. Nesse caso, μ assume o nome de *medida multifractal*.

$$E_\alpha = \{x \in \mathbb{R}^n : dim_{loc}\mu(x) = \alpha\} \tag{2.8}$$

$$= \left\{x \in \mathbb{R}^n : \lim_{r\to 0} \log \mu(B(x,r)) \,/\log r = \alpha\right\} \tag{2.9}$$

As medidas multifractais são caracterizadas através do espectro multifractal ou espectro de singularidades, definido como $f(\alpha) \equiv \dim E_\alpha$.

Existem dois enfoques para a análise multifractal: a *teoria fina* (do inglês, *fine theory*), na qual se estuda o comportamento local de $\mu(B(x,r))$ quando $r \to 0$, e a *teoria grosseira* (do inglês, *coarse theory*), em que se quantificam as irregularidades globais de $\mu(B(x,r))$ para r pequeno e, em seguida, é avaliado o limite quando $r \to 0$. Assim, a teoria fina talvez seja mais adequada para a análise matemática, exigindo idéias próximas às utilizadas no estudo da dimensão de Hausdorff de conjuntos. Por outro lado, a teoria grosseira é mais conveniente quando se trata de encontrar espectros multifractais dos exemplos da física ou estimar espectros a partir de experimentos de computador. Esta abordagem lembra o cálculo da dimensão através do método da contagem de caixa (do inglês, *Box-counting*) (Falconer, J. K. 2003). A seguir, será discutida a definição de $f(\alpha)$ a partir das duas perspectivas.

O objetivo básico da abordagem fina para análise multifractal é encontrar $\dim E_\alpha$ para $\alpha \geq 0$,

$$f_H(\alpha) \coloneqq \dim(E_\alpha) \tag{2.10}$$

onde dim (E_α) é a dimensão Hausdorff do conjunto E_α.

Para a estimação do espectro de singularidades através da teoria grosseira, são consideradas as irregularidades da distribuição da medida para $r > 0$ e $r \to 0$. Normalmente, nos processos multifractais, existe um valor de expoente de Hölder α_0 mais frequente, mas outros valores também ocorrem. Esses expoentes de Hölder, com valores diferentes de α_0, são bastante importantes, uma vez que a maior parte das variações em uma função multifractal encontra-se em instantes com tais expoentes. Tal característica permite discriminar multifractais de monofractais, dando origem a definição 2.2.3.

Definição 2.2.3 *Seja $N_r(\alpha)$ o número de expoentes de Hölder aproximadamente iguais a α que ocorrem ao se subdividir um processo em r partes de mesmo tamanho. Então, o espectro multifractal, representado por $f(\alpha)$, é definido por:*

$$f(\alpha) \equiv \lim\left\{\frac{log\, N_r(\alpha)}{log\, r}\right\} para\ r \to 0 \tag{2.11}$$

Para processos multifractais, o espectro apresenta uma forma parabólica côncava, onde $f(\alpha) \leq \alpha(t)$, para todo $\alpha(t)$ e $f(\alpha) \leq f(\alpha_0)$, onde $f(\alpha_0)$ é o valor máximo de $f(\alpha)$ (Riedi, R. H. 2002).

2.5 Estimação de Características Multifractais

Dadas as definições anteriores dos processos multifractais, existem duas abordagens diferentes para se estudar o comportamento multifractal de uma série temporal, baseadas no "espectro multifractal". A primeira se fundamenta na estimação da função de partição do processo usando o método dos momentos, e a segunda na análise de regularidade do processo através de seu "espectro multifractal".

2.5.1 MÉTODO DOS MOMENTOS

Este método está baseado no formalismo de processos multifractais, acompanhando a primeira definição apresentada na Equação (2.4). O método dos momentos tem como hipótese a presença de uma cascata multiplicativa. Fundamenta-se na estimação do espectro multifractal através do estudo das propriedades de singularidade desta cascata, a fim de se ter uma idéia da distribuição dos expoentes de Hölder (Krishna, M. P., Gadre, V. M., & Dessay, U. B. 2003). Esta estimação emprega o conceito de função de partição para destacar as singularidades da distribuição medida na cascata. O processo consiste na reconstrução de qualquer estágio anterior da cascata, partindo de agregações em intervalos de tamanhos 2^{-N}. Assim, uma série temporal $\{\boldsymbol{X_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{N}}$ é considerada como uma amostra de um nível da cascata, com uma medida no intervalo [0,1] e escala $1/2^N$.

Define-se a soma de partição como (Krishna, M. P., Gadre, V. M., & Dessay, U. B. 2003):

$$\mathcal{X}_m^X(q) \coloneqq \sum_{k=1}^{N/m} \left(\overline{\boldsymbol{X}_k^{(m)}}\right)^q \tag{2.12}$$

onde

$$\overline{\boldsymbol{X}_k^{(m)}} \coloneqq \sum_{i=1}^{m} \boldsymbol{X}_{(k-1)m+i}^m \tag{2.13}$$

com um valor fixo de m. A Equação (2.12) apresenta a maneira como a função de partição exibe a natureza de escala dependendo do valor de m.

$$\mathcal{X}_m^X(q) \sim m^{\tau(q)} \tag{2.14}$$

Partindo da expressão acima, pode-se estimar a relação de escala $\tau(q)$ aplicando a função logaritmo, obtendo-se:

$$\log \mathcal{X}_{\mathrm{m}}^{\mathrm{X}}(q) = \tau(q) \log m + \log c(q) \tag{2.15}$$

onde $\log c(q)$ é constante. Quando $\log \mathcal{X}_{\mathrm{m}}^{\mathrm{X}}$ exibe linearidade em relação a $\log m$, para um valor fixo de q_i, tem-se que a série temporal apresenta natureza fractal. Na Figura 2.4, ilustra-se melhor a interpretação dos parâmetros $\tau(q)$ e $c(q)$ da Equação (2.15). Assim, os parâmetros $\tau(q_i)$ e $c(q_i)$ podem ser determinados pela regressão do logaritmo da função de partição.

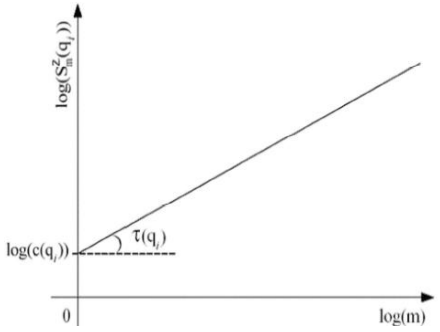


Figura 2.4 – Estimação das funções τ(q) e c(q).

2.5.2 ESPECTRO MULTIFRACTAL

O espectro multifractal $f(\alpha)$ é uma representação da distribuição do expoente de Hölder num processo. Esta função espectral pode ser encontrada usando algumas técnicas tais como: espectro de granularidade grosseira (*coarse graining spectrum*), espectro de Hausdorff e espectro de Legendre. Na seção 2.2.2, é descrito o procedimento para a obtenção tanto do espectro de granularidade grosseira (*coarse graining spectrum*) ou espectro de grandes desvios (*large deviation spectrum*), quanto do espectro de Hausdorff. Provavelmente, a dimensão de Hausdorff é a mais importante. O espectro de Hausdorff provê uma informação geométrica pertinente à dimensão fractal dos conjuntos de pontos em um sinal que possui um dado expoente de Hölder. Do ponto de vista matemático, este é o espectro multifractal mais preciso, sendo também o mais difícil de ser estimado (Falconer, J. K. 2003).

O espectro de granularidade grosseira provê informações estatísticas relacionadas à probabilidade de encontrar no sinal um ponto com um dado expoente de Hölder. Além de permitir a medição do comportamento desta probabilidade, quando é submetida a mudanças de resolução. Embora este espectro não seja exatamente a densidade correspondente aos α´s, mas sim uma dupla normalização logarítmica desta densidade, a estimação deste espectro exige a aplicação de ferramentas de estimação de densidade de probabilidade. Neste caso, para a estimação da densidade de probabilidade, normalmente são empregadas ferramentas clássicas como o método de kernel duplo (Devroye, L. 1989).

O espectro de Legendre é uma aproximação côncava do espectro de grandes desvios. Este espectro é de grande interesse, pois normalmente permite estimações robustas, embora para alguns sinais específicos (Riedi, R. H. & Véhel, J. L. 1997) omita algumas informações possíveis de serem obtidas através do espectro de grandes desvios. A robustez e a simplicidade de estimação do espectro de Legendre o tornam o mais atrativo para o espectro multifractal. Este trabalho centrou sua atenção no uso de espectro de Legendre, pelas qualidades acima (Stênico, J. W. e Lee, L. L. 2009).

O espectro multifractal pode ser obtido através da transformada de Legendre de $\tau(q)$ (função de escalonamento) (Krishna, M. P., Gadre, V. M., & Dessay, U. B. 2003), com a seguinte relação:

$$f(\alpha) = \min_q\{q\alpha - \tau(q)\} \tag{2.16}$$

Basicamente, o espectro provê informação das singularidades do sinal e quais delas predominam. Em particular, para o caso de sinal com natureza monofractal, $\tau(q)$ varia linearmente ($\tau(q) = nq - 1$), fazendo com que o expoente de Hölder assuma um valor único. Daí resulta que a representação gráfica dos processos monofractais se resume a um ponto ou uma reta.

3 Multifractais Multiplicativos

Neste capítulo, será estudado o modelo multifractal **cascata multiplicativa**. O uso deste modelo surgiu da física, especificamente da modelagem da turbulência por (Kolmogorov 1962), onde é pesquisada a intermitência e invariância de escala. Este modelo é baseado na tendência da turbulência a se concentrar localmente enquanto a escala diminui, dando lugar ao aumento de heterogeneidade. Assim, a energia ingressa em um sistema com turbulência em grande escala, tanto em termos de espaço quanto em quantidade de energia. Esta energia é dissipada de uma maneira não uniforme, devido a presença de diferentes fenômenos de dissipação. Partes do espaço podem apresentar redemoinhos com comportamentos violentos, enquanto outras partes encontram-se relativamente calmas. Estes fenômenos se repetem em escalas cada vez menores, até o ponto em que a energia é dissipada como calor (Harte, D. 2001). Dado isso, a turbulência pode ser estimada em função da energia transferida em escalas menores.

Na atualidade, diversas áreas aplicam as cascatas multiplicativas para modelar fenômenos não-lineares que apresentam estrutura multiplicativa, como é o caso da modelagem de tráfego (Riedi et al., 1999), fenômenos geofísicos (Gupta, V. & Waymire, E. 1993), estudo de finanças (Mandelbrot, B. B. 1997), entre outros.

O capítulo está organizado da seguinte forma: Na seção 3.1, é apresentada a definição formal de cascata multiplicativa. Na seção 3.2, é observada a construção de um tipo de cascata particular, chamada cascata binomial. Finalmente, na seção 3.4, é feita a descrição do modelo VVGM (*Variable Variance Gaussian Model*), o qual é empregado neste trabalho para a obtenção de parâmetros característicos de sinais de fala propostos.

3.1 Definição

A cascata multiplicativa é um processo iterativo, que se inicia assumindo um conjunto de tamanho finito fechado com uma massa definida unitária. Para cada iteração, esse conjunto é dividido em subconjuntos menores de comprimento b^{-j}, onde b é um número inteiro e j representa a iteração corrente. A massa também é distribuída entre os subconjuntos, com uma probabilidade $\{m_i\}$, onde cada m_i está relacionado com o i-ésimo intervalo gerado na atual iteração. Considera-se que, em cada estágio do processo da cascata, a medida da massa total é preservada, satisfazendo a expressão $\sum_i m_i = 1$, onde $i = 1, \dots, b^j$ (Gao, J ., Cao, Y., Hu, J. & Tung, W. 2007).

Usando partições diádicas com $b=2$, é possível apreciar melhor as regras de construção da cascata multiplicativa. Na seção 3.2, será analisado o procedimento de construção da cascata multiplicativa binomial. Os estágios da cascata podem também ser divididos em números de subintervalos maiores, quando $b>2$, processo este chamado de cascata multinomial.

As cascatas, dependendo de sua estrutura e comportamento estatístico, podem gerar diferentes tipos de processos multifractais:

- **Nús e vestidos (do inglês *bare* and *dressed*)**. Os **nús** são obtidos após um número de iterações finito. A cascata é desenvolvida começando com as escalas maiores e determinando as menores. Os **vestidos** são obtidos experimentalmente, partindo de um processo físico com valores médios (temporais ou espaciais) para uma determinada resolução.
- **Fortes e suaves (do *inglês hard* and *soft*)**. Os multifractais **fortes** são caracterizados pela presença de singularidades elevadas, portanto os momentos estatísticos de maior ordem divergem; pelo contrário, os multifractais **suaves** possuem flutuações suficientemente pequenas, que evitam a divergência dos momentos.
- **Microcanônicos e canônicos (do inglês *microcanonical* and *canonical*)**. Uma cascata é considerada **microcanônica** quando o fluxo de energia transferida é conservado exatamente em cada iteração. É conhecida como **canônica** quando a energia é conservada na média (Mandelbrot 1982).

- **Calmos e Selvagens (do inglês *calm and wild*)**. É denominada **calma** quando as singularidades não afetam a conservação da energia no processo microcanônico. Caso afetem, é denominada **selvagem**.

3.2 Cascata Multiplicativa Binomial

Uma Cascata Multiplicativa Binomial é construída através de um processo iterativo em que cada intervalo é dividido em dois novos subintervalos. Para tal, considera-se um intervalo unitário inicial [0,1], com uma medida de massa unitária associada. Para a primeira iteração, o intervalo é dividido em dois subintervalos de igual comprimento, e a massa é distribuída entre os dois novos intervalos, com valores m_1 e m_2 respectivamente, onde é suposto que m_1 satisfaz $0<m_1<1$, $m_1 \neq 1/2$, e $m_2 = 1 - m_1$. Para o desenvolvimento matemático são considerados valores de m_1 superiores aos de m_2, ou seja, $1/2 < m_1 < 1$.

Desta forma geral, a medida da massa é particionada, sobre as duas metades de cada intervalo diádico, com a relativa proporção de m_1 e m_2. Assim, a medida da metade esquerda é determinada multiplicando a massa do intervalo atual por m_1, e a da metade direita por m_2.

Geralmente, o parâmetro m_1 é uma variável aleatória dada por uma distribuição escolhida e é chamado de multiplicador. Cada estado é dividido seguindo as mesmas regras de construção. Este procedimento é apresentado na Figura 3.1 (Krishna, M. P., Gadre, V. M., & Dessay, U. B. 2003).

Como é observado na Figura 3.1, no início do processo, $j = 0$ e tem-se o conjunto I_0 denotado pelo intervalo unitário [0,1] com medida da massa $\mu(I_0) = 1$. O intervalo é dividido em duas partes, gerando os subintervalos $I_{1.1}$ de $\left[0, \frac{1}{2}\right]$ e $I_{1.2}$ de $\left[\frac{1}{2}, 1\right]$, ambos de comprimento 0,5. As massas são atribuídas seguindo $\mu(I_0) * m_1$ para a metade esquerda e $\mu(I_0) * m_2$ para a metade direita. Para a segunda iteração, devido ao uso de intervalos diádicos, são gerados quatro subintervalos denotados $I_{2.1}$ de $\left[0, \frac{1}{4}\right]$, $I_{2.2}$ de $\left[\frac{1}{4}, \frac{1}{2}\right]$, $I_{2.3}$ de $\left[\frac{1}{2}, \frac{3}{4}\right]$ e $I_{2.4}$ de $\left[\frac{3}{4}, 1\right]$.

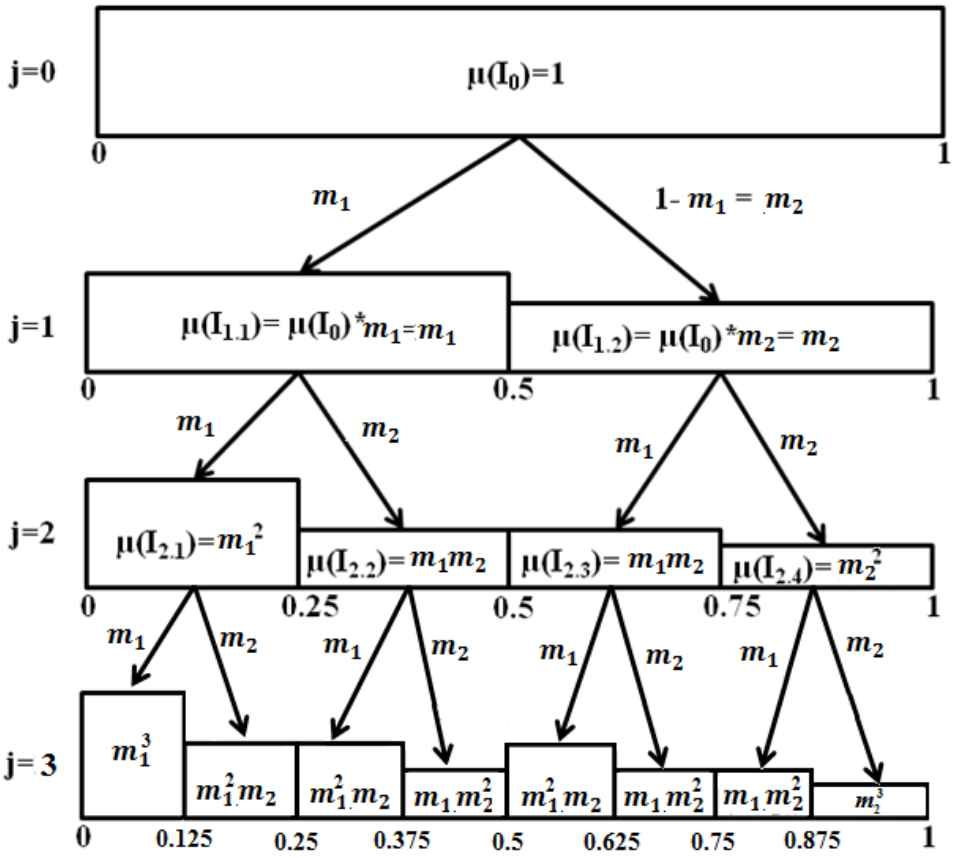


Figura 3.1 – Processo de construção da cascata binominal.

A massa também é distribuída seguindo as mesmas regras de construção (Mandelbrot 1982), obtendo-se

$$\mu(I_{2.1}) = m_1 m_1 \quad \mu(I_{2.2}) = m_1 m_2$$
$$\mu(I_{2.3}) = m_1 m_2 \quad \mu(I_{2.4}) = m_2 m_2 \tag{3.1}$$

De forma geral, para uma cascata na $j^{n-ésima}$ iteração, é gerada a sequência de medidas μ_j, que converge para o processo multifractal μ. Os intervalos diádicos são definidos por $[i2^{-j}, (i+1)2^{-j}]$, onde $i = 0, \dots, 2^j - 1$, com uma escala (comprimento dos intervalos) de 2^{-j}. As massas de cada intervalo são definidas por

$$\mu(I_j) = \prod_{i=1}^{j} m_{\beta_i} = {m_1}^{n_1} {m_2}^{n_2} \tag{3.2}$$

onde n_1 e n_2 (Krishna, M. P., Gadre, V. M., & $Dessay, U. B.$ 2003) são o número de vezes que m_1 e m_2 são multiplicadas e satisfazem a expressão $n_1 + n_2 = j$. Como caso particular, esta cascata apresentada conserva o valor da medida m_1 constante durante toda a construção, portanto, recebe o nome de determinística.

As principais características dos fenômenos tipo cascata são a invariância de escala e a conservação dos fluxos desde escalas maiores até escalas menores (García, A. P. M., Jiménez, F. J. & Ayuso, J. L 2007). Estas características podem ser vistas na Figura 3.1. Assim, os intervalos horizontais com comprimento b^{-j} são uma réplica em tamanho reduzido do conjunto de tamanho unitário original e a redução vertical representa a medida do intervalo $\mu(I_j)$ transferida.

O expoente Hölder para um intervalo diádico I_j de comprimento 2^{-j} é definido por:

$$\alpha(I_j) = \frac{\log \mu(I_j)}{\log (2^{-j})} = \frac{\log\left[m_1^{n_1} m_2^{n_2}\right]}{-j \log 2} \tag{3.3}$$

$$= -\frac{n_1}{j} \log_2 m_1 - \frac{n_2}{j} \log_2 m_2 \tag{3.4}$$

Considerando a substituição na Equação (3.4), das variáveis $\varphi_1 = \frac{n_1}{j}$ e $\varphi_2 = \frac{n_2}{j}$, onde φ_1 e φ_2 denotam a frequência relativa de 0's e 1's no desenvolvimento binário de cada intervalo do estágio j, e das expressões $0 < \alpha_{min} = -\log_2 m_1 \leq \alpha \leq \alpha_{max} = -\log_2 m_2 < \infty$, pode-se reescrever a Equação (3.4) como:

$$\alpha = \alpha(\varphi_1) = \varphi_1 \alpha_{min} + (1 - \varphi_1)\, \alpha_{max} \tag{3.5}$$

onde α é função somente da variável φ_1 (Krishna, M. P., Gadre, V. M., & Dessay, U. B. 2003).

3.3 Derivação do Espectro Multifractal

O processo de cascata pode ser caracterizado através da curva do espectro multifractal $f(\alpha)$. No caso da cascata binomial, define-se $N_j(\alpha)$ como o número de intervalos de comprimento 2^{-j} com expoente de Hölder α. Além disso, como apresentado na Equação (3.5), o expoente α depende da variável φ_1. Portanto, o número de intervalos com expoente α é o mesmo do número de modos de distribuir $n_1 = \varphi_1 j$ zeros entre j posições.

$$N_j(\alpha) = \binom{j}{\varphi_1 j} \tag{3.6}$$

Através do desenvolvimento matemático apresentado por (Krishna, M. P., Gadre, V. M., & Dessay, U. B. 2003), chega-se a expressão

$$N_j(\alpha) \sim \left[2^{-j}\right]^{-g(\varphi_1)} \tag{3.7}$$

onde,

$$g(\varphi_1) = -\log_2[{\varphi_1}^{\varphi_1}(1-\varphi_1)^{1-\varphi_1}] \tag{3.8}$$

Da Equação (3.5), a variável φ_1é isolada:

$$\varphi_1 = \frac{\alpha_{max} - \alpha}{\alpha_{max} - \alpha_{min}} \tag{3.9}$$

Substituindo as expressões definidas anteriormente na Equação (2.11), é definido o espectro multifractal da cascata determinística binomial como:

$$f(\alpha) = -\left(\frac{\alpha_{max} - \alpha}{\alpha_{max} - \alpha_{min}}\right)\log_2\left(\frac{\alpha_{max} - \alpha}{\alpha_{max} - \alpha_{min}}\right)$$

$$-\left(\frac{\alpha-\alpha_{min}}{\alpha_{max}-\alpha_{min}}\right)\log_2\left(\frac{\alpha-\alpha_{min}}{\alpha_{max}-\alpha_{min}}\right) \tag{3.10}$$

Considerando $\alpha_0 = \frac{\alpha_{min}+\alpha_{max}}{2}$ e usando expansão em série de Taylor para o $\log_2 x$, em torno de $x = \alpha_0$, considerando $|\alpha - \alpha_0| \to 0$, pode-se reescrever a Equação (3.10) como:

$$f(\alpha) = 1 - \frac{2}{\ln 2}\left(\frac{\alpha-\alpha_0}{\alpha_{max}-\alpha_{min}}\right)^2 \tag{3.11}$$

Das expressões 3.10 e 3.11, podem-se verificar algumas propriedades do espectro multifractal:

- A função $f(\alpha)$ pode assumir um valor máximo de 1; isso acontece para $\alpha = \alpha_0$.
- A função $f(\alpha)$ apresenta um comportamento quadrático perto de α_0.
- A função $f(\alpha) = \alpha$ para $\alpha = -\ m_1 \log_2\ m_1\ -\ m_2 \log_2\ m_2$.
- A função $f(\alpha)$ apresenta simetria par em torno de α_0.

3.4 Modelo Multifractal VVGM

O modelo multiplicador Gaussiano de variância variável (*Variable Variance Gaussian Multiplier*, VVGM) foi proposto por (Krishna, M. P., Gadre, V. M., & Dessay, U. B. 2003) para a modelagem de intervalos de tempo de chegada de tráfego LAN em banda larga. Este modelo é baseado no modelo de cascata multiplicativa binomial, descrito na seção 3.2, quando foi apresentado o processo de construção da cascada empregando os multiplicadores m_1 e m_2 fixos. O modelo VVGM assume que os multiplicadores são variáveis aleatórias independentes em [0,1], com densidade de probabilidade $f_{R_r}(r)$ (para simplificação, na explicação será substituída a variável *m* por *r*). Este modelo permite obter uma estrutura mais geral que a determinística obtida pelos multiplicadores fixos.

Para efetuar a modelagem de um sinal usando VVGM é preciso que esse sinal satisfaça algumas condições (Krishna, M. P., Gadre, V. M., & Dessay, U. B. 2003):

- O sinal deve ser positivo, uma vez que o modelo foi desenvolvido para representar tempos de chegada de tráfego e, portanto, não pode assumir valores negativos;
- O sinal deve apresentar múltiplas escalas; por conseguinte sua auto-similaridade não é estacionária;
- O sinal deve exibir distribuição não-Gaussiana.

O multiplicador r é uma variável aleatória escolhida de uma distribuição de probabilidades $f_{R_j}(r)$, $0 \leq r \leq 1$, onde j indica o estágio da cascata. Assume-se que $f_{R_j}(r)$ é simétrica em torno de $r = 1/2$, logo, tanto r quanto $(1 - r)$ têm a mesma distribuição de probabilidade.

3.4.1 ESTIMAÇÃO DA DENSIDADE DE PROBABILIDADE DOS MULTIPLICADORES

Dado que $X_i^N, i = 1, \ldots, 2^N$, representa o estágio N obtido no processo de construção de uma cascata (com tempo de resolução de 2^{-N}), esta cascata pode ser restaurada através do processo inverso da construção, ou seja, determinando-se os níveis anteriores. Cada nível é determinado baseando-se no estágio posterior. Logo, o nível $(N - 1)$ é obtido por um processo de agregação do estado N. Este processo de agregação consiste na adição de valores consecutivos em blocos não-sobrepostos de tamanho 2. Este processo é ilustrado na Figura 3.2.

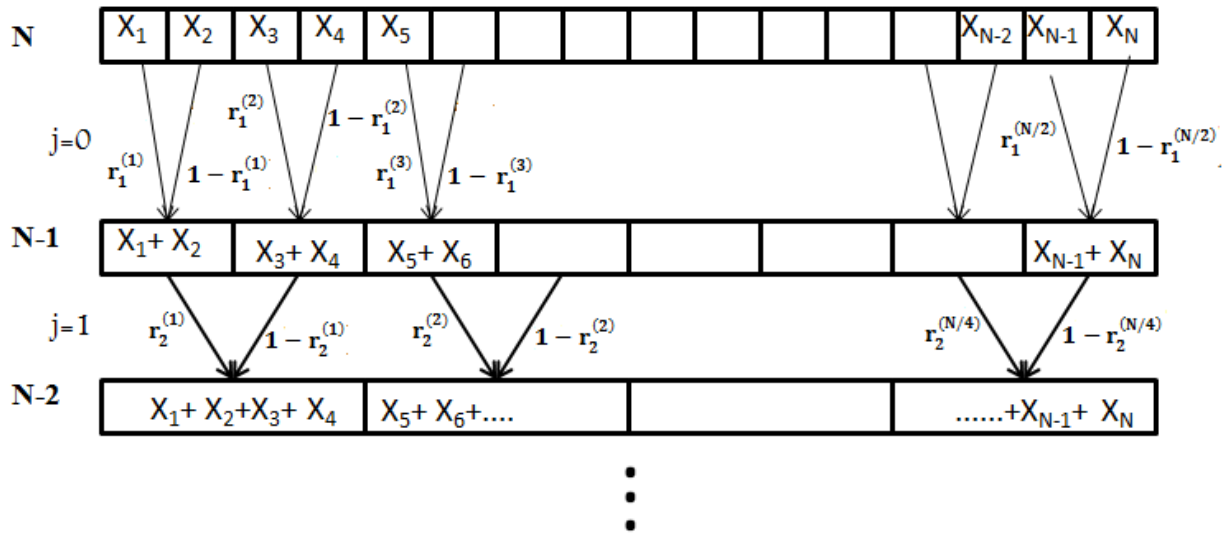


Figura 3.2 – Diagrama do processo de estimação dos multiplicadores.

De forma geral, dada uma série na escala $(N-j)$, $X_i^{N-j}(i=1,\dots,2^{N-j})$, são obtidos os dados no estágio $(N-j-1)$ pela soma dos valores do estágio $(N-j)$, chegando a expressão:

$$X_i^{N-j-1} = X_{2i-1}^{N-j} + X_{2i}^{N-j} \tag{3.12}$$

para $i=1,\dots,2^{N-j-1}$. Este procedimento termina quando a agregação dos valores forma apenas um ponto na última escala da cascata. Uma estimativa dos multiplicadores $r_j^{(i)}$ pode ser obtida tendo em conta a transição do estado j para o estado $j+1$, dado pela seguinte equação:

$$r_j^{(i)} = \frac{X_{2i-1}^{N-j}}{X_{2i}^{N-j-1}} \tag{3.13}$$

para $i = 1,\dots,2^{N-j-1}$. Os $r_j^{(i)}$ podem ser considerados amostras da distribuição $f_{R_r}(r)$ no estágio j. A distribuição dos multiplicadores na escala j pode ser obtida pelos histogramas de $r_j^{(i)}$. Este método assume que os multiplicadores possuem uma distribuição de probabilidade Gaussiana com média *r=0,5* e variância variável para cada nível da cascata. Nas Figura 3.3 e 3.4, são mostrados dois histogramas obtidos de um processo de agregação, para um trecho de sinal de

fala de 4096 amostras (f_s=11,025 kHz). No Capítulo 5, será discutido com maior profundidade este procedimento (algoritmos e descrição das bases de dados).

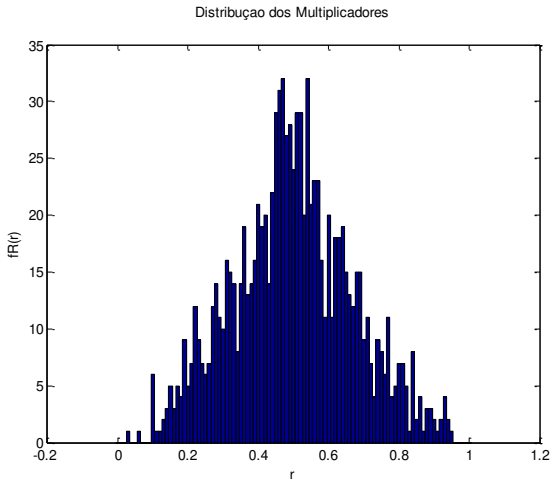


Figura 3.3 – Histograma estágio 2.

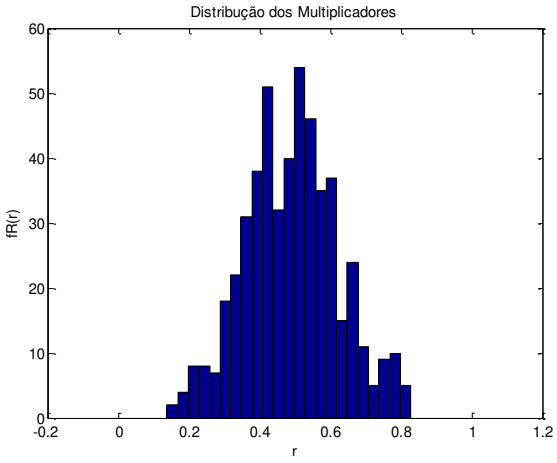


Figura 3.4 – Histograma estágio 3.

Da distribuição $f_{R_r}(r)$ obtida em cada estágio, é estimada a variância para cada nível da cascata. Este método sugere que a mudança das variâncias da distribuição do multiplicador seja parametrizada, utilizando um ajuste de curva. Assim, é determinada uma equação paramétrica para a variação da variância, a qual depende somente do estágio da cascata *j*.

Na análise acima, foi considerada uma cascata com os multiplicadores m_1 e m_2 fixos e $m_1 + m_2 = 1$. Ao se permitir que os multiplicadores da cascata sejam variáveis aleatórias independentes em [0,1] com densidade de probabilidade $f_{R_r}(r)$, obtém-se uma estrutura mais geral do que a determinística que emprega multiplicadores de valor fixo. Desta forma, para o estágio l da cascata obtida com intervalo diádicos de comprimento $\Delta t_l = 2^{-l}$, onde o começo de t está definido pela combinação binária $t = 0.\eta_1 \ldots . \eta_l = \sum_{k=1}^{l} \eta_k 2^{-k}$,onde $\eta_k = 0$ ou 1 a medida μ está definida como:

$$\mu(\Delta t_l) = R(\eta_1).R(\eta_1, \eta_2), \ldots, R(\eta_1 \ldots . \eta_l), \tag{3.14}$$

onde $R(\eta_1 \ldots . \eta_k)$ representa o multiplicador no estágio k da cascata. Através da Figura 3.2 pode-se visualizar este conceito. Considerando o estágio $l = N - 1$, o qual está composto por 2^{N-1} intervalos diádicos de tamanho $2^{-(N-1)}$, a medida μ para o primeiro intervalo é dada por:

$$\mu\left(\left[0, \frac{1}{2^{N-1}}\right]\right) = r_{N-1}^{(1)}.r_{N-2}^{(1)}.\ldots.r_2^{(1)}r_1^{(1)}$$

De igual forma pode ser determinada a medida μ para qualquer intervalo do estágio l. Considerando que os multiplicadores são i.i.d, a medida μ atende a relação de escala (Mandelbrot 1982):

$$E(\mu(\Delta t_l)^q) = \left(E(R^q)\right)^l = \Delta \mathrm{t_j}^{-\log \mathrm{E_2(R^q)}} \tag{3.15}$$

que define um processo multifractal com função de escala $\tau(q) = -\log_2 E(R^q) - 1$. Comparando a Equação (3.15) com a definição de processo multifractal descrito pela Definição 2.2.1, pode-se observar que a cascata binomial satisfaz esta condição (Vieira, F.H.T. & Lee L.L. 2006).

4 Reconhecimento Automático de Locutor

Neste capítulo serão introduzidos alguns conceitos que facilitam a compreensão do funcionamento de um sistema de reconhecimento automático de locutor. Na seção 4.1, são apresentados os conceitos gerais dos ASR. As quatro seções seguintes se referem a primeira fase do ASR, quando o sinal acústico de entrada é convertido em uma sequência de vetores de características. Assim, na seção 4.2, apresenta-se o pré-processamento, o qual consiste na preparação prévia do sinal de entrada para ser usado posteriormente; na seção 4.3, é apresentada a extração dos coeficientes MFCCs a partir de um banco de filtros na escala Mel; outros parâmetros característicos são analisados na seção 4.4. Para finalizar, é apresentada nas seções 4.5 e 4.6, a segunda fase do ASR, na qual se efetua a modelagem de cada locutor através de uma mistura de gaussianas (GMM) e o sistema de identificação de locutor.

4.1 Introdução

O ASR é um dos métodos mais naturais e econômicos para resolver problemas de autorização/senha. Como sugere (Campbell, J. 1997), a combinação entre a anatomia inerente ao trato vocal e os hábitos de diferentes indivíduos faz com que o sinal de fala contenha uma grande quantidade de informações da identidade do locutor, tornando o sistema de reconhecimento de locutor um método bastante eficaz.

Na Figura 4.1, são mostradas as áreas de aplicação do processamento de fala. Ao se concentrar nos sistemas de reconhecimento de locutor, é importante destacar mais uma vez que esta área é classificada em duas categorias, identificação e verificação.

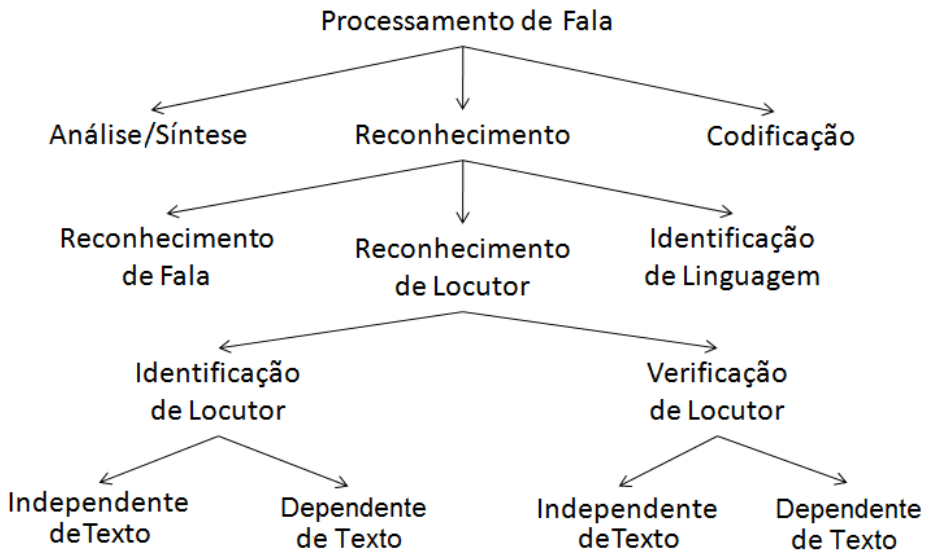


Figura 4.1– Processamento de fala (Campbell, J. 1997).

Nos sistemas de identificação automática de locutor (ASI), o usuário não fornece nenhuma informação sobre sua identidade e, assim, o sistema é responsável por determinar quem é o locutor, dentro de um grupo de indivíduos previamente cadastrados. Nos sistemas de verificação automática de locutor (ASV), o locutor fornece sua identidade (senha específica) e o sistema decide aceitar ou recusar o usuário, dependendo da comparação com o seu padrão armazenado. Adicionalmente, os ASR podem operar de dois modos: dependente de texto e independente de texto. No modo dependente de texto, tanto na etapa de treinamento quanto na etapa de reconhecimento, é usado um texto predeterminado fornecido ao usuário (senha fixa). Embora este modo ofereça um melhor desempenho devido a prover informação adicional (transcrição de texto) (Reynolds, D. 2002), ele apresenta algumas desvantagens, tais como precisar de novos treinamentos cada vez que a senha seja alterada e ter uma maior

probabilidade de ataque por impostores ao utilizar uma frase fixa. Por outro lado, no modo independente de texto, tanto as locuções de treinamento quanto as de teste são diferentes, permitindo ao usuário falar livremente, o que torna este modo mais seguro. Os dois modos de operação têm tarefas e objetivos diferentes e, portanto, podem empregar técnicas diferentes. Os sistemas independentes de texto são usualmente tratados com técnicas baseadas em GMMs (Gaussian *Mixture Models*), enquanto os sistemas dependentes de texto com técnicas como DTW (*Dynamic Time Warping*) ou HMMs (*Hidden Markov Models*).

Cabe destacar que este projeto é focado em um sistema de identificação do locutor operando em modo independente de texto. Na Figura 4.2, é apresentado o diagrama de um sistema de identificação tradicional. Basicamente, um sistema de identificação de locutor é composto por 3 módulos: aquisição do sinal digital (conversão do sinal analógico em digital), extração de parâmetros e comparação com um modelo ("pattern matching"). A descrição detalhada de cada módulo será dada nas seções seguintes.

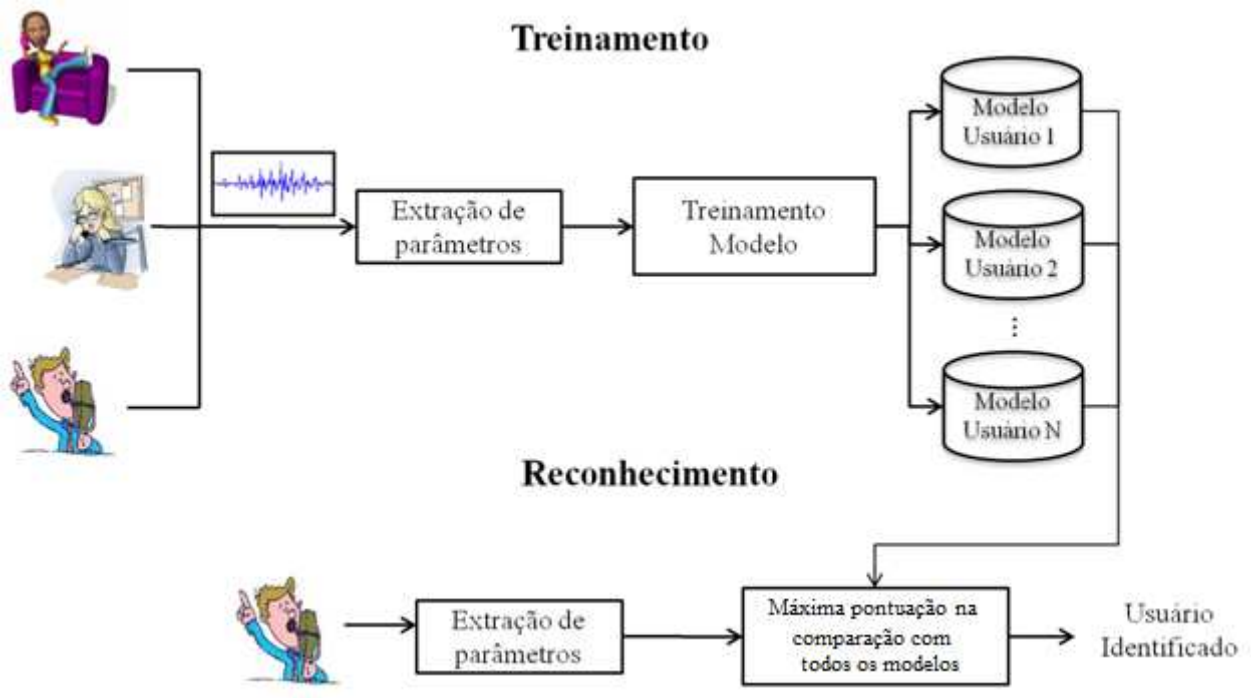


Figura 4.2 – Sistema de identificação de locutor convencional. O sistema seleciona o modelo do locutor que tenha maior semelhança.

4.2 Pré-Processamento

Este processo é realizado sobre o sinal de fala de entrada, adequando-o para a extração de parâmetros característicos. Usualmente, no pré-processamento são efetuadas tarefas tais como normalização, filtragem, remoção de trechos de sinal indesejados, redução de taxa de amostragem, entre outras.

4.2.1 PRÉ-ÊNFASE

O espectro da fala é caracterizado por uma tendência descendente, pela qual as frequências na parte superior do espectro são atenuadas em cerca de -6 dB/oitava. Esta tendência deve-se a combinação da inclinação negativa de -12 dB/oitava do espectro da fonte glotal com a elevação de +6 dB/oitava dada pelo efeito da radiação devido aos lábios (Holmes, J. & Holmes, W. 2001). Para compensar esse efeito, é comum se aplicar uma pré-ênfase de +6 dB/oitava. Normalmente essa pré-ênfase é feita empregando-se um filtro FIR definido segundo a Equação (4.1) (Picone, J. 1993):

$$H_{pres}(z) = 1 - \alpha z^{-1} \tag{4.1}$$

Normalmente, o valor usado para α está em torno de 0,95. Existe outra motivação para o uso deste filtro no pré-processamento, dado que a audição humana apresenta uma maior sensibilidade nas frequências entorno a 1 kHz. O filtro de pré-ênfase amplifica esta região, ressaltando os aspectos perceptualmente importantes do espectro do sinal de fala (Picone, J. 1993).

4.3 Coeficientes Mel-Cepstrais

Para o desenvolvimento de um ASR, é necessário converter o sinal de fala de cada locutor numa representação paramétrica, que contenha informação relevante da fonte geradora e possa ser interpretada pelo sistema. Nesse processamento, é amplamente usada a análise

espectral de curto tempo para sua caracterização. Isso se deve ao fato de que a caixa de ressonância, composta por laringe, faringe, boca e cavidade nasal, é um filtro mecânico, com movimentos lentos. Portanto, o sinal de fala pode ser considerado estacionário em curtos intervalos de tempo (da ordem de 20ms). Existem diversos parâmetros para caracterizar o sinal de fala, tais como os coeficientes LPC (*Linear Prediction Coding*), MFCCs (*Mel-Frequency Cepstrum Coefficients*), parâmetros prosódicos, entre outros. Este capítulo se concentra no uso dos MFCCs, já que é o método clássico encontrado literatura, como referencia dados seus bons resultados (Reynolds, D. A 1994) (Kinnunen, T. & Li, Haizhou. 2010).

Os MFCCs são parâmetros baseados na percepção auditiva humana. Estudos empíricos demonstraram que o sistema de audição humana responde a frequências de uma maneira não-linear. Adicionalmente, esses estudos mostraram que a resolução em frequência do sistema de audição apresenta linearidade para frequências inferiores a 1000 Hz e logarítmicas acima deste valor (Volkmann, J., Stevens, S. & Newman, E 1937). Essa resolução não linear pode ser aproximada na escala Mel como:

$$M(f) = 1127{,}01048 \log_{\mathrm{e}}\left(1 + \frac{f}{700}\right) \tag{4.3}$$

onde f é a frequência em Hz (Picone, J. 1993).

A idéia principal para calcular os MFCCs é realizar uma análise em frequência com base num banco de filtros triangulares. Para determinar o espaçamento e a largura de banda destes filtros, é usado o conceito de banda crítica[1], o qual fornece uma indicação da banda efetiva do filtro auditivo (Holmes, J. & Holmes, W. 2001). Além disso, a largura de banda dos filtros varia com a frequência. Assim, para frequências inferiores a 1000 Hz, são empregados filtros com largura de banda da ordem de 100 Hz; para frequências superiores a largura aumenta logaritmicamente. O número de filtros pode mudar dependendo da aplicação desejada e da

[1] *Informações mais detalhadas sobre a banda crítica são apresentadas por (Holmes, J. & Holmes, W. 2001) no capítulo 3 "Mechanisms and Models of the Human Auditory System".*

frequência de amostragem empregada (f_s) . Assim, para uma $f_s = 8\,kHz$, são usados em torno de 18 filtros; para $f_s = 11{,}025\,kHz$, em torno de 21 filtros; para $f_s = 16\,kHz$, em torno de 23; e para $f_s = 22{,}05\,kHz$, em torno de 26 filtros.

Na Figura 4.3, observa-se o diagrama de blocos do processo de extração dos parâmetros MFCCs presente na maioria das implementações. Assim, a análise de Fourier de curto tempo é aplicada a um sinal de entrada através da DFT (*Discrete Fourier Transform*). Em seguida, os valores do módulo da DFT são agrupados em bandas críticas e ponderados por uma função triangular. Na saída de cada um destes filtros é calculada a energia. Finalmente se efetua o cálculo da DCT (*Discrete Cosine Transform*) do logaritmo da energia calculada na saída de cada filtro. Esta transformação possui a propriedade de comprimir a informação nos coeficientes de baixa ordem e também produz uma descorrelação entre os coeficientes. A dimensão M dos coeficientes MFCCs extraídos é definida tomando-se os M coeficientes iniciais da DCT. Cabe mencionar que o coeficiente de ordem zero não é usado na prática, já que depende do ganho do sinal.

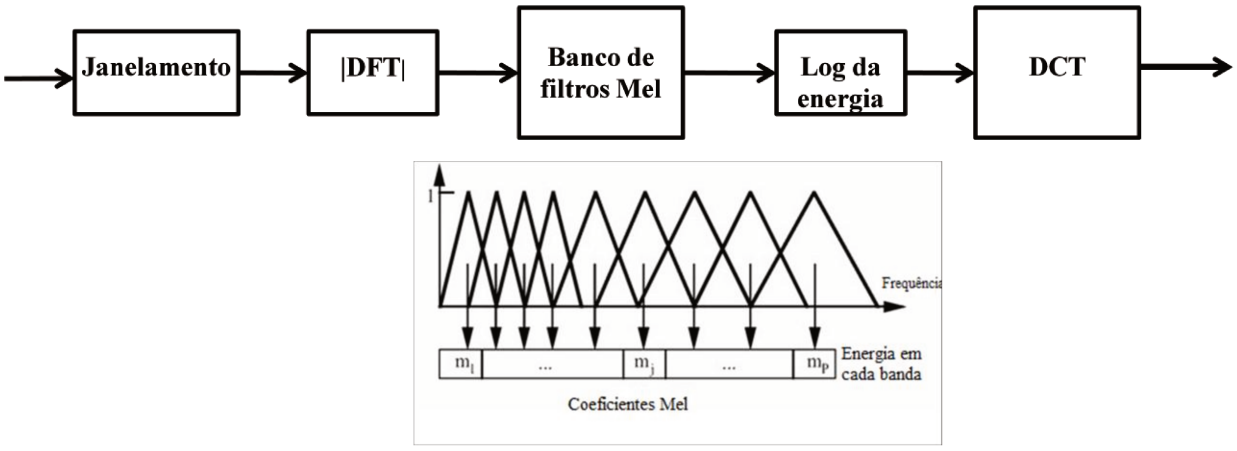


Figura 4.3 – Diagrama do processo de extração dos MFCCs.

4.4 Parâmetros Adicionais

Para o desenvolvimento de sistemas de reconhecimento de locutor, são usados, além dos parâmetros analisados na seção anterior, outros parâmetros alternativos, tais como a energia e a

taxa de cruzamentos por zero. Estes dois parâmetros são igualmente analisados em intervalos de tempo curto (da ordem de 20ms). O uso mais comum destes parâmetros é na supressão de silêncios.

A energia é uma informação simples de determinar num sinal, e pode ser calculadas através da Equação (4.4), para um sinal N amostras.

$$E = \sum_{n=1}^{N} x^2(n) \tag{4.4}$$

Usualmente se trabalha com a função logaritmo da energia para se acentuar as baixas mudanças. A partir deste parâmetro, são eliminados trechos do sinal de baixa energia, que correspondem a silêncio ou ruído, e não contêm informação relevante da identidade locutor.

A taxa de cruzamentos por zero é definida pelo número de vezes que um sinal de fala troca de polaridade no intervalo de tempo analisado. Trechos de sinal que possuem taxas de cruzamento maiores que certo limiar estabelecido são eliminados. Este limiar é determinado empiricamente testando taxas de cruzamentos de diferentes fonemas com alto conteúdo nas altas frequências, como é o caso das fricativas, tais como "*s*", "*f*" e "*x*".

4.5 Classificador

Nesta seção, será apresentado o método de classificação para identificação de locutor independente de texto adotado neste trabalho. A técnica é conhecida como o modelo estatístico *GMM* (*Gaussian Mixture Models*), introduzido por Reynolds, 1992, e que, na atualidade, é o método mais usado por ter demonstrado os melhores resultados.

O GMM pode ser entendido como um HMM de um único estado, com densidade de probabilidade modelada como uma mistura de gaussianas multidimensionais, onde cada uma destas misturas pode representar uma ou várias classes fonéticas que compõem o som produzido por uma pessoa.

4.5.1 MODELOS DE MISTURA DE GAUSSIANAS (GMM)

Uma mistura de Gaussianas é a soma ponderada de M densidades gaussianas multidimensionais, e pode ser descrita matematicamente pela seguinte equação:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i b_i(\vec{x}) \tag{4.5}$$

onde $\vec{x}$ é um vetor aleatório de dimensão D (vetor de características do locutor), $b_i(\vec{x}), i = 1, \dots, M$ são as M densidades de dimensão D e $w_i, i = 1, \dots, M$, são os pesos dessas densidades. Cada componente Gaussiana da mistura com vetor de médias $\vec{\mu}_i$ e matriz de covariância Σ_i é dada por:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x} - \vec{\mu}_i)\right\} \tag{4.6}$$

onde $|.|$ representa determinante e $()^T$ indica transposta. A ponderação das misturas deve satisfazer a condição $\sum_{i=1}^{M} w_i = 1$. Nos sistemas de identificação de locutor, cada pessoa está caracterizada por um GMM, chamado λ, representado pela notação

$$\lambda = \{w_i, \vec{\mu}_i, \Sigma_i\} \quad i = 1, \dots, M \tag{4.7}$$

Esta modelagem utiliza uma matriz de covariância para cada componente da mistura. Adicionalmente a isso, a matriz de covariância pode ser também cheia ou diagonal, dependendo da correlação entre as componentes do vetor de características. Normalmente, para aplicações com pouco material de treinamento, são mais empregadas matrizes diagonais (componentes independentes entre si).

O GMM tem a capacidade de modelar densidades de probabilidades arbitrárias (Reynolds, D. 2002) (Vuuren, V. S. 1999), especificamente a distribuição dos vetores de características extraídos de uma locução.

4.5.2 ESTIMAÇÃO DE PARÂMETROS DO MODELO

No sistema de identificação de locutor, cada locutor está representado por um GMM λ. Os parâmetros desse modelo são estimados na etapa de treinamento, onde se tem como objetivo principal encontrar uma representação mais adequada para vetores característicos de fala. Existem diversos métodos disponíveis para estimação dos parâmetros do GMM (McLachlan, G. & Peel, D. 2000). Um método amplamente difundido e que apresenta bom desempenho é a estimação da máxima verossimilhança (Maximum Likelihood - ML).

Este método de estimação tem como princípio escolher os parâmetros do modelo λ que maximizam a função de verossimilhança de um conjunto de observações. Para uma sequência de entrada de T vetores de treinamento $\boldsymbol{X} = \{\vec{x}_i,\ i = 1, \dots, T\}$, pode ser definida a função de verossimilhança para modelo λ como a função de densidade de probabilidade conjunta de $\boldsymbol{X}$ dado o modelo λ, descrita por $p(\boldsymbol{X}|\lambda)$. Assumindo independência entre os vetores de entrada, a função de verossimilhança pode ser escrita como:

$$p(\boldsymbol{X}|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda) \tag{4.8}$$

Normalizando pelo número total de vetores T e usando o logaritmo, chega-se a

$$\log p(\boldsymbol{X}|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|\lambda) \tag{4.9}$$

No entanto, a Equação (4.7) é uma função não linear dos parâmetros λ. Portanto, obtém-se um conjunto não fechado de soluções, sendo impossível a maximização direta (Reynolds, D. 2002). Contudo, os parâmetros do modelo λ podem ser obtidos iterativamente através do

algoritmo EM (*Expectation Maximization*) descrito por (Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977).

O EM é um algoritmo iterativo, o qual assume um modelo λ inicial como base para a estimação de um novo modelo $\bar{\lambda}$, tal que $p(\boldsymbol{X}|\bar{\lambda}) \geq p(\boldsymbol{X}|\lambda)$. O novo modelo torna-se o modelo inicial para a seguinte iteração, e esse processo é repetido até que um limiar de convergência seja alcançado. O algoritmo EM funciona alternando iterativamente entre duas etapas distintas. Na primeira etapa, chamada E (expectation), são calculados os valores de $b_i(\vec{x}_t)$ de cada uma das componentes do GMM. Na segunda etapa, chamada de M ("maximization"), são atualizados os parâmetros do modelo, tendo por base as seguintes expressões:

Pesos da mistura:

$$\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda) \tag{4.11}$$

Médias:

$$\vec{\bar{\mu}}_i = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} \tag{4.12}$$

Variâncias:

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} - \bar{\mu}_i^2 \tag{4.13}$$

onde σ_j^2, x_t e μ_j, $j = 1, \ldots D$, referem-se aos elementos dos vetores $\vec{\sigma}_i^2$, $\vec{x}_t$, $\vec{\mu}_i$ respectivamente.

A probabilidade *a posteriori* para uma classe acústica *i* é dada por

$$p(i|\vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \tag{4.14}$$

4.6 Sistema de Identificação de Locutor

No processo de treinamento, o sistema gera e armazena modelos $(\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_N)$ para um grupo de N locutores. No processo de teste ou identificação de locutor, o sistema recebe novas locuções (vetores de características) pertencentes a qualquer um dos locutores treinados com o objetivo de encontrar o modelo λ que assegura a máxima probabilidade *a posteriori*. Assim,

$$\widehat{N} = \arg \max_{1 \le k \le N} \Pr(\lambda_k | \boldsymbol{X}) = \arg \max_{1 \le k \le N} \frac{p(\boldsymbol{X}|\lambda_k) \Pr(\lambda_k)}{p(\boldsymbol{X})} \tag{4.15}$$

onde a segunda equação é obtida através da regra de Bayes. Assumindo que todos os modelos treinados têm a mesma probabilidade de ocorrerem, tem-se $\Pr(\lambda_k) = 1/N$. Adicionalmente, presume-se que $p(\boldsymbol{X})$ é igual para todos os modelos, já que depende unicamente da locução testada. Daí resulta que a identificação do locutor pode ser simplificada como

$$\widehat{N} = \arg \max_{1 \le k \le N} p(\boldsymbol{X}|\lambda_k) \tag{4.16}$$

Utilizando o logaritmo e assumindo independência entre as observações, tem-se

$$\widehat{N} = \arg \max_{1 \le k \le N} \sum_{t=1}^{T} \log p(\vec{\boldsymbol{x}}_t|\lambda_k) \tag{4.17}$$

onde $p(\vec{\boldsymbol{x}}_t|\lambda_k)$ é definido na Equação (4.5). O sistema de identificação aceita o locutor que possui o modelo que maximize a verossimilhança. Na Figura 4.4, se observa o diagrama de blocos do sistema de identificação de locutor.

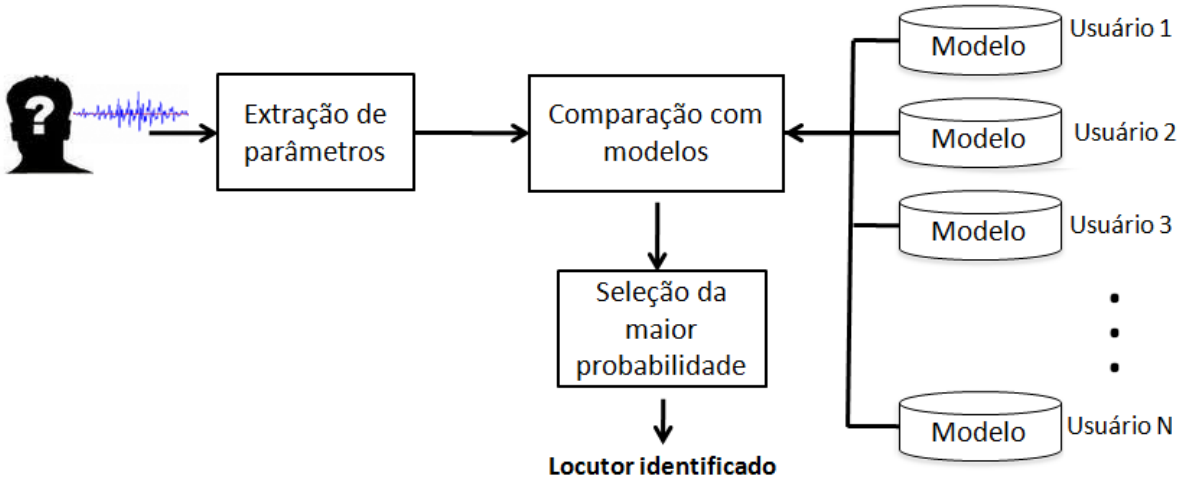


Figura 4.4 – Diagrama do sistema de identificação de locutor.

5 Sistema Desenvolvido

A estrutura de um sistema automático básico de identificação de locutor é formada por três módulos principais, como ilustrado na Figura 4.2, e listado abaixo:

- Módulo de extração de parâmetros
- Módulo de treinamento
- Módulo de reconhecimento

Na primeira fase deste trabalho, um sistema ASI tradicional independente de texto foi desenvolvido com o objetivo de criar um sistema padrão de referência para a avaliação dos novos sistemas a serem desenvolvidos. Este sistema empregou parâmetros MFFCs como vetor de características, os quais foram modelados através de GMM.

Em uma segunda fase, os novos parâmetros obtidos através do modelo multifractal VVGM foram usados como vetor de características. Os módulos de treinamento e reconhecimento do sistema desta fase foram dimensionados analogamente, baseando-se na técnica GMM.

Na terceira fase do trabalho, o sistema ASI desenvolvido conjugou os dois tipos de características com o objetivo de atingir um ASI mais robusto e taxas de reconhecimento superiores. A seguir, serão apresentadas algumas informações técnicas dos três módulos.

5.1 Módulo de Extração de Parâmetros

O módulo de extração de parâmetros transforma as locuções de entrada em parâmetros que possam ser interpretados pelos módulos seguintes. Este módulo tem por entrada um sinal de voz em formato WAV com 16 bits de resolução e amostrado a 8 kHz, 11,025 kHz ou 16 kHz, dependendo da base de fala empregada. Os parâmetros característicos (MFCCs e VVGM) dos sistemas desenvolvidos são calculados usando módulos de extração independentes. As subseções a seguir dedicam-se a análise destes parâmetros.

5.1.1 DESCRIÇÃO DA EXTRAÇÃO DOS PARÂMETROS MFCCS

Os parâmetros MFCCs são calculados utilizando informação de janelas de 20ms e deslocadas a cada 10ms. Esta segmentação em trechos dessa ordem se deve ao fato de que a caixa de ressonância pode ser considerada quase-estacionária em curtos intervalos de tempo, já que é um filtro mecânico com movimentos lentos. Na Figura 4.3, pode-se observar a arquitetura da extração dos parâmetros MFCCs. Antes da extração, o sinal é submetido a alguns pré-processamentos: retirada do nível DC, pré-ênfase com um filtro passa altas ($1 - 0{,}95\, z^{-1}$), e janelamento através de uma janela de Hamming. Além disso, o parâmetro log-energia é calculado para cada janela e normalizado, tomando como referência o quadro de maior energia em toda a locução sob análise, gerando um limiar para detecção e eliminação de silêncios por baixa energia. Ao normalizar a energia, seu valor máximo é 0 dB, e o limiar de energia foi escolhido em – 30 dB. Este limiar foi estabelecido a partir de medições da energia em algumas locuções de sons de fricativas de teste.

A partir do sinal janelado, são calculados os MFCCs. O sistema opera com 12 coeficientes Mel-Cepstrais. Note-se que o coeficiente de ordem zero do vetor de características não é usado, pois possui informação do ganho.

5.1.2 DESCRIÇÃO DA EXTRAÇÃO DOS PARÂMETROS CARACTERÍSTICOS ATRAVÉS DA MODELAGEM VVGM

Os novos parâmetros característicos de fala propostos para serem usados pelo ASR são baseados no modelo VVGM apresentado no capítulo 3. A seguir, será descrito o procedimento de extração destes parâmetros:

- *Pré-processamento:* Nesta etapa, varias operações preliminares são executadas, as quais foram resultado de investigação intensa e são os procedimentos necessários para adequar os sinas de entrada:
    1. *Pré-ênfase:* Emprega-se o mesmo filtro de pré-ênfase usado para o cálculo dos parâmetros MFCCs.
    2. *Normalização:* Normaliza-se a amplitude do sinal limitado entre +1 e -1, com o maior pico positivo (ou negativo) atingindo a amplitude de +1 (ou -1), aproveitando melhor a faixa dinâmica da locução sem chegar à saturação ou distorção. Este procedimento de normalização visa reduzir a influência da amplitude do sinal, como o volume do microfone e a distância do microfone ao locutor, homogeneizando todas as locuções de entrada.
    3. *Eliminação de silêncios:* Períodos de silêncio entre palavras, assim como no início e final das locuções, são removidos, usando o mesmo detector de silêncio implementado para os parâmetros MFCCs.
    4. *Adequação do sinal:* De acordo com a primeira condição fornecida na seção 3.4 para a implementação do modelo VVGM, o sinal a ser processado deve ser positivo. Dado que o sinal de fala apresenta amplitudes positivas e negativas, faz-se necessário efetuar um tratamento para que todas as amostras sejam positivas sem perda de informação. Para isso, foram testadas duas abordagens: deslocamento e retificação do sinal. Verificou-se que o sistema tem melhor desempenho usando sinal

retificado. Ao concluir esta adequação, foi necessária a agregação de um pequeno nível DC ao sinal para evitar sinais resultantes próximos a zero, os quais poderiam causar problemas no processo de agregação. Através de testes, observou-se que a etapa de eliminação de silêncios é essencial no processo de extração de parâmetros, pois a presença de silêncios polariza a variância (em torno de $1/2$ para sistemas deslocados e em torno de 0 para sistemas retificados), provocando perda de informação relevante.

- *Janelamento:* Usam-se janelas retangulares com 100ms de duração e atualização a cada 10ms. A escolha do tamanho da janela foi baseada em vários testes realizados, dos quais se concluiu que, para essa escala, a modelagem VVGM consegue caracterizar adequadamente os sinais de fala.
- *Adequação de duração:* Embora sejam usadas janelas de 100ms, o algoritmo trabalha com comprimentos da ordem de potências de 2 (2^N), onde N é o número de estágios possível da cascata. Portanto, é preciso limitar cada quadro ao número de amostras máximo que seja potência de 2. Por exemplo, dado um sinal de entrada com uma taxa de amostragem de 11,025 kHz, uma janela de 100ms contém 1102 amostras: no entanto, serão usadas só 1024 amostras ($2^{10} = 1024$).
- *Processo de agregação:* A reconstrução da cascata é levada a cabo como foi apresentado na sub-seção 3.4.1 e seguindo a Figura 3.2.
- *Histogramas:* Nesta etapa, são obtidos histogramas dos multiplicadores para os N estágios. A distribuição probabilística dos multiplicadores $f_r(R)$ tem o comportamento de função Gaussiana com média $\mu = 1/2$, e variância variável para diferentes estágios. Estas variâncias são determinadas para cada janela e armazenadas, criando uma matriz de parâmetros característicos de cada locutor. A ordem do vetor de características usado depende do número de estágios presente na cascata. Esse valor é função do número de amostras do quadro

analisado e da frequência de amostragem da locução processada. Assim, ao gerar uma cascata com N níveis, são usadas as variâncias dos primeiros $N-2$ níveis como parâmetros característicos.

5.2 Módulo de Treinamento

Este módulo é responsável por modelar cada um dos locutores em treinamento. Para todos os sistemas ASI desenvolvidos, este módulo é baseado em Modelos de Mistura de Gaussianas (*GMM*). O módulo recebe a matriz de características de cada locutor, onde o número de linhas representa o total de janelas analisadas em todo o material de treinamento e as *D* colunas dependem do algoritmo empregado para a extração de parâmetros (VVGM, MFCCs ou VVGM+MFCCs). Com estes parâmetros, o locutor é caracterizado por um GMM, chamado λ, e representado pela notação $\lambda = \{p_i, \vec{u}_i, \Sigma_i\},\ i = 1, \dots, M$, onde M é o número de densidades gaussianas empregadas, que é pré-fixado no início do treinamento. A ordem do modelo foi determinada experimentalmente, testando diversos valores. É muito importante escolher um número apropriado de componentes da mistura, pois, para este tipo de aplicação, o material de treinamento usado é reduzido e um número elevado de misturas poderia ocasionar uma partição excessiva do espaço de dados (*over-fitting*), enquanto que um número muito reduzido não seria suficientemente flexível para se aproximar ao modelo real seguido pelos dados.

O treinamento foi usado o algoritmo EM (*Expectation Maximization*) apresentado na seção 4.5.2. Para a aplicação deste algoritmo, é requerido um valor inicial dos parâmetros do modelo. Assim, os parâmetros $\{p, \vec{u}, \Sigma\}$ foram inicializados através do algoritmo *Segmental K-Means*, realizando um clustering prévio com M classes e gerando um λ inicial para cada locutor. Baseando-se no λ inicial, os parâmetros são ajustados iterativamente por meio do algoritmo EM. Este algoritmo converge quando a probabilidade de que as observações $\boldsymbol{X}$ tenham sido geradas pelo novo modelo $\bar{\lambda}$ é muito próxima ao do modelo anterior λ, $p(\boldsymbol{X}|\lambda) \approx p\left(\boldsymbol{X}\middle|\bar{\lambda}\right)$. Este processo foi repetido para cada locutor e armazenado em uma base de modelos.

É importante ressaltar que, nesta modelagem, foram usadas matrizes de covariância diagonais, devido a pouca quantidade de material de treinamento disponível.

5.3 Módulo de Reconhecimento

O módulo de reconhecimento é o responsável pelo mapeamento dos parâmetros acústicos correspondentes à locução de entrada de teste. Ele recebe como entrada a matriz de características de uma locução nova, pertencente a qualquer um dos locutores previamente treinados. Com o objetivo de encontrar o modelo λ que assegura a máxima probabilidade a posteriori, os novos parâmetros são comparados com cada um dos modelos dos locutores treinados, e escolhido o de maior probabilidade de ter gerado a nova locução, como é discutida na seção 4.6

.

5.4 Fusão de Sistemas

Os sistemas desenvolvidos com parâmetros VVGM e MFCCs são integrados, a fim de aperfeiçoar o sistema ASI. Esta combinação é executada através de dois métodos propostos, descritos a seguir.

5.4.1 FUSÃO NO NÍVEL DE PONTUAÇÃO

No primeiro método, os sistemas são integrados como se observa no diagrama de blocos da Figura 5.1. Os módulos de treinamento operam de forma independente para cada parâmetro. A modelagem é feita individualmente, pois cada parâmetro utiliza janelas de diferentes comprimentos para sua extração: 20ms para MFCCs e 100ms para VVGM.

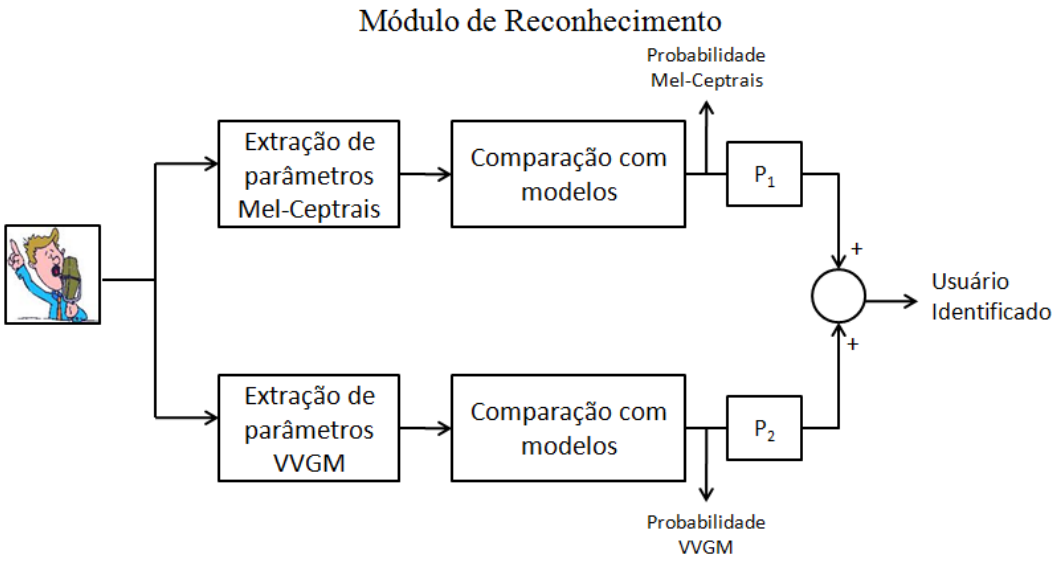


Figura 5.1 – Diagrama de blocos do ASI misturando os parâmetros MFCCs e VVGM.

No módulo de reconhecimento, são determinadas as probabilidades de cada locutor ter falado a nova locução, tanto para o sistema que emprega coeficientes VGGM quanto para o que usa MFCCs. Deste processo, são gerados dois vetores de probabilidades com comprimento igual ao número de locutores treinados. Cada um destes vetores de probabilidades é multiplicado por um peso de ponderação ($P_1 e\, P_2$), e os resultados são somados. Os pesos devem satisfazer a restrição $P_1 + P_2 = 1$. Os valores assumidos para os pesos foram determinados experimentalmente adotando o melhor resultado, ou seja, $P_1 = 0{,}6$ para os parâmetros MFCCs e $P_2 = 0{,}4$ para os parâmetros VVGM.

5.4.2 FUSÃO NO NÍVEL DE CARACTERÍSTICAS

Na segunda abordagem de fusão biométrica, os parâmetros VVGM e MFCCs foram combinados resultando em um único vetor de características, gerando um único GMM, como é ilustrado no esquema da Figura 5.2.

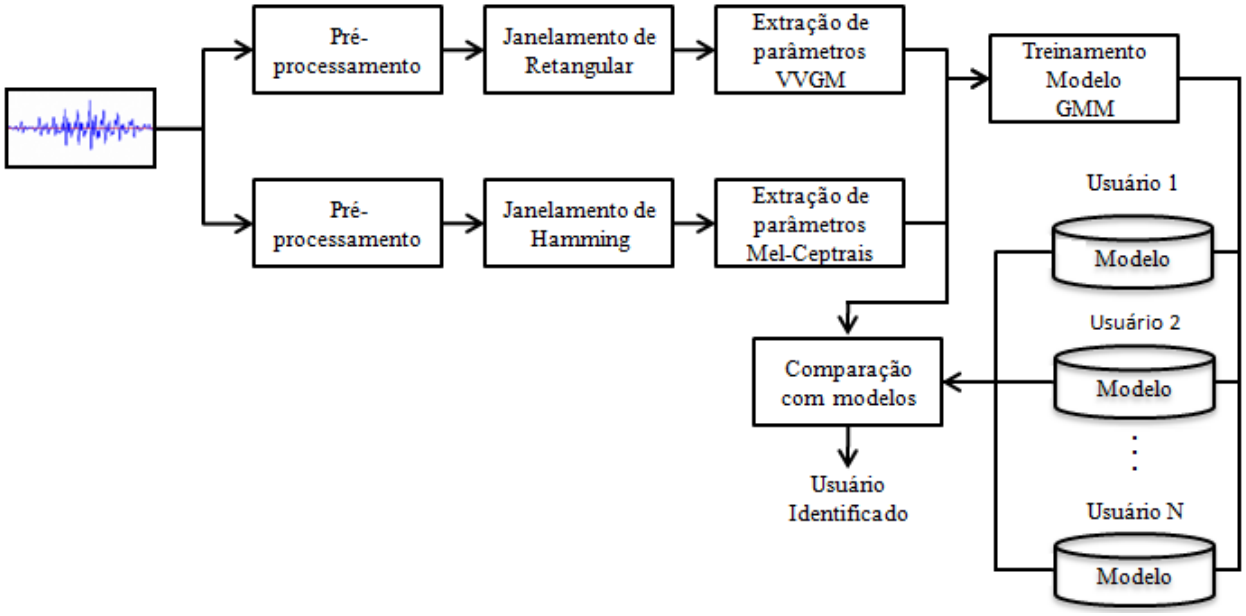


Figura 5.2 – Diagrama de blocos do ASI empregando fusão no nível de características.

Após o pré-processamento, os parâmetros MFCCs e VVGM são extraídos de forma independente. Seguido a isso, é efetuada a subdivisão do sinal em quadros usando janela de Hamming para os parâmetros MFCCs, e janela retangular para os parâmetros VVGM, com duração de 20 ou 30ms. Note que, para este método, o intervalo de análise é único para a extração de ambos os parâmetros. Devido a dimensão das janelas, os parâmetros VVGM não apresentam um comportamento ótimo, o que só é alcançado com quadros da ordem de 100ms, com os quais é possível gerar cascatas maiores. No entanto, o fato de misturar diretamente os dois parâmetros e modelar conjuntamente através de um GMM pode ser vantajoso para o sistema de identificação.

A dimensão do vetor de características empregado é de 12 parâmetros MFCCs mais $N - 2$ componentes gerados na cascata do VVGM, dependendo da frequência de amostragem da base de dados empregada.

5.5 Bases de Dados

Os experimentos foram realizados usando três bases de dados com diferentes configurações, tais como frequência de amostragem, número de locutores e de locuções, duração de treinamento e teste, entre outras. Os experimentos com diferentes configurações permitem determinar as características de sinal de fala requeridas pelo sistema.

A primeira base de fala, denominada "*Ynoguti 1*", foi criada originalmente por Carlos Alberto Ynoguti no Laboratório de Processamento Digital de Fala do DECOM/FEEC/UINICAMP (Ynoguti, C. & Violaro, F 1999) para aplicação em reconhecimento de fala. As gravações foram realizadas em ambiente relativamente silencioso, com um microfone direcional de boa qualidade, utilizando uma placa de som SoundBlaster AWE 64. As locuções estão armazenadas em formato Windows PCM (WAV). Esta base de dados emprega frases foneticamente balanceadas que, portanto, têm uma distribuição fonética similar àquela encontrada na fala espontânea. Ela foi implementada com uma frequência de amostragem de 11,025 kHz, com 16 bits/amostra, e é constituída por 30 locutores, 15 homens e 15 mulheres. Cada locutor leu quarenta frases diferentes, das quais trinta são usadas como amostras de treinamento, com uma duração total aproximada de 60 s, e as outras dez como amostras de teste do sistema, com comprimento variando entre 2,5 e 3 s.

A segunda base, denominada "*Ynoguti 2*" (Ynoguti, C. A. & Violaro, F. 2008) é composta por sinais de fala de 71 locutores (50 homens e 21 mulheres), digitalizados a 22,05 kHz e com 16 bits/amostra. O treinamento foi feito com 20 locuções de cada locutor, totalizando 70 s de duração na média. O sistema foi testado usando 10 locuções de cada locutor, cada uma com duração entre 3 e 4 s.

A terceira base, "*corpus ELSDSR*", foi desenvolvida no "Department of Informatics and Mathematical Modeling, Technical University of Denmark". Este corpus foi concebido para fornecer dados de fala para o desenvolvimento e avaliação de sistemas automáticos de reconhecimento de locutor em ambiente controlado. As locuções estão registradas em arquivo tipo WAV (PCM). A base é feita via gravação direta com microfone de alta qualidade (Feng, L. & Hansen, L. K. 2005), e é composta por sinais de fala de 22 locutores (11 homens e 11 mulheres)

com frequência de amostragem de 16 kHz e com 16 bits/amostra. Em média, a duração do material do treinamento é de 83 s. Para o teste são empregadas 2 locuções com duração, em média, de 17,6 s cada uma.

6 Análise da Natureza Multifractal em Sinais de Fala

Neste capítulo, é realizado um estudo sobre características multifractais presentes em sinais de fala, através de curvas multifractais como espectro multifractal $f(\alpha)$ ou funções de escalonamento. Estas curvas (curvas de singularidade) fornecem informações e orientações importantes para o processamento, como decomposição, representação e caracterização do espectro, de forma análoga a análise de Fourier em abordagens tradicionais (Langi, A. & Kinsner, W. 1995). Para este estudo, são avaliadas locuções das três bases de dados mencionadas anteriormente. Este estudo aproveita as diferentes taxas de amostragem da base, além de experimentar locuções de diferentes comprimentos a fim de observar e determinar a natureza multifractal para sinais em diferentes condições. Com isso, pretende-se abrir as portas para o uso de ferramentas multifractais em processamento de fala, como alternativa ou complemento aos métodos tradicionais. Para detectar a natureza multifractal do sinal de fala, foi analisado seu comportamento, tendo como base a teoria apresentada nos Capítulos 2 e 3 sobre o formalismo multifractal.

6.1 Testes

6.1.1 DESCRIÇÃO DOS SINAIS DE FALA

Os sinais de fala empregados nas simulações são locuções das três bases de fala citadas no capítulo 5. Foram selecionadas aleatoriamente 30 locuções de alguns locutores de cada base de fala, para a formação do subconjunto de teste. Deve ser lembrado que as locuções provenientes de bases de fala diferentes possuem diferentes frequências de amostragem.

Antes da análise multifractal do sinal de fala proposto, as locuções são submetidas a um pré-processamento, com o fim de corrigir alguns fatores que podem atrapalhar ou alterar o comportamento real do sinal. Nesta etapa são efetuados: filtragem de pré-ênfase, normalização e eliminação de trechos de silêncio. Esses procedimentos são implementados da mesma forma como apresentado em capítulos anteriores.

6.1.2 INVESTIGAÇÃO EXPERIMENTAL

Nesta subseção, é apresentado um trabalho experimental para a estimação de parâmetros multifractais em sinais de fala, baseado nos métodos de análise multifractal descritos no Capítulo 2. Graficamente, foi avaliado o comportamento multifractal de sinais de fala, encontrando características multifractais similares para todas as locuções testadas. A metodologia adotada para o desenvolvimento desta avaliação é formulada em duas etapas sequenciais de processamento: (a) inicialmente, é aplicado o método dos momentos para obter a função de partição $\mathcal{X}_m^X(q)$ e a função de escala $\tau(q)$, abordadas na seção 2.3.1; (b) é feita a análise do comportamento de escala para diferentes classes fonéticas (vogais, fricativas, etc.) presentes no sinal de fala espontânea, através do espectro de Legendre tratado na seção 2.3.2.

Teste Experimental 1: O método dos momentos foi implementado e desenvolvido no software MATLAB. O algoritmo determina a soma de partição e a função de partição através da variação da ordem do momento q. Para fins de ilustração, as Figura 6.1.a e 6.1.c mostram as curvas das funções de partição ($\log \mathcal{X}_{\mathrm{m}}^{\mathrm{X}}$ *versus* $\log m$) de duas locuções arbitrariamente selecionadas das bases ELSDSR e Ynoguti 2, respectivamente. De fato, a maioria das locuções avaliadas apresentam comportamentos similares aos ilustrados nas figuras. Observe que estas

curvas das funções de partição são bastante lineares relativamente, independentemente da frequência de amostragem e duração do sinal. Isso significa que os sinais de fala podem ter um comportamento ou característica fractal. Entretanto, existem pequenas irregularidades, que indicam que os sinais podem apresentar diferentes propriedades de escala, ou seja, comportamento não uniforme em diferentes escalas.

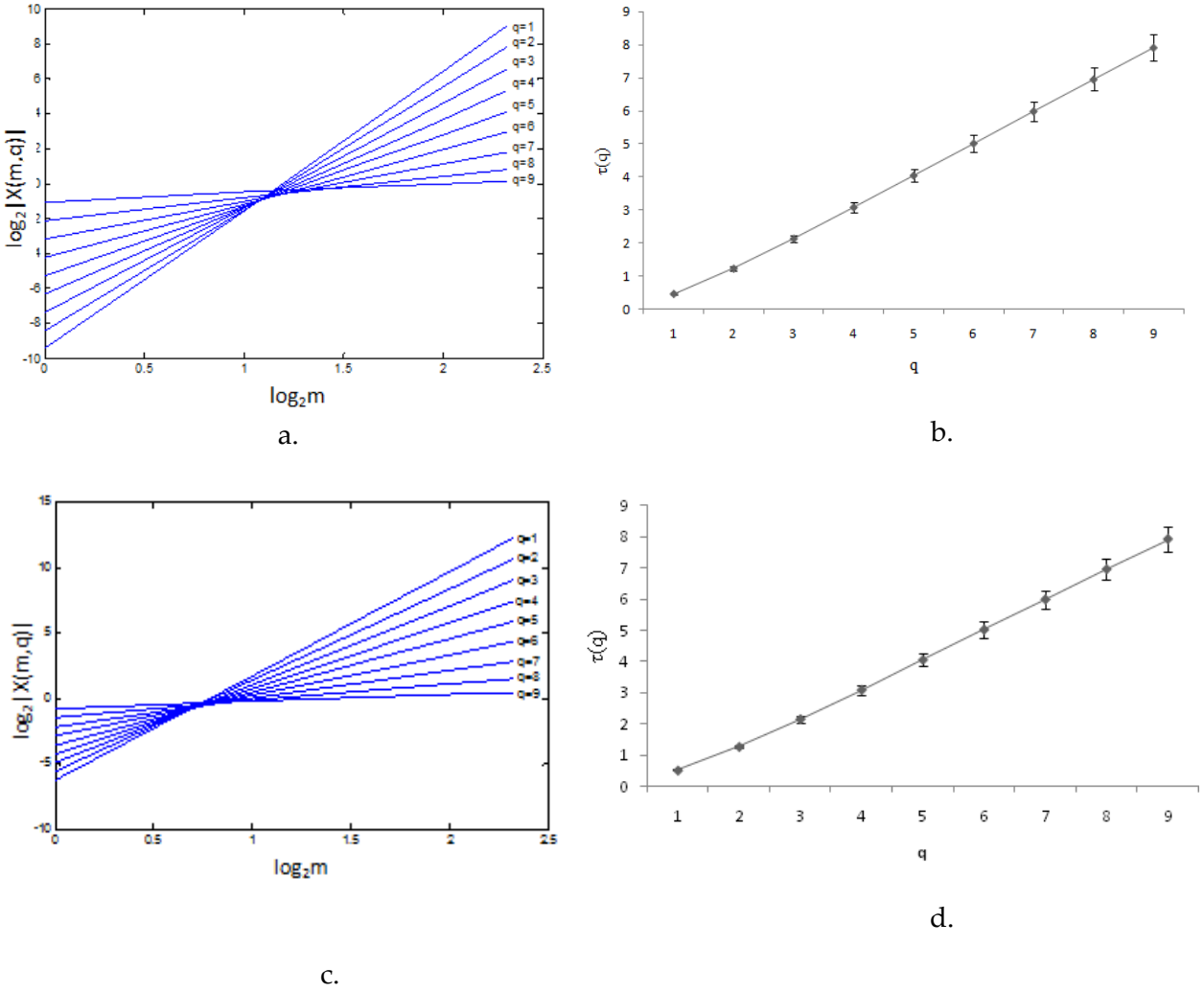


Figura 6.1 – Função de Partição: (a e c); Função de escalonamento $\tau(q)$ vrs q: (b e d).

As curvas da função de escalonamento $\tau(q)$ ilustradas na Figura 6.1.b e 6.1.d mostram que $\tau(q)$ não é linear em função de q, o que sugere a existência de propriedades de multi-escala, ou seja, multifractal. Nestes gráficos as barras verticais representam os intervalos de confiança

de 95%, dos valores estimados da função de escalonamento $\tau(q)$ em relação a cada ordem do momento q, considerando todas as locuções do conjunto do teste tanto para a base *Ynoguti 2* como para a base *ELSDSR*.

Para as três bases analisadas, quase todos os intervalos de confiança são pequenos e apresentam dinâmicas semelhantes em q. Além disso, foi avaliada a correlação entre as diferentes funções de escalonamento estimadas a partir das distintas locuções (Figura 6.1.b e 6.1.d). Foi observado que, para cada ordem do momento q, a função de correlação tende a 1, o que implica forte correlação entre as locuções utilizadas nos experimentos.

A análise do sinal da fala realizada neste *Teste Experimental 1* é do tipo inspeção visual sobre a função $\tau(q)$ e não pode ser definitiva ou conclusiva, ainda que sugira a presença ou não de propriedades de escala diferentes. Portanto, adicionalmente, foi adotada uma abordagem de análise complementar através da ferramenta espectro multifractal (espectro de Legendre). Essa abordagem é geralmente muito mais confiável, informativa, e definitivamente, conclusiva.

*Teste Experimental 2: Ne*ste teste experimental, sinais de fala são analisados por meio de fonemas, a menor unidade sonora das quais as palavras são compostas, e do relacionamento deles com outros fonemas vizinhos. Conforme relatado no Capítulo 2, o espectro multifractal fornece informação do grau de singularidade de um sinal no tempo e, portanto, da mudança do expoente de Hölder. Esta variação do expoente de singularidade ao longo do tempo permite determinar o comportamento multifractal de uma série temporal. O espectro multifractal do sinal de fala é obtido a partir da aplicação da transformada de Legendre através do software MATLAB e da ferramenta FRACLAB, desenvolvida pelo **centro de pesquisa INRIA Saclay - Île-de-France** (Institut National de Recherche en Informatique et en Automatique) e IRCCyN (L'Institut de Recherche en Communications et Cybernétique, Nantes).

Neste teste, foi usado o mesmo conjunto de locuções do *Teste Experimental 1*, mas focalizando-se nas bases de dados "*Ynoguti 1*" e "*Ynoguti 2*", pois os fonemas empregados variam para cada língua, e estas duas bases estão compostas por locuções em português nativo. Para o desenvolvimento destes experimentos foram considerados 36 fones do português falado no Brasil, mostrados na Tabela 6.2, os quais são associados a diferentes classes fonéticas. Na Tabela 6.1, são listadas as classes fonéticas e os fones que as compõem:

Tabela 6.1: Classes fonéticas com seus respectivos fones.

Classes	Fones
Silêncio (s)	#
Vogais orais (v)	a, e, ɛ, i, j, o, ɔ, u
Vogais nasais (vn)	ɑ̃, ẽ, ĩ, õ, ũ
Consoantes plosivas (p)	p, t, tʃ, k, b, d, dʒ, g
Consoantes fricativas (f)	f, s, ʃ, v, z, ʒ
Consoantes laterais (l)	l, ʎ
Consoantes nasais (n)	n, m, ɲ
Consoantes vibrantes (vb)	r, r̄, R

O dicionário fonético empregado nesta análise foi o adotado por (Ynoguti, C. & Violaro, F 1999), e suas sub-unidades acústicas são apresentadas na Tabela 6.2.

Para visualizar os comportamentos obtidos foram analisados trechos de fala que incluem diferentes classes fonéticas (*v*, *p*, *f* e *n*) de 4 locuções do conjunto de teste. Para cada trecho é estimada a distribuição de singularidades em diferentes escalas de tempo (20ms, 50ms, 100ms, 200ms, 400ms), a fim de examinar as dinâmicas da fala espontânea nas diferentes escalas. As escalas menores cobrem parte do fonema estudado e, portanto, a análise focaliza o comportamento do fonema quase isolado. Nas escalas de tempo maiores, os intervalos de fala estudados incluem tanto o fonema em questão assim como fonemas vizinhos, observando o comportamento da interação entre diferentes fonemas.

Tabela 6.2: Sub-unidades acústicas utilizadas na transcrição fonética das locuções, com exemplos.

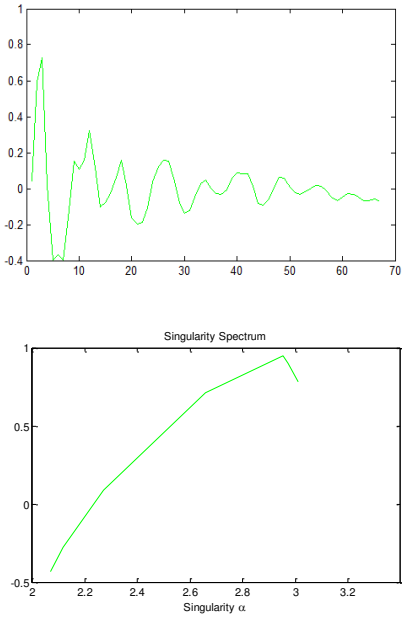
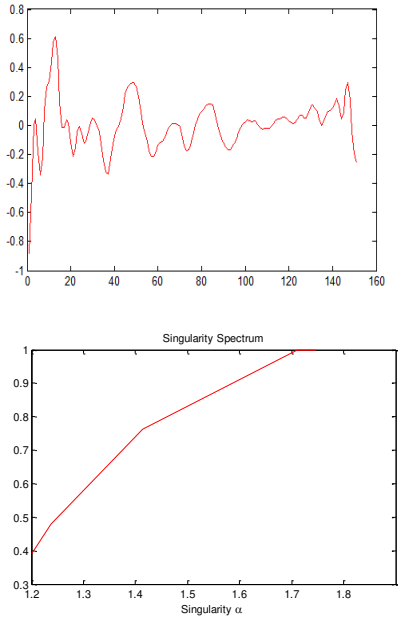
Fone	Símbolo utilizado	Exemplo
#	#	silêncio
a	a	***a*** çafrão
e	e	***e*** levador
ɛ	E	p ***e*** le
i	i	s ***i*** no
j	y	fu ***i***
o	o	b ***o*** lo
ɔ	O	b ***o*** la
u	u	l ***u*** a
ɑ̃	an	maç ***ã***
ẽ	en	s ***en*** ta
ĩ	in	p ***in*** to
õ	on	s ***om*** bra
ũ	un	um
b	b	***b*** ela
d	d	***d*** ádiva
ʤ	D	***d*** iferente
f	f	***f*** eira
g	g	***g*** *orila*
ʒ	j	***j*** iló
k	k	***c*** achoeira
l	l	***l*** eão
ʎ	L	***lh*** ama
m	m	***m*** ontanha
n	n	***n*** évoa
ɲ	N	i ***nh*** ame
p	p	***p*** oente
r	r	ce ***r*** a
r̄	rr	ce ***rr*** ado
R	R	ca ***r*** ta
s	s	***s*** apo
t	t	***t*** empes ***t*** ade
ʧ	T	***t*** igela
v	v	***v*** erão
ʃ		***ch*** ave
z	z	***z*** abumba

Na Tabela 6.3, são listadas as descrições das frases usadas, assim como uma cor associadas a cada uma delas.

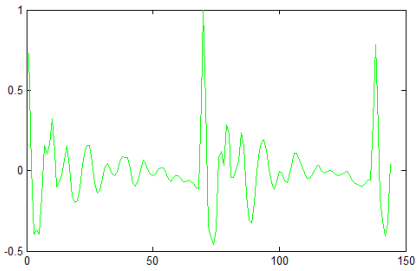
Tabela 6.3: Descrição das frases analisadas.

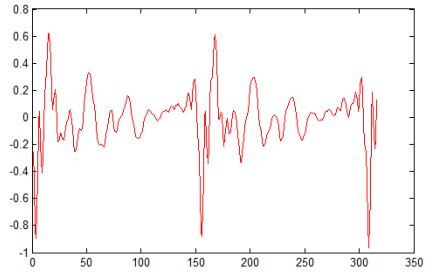
Frase	Gênero	Base de dados	Cor
A vitória foi paga, com muito sangue / *a / v i t O r y a /f o y /p a g a / k on / m u y t o /s an g i /*	Homem	Ynoguti 1	1
A cabine telefônica fica na próxima rua / *a / k a b i n y / t e l e f o n i k a /f i k a /n a / p r O s i m a /rr u a /*	Homem	Ynoguti 1	2
Tudo para incentivar o turismo na região / *t u d u / p a r a / in s en t i v a r / o / t u r i s m u / na / rr e j i an un /*	Homem	Ynoguti 2	3

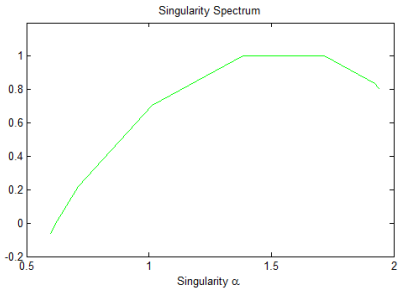
Primeiro será analisada a classe fonética "*Vogais orais*". Para isto são escolhidos alguns fonemas em diferentes condições. Por exemplo, na Figura 6.2, é ilustrada a distribuição dos expoentes de Hölder para a vogal "*a*", tanto isolada quanto no meio de uma palavra.

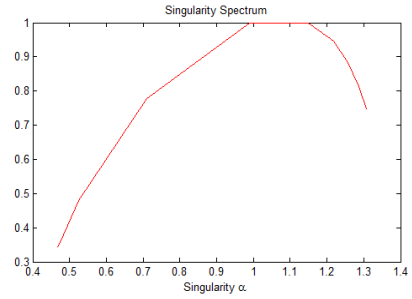
A cabine	Para
a. Duração = 1 período de pitch	
	

b. Duração = 2 período de pitch

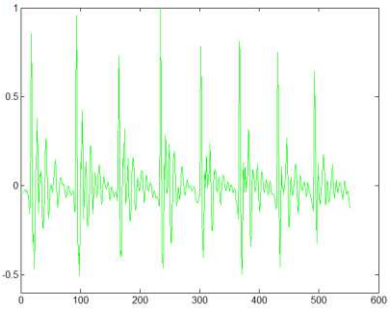


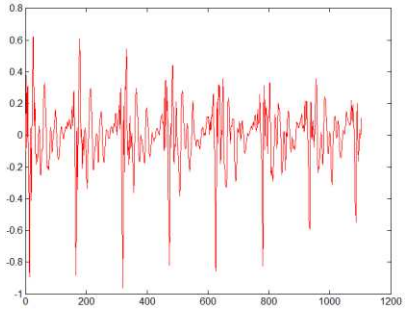


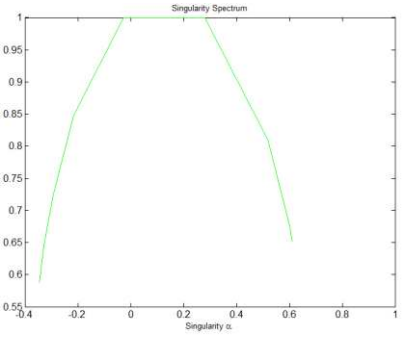


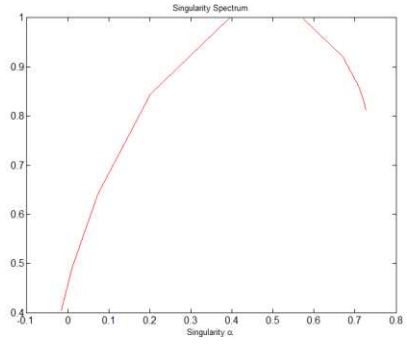


c. Duração = 50ms

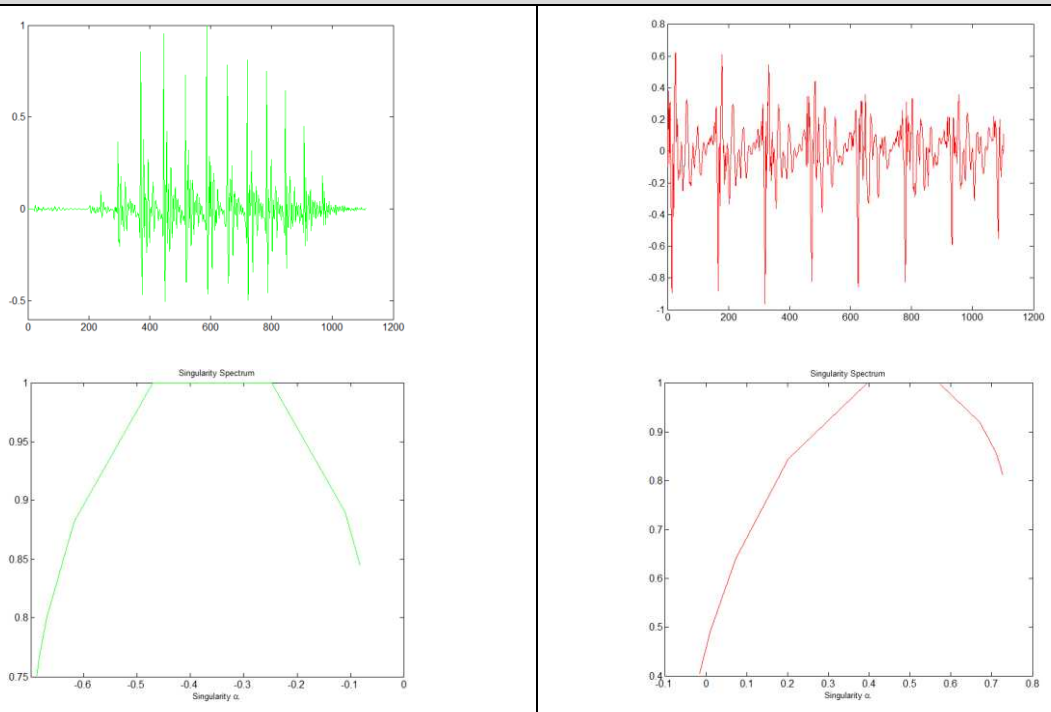








d. Duração = 100ms



e. Duração = 500ms

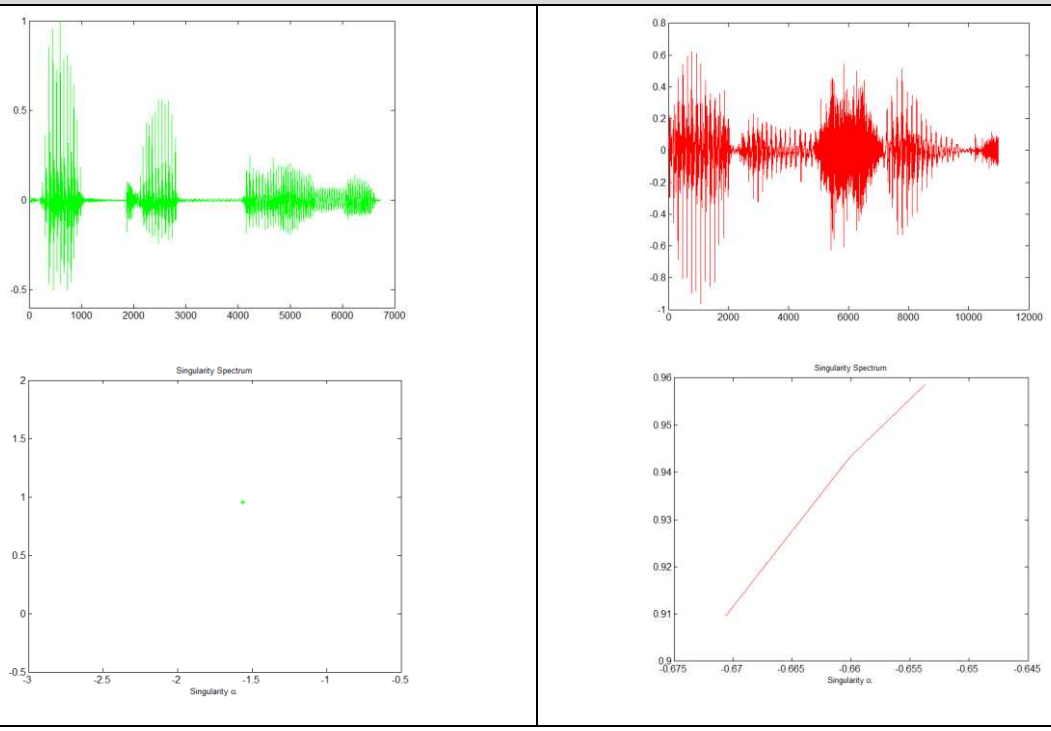


Figura 6.2 – Espectro de Legendre de segmentos das frases 2 e 3, iniciando na vogal "a": a. Um período de pitch do fonema *"a"*, b. dois períodos de pitch do fonema *"a"*, c. 50ms, d. 100ms, e. 500ms.

A Figura 6.2.a mostra o espectro da singularidade para o sinal de fala (vogal *"a"*) com duração de 1 período de pitch. Observa-se que, nesta escala, o sinal de fala tem característica monofractal. Na Figura 6.2.b, considerando uma duração de 2 períodos de pitch, já se nota um comportamento multifractal (curva com concatividade negativa). Este mesmo comportamento persiste até a escala de tempo de 100ms, onde são incluídos vários períodos de pitch, Figura 6.2.d e c. Para escalas maiores o sinal de fala perde o comportamento multifractal; assim, para intervalos em torno de 500ms de duração e superiores, o sinal da fala revela novamente características monofractais, como ilustrado pela Figura 6.2.e. A maioria dos testes efetuados apresentou um comportamento semelhante para segmentos de fala compostos por vogais orais, nasais e algumas consoantes sonoras

A segunda classe fonética estudada foi "*Consoantes plosivas*". Estas consoantes são caracterizadas por terem muita curta duração, pois são resultado da liberação explosiva de um acúmulo de pressão que ocorre quando o trato vocal está fechado em algum ponto por um obstáculo bucal. Além disso, usualmente são seguidas por vogais, o que ocasiona uma mudança não significativa do som da vogal. Portanto, no entorno das plosivas, o comportamento multifractal se mantém de maneira semelhante ao dos fonemas vocálicos, como é registrado na Figura 6.3.

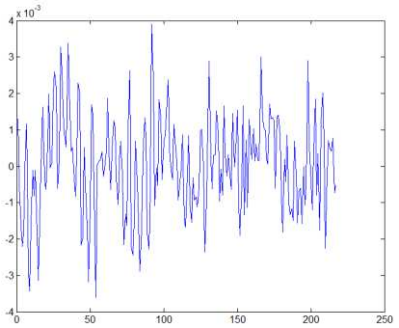
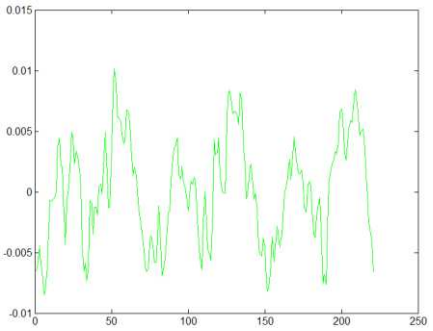
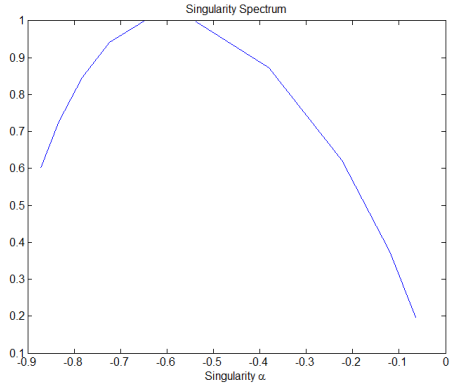
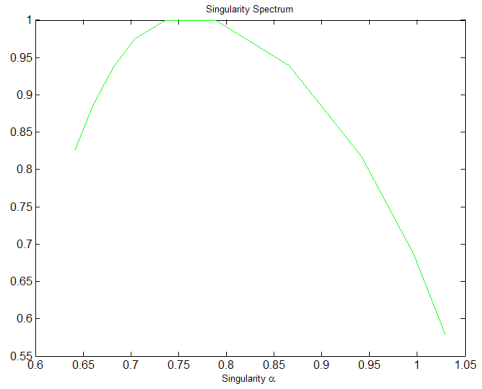
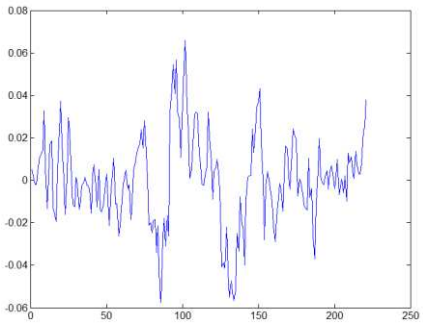
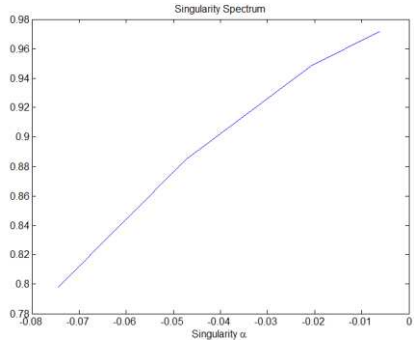
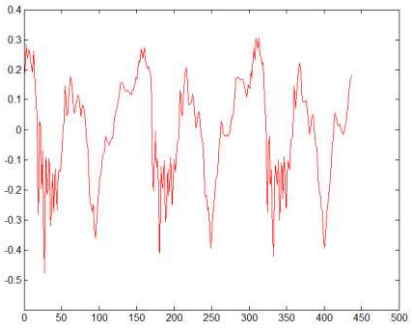
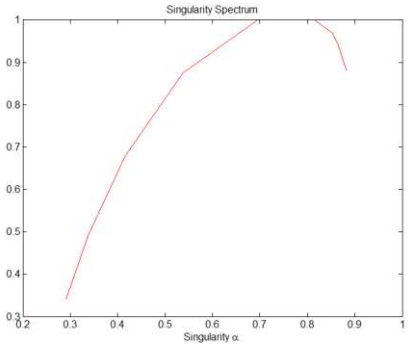
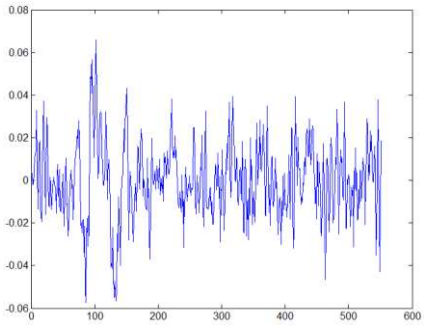
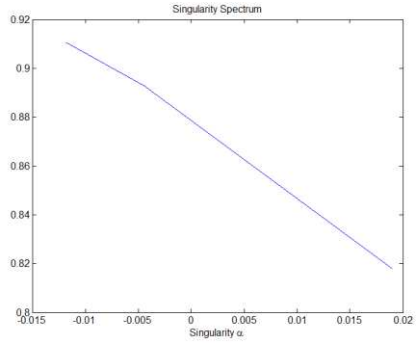
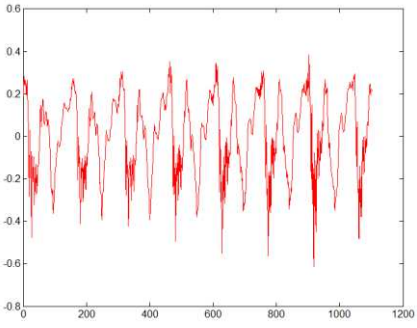
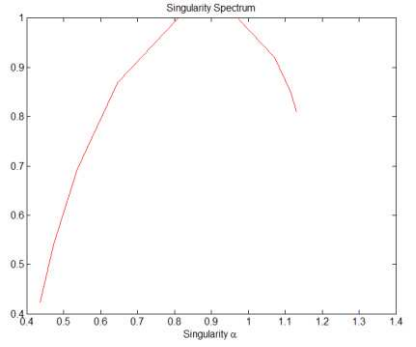
A vitória	A cabine
Duração = 20ms	
	
	

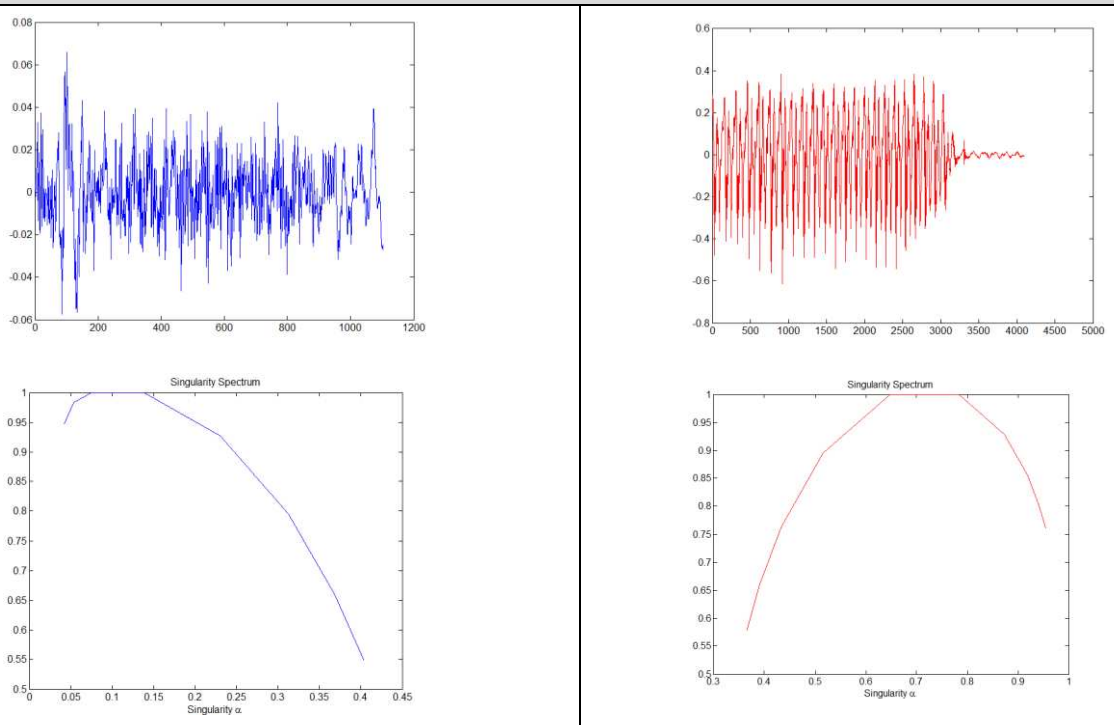
Figura 6.3 – Espectro de Legendre de segmentos das frases 1 e 2, em torno das plosivas *"t"* e *"b"*.

A terceira classe fonética analisada é "*Consoantes fricativas*". Desta classe, são tomados como exemplo os fonemas, "*f*" e "*x*", e ilustrados seus espectros multifractais, para diferentes de escalas de tempo, na Figura 6.4.

As consoantes fricativas são geradas pela turbulência produzida quando o ar dos pulmões é forçado a passar através de uma constrição no trato vocal (Holmes, J. & Holmes, W. 2001). Esta constrição pode ser causada pelos dentes, língua, entre outros. Este fluxo turbulento de ar é chamado de fricção. No caso da análise das consoantes fricativas se apresentaram dois comportamentos diferentes, exemplificados pelos dois fonemas selecionados.

Foi	Pró**x**ima
a. Duração = 20ms	
 	 
b. Duração = 50ms	
 	 

c. Duração = 100ms



d. Duração = 200ms

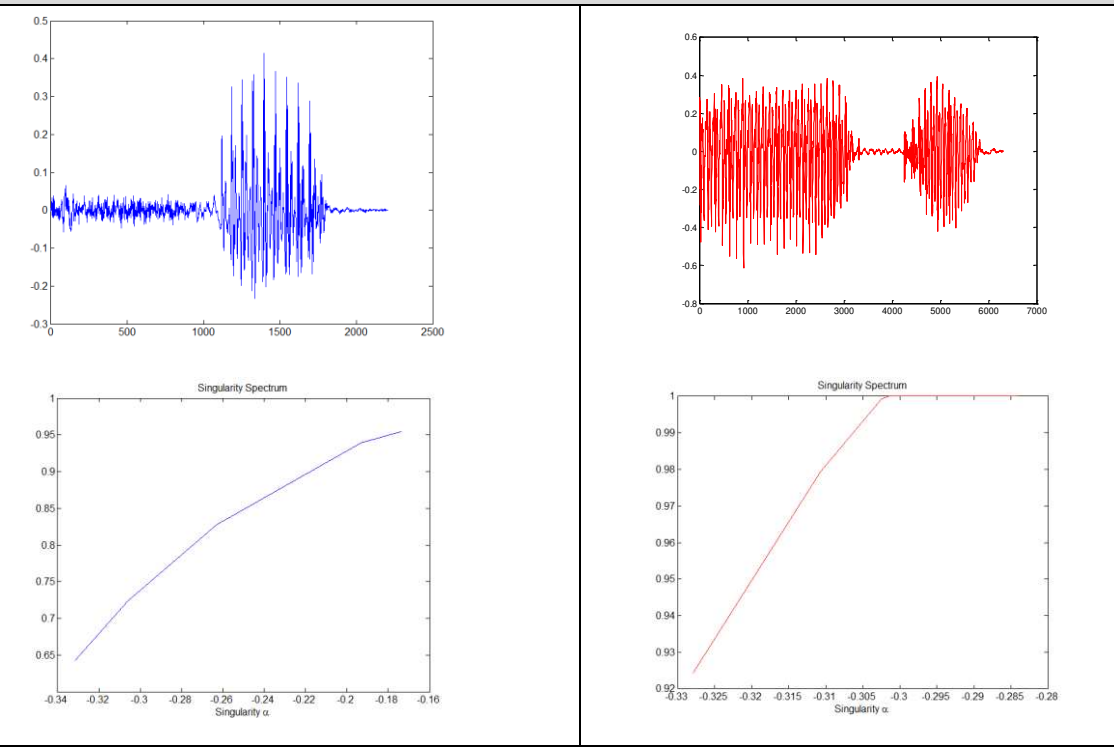


Figura 6.4 – Espectro de Legendre do fonema *"f"* da frase 1 e do fonema *"x"* da frase 3 para as escalas: a. 20ms, b. 50ms, c. 100ms e d. 200ms.

Como é visto na palavra *"foi"*, o fonema "*f* "nas Figura 6.4.a e 6.4.b de cor azul, tem um comportamento similar ao de um sinal aleatório, que normalmente é caracterizado por processos monofractais. Na fala espontânea, esta letra é acompanhada por sons sonoros, precisando deles para alcançar características multifractais em escalas próximas a 100ms, como é observado na Figura 6.4.c. A maioria de fricativas têm comportamentos similares a este.

O seguinte caso é visualizado analisando o fonema "*x*". Segundo a transição fonética este fonema tem som de *s*. Na língua portuguesa, quando o *s* está localizado entre duas vogais, é produzida uma pequena vibração das cordas vocais, como no caso de sons sonoros. Como se pode reconhecer nas Figura 6.4.a e 6.4.b de cor vermelha, para escalas de tempo de 20ms e 50ms, existe periodicidade que faz com que o sinal possua um comportamento semelhante ao encontrado nos sons vocálicos, exibindo características multifractal em escalas inferiores a 100ms.

De forma geral, todos os segmentos de fala estudados mostraram comportamento monofractal em grandes escalas (Figura 6.4.d). Outra conduta encontrada na comparação das classes fonéticas é a variedade de comportamentos para escalas menores, apresentando características multifractais em alguns casos e em outros não. Este fato pode ser verificado na comparação da análise do fonema *b* na Figura 6.3 com o fonema *f n*a Figura 6.4.a. O fonema *b* apresenta comportamento multifractal desde escalas de tempo pequenas, mas o fonema *f* só o apresenta para escalas superiores aos 50ms. Por conseguinte, foi determinado que, para sinais de fala espontânea, a gama de escalas entre os 50ms e os 100ms garante o comportamento multifractal.

6.2 Deslocamento versus Retificação

Na descrição do modelo multifractal VVGM no Capitulo 3, foi estipulada a condição de que o sinal a ser modelado necessariamente deveria ser positivo. Dado que o sinal de fala apresenta amplitudes positivas e negativas, faz-se necessário efetuar um tratamento para que todas as amostras sejam positivas sem perda de informação. Nesta seção, é realizada uma

análise experimental de dois possíveis métodos que satisfazem essa condição: deslocamento e retificação.

Tendo por referência os resultados obtidos nas seções anteriores sobre o comportamento multifractal para sinais de fala, foram realizados vários testes alterando os sinais por meio dos métodos propostos e fazendo a análise em diferentes escalas de tempo, a fim de avaliar as mudanças no comportamento fractal produzidas em relação aos sinais originais. Baseando-se no *Teste Experimental 2*, sinais de fala escolhidos aleatoriamente foram modificados e verificadas suas distribuições de singularidade (expoente Hölder) através da transformada de Legendre para diferentes escalas de tempo.

Após realizar diversos testes com diferentes sinais de fala, foi observado, para a maioria dos casos, que os sinais de fala submetidos a processo de retificação apresentam um comportamento multifractal mais acentuado para uma gama mais ampla de escalas de análise do que os sinais deslocados. Com objetivo de acompanhar melhor os resultados obtidos, na Figura 6.5, é apresentado um exemplo gráfico das distribuições de singularidade de um trecho de sinal de fala original, escolhido aleatoriamente, e dos sinais modificados, para diferentes escalas (10ms, 20ms, 50ms, 100ms, 200ms, 400ms). Este gráfico serve de comparação entre os métodos de correção.

A análise multifractal do sinal de fala para escalas entre 50ms e 100ms é ilustrada nas Figura 6.5.c e 6.5.d. Em concordância com o estudo apresentado na seção anterior, no qual foi concluído que, de forma geral para esta faixa de escalas, os sinais de fala têm comportamento multifractal, tanto o sinal deslocado quanto o retificado mostram espectros multifractais com concavidade negativa semelhantes ao original. Para as escalas menores (Figura 6.5.a e 6.5.b) e maiores (Figura 6.5.e, e 6.5f), pode-se observar que o sinal retificado mantém as propriedades multifractais para uma gama de escalas mais ampla que o sinal deslocado e, até mesmo, que o sinal original, como é caso das escalas 10ms e 200ms. Este fenômeno pode acontecer, já que, ao retificar-se o sinal, são introduzidos transientes de alta frequência e, portanto, enfatizadas as singularidades, havendo necessidade de um maior número de expoentes de Hölder para sua caracterização. No caso do sistema de identificação desenvolvido, foram empregados sinais de fala retificados, já que, para a combinação dos parâmetros característicos de fala (MFCCs e

VVGM) tanto pelo método da ponderação da probabilidade a posteriori (janelas de 100ms) quanto no método de fusão em um só vetor (janelas de 20ms ou 30ms), foi obtido um melhor desempenho.

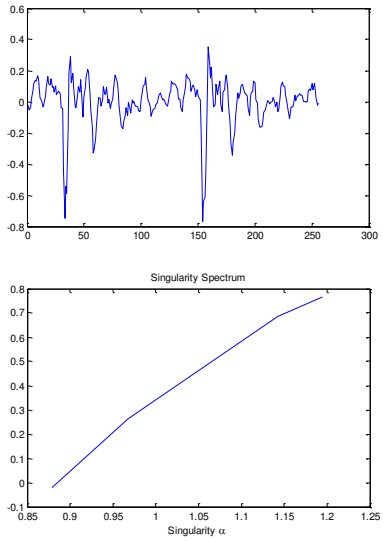
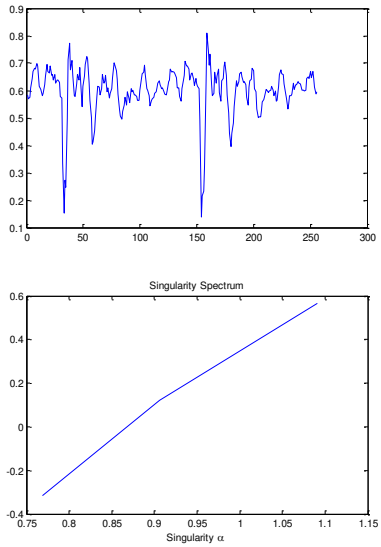
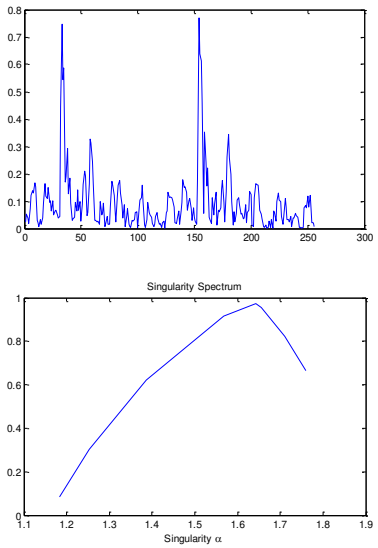
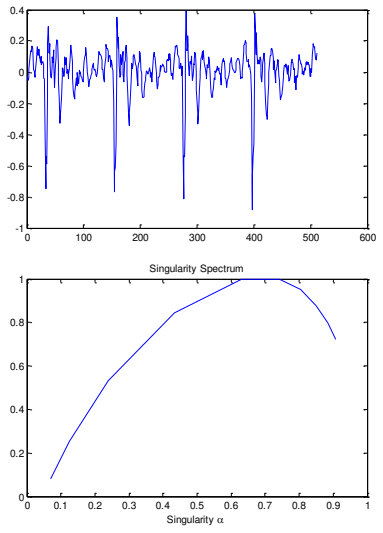
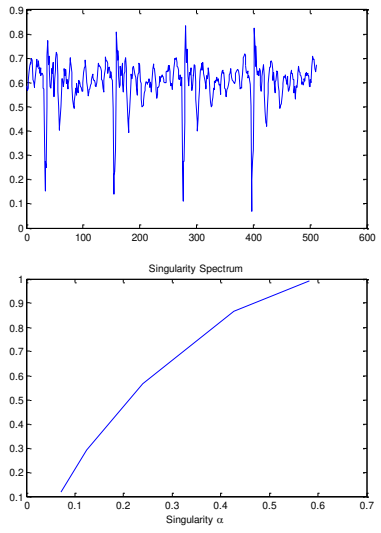
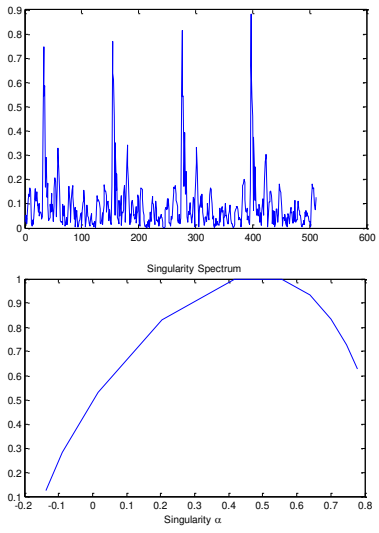
Sinal Original	Sinal Deslocado	Sinal Retificado
a. Duração = 10ms		
		
a. Duração = 20ms		
		

Figura. 6.5 – Comparação do espectro multifractal de trecho de sinal de fala entre a faixa de escalas a. 10ms e b. 20ms.

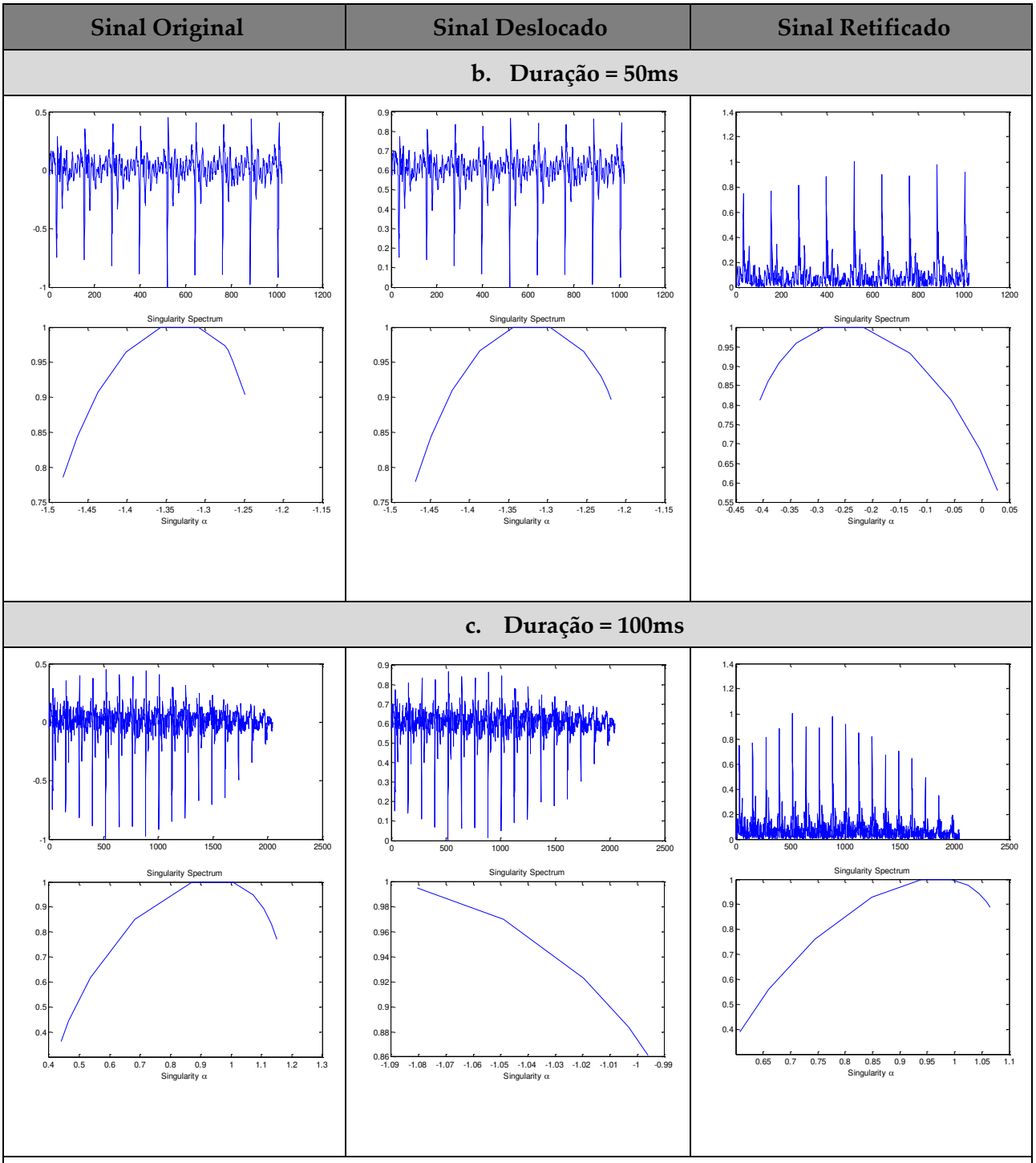


Figura. 6.5 – Comparação do espectro multifractal de trecho de sinal de fala, entre a faixa de escalas c. 50ms e d. 100ms.

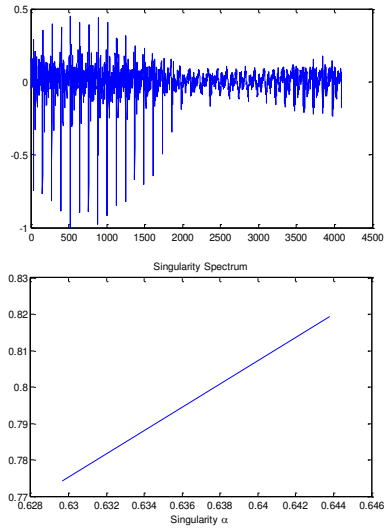
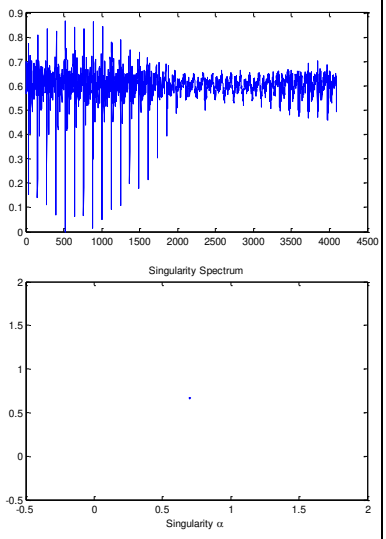
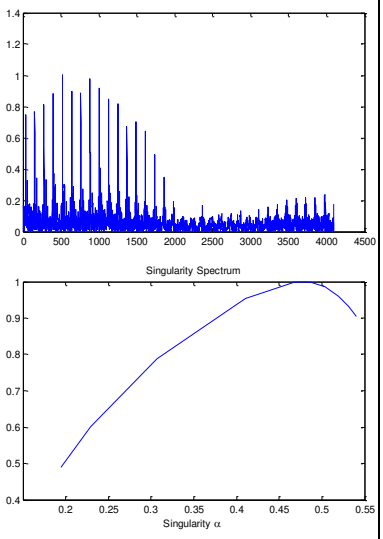
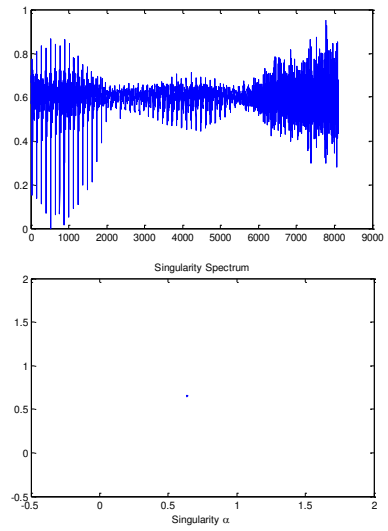
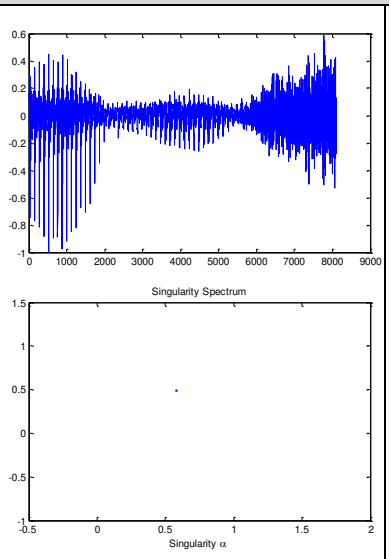
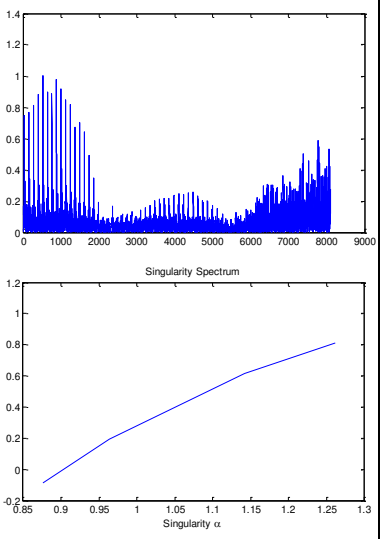
Sinal Original	Sinal Deslocado	Sinal Retificado
d. Duração = 200ms		
		
e. Duração = 400ms		
		

Figura 6.5 – Comparação do espectro multifractal de trecho de sinal de fala entre a faixa de escalas e. 200ms e f. 400ms.

.

7 Teste e Análise de Resultados

Neste capítulo, são descritos os resultados de avaliação dos sistemas desenvolvidos utilizando as bases de dados expostas no Capítulo 5. O estudo experimental está dividido em três conjuntos de testes. No primeiro, é discutido o desempenho do sistema ASI empregando os parâmetros MFCCs e VVGM's individualmente. No segundo, os dois sistemas são fundidos no nível de pontuação descritos na seção 5.2. Finalmente, é avaliado o método de fusão no nível de características apresentado na seção 5.3.

Antes de apresentar os referidos testes, são feitas algumas especificações necessárias para a avaliação do sistema de identificação. A motivação e o procedimento são descritos na seção 7.1.

7.1 Avaliação do Desempenho

7.1.1 Locuções de Teste

O sistema de identificação foi avaliado com duas estruturas de locuções de teste. Com a primeira estrutura, o sistema trabalha com frases completas, as quais possuem comprimentos diferentes. Através deste enfoque, foram examinadas as melhores configurações dos parâmetros envolvidos no sistema, tais como comprimento da janela, ordem do modelo, entre outros, em relação a carga computacional versus taxas de reconhecimento do sistema. A segunda estrutura avalia o sistema de identificação utilizando locuções de teste com diferentes comprimentos (1 s, 2 s, 3 s). As locuções são segmentadas nos comprimentos específicos através do procedimento descrito a seguir. Denotando os vetores de características da locução completa como $\{\vec{x}_1, \ldots, \vec{x}_t\}$,

são empregados T vetores consecutivos extraídos das locuções. Este valor depende do deslocamento da janela do processamento. Assim, para um teste de 1 s e uma locução processada com vetor de parâmetros calculados cada 10ms, T será igual a 100. Conhecendo o número de vetores que representa o tempo escolhido, são selecionados vários segmentos para serem avaliados. O primeiro segmento é composto pelos vetores $\vec{x}_1$ a $\vec{x}_T$. O segundo segmento é composto pelos vetores $\vec{x}_n$ a $\vec{x}_{n+T-1}$. O terceiro segmento é composto pelos vetores $\vec{x}_{2n}$ a $\vec{x}_{2n+T-1}$, e assim por diante.

Os experimentos foram realizados empregando $n = 25$ (deslocamentos de 250ms), valor escolhido para experimentar locuções com diferenças consideráveis entre elas, enquanto aproveitando o maior número de segmentos para avaliar.

Em ambas as estruturas, trechos de silêncio são eliminados, sendo considerados só os vetores de características que possuem informação relevante.

7.1.2 CÁLCULO DE DESEMPENHO

Para o cálculo de desempenho do sistema de identificação, cada locutor treinado é testado com um número de locuções definido pela base de dados usada. Cada identificação correta do sistema é contabilizada.

Este processo é feito com todos os locutores. Ao fina do processo, tem-se a soma de todos os acertos do sistema. A porcentagem de identificação do sistema é determinada usando-se a seguinte relação:

$$\%\ identificação\ = \frac{\#\ Locuções\ identificadas}{\#\ Locuções\ testadas}\ x\ 100 \qquad (6.1)$$

Cada locutor tem, aproximadamente, um igual número de locuções de teste. Portanto, a avaliação do desempenho do sistema pode ser vista como não-tendenciosa (“unbiased”).

7.2 Características dos Parâmetros VVGM

Baseando-se no procedimento de extração dos parâmetros através do VVGM discutido na seção 5.1.2, é determinada a ordem do vetor de características usado pelo sistema de identificação. Esta ordem é determinada dependendo da frequência de amostragem e do comprimento da janela de processamento. Por exemplo, um intervalo de 100ms de uma locução com frêquencia de amostragem 22,05 kHz pode gerar uma cascata de 11 níveis, definindo um vetor de características de dimensão 9. Na Tabela 7.1, apresenta-se a descrição dos parâmetros VVGM dos diferentes experimentos desenvolvidos, para cada uma das bases de dados testadas.

Tabela 7.1: Descrição dos parâmetros VVGM, para as três bases empregadas.

Primeira Base de Dados "*Ynoguti 1*" Frequência de amostragem: 11,025 kHz			
Janela (ms)	Amostras	# níveis da cascata *N*	Parâmetros VVGM *N-2*
100	1102	10	8
30	330	8	6
20	220	8	6

a.

Segunda Base de Dados "*Ynoguti 2*" Frequência de amostragem: 22,05 kHz			
Janela (ms)	Amostras	# níveis da cascata N	Parâmetros VVGM N-2
100	2205	11	9
30	661	9	7
20	445	9	7

b.

Terceira Base de Dados "*corpus ELSDSR*" Frequência de amostragem: 16 kHz			
Janelas (ms)	Amostras	# níveis da cascata N	Parâmetros VVGM N-2
100	1600	10	8
30	480	8	7
20	320	8	7

c.

Deve ser lembrado que o módulo de extração de parâmetros VVGM trabalha com um número de amostras 2^N, onde N é o número de estágios possíveis da cascata. Assim, para janelas de 30 e 100ms, é utilizado o máximo número de amostras que satisfaz esta condição. Por exemplo, ao analisar a base de dados 1 com janelas de 100ms (Tabela 7.1.a), são usadas apenas 1024 amostras com $N = 10$, mesmo que cada janela possua 1102 amostras. Para o caso de análise com janelas de 20ms, por serem intervalos muito curtos, é preciso usar um N superior ao obtido pelo número de amostras da janela, esticando com amostras da janela seguinte até o N mais próximo que satisfaça a condição. Observando a base 1, uma janela de 20ms está composta por 220 amostras: logo, serão acrescentadas amostras da janela seguinte até se atingir o total de 256 amostras, com $N = 8$.

7.3 Primeiro Conjunto de Testes: Sistema Usando Parâmetros MFCCs e VVGM Individualmente

Para a formação do subconjunto de testes, foram utilizadas as três bases de dados apresentadas na seção 5.5: "*Ynoguti 1*", "*Ynoguti 2*" e "*Elsdsr*", compostas por 30, 71, e 22 locutores, respectivamente. Os primeiros testes utilizaram frases completas e individuais com comprimentos diferentes. Para as primeiras duas bases de dados, foram empregadas as 10 locuções de teste de cada locutor originalmente gravadas. Para avaliar a terceira base de dados, foram segmentadas manualmente as duas locuções de teste originalmente gravadas por cada locutor. Desta segmentação, foram geradas entre 5 e 9 locuções por pessoa, com comprimentos na faixa entre 2,0 e 3,5 s.

O treinamento para o sistema baseado nos parâmetros VVGM usou janelas de 100ms com deslocamentos a cada 10ms, e a dimensão do vetor de características é definida na Tabela 7.1 para cada base de dados. Já o treinamento do sistema de parâmetros MFCCs usou janelas de 20ms com deslocamentos de 10ms, e vetor de características de dimensão 12. Além disso, foram experimentadas misturas com diferentes números de gaussianas, no entanto, a melhor configuração em termos de tempo de processamento e desempenho foi obtida com 8 gaussianas.

Para um maior número de gaussianas, o tempo de processamento aumentou consideravelmente, enquanto a taxa de reconhecimento não mostrou melhoria significativa. Na

Tabela 7.2, têm-se os resultados deste primeiro conjunto de testes.

Tabela 7.2: Taxa de reconhecimento (%) dos sistemas de identificação baseados em VVGM e MFCCs.

Base de dado	**Duração das locuções de treinamento Aprox.**	**Ordem do modelo**	**VVGM**	**MFCCs**
Ynoguti 1	60s	6	71,03 %	98,70 %
		8	75,80 %	99,30 %
Ynoguti 2	70s	6	88,90 %	99,01 %
		8	91,30 %	99,57 %
Elsdsr	83s	6	63,20 %	95,90 %
		8	70,01 %	97,30 %

7.4 Segundo Conjunto de Testes: Sistema de Identificação Empregando Fusão no Nível de Pontuação dos Sistemas VVGM e MFCCs

Este conjunto de testes avalia o desempenho do sistema de identificação apresentado na seção 5.4.1, no qual os sistemas ASI são desenvolvidos de maneira independente para cada tipo de parâmetro (VVGM e MFCCs) e misturados no nível de pontuação ao ponderar a probabilidade *a posteriori* de cada locutor, baseando-se no esquema da Figura 5.1.

Uma vez estabelecida uma referência da configuração para o sistema de identificação através do primeiro conjunto de testes, são escolhidos os sistemas com configurações que alcançaram as melhores respostas para serem misturados. O resultado da taxa de reconhecimento deste método é apresentado na Tabela 7.3. Para as três bases de dados, a melhor

taxa foi obtida com um peso de ponderação de 0,6 para parâmetros MFCCs e 0,4 para os parâmetros VVGM.

Tabela 7.3: Taxa de reconhecimento (%) do sistema de identificação misturando as probabilidades a posteriori dos sistemas que empregam parâmetros VVGM e MFCCs.

Base de dados	MFCCs +VVGM	MFCCs	VVGM
Ynoguti 1	99,70 %	99,30 %	75,80 %
Ynoguti 2	99,89 %	99,57 %	91,30 %
Elsdsr	99,01 %	97,30 %	70,01 %

7.5 Terceiro Conjunto de Testes: Sistema de Identificação Empregando Fusão no Nível de Características.

Este conjunto de testes é fundamentado no sistema de identificação descrito na seção 5.4.2. Neste sistema, são acoplados os parâmetros MFCCs e VVGM em um só vetor de características, sendo efectuada uma única modelagem através de GMM para cada locutor, conforme indicado no diagrama da Figura 5.2. Estes testes envolvem as três bases de dados.

O desempenho do sistema foi avaliado inicialmente através da abordagem com locuções de frases completas. Além disso, foram experimentados dois tamanhos de janela para o processamento: 20 e 30ms (usadas no módulo de extração de características). O tamanho do vetor de características obtido desta fusão está sujeito ao comprimento desta janela, pois, como se apresenta na Tabela 7.1, cada base gera um número diferente de parâmetros VVGM. Por exemplo, ao analisar uma locução da primeira base usando janelas de 30ms, são gerados 6 parâmetros. Estes, concatenados aos 12 parâmetros MFCCs, geram um vetor de dimensão 18. A Tabela 7.4 expõe as taxas de reconhecimento obtidas pela nova configuração. Apenas as configurações com melhores taxas foram apresentadas.

Tabela 7.4: Taxa de reconhecimento (%) do sistema de identificação, fundindo os parâmetros VVGM e MFCCs.

Base de fala	Janelas	MFCCs	VVGM	Ordem do modelo	MFCCs+VVGM
Ynoguti 1	20	99,30 %	62,70 %	10	100,00 %
	30	99,30 %	62,70 %	10	100,00 %
Ynoguti 2	20	99,57 %	81,90 %	9	99,85 %
	30	98,59 %	84,78 %	8	100,00 %
Elsdsr	20	95,90 %	60,89 %	8	97,30 %
	30	95,30 %	63,08 %	8	98,65 %

Na Tabela 7.4, pode-se observar que o fato de empregar comprimentos de 30ms degrada o desempenho do sistema de reconhecimento baseado em parâmetros MFCCs, mas permite que o sistema VVGM consiga representar melhor o locutor, pois determina as variâncias dos multiplicadores de forma mais precisa. Assim, o sistema combinado alcança taxas de reconhecimento mais altas.

Depois de estabelecer os melhores parâmetros de funcionamento do sistema, foi realizado um último conjunto de experimentos enfatizado na avaliação dos comprimentos das locuções de teste. As três bases de dados são avaliadas com locuções de 1, 2 e 3 s de duração. Como listado na Tabela 7.4, ao serem usados janelas de 30ms são registradas as melhores taxas de reconhecimento. Por isso, este comprimento é adotado nos atuais testes. Os resultados para este procedimento são mostrados na Tabela 7.5

Tabela 7.5: Taxa de reconhecimento (%) do sistema de identificação, combinando os parâmetros VVGM e MFCCs com locuções de teste de diferentes durações.

Base de fala	Ordem do modelo	Duração das Locuções de Teste	MFCCs	MFCCs + VVGM	Número de testes
Ynoguti 1	10	1 s	87,47 %	90,64 %	1876
		2 s	97,50 %	98,75 %	1082
		3 s	98,73 %	99,71 %	395
Ynoguti 2	8	1 s	90,60 %	94,37 %	7891
		2 s	96,31 %	98,89 %	6939
		3 s	98,76 %	99,43 %	4129
Elsdsr	8	1 s	92,76 %	94.30 %	815
		2 s	94,95 %	96,87 %	615
		3 s	98,40 %	99,31 %	300

7.5.1 EXPERIMENTO COM BASE RUIDOSA

Adicionalmente aos testes apresentados, foi realizado um experimento preliminar empregando uma base de fala gravada através de telefone fixo, a fim de se ter uma perspectiva do comportamento do sistema de reconhecimento com a introdução de ruído de canal. A base usada foi *"BaseIME"*, desenvolvida pelo "Departamento de Engenharia Elétrica de Instituto Militar de Engenharia (IME)". Esta base é composta por 75 pessoas (50 homens e 25 mulheres), locuções com frequência de amostragem de 8 kHz e com codificação linear com 8 bits/amostra. O material de treinamento e teste de cada locutor tem, em média, 140 s de duração.

Neste experimento, foi calculada a taxa de reconhecimento obtida pelo sistema com parâmetros MFCCs e com o sistema baseado no método de fusão (MFCCs + VVGM) com janelas de 30ms de comprimento. Para o primeiro, foi empregado um vetor de características de dimensão 12, e, para o segundo, um vetor de dimensão 18 (12 parâmetros MFCCs + 6

parâmetros VVGM). Este experimento foi avaliado com locuções de teste com duração de 5 s. Na Tabela 7.6, são apresentadas as taxas de reconhecimento obtidas pelos sistemas.

Foi observado que as locuções do **locutor 14** apresentaram erros de gravação, gerando uma grande quantidade de erros de reconhecimento. Ao excluir as locuções de teste deste locutor o sistema aumentou significativamente a taxa de reconhecimento (Tabela 7.6).

Tabela 7.6: Taxa de reconhecimento (%) do sistema de identificação, combinando os parâmetros VVGM MFCCs

Base de fala	Ordem modelo GMM	Duração Locuções Teste	MFCCs	MFCCs + VVGM	Número de testes
BaseIME	9	5 s	96.73 %	97.27 %	35041
BaseIME Excluído Locutor 14	9	5 s	97.96 %	98, 30 %	34601

7.6 Análise dos Resultados

Neste capítulo, foram apresentados os testes de avaliação dos sistemas implementados, utilizando as bases de dados descritas no Capítulo 5. Os seguintes itens foram analisados baseando-se nos resultados obtidos dos testes:

- Desempenho do sistema de identificação baseado nos parâmetros VVGM;
- Influência do comprimento da janela de processamento no desempenho dos parâmetros VVGM;
- Avaliação de sistema com adição de ruído de canal e comparação com sistemas de reconhecimento atuais.
- Desempenho final do sistema.

A seguir, cada um destes itens será analisado com maiores detalhes.

7.6.1 DESEMPENHO DO SISTEMA DE IDENTIFICAÇÃO BASEADO NOS PARÂMETROS VVGM

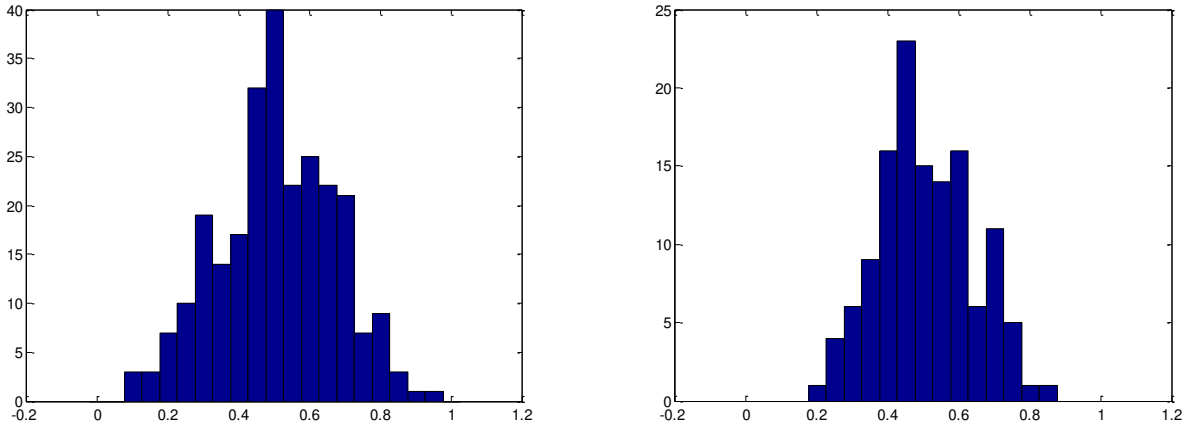
Estes testes iniciais mostraram taxas de reconhecimento relativamente elevadas, com porcentagens aproximadas entre 70 e 90% de acerto para as três bases experimentadas. Com isso, foi possível corroborar a idéia de modelar sinais de fala como processos multifractais. Assim, por meio da análise das variâncias dos multiplicadores da cascata multiplicativa, representa-se a distribuição da medida μ apresentada no Capítulo 2, a qual mostra a forma do crescimento do processo multifractal.

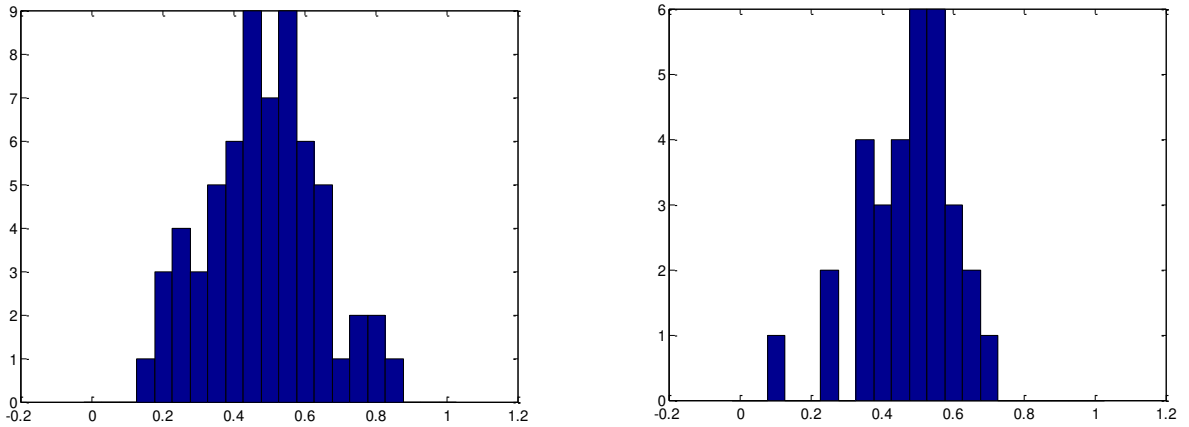
Além disso, uma vez que o modelo VVGM faz uma análise focada nas partes não-estacionárias do sinal, é possível distinguir a importância e a quantidade de informação da identidade de cada locutor presente nestas áreas.

7.6.2 INFLUÊNCIA DO COMPRIMENTO DA JANELA DE PROCESSAMENTO NO DESEMPENHO DOS PARÂMETROS VVGM

No Capítulo 5, foi mencionado que o comprimento das janelas retangulares adequado ao procedimento de extração dos parâmetros VVGM era 100ms. Isso foi verificado aplicando este tamanho de janela no sistema de identificação e obtendo as melhores taxas de reconhecimento, como pode ser visto na Tabela 7.2, em relação as taxas alcançadas com o uso de outros comprimentos de janelas menores, listadas na Tabela 7.4.

Dado que os parâmetros VVGM representam a análise da variância dos multiplicadores nos estágios da cascata gerada para um trecho de fala, ao considerar janelas de 20ms ou 30ms, têm-se muito menos amostras em comparação com uma janela de 100ms. Isto implica que a cascata gerada em trechos curtos vai ter menos estágios, assim como variâncias estimadas de forma menos precisa. Este fenômeno pode ser observado na Figura 7.1, onde são ilustrados os histogramas dos estágios 2 e 3 para uma locução da primeira base de dados. Nas Figura 7.1.a e 7.1.b, é analisado um trecho de 100ms de duração, e se observa que a distribuição dos multiplicadores tende a ser gaussiana. Esta distribuição se degenera significativamente nas Figura 7.1.c e 7.1.d. para 30ms e 7.1.e e 7.1.f. para 20ms.





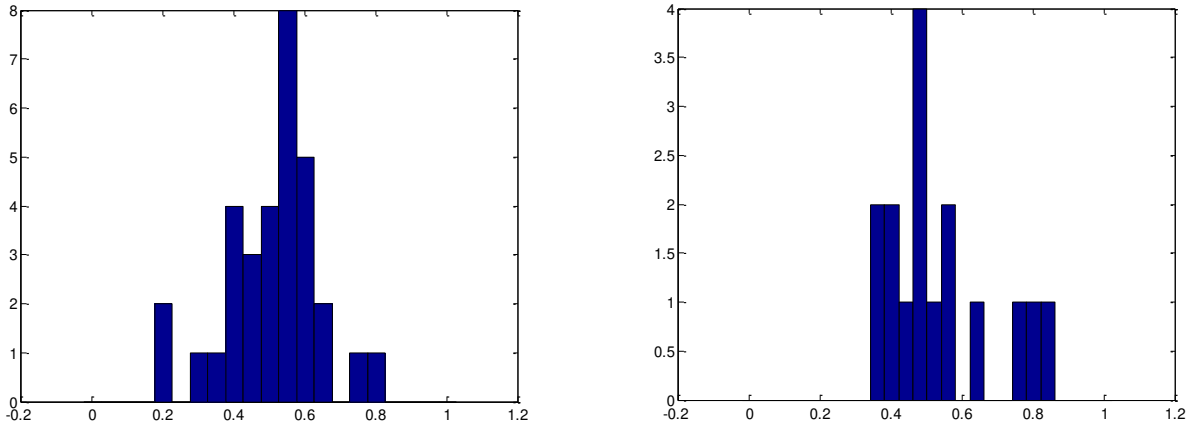


Figura 7.1 – Histogramas dos multiplicadores para os estágios 2 e 3 de uma locução da primeira base de dados: a. Estágio 2 para janela de 100ms; b. Estágio 3 para janela de 100ms; c. Estágio 2 para janela de 30ms; d. Estágio 3 para janela de 30ms; e. Estágio 2 para janela de 20ms; f. Estágio 3 para janela de 20ms.

Apesar de que dessa degeneração não pode ser corrigida, a fusão dos dois parâmetros em um só vetor representa grandes vantagens nas taxas de reconhecimento e na robustez do sistema, confirmando a hipótese previamente apresentada. Em resumo, para estes casos, sugere-se a implementação de um sistema com o mesmo tamanho de janelas para o processamento dos dois parâmetros VVGM e MFCCs. Analisando a Tabela 7.4, verifica-se que esta configuração obteve melhores resultados que a fusão no nível de pontuação da Tabela 7.3. Outro evento que mostrou a superioridade do método de fusão no nível de características foi que, em várias ocasiões, os dois subsistemas isolados (método de fusão no nível de pontuação), escolhem erradamente um mesmo locutor, impossibilitando sua correção.

7.6.3 AVALIAÇÃO DE SISTEMA COM RUÍDO DE CANAL E COMPARAÇÃO COM SISTEMAS DE RECONHECIMENTO ATUAIS.

Embora, esta avaliação não seja tão aprofundada (Tabela 7.6), foi possível observar que o comportamento do sistema foi coerente com os experimentos realizados usando as três bases de fala gravadas através de microfone, apresentadas no Capitulo 5, e as taxas de reconhecimento obtidas foram superiores para o sistema com parâmetros combinados.

A base *"BaseIME"* foi empregada para a avaliação do sistema de reconhecimento de locutor implementado por (Sant'Ana, R., Coelho, R. & Alcaim, A. 2006), citado no Capítulo de Introdução. Embora, na referida pesquisa, não seja assumido que o sinal de fala é fractal, usam-se fundamentos fractais tanto para a extração de parâmetros quanto para o classificador. O sistema implementado na referência acima também sugere a combinação dos parâmetros tradicionais MFCCs com seu vetor de parâmetros proposto Hurst e obteve uma taxa de reconhecimento de 97,46% para locuções de teste de comprimento 5 s, empregando 15 parâmetros MFCCs e 97,66% combinando (MFCCs+HURST). No sistema acima também foi usado GMM como classificador.

Não foi possível realizar uma comparação direta do sistema apresentado acima com o proposto nesta dissertação, uma vez que a referência não fornece o número de gaussianas empregado para a modelagem e, portanto, não é possível replicar a configuração do sistema

com os parâmetros MFCCs. Assim, as taxas de reconhecimento obtidas em cada sistema são diferentes.

Adicionalmente, observou-se que a base foi gravada com PCM linear a 8 kHz com 8 bits/amostra (elevado ruído de quantização), além de apresentar picos de amplitude elevada, que interferem com o sistema proposto no processo de normalização e adequação do sinal. Mesmo nestas condições, as taxas de reconhecimento de ambos os sistema são muito próximas. Embora o sistema proposto não supere as taxas de reconhecimento do sistema citado, ao excluir as locuções da pessoa 14 o sistema com VVGM+MFCCs obteve taxas superiores como se observa na Tabela 7.6. Uma vantagem do sistema proposto nesta dissertação, em relação ao outro, é a baixa carga computacional, devido ao uso de funções básicas aritméticas (VVGM) enquanto que o sistema citado emprega funções wavelet (HURST).

7.6.4 DESEMPENHO FINAL DO SISTEMA

Em todos os testes, foi observado um incremento no desempenho do sistema de identificação ao misturar os parâmetros.

- Comparando os resultados dos primeiro conjunto de testes (parâmetros individuais), mostrados na seção 6.3, com os resultados do segundo conjunto de testes finais (fusão no nível de pontuação), descritos na seção 6.4, pode-se verificar que a taxa de reconhecimento subiu 0,40% para a primeira base, 0,32% para a segunda base e 1,70% para a terceira base.
- Comparando os resultados dos primeiro conjunto de testes (parâmetros individuais), mostrados na seção 6.3, com os resultados do terceiro conjunto de testes finais (fusão no nível de características), descritos na seção 6.5, pode-se verificar que a taxa de reconhecimento subiu 0,70% para a primeira base, 1,42% para a segunda base e 3,33% para a terceira base.
- Analisando a Tabela 7.5, pode-se observar que, para todos os comprimentos testados, o sistema fusionado apresentou taxas de reconhecimento mais altas. Tendo em vista que, o teste foi realizado com uma ampla quantidade de locuções, esta diferença de porcentagem de acerto representa um incremento importante no desempenho.

8 Conclusões

Neste trabalho, foi proposto e implementado um sistema de reconhecimento de locutor independente de texto, baseado em misturas de gaussianas, que utiliza como vetor de características os parâmetros multifractais VVGM e os parâmetros clássicos MFCCs. O sistema desenvolvido tem uma estrutura composta pelos módulos de extração de parâmetros, treinamento e reconhecimento. A partir desta estrutura, são implementados dois sistemas novos que combinam os dois parâmetros de características. No primeiro sistema, é efetuada ponderação da probabilidade *a posteriori* obtida a partir de cada parâmetro individualmente no módulo de reconhecimento. No segundo, é feita a fusão dos parâmetros em um só vetor no módulo de extração de parâmetros. Foram realizados diferentes testes de avaliação para cada sistema, empregando três bases de fala: *"Ynoguti 1"*, *"Ynoguti 2"* e *"Elsdsr"*. Destes testes foi possível concluir:

- Ao se avaliar o sistema ASR empregando unicamente os parâmetros VVGM, foram obtidas taxas de reconhecimento de 75,80%, 91,30% e 70,01% nas bases *"Ynoguti 1"*, *"Ynoguti 2"* e *"Elsdsr"* respectivamente. Embora estas taxas não sejam tão elevadas quanto as obtidas pelos sistemas com MFCCs, são suficientes para mostrar que estes parâmetros possuem informação relevante da identidade do locutor que pode ser usada como informação complementar.
- Foi observado um melhor desempenho nos sistemas com fusão dos parâmetros VVGM e MFCCs que no sistema baseado unicamente nos parâmetros MFCCs. Assim, para as bases de dados *"Ynoguti 1"*, *"Ynoguti 2"* e *"Elsdsr"*, a taxa de reconhecimento aumentou 0,4% , 0,33% e 1,71% respectivamente para o método de fusão no nível de pontuação de probabilidades e 0,7%, 1,43% e 3,35%

respectivamente para o método de fusão no nível de características. Esses aumentos da taxa de reconhecimento mostram também a superioridade em relação ao método de ponderação.

Por outro lado, dado que o modelo multifractal VVGM é baseado em cascatas multiplicativas conservativas, é necessário trabalhar com sinais positivos. Para cumprir este requisito, foram experimentados dois métodos de adequação: deslocamento e retificação. A implementação final foi desenvolvida usando retificação, pois, com este método, foi obtido um melhor desempenho do sistema. Quando são usados sinais deslocados, os intervalos de baixo nível concentram os valores dos multiplicadores próximos a ½, influenciando sua variância. Por outro lado, o fato de retificar o sinal intensifica as singularidades presentes no sinal de fala, ao introduzir mudanças de alta frequência, o que pode ser aproveitado para uma melhor caracterização por meio de processos multifractais.

Além do sistema de reconhecimento de locutor, neste trabalho, foi feita uma análise das características multifractais em sinas de fala das três diferentes bases de dados. A partir de extensos testes e avaliações, concluiu-se que os sinais de fala podem apresentar comportamentos monofractal ou multifractal, dependendo do tipo de fonema considerado e da escala de tempo empregada. Assim:

- Os resultados experimentais mostram que alguns fonemas, tais como algumas fricativas, têm usualmente um comportamento monofractal, enquanto os fonemas vocálicos apresentam um comportamento multifractal. Este análise é feita sob resolução de escalas de tempo menores que tentam cobrir o fonema quase isolado (10s, 20ms, 30ms). Por esta razão, não é possível estabelecer um comportamento fractal único para estas escalas.
- De forma geral, quando os intervalos de análise têm duração superior aos 200ms, o comportamento tende a ser monofractal (pontos e linhas retas).
- Embora não seja possível definir uma fronteira rígida de separação entre as escalas de tempo nas quais o sinal apresente sempre um comportamento multifractal, observou-se que intervalos de fala com duração entre 50ms e 100ms

revelam um comportamento multifractal de forma geral. Nestas escalas, normalmente os fonemas aparecem ou isolados (vogais) ou combinados (plosivas e vogais ou fricativas e vogais), o que garante um comportamento multifractal.

Trabalhos futuros

- Validação dos sistemas propostos em situações adversas, como presença de ruído telefônico.
- Validação dos sistemas empregando menos material de treinamento, o que pode ser interessante para aplicações comerciais.
- Experimentar o sistema com outros classificadores alternativos, tais como redes neurais, SVM, entre outros.
- Abordagem de outras aplicações de processamento de fala tais como segmentação de fala, onde a análise das transições pode proporcionar informação útil e complementar as abordagens tradicionais.

9 Referências Bibliográficas

Barnsley, M.F. *Fractals Everywhere.* 2nd ed. Boston: Academic Press, 1993.

Bund, A. & Havlin, S. *Fractals and Disordered Systems.* 2nd ed. Cambridge: Cambridge University Press, 2000.

Campbell, J. "Speaker Recognition: A Tutorial." *in Proc. IEEE* Vol. 85, no. 9 (September 1997): 1437-1462.

Cirigliano, R. J. da R. "Identificação de Locutor: Otimização do Número de Componentes." Dissertação de Mestrado, UFRJ, Rio de Janeiro, 2007, 61.

de Lima, MIP. "Multifractals and the Temporal Structure of Rainfall." Ph.D Thesis, Wageningen Agricultural Univ, 1998, 229 pp.

Dempster, A. P., Laird, N. M. & Rubin, D. B. "Maximum Likelihood from Incomplete Data Via the EM Algorithm." *Journal of the Royal Statistical Society* Vol. 39, no. 1 (1977): 1-38.

Devroye, L. "The Double Kernel Method in Density Estimation." *In Anais do Instituto Henri Poincaré* Vol. 25 (1989): 533–580.

Elizalde, C. E. & Torre, D. *Reconocimiento de Locutor Dependiente de Texto Mediante Adaptación de Modelos Ocultos de Markov fonéticos.* Proyecto fin de carrera, Madrid: Universidad Autónoma de Madrid, 2007, 89.

Falconer, J. K. *Fractal Geometry: Mathematical Foundations and Applications.* 2nd ed. Chichester: John Wiley & Sons, 2003.

Feder, J. *Fractals.* New York and London: Plenum Press, 1988.

Feng, L. & Hansen, L. K. "A New Database for Speaker Recognition." *IMM-Technical Report*, 2005.

Gao, J ., Cao, Y., Hu, J. & Tung, W. *Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond.* New Jersey: A John Wiley & Sons, 2007.

García, A. P. M., Jiménez, F. J. & Ayuso, J. L. "Análisis Multifractal de Series de Datos Pluviométricos en Andalucía." Tesis Doctoral, Universidad de Córdoba, Córdoba, 2007, 163.

Gupta, V. & Waymire, E. "A Statistical Analysis of Mesoscale Rainfall as a Random Cascade." *Journal of Applied Meteorology* Vol.32 (February 1993): 251–267.

Harte, D. *Multifractals : Theory and Applications.* Boca Raton: Chapman & Hall/CRC, 2001.

Holmes, J. & Holmes, W. *Speech Synthesis and Recognition.* 2nd ed. London: Tayor & Francis, 2001.

Ivanov, P. Ch. *Long-Range Dependence in Heartbeat Dynamics.* Vol. Vol.621, in *Processes with Long-Range Correlations: Theory and Applications*, by G. Rangarajan and M. Ding, 339-372. Berlin: Springer, 2003.

Kinnunen, T. & Li, Haizhou. "An Overview of Text-Independent Speaker Recognition: From Fatures to Supervectors." *ScienceDirect* Speech Communication , no. 52 (2010): 12-40.

Kinsner, W. & Grieder, W. "Speech Segmentation Using Multifractal Measures and Amplification of Signal Features." *in Proc. 7th IEEE International Conference on Cognitive Informatics*, October 2008: 351-357.

Kolmogorov, A. N. "A Refinement of Previous Hypotheses Concerning the Local Structure of Turbulence in a Viscous Incompressible Fluid at High Reynolds Number." *Journal of Fluid Mechanics* (Cambridge University Press) Vol. 13, no. 1 (Decembro 1962): 82–85.

Krishna, M. P., Gadre, V. M., & Dessay, U. B. *Multifractal Based Network Traffic Modeling.* Bombay: Kluwer Academic Publishers, 2003.

Langi, A. & Kinsner, W. "Consonant Characterization Using Correlation Fractal Dimension for Speech Recognition." *in Proc. IEEE Western Canada Conference on Communications, Computer, and Power in the Modern Environment* Vol. 1 (May 1995): 208-213.

Langit, A. Z. R., Soemintapurat, K. & Kinsners, W. "Multifractal Processing of Speech Signals." *in Proc. IEEE International Conference on Information, Communications and Signal Processing* Vol. 1 (September 1997): 527-531.

Mandelbrot, B. B. *Fractals and Scaling in Finance.* New York: Springer, 1997.

Mandelbrot, B. *The Fractal Geometry of Nature.* New York: WH Freeman, 1982.

McLachlan, G. & Peel, D. *Mixture Models.* New York: John Wiley & Sons, Inc, 2000.

Petry, A. & Barone, D. A. C. "Fractal Dimension Applied to Speaker Identification." *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* Vol. 1 (May 2001): 405-408.

Picone, J. "Signal Modeling Techniques In Speech Recognition." *in Proc. IEEE* Vol. 81, no. 9 (June 1993): 1215 - 1247.

Quatieri, T. F. *Discrete- Time Speech Signal Processing Principles and Practice.* New Jersey: Prentice Hall PTR, 2001.

Rabiner, L. R. & Schafer R. W. *Introduction to Digital Speech Processing.* Boston-Delf: Now Publishers Inc., 2007.

Reynolds, D. & Rose, R. C. "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification." *IEEE Transactions on Speech and Audio Processing* Vol. 3, no. 1 (January 1992): 72-83.

Reynolds, D. A. "Experimental Evaluation of Features for Robust Speaker Identification." *IEEE Transactions on Speeach and Audio Processing* Vol.2, no. 4 (October 1994): 639-643.

Reynolds, D. "An Overview of Automatic Speaker Recognition Technology." *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* Vol. 4 (May 2002): 4072-4075.

Riedi, R. H. & Véhel, J. L. "Tcp Traffic is Multifractal: a Numerical Study." Technical Report 3129, INRIA Research report, 1997.

Riedi, R. H. "Introduction to Multifractals." Chap. 28 in *Long Range Dependence : Theory and Applications*, by eds. Doukhan, Oppenheim and Taqqu, 625-715. Birkhäuser: RICE University Technical Report, 2002.

Riedi, R. H., Crouse, M. S., Ribeiro, V. J. & Baraniuk, R. G. "A Multifractal Wavelet Model with Application to Network Traffic." *IEEE Transactions on Information Theory* Vol. 45, no. 3 (April 1999): 992-1018.

Sant'Ana, R., Coelho, R. & Alcaim, A. "Text-Independent Speaker Recognition Based on the Hurst Parameter and the Multidimensional Fractional Brownian Motion Model." *IEEE Transactions on Audio, Speech and Language Processing* Vol. 14, no. 3 (May 2006): 931-940.

Scarborough, J.B. *Numerical Mathematical Analysis.* 5nd ed. Boston: Johns Hopkins Press, 1966.

Stanley, H.E. "Powerlaws and Universality." *Nature 378*, 1995.

Stênico, J. W. e Lee, L. L. "Estimação da Probabilidade de Perda e um Esquema de Controle de Admissão para Tráfego Multifractal de Redes." Dissertação de Mestrado, FEEC, UNCAMP, Campinas, 2009.

Veith, D. & Abry, P. "A Wavelet-Based Joint Estimator of the Parameters." *IEEE Transactions on Information Theory* Vol. 45, no. 3 (Mar 1998): 878–897.

Vicsek, T. *Fractal Growth Phenomenon.* 2nd ed. Singapore: World Scientific Pub Co Inc, 1993.

Vieira, F.H.T. & Lee L.L. "Contribuições ao Cálculo de Banda e de Probabilidade de Perda para Tráfego Multifractal de Redes." Tese de Doutorado, UNICAMP, Campinas, 2006.

Volkmann, J., Stevens, S. & Newman, E. "A Scale for the Measurement of the Psychological Magnitude Pitch." *The Journal of the Acoustical Society of America* Vol. 8, no. 3 (January 1937): 185-190.

Vuuren, V. S. "Speaker Verification in a Time-Feature Space." Ph.D Thesis, Oregon Graduate Institute of Science and Technology, Pretoria, 1999.

Wang, L. & Geng, X. *Behavioral Biometrics for Human Identification: Intelligent Applications.* Hershey-New York: Medical Information Science Reference, 2009.

Ynoguti, C. & Violaro, F. "Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov." Tese de Doutorado, FEEC, UNICAMP, Campinas, 1999.

Ynoguti, C. A. & Violaro, F. "A Brazilian Portuguese Speech Database-DVD." *XXVI Simpósio Brasileiro de Telecomunicações .* Rio de Janeiro , 2008.

Zhou, Y., Wang, J. & Zhang, X. "Research on Speaker Recognition Based on Multifractal Spectrum Feature." *Second International Conference on Computer Modeling and Simulation*, January 2010: 463-466.