3
Extração de Atributos

No capítulo 2, a sequência de vetores $O = \{O_1, O_2, ... , O_T\}$ representa as informações (também chamadas de atributos) extraídas do sinal de voz. Nada foi dito sobre como encontrá-las. Infelizmente, não existe nenhuma equação ou modelo que indique a melhor maneira de obter esses dados, e nem como minimizar o problema do descasamento citado em 2.4.3.

Vários autores apresentaram diferentes tipos de atributos. Este capítulo fará uma comparação teórica e prática dos seguintes métodos: *Mel-Frequency Cepstral Coefficients* (MFCC), *Subband Spectral Centroid Histogram* (SSCH) e *Power Normalized Cepstral Coefficients* (PNCC). Esses dois últimos métodos têm mostrado bons resultados na literatura, mas não foi encontrada nenhuma comparação direta entre eles.

As extrações de MFCC, SSCH e PNCC possuem etapas em comum, como mostra a Figura 18. A diferença está na implementação do bloco Informações do Espectro.

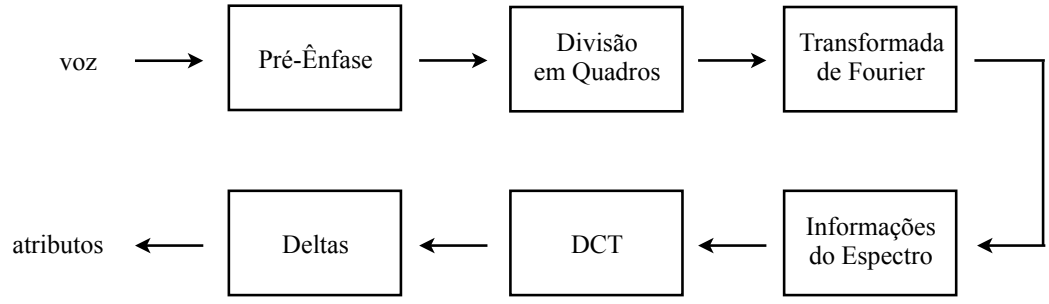

Figura 18:Esquema geral da extração de atributos.

Cada uma das etapas será explicada nas seções a seguir.

3.1
Pré-Ênfase

As baixas frequências concentram a maior parte dos dados da voz. No entanto, a informação das altas frequências também é importante para o

reconhecimento. Para enfatizá-la um pouco mais, é aplicado o filtro passa-alta de primeira ordem [18], dado por

$$H(z) = 1 - \alpha z^{-1} \quad 0 \le \alpha \le 1 \tag{16}$$

O efeito deste primeiro estágio é ilustrado na Figura 19.

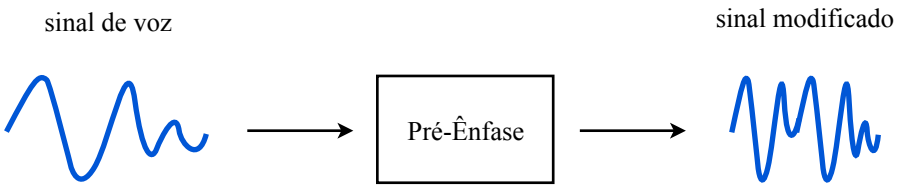

Figura 19:Entrada e saída do bloco de Pré-Ênfase.

3.2
Divisão em Quadros

Ao longo das frases, as características do sinal se alteram de acordo com as vogais e consoantes. Logo, os atributos precisam ser extraídos de vários trechos, a fim de acompanhar essas mudanças. Por isso, existe o segundo bloco, para dividir o sinal nos chamados "quadros", conforme mostra a Figura 20.

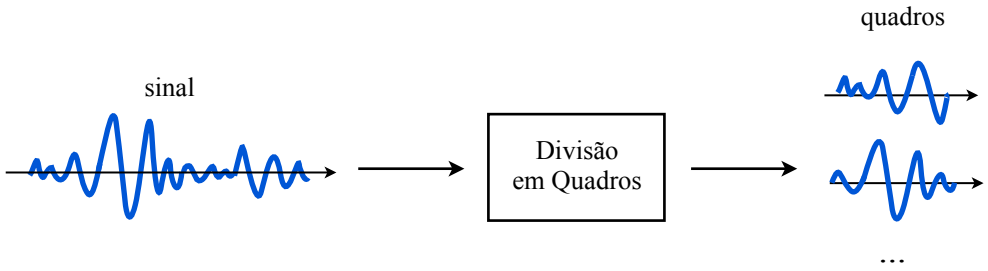

Figura 20:Entrada e saída do bloco da Divisão em Quadros.

Um modo de fazer a divisão é usar quadros consecutivos de mesmo tamanho, como na Figura 21. Mas o problema dessa abordagem é a quebra de ondulações em quadros vizinhos, gerando perda de informação.

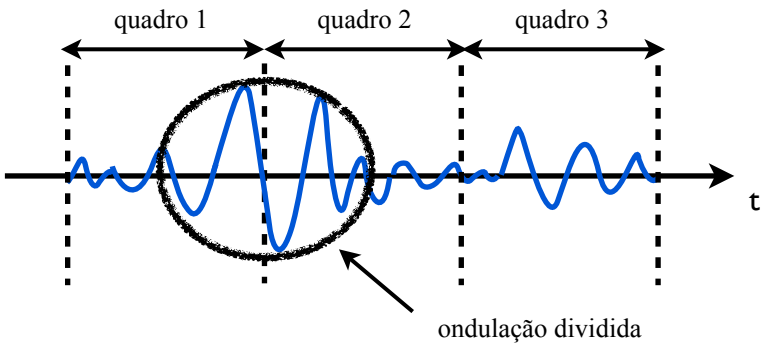

Figura 21:Divisão em quadros consecutivos, causando perda de informação.

Por esse motivo, a divisão costuma ser feita com superposições. Em outras palavras, um quadro começa um pouco antes de o anterior terminar, tal qual se vê na Figura 22. Desse modo, mesmo que um trecho importante seja cortado no fim de um quadro, ele estará inteiro no quadro seguinte.

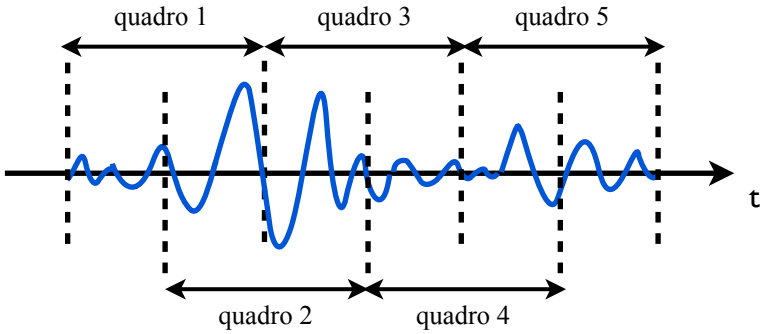

Figura 22:Divisão em quadros com superposição.

A divisão do sinal traz também um segundo problema: a descontinuidade. Por causa da partição, cada trecho começa e termina bruscamente, o que prejudica a extração de atributos. É necessário então suavizar o quadro, multiplicando-o por uma função janela como na Figura 23.

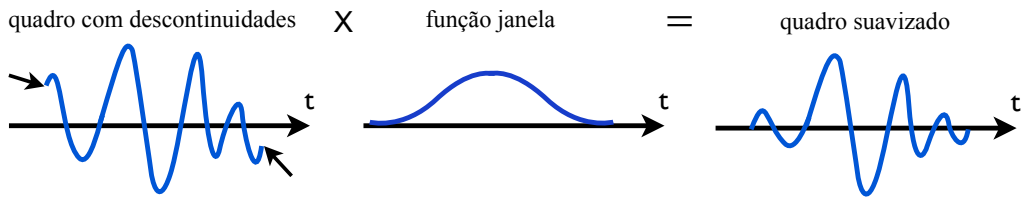

Figura 23:Suavização de um quadro através da multiplicação por uma função janela.

Existem diversos tipos de função janela. Uma das mais utilizadas no campo de processamento de sinais de voz é a janela de Hamming [19], dada pela expressão

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{17}$$

onde N é o número total de amostras do quadro.

3.3 Transformada de Fourier

Os atributos são utilizados para o treinamento do HMM de cada fone. É importante que esses valores representem bem um som e diferenciem-no dos demais. Ou seja, se os dados numéricos do fone "e" forem similares aos do fone "a", seus Modelos de Markov serão parecidos e o reconhecedor não terá precisão para identificar qual deles foi pronunciado.

Por isso, não se usam os valores do sinal no domínio do tempo diretamente. A Figura 24 mostra o motivo com as formas de onda das vogais "a" e "e" (aberto, como em "época"), respectivamente. Como é difícil diferenciar uma da outra até mesmo visualmente, fica claro que elas não são adequadas para definir os modelos.

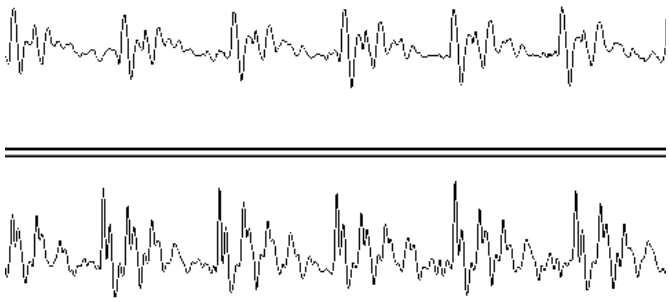

Figura 24:Sinal sonoro no domínio do tempo das vogais "a" e "e", respectivamente.

No entanto, é fácil notar que os fones são formados por vibrações que se repetem ao longo do tempo. Daí surge a ideia de extrair características sobre essas oscilações. Isso pode ser feito com a transformada discreta de Fourier [20], dada por

$$G(k) = \sum_{k=0}^{N-1} g(n) e^{-\frac{j2\pi kn}{N}} \tag{18}$$

onde N é o total de amostras do quadro.

Aplicando a transformada nos sinais da Figura 24, e utilizando a escala em dB (isto é, 10 log|$G(k)$|), os espectros das vogais "a" e "e" (aberto) são obtidos

como na Figura 25. A diferença entre dos dois sons fica mais clara agora: o “e” é mais agudo e, portanto, apresenta vibrações em frequências mais altas.

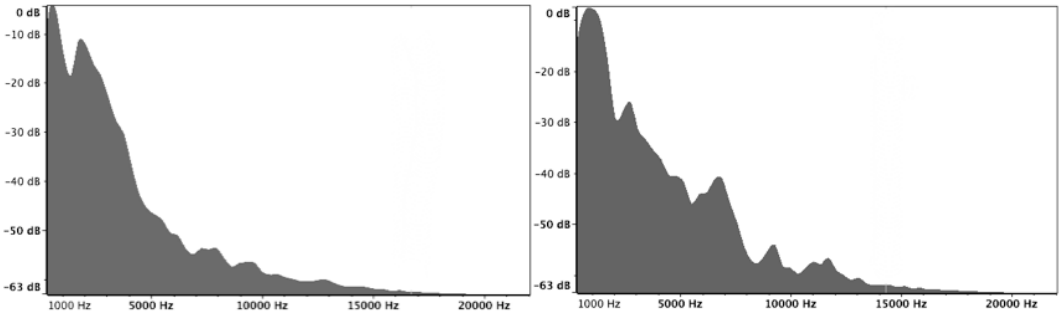

Figura 25:Espectro das vogais “a” e “e”, respectivamente.

Aqui se percebe a necessidade do primeiro estágio da extração de atributos (a Pré-Ênfase). Existe uma queda natural do espectro, diminuindo os detalhes nas altas frequências. O filtro passa alta mostrado em 3.1 provoca uma elevação para compensar essa queda, aumentando as informações das bandas altas.

Assim, a transformada de Fourier é extraída de cada quadro do sinal, como mostra a Figura 26.

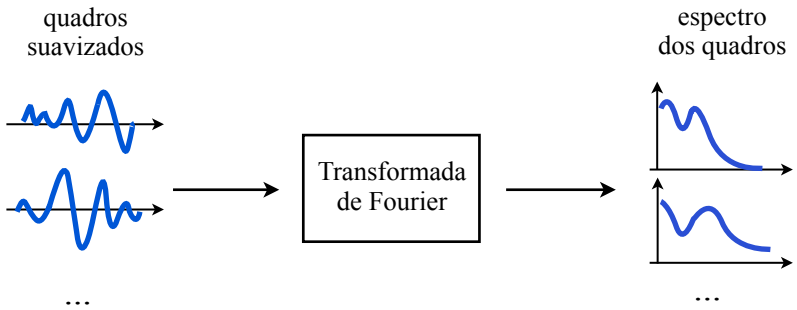

Figura 26:Entrada e saída do bloco correspondente à transformada de Fourier.

3.4
Informações do Espectro

Os três métodos investigados – MFCC, SSCH e PNCC – extraem informações numéricas do espectro $G(k)$ de cada quadro do sinal, conforme mostra a Figura 27.

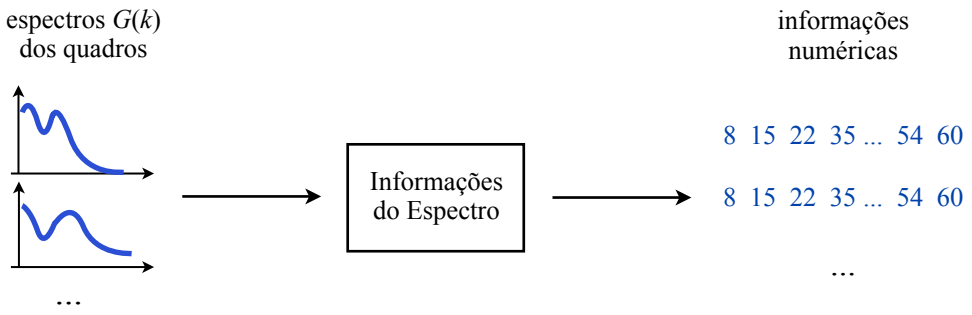

Figura 27:Entrada e saída do bloco correspondente às técnicas específicas.

O procedimento acima envolve dividir o espectro em *B* bandas (trechos) e extrair um valor de cada um deles separadamente. Portanto, a Figura 27 pode ser detalhada na Figura 28.

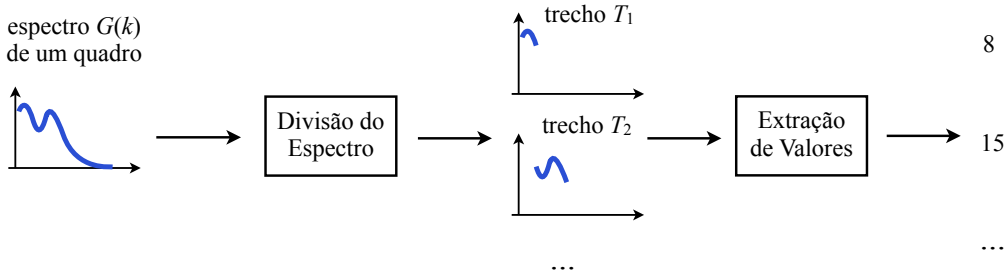

Figura 28:Detalhamento do bloco “técnicas gerais” para o espectro de um quadro.

Nas subseções 3.4.1, 3.4.2 e 3.4.3, é explicado em detalhes como cada um dos três métodos realiza a divisão do espectro e a extração de valores.

3.4.1. MFCC

O método MFCC (*Mel-Frequency Cepstral Coefficients*) [11, p. 314][21] faz a divisão do espectro por filtros, com os mesmos princípios da divisão do sinal em quadros (vista em 3.2). Ou seja, o começo de um filtro está sempre um pouco antes do final do filtro anterior, para evitar a perda de informação. E, similares às janelas de Hamming, são aplicados filtros triangulares, para evitar o problema da descontinuidade. Em seguida, a extração de valores é feita com o logaritmo da energia de cada trecho. A Figura 29 ilustra o processo.

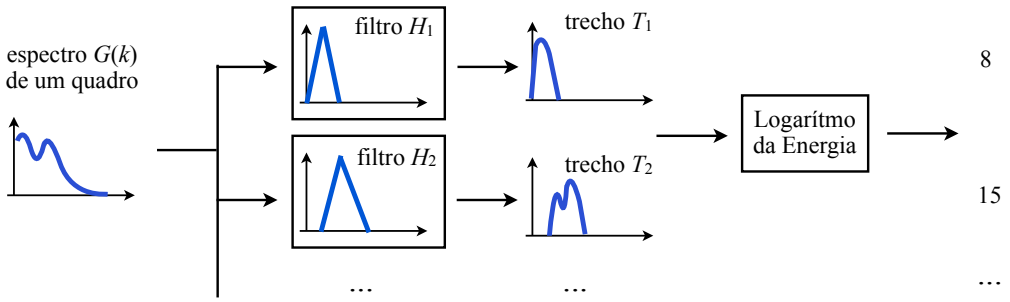

Figura 29:Detalhamento da Figura 28 para o método MFCC.

Além de superpostos e triangulares, os filtros do método MFCC têm larguras crescentes, como mostra a Figura 30.

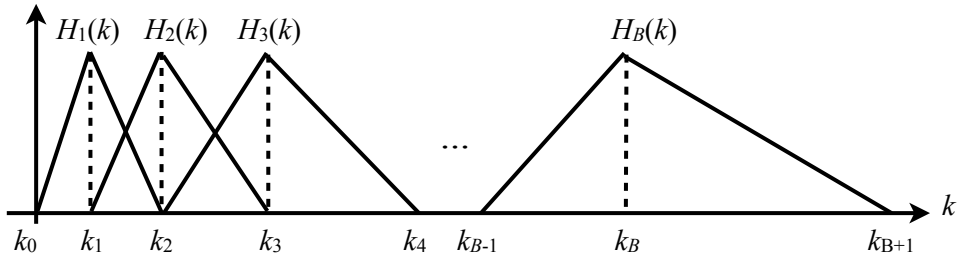

Figura 30:Filtros triangulares do método MFCC.

Essa diferença de larguras ocorre porque outra escala de frequências é utilizada, já que os valores em Hertz não refletem bem a percepção auditiva humana. Por exemplo, um sinal senoidal de 880 Hz não soa duas vezes mais agudo que um de 440 Hz e nem quatro vezes mais agudo que um de 220 Hz. Para se ter uma representação melhor, foi criada a escala mel, cujo nome veio da palavra melodia [22]. Através de experimentos com diversos ouvintes, chegou-se à conversão de f (em Hertz) para m (em mel) dada por

$$m = M(f) = 1125 \ln\left(1 + \frac{f}{700}\right)$$
$$f = M^{-1}(m) = 700\left(e^{\frac{m}{1125}} - 1\right) \tag{19}$$

A conversão se mostrou muito eficaz para extrair dados do sinal de voz, já que os filtros seguem o comportamento da audição humana – 2 mel soa duas vezes mais agudo que 1 mel. Daí vem o nome "Mel-Frequency" da sigla MFCC.

Com a expressão (27), os filtros $H_b(k)$ da Figura 29 são definidos matematicamente no apêndice 9.6. Multiplicando cada filtro pelo módulo de $G(k)$ ao quadrado, obtém-se os trechos $T_b(k)$, ou seja,

$$T_b(k) = |G(k)|^2 H_b(k) \tag{20}$$

Calculando o logaritmo da energia de cada $T_b(k)$, são obtidos os B valores finais a_b do bloco de técnicas específicas, dados por

$$a_b = \log\left(\sum_{k=1}^{N} T_b(k)\right) \tag{21}$$

3.4.2. SSCH

Embora tenha um desempenho bom com sinais de voz limpos, o método MFCC se mostrou insatisfatório quando há ruído. Por isso, o método SSCH (*Subband Spectral Centroid Histograms*) foi proposto em [3][23] como uma alternativa mais robusta.

Seu argumento é que o formato do espectro do sinal de voz não se altera tanto com a adição de um ruído moderado. Mais especificamente, a posição horizontal dos picos não é muito afetada, conforme mostra a Figura 31.

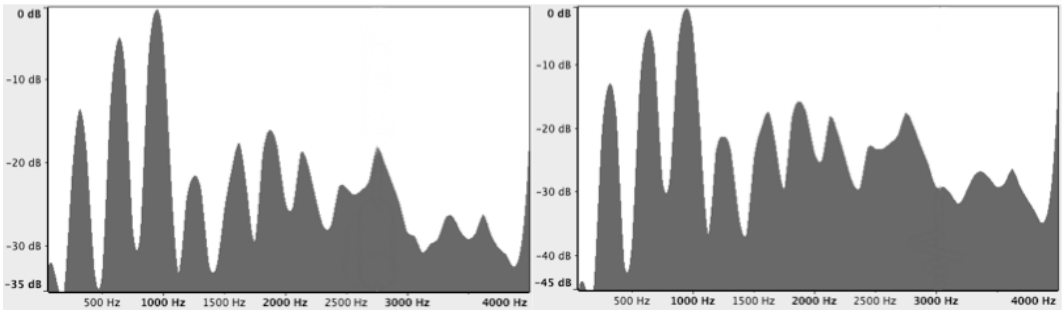

Figura 31:Comparação entre os espectros de um sinal limpo e do mesmo sinal com ruído branco de razão sinal-ruído de 10 dB.

Para representar a posição horizontal dos picos, o centróide de cada trecho do espectro é calculado – trata-se de uma medida equivalente ao centro de massa. Mesmo que o trecho sofra certa distorção por causa do ruído, seu centróide será pouco afetado. A Figura 32 ilustra a propriedade.

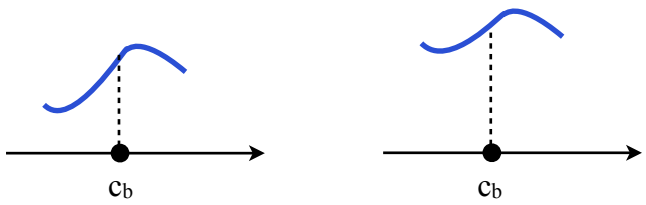

Figura 32:Centróide de um trecho limpo do espectro comparado com o centróide desse mesmo trecho com ruído.

Portanto, é interessante combinar o conceito de energia visto em 3.4.1 com os centróides, obtendo assim valores numéricos mais robustos ao ruído. A combinação é feita por um histograma que será explicado mais adiante.

Estabelecida a ideia do método SSCH, a Figura 33 mostra seu esquema geral em contraste com o MFCC.

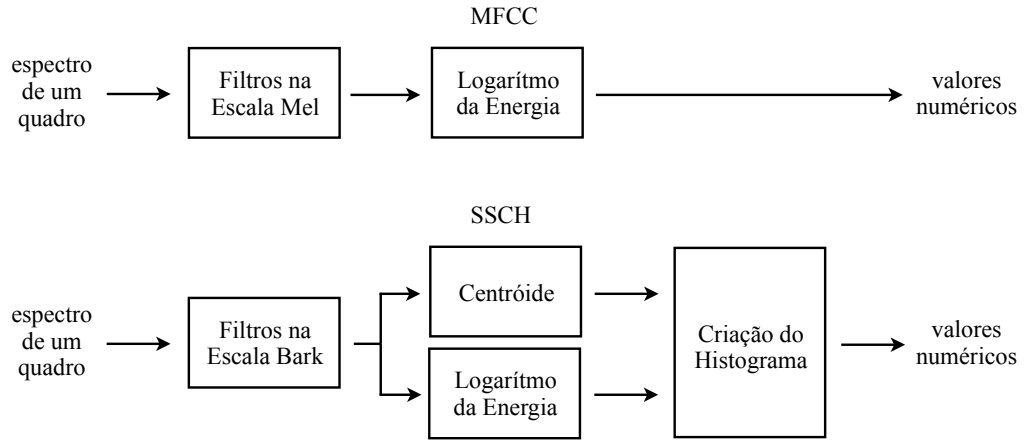

Figura 33:Comparação geral entre os métodos MFCC e SSCH.

Os procedimentos se iniciam de modo similar, com a criação dos filtros triangulares $H_b(k)$ da Figura 30. A diferença agora é que, para definir a posição e a largura dos filtros, a escala Bark é usada no lugar da Mel. A conversão de Bark para Hertz é dada por

$$r = R(f) = \frac{26.81 f}{f + 1960} \tag{22}$$

Aplicando (20), são geradas os trechos $T_b(k)$. Para cada um deles, o centróide c_b é obtido pela equação

$$c_b = \frac{\sum_{k=1}^{N} k T_b(k)}{\sum_{k=1}^{N} T_b(k)} \tag{23}$$

Já a energia e_b da banda é calculada com

$$e_b = \log\left(\sum_{k=1}^{N} T_b(k) \right) \tag{24}$$

Com o centróide e a energia de cada trecho do espectro, resta apenas combiná-los. O eixo das frequências é dividido em I intervalos de mesmo comprimento e sem sobreposição. Para cada um deles, é somada todas as energias

dos centróides que estiverem no intervalo. A coletânea das somas forma o histograma com valores $a_1, a_2, \ldots, a_I$. A Figura 34 ilustra o procedimento.

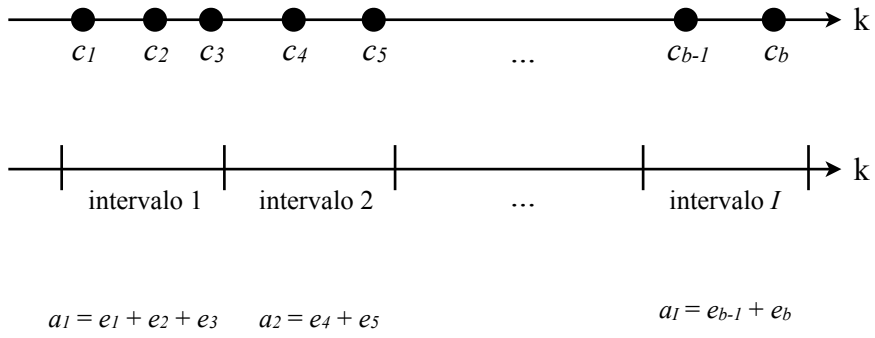

Figura 34:Criação do histograma.

3.4.3. PNCC

Os atributos PNCC (*Power Normalized Cepstral Coefficients*) foram apresentados em [4][24] como uma evolução dos MFCC, alterando algumas de suas etapas para torná-lo mais robusto. A Figura 35 mostra a comparação entre os dois.

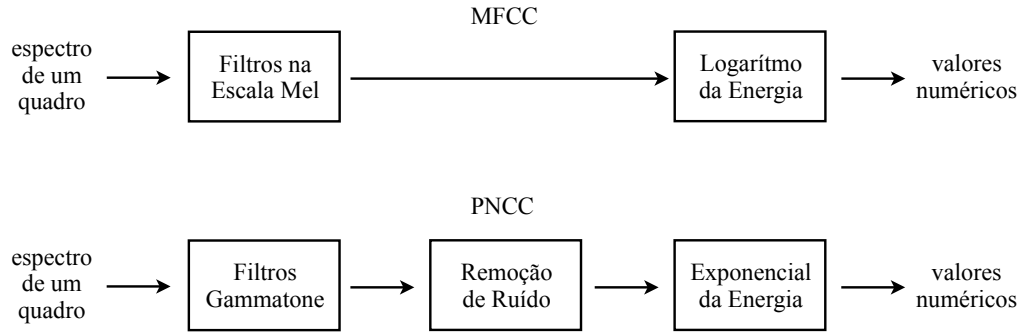

Figura 35:Comparação geral entre os métodos MFCC e PNCC.

A primeira modificação se dá na divisão do espectro. Em vez de B filtros triangulares baseados na escala mel, são aplicados B filtros gammatone [25]. Eles representam bem a resposta impulsional da membrana basilar do ouvido humano, cuja expressão no domínio do tempo (resposta impulsional) é dada por

$$g(t) = at^{n-1}e^{-2\pi c_b t}\cos\left(2\pi f_b t + \phi\right) \tag{25}$$

onde a é a amplitude, n é a ordem do filtro, c_b é o comprimento de banda, f_b é a frequência central da banda e ϕ é a fase. Baseado nesse comportamento, a escala Equivalent Rectangular Bandwidth (ERB) foi criada, e seus valores em função de f (em Hertz) são iguais a

$$r = ERB(f) = 24.7(1 + 0.00437f) \tag{26}$$

Com a mesma lógica utilizada no MFCC, as frequências centrais f_b são espaçadas em intervalos de mesmo comprimento na escala ERB, como mostra a Figura 36.

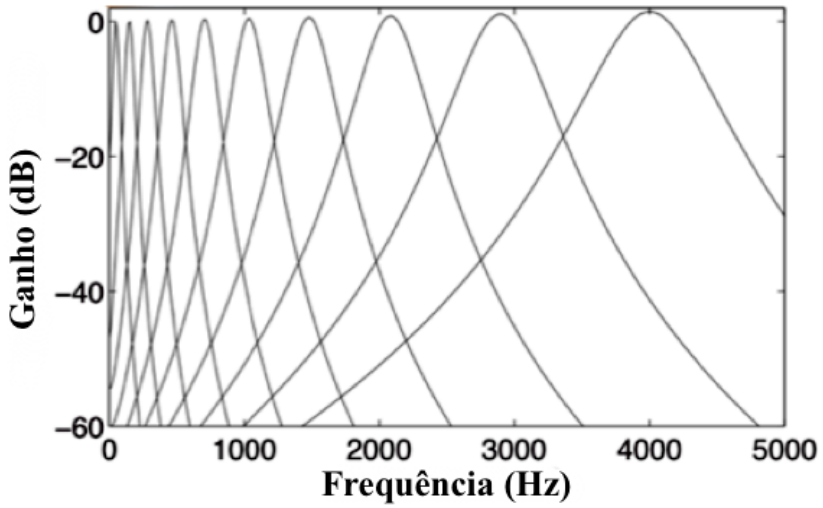

Figura 36:Filtros gammatone igualmente espaçados na escala ERB.

A implementação detalhada dos filtros pode ser vista em [26].

A segunda modificação adiciona uma nova etapa. A Figura 31 mostrou o efeito do ruído no espectro de voz. Nota-se que houve uma elevação geral na curva. Portanto, é interessante remover esse acréscimo após a divisão do sinal em bandas, aprofundando seus vales. Isso é feito com a média das energias de uma banda ao longo de alguns quadros consecutivos, pois o ruído costuma ser mais estacionário que a onda de voz. A implementação detalhada é descrita em [4].

A terceira e última modificação foi feita na operação não-linear sobre a energia da banda. A função logarítmica apresenta uma grande inclinação para valores próximos de zero. Isso altera bastante os atributos MFCC quando se adiciona ruído a pequenos valores de energia. Por isso, foi escolhida a função de potenciação, que cresce mais suavemente. A energia então é elevada a uma constante a_0 determinada experimentalmente (vide seção 6.3).

3.5
Transformada Discreta do Cosseno (DCT)

Após as técnicas específicas, gera-se uma determinada quantidade de valores para cada quadro. Em geral, quanto mais valores existirem, mais detalhes do sinal são capturados. Entretanto, a quantidade não deve ser muito grande, pois haveria um aumento de processamento no reconhecedor de voz.

Felizmente, existem métodos de compressão que reduzem o total de valores sem perda de informação. Um desses métodos é bastante conhecido e se chama transformada discreta do cosseno (em inglês, *discrete cosine transform*, ou DCT) [27]. Aplicando a sequência de vetores das técnicas específicas, a transformada gera novos valores em menor quantidade, como mostra a Figura 37.

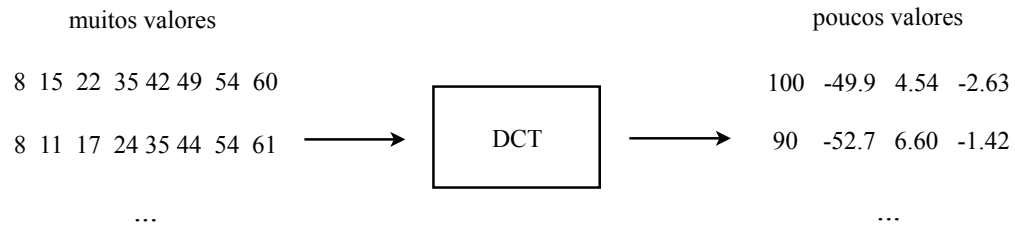

Figura 37:Entrada e saída do bloco correspondente à transformada discreta do cosseno.

Dada uma sequência de valores $a = \{a_1, a_2, \ldots, a_N\}$, a transformada y é dada pela primeira linha da expressão (27). Logo abaixo, está a expressão da transformada inversa (para converter os valores de y de volta para a).

$$y_k = C(a) = \sum_{n=1}^{N} w_k a_n \cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right)$$

$$a_n = C^{-1}(y) = \sum_{k=1}^{N} w_k y_k \cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right)$$

$$w_k = \begin{cases} \frac{1}{\sqrt{N}} & k = 1 \\ \frac{2}{\sqrt{N}} & 2 \le k \le N \end{cases} \tag{27}$$

A aplicação direta da primeira expressão gera $y = \{y_1, y_2, \ldots, y_N\}$, com a mesma quantidade de termos que a. A redução do total de números ocorre em seguida: como a transformada tende a concentrar os valores mais significativos nos primeiros termos de y, os últimos podem ser descartados. Para ilustrar essa propriedade, a Figura 38 mostra um exemplo em que se calcula a DCT, eliminam-se os últimos coeficientes (são substituídos por zero) e aplica-se a inversa. A sequência final â fica bem similar à inicial, comprovando que a maior parte da informação está no começo de y.

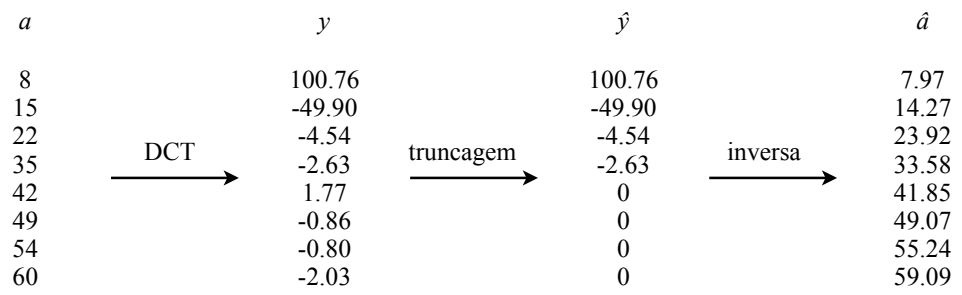

Figura 38:Exemplo da compressão de informação através da DCT.

A quantidade de termos descartados é ajustada empiricamente.

3.6
Coeficientes Delta e de Aceleração

Na seção 2.2, os trifones foram criados para que cada fone pudesse guardar a influência de seus vizinhos, melhorando assim a performance do reconhecimento. O mesmo princípio pode ser usado na extração de atributos: acrescentar alguma informação sobre os quadros adjacentes. A Figura 39 mostra um exemplo hipotético.

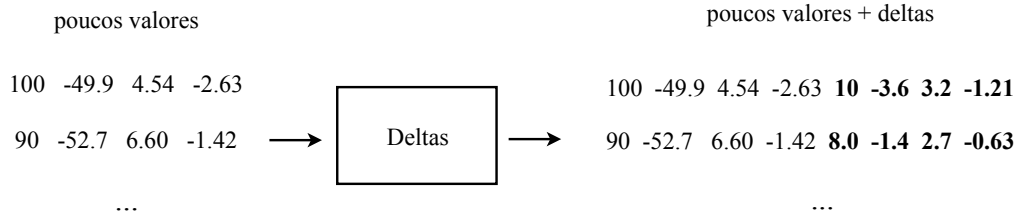

Figura 39:Entrada e saída do bloco correspondente aos coeficientes delta.

Em análise de funções contínuas, a derivada fornece o comportamento da curva nas vizinhanças de um ponto. No domínio discreto, isso se traduz na diferença entre um elemento da sequência e seus vizinhos. Através da análise de regressão, os coeficientes delta d_n [28] são obtidos considerando a diferença entre a amostra n e as Δ anteriores e posteriores.

$$d_n = \frac{\sum_{\delta=1}^{\Delta}\left(x_{n+\delta} - x_{n-\delta}\right)}{2\sum_{\delta=1}^{\Delta}\delta^2} \tag{28}$$

Acrescentando os coeficientes d_n, o total de atributos por quadro dobra. Opcionalmente, coeficientes de aceleração também podem ser inclusos aplicando (28) nos próprios coeficientes delta d_n, triplicando a total inicial de valores.

Vale notar que o ganho de desempenho com esses coeficientes vem às custas de um aumento na carga computacional, já que os HMMs passam a processar mais entradas numéricas.