4
Atributos para Reconhecimento de Voz Distribuído

Os esquemas de codificação de voz usados operam a baixas taxas de bits e utilizam, em geral, codificação preditiva linear (LPC – *Linear Predictive Coding*), com base em um modelo de produção da fala. Nesse modelo, um sinal de excitação é aplicado a um filtro só de pólos (caracterizado por parâmetros LPC), que representa a informação da envoltória espectral do sinal de voz. Usualmente os parâmetros LPC são transformados para LSF (*Line Spectral Frequencies*) ou ISF (*Immittance Spectral Frequencies*), devido às propriedades atraentes destes últimos para os processos de quantização e interpolação. No caso de sistemas de RAV distribuídos é preferível utilizar diretamente os parâmetros do *codec* do que extraí-los a partir do sinal decodificado [1]. A realização desse processamento envolve um grande número de aspectos e estratégias para concepção de reconhecedores de voz eficientes.

Para isso, conforme mencionado na caracterização do problema, diversos aspectos e estratégias deverão ser considerados. Primeiramente, os parâmetros LSF do *codec* não são necessariamente as melhores opções de atributos a serem usadas no RAV [1]. Portanto, transformações desses parâmetros são estratégias importantes a serem consideradas.

Neste capítulo serão apresentadas as deduções matemáticas dos atributos de reconhecimento de voz distribuído a serem utilizados nesta tese.

Na seção 4.1 deste capítulo será feita a apresentação dos atributos obtidos a partir dos parâmetros LPC do decodificador e na seção 4.2 serão apresentados os atributos obtidos dos parâmetros LSF. A seção 4.3 descreve os parâmetros ISF. A seção 4.4 é dedicada ao atributo mais amplamente usado, obtido da voz reconstruída pelos codificadores (MFCC - *Mel-Frequency Cepstral Coefficients*) . Finalmente, a seção 4.5 contém uma breve conclusão.

4.1.
Atributos Extraídos de LPCs

Nesta seção são analisados os parâmetros de reconhecimento que podem ser extraídos diretamente dos parâmetros LPC (*Linear Predictive Coefficients*), sem a necessidade de reconstrução do sinal de voz para obtenção dos atributos. Esta abordagem se deve ao fato de que, dentro dos decodificadores de voz utilizados para telefonia celular e voz sobre IP, já serem produzidos naturalmente, no seu processo de recuperação de voz, os parâmetros LPC, em um estágio anterior à reconstrução da voz. Sendo assim, parâmetros de reconhecimento de voz, obtidos neste estágio, são menos complexos computacionalmente do que os obtidos de voz reconstruída, pois evitam a necessidade de recuperação da mesma. Além disso, a MFCC de voz reconstruída apresenta resultado pior [6].

Primeiramente, será feita, na seção 4.1.1, uma apresentação matemática do método de obtenção dos parâmetros LPC que serão a base dos atributos de reconhecimento apresentados nas seções 4.1.2 e 4.1.3.

4.1.1.
Linear Predictive Coding (LPC)

A idéia básica da análise LPC consiste em uma amostra do sinal de fala ser modelada por uma combinação linear de suas p amostras passadas, dada por

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p) \tag{4.1}$$

onde os coeficientes $a_1, a_2, \ldots, a_p$ são recalculados para cada janela do sinal, pois, em pequenos trechos, o sinal pode ser assumido como sendo estacionário.

A equação anterior pode ser convertida em uma igualdade, incluindo um termo de excitação do sinal, $Gu(n)$, onde $u(n)$ é a excitação normalizada e G é o seu ganho.

$$s(n) = Gu(n) + \sum_{i=1}^{p} a_i s(n-i) \tag{4.2}$$

Isso nos leva a uma função de transferência do trato vocal

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4.3}$$

Os coeficientes $a_1, a_2, \ldots, a_p$ são os parâmetros LPC do sinal. Eles são calculados considerando-se a predição linear dada por

$$\tilde{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{4.4}$$

e o seu erro de predição é

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{4.5}$$

Os coeficientes são escolhidos a fim de minimizar uma função do erro de predição. Para isso, dentro de uma janela de sinal de tamanho N, o erro médio quadrático definido por

$$E_l = \sum_{n=0}^{N-1} (e(n))^2 = \sum_{n=0}^{N-1} \left(s(n) - \sum_{k=1}^{p} a_k s(n-k) \right)^2 \tag{4.6}$$

deve ser derivado em função de cada coeficiente a_i e igualado a zero

$$\frac{\partial E_l}{\partial a_i} = 0, \qquad i = 1,2,\ldots,p \tag{4.7}$$

onde l é o índice do segmento considerado.

Obtendo

$$\sum_{n=0}^{N-1} s(n-i)s(n) = \sum_{k=1}^{p} a_k \left(\sum_{n=0}^{N-1} s(n-i)s(n-k) \right), \qquad i = 1,2,\ldots,p \tag{4.8}$$

Pode-se definir os coeficientes de correlação como

$$\varphi_l(i,k) = \sum_{n=0}^{N-1} s(n-i)s(n-k) \tag{4.9}$$

e então

$$\sum_{k=1}^{p} a_k \varphi_l(i,k) = \varphi_l(i,0), \qquad i = 1,2,\ldots,p \tag{4.10}$$

A solução de p equações lineares resulta em p coeficientes LPC que minimizam o erro de predição. Com a_i satisfazendo a (4.10), o erro de predição total em (4.6) assume o seguinte valor

$$E_l = \sum_{n=0}^{N-1} s^2(n) - \sum_{k=1}^{p} a_k \sum_{n=0}^{N-1} s(n)s(n-k) = \varphi(0,0) - \sum_{k=1}^{p} a_k \varphi(0,k) \tag{4.11}$$

Com uma simples substituição de variáveis, (4.9) pode ser reescrita como

$$\varphi_l(i,k) = \sum_{n=-i}^{N-1-i} s(n)s(n+i-k) = \sum_{n=-k}^{N-1-k} s(n)s(n+k-i) \tag{4.12}$$

Como o sinal é processado em janelas de duração finita $(0 \le n \le N-1)$, sendo o sinal zero fora da janela, os limites do somatório podem ser alterados

$$\varphi_l(i,k) = \sum_{n=0}^{N-1-(i-k)} s(n)s(n+i-k) = \sum_{n=0}^{N-1-(k-i)} s(n)s(n+k-i) \equiv r(|i-k|) \tag{4.13}$$

Com a alteração dos limites do somatório em (4.13) temos a autocorrelação do sinal. Neste caso, a equação (4.10) torna-se

$$\sum_{k=1}^{p} a_k r(|i-k|) = r(i), \qquad i = 1,2,\ldots,p \tag{4.14}$$

Este é chamado de método da autocorrelação e é utilizado pelos codificadores aqui utilizados. O sistema de equações pode ser visto na sua forma matricial

$$\begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(p-1) \\ r(1) & r(0) & r(1) & \cdots & r(p-2) \\ r(2) & r(1) & r(0) & \cdots & r(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & r(p-3) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ a(3) \\ \vdots \\ a(p) \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ r(p) \end{bmatrix} \tag{4.15}$$

Como a matriz é do tipo Toeplitz [64-66], o melhor método para resolvê-la é utilizar o algoritmo de Levinson-Durbin [64-66], que também é utilizado pelos codificadores para resolver o sistema de equações, e é dado por

$$\begin{aligned} &\textit{Valores iniciais}: E^{(0)} = r(0), \qquad k_0 = 0 \\ &\textit{Iteração}: \quad 1 \le i \le p \\ &E^{(i)} = \left(1 - k_{i-1}^2\right) E^{(i-1)} \\ &k_i = \left\{ r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|) \right\} \Big/ E^{(i)} \\ &\alpha_i^{(i)} = k_i \\ &\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, \qquad 1 \le j < i \\ &\textit{Resultado}: \quad a_i = LPC \quad \textit{coeficients} = \alpha_i^{(p)}, \qquad 1 \le i \le p \end{aligned} \tag{4.16}$$

Os parâmetros de reconhecimento que podem ser obtidos dos parâmetros LPC são os parâmetros LPCC (*LPC Cepstrum*) e MLPCC (*Mel-Frequency* LPCC). Os parâmetros LPCC serão obtidos a partir dos parâmetros LPC por uma fórmula recursiva a ser deduzida na seção 4.1.2., sendo aqui apresentados, pois são a base da obtenção dos atributos MLPCC. Porém não serão utilizados em simulações, pois têm um pior desempenho no reconhecimento do que o MLPCC

[5]. Já os parâmetros MLPCC serão obtidos dos LPCCs através de uma rede de filtros passa-tudo a ser apresentada na seção 4.1.3.

4.1.2. LPC *Cepstrum* (LPCC)

O processo de obtenção dos parâmetros LPCC a partir dos coeficientes LPC será formulado no domínio da Transformada-Z, com o cálculo da resposta ao impulso do logaritmo complexo do sistema LPC, o que é análogo ao cálculo do Cepestro no domínio da Transformada Discreta de Fourier [5].

Primeiramente, se constrói a função de transferência do sistema LPC de ordem p, que é dada por

$$H(z) = \sum_{n=0}^{+\infty} h[n] z^{-n} = \frac{G}{A(Z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4.17}$$

onde a_i é o i-ésimo parâmetro LPC e G é o fator de ganho.

Calculando a derivada do polinômio complexo $\ln(H(z))$, em relação a $\rho = z^{-1}$, obtém-se

$$\frac{\partial}{\partial \rho} \ln(H(\rho)) = \frac{\partial}{\partial \rho} [\ln(G) - \ln(A(\rho))] = \frac{\sum_{l=1}^{p} l a_l \rho^{l-1}}{1 - \sum_{i=1}^{p} a_i \rho^{i}} \tag{4.18}$$

Como $H(z)$ é a função de transferência do sistema LPC obtido no transmissor, onde são utilizados métodos para garantir a estabilidade da função $H(z)$, a mesma deverá ter todos os seus pólos dentro do círculo unitário, então $\ln(H(z))$ é unilateral, o que leva a escrever

$$C(z) = \sum_{i=0}^{+\infty} c_i z^{-i} \tag{4.19}$$

onde c_i é o i-ésimo parâmetro LPCC e $C(z)$ é o logaritmo complexo da função de transferência do sistema LPC.

Derivando $C(z)$ em relação a ρ e igualando a (4.18), obtém-se a equação

$$\sum_{j=1}^{+\infty} jc_j \rho^{j-1} = \frac{\sum_{l=1}^{p} la_l \rho^{l-1}}{1 - \sum_{i=1}^{p} a_i \rho^i} \tag{4.20}$$

que pode ser reescrita na forma

$$\left(\sum_{j=1}^{+\infty} jc_j \rho^{j-1} \right) \left(1 - \sum_{i=1}^{p} a_i \rho^i \right) = \sum_{l=1}^{p} la_l \rho^{l-1} \tag{4.21}$$

Comparando os coeficientes das séries de ρ em ambos os lados, chega-se a uma equação recursiva que permite a obtenção dos parâmetros LPCC, onde o parâmetro c_0 é determinado pelo termo constante da definição original de $H(z)$. Essa equação é dada por

$$c_i = \begin{cases} \ln(G) & i = 0 \\ a_1 & i = 1 \\ a_i + \sum_{j=1}^{i-1} \frac{i-j}{i} c_{i-j} a_j & 1 < i \le p \\ \sum_{j=1}^{p} \frac{i-j}{i} c_{i-j} a_j & i > p \end{cases} \tag{4.22}$$

4.1.3. *Mel-Frequency* LPCC (MLPCC)

O processo de obtenção do parâmetro MLPCC passa pela transformação do eixo de frequência real para o eixo de frequência na escala mel dos parâmetros LPCC [4]. Para ser realizada esta transformação, utiliza-se um banco de n filtros passa-tudo de primeira ordem que permite efetuar a transformação do eixo de frequência real para o eixo de frequência na escala mel - onde n é o número de parâmetros LPCC obtidos através de (4.22) - [67]. Todos os filtros deste banco

terão sua função de transferência $\psi(z)$ passa-tudo de primeira ordem [68] dada pela expressão

$$\psi(z) = \frac{z^{-1} - a^*}{1 - az^{-1}} \tag{4.23}$$

devendo cada coeficiente cepestral c_i passar por um filtro diferente deste banco de filtros, onde a é o coeficiente deste filtro passa-tudo e a^* é o conjugado de a.

Como o objetivo de cada filtro é realizar a aproximação da escala mel de frequências, tem-se que analisar o que a função de transferência em (4.23) está realizando com os eixos das frequências. Para isto, será considerado a real, o que facilitará a implementação do filtro [69].

Para que seja feita a análise do que está sendo feito com os eixos de frequência, deve-se reescrever ψ, em função de $e^{j\Omega}$, como

$$\psi\left(e^{j\Omega}\right) = e^{-j\theta(\Omega)} \tag{4.24}$$

onde Ω é a frequência real e

$$\theta(\Omega) = \arctan\left[\frac{\left(1-a^2\right)\operatorname{sen}\Omega}{\left(1+a^2\right)\cos\Omega - 2a}\right] \tag{4.25}$$

é a frequência na escala mel expressa em função da frequência real Ω.

Ao se ajustar a curva de $\theta(\Omega)$ à curva da escala mel, para a frequência de amostragem de 8 kHz, por meio da variação do termo a real, obtém-se $a = 0{,}3624$ [69] e para a frequência de amostragem de 16 kHz, $a = 0{,}6$ [5].

As saídas do banco de filtros serão os parâmetros MLPCC.

4.2.
Atributos Extraídos de LSFs

As *Line spectral frequencies* (LSFs) são usualmente utilizadas para codificação de voz, devido à sua grande eficiência de codificação e suas propriedades atraentes para interpolação [70]. Porém, as LSFs não apresentam bom desempenho quando utilizadas como atributos para reconhecimento de voz [1].

Primeiramente, será desenvolvido matematicamente, na seção 4.2.1., o método de obtenção dos parâmetros LSF a partir dos parâmetros LPC, sendo este desenvolvimento justificado, pois as LSFs serão a base dos atributos de reconhecimento apresentados nas seções 4.2.2., 4.2.3., 4.2.4. e 4.2.5. e para que, quando da apresentação da obtenção dos parâmetros ISFs dos parâmetros LPC na seção 4.3., fique clara a semelhança entre os parâmetros ISF e LSF.

4.2.1.
Line Spectral Frequencies (LSF)

Os coeficientes LSF constituem uma das várias representações possíveis para os coeficientes de predição a_i do filtro de síntese utilizado na análise LPC.

Este filtro é definido por

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4.26}$$

onde $A(z)$ é o filtro inverso de ordem p.

Para o cálculo dos coeficientes LSF é necessário definir dois polinômios auxiliares $P(z)$ e $Q(z)$ obtidos a partir de $A(z)$ da seguinte forma [71]:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{4.27}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{4.28}$$

onde, $P(z)$ é um polinômio simétrico e $Q(z)$ é um polinômio antissimétrico.

As raízes de $P(z)$ e $Q(z)$ determinam os coeficientes LSF. Estes polinômios possuem ligação direta com o modelo acústico do trato vocal e com os estágios do filtro preditor com estrutura em treliça.

Se $A(z)$ é de fase mínima, ou seja, se $H(z)$ é estável, então:

1. As raízes de $P(z)$ e $Q(z)$ estão sobre o círculo unitário.
2. As raízes de $P(z)$ estão alternadas com as raízes de $Q(z)$, ou seja, $r_1 < q_1 < r_2 < q_2 < \ldots < q_{p+1}$, onde r_i e q_i representam a posição angular da i-ésima raiz de $P(z)$ e $Q(z)$, respectivamente.
3. O filtro $H(z)$ continuará estável após a quantização das raízes de $P(z)$ e $Q(z)$ desde que (1) e (2) sejam respeitados pelos valores quantizados.
4. Sendo $H(z)$ estável, a mesma permanecerá estável após a interpolação.

Além disso, os LSF possuem uma faixa dinâmica bem comportada, possibilitando uma quantização mais eficiente do que as outras formas de representar os coeficientes LPC.

Finalmente, os coeficientes de predição de $A(z)$ são obtidos a partir dos coeficientes de $P(z)$ e $Q(z)$ através da seguinte igualdade polinomial:

$$A(z) = \frac{P(z) + Q(z)}{2} \tag{4.29}$$

Os coeficientes LSF apresentam a importante propriedade de robustez à distorção. De acordo com ela, qualquer alteração sofrida por um desses coeficientes não terá um efeito global. Apenas será afetada a região do espectro próxima a esta frequência. Esta propriedade pode ser explorada em sistemas de codificação da voz, uma vez que o ouvido humano não é muito sensível a variações em frequências elevadas. Nesses sistemas, é possível representar os coeficientes LSF de elevadas frequências com um menor número de bits (o que é realizado normalmente pelos codificadores de voz), o que possibilita uma diminuição da taxa de bits do sistema.

De acordo com as propriedades acima, pode-se concluir que a utilização de coeficientes LSF apresenta vantagens em relação aos coeficientes LPC em termos de transmissão, quantização e interpolação. No entanto, o cálculo direto dos coeficientes LSF exige uma elevada capacidade computacional. A alternativa é calcular os coeficientes LPC e depois transformá-los em LSF como é feito nos codificadores aqui apresentados.

A obtenção de parâmetros de reconhecimento a partir das LSFs evita a necessidade de utilização de um decodificador de voz, ou da transformação para LPC, no receptor para a realização do reconhecimento. O sistema de reconhecimento de voz distribuído que evita tal utilização se torna mais leve computacionalmente que quaisquer outros baseados em parâmetros que dependam da reconstrução da voz ou dos parâmetros LPC. Os parâmetros de reconhecimento que podem ser obtidos desta forma (diretamente de LSFs) são os parâmetros PCC (*Pseudo-Cepstral Coefficients*), PCEP (*Pseudo-Cepstrum*), MPCC (Mel-Frequency PCC) e MPCEP (Mel-Frequency PCEP). Apenas os atributos MPCC e MPCEP serão utilizados nesta tese para obter os resultados de simulação, pois já foi demonstrado que os mesmos têm melhor desempenho que os atributos PCC, PCEP [72]. Porém a apresentação da dedução matemática dos mesmos se justifica por dela derivar a dedução dos atributos na escala mel.

Cabe ressaltar que estes parâmetros, obtidos diretamente de LSF, são aproximações da obtenção dos parâmetros LPCC e MLPCC, anteriormente apresentados. Estas aproximações têm como finalidade evitar a necessidade de recuperação dos parâmetros LPC, reduzindo a complexidade computacional do sistema e, ao mesmo tempo, buscando não perder o desempenho no reconhecimento.

4.2.2. Pseudo-Cepstral Coefficients (PCC)

O parâmetro PCC é obtido diretamente de LSF, porém a sua dedução passa pela obtenção do parâmetro LPCC a partir de LPC, com manipulações matemáticas e aproximações que permitem obtê-lo diretamente de LSF sem necessitar dos parâmetros LPC. Esses procedimentos serão apresentados em seguida.

Um filtro inverso de ordem p estável, onde todas as raízes se encontram dentro do círculo unitário, pode ser definido por

$$A_p(z) = \sum_{i=0}^{p} a_i z^{-i} \tag{4.30}$$

onde $a_0 = 1$ e a_i é o i-ésimo coeficiente de predição linear (LPCs).

As LSFs de ordem p são definidas como sendo as raízes complexas dos polinômios $P(z)$ e $Q(z)$, as quais são expressas por

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{4.31}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{4.32}$$

Para obter a relação entre LPCC e LSF é preciso realizar a multiplicação de (4.31) e (4.32), resultando em

$$P(z)Q(z) = A^2(z)\left[1 - R^2(z)\right] = (1 - z^{-2}) \prod_{i=1}^{p} \left(1 - e^{jw_i} z^{-1}\right)\left(1 - e^{-jw_i} z^{-1}\right) \tag{4.33}$$

para p par e maior que 2, onde w_i é o i-ésimo parâmetro LSF e

$$R(z) = \frac{z^{-(p+1)} A(z^{-1})}{A(z)} \tag{4.34}$$

e aplicando o logaritmo nos dois lados de (4.33) chega-se a

$$\begin{aligned} 2\log A_p(z) + \log\left(1 - R^2(z)\right) &= \log\left(1 - z^{-2}\right) \\ &+ \sum_{i=1}^{p} \left(\log\left(1 - e^{jw_i} z^{-1}\right) + \log\left(1 - e^{-jw_i} z^{-1}\right)\right) \end{aligned} \tag{4.35}$$

Fazendo, agora, a expansão em série em ambos os lados de (4.35), obtém-se

$$-2\sum_{n=1}^{\infty} c_n e^{-jwn} + \sum_{n=1}^{\infty} R_n e^{-jwn} = -\sum_{n=1}^{\infty} \frac{1}{n}\left(1+(-1)^n\right)e^{-jwn}$$
$$-\sum_{n=1}^{\infty} \frac{1}{n}\sum_{i=1}^{p}\left(e^{jnw_i} + e^{-jnw_i}\right)e^{-jwn} \qquad (4.36)$$

onde c_n é o n-ésimo parâmetro LPCC que satisfaz a relação

$$\log A_p(e^{jw}) = -\sum_{n=1}^{\infty} c_n e^{-jwn} \qquad (4.37)$$

e R_n é a transformada inversa de Fourier de $\log(1-R^2(z))$. Pode-se mostrar que a expansão dada pela equação (4.36) converge [4]. De (4.36) pode-se obter a partir de algumas manipulações matemáticas que

$$c_n = \frac{1}{2n}\left(1+(-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p}\cos nw_i + R_n \qquad (4.38)$$

Observando-se a equação (4.38), percebe-se que ainda existe o termo R_n que depende dos parâmetros LPC e que os demais só dependem das LSFs. Sendo assim, será desconsiderado este termo, dando origem à expressão do parâmetro PCC definido por

$$\hat{c}_n = \frac{1}{2n}\left(1+(-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p}\cos nw_i \qquad (4.39)$$

É razoável esperar que desprezar o fator $R_n/2$ não venha a prejudicar o desempenho, pois este fator será zero, ou assumirá valores muito pequenos, para a maioria dos casos [1].

4.2.3.
Pseudo-Cepstrum (PCEP)

Com base na dedução matemática dos parâmetros PCC, se torna bastante trivial a obtenção dos parâmetros PCEP. Esses parâmetros são obtidos a partir dos parâmetros PCC, eliminando-se o termo $\frac{1}{2n}\left(1+(-1)^n\right)$ que não depende da voz, ou seja, não depende dos parâmetros LSF. A expressão dos parâmetros PCEP é dada por

$$\hat{d}_n = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i \tag{4.40}$$

Pode-se esperar um bom desempenho espectral dos parâmetros PCEP, pois os mesmos fornecem uma envoltória espectral bastante parecida com a do Cepestro obtido diretamente de voz [1]. O PCEP possui a vantagem de apresentar ainda uma carga computacional mais baixa do que o parâmetro PCC obtido anteriormente.

4.2.4.
Mel-Frequency PCC (MPCC)

Para obter os parâmetros MPCC a partir dos parâmetros PCC basta manipular as LSFs a serem utilizadas em (4.39), onde w_i é substituído por w_i^m, definido pela transformação

$$w_i^m = w_i + 2\tan^{-1}\left(\frac{0{,}45\,sin\,w_i}{1-0{,}45\cos w_i}\right) \tag{4.41}$$

Essa equação consiste em uma forma de se transformar os eixos de frequência de um determinado conjunto de parâmetros nos eixos de frequência da escala mel [73]. Com esta alteração de eixo, obtém-se os parâmetros MPCC, dados pela expressão

$$\hat{c}_n^m = \frac{1}{2n}\left(1+(-1)^n\right)+\frac{1}{n}\sum_{i=1}^{p}\cos nw_i^m \tag{4.42}$$

onde $\hat{c}_n^m$ é o n-ésimo parâmetro MPCC.

4.2.5. Mel-Frequency PCEP (MPCEP)

Para se chegar aos parâmetros MPCEP, basta repetir o procedimento descrito para os parâmetros MPCC, obtendo a seguinte expressão

$$\hat{d}_n^m = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i^m \tag{4.43}$$

onde $\hat{d}_n^m$ é o n-ésimo parâmetro MPCEP.

4.3. Atributos Extraídos de ISFs

As *Immittance spectral frequencies* (ISFs) são utilizadas para codificação de voz no codificador AMR-WB, o codec recomendado para a terceira geração de telefonia celular que está sendo implantado em todo o mundo.

Primeiramente, na seção 4.3.1, será deduzido matematicamente o método de obtenção dos parâmetros ISF a partir dos parâmetros LPC, sendo esta apresentação justificada, pois as ISFs são os parâmetros do codificador utilizado na terceira geração de telefonia celular, sendo assim interessante se pesquisar também atributos que possam ser extraídos diretamente dos mesmos (o que será colocado nesta tese como proposta para futuros trabalhos).

4.3.1. *Immittance Spectral Frequencies* (ISF)

Uma outra representação dos parâmetros LPC são os parâmetros ISF que estão proximamente relacionados aos parâmetros LSF. O modelo ISF [60] é definido usando os polinômios $F(z)$ e $G(z)$ que são definidos como

$$F(z) = A(z) + z^{-p} A(z^{-1}) \tag{4.44}$$

$$G(z) = A(z) - z^{-p} A(z^{-1}) \tag{4.45}$$

onde $F(z)$ é um polinômio simétrico e $G(z)$ é um polinômio antissimétrico.

Note a similaridade entre os polinômios $F(z)$ e $G(z)$ definidos em (4.44) e (4.45) com os polinômios $P(z)$ e $Q(z)$ definidos em (4.27) e (4.28), estando a diferença no fator que multiplica $A(z^{-1})$. Nas LSF tem $z^{-(p+1)}$ e nos ISF tem z^{-p} como este fator. As raízes de $F(z)$ e $G(z)$ determinam os coeficientes ISF [42].

Se $A(z)$ é de fase mínima, ou seja, se $H(z)$ é estável, então:

1. As raízes de $F(z)$ e $G(z)$ estão sobre o círculo unitário.
2. As raízes de $F(z)$ estão alternadas com as raízes de $G(z)$, ou seja, $r_1 < q_1 < r_2 < q_2 < \ldots < q_{p+1}$, onde r_i e q_i representam a posição angular da i-ésima raiz de $F(z)$ e $G(z)$, respectivamente.
3. O filtro H(z) continuará estável após a quantização das raízes de $F(z)$ e $G(z)$ desde que (1) e (2) sejam respeitados pelos valores quantizados.
4. Sendo $H(z)$ estável, a mesma permanecerá estável após a interpolação.

Além disso, os ISF possuem uma faixa dinâmica bem comportada, possibilitando uma quantização mais eficiente, do que as outras formas de representar os coeficientes LPC, o que foi verificado em resultados experimentais em [60].

Finalmente, os coeficientes de predição de $A(z)$ são obtidos a partir dos coeficientes de $F(z)$ e $G(z)$ através da seguinte igualdade polinomial:

$$A(z) = \frac{F(z) + G(z)}{2} \tag{4.46}$$

Os coeficientes ISF apresentam a importante propriedade de robustez à distorção. De acordo com ela, qualquer alteração sofrida por um desses coeficientes não terá um efeito global. Apenas será afetada a região do espectro próxima a esta frequência. Assim como ocorre com as LSFs, esta propriedade pode ser explorada em sistemas de codificação da voz, uma vez que o ouvido humano não é muito sensível a variações em frequências elevadas. Nesses sistemas, é possível representar os coeficientes ISF de elevadas frequências com um menor número de bits (o que é realizado normalmente pelos codificadores de voz), o que possibilita uma diminuição da taxa de bits do sistema. Foi observado em [60] que quando se comparam ISFs às LSFs esta compressão é ainda maior, quando se mantem a qualidade de voz desejada depois da decodificação, ou seja, a ISF tem maior capacidade de armazenamento de informação e proteção da mesma.

De acordo com as propriedades acima, pode-se concluir que a utilização de coeficientes ISF apresenta vantagens em relação aos coeficientes LPC em termos de transmissão, quantização e interpolação. No entanto, o cálculo direto dos coeficientes ISF exige uma elevada capacidade computacional. A alternativa é calcular os coeficientes LPC e depois transformá-los em ISF como é feito no codificador AMR-WB.

Os coeficientes ISFs ainda têm como vantagem sobre as LSFs a redução de carga computacional, pois reduzem em um o número de raízes que precisam ser calculadas no processo de obtenção das ISFs, em detrimento das LSFs [60].

Será apresentado, no capítulo 7 desta tese, que a dedução matemática dos atributos de reconhecimento a partir dos ISFs é uma das propostas de trabalhos futuros.

4.4. Atributo Extraído de Voz Reconstruída (MFCC)

Nesta seção é considerado o atributo (MFCC - *Mel-Frequency Cepstral Coefficients*) que necessita ser obtido a partir de voz. No sistema aqui considerado, o atributo será obtido a partir da voz recuperada no decodificador localizado no receptor do sistema celular ou de voz sobre IP. Por esse motivo, ele foi classificado como atributo extraído de voz reconstruída.

Os coeficientes Mel-cepestrais surgiram devido aos estudos na área de psicoacústica (ciência que estuda a percepção auditiva humana), que mostraram que a percepção humana das frequências de tons puros ou de sinais de voz não segue uma escala linear. Isto estimulou a idéia de serem definidas frequências subjetivas de tons puros, da seguinte forma: para cada tom com frequência *f*, medida em Hz, define-se um tom subjetivo medido em uma escala que se chama escala mel. O mel, então, é uma unidade de medida da frequência percebida de um tom.

Como referência, definiu-se a frequência de 1 kHz, com potência 40 dB acima do limiar mínimo de audição do ouvido humano, como 1000 mels [66]. Os outros valores subjetivos foram obtidos através de experimentos, onde se pedia a ouvintes que ajustassem a frequência física de um tom, até que a frequência percebida fosse igual a duas vezes a frequência de referência; depois, 10 vezes a frequência de referência e assim por diante. Essas frequências teriam os valores de 2000 mels, 10000 mels e assim sucessivamente. O mesmo processo era efetuado na outra direção, ou seja, metade do tom de referência, um décimo do tom de referência, etc. Essas frequências teriam valores de 500 mels, 100 mels, etc. Isto permitiu verificar que o mapeamento entre a escala de frequência real, em Hz, e a escala de frequências percebida, em mel, é aproximadamente linear abaixo de 1000 Hz e, logarítmica, acima.

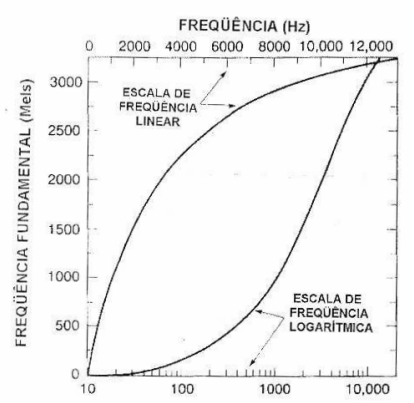

Figura 4.1 – Percepção subjetiva da frequência fundamental de sons sonoros

A Fig. 4.1 apresenta um gráfico da frequência fundamental subjetiva de tons em função da frequência [74]. A curva superior mostra a relação entre aquela e esta em uma escala linear. Pode-se observar que a frequência fundamental subjetiva, em mels, cresce menos e menos rapidamente à medida que há um

aumento linear na frequência. A curva inferior, por outro lado, mostra a frequência fundamental subjetiva em função da frequência em uma escala logarítmica. Pode-se notar na Fig. 4.1, que a frequência fundamental subjetiva é essencialmente linear para frequências inferiores a 1000 Hz.

Um outro importante critério subjetivo de conteúdo de frequência de um sinal é a banda crítica. Alguns experimentos demonstraram que a percepção humana de algumas frequências de sons complexos não pode ser individualmente identificada, dentro de certas bandas. Quando uma componente cai fora da banda, chamada de banda crítica, ela pode ser identificada. Uma explicação apresentada para esse fato foi que a percepção de uma frequência particular pelo sistema auditivo, por exemplo f_0, é influenciada pela energia da banda crítica das frequências em torno de f_0. O valor dessa banda varia nominalmente de 10 a 20 % da frequência central do som, começando em torno de 100 Hz para frequências abaixo de 1 kHz e aumentando em escala logarítmica, acima.

Esses fenômenos (escala mel e banda crítica) sugeriram que seria mais interessante fazer algumas modificações na representação e nas medidas de distâncias espectrais. Tais modificações consistiram, primeiramente, em fazer uma ponderação da escala de frequência para a escala mel e, além disso, incorporar a noção de banda crítica na definição de distorção espectral. Ou seja, ao invés de se usar simplesmente o logaritmo da magnitude das frequências, passou-se a utilizar o logaritmo da energia total das bandas críticas em torno das frequências mel. A aproximação mais utilizada para esse cálculo é a utilização de um banco de filtros triangulares, espaçados uniformemente em uma escala não linear (escala mel).

A técnica de ponderação mel pode ser aplicada a vários tipos de representação espectral. Cabe destaque a representação cepestral, devido à combinação da mesma com a técnica mencionada (mel), ser a mais utilizada e apresentar maior eficácia computacional, sendo chamada de Mel-Cepestral [66].

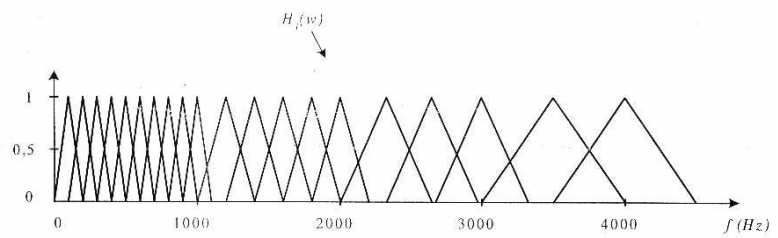

Figura 4.2 – Magnitude do espectro dos filtros de banda crítica

A Fig. 4.2 apresenta a configuração para o cálculo dos coeficientes Mel-Cepestrais. Para a faixa de frequências de interesse da voz humana, utilizam-se 20 filtros centrados nas frequências da escala mel. O espaçamento é de aproximadamente 150 mels e a largura de banda de cada filtro triangular é de 300 mels. Os valores dos centros são apresentados na Tab. 4.1. Como os valores calculados pela Transformada Rápida de Fourier (*Fast Fourier Transform – FFT*) são discretos, a tabela também mostra as aproximações para esses centros quando se utiliza FFT de 1024 pontos e frequência de amostragem de 8 kHz [68].

Filtro i	Centro Desejado (Hz)	Centro Aproximado (Hz)	Banda Crítica (Hz)
1	100	102	100
2	200	203	100
3	300	305	100
4	400	406	100
5	500	500	100
6	600	602	100
7	700	703	100
8	800	805	100
9	900	906	100
10	1000	1000	124
11	1148	1148	160
12	1318	1320	184
13	1514	1516	211
14	1737	1742	242
15	1995	2000	278
16	2291	2297	320
17	2630	2633	367
18	3020	3023	422
19	3467	3469	484
20	4000	4000	556

Tabela 4.1 – Frequências dos centros e banda crítica dos filtros utilizados para cálculo dos coeficientes mel-cepestrais

Inicialmente, divide-se o sinal de voz $s(n)$ em janelas. Para cada janela m estima-se o espectro $S(w,m)$, utilizando-se FFT, cujo espectro de magnitude é dado por

$$|S(w,m)| = (\mathrm{Re}[S(w,m)]^2 + \mathrm{Im}[S(w,m)]^2)^{1/2} \tag{4.47}$$

O espectro modificado $P(i), i = 1,2,\ldots,N_f$, consistirá na energia de saída de cada filtro, expresso por

$$P(i) = \sum_{k=0}^{N/2} |S(k,m)|^2 H_i\left(k\frac{2\pi}{N}\right) \tag{4.48}$$

onde N é o número de pontos da FFT, N_f é o número de filtros triangulares, $|S(k,m)|$ é o módulo da amplitude na frequência do k-ésimo ponto da m-ésima janela e $H_i(w)$ é a função de transferência do i-ésimo filtro triangular, definido por

$$H_i(w) = \begin{cases} \dfrac{1}{k_i - k_{i-1}}(w - k_{i-1}) & k_{i-1} \le w \le k_i \\ \dfrac{1}{k_i - k_{i+1}}(w - k_{i+1}) & k_i \le w \le k_{i+1} \end{cases} \tag{4.49}$$

onde, k_i é o i-ésimo centro, cujos valores estão mostrados na Tab. 4.1, $k_0 = 0$, e w é uma escala ajustada de acordo com o número de pontos da FFT, e expressa por

$$w = k\frac{2\pi}{N} \qquad 0 \le k \le N/2 \tag{4.50}$$

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

$$E(k) = \begin{cases} \log[P(i)] & k = k_i \\ 0 & qq\ outro\ k \in [0, N-1] \end{cases} \tag{4.51}$$

Os coeficientes mel-cepestrais $c_{mel}(n)$ são então obtidos com o uso da Transformada Inversa de Fourier (IFFT), usando-se a seguinte equação:

$$c_{mel}(n) = \frac{1}{N}\sum_{k=0}^{N-1} E(k) e^{j\left(\frac{2\pi}{N}\right)kn} \qquad n = 1,2,\ldots,N_c \tag{4.52}$$

onde N_c é o número de coeficientes desejado.

Como $E(k)$ é simétrico em relação a $N/2$ (ou $\pi/2$) e lembrando que

$$e^{j\left(\frac{2\pi}{N}\right)kn} = \cos\left(\frac{2\pi}{N}kn\right) + j\,\mathrm{sen}\left(\frac{2\pi}{N}kn\right) \tag{4.53}$$

resulta que os termos em seno da (4.52) se cancelam, gerando a equação

$$c_{mel}(n) = \frac{1}{N}\sum_{k=0}^{N-1} E(k)\cos\left(\frac{2\pi}{N}kn\right) \tag{4.54}$$

Ainda usando a simetria e observando que

$$E(0) = E(N/2) \tag{4.55}$$

obtém-se a expressão

$$c_{mel}(n) = \frac{2}{N}\sum_{k=1}^{\frac{N}{2}-1} E(k)\cos\left(\frac{2\pi}{N}kn\right) \tag{4.56}$$

Sabendo-se que no intervalo $0 \le k \le (N-2)/2$ existirão apenas N_f termos diferentes de zero, que são os correspondentes aos centros dos filtros, e

eliminando-se o fator de escala $2/N$, a equação (4.56) pode ser simplificada, chegando-se à expressão final para os coeficientes MFCC, dado por

$$c_{mel}(n) = \sum_{i=1}^{N_f} E(k_i) \cos\left(\frac{2\pi}{N} k_i n\right) \qquad n = 1,2,\ldots,N_c \qquad (4.57)$$

onde N_c é o número de coeficientes mel-cepestrais desejado, N_f é o número de filtros e k_i é o centro do i-ésimo filtro.

4.5. Conclusão

Neste capítulo foram apresentados a base teórica e os parâmetros/atributos que serão utilizados para a implementação do sistema de reconhecimento de voz distribuído no ambiente celular/voz sobre IP desta tese.

O capítulo seguinte descreve uma nova técnica de interpolação de parâmetros que visa melhorar o desempenho do reconhecedor de voz distribuída quando comparada com a interpolação linear.