3
Os Atributos MFCC e PNCC do Sinal de Voz

No sinal de voz incorporam-se informações importantes do falante, que são altamente perceptíveis, tais como, dialeto, contexto, estilo de falar, estado emocional, etc. Há outras características que dificilmente são perceptíveis, tais como tons, frequências dos formantes, intensidade, entre outros [29].

Uma grande quantidade de dados é gerada durante a produção de voz, enquanto as características fundamentais das mudanças da voz são geradas lentamente. Por conseguinte, requer-se menos dados para representar as características da voz e da pessoa que falou. Com a parametrização da voz consegue reduzir a quantidade de dados a processar.

Por exemplo, se o sinal de voz é adquirido a 8000 amostras/s quantificadas a 16 bits/amostra, parametrizando quadros de 20 ms a 14 coeficientes, obtêm-se uma redução de dados de 11.4 (160/14). Desta forma, reduz-se a complexidade computacional do processo [30].

Além disso, através da parametrização consegue-se vetores de características, como foi mencionado em 2.4.2, nos quais obtêm-se elementos relevantes de classificação que apresentam características distintivas respeito aos gerados por uma outra palavra ou som, contribuindo significativamente para o seu reconhecimento.

Na atualidade existem diferentes metodologias e procedimentos de análise para extração de atributos do sinal de voz, que se centram em diferentes aspectos representativos. Neste capítulo apresentam-se e analisam-se as técnicas de parametrização utilizadas nos experimentos desta dissertação, as quais são:

- Mel-Frequency Cepstral Coefficients (MFCC).
- Power Normalized Cepstral Coefficients (PNCC).

Essas duas técnicas possuem etapas em comum, tais como: a pré-ênfases, a análise localizada do sinal de voz (divisão em quadros) e a transformada de Fourier, que foram apresentadas na seção 2.4, além das etapas de transformada discreta do cosseno, coeficientes delta e os coeficientes de aceleração, que serão apresentados neste capítulo.

A Fig. 3.1 faz uma comparação das duas técnicas, destacando o bloco que faz diferença entre elas, chamado de *informação do espectro* [15]. Esse bloco consiste em dividir cada quadro fornecido pela DFT em B bandas de frequência e extrair um valor de cada um deles separadamente. Esse procedimento será explicado para cada um dos métodos nas seções a seguir.

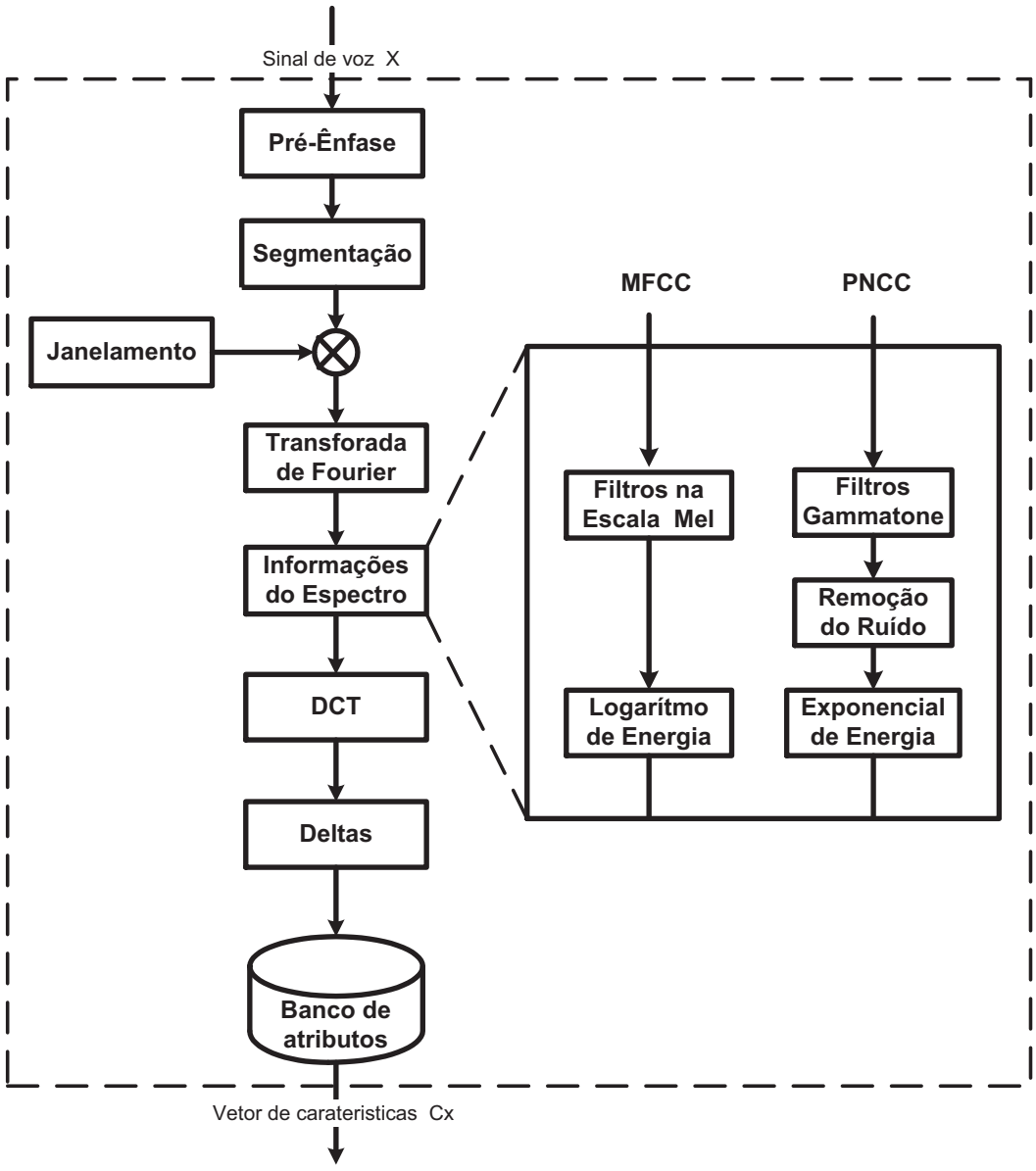

Figura 3.1: Comparação dos métodos de extração de atributos

Cabe ressaltar que para realizar a extração de atributos foram utilizadas as ferramentas *Matlab* [31] [32] e *HMM Tool Kit (HTK)*[33], apresentadas nos Anexos C e D, respectivamente.

3.1 Mel-Frequency Cepstral Coefficients (MFCC)

A técnica de extração de atributos Mel-Frequency Cepstral Coefficients (MFCC)[34] faz uma análise de características espectrais de tempo curto, baseando-se no uso do espectro da voz convertido para uma escala de frequências denominada MEL que é uma escala que visa imitar as características únicas perceptíveis pelo ouvido humano. Estes coeficientes são uma representação definida como o cepstrum de um sinal janelado no tempo, que tem sido derivado da aplicação da DFT, em escalas de frequência não lineares.

Para a extração dos vetores de características MFCC, são necessárias as etapas mencionadas na seção 2.4.1, sendo o processo explicado brevemente a seguir.

- O sinal de voz *x(n)* a parametrizar é passado através do filtro de pre-ênfase da equação 2-2 com $\alpha = 0.97$. Essa etapa é recomendável para compensar a atenuação das componentes de alta frequência causadas pelo mecanismo da produção de voz.
- Depois do sinal ser filtrado, é necessário atenuar as descontinuidades causadas no início e no final do sinal de cada segmento, aplicando uma janela Hamming de 25 ms de comprimento, com deslocamento entre janelas de 10 ms, obtendo-se assim vetores MFCC a cada 10 ms (equações 2-4 e 2-5.
- Após a etapa de janelamento do sinal, aplica-se a DFT da equação 2-6 para obter o espectro.
- Uma vez calculada a DFT obtém-se a potência espectral, utilizando a equação

$$S[k] =\mid X[k] \mid^2 = (real(X[k]))^2 + (imag(X[k]))^2 \tag{3-1}$$

- A etapa a seguir é a chamada **informações de espectro**, que faz a distinção entre as técnicas de extração, na qual aplica-se um banco de M filtros à potência espectral.

  O banco de filtros está formado por filtros triangulares, espaçados de acordo com a escala de frequência MEL, representada pela equação

$$Mel(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{3-2}$$

que, como foi mencionado, imita a resposta em frequência do sistema auditivo humano.

Matematicamente, os filtros MEL, são definidos pela seguinte resposta em frequência

$$H_m[k] = \begin{cases} 0 & k < k[m-1] \\ \frac{2(k-k[m-1])}{(k[m+1]-k[m-1])(k[m]-k[m-1])}\ , & k[m-1] \leq k \leq k[m] \\ \frac{2(k[m+1]-k)}{(k[m+1]-k[m-1])(k[m+1]-k[m])}\ , & k[m] \leq k \leq k[m+1] \\ 0 & k > k[m+1] \end{cases} \tag{3-3}$$

Cada filtro calcula a média do espectro em torno da frequência central, e têm diferentes larguras de banda. Quanto maior é a frequência maior é a largura de banda, como mostra a Fig. 3.2.

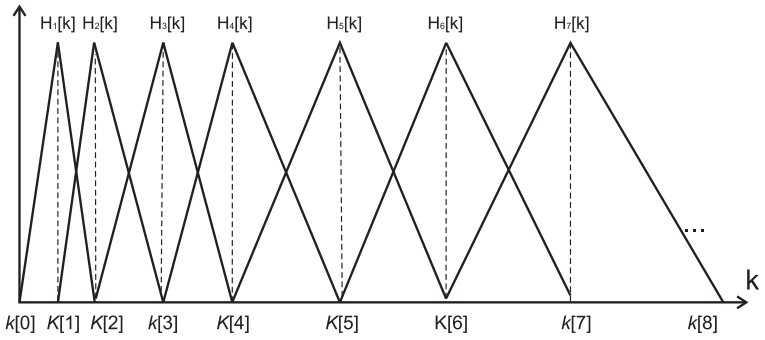

Figura 3.2: Banco de filtros usado na técnica MFCC

Para determinar matematicamente os segmentos, parte-se das frequências extremas f_l e f_h que são as frequências de corte do banco de filtros em Hz. Esses valores são usados para dividir o intervalo em $B+1$ partes iguais. Para obter os valores em Hz, basta aplicar a função inversa

$$k[m] = \left(\frac{N}{F_s} \right) Mel^{-1} \left(\ Mel(f_l) + m\frac{Mel(f_h) - Mel(f_l)}{M+1} \right) \tag{3-4}$$

onde F_s é a frequência de amostragem em Hz, M é o número de filtros e N o número de amostras da DFT. $k[m]$ são as frequências digitais e Mel^{-1} determina a largura do banco de filtros e é dado por

$$Mel^{-1}(m) = 700 \left(\ e^{\frac{m}{1125}} - 1 \right) \tag{3-5}$$

– Em seguida, obtém-se a log-energia da saída de cada um dos filtros MEL.

$$\widehat{S}(m) = \ln\left(\sum_{m=0}^{\frac{N}{2}-1} S[k]H_m[k]\right), \qquad 1 < m < M \tag{3-6}$$

– Finalmente, os coeficientes MFCC são obtidos aplicando a transformada inversa do cosseno (DCT) ao logaritmo dos coeficientes de energia obtidos no item anterior

$$c[n] = \sum_{m=0}^{M-1} \widehat{S}[m]\cos\left(\frac{\pi n(m+0,5)}{M}\right), \qquad 0 < n < M-1 \tag{3-7}$$

Por exemplo, se M = 13 tem-se um vetor como é mostrado a seguir:

$$C_{mel} = c_0, c_1, c_2, ..., c_{12}.$$

O primeiro coeficiente do vetor C_{mel}, denotado por c_0, pode carregar muita informação do meio de transmissão [35]. Este coeficiente por vezes é considerado e por vezes não; isto vai depender do tipo de reconhecimento desejado, que pode ser reconhecimento de voz, ou reconhecimento de locutor.

A avantagem de utilizar DCT no lugar da IFFT (transformada inversa de Fourier),é que a DCT reduz o número de coeficientes gerados após utilizar as técnicas de parametrização especificadas (MFCC ou PNCC). Esta redução é feita através de uma propriedade da DCT conhecida como compactação da energia, concentrando os valores mais significativos nos primeiros termos do vetor, e descartando os últimos, melhorando assim a eficiência computacional.

– A ideia principal da extração de atributos é captar as mudanças temporais bruscas presentes no espectro. Devido a isto, utilizam-se além, dos coeficientes extraídos até agora, chamados coeficientes "estáticos", os coeficientes delta e de aceleração, chamados coeficientes "dinâmicos", que capturam essas mudanças e incorporam informação relativa à transição dos coeficientes estáticos entre quadros vizinhos.

O cálculo dos coeficientes dinâmicos faz-se através de regressão linear sobre uma janela, cobrindo dois vetores antes e dois após o vetor calculado [36], ou seja,

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{3-8}$$

onde d_t é o coeficiente (diferencial) delta (Δ) computado no tempo t, calculado em termos dos correspondentes coeficientes estáticos $C_{t-\theta}$ até $C_{t+\theta}$. O parâmetro Θ indica o tamanho da janela de regressão, usualmente igual a 2.

Os parâmetros de segunda ordem chamados delta-delta, são obtidos replicando a derivada sobre os resultados obtidos na primeira derivação [37].

3.2 Power-Normalized Cepstral Coefficients (PNCC)

Diferentes experimentos ao longo do tempo têm mostrado que os tons não são representados em escalas lineares de frequência. Por isso, tenta-se aproximar, através de escalas de frequências não lineares, o comportamento auditivo humano, tal como acontece com os atributos MFCC. Porém, nestes atributos a eficácia do reconhecimento cai rapidamente com a presença do ruído.

Recentemente, [38] introduziu um método mais eficiente para extração de atributos, chamado Power-Normalized Cepstral Coefficients. Sua eficiência é devida à adição de uma nova etapa de remoção de ruído, a qual, através da média das energias de uma banda ao longo de alguns quadros consecutivos, consegue remover a adição do ruído do sinal. Esse procedimento é feito após a divisão do sinal em bandas de frequência superpostas, similar ao utilizado nos MFCC. A diferença é o uso de um novo tipo de escala que imita a resposta em frequência do sistema auditivo. De acordo com isso, os atributos PNCC são considerados uma evolução dos atributos MFCC.

Os PNCC utilizam a mesma metodologia de análise de tempo curto que os MFCC, visando desenvolver conjuntos de atributos baseados em critérios perceptuais. A estrutura do método PNCC é mostrada na Fig 3.1, onde pode-se ver que é similar à estrutura MFCC descrita na seção anterior, com algumas variações, especialmente na etapa de *informações de espectro*.

O pré-processamento para a extração de atributos PNCC é o mesmo para os MFCC, que consiste em um filtro de pré-ênfase e a análise de Fourier utilizando o mesmo janelamento Hamming de 25 ms de comprimento com deslocamento entre janelas de 10 ms.

Uma vez obtida esta informação, procede-se à análise espectral constituída por três partes explicadas a seguir.

– A primeira parte consiste na utilização de filtros Gammatome baseados na escala de Bandas Retangulares Equivalentes (ERB) [39].

Esses filtros possuem bandas de passo não uniformes e sobrepostos, como é mostrado na Fig. 3.3, onde cada filtro representa a resposta em frequência relacionada com um ponto particular da membrana basilar [40].

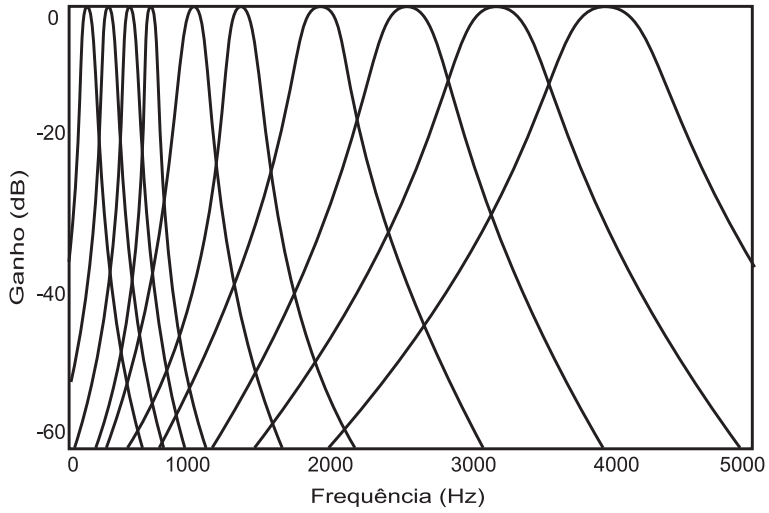

Figura 3.3: Banco de filtros Gammatone.

O modelo utilizado para a construção dos filtros é o proposto em [41]. A resposta ao impulso de cada filtro é dada por

$$g_t = t^{l-1} e^{-2\Pi t 1{,}019 ERB} \cos(2\pi f_c t) \qquad com \qquad t \geq 0 \qquad (3\text{-}9)$$

onde l é a ordem do filtro ERB e f_c é a frequência central associada a ela. Assim, a largura de banda de cada filtro é ajustada conforme as medidas da largura de ERB dos filtros auditivos humanos dados pela equação

$$ERB(f_c) = 24,7 \left(4,37 \frac{f_c}{1000} + 1 \right) Hz \qquad (3\text{-}10)$$

A largura da anda ERB corresponde aproximadamente 11% de sua frequência central, pelo que os filtros auditivos equivalentes têm uma largura de banda inferior à apresentadas pelas bandas criticas (20% de

f_c em 500 Hz). Devido a isso, é necessário uma maior quantidade de filtros *ERB*, a fim de caracterizar a faixa completa de frequências do sistema auditivo humano (de 20 Hz a 22.050 Hz)

Estas frequências centrais são distribuídas uniformemente em uma escala auditiva de frequências ERB 3-11, representadas por uma função quase logarítmica que relaciona a frequência com o número de canais do banco de filtros.

$$ERB_N = 21,4 \log_{10}(0,00437f + 1) \tag{3-11}$$

onde f é a frequência em Hz e ERB_N é o número ERB(razão ERB)

A implementação completa dos filtros Gammatone pode ser encontrada em [42].

– A segunda modificação é a implementação da etapa de remoção de ruído acima referida, afim de estimar a redução da qualidade da fala causada pelo ruído [43], já que este costuma ser mais estacionário que o sinal de voz.

  Este procedimento é motivado pelo fato de que o sistema auditivo humano é mais sensível a alterações na frequência ao longo do tempo, do que a excitação relativamente constante de fundo [38].

  A implementação detalhada é descrita em [44].

– A terceira modificação está relacionada com a mudança da função logarítmica utilizada na saída dos bancos MEL, por uma função de potenciação aplicada na operação não linear sobre a energia de cada banda. A utilização desta nova função tenta evitar que os valores de saída das bandas estejam perto de zero, já que as regiões onde o sinal possui pequenas energias serão mas vulneráveis à adição de ruído aditivo, como acontece com a função logarítmica alterando os atributos MFCC. É por isso que se utiliza uma função de potenciação, que vai crescer mais suave, reduzindo assim a distorção espectral.

Uma vez obtida a informação do espectro, o cepstrum em escala ERB é a transformada discreta do cosseno das saídas dos bancos de filtros, similar ao utilizado para os MFCC, (equação 3-7).

Finalmente, o vetor de atributos é constituído pelos n coeficientes determinados da DCT, além da adição dos correspondentes coeficientes de regressão (delta e delta-delta) obtidos através da equação 3-8.