2
Reconhecimento Automático De Locutor

A voz é o produto resultante de uma sequência complexa de transformações que ocorrem em diferentes níveis, quais sejam: semântico, linguístico, articulatório e acústico. É razoável que as diferenças nestas transformações apareçam como diferenças nas propriedades acústicas do sinal de voz. As variações na voz relativas ao locutor são causadas pelas diferenças anatômicas no trato vocal e pela diferenças nos hábitos de falar de diferentes indivíduos [2].

Primariamente o sinal de voz é composto por palavras que são unidas para formarem a mensagem que se quer transmitir; em um nível secundário, o sinal de voz contém informações sobre a identidade do locutor. Enquanto o reconhecimento de voz é baseado na linguística do sinal, o reconhecimento de locutor é baseado na extração de informações sobre a identidade do locutor [19].

No Reconhecimento de locutor podem acontecer alterações as quais podem afetar seu desempenho sendo por exemplo:

- Inter locutor (locutores diferentes): fisiologia, sotaque, dialeto.
- Intra locutor (mesmo locutor): estado de saúde, emocional.
- Estilo da fala: espontâneo, formal, casual.
- Distorções acústicas: meio de gravação das locuções, meio de transmissão, ruídos aditivos .

Como foi comentado no capitulo 1 o reconhecimento de locutor pode ser dividido em duas áreas clássicas: verificação e identificação.

No caso de verificação busca-se determinar se um locutor é ou não a pessoa com a qual ela se identifica ao sistema. Nesse caso como a combinação é binária, o desempenho do sistema(probabilidade de acerto) independe do número de usuários.

No caso da identificação do locutor, o sistema deve decidir a que pessoa, dentre um conjunto de pessoas pré-especificado, pertence uma voz

desconhecida. Como esta tarefa envolve N comparações, o desempenho obviamente tende a piorar com o aumento do tamanho da população de usuários [1].

Com relação ao tipo de texto o reconhecimento automático de locutor pode ser dividido em dependente ou independente do texto.

Os sistemas dependentes do texto exigem uma fala pré-determinada. Com estes sistemas são realizadas comparações entre duas locuções com o mesmo texto. Estes sistemas são particularmente utilizados em aplicações onde o locutor é cooperativo (sabe que está sendo reconhecido) e a amostra de voz questionada pode ser obtida sob demanda. Dentre estas aplicações temos por exemplo banco por telefone e o controle de acesso [15].

Os sistemas independentes do texto trabalham com amostras de voz independentemente de seu conteúdo textual. Esses sistemas permitem uma maior flexibilidade e se adéquam propriamente a aplicações forenses e investigativas, as quais independem da pronúncia de uma frase específica nas amostras questionadas. Por aceitarem qualquer conteúdo textual, esses sistemas exigem que as amostras de voz sejam relativamente longas, aproximadamente 30 segundos. Com amostras mais longas, obtém-se uma melhor caracterização acústica da voz, e uma modelagem mais completa da voz do locutor [19] [6].

Uma visão geral de um sistema de reconhecimento de locutor consiste basicamente de 3 etapas: aquisição do sinal de voz (conversão analógico-digital), extração das características do sinal de voz e sistema classificador. Essas etapas estão ilustradas na fig.2.1.

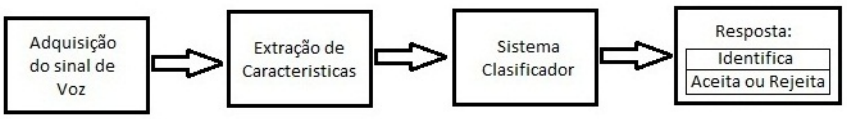

Figura 2.1: Etapas de um sistema genérico para reconhecimento de locutor

Na aquisição do sinal, a voz é transformada em um sinal elétrico (analógico) por meio de um microfone, sendo logo convertido num sinal digital através de um conversor analógico/digital. Esse sinal é então pre-processado, para supressão de informações desnecessárias ou ênfase nas importantes [18]. Por último, são extraídas as características desejadas, gerando desta maneira

vetores característicos(atributos), os quais darão os padrões para o classificador. Este depois do treinamento, visará separar classes distintas.Na identificação de locutor, todos os modelos treinados são avaliados com uma locução de teste. O modelo que apresentar melhor resultado (verossimilhança) é aceito como verdadeiro ou seja a locução de teste pertence ao locutor cujo modelo apresentou o melhor resultado. Na verificação de locutor, o modelo de um pretenso locutor determinará se a locução de teste pertence ou não ao locutor.

2.1 Pré-Processamento

O pré-processamento é realizado no o sinal de voz antes da extração de características, adequando o sinal para o processamento que vai ser efetuado. No presente trabalho foram utilizados a pré-ênfase e a normalização .

2.1.1 Pré-ênfase

A filtragem de pré-ênfase serve para atenuar as componentes de baixa frequência e incrementar as componentes de alta frequência do sinal de voz, além de prevenir contra instabilidade numérica [7].
Sua função de transferência no domino z é:

$$H(z) = 1 - az^{-1} \tag{2-1}$$

Onde o valor "a"é o coeficiente de pré-ênfase e tem valores compreendidos entre 0.9 e 1. Para os experimentos desta dissertação manteve-se o valor "a"constante em 0.95.

2.1.2 Normalização

A normalização do sinal de voz é feita com a finalidade de que o sinal de voz assuma valores dentro de uma faixa especifica, tal como [-1, 1] e [-0.5,0.5]. Essa normalização visa a corrigir a intensidade de sons gravados em seções diferentes ou por pessoas diferentes.

2.2 Extração de Características

Por meio da extração das características relevantes do sinal de voz se faz uma compressão dela para, por exemplo, uma posterior transmissão, síntese, armazenamento ou reconhecimento automático. Estas características deverão atender, na medida do possível, as seguintes condições: eficiência na representação do locutor, fácil de determinar, estável ao longo do tempo, ocorrer naturalmente e frequentemente na voz, etc [2]. Para o uso das técnicas convencionais da análise aplicadas ao sinal de voz, é necessário trabalhar com pequenos intervalos do sinal chamadas quadros entre 10 a 40 ms, em cuja duração o sinal de voz pode ser considerado aproximadamente estacionário [18].

2.2.1 Janelamento

O janelamento consiste em multiplicar no tempo o sinal do quadro x(n) por uma função chamada janela w(n), conforme a seguinte equação:

$$\tilde{x}(n) = x(n).w(n) \tag{2-2}$$

onde n é o índice da amostra. Uma janela de comprimento longo tende a produzir uma melhor representação espectral do sinal enquanto uma janela de comprimento pequeno tende a ser melhor em análises no domínio do tempo. As janelas mais utilizadas e que procuram evitar o fenômeno de Gibbs (ripple que apresenta a amplitude na resposta em frequência da janela retangular) são: Hamming, Hanning, Bartlett, Blacknam e Kaiser. Para compensar o efeito produzido pelo amaciamento da janela temporal, uma sobreposição entre janelas é efetuada aumentando a correlação entre janelas adjacentes.
Na Fig.2.2 é apresentado o janelamento com sobreposição.

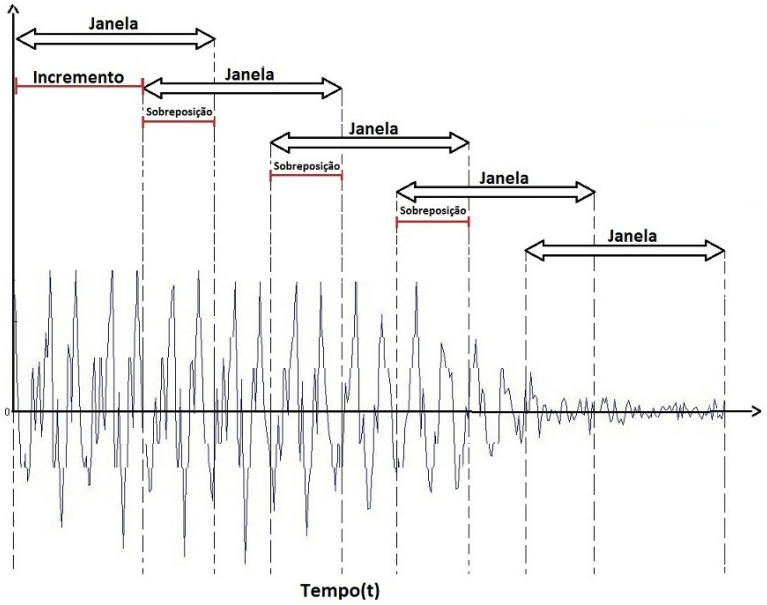

Figura 2.2: Janelamento no domínio do tempo

A sobreposição é dada pela seguinte formula [17]:

$$Sbpos = \frac{T_J - T_I}{T_J} \times 100\% \tag{2-3}$$

Onde T_J e T_I são o tamanho da janela e o tamanho do incremento, respectivamente, ambos em unidades de tempo ou de amostras.Sobreposições típicas encontram-se na ordem de 50% ou mais.

Entre algumas características de voz que pode-se utilizar no processamento de voz, e que são utilizadas no presente trabalho pode-se destacar:

2.2.2 Cepestro

O cepestro de um sinal é a transformada de Fourier inversa do logaritmo do modulo do espectro e é calculado como:

$$C_s(n) = F^{-1}\{log|F[s(n)]|\} \tag{2-4}$$

A análise Cepestral tem sua maior aplicação nos segmentos sonoros do sinal de voz, tendo em vista que o seu principal objetivo é separar linearmente no domínio da quefrência as influências da excitação (responsável pelas variações rápidas do sinal) e do trato vocal (responsável pelas variações lentas do sinal).

O nome Cepestro deriva da palavra "espectro"em virtude da analogia (mesmas regras) do cepestro no domínio "quefrência"com o "espectro"no domínio da frequência.

O cepestro pode também ser utilizado para eliminar ruído convolucional produzido por um filtro qualquer, cálculo de coeficientes LPC e calculo da pitch, o qual faz do cepestro uma característica muito utilizada no processamento de voz [4] [7].

Extração de Atributos utilizados no Reconhecimento de Locutor Independente do Texto

A seguir são apresentados os atributos utilizados nesta dissertação, os quais são explicados brevemente.

2.2.3 Coeficientes Cepestrais de Frequência Mel (MFCC)

A técnica MFCC apareceu devido a estudos na área de psicoacústica (estudo da percepção auditiva humana). Estes estudos mostraram que a escala de frequências da percepção da voz humana é não-linear; surgindo assim a idéia de criar una nova escala (escala mel).

Escala Mel

Para cada tom com uma frequência medida em Hertz, há uma relação com uma frequência de percepção medida na escala chamada mel. Definiu-se como referência a frequência de 1kHz, com potência de 40dB como 1000 mels [17]. Um mel é a unidade de medida da frequência de um tom percebida pelo ser humano. Esta frequência não corresponde linearmente à frequência física de um tom, da mesma forma que o sistema auditivo humano não percebe um tom de modo linear.

Os trabalhos desenvolvidos por Stevens e Volkmans [22] mostraram que o ouvido tem uma resolução de frequência aproximadamente linear abaixo dos 1000 Hz e, logaritmica acima deste. Uma aproximação analítica utilizada comumente deste mapeamento é dado por [17]:

$$Mel(f) = 2595.log_{10}\left(1 + \frac{f}{700}\right) \quad (2\text{-}5)$$

Banda Crítica

Experimentos desenvolvidos na percepção humana mostram que, dentro de algumas bandas não podem ser identificadas independentemente algumas frequências de som complexo; mas, se uma dessas frequências fica fora dessas bandas [20], ela pode ser identificada. Cada um destas bandas é chamada de banda crítica e varia nominalmente de 10 a 20 porcento da frequência central daquele som [20] [17]. A banda critica é comumente aproximada pela seguinte expressão :

$$Bw = 25 + 75\left[1 + 1.4(\frac{f}{1000})^2\right]^{0.69} \tag{2-6}$$

Cálculo dos MFCC

Para a obtenção dos coeficientes cepestrais de frequência mel, primeiramente o sinal de voz digitalizado s(n) passa por um filtro de pre-ênfase, e em seguida é dividido em janelas (ver capítulo anterior). Para cada janela m extrai-se o espectro S*(k,m)* do sinal , utilizando a FFT. O espectro obtido P(i),i=1,2,....,N_f. consistirá na energia de saída de cada um dos filtros, e é expressa por:

$$P(i) = \sum_{k=1}^{N/2} |S(k,m)|^2 H_i\left(k\frac{2\Pi}{N}\right), \qquad i = 1, 2, ..., N_f. \tag{2-7}$$

onde N são os números de pontos da FFT, H_i (w) é a função de transferência do i-ésimo filtro triangular[1] , $|S(k,m)|$ é o módulo da amplitude da frequência do k-ésimo ponto da m-ésima janela e N_f é o número de filtros.

Em seguida, define-se o conjunto de pontos *E(k)* por:

$$E(k) = \begin{cases} log[P(i)], & k{=}K_i \\ 0, & \text{qq outro } k \ \in [0,N\text{-}1]. \end{cases} \tag{2-8}$$

onde K_i é o ponto máximo do i-ésimo filtro [20]. Os coeficientes mel-cepestrais C_{mel}(n) são obtidos utilizando a Transformada Discreta de Coseno (DCT), dado por:

[1] *costuma-se utilizar filtros triangulares conforme ilustrado na Fig.2.3*

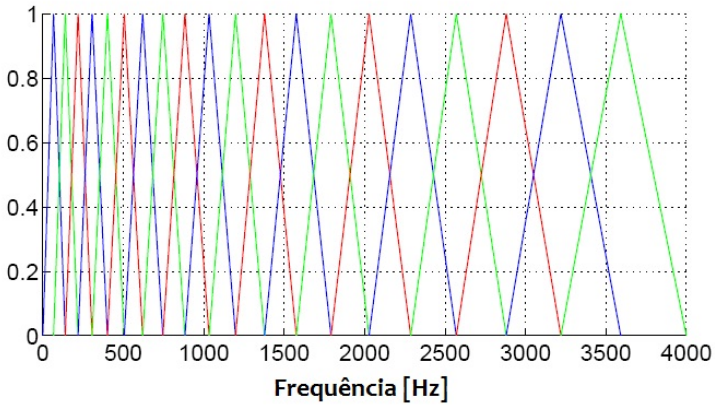

Figura 2.3: Banco de Filtros triangulares

$$C_{mel} = \sum_{i=1}^{Nf} E(k_i) cos\left(\frac{2\Pi}{N} k_i n\right), \quad n = 0, 1, 2, ..., N_c - 1 \tag{2-9}$$

onde Nc são os números de coeficientes mel-cepestrais desejados, N_f é o número de filtros e k_i é o ponto máximo do i-ésimo filtro.A Fig.2.4 ilustra o procedimento de obtenção dos coeficientes MFCC.

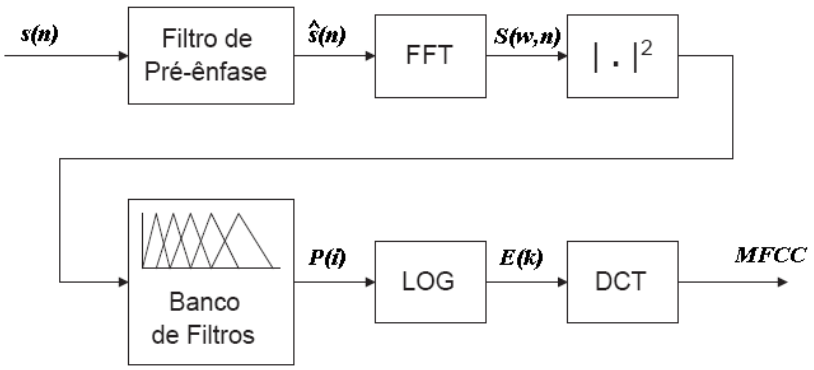

Figura 2.4: Diagrama em blocos para o cálculo dos MFCCs

2.2.4 Histogramas de Centróides de Sub-Bandas Espectrais(SSCH)

SSCH surge como a idéia de construir um método de extração de características que combina a informação das sub-bandas de potência utilizada

por métodos convencionais com a informação da frequência dominante proporcionada pelo SSC(Centróides de Sub-Bandas Espectrais) de uma maneira simples e computacionalmente eficiente [9]. A seguir descreve-se este método, que tem proporcionado bons resultados na área de reconhecimento de voz.

Estimação do espectro de potência e Calculo do centróide

Primeiro calcula-se o espectro de potência de cada quadro de voz utilizando a transformada rápida de Fourier S(f), e em seguida esse espectro é passado através de um conjunto de K filtros passa faixa com respostas de amplitude igual a H_k(f), k=1,...,K. Os centróides de Sub-Bandas Espectrais são calculados como:

$$C_k = \frac{\sum f H_k(f) S^{\gamma}(f)}{\sum H_k(f) S^{\gamma}(f)}, \qquad k = 1, ..., K \tag{2-10}$$

Onde γ é a variação dinâmica constante a qual não pode ser muito pequena ($\cong$0) já que não poderia conter informação e, nem muito grande, já que pode ser uma estimativa de ruído, e o somatório é realizado sobre todas as amostras de frequência na FFT [9] [10].

Cálculo das Sub-bandas de Potência

As Sub-bandas de Potência são calculadas como:

$$P_k = \sum H_k(f) S^{\gamma}(f) \tag{2-11}$$

Onde a soma é realizada sobre toda a sub-bandas ou sobre pequenas faixas da frequência centradas ao redor do Centróide de Sub-Bandas Espectrais, que deve fornecer estimações mais robustas, já que, a área da frequência ao redor da frequência dominante é menos influenciada pelo ruído que as outras partes da sub-banda [9].

Construção do Histograma

O histograma dos Centróides de Sub-Bandas Espectrais é construído a partir da divisão da faixa das frequências da voz em bins R_j, j=1,...,J, e calculando os número de bins como:

$$count(j) = \sum_{k=1}^{K} \Psi_j\{C_k\}, \qquad j = 1, ..., J \tag{2-12}$$

Onde :

$$\Psi_j\{C_k\} = \begin{cases} ln(\frac{p_k}{N_k}), & C_k \in R_j \\ 0, & \text{outro} \end{cases} \tag{2-13}$$

Onde N_k é o número de amostras das frequências de corte superior e inferior do k-iésimo filtro passa faixa. Para cada Centróide de Sub-Bandas Espectrais, a correspondente contagem de bin é, portanto, aumentada pelo logaritmo da Sub-banda de Potência normalizado pela largura de banda da sub-banda. A normalização é feita a fim de evitar polarização do histograma em direção às altas frequências devido ao aumento das larguras de banda dos filtros. Finalmente a transformada discreta do coseno (DCT) do histograma é calculado [9].A Fig.2.5 mostra o procedimento para o calculo do SSCH.

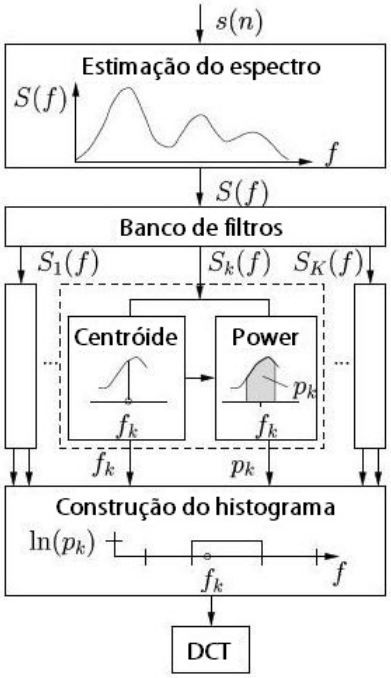

Figura 2.5: diagrama do SSCH

2.2.5
Coeficientes Cepestrais da Energia Teager

Esta técnica foi motivada pelas considerações da percepção da voz e o operador não linear teager-kaiser que calcula a energia de um sinal de ressonância. Os coeficientes cepestrais de energia teger(TECC) foram avaliados em reconhecimento de voz com ruído e mostraram-se mais robustos do que o MFCC. A seguir é descrita esta técnica [8].

Operador de Energia Teager-Kaiser

Para lei do movimento do Newton para um oscilador com massa m e elasticidade constante k tem-se que :

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0 \tag{2-14}$$

a solução consiste de um sinal x(t)=a.cos(ϕ(t)). A energia total E é a soma da energia cinética e potencial dada por :

$$E = \frac{1}{2}kx^2 + \frac{1}{2}m\dot{x}^2 \Rightarrow E \approx \omega^2 A^2 \tag{2-15}$$

onde ω é a frequência de oscilação e A é a amplitude. A partir desta análise em consideração, Teager e logo Kaiser [8], propuseram o Operador Teager-Kaiser Ψ :

$$\Psi[x(t)] = \dot{x}^2(t) - x(t)\ddot{x}(t) \approx x^2(t) - x(t+1)x(t-1) \tag{2-16}$$

Assim, além de utilizar o tradicional aproximação da energia do sinal de x^2(que só leva em conta a energia cinética do sinal), é utilizado o operador de energia Teager-Kaiser para calcular o conteúdo de energia. A média dos quadros da saída do operador de energia pode ser utilizada para a estimação das características.

Banco de filtros auditivos

O processo da audição humana, pode ser modelado por um conjunto filtros assimétricos que estimam a atividade em cada faixa de frequência. O conceito do Equivalent Rectangular Bandwidth (ERB) pode ser utilizado para quantificar a largura de faixa de filtros assimétricos como dos auditivos. Especificamente, dada a magnitude de um filtro de resposta de frequência

$|H(f)|$ e o ganho máximo do filtro $|H(f_{max})|$, a frequência f_{max} do filtro ERB (em Hz) é definida como [8]:

$$ERB = \frac{\int |H(f)|^2 df}{|H(f_{max})|^2} \tag{2-17}$$

A ERB é a passa faixa equivalente de um filtro ortogonal com ganho constante $|H(f_{max})|$ e energia igual ao original filtro de energia (o filtro de energia é definido como a integral do filtro de resposta da frequência ao quadrado). Estudos recentes [11, 14] têm mostrado que a fisiologia humana dita que as larguras de banda do filtro auditivo são dadas pela função ERB(f):

$$ERB = 6.23(f/1000)^2 + 93.39(f/1000) + 28.52 \tag{2-18}$$

onde f é a frequência central do filtro em hertz. Além disso, o filtro utilizado é equidistante na escala de frequência crítica(bark).

$$bark(f) = \frac{26.81f}{f + 3920} - 0.53 \tag{2-19}$$

onde $0 \leq$ f $\leq F_s/2$ e F_s é a frequência de amostragem do sinal. Uma boa aproximação dos filtros auditivos são filtros Gammatone assimétrica com resposta ao impulso :

$$g(t) = At^{n-1}exp(-2\pi bERB(f_c)t)\cos(2\pi f_c t) \tag{2-20}$$

Onde A, b, n são os parâmetros de projeto do filtro Gammatone e fc é a frequência central do filtro. Em [14], propõe-se que os filtros auditivos devem ter b= 1.019 e n=4, e sua magnitude espectral é mostrada na Fig.2.6.

O banco de filtros Gammatone, com filtros com largura de banda dada pela ERB(f) é uma boa aproximação do sistema auditivo humano [11].

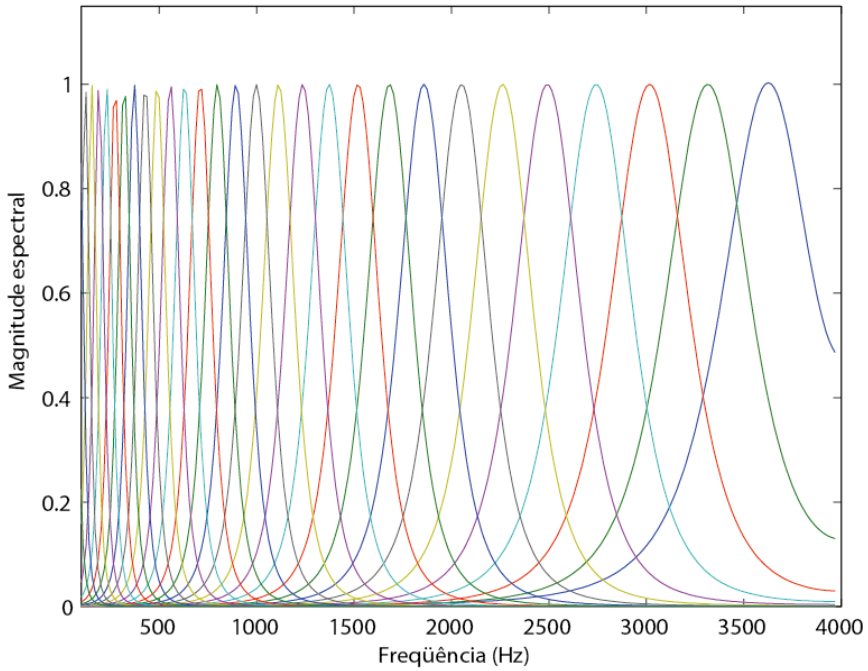

Figura 2.6: Banco de filtros de 30 filtros Gammatone

Coeficientes Cepestrais da Energia Teager

Os Coeficientes Cepestrais da Energia Teager (TECCs) são extraídos do sinal de voz da seguinte maneira: primeiro o sinal de voz é passado pelo banco de filtros Gammatone(neste trabalho foi usado entre 20-40 filtros) utilizando a equação (2-20). Depois estima-se o logaritmo da media dos quadros do operador de energia Teager-Kaiser para cada um das faixas do sinal. Logo estima-se os coeficientes cepestrais da média dos quadros da energia Teager utilizando a transformada discreta do coseno (DCT).

2.2.6 Coeficientes Delta e Delta-Delta

Estes parâmetros são utilizados para representar as mudanças dinâmicas no espectro de voz e assim perceber variações bruscas dentro do espectro, como também a variação dinâmica das caracteristicas. Os Coeficientes delta e delta-delta são obtidos através das derivadas de primeira e segunda ordem das características de voz, respectivamente. Os coeficientes delta podem ser calculados sobre os coeficientes cepestrais. Se a representação das amostras da voz no tempo t é dada por o vetor de atributos (por exemplo: MFCC, SSCH,

TECC, PAC-MFCC, etc) c_t, o vetor de parâmetros delta correspondentes é dado por [9] [24]:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta \left(c_{t+\theta} - c_{t-\theta}\right)}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2-21}$$

Onde d_t é o coeficiente delta($\triangle$) e Θ é o número de amostras necessárias para o cálculo dos coeficientes delta (tipicamente Θ=2) [9].

2.3 Sistemas de Classificação

A partir das características que são extraídas do sinal de voz é que são construídos os modelos dos locutores. Quando um locutor novo entra no sistema, se gera e armazena um modelo de sua voz a partir de suas características extraídas. Na identificação de locutor, o modelo que apresenta maior similaridade indica a quem pertence o sinal de entrada.

Existem dois tipos clássicos de modelos: O modelo baseado em casamento de padrões característicos e o modelo estatístico ou estocástico.

No modelo baseado em casamento de padrões característicos o sistema de classificação faz comparações. É asumido que a observaçao é uma replica imperfeita do modelo armazenado e geralmente o alinhamento dos quadros do modelo para os quadros observados é feito para minimizar uma medida de distancia d [5]. Dentro deste método temos: Alinhamento Temporal Dinâmico (DTW), Quantização Vetorial(VQ), Modelo Autoregressivo Vetorial (AR Vetorial).

No modelo estatístico ou estocástico, o sistema de classificação é probabilística e resulta numa medida de verossimilhança, ou probabilidade condicional, dada pela observação do modelo [5]. Dentro deste modelo temos: Modelo de Markov Escondido(HMM) e o Modelo de Mistura de Gaussianas (GMM).