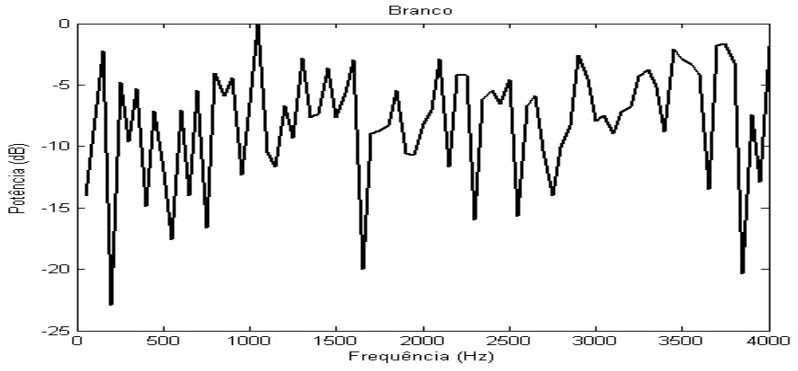
2
Reconhecimento Automático de Locutor em Ambientes Ruidosos

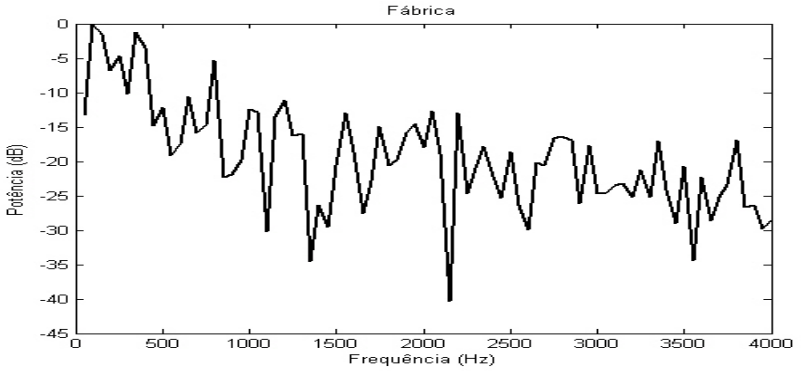
Neste capítulo são apresentadas, de forma resumida, as técnicas de classificação e de extração de atributos mais usadas recentemente para aplicações de reconhecimento de locutor, visando permitir um melhor entendimento dos próximos capítulos.

Em muitas aplicações os sistemas de reconhecimento de locutor têm o seu desempenho, em termos da taxa de acerto, afetado pelo ruído no sinal de fala. Os sons em ambientes de fábrica, de carros e até de várias pessoas falando ao mesmo tempo são fontes de ruído que prejudicam o reconhecimento. A obtenção de elevadas taxas de acerto requer primeiramente que os atributos sejam uma boa representação das propriedades da fala mais associadas à identificação do locutor. Nesse contexto, na próxima seção, são apresentadas técnicas de extração de atributos sendo que as mais eficientes, em termos de melhoria na taxa de acerto, são aquelas que baseiam-se no comportamento auditivo humano. Além disso, um sistema de reconhecimento eficaz, em ambientes ruidosos, requer que o efeito da contribuição do ruído seja minimizado. Para essa finalidade, as técnicas mais robustas de extração de atributos possuem estratégias que buscam eliminar a contribuição dos fatores que pioram a taxa de acerto, como os ruídos.

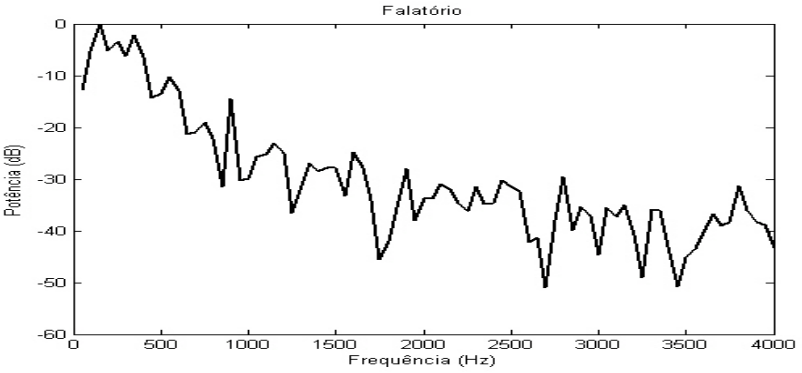
Nessa tese são mostrados experimentos de identificação de locutor na presença de ruído branco e de ruído não-branco (ou colorido). O ruído branco caracteriza-se por possuir densidade espectral de potência plana em ampla faixa do espectro, como mostrado na Figura 4(a). Já os ruídos não-brancos têm suas contribuições mais concentradas em certas regiões do espectro. Isso pode ser visto, por exemplo, na Figura 4(b)-(d), para ruídos da base NOISEX-92 [57]. O ruído de fábrica, mostrado nessa Figura, foi coletado próximo a uma máquina cortadora de metais e a um equipamento elétrico de solda, o ruído falatório foi coletado próximo a 100 pessoas falando em uma cantina, e o ruído de carro foi coletado em um veículo Volvo 340 deslocando-se a 120 Km/h, em quarta marcha, numa pista de asfalto em condições de chuva. O ruído branco foi gerado artificialmente, utilizando o Matlab.



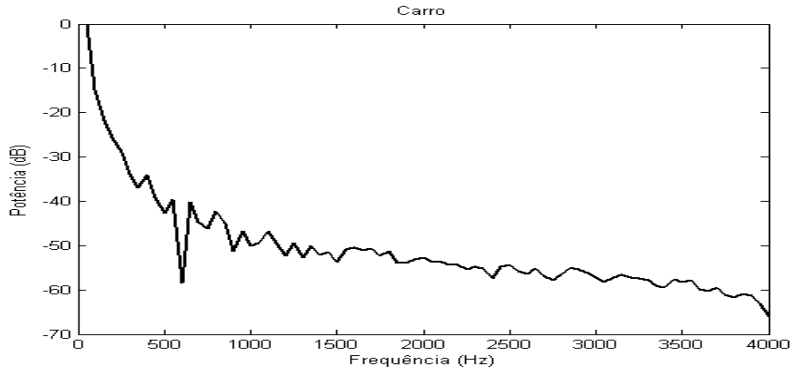
(a)



(b)



(c)



(d)

Figura 4 – Espectro de potência dos ruídos; (a) ruído branco. (b) fábrica, (c) falatório, e (d) carro

2.1
Os Atributos Utilizados

Nesta seção, são apresentadas diversas técnicas de extração de atributos do sinal de voz. Esses atributos são os dados de entrada do classificador usado no reconhecimento, representando as características do locutor.

Quando uma pessoa está falando, o seu aparelho vocal se movimenta de acordo com a constituição física própria de cada indivíduo, produzindo sinais acústicos formados por variações de pressão do ar. Esses sinais são percebidos pelo sistema auditivo de acordo com a sensibilidade humana às características de amplitude e frequência.

Os sinais de voz usados no reconhecimento de locutor, são, primeiramente, coletados por microfones e amostrados. As amostras são divididas em janelas (ou quadros). Um conjunto de atributos (parâmetros, coeficientes) é obtido das amostras de cada janela por meio de um algoritmo de extração. Esse conjunto é organizado sob a forma de um vetor, em que cada componente do vetor representa um valor de atributo.

Muitas das técnicas de extração de atributos são baseadas em modelos de produção da fala e da sensibilidade auditiva humana [5]. A introdução dessas propriedades biológicas no projeto do atributo acrescenta mais informação sobre a identidade do locutor. Basicamente, tais propriedades são caracterizadas pela sensibilidade auditiva em relação à freqüência e à amplitude, como também pelas frequências de formação da voz no aparelho vocal. No espectro do sinal de voz existem regiões de elevada energia que correspondem às frequências de ressonância (formantes) do aparelho vocal [58],[59]. As formantes são as principais frequências de operação desse aparelho, trazendo informação essencial sobre a identidade de uma pessoa.

Uma técnica bem conhecida é a que extrai parâmetros de sinais de voz baseando-se no modelo de predição linear usado para fazer predição de sinais de voz. Esse modelo consiste em representar um sinal por meio de suas amostras passadas e pode ser expresso por

$$\hat{s}_n = -\sum_{k=1}^{p} a_k s_{n-k} \qquad (1)$$

onde s_{n-k} , sendo k de 1 a p, são as amostras passadas e a_k são os parâmetros. Os atributos que são obtidos baseando-se neste princípio, são chamados de parâmetros LPC (*Linear Prediction Coefficients*) [60]-[62] e podem ser usados para gerar um modelo de produção da fala. Esse modelo, ilustrado na Figura 5, é representado por uma função de transferência formada somente por pólos e é uma representação matemática aproximada do aparelho vocal. Nesse sentido, o erro de predição linear pode ser interpretado como uma excitação para este modelo. A partir da expressão do erro de predição dada por

$$e_n = s_n - \hat{s}_n = s_n + \sum_{k=1}^{p} a_k s_{n-k} \,, \tag{2}$$

que na Transformada-Z pode ser expressa por

$$E(z) = S(z)\left[1 + \sum_{k=1}^{p} a_k z^{-k}\right], \tag{3}$$

chega-se à função de transferência que modela o aparelho vocal,

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \,. \tag{4}$$

Basicamente, existem dois métodos para a estimação dos parâmetros LPC: o método das autocorrelações e o método das covariâncias. No primeiro, os parâmetros LPC são obtidos pela função de autocorrelação das amostras do sinal de voz que é calculada assumindo-se que o valor médio quadrático do erro de predição linear E é minimizado, em função de a_k , em - ∞< n < ∞,

$$\frac{\partial E}{\partial a_k} = 0 \tag{5}$$

onde n é a n-ésima amostra. Essa função de autocorrelação, para um sinal de infinitas amostras, é definida por

$$R(i) = \sum_{n=-\infty}^{+\infty} s_n s_{n+i} \tag{6}$$

e das soluções de (5), chamadas de equações de Yule-Walker, obtém-se os parâmetros que são determinados por

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i), 1 \le i \le p. \quad (7)$$

Na prática, o cálculo de (7) baseia-se no conceito de função de autocorrelação de intervalo curto, no qual é realizada uma aproximação aplicando-se uma janela ao sinal amostrado. Fora da janela o sinal é feito igual a zero. No segundo método, assume-se que o valor médio quadrático do erro de predição linear é minimizado em um intervalo finito 0 ≤n≤ *N*-1. Os parâmetros são calculados usando-se a covariância das amostras do sinal presentes neste intervalo. A covariância é expressa por

$$\varphi_{ki} = \sum_{n=0}^{N-1} s_{n-i} s_{n-k} \quad (8)$$

e os parâmetros são determinados por

$$\sum_{k=1}^{p} a_k \varphi_{ki} = -\varphi_{0i}, 1 \le i \le p. \quad (9)$$

Quando o intervalo de *n* tende a infinito, o método da covariância tende ao da autocorrelação.

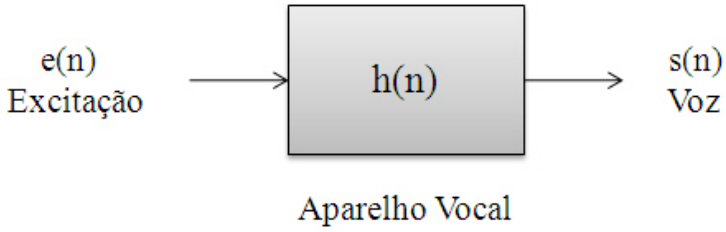


Figura 5 – Ilustração simplificada do modelo do aparelho vocal

O reconhecimento de locutor pode ter o seu desempenho melhorado se for feito um pré-processamento no sinal de voz – de acordo com a percepção auditiva – antes da extração dos parâmetros LPC, que consideram igualmente todas as frequências do espectro. Ressalta-se que, a partir de aproximadamente 800 Hz, a percepção auditiva humana diminui com a freqüência [63]. Os parâmetros obtidos por meio desse pré-processamento são chamados parâmetros PLP (*Perceptual Linear Predictive*) [63]-[65]. Nessa técnica de extração, inicialmente a DFT (*Discrete Fourier Transform*) é aplicada ao sinal. Em seguida, o espectro de potência resultante é mapeado na escala Bark [64],[66]-[68] que é

aproximadamente logarítmica. Essa escala, em termos da frequência angular ω expressa em rad/s, é dada pela função de mapeamento

$$\Omega(\omega) = 6\ln\left\{\omega/1200\pi + \left[\left(\omega/1200\pi\right)^2 + 1\right]^{0.5}\right\}. \tag{10}$$

O espectro de potência P(ω) do sinal, nesta escala, é aplicado em um esquema de curvas

$$\Psi(\Omega) = \begin{cases} 0 & \Rightarrow \quad \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & \Rightarrow \quad -1.3 \le \Omega \le -0.5, \\ 1 & \Rightarrow \quad -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)} & \Rightarrow \quad 0.5 \le \Omega \le 2.5, \\ 0 & \Rightarrow \quad \Omega > 2.5. \end{cases} \tag{11}$$

que modela os filtros auditivos. As componentes resultantes da convolução expressa por

$$\Theta(\Omega) = \sum_{\Omega_1} P(\Omega_1 - \Omega\ )\Psi(\Omega_1), \tag{12}$$

passam por uma pré-ênfase $\in(\omega)$ que modela a sensibilidade auditiva humana em função da frequência,

$$\Xi\left[\Omega(\omega)\right] = \in(\omega)\Theta\left[\Omega(\omega)\right] \tag{13}$$

onde essa pré-ênfase é expressa por

$$\in(\omega) = \left[\left(\omega^2 + \ 56.8\times10^6\right)\omega^4\right] / \left[\left(\omega^2 + 6.3\times10^6\right)^2 \times \left(\omega^2 + 0.38\times10^9\right)\right]. \tag{14}$$

Em seguida, as componentes resultantes da pré-ênfase passam por uma transformação

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \tag{15}$$

que modela a sensibilidade auditiva com relação à potência percebida. Após essa operação, é aplicada a transformada inversa de Fourier. A última etapa consiste na extração dos parâmetros através da análise LPC. O uso desses procedimentos de extração causa uma melhoria no desempenho do reconhecimento porque os parâmetros PLP têm como propriedades intrínsecas as características da sensibilidade auditiva humana que foram incluídas nele durante este pré-processamento.

Em uma outra estratégia de extração de atributos, foi utilizada a representação do espectro do sinal de voz obtido a partir da DFT em uma escala chamada Mel [5],[69]-[72], que modela a sensibilidade auditiva humana. Essa escala assume um comportamento quase linear para frequências abaixo de 1 kHz e logarítmico para frequências acima de 1 kHz. Em frequências próximas a 1 kHz, o valor em Mel é aproximadamente igual ao valor em Hz, isto é: 1 kHz ~1 kMel. Para frequências abaixo de 1 kHz o valor em Mel é maior do que o valor em Hz (por exemplo, 500 Hz equivale a 607 Mel). Para frequências acima de 1 kHz o valor em Mel é menor do que o valor em Hz (por exemplo, 2000 Hz equivale a 1521 Mel). A diferença entre a escala Mel e a escala Hz cresce com o afastamento do ponto de 1 kHz. Os atributos que são obtidos baseando-se nesses conceitos são chamados de Atributos Mel Cepestrais ou MFCC (*Mel Frequency Cepstral Coefficients*) [4],[73]-[75]. Nessa técnica de extração, é usado um banco de filtros triangulares [4],[76] para realizar a decomposição do espectro na escala Mel, como mostrado na Figura 6. O logaritmo é aplicado na saída de cada filtro do banco e em seguida é aplicada a DCT (*Discrete Cosine Transform*) para a obtenção dos atributos MFCC,

$$MFCC_k = \sum_{n=0}^{N-1} \zeta(n) \cos((\mathrm{n} + 1/2)), \quad \mathrm{k}=1,2,\ldots,\mathrm{M}, \quad \mathrm{M}<\mathrm{N} \tag{16}$$

onde os $\zeta(n)$ representam os valores do logaritmo aplicado à saída de cada filtro do banco. A análise Mel representa com maior resolução as componentes de baixas frequências do sinal de voz, em concordância com o fato de que a percepção auditiva é maior nessas frequências.

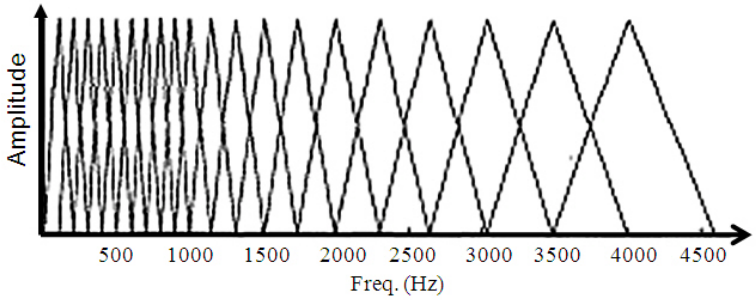


Figura 6 – Banco de filtros na escala Mel

Certos tipos de atributos são mais adequados para aplicações que exijam um menor grau de processamento, como por exemplo, os atributos LFCC (*Linear Frequency Cepstrum Coefficients*) [4], que são obtidos aplicando-se o logaritmo diretamente ao espectro gerado pela DFT sem a decomposição na escala Mel, e os LPCC (*Linear Prediction Cepstrum Coefficients*) [4],[77]-[80], que são obtidos diretamente dos parâmetros LPC por meio de simples equações recursivas expressas por

$$c_1 = a_1$$

$$c_n = \sum_{k=1}^{N-1} (1 - \mathrm{k}/\mathrm{n})\, a_k c_{n-k} + a_n \,,\ 1 < \mathrm{n} \le \mathrm{p}$$

$$c_n = \sum_{k=1}^{N-1} (1 - \mathrm{k}/\mathrm{n})\, a_k c_{n-k} \,,\ \mathrm{n} > \mathrm{p}. \qquad (17)$$

Essas equações fornecem a representação cepestral c_n dos parâmetros LPC, e baseiam-se no modelo da audição humana. Os LFCC, tal como os LPC, consideram igualmente todas as frequências do espectro. Porém, os LFCC fazem uso da representação cepestral, já que aplicam o logaritmo ao espectro gerado pela DFT. Quanto aos LPCC, foi verificado experimentalmente em [4] que esses atributos e também os LPC preservam menos componentes do sinal de voz do que os obtidos usando-se a DFT, como por exemplo, os MFCC e os LFCC. Dois outros tipos de atributos que também podem ser usados para aplicações de

reconhecimento são os delta e os delta-delta [6],[81],[82], que expressam variações dinâmicas no espectro e são mais adequados para o reconhecimento dependente do texto. Os parâmetros delta são representados por

$$\Delta\chi_i = \chi_{i+l} - \chi_{i-l}, \tag{18}$$

e os delta-delta por

$$\Delta\Delta\chi_i = \Delta\chi_{i+L} - \Delta\chi_{i-L}. \tag{19}$$

Onde χ_i é o i-ésimo atributo estático. Por exemplo, em (18) são mostrados atributos delta estimados a partir de outros tipos de atributos, representados por χ, como por exemplo, os MFCC, extraídos das janelas indicadas pelos índices subscritos. Em (19) é mostrado que os atributos delta-delta na janela (*i*) estão sendo estimados a partir dos delta nas janelas (*i*+L) e (*i*-L). Esses tipos de atributos, baseados em diferenças dinâmicas, são robustos às variações provocadas pelo canal de comunicação no sinal de voz, já que removem informações espectrais invariantes no tempo associadas ao canal [6]. Essas técnicas de extração de atributos requerem menores graus de processamento.

A técnica que extrai os atributos chamados de ZCPA (*Zero Crossings with Peak Amplitudes*) [7],[8],[83],[84] baseia-se nas propriedades auditivas e nos cruzamentos por zero efetuados pelo sinal de voz. Essa estratégia considera a amplitude de pico percebida entre dois cruzamentos por zero consecutivos do sinal, durante a subida (passagem de um valor negativo de intensidade do sinal, a um valor positivo). Os valores dessas amplitudes de pico entre os cruzamentos por zero são obtidos na saída de um banco de filtros cocleares [85], que modelam a sensibilidade auditiva com relação à frequência. A estratégia EIH (*Ensemble Interval Histogram*) [83], que modela a audição, usa detectores de cruzamento de nível (usado como uma espécie de referência) na saída de cada filtro do banco para detectar os cruzamentos. Os valores dos níveis (intensidades) são distribuídos em uma escala logarítmica ao longo de todos os valores positivos de intensidade, que o sinal na saída de cada filtro pode assumir. A Figura 7 ilustra esta situação.

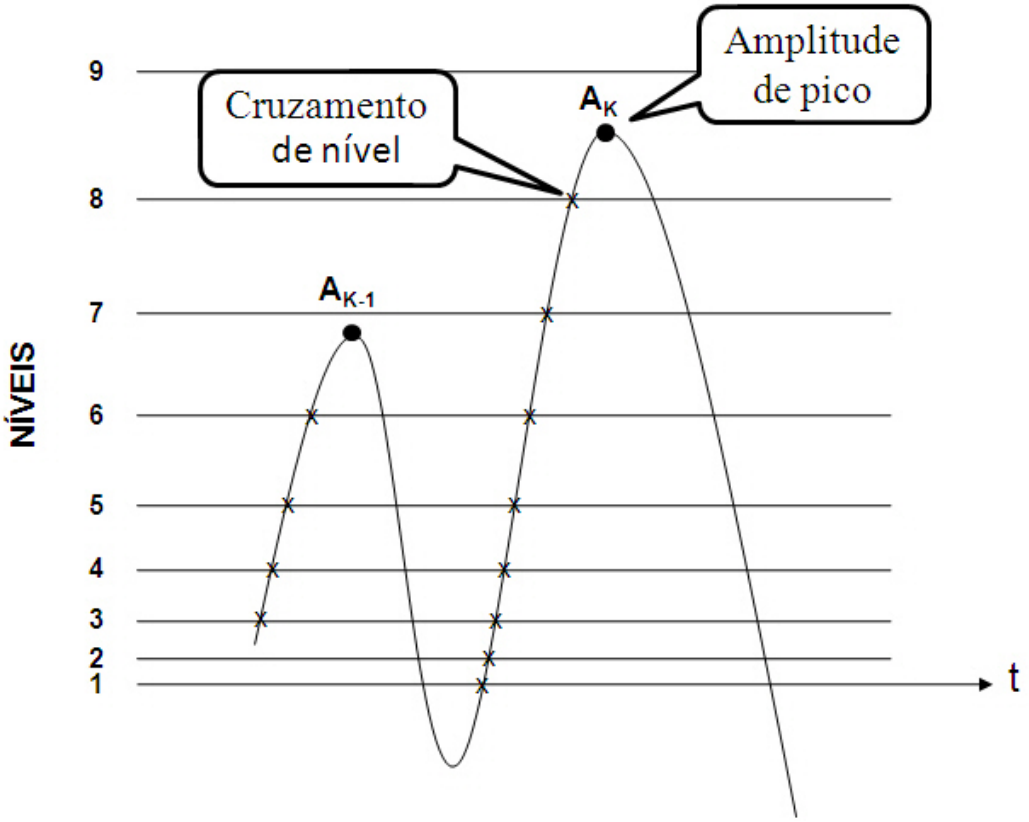


Figura 7 – Cruzamentos de nível

Os valores de frequência utilizados nesse esquema de extração de atributos são estimados calculando-se o inverso do tempo decorrido entre dois cruzamentos por zero consecutivos. Foi escolhido o nível de cruzamento em zero porque a estimação de frequências usando níveis em valores baixos é menos afetada por ruído aditivo do que quando são utilizados níveis em valores mais altos [7]. Nessa técnica de extração, a função logaritmo, que é aplicada nos valores das amplitudes de pico entre dois cruzamentos por zero na saída de cada filtro do banco, é usada para modelar a sensibilidade do nervo auditivo em relação à intensidade do estímulo sonoro. Foi mostrado experimentalmente que estes atributos ZCPA são robustos a ruído, para aplicações de reconhecimento de voz [7]. Os atributos ZCPA podem ser expressos por

$$\gamma(t,i) = \sum_{\varsigma} \sum_{k=1}^{K-1} \delta_{iJ_k} f(A_k), 1 \le i \le \mathrm{N} \qquad (20)$$

onde A_k representa a amplitude de pico na saída de cada filtro do banco, k é o índice que representa o k-ésimo cruzamento por zero, $f(A_k)=\log(1+20A_k)$ representa a função utilizada para modelar a sensibilidade do nervo em relação à

intensidade, N representa o número de componentes de frequência na saída ς de cada filtro, J_k é o índice que representa cada uma dessas componentes e δ é o Delta de Kronecker. O valor γ (t,i) do atributo é a soma das contribuições de cada filtro do banco.

Recentemente, em [8], foi feita a representação cepestral dos atributos ZCPA e utilizada a DCT para reduzir a correlação dos dados. Esses atributos chamados de ZCPAC (*Zero Crossing with Peak Amplitude Cepstrum*) foram usados em reconhecimento robusto de locutor dependente do texto e superaram os MFCC em desempenho de reconhecimento nesta aplicação [8].

O parâmetro de Hurst (pH) [29],[86] foi apresentado em 2006 e tem contribuído significativamente para aumentar a eficiência do reconhecimento automático de locutor. Os resultados apresentados até o momento já mostram a eficácia desta estratégia. Diferentemente dos demais atributos, o pH é uma característica estatística do sinal de voz. O parâmetro de Hurst modela o comportamento estocástico do sinal de voz através da correlação ou dependência temporal entre as amostras do sinal de voz, no infinito. Nessa abordagem de extração de atributos, aplica-se a DWT (*Discrete Wavelet Transform*) às amostras do sinal de voz. A DWT pode ser interpretada como uma filtragem (ou decomposição) em sub-bandas (ou escalas) de frequências. Ou seja, o sinal é passado em um banco de filtros resultando em um conjunto de coeficientes agrupados por sub-bandas de frequências. Esses coeficientes representam as contribuições de cada sub-banda para a formação do sinal. As contribuições de altas frequências são responsáveis pelos detalhes do sinal. Após a aplicação da DWT, para cada escala *j*, calcula-se a variância dos coeficientes wavelet. Em seguida, plota-se os valores do logaritmo na base dois de cada valor de variância *versus j*. No gráfico obtido, faz-se uma regressão linear, como mostrado na Figura 8, e calcula-se a inclinação α da reta gerada. O valor do parâmetro de Hurst (pH) é $h=(1+\alpha)/2$.

Os parâmetros de Hurst, como modelam o comportamento estocástico do sinal de voz, tendem a ser robustos às distorções provocadas pelo canal de comunicação [87],[88]. Alternativamente, os parâmetros de Hurst podem ser obtidos não diretamente das amostras de voz, mas também de outros tipos de atributos, como por exemplo, os parâmetros mel-cepestrais.

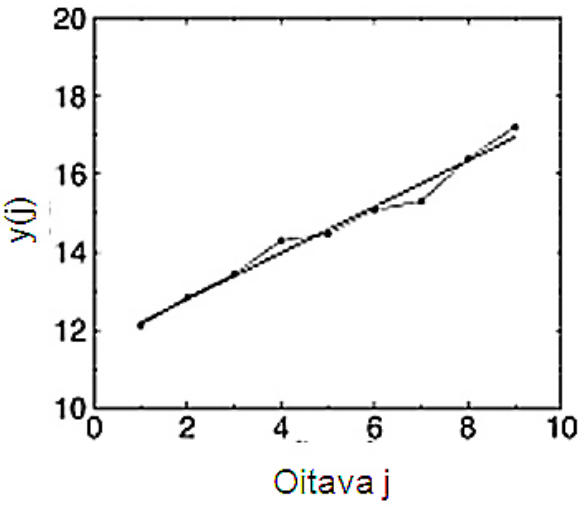


Figura 8 – Regressão usada para obter os parâmetros de Hurst

2.2
Os Classificadores

Esta seção apresenta os principais classificadores usados no reconhecimento automático de locutor. Os métodos de classificação modelam os locutores com base na matriz de atributos da voz.

A etapa de classificação é realizada após a extração dos atributos do sinal de voz. O classificador, geralmente, faz uma modelagem probabilística da distribuição dos atributos associados aos locutores. Basicamente, consideram-se duas fases de operação do classificador: a fase de treinamento e a fase de teste. Na primeira, o classificador recebe a matriz de atributos e gera o modelo probabilístico (histogramas) relacionado a cada locutor. Na segunda, o sistema de classificação recebe os atributos da voz de um pretenso locutor e, em seguida, faz uma comparação, usando o modelo armazenado, para reconhecer, ou não, a pessoa associada aos atributos recebidos. A decisão baseia-se em uma medida de distância (verossimilhança) que é calculada utilizando as estatísticas associadas aos atributos recebidos no teste e obtidas do modelo probabilístico relacionado ao locutor. No caso da verificação de locutor, também é usado um modelo só de locutores falsos chamado modelo de *background* (*Universal Background Model,*

UBM) [6],[29],[89]. Nessa aplicação, o locutor só é reconhecido (verificado) quando a diferença entre a sua verossimilhança associada (em relação ao seu modelo) e a associada aos locutores falsos (*background*) é maior do que um determinado limiar de aceitação que é determinado subjetivamente (experimentalmente). O uso desse modelo de locutores falsos, considerado como uma espécie de referência, facilita a escolha do limiar de decisão. A Figura 9 ilustra o esquema de operação do sistema de classificação quando, por exemplo, é desejado verificar o locutor José. Nessa figura, o sistema recebe os parâmetros e apresenta na saída o resultado da decisão. A decisão na verificação de locutor é um critério binário: ou é verificado, ou não.

A Figura 10 ilustra a identificação do locutor José, em um grupo de pessoas formado pelo próprio José, Pedro e outros locutores.

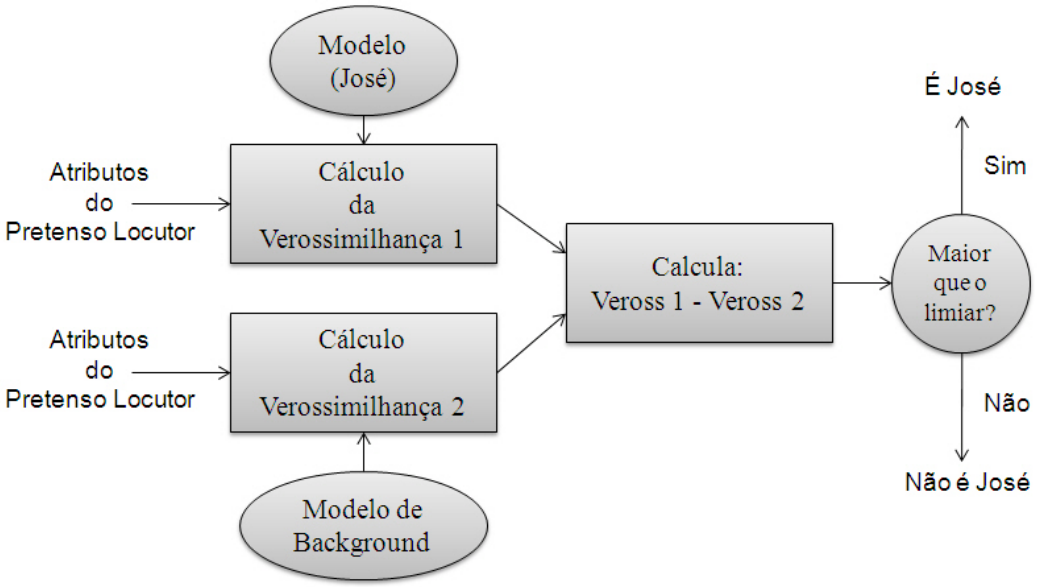


Figura 9 – Sistema de classificação baseado na verificação de locutor

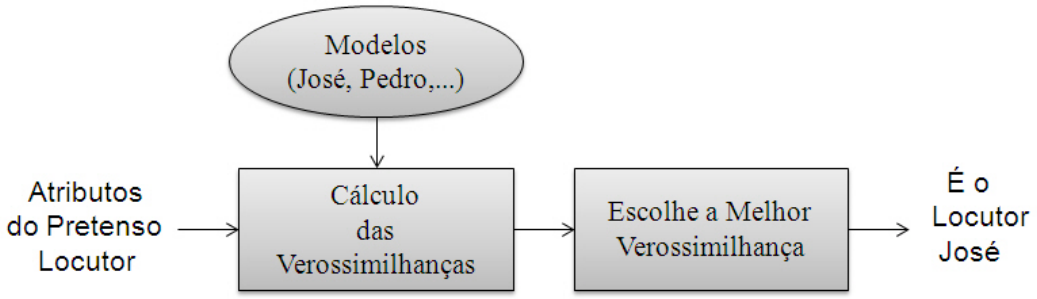


Figura 10 – Sistema de classificação baseado na identificação de locutor

Existem várias formas de gerar um modelo estatístico para o locutor. O classificador GMM (*Gaussian Mixture Models*) [6] é bastante difundido na literatura por possibilitar uma modelagem eficiente para o locutor em várias aplicações, porém é uma técnica que tende a requerer elevados níveis de processamento.

Uma técnica simples de classificação é a *Bhattacharyya Distance* (dB) [29]. Esse classificador mede a distância (separação) entre duas distribuições gaussianas. A distância total calculada utilizando as estatísticas dos padrões (vetores de atributos) é dada por

$$dB = (1/2)\ln\left[\left(\left|\boldsymbol{C}_i + \boldsymbol{C}_J\right|/2\right)/\left(\left|\boldsymbol{C}_i\right|^{1/2}\left|\boldsymbol{C}_J\right|^{1/2}\right)\right] + (1/8)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_J\right)^T\left[\left(\boldsymbol{C}_i + \boldsymbol{C}_J\right)/2\right]^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_J\right) \tag{21}$$

A expressão (21) pode ser representada por

$$dB = d_C + d_M \tag{22}$$

onde d_C representa a distância devido à diferença entre as matrizes de covariância e d_M caracteriza a distância devido à diferença entre os vetores de média. O locutor identificado é aquele que possui o menor valor da distância total. Em (21), para duas classes (locutores) representadas pelos índices (*i,J*), o vetor de média da classe *i* é $\boldsymbol{\mu}_i$ e o vetor de média da classe *J* é $\boldsymbol{\mu}_J$. Analogamente, as matrizes de covariância das duas classes são representadas por $\boldsymbol{C}_i$ e $\boldsymbol{C}_J$, respectivamente. Nesse tipo de classificador, se na etapa de treinamento forem armazenadas as estatísticas ($\boldsymbol{C}$, $\boldsymbol{\mu}$) dos parâmetros da classe *i*, no teste o classificador calcula as estatísticas da classe *J*, ou *vice-versa*. Essas estatísticas são calculadas sobre os vetores de parâmetros de cada classe. Esse classificador compara diretamente as características da voz do locutor testado com as características do sinal do locutor modelado no treinamento. Quanto menor for o valor de *dB* maior será a probabilidade do locutor testado ser o mesmo que aquele modelado no treinamento.

Uma outra técnica (*AR-vector*) [29],[90],[91] baseia-se na análise LPC para vetores. Esse tipo de análise LPC utiliza os conceitos da predição aplicada a vetores. Essa predição pode ser expressa por

$$\hat{\boldsymbol{S}}_n = \sum_{k=1}^{p} \tilde{\boldsymbol{A}}_k \boldsymbol{S}_{n-k} + \boldsymbol{E}_n \tag{23}$$

onde $\hat{S}_n$ é um vetor de amostras, $\boldsymbol{S}_{n-k}$ representa os vetores de amostras passadas (sendo *k* de 1 a *p*), $\boldsymbol{E}_n$ é o vetor erro de predição e $\tilde{\boldsymbol{A}}_k$ é a matriz de predição. Nessa técnica de classificação, as autocorrelações são expressas por

$$\boldsymbol{R}_k = \sum_{n=1}^{N-k} \boldsymbol{S}_n \boldsymbol{S}^T{}_{n+k} \tag{24}$$

onde N representa o número de vetores de parâmetros representados por $\boldsymbol{S}_n$. As matrizes de coeficientes $\tilde{\boldsymbol{A}}_k$ associadas aos vetores de atributos são calculadas por

$$\begin{pmatrix} \mathrm{R}_0 & \mathrm{R}^{\mathrm{T}}{}_1 & \cdots\cdots & \mathrm{R}^{\mathrm{T}}{}_{\mathrm{p-1}} \\ \mathrm{R}_1 & \mathrm{R}_0 & \cdots\cdots & \mathrm{R}_{\mathrm{p-2}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{R}_{\mathrm{p-1}} & \mathrm{R}_{\mathrm{p-2}} & \cdots\cdots & \mathrm{R}_0 \end{pmatrix} \begin{pmatrix} \tilde{A}_1 \\ \tilde{A}_2 \\ \vdots \\ \tilde{A}_p \end{pmatrix} = \begin{pmatrix} \mathrm{R}_1 \\ \mathrm{R}_2 \\ \vdots \\ \mathrm{R}_{\mathrm{p}} \end{pmatrix} . \tag{25}$$

Cada locutor tem uma matriz de autocorrelação e uma matriz de coeficientes. A distância de similaridade (*Itakura distance*) é dada por

$$d = (1/2)\left[\log\left(\left[\ \mathrm{tr}\left(\tilde{\boldsymbol{A}}\boldsymbol{R}_B\tilde{\boldsymbol{A}}^T\right)/\mathrm{tr}\left(\boldsymbol{B}\boldsymbol{R}_B\boldsymbol{B}^T\right)\right]\right) + \log\left(\left[\ \mathrm{tr}\left(\boldsymbol{B}\boldsymbol{R}_{\tilde{A}}\boldsymbol{B}^T\right)/\mathrm{tr}\left(\tilde{\boldsymbol{A}}\boldsymbol{R}_{\tilde{A}}\tilde{\boldsymbol{A}}^T\right)\right]\right)\right] \tag{26}$$

onde ($\tilde{\boldsymbol{A}}$, $\boldsymbol{R}_{\tilde{A}}$) é o modelo associado ao locutor armazenado no treinamento, ($\boldsymbol{B}$, $\boldsymbol{R}_B$) é o modelo associado ao pretenso locutor a ser testado e tr é a função traço de uma matriz. Nessa técnica, $\tilde{\boldsymbol{A}} = [\ \tilde{\boldsymbol{A}}_0\ \tilde{\boldsymbol{A}}_1\ \tilde{\boldsymbol{A}}_2 \ldots\ \tilde{\boldsymbol{A}}_p\ ]$ e $\tilde{\boldsymbol{A}}_0$ é a matriz identidade. O locutor identificado é aquele que possui o menor valor de distância.

O classificador GMM [6] baseia-se em modelar uma distribuição usando uma combinação linear de M gaussianas expressa por

$$p(\boldsymbol{\chi} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\boldsymbol{\chi}) \tag{27}$$

onde cada gaussiana é dada por

$$b_i(\chi) = (1/(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}) \exp\{-(1/2)(\chi - \mu_i)^T \Sigma_i^{-1} (\chi - \mu_i)\} . \qquad (28)$$

Nessa expressão, o vetor de atributos é representado por χ. As médias e as covariâncias que definem a distribuição gaussiana são estatísticas obtidas dos vetores de atributos associados ao locutor. Todas as funções gaussianas da combinação linear são ponderadas por um peso p_i de forma a representar a distribuição associada ao locutor. O efeito dessa combinação pode ser observado na Figura 11, onde o primeiro gráfico (a) é o histograma de um coeficiente cepstral obtido de 25 segundos de fala de um locutor feminino. O segundo gráfico (b) mostra a representação desse histograma usando apenas uma gaussiana. O último gráfico (c) mostra a representação feita usando-se uma combinação de 10 gaussianas.

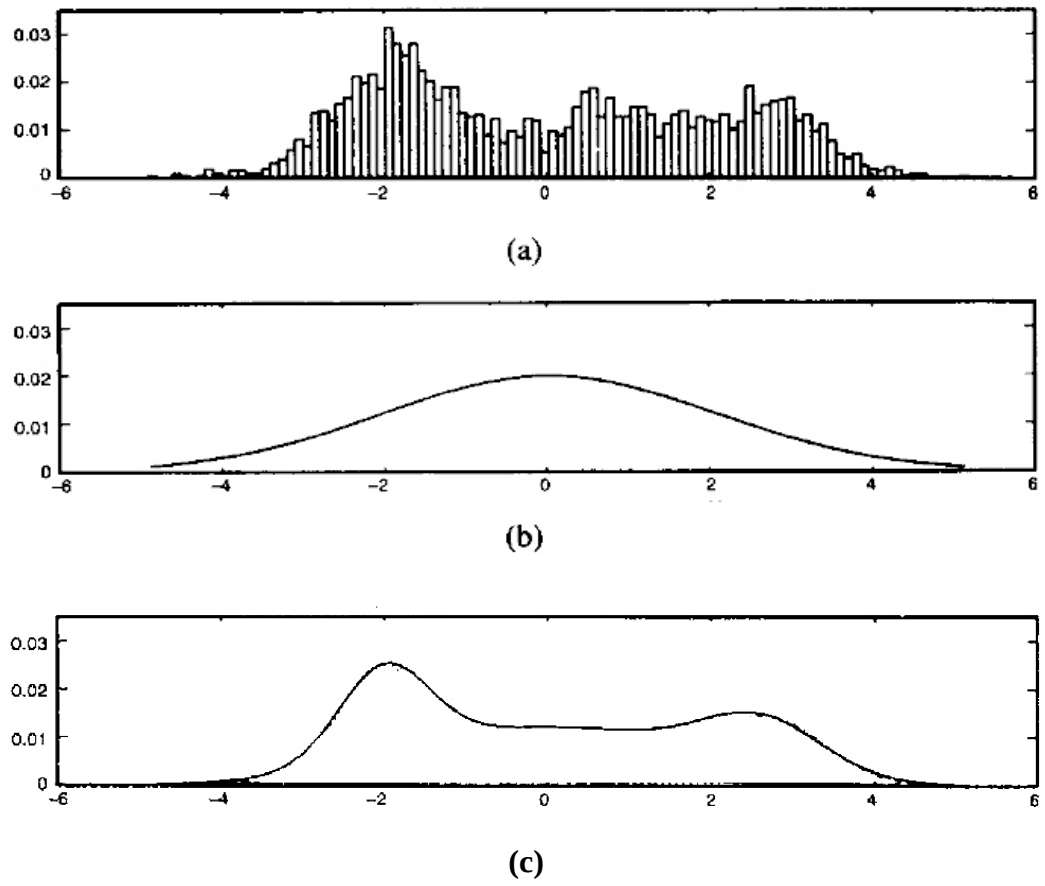


Figura 11 – Histograma dos padrões: (a) distribuição de um coeficiente cepstral obtido de 25s de sinal de fala; (b) distribuição modelada usando uma gaussiana; (c) distribuição modelada usando 10 gaussianas

No treinamento, a medida em que os vetores de atributos do locutor são apresentados ao classificador, as médias, covariâncias e pesos são recalculados pelo algoritmo EM (*Expectation-Maximization*) de forma a modelar a distribuição probabilística de todos os atributos recebidos até o momento. Nesse algoritmo, essas constantes são obtidas por

$$\overline{p}_i = \frac{1}{T}\sum_{t=1}^{T} p\left(i|\boldsymbol{\chi}_t, \lambda\right) \tag{29}$$

$$\overline{\mu}_{\mathrm{i}} = \frac{\sum_{t=1}^{T} p(i \mid \boldsymbol{\chi}_{\mathrm{t}}, \lambda)\boldsymbol{\chi}_t}{\sum_{t=1}^{T} p(i \mid \boldsymbol{\chi}_{\mathrm{t}}, \lambda)} \tag{30}$$

$$\overline{\sigma}_{\mathrm{i}}^2 = \frac{\sum_{t=1}^{T} p(i \mid \boldsymbol{\chi}_{\mathrm{t}}, \lambda)\boldsymbol{\chi}_t^2}{\sum_{t=1}^{T} p(i \mid \boldsymbol{\chi}_{\mathrm{t}}, \lambda) - \overline{\mu}_i^2} \tag{31}$$

com

$$p(i \mid \boldsymbol{\chi}_{\mathrm{t}}, \lambda) = \frac{p_i b_i(\boldsymbol{\chi}_{\mathrm{t}})}{\sum_{k=1}^{M} p_k b_k(\boldsymbol{\chi}_{\mathrm{t}})}, \tag{32}$$

onde *t* representa o t-ésimo vetor de atributos.

No início do treinamento do GMM pode-se usar um conjunto de vetores de atributos do locutor para gerar a média inicial. Além disso, pode-se utilizar como matriz de covariância inicial a matriz identidade. Para cada locutor é gerado e armazenado no treinamento, após a apresentação de todos os padrões, um modelo estatístico que pode ser representado simbolicamente por λ, onde

$$\lambda = \{p_i, \boldsymbol{\mu}_{\mathrm{i}}, \boldsymbol{\Sigma}_{\mathrm{i}}\} \quad \mathrm{i} = 1, \ldots, \mathrm{M}\,. \tag{33}$$

A fase de testes consiste em primeiramente apresentar ao classificador uma coleção de vetores de atributos associados a um pretenso locutor. Em seguida, são calculadas as probabilidades desses vetores de atributos condicionado a cada modelo λ, ou seja, são calculadas as probabilidades de que os vetores de atributos tenham sido originados do locutor representado por λ. O próximo passo é calcular a medida de distância com base nesses valores de probabilidades. O

locutor identificado é aquele que possui o maior valor dessa medida de distância, isto é,

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\chi_t \mid \lambda_k) \tag{34}$$

onde S é o número de locutores e k é o k-ésimo locutor.

O classificador M_dim_fBm (*Multidimensional fractional Brownian motion*) [29] é um dos mais recentes e baseia-se nos processos de movimento browniano fracionário. Esse classificador visa modelar o comportamento estocástico do sinal de voz. Em seu funcionamento, os parâmetros da voz são organizados na forma matricial, onde o número de linhas, c, é igual ao número de coeficientes contidos em cada vetor de atributos, e o número de colunas N é igual ao número de vetores de atributos. A transformada *wavelet* é aplicada em cada linha da matriz. Para cada conjunto de detalhes obtido pela aplicação da transformada em cada linha, é estimada a média, a variância e os atributos pH. Em seguida, são gerados os processos fBm de cada linha usando os valores (de média, de variância e de parâmetro de Hurst) estimados na etapa anterior. Consequentemente, são gerados c processos fBm. O próximo passo consiste em gerar o histograma de cada processo fBm da matriz. O conjunto de histogramas define o modelo c-dimensional do locutor associado àquela matriz. Na fase de testes, os histogramas do locutor são usados para calcular a probabilidade de que um vetor de atributos c-dimensional pertença àquele locutor. Isso é feito para os N vetores de atributos, resultando em N valores de probabilidades. Somando-se estes valores, é obtida uma medida que permite saber se o conjunto de vetores de atributos analisados pertence àquele locutor representado pelo modelo. Nesse classificador também existe a possibilidade de a matriz de atributos ser dividida em r regiões. O algoritmo então é aplicado a cada uma das regiões gerando um processo fBm r.c-dimensional que define o modelo do locutor. As outras etapas são as mesmas que as usadas quando a matriz não é dividida (r=1).