4
Atributos de Voz para Reconhecimento

Os sistemas de reconhecimento de voz distribuído podem ser divididos em dois blocos principais: pré-processamento e obtenção dos atributos de reconhecimento.

O bloco de pré-processamento é composto por:

- Microfone, responsável pela conversão do sinal acústico em um sinal elétrico analógico;
- Conversor A/D, responsável pela conversão analógica do sinal elétrico gerado pelo microfone;
- *Endpoints,* a estimação dos pontos terminais;
- *Codec* de voz, que será responsável pela codificação paramétrica do sinal de voz para o caso de telefonia celular e de voz sobre IP, que será transmitido ao sistema remoto.

As funções do pré-processamento e suas principais características serão brevemente apresentadas na Seção 4.1.

Na obtenção dos atributos de reconhecimento é feita a manipulação dos parâmetros recebidos no sistema remoto, com a finalidade de obter novos parâmetros que possuam desempenho superior aos parâmetros transmitidos pelo *front-end* local.

Será dada atenção especial na obtenção dos atributos de voz para reconhecimento, pois é de extrema importância encontrar atributos de reconhecimento de grande simplicidade computacional. Os atributos de reconhecimento serão apresentados nas Seções 4.2, 4.3 e 4.4, de acordo com o parâmetro original utilizado para sua obtenção ou se obtido diretamente de voz reconstruída.

Cabe-se ressaltar que os atributos de reconhecimento obtidos dos parâmetros LPC e LSF são de maior simplicidade quando comparados com os obtidos de voz reconstruída, representando assim uma menor carga computacional

ao sistema, fator este de grande importância para sistemas de reconhecimento distribuídos.

4.1. Pré-Processamento

Para se ter um bom desempenho do reconhecedor de voz, é de extrema importância que sejam determinados de forma eficiente e precisa, o início e o final de uma locução, com a finalidade de excluir os silêncios que não trazem nenhuma informação adicional sobre a locução a ser reconhecida.

Num sistema típico de reconhecimento automático de voz, esta é, inicialmente, amostrada a uma taxa que varia de 6,67 a 16 kHz [4]. Essas taxas são compatíveis com os conversores analógicos digitais utilizados em sistemas celulares e de voz sobre IP, onde a taxa de amostragem é de 8 kHz.

O sinal de voz amostrado, $s(n)$, atravessa, em seguida, um sistema digital de baixa ordem (pré-ênfase). O objetivo da pré-ênfase é atenuar as componentes de baixa freqüência do sinal, prevenindo contra instabilidade numérica e, também, minimizando o efeito dos lábios e da glote [22].

A função de transferência de pré-ênfase mais largamente empregada [23] consiste de um sistema de primeira ordem fixo, cuja função de transferência é dada por

$$H(z) = 1 - \tilde{a}z^{-1}, \qquad 0{,}9 \le \tilde{a} \le 1{,}0 \tag{4.1}$$

Neste caso, a saída da pré-ênfase, $\tilde{s}(n)$, está relacionada à entrada, $s(n)$, pela equação diferença

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \tag{4.2}$$

O valor mais comum de $\tilde{a}$ é aproximadamente 0,95.

É importante destacar que os *codecs* usados em telefonia celular e voz sobre IP possuem este filtro de pré-ênfase em seu sistema de codificação de voz. Logo, a necessidade do mesmo no bloco de determinação de *endpoints,* necessário ao

sistema de reconhecimento, não representa sobrecarga de processamento para o transmissor.

Após a pré-ênfase, passa-se à etapa da "blocagem" do sinal de voz. Nesta etapa são extraídos quadros de N amostras a partir do sinal $\tilde{s}(n)$, sendo os quadros adjacentes separados por M amostras. Tal divisão é extremamente importante devido ao fato de um sinal de fala ser estatisticamente variante no tempo. A divisão em pequenos segmentos, que variam de 10 a 45 ms, possibilita admitir que ele seja aproximadamente estacionário nesses intervalos [1].

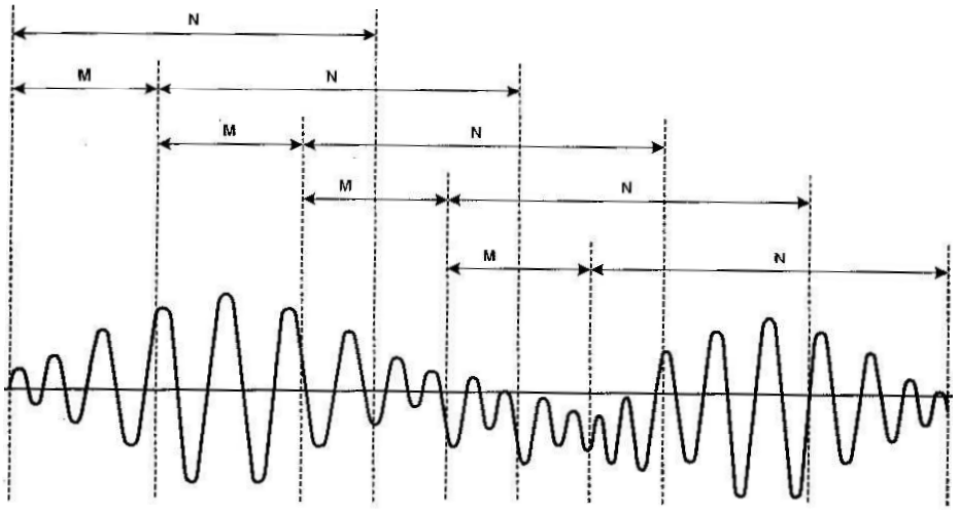

Figura 4.1 – Divisão em quadros do sinal de voz.

Ressalta-se que, se $M < N$, então os quadros adjacentes sobrepõem-se, como na Fig. 4.1, e os valores estimados serão correlatados de quadro para quadro; se $M \langle\langle N$, então as estimativas de um quadro para o outro serão bem suaves. Por outro lado, se $M \geq N$, não há sobreposição entre quadros adjacentes; na verdade, parte do sinal será totalmente perdida (ou seja, não aparecerá em nenhum quadro de análise), e a correlação entre os valores estimados dos quadros adjacentes conterá uma componente ruidosa, cuja amplitude é diretamente proporcional a M. Esta situação é inaceitável em qualquer análise para reconhecimento da fala. Denotando-se por $x_l(n)$ o l-ésimo quadro da fala, e o sinal completo apresente L quadros, então tem-se

$$x_l(n) = \tilde{s}(Ml + n), \qquad n = 0,1,\ldots,N-1 \qquad l = 0,1,\ldots,L-1 \qquad (4.3)$$

No caso dos codificadores de voz para telefonia celular e voz sobre IP, este processo de "blocagem" também é realizado pelo codificador. Sendo assim, sempre se utilizará valores de N e M para o processo de determinação de *endpoints* idênticos ao do codificador, para que isto não provoque sobrecarga de memória e de processamento para o transmissor.

Em seguida, a voz é processada para detecção dos pontos terminais (*endpoints*). O sucesso na determinação dos limites de uma palavra representa um aspecto crucial na perfomance de um sistema de reconhecimento de palavras isoladas. De fato, trata-se de um problema extremamente complicado, particularmente para palavras que começam ou acabam em fonemas de baixa energia, como fricativas ou nasais ou, ainda, palavras que possuem oclusivas, pois o silêncio que precede a oclusiva pode ser confundido com o fim da palavra.

Os *endpoints* são determinados pelo primeiro quadro onde o sinal de voz realmente se inicia e pelo último quadro do sinal de voz. Eles são importantes porque evitam o processamento dos segmentos onde não há voz, antes e depois do sinal com voz, evitando carga computacional e economizando tempo, além de servir como marco de início e fim de um segmento de voz [4]. Estes fatores são de grande importância, pois minimizam a carga do reconhecedor automático de voz, visto que o mesmo não terá que processar atributos de reconhecimento de trechos sem informação de voz.

A determinação dos *endpoints* deve ser feita de forma cuidadosa, pois os mínimos erros nesta estimação podem degradar o reconhecimento, como apresentado por Santos [24].

A determinação dos *endpoints* é realizada através de um classificador de voz que pode diferenciar entre sons sonoros, surdos ou silêncio. Neste trabalho foi utilizado um classificador baseado nas características temporais do sinal [22].

Um dos algoritmos mais simples apresentado por Sotomayor [22] é baseado nas seguintes características:

- Energia do sinal;
- Taxa de cruzamentos por zero.

Um dos algoritmos mais empregados é o proposto por Rabiner [25]. Nesta dissertação foi utilizada uma modificação deste algoritmo, proposta por Lima [26], e que pode ser resumida como segue.

O algoritmo é baseado na estimação da amplitude média do sinal. Os 100 ms iniciais e 30 ms finais da locução são considerados como ruído de fundo. Desta estimação inicial e final, é calculada a média e o desvio padrão. Um limiar do valor da amplitude média é estipulado e todo sinal abaixo deste limiar é considerado como ruído de fundo. Considera-se sempre três quadros adjacentes para evitar ruídos espúrios.

O inconveniente deste método é que ele necessita ter no começo e no final do sinal somente ruído, para daí extrair as características estatísticas dele. Sua vantagem é seu baixo custo computacional.

Os passos do algoritmo são:

1. Cálculo da amplitude média e desvio padrão das primeiras e últimas janelas da voz;
2. Com a média e o desvio padrão do passo anterior, se define o limiar: limiar = média + 0,5 do desvio padrão;
3. Comparação das amplitudes médias de três janelas consecutivas com o limiar estimado: se as amplitudes médias das três janelas estiverem acima do limiar, a primeira janela é marcada como trecho de voz; ao contrário, se as amplitudes médias de três janelas consecutivas não estiverem acima do limiar, a primeira janela é marcada como trecho de silêncio. Este processo é realizado até se achar a primeira janela que possa ser marcada como voz (início da locução).
4. O passo anterior é repetido varrendo-se o sinal de trás para frente para determinar o fim da locução.

Com este algoritmo, determinou-se o final e início da locução a ser processada pelo sistema de reconhecimento, ou seja, os *endpoints*.

4.2.
Atributos Extraídos de Voz Reconstruída

Nesta Seção serão considerados os parâmetros que necessitam ser obtidos a partir de voz original. No sistema aqui considerado, porém, os parâmetros serão obtidos a partir da voz recuperada no decodificador localizado no receptor do sistema celular ou de voz sobre IP. Por esse motivo, eles foram classificados como atributos extraídos de voz reconstruída.

Pode-se citar como atributos obtidos de voz, os coeficientes cepestrais (CC - *Cepstral Coefficient*) [24], os Mel-Cepestrais (MFCC - *Mel-Frequency Cepstral Coefficient*), os PLP (*Perceptual Linear Predictive*) [24] e os parâmetros ZCPA (*Zero-Crossings with Peak Amplitudes*) [27], dentre outros.

Nesta dissertação será considerada apenas a MFCC, obtida tanto de voz original como de voz reconstruída. Será apresentado aqui o desenvolvimento matemático deste parâmetro, deixando para o leitor as referências para obtenção dos demais parâmetros. A escolha da MFCC se deve ao fato da mesma ter elevado desempenho no cenário de reconhecimento de voz, cuja vasta gama de resultados servem como referência para guiar e apoiar as conclusões desta dissertação.

4.2.1.
Coeficientes Mel-Cepestrais (MFCC)

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 freqüências de tons puros ou de sinais de voz, não seguem uma escala linear. Isto estimulou a idéia de serem definidas freqüências subjetivas de tons puros, da seguinte forma: para cada tom com freqüência 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 freqüência percebida de um tom.

Como referência, definiu-se a freqüência de 1 kHz, com potência 40 dB acima do limiar mínimo de audição do ouvido humano, como 1000 mels [1]. Os outros valores subjetivos foram obtidos através de experimentos, onde pedia-se a ouvintes que ajustassem a freqüência física de um tom, até que a freqüência percebida fosse igual a duas vezes a freqüência de referência; depois, 10 vezes a

freqüência de referência e assim por diante. Essas freqüências 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 freqüências teriam valores de 500 mels, 100 mels, etc. Isto permitiu verificar que o mapeamento entre a escala de freqüência real, em Hz, e a escala de freqüências percebida, em mel, é aproximadamente linear abaixo de 1000 Hz e, logarítmica, acima.

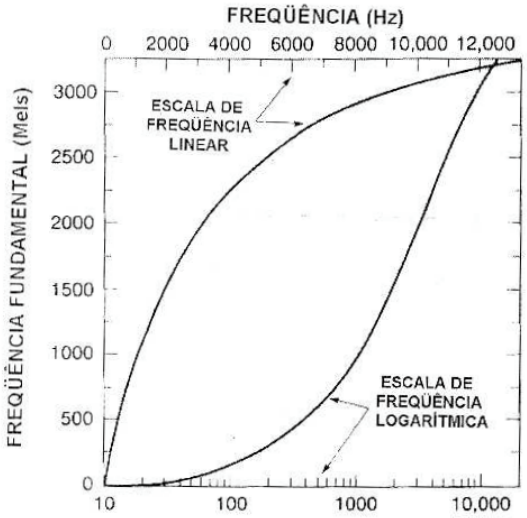

Figura 4.2 – Percepção subjetiva da freqüência fundamental de sons sonoros.

A Fig. 4.2 apresenta um gráfico da freqüência fundamental subjetiva de tons em função da freqüência [28]. A curva superior mostra a relação entre aquela e esta em uma escala linear. Pode-se observar que a freqüência fundamental subjetiva, em mels, cresce menos e menos rapidamente à medida que há um aumento linear na freqüência. A curva inferior, por outro lado, mostra a freqüência fundamental subjetiva em função da freqüência em uma escala logarítmica. Pode-se notar na Fig. 4.2, que a freqüência fundamental subjetiva é essencialmente linear para freqüências inferiores a 1000 Hz.

Um outro importante critério subjetivo de conteúdo de freqüência de um sinal é a banda crítica. Alguns experimentos demonstraram que a percepção humana de algumas freqüências de sons complexos não podem ser individualmente identificadas, 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 freqüência

particular pelo sistema auditivo, por exemplo f_0, é influenciada pela energia da banda crítica das freqüências em torno de f_0. O valor dessa banda varia nominalmente de 10 a 20 % da freqüência central do som, começando em torno de 100 Hz para freqüências 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 freqüência 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 freqüências, passou-se a utilizar o logaritmo da energia total das bandas críticas em torno das freqüências 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 [1].

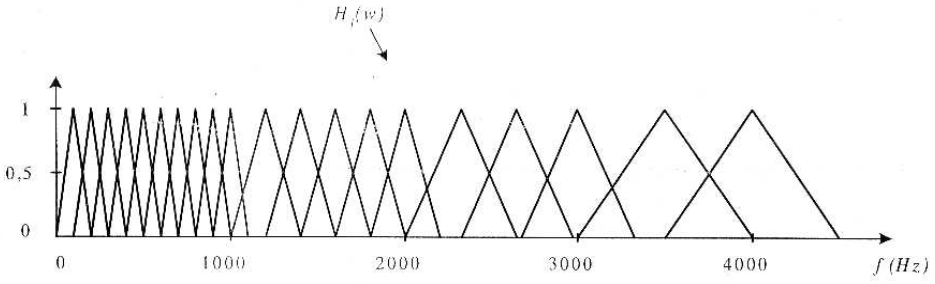

Figura 4.3 – Magnitude do espectro dos filtros de banda crítica.

A Fig. 4.3 apresenta a configuração para o cálculo dos coeficientes Mel-Cepestrais. Para a faixa de freqüências de interesse da voz humana, utilizam-se 20 filtros centrados nas freqüências 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 freqüência de amostragem de 8 kHz [21].

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 – Freqüências 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

$$|S(w,m)| = (\mathrm{Re}[S(w,m)]^2 + \mathrm{Im}[S(w,m)]^2)^{1/2} \qquad (4.4)$$

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

$$P(i) = \sum_{k=0}^{N/2} \left|S(k,m)\right|^2 H_i\left(k\frac{2\pi}{N}\right) \tag{4.5}$$

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

$$H_i(w) = \begin{cases} \dfrac{1}{k_i - k_{i-1}}(w - k_{i-1}) & k_{i-1} \le w \le k_i \\ \dfrac{1}{k_i - k_{i+1}}(w - k_{i+1}) & k_i \le w \le k_{i+1} \end{cases} \tag{4.6}$$

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

$$w = k\frac{2\pi}{N} \qquad 0 \le k \le N/2 \tag{4.7}$$

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

$$E(k) = \begin{cases} \log[P(i)] & k = k_i \\ 0 & qq\ outro\ k \in [0, N-1] \end{cases} \tag{4.8}$$

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

$$c_{mel}(n) = \frac{1}{N}\sum_{k=0}^{N-1} E(k) e^{j\left(\frac{2\pi}{N}\right)kn} \qquad n = 1,2,\ldots,N_c \tag{4.9}$$

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

$$e^{j\left(\frac{2\pi}{N}\right)kn} = \cos\left(\frac{2\pi}{N}kn\right) + j\,\mathrm{sen}\left(\frac{2\pi}{N}kn\right) \tag{4.10}$$

resulta que os termos em seno da (4.9) se cancelam, gerando a equação

$$c_{mel}(n) = \frac{1}{N}\sum_{k=0}^{N-1} E(k)\cos\left(\frac{2\pi}{N}kn\right) \tag{4.11}$$

Ainda usando a simetria e observando que

$$E(0) = E\left(N/2\right) \tag{4.12}$$

obtém-se a expressão

$$c_{mel}(n) = \frac{2}{N}\sum_{k=1}^{\frac{N}{2}-1} E(k)\cos\left(\frac{2\pi}{N}kn\right) \tag{4.13}$$

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 (4.13) pode ser simplificada, chegando-se à expressão final para os coeficientes MFCC, dado por

$$c_{mel}(n) = \sum_{i=1}^{N_f} E(k_i) \cos\left(\frac{2\pi}{N} k_i n\right) \qquad n = 1,2,\ldots,N_c \qquad (4.14)$$

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.

4.3. Atributos Extraídos dos Parâmetros LPC

Nesta seção será feita a análise dos 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.

Como os parâmetros LPC são obtidos pelo decodificador de voz no receptor e pelo fato do sistema proposto nesta dissertação não estar alterando esta etapa de decodificação, não será exposta aqui a obtenção dos parâmetros LPC.

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 4.3.1. Já os parâmetros MLPCC serão obtidos dos LPCCs através de uma filtragem passa-tudo de primeira ordem a ser apresentada na Seção 4.3.2.

4.3.1.
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 [29].

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

$$H(z) = \sum_{n=0}^{+\infty} h[n] z^{-n} = \frac{G}{A(Z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4.15}$$

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

$$\frac{\partial}{\partial \rho} \ln(H(\rho)) = \frac{\partial}{\partial \rho} [\ln(G) - \ln(A(\rho))] = \frac{\sum_{l=1}^{p} l a_l \rho^{l-1}}{1 - \sum_{i=1}^{p} a_i \rho^{i}} \tag{4.16}$$

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

$$C(z) = \sum_{i=0}^{+\infty} c_i z^{-i} \tag{4.17}$$

onde c_i é o i-ésimo parâmetro LPCC e $C(z)$ é resposta ao impulso do logaritmo complexo do sistema LPC.

Derivando $C(z)$ em relação a ρ e igualando a (4.16), obtém-se a equação

$$\sum_{j=1}^{+\infty} jc_j \rho^{j-1} = \frac{\sum_{l=1}^{p} la_l \rho^{l-1}}{1 - \sum_{i=1}^{p} a_i \rho^i} \tag{4.18}$$

que pode ser reescrita na forma

$$\left(\sum_{j=1}^{+\infty} jc_j \rho^{j-1} \right) \left(1 - \sum_{i=1}^{p} a_i \rho^i \right) = \sum_{l=1}^{p} la_l \rho^{l-1} \tag{4.19}$$

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

$$c_i = \begin{cases} \ln(G) & i = 0 \\ a_1 & i = 1 \\ a_i + \sum_{j=1}^{i-1} \frac{i-j}{i} c_{i-j} a_j & 1 < i \leq p \\ \sum_{j=1}^{p} \frac{i-j}{i} c_{i-j} a_j & i > p \end{cases} \tag{4.20}$$

4.3.2. *Mel-Frequency* LPCC (MLPCC)

O processo de obtenção do parâmetro MLPCC passa pela transformação do eixo de freqüência real para o eixo de freqüência na escala mel dos parâmetros LPCC [15]. 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 freqüência real para o eixo de freqüência na escala mel - onde n é o número de parâmetros LPCC obtidos através de (4.20) - [30]. Todos os filtros deste banco terão sua função de transferência $\psi(z)$ passa-tudo de primeira ordem [21] dada pela expressão

$$\psi(z) = \frac{z^{-1} - a^*}{1 - az^{-1}} \tag{4.21}$$

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 freqüências, tem-se que analisar o que a função de transferência em (4.21) está realizando com os eixos das freqüências. Para isto, será considerado a real, o que facilitará a implementação do filtro [31].

Para que seja feita esta análise, deve-se reescrever ψ, em função de $e^{j\Omega}$, como

$$\psi\left(e^{j\Omega}\right) = e^{-j\theta(\Omega)} \tag{4.22}$$

pois isto permite analisar o que está sendo feito com os eixos de freqüência, onde Ω é a freqüência real e

$$\theta(\Omega) = \arctan\left[\frac{\left(1-a^2\right)\operatorname{sen}\Omega}{\left(1+a^2\right)\cos\Omega - 2a}\right] \tag{4.23}$$

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

Ao se ajustar a curva de $\theta(\Omega)$ à curva da escala mel, para a freqüência de amostragem de 8 kHz, por meio da variação do termo a, obtém-se a curva da Fig. 4.4 [31].

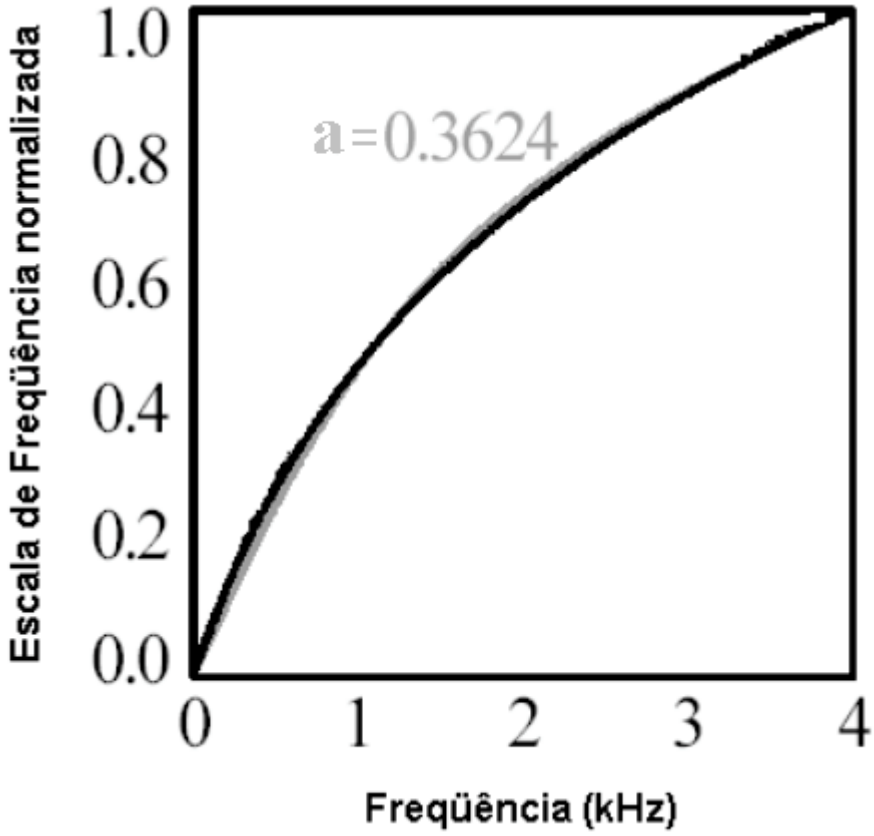

Figura 4.4 – Aproximação da escala mel pela transformação bilinear, escala mel em preto e aproximação em cinza.

Note que o valor de a real que melhor aproxima a curva da escala mel da Fig. 4.2 é 0,3624 [31], como mostrado na Fig. 4.4.

As saídas do banco de filtros serão os parâmetros MLPCC.

4.4. Atributos Extraídos das 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 [32].

Os padrões de codificadores para redes IP e ambientes móveis celulares são enumerados abaixo, colocando-se na seguinte ordem as informações – nome do padrão / tipo do *codec* – ano – taxa – tipo de rede em que é utilizado:

- ITU-T G.723.1 / CELP – 1995 – 5,3 ou 6,3 kb/s – IP
- ITU-T G.729 / CS-ACELP – 1995 – 8 kb/s – IP
- TIA – IS-54 / VSELP - 1992 – 8 kb/s – Cel USA TDMA

- VSELP – Japão – 1993 – 6,7 kb/s – Cel Japonês TDMA
- TIA – IS-95 / QCELP – 1993 – 1 / 2 / 4 ou 8 kb/s – Cel USA CDMA
- TIA – IS-96-A / QCELP – 1995 – 1,2 / 2,4 / 4,8 / 9,6 kb/s – Cel USA CDMA
- GSM – HR (Half Rate) / VSELP – 1995 – 5,6 kb/s – Cel Europeu TDMA
- GSM – EFR (Enhanced Full Rate) / ACELP – 1997 – 12,2 kb/s – Cel Europeu TDMA
- TIA – IS-641 (substitui o IS-54) / ACELP – 1997 – 7,4 kb/s – Cel USA TDMA
- TIA – IS-733 / QCELP – 1998 – 1,8 / 3,6 / 7,8 / 14,4 kb/s – Cel USA CDMA
- TIA – IS-127 EVRC (Enhanced Variable Rate Coder) / ACELP – 1998 – 1,2 / 4,8 / 9,6 kb/s – Cel USA CDMA
- AMR-WB (Adaptive Multi Rate – Wide Band) / ACELP – 2001 – 6,6 / 8,85 / 12,65 / 14,25 / 15,85 / 18,25 / 19,85 / 23,05 / 23,85 kb/s – Cel 3G Europa, Japão, USA, Coréia – WCDMA

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 são os parâmetros PCC (*Pseudo-Cepstral Coefficients*), PCEP (*Pseudo-Cepstrum*), MPCC (Mel-Frequency PCC) e MPCEP (Mel-Frequency PCEP).

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.

4.4.1.
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, é definido por

$$A_p(z) = \sum_{i=0}^{p} a_i z^{-i} \tag{4.24}$$

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 expressos por

$$P(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \tag{4.25}$$

$$Q(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \tag{4.26}$$

Para obter a relação entre LPCC e LSF é preciso realizar a multiplicação de (4.25) e (4.26), resultando em

$$P(z)Q(z) = A_p^2(z)\left[1 - R^2(z)\right] = (1 - z^{-2}) \prod_{i=1}^{p} \left(1 - e^{jw_i} z^{-1}\right)\left(1 - e^{-jw_i} z^{-1}\right) \tag{4.27}$$

onde w_i é o i-ésimo parâmetro LSF. Definindo

$$R(z) = \frac{z^{-(p+1)} A_p(z^{-1})}{A_p(z)} \tag{4.28}$$

e aplicando o logaritmo nos dois lados de (4.27) chega-se a

$$2\log A_p(z) + \log\left(1 - R^2(z)\right) = \log\left(1 - z^{-2}\right) + \sum_{i=1}^{p}\left(\log\left(1 - e^{jw_i} z^{-1}\right) + \log\left(1 - e^{-jw_i} z^{-1}\right)\right) \tag{4.29}$$

Fazendo, agora, a expansão em série de *Fourier* em ambos os lados de (4.29), obtém-se

$$-2\sum_{n=1}^{\infty} c_n e^{-jwn} + \sum_{n=1}^{\infty} R_n e^{-jwn} = -\sum_{n=1}^{\infty} \frac{1}{n}\left(1 + (-1)^n\right) e^{-jwn} - \sum_{n=1}^{\infty} \frac{1}{n} \sum_{i=1}^{p}\left(e^{jnw_i} + e^{-jnw_i}\right) e^{jwn} \tag{4.30}$$

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

$$\log A_p(e^{jw}) = -\sum_{n=1}^{\infty} c_n e^{-jwn} \tag{4.31}$$

e R_n é a transformada inversa de fourier de $\log(1 - R^2(z))$. Pode-se mostrar que a expansão dada pela equação (4.30) converge [15]. De (4.30) pode-se concluir que

$$c_n = \frac{1}{2n}\left(1 + (-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p} \cos nw_i + \frac{R_n}{2} \tag{4.32}$$

Observando-se a equação (4.32), percebe-se que ainda existe o termo $R_n/2$ 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

$$\hat{c}_n = \frac{1}{2n}\left(1 + (-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p} \cos nw_i \tag{4.33}$$

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 [33].

4.4.2. 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

$$\hat{d}_n = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i \tag{4.34}$$

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 [33]. O PCEP possui a vantagem de apresentar ainda uma carga computacional mais baixa do que o parâmetro PCC obtido anteriormente.

4.4.3. Mel-Frequency PCC (MPCC)

Para obter os parâmetros MPCC a partir dos parâmetros PCC basta manipular as LSFs a serem utilizadas em (4.33), onde w_i é substituído por w_i^m, definido pela transformação

$$w_i^m = w_i + 2\tan^{-1}\left(\frac{0{,}45\,sin\,w_i}{1-0{,}45\cos w_i}\right) \tag{4.35}$$

Essa equação consiste em uma forma de se transformar os eixos de freqüência de um determinado conjunto de parâmetros nos eixos de freqüência da escala mel [34]. Com esta alteração de eixo, obtém-se os parâmetros MPCC, dados pela expressão

$$\hat{c}_n^m = \frac{1}{2n}\left(1+(-1)^n\right)+\frac{1}{n}\sum_{i=1}^{p}\cos nw_i^m \tag{4.36}$$

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

4.4.4. Mel-Frequency PCEP (MPCEP)

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

$$\hat{d}_n^m = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i^m \tag{4.37}$$

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

4.5. Conclusão

Neste capítulo foram apresentados a base teórica e todos os parâmetros necessários para a implementação do sistema de reconhecimento de voz distribuído no ambiente celular/voz sobre IP.

No próximo capítulo, será analisado o desempenho destes parâmetros em diversos cenários, bem como as formas de aumentar o desempenho do reconhecimento na utilização dos mesmos.