Capítulo 3

Análise de Fala

No capítulo anterior foi descrito o processo de produção de fala e examinadas as características mais importantes dos sinais de fala utilizadas na maior parte das aplicações em fala. Este capítulo apresenta alguns dos métodos de análise dos sinais de fala que extraem essas características, convertendo o sinal num outro ou num conjunto de parâmetros, capazes de o descrever de um modo simplificado.

Os métodos de análise podem ser realizados no domínio do tempo, processando directamente o sinal de fala, ou no domínio da frequência, depois de efectuada uma transformação espectral. Em qualquer dos casos, o objectivo será o de obter uma representação do sinal que contenham a informação relevante no formato mais eficiente. Relativamente à codificação e armazenamento, o objectivo é eliminar a redundância. Em termos do reconhecimento, o objectivo principal é extrair um conjunto de parâmetros que sejam consistentes para

oradores diferentes e com pouca dispersão para a mesma entidade a reconhecer (*e.g.*, fonemas, classes de fonemas, palavras isoladas, etc.), ao mesmo tempo que exibam variações suficientes entre essas entidades.

Começaremos esta análise por um dos métodos mais poderosos de análise e caracterização de sinais de fala, o método da predição linear. Através deste método estimaremos a envolvente espectral e deste modo um modelo do tracto vocal. Seguem-se estimativas de outros parâmetros alternativos na modelação do tracto vocal: os coeficientes LSF (*Line Spectrum Frequency*), os formantes e os *cepstra*. Seguidamente apresentaremos métodos de estimação de parâmetros da excitação do tracto vocal, embora na sua forma mais simples: decisão de vozeamento e estimação da frequência fundamental. Finalmente apresentaremos métodos de detecção do género do orador e de actividade de voz quando envolvido em ruído ou silêncio.

3.1 Predição Linear

A predição linear tornou-se num dos métodos dominantes na estimação de parâmetros do sinal de fala, numa trama em que se considera o sinal estacionário. A ideia básica por detrás da predição linear é a de que o valor de uma amostra pode ser aproximado (predito), por combinação linear dos valores das amostras anteriores, tirando partido da correlação entre estas. Os coeficientes de predição linear ou coeficientes LPC (*Linear Predictive Coding*) são estimados por minimização do erro quadrático entre a amostra actual e a sua predição. Será apresentada a formulação desta estimação, baseada na codificação de sinais, e os resultados interpretados no domínio da frequência e através da função de autocorrelação. O filtro resultante modela o tracto vocal, pelo que são apresentadas analogias com o processo natural de produção de fala.

A utilização da predição linear não se limita à codificação, sendo exemplos de outras aplicações no processamento de fala o reconhecimento e síntese e a identificação e verificação do orador. Exemplos de aplicações noutros campos são: a prospecção de petróleo através da análise da vibração da terra causada pela explosão de cargas de dinamite; o diagnóstico médico do cérebro através da análise de sinais do electroencefalograma; e a codificação digital de imagem. Uma visão mais general da predição linear pode ser encontrada por exemplo em [Makhoul (75)].

3.1.1 Codificação por predição - princípios gerais

Na figura 3.1 é apresentado o esquema básico de um codificador, emissor e receptor, baseado em predição. O sinal de entrada *s*[*n*] está codificado em modulação por código de impulsos (PCM - *Pulse Code Modulation*) uniforme com um número suficiente de bits por amostra (12-16) para se considerar com qualidade indistinguível do original. É este sinal que será processado de modo a ser representado com um débito binário inferior.

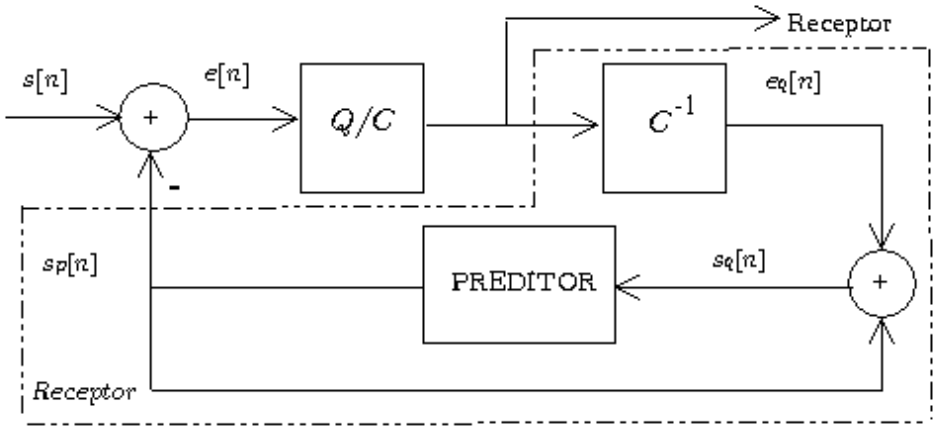

Figura 3.1
Emissor de um codificador por predição. *Q/C* e C^{-1} representam respectivamente um quantificador seguido de um codificador e um codificador inverso PCM. O emissor inclui um receptor apresentado dentro do tracejado.

O sinal é predito através das amostras anteriores, mas de modo a obter-se uma réplica desta predição no receptor são utilizadas as amostras quantificadas $s_q[n]$ e não o sinal original. É transmitido para o receptor, após quantificação, o erro ou resíduo de predição $e[n]$, dado pela diferença entre a amostra presente e a sua predição $s_p[n]$. O sinal de saída é sintetizado somando a predição com o resíduo após quantificação inversa. Na hipótese deste resíduo não sofrer qualquer degradação com a quantificação, o sinal de saída $s_q[n]$ é igual ao sinal de entrada $s[n]$. Como existe sempre um erro de quantificação dado pela diferença $e[n]$-$e_q[n]$, o ruído de quantificação $n[n]$ do sinal de saída é afectado exactamente pela mesma quantidade,

$$n[n] = s[n] - s_q[n] = (s_p[n] + e[n]) - (e_q[n] + s_p[n]) = e[n] - e_q[n] \qquad (3.1)$$

A relação sinal-ruído (SNR - *Signal to Noise Ratio*) de quantificação, corresponde por definição à relação entre as potências do sinal de entrada P_s e do ruído de quantificação P_n, virá,

$$SNR = \frac{P_s}{P_n} = \frac{P_s}{P_e}\frac{P_e}{P_n} = G_p \frac{P_e}{P_n} \qquad (3.2)$$

em que P_e é a potência do resíduo de predição e G_p é o ganho de predição. P_e/ P_n é a relação sinal-ruído de quantificação em PCM do resíduo de predição. A relação sinal-ruído é então calculada através do produto do ganho de predição pela relação sinal-ruído de quantificação em PCM do resíduo. Assumindo que esta relação sinal-ruído é igual à conseguida pela codificação directa em PCM do sinal de entrada utilizando o mesmo número de bits de codificação (por exemplo, como será visto na secção 6.2, utilizando PCM *companding*), resulta do esquema preditor uma melhoria na relação sinal-ruído desde que,

$$G_p = \frac{P_s}{P_e} > 1 \tag{3.3}$$

Para um dado sinal de entrada, G_p será tanto maior quanto menor for a potência do resíduo de predição. Os sinais de fala são não estacionários mas podem ser considerados quase estacionários (localmente estacionários) devido à variação lenta do tracto vocal. Assim, o ganho de predição será dependente das características acústicas locais do sinal de fala e a qualidade do sinal de saída terá flutuações temporais. O ganho de predição G_p pode ser calculado localmente dividindo o sinal em tramas de dimensão típica entre os 5 e os 30 ms. G_p em cada trama é calculado através da relação de potências ou da relação de energias localizadas, entre o sinal de entrada e o resíduo de predição.

3.1.2 Preditor de primeira ordem unitário

Uma vez que o sinal de fala tem uma forte característica passa-baixo, espera-se não haver uma variação muito forte entre duas amostras consecutivas, pelo que um preditor possível assume a forma,

$$s_p[n] = s_q[n-1], \tag{3.4}$$

e uma estimativa da energia localizada do respectivo resíduo de predição virá:

$$E_{em} = \sum_{n=m}^{m+N-1} e^2[n] = \sum_{n=m}^{m+N-1} \left(s[n] - s_q[n-1]\right)^2 , \tag{3.5}$$

em que m é a primeira amostra da trama e N a sua dimensão em número de amostras. Por simplicidade de notação deixaremos de utilizar o índice que referencia a trama, pelo que E_e representará a

estimativa da energia do resíduo de predição na trama em análise. Esta simplificação na notação será estendida a outros parâmetros.

Desenvolvendo o quadrado da equação (3.5) e assumindo que a energia do sinal após quantificação é igual à do sinal original E_s, virá,

$$E_e = \sum_{n=m}^{m+N-1} \left(s^2[n] + s_q^2[n-1] - 2s[n]s_q[n-1] \right) \approx 2E_s - 2R_s[1], \quad (3.6)$$

em que $R_s[k]$ é uma estimativa, na janela de análise, da função de autocorrelação do sinal de entrada. Designando por $r_s[k]$ a função de autocorrelação normalizada pela energia, o ganho de predição virá,

$$G_p = \frac{E_s}{E_e} = \frac{1}{2(1 - r_s[1])}, \quad (3.7)$$

pelo que haverá um ganho de predição desde que,

$$r_s[1] > 0{,}5. \quad (3.8)$$

Na figura 3.2 são ilustradas as formas de onda de uma trama de sinal de fala com a duração de 22,5 ms e o respectivo resíduo de predição. O sinal é amostrado a 8000 amostras por segundo, valor normalizado para a denominada banda telefónica. Pode verificar-se a diminuição da gama dinâmica do resíduo em relação ao sinal original, sendo susceptível de uma melhor quantificação. O ganho de predição é de 14,8 dB, o equivalente a mais de 2 bits de quantificação (6,02 dB por bit de quantificação, como se verá através da equação 6.4). Em relação a um codificador PCM pode então codificar-se o sinal com melhor qualidade com o mesmo débito binário, ou assumindo a mesma qualidade reduzir neste caso 2 bits por amostra, o que equivale a uma compressão de 16 kbit/s (2x8000). É no entanto evidente que o codificador não tem este ganho em todas as tramas.

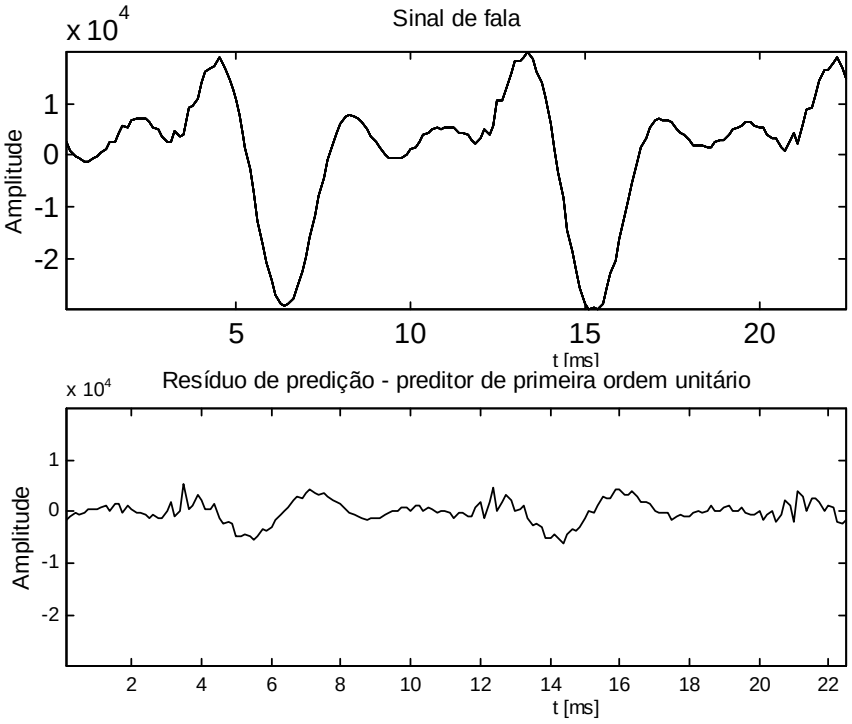

Figura 3.2
Em cima, trama de 22,5 ms de um sinal de fala.
Em baixo, resíduo de predição calculado com um preditor de primeira ordem unitário.

3.1.3 Preditor de primeira ordem adaptativo

Um outro possível preditor de primeira ordem será descrito por,

$$s_p[n] = as_q[n-1], \tag{3.9}$$

sendo *a* o coeficiente de predição. Este coeficiente é estimado de modo a maximizar o ganho de predição, o que para um dado sinal de entrada é equivalente a minimizar a energia do resíduo de predição,

$$E_e = \sum_{n=m}^{m+N-1} \left(s[n] - as_q[n-1]\right)^2 \approx E_s + a^2 E_s - 2aR_s[1], \tag{3.10}$$

em que novamente se assume que o sinal após quantificação é muito próximo do sinal de entrada.

Para calcular o coeficiente de predição *a* que minimiza E_e,

$$\frac{\partial E_e}{\partial a} = 2aE_s - 2R_s[1] = 0, \tag{3.11}$$

pelo que o coeficiente de predição *a* virá,

$$a = \frac{R_s[1]}{E_s} = r_s[1]. \tag{3.12}$$

De notar que o coeficiente de predição deverá ser recalculado no emissor trama-a-trama e enviado para o receptor, pelo que o número de bits que se poupa utilizando predição adaptativa deverá ser suficiente para quantificar este parâmetro e ainda obter uma compressão na representação do sinal. Das equações (3.10) e (3.12), o ganho de predição virá,

$$G_p = \frac{1}{1 - r_s[1]^2}. \tag{3.13}$$

Quando a correlação é nula, o ganho de predição atinge o valor mínimo igual à unidade, o coeficiente de predição é nulo e o erro de predição é igual ao sinal de entrada, pelo que o codificador degenera num codificador PCM. Ao contrário do que quando se utiliza um preditor unitário, o codificador com este preditor nunca tem um desempenho inferior a um codificador PCM.

3.1.4 Predição linear de sinais de fala

A ordem de predição pode ser estendida até um valor arbitrário *p*, sendo o preditor definido pela combinação linear das últimas *p* amostras[1],

$$s_p[n] = -\sum_{k=1}^{p} a_k s[n-k], \tag{3.14}$$

e a energia do erro de predição virá,

$$E_e = \sum_{n=m}^{m+N-1} \left(s[n] + \sum_{k=1}^{p} a_k s[n-k] \right)^2 . \tag{3.15}$$

Os coeficientes de predição a_k, k=1,..,p, são novamente estimados minimizando a energia do resíduo de predição,

$$\frac{\partial E_e}{\partial a_k} = 0 \qquad k=1,...,p \ , \tag{3.16}$$

pelo que resulta depois de alguma manipulação algébrica, juntamente com a equação (3.15), no sistema de p equações a p incógnitas

$$\sum_{k=1}^{p} a_k \sum_{n=m}^{m+N-1} s[n-k]s[n-i] = -\sum_{n=m}^{m+N-1} s[n]s[n-i] \qquad i=1,...,p, \tag{3.17}$$

que é equivalente a,

$$R_s[i] = -\sum_{k=1}^{p} a_k R_s[k-i] \qquad i=1,...,p. \tag{3.18}$$

Na figura 3.3 é apresentada a mesma trama da figura 3.2, sendo o resíduo de predição agora calculado com um preditor de ordem 10, típico das aplicações de codificação de sinais de fala amostrados a 8

[1] O sinal (-) que afecta a equação (3.14) provém da representação da equação às diferenças através de $\sum_{k=0}^{p1} b_k x[n-k] = \sum_{k=0}^{p2} a_k y[n-k]$ com a_0=1.

kHz. O ganho de predição aumenta, pelo que o resíduo de predição exibe uma menor correlação entre amostras consecutivas e uma menor gama dinâmica do que quando utilizado um preditor unitário.

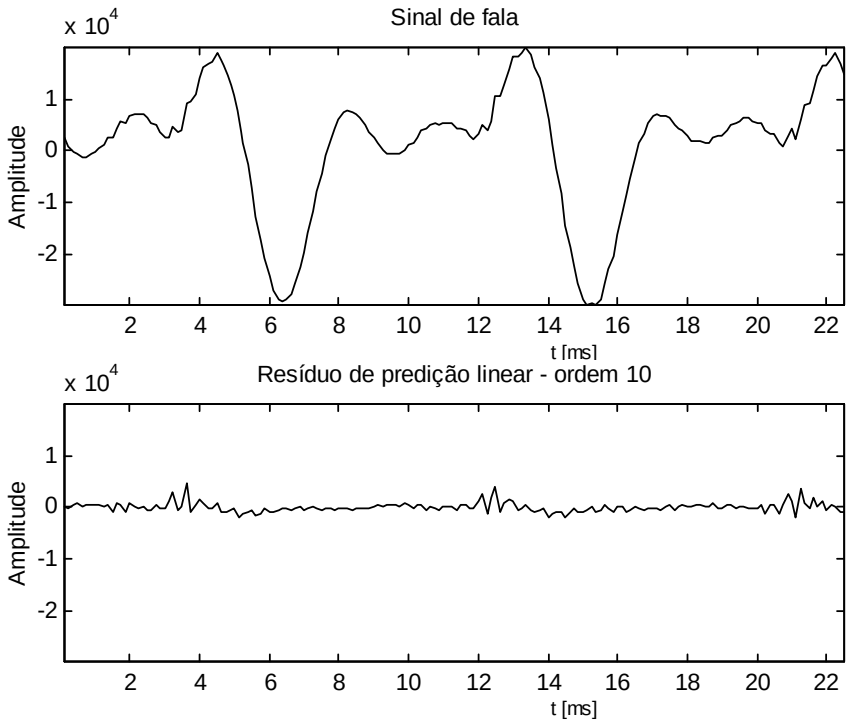

Figura 3.3
Em cima, trama de 22,5 ms de um sinal de fala. Em baixo, respectivo resíduo de predição calculado com um preditor linear adaptado de ordem 10.

Desenvolvendo o quadrado da equação (3.15) e tendo em atenção a equação (3.18), o ganho de predição virá,

$$G_p = \frac{1}{1 + \sum_{k=1}^{p} a_k r_s[k]}, \qquad (3.19)$$

ilustrado na figura 3.4 função da ordem de predição, para a mesma trama das figuras 3.2 e 3.3.

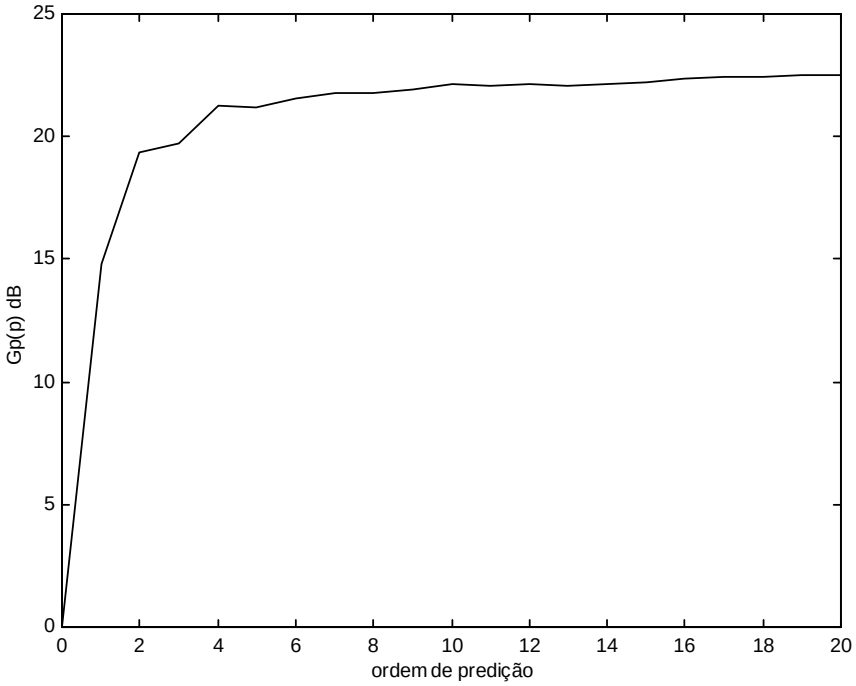

Figura 3.4
Andamento do ganho de predição para uma trama vozeada de 22,5 ms, em função da ordem de predição.

3.1.5 Filtro de síntese de predição linear

Através do esquema de blocos da figura 3.1 e da equação (3.14), o sinal sintetizado é descrito através da equação de filtragem,

$$s[n] = -\sum_{k=1}^{p} a_k s_q[n-k] + e_q[n]. \tag{3.20}$$

Assumindo novamente que a energia do sinal quantificado é igual à do sinal de entrada e aplicando transformada *z* a ambos os membros da equação (3.20), a função de transferência do filtro correspondente será descrita por,

$$H(z) = \frac{S(z)}{U(z)} = \frac{1}{A(z)} = \frac{G}{1+\sum_{k=1}^{p} a_k z^{-k}}, \tag{3.21}$$

em que se assume um sinal de entrada *u*[*n*] com potência unitária. O filtro resultante só tem pólos e a respectiva resposta impulsiva *h*[*n*] é dada por,

$$h[n] = -\sum_{k=1}^{p} a_k h[n-k] + G\delta[n]. \tag{3.22}$$

Multiplicando ambos os termos da equação (3.22) por *h*[*n-i*] e somando em *n*, resulta a seguinte recursão para a autocorrelação $R_h(i)$ da resposta impulsiva do filtro LPC,

$$R_h[i] = -\sum_{k=1}^{p} a_k R_h[i-k] \qquad i \neq 0. \tag{3.23}$$

sendo o valor da autocorrelação de ordem zero dado por,

$$R_h[0] = -\sum_{k=1}^{p} a_k R_h[k] + G^2 \tag{3.24}$$

Impondo a condição de que a energia da resposta impulsiva do filtro deverá ser igual à energia do sinal de entrada e tomando em consideração a parecença entre (3.18) e (3.23), conclui-se que,

$$R_h[i] = R_s[i] \quad 0 \le i \le p\,, \tag{3.25}$$

pelo que se pode reinterpretar a estimação dos parâmetros de predição linear como a estimação dos coeficientes de um filtro só de pólos (modelo auto-regressivo), tal que os p+1 primeiros valores da autocorrelação da respectiva resposta impulsiva sejam iguais aos do sinal que se quer modelar. Das equações (3.24) e (3.25) obtêm-se ainda o ganho *G*,

$$G = \sqrt{R_s[0] + \sum_{k=1}^{p} a_k R_s[k]}\,. \tag{3.26}$$

3.1.6 Estimação da envolvente espectral

Através da equação 3.25 demonstrou-se que os primeiros p+1 valores da função de autocorrelação da resposta impulsiva do filtro de predição linear são idênticos aos da função de autocorrelação do sinal que se quer modelar. Dado que são os primeiros valores da função de autocorrelação que contribuem primordialmente para a definição da envolvente espectral, sendo perdida a eventual estrutura harmónica conferida pela periodicidade do sinal de entrada, existe uma aproximação da resposta em frequência do filtro de predição linear à envolvente espectral do sinal a modelar.

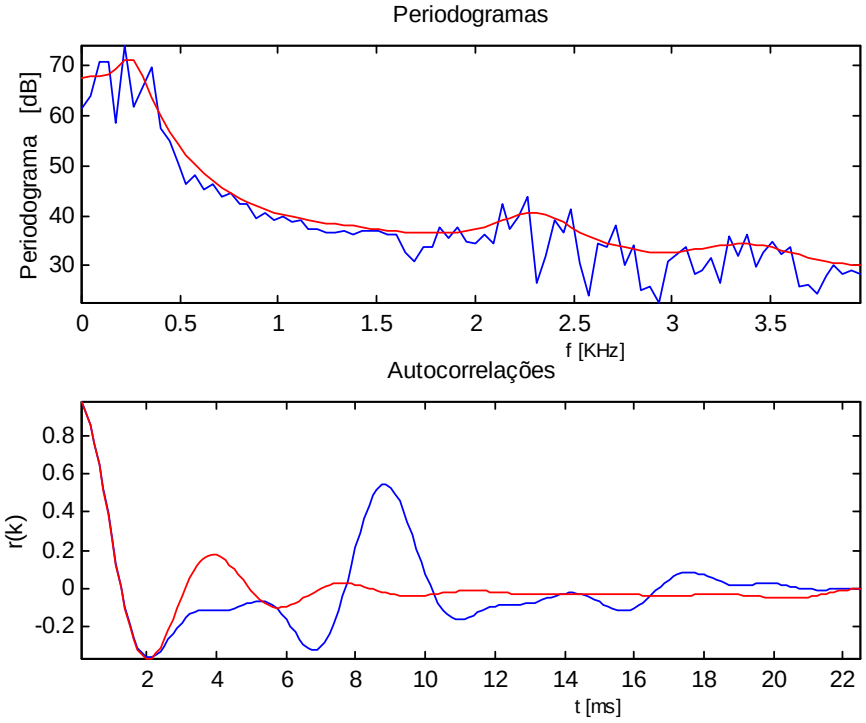

Figura 3.5
Em cima, periodograma do sinal de entrada e da resposta do filtro de predição linear de ordem 10. Em baixo, função de autocorrelação normalizada do sinal de entrada e da correspondente resposta do filtro de predição linear.

Esta característica do filtro LPC é ilustrada na figura 3.5 para a mesma trama das figuras anteriores, sendo também ilustrada a

coincidência, para valores até à ordem de predição, entre a função de autocorrelação normalizada do sinal de entrada e da resposta do filtro LPC. Para valores de ordem superior a autocorrelação da resposta impulsiva tende para zero, enquanto a autocorrelação do sinal reproduz a mesma periodicidade da entrada.

3.1.7 Modelo do tracto vocal

O filtro de predição linear, ao modelar a envolvente espectral do sinal, não entra em conta com a eventual estrutura harmónica da excitação vozeada. Esta tem uma característica essencialmente de baixa frequência que pode ser modelada com dois pólos reais próximos de $z=1$, que como mostrado na figura 3.6 produz um declive passa-baixo ao longo do espectro do sinal. A excitação não vozeada é contudo do tipo ruído branco, com espectro plano. Por outro lado a radiação nos lábios tem uma característica essencialmente passa-alto de primeira ordem. Esta poderá ser modelada por um zero real muito perto de $z=1$, eliminando a contribuição de um dos pólos da excitação nas zonas vozeadas. De modo a eliminar a contribuição do segundo pólo, diminuindo a gama dinâmica da envolvente espectral, é normalmente introduzido antes da estimação dos parâmetros do modelo um filtro de pré-ênfase das altas frequências, sendo estas melhor modeladas. Este filtro de pré-ênfase tem um zero real muito próximo da origem e é dado por,

$$P(z) = 1 - \mu z^{-1}, \tag{3.27}$$

em que μ tem um valor típico entre 0,9 e 1. Na síntese, é colocado à saída um filtro de de-ênfase com a característica inversa da representada na equação 3.27.

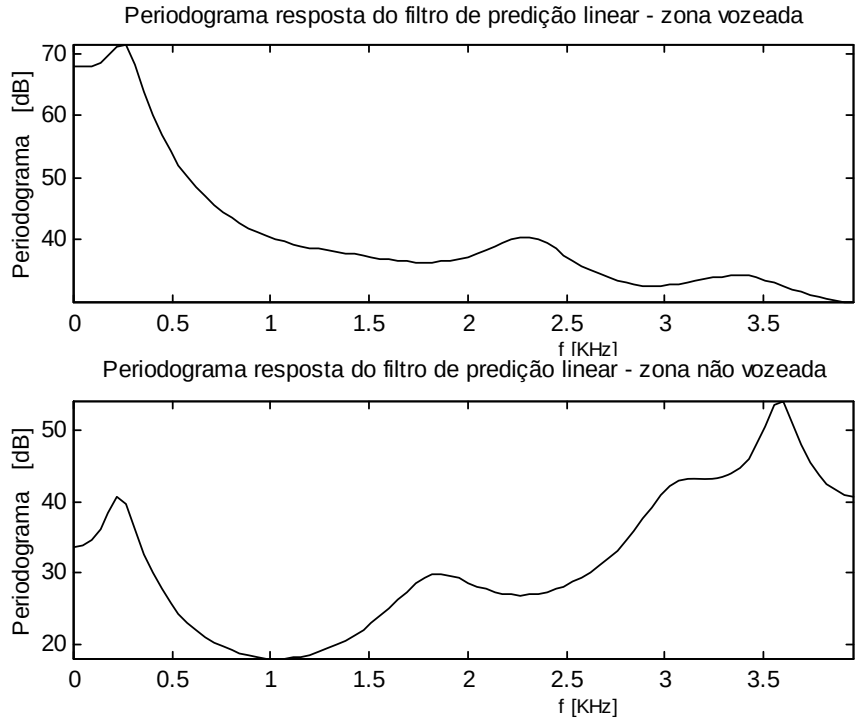

Figura 3.6
Periodograma da resposta impulsiva do filtro LPC.
Em cima, trama vozeada. Em baixo, trama não vozeada.

Dado que a acção conjunta da radiação nos lábios e do filtro de pré-ênfase tendem a anular o declive espectral provocado pela excitação vozeada, o filtro de predição linear modela apenas o tracto vocal.

3.1.8 Métodos de análise

De modo a se estimar os coeficientes de predição linear deve-se resolver o sistema de p equações a p incógnitas descrito pela equação 3.18, havendo para isso o método *lattice* e dois métodos baseados em definições diversas dos limites dos somatórios: o método da autocorrelação e o método da covariância.

3.1.8.1 Método da autocorrelação

Considere-se que o sinal $s[n]$ é nulo fora da janela de análise, o que pode ser descrito por,

$$s_m[n] = s[m+n]\,w[n] \tag{3.28}$$

sendo $w[n]$ uma janela de duração N, com valores zero fora do intervalo $0 \leq n \leq N\text{-}1$. Esta suposição leva a que o erro de predição no início da janela seja grande, já que se tenta predizer uma amostra à custa de amostras cujo valor é zero e estende-se o erro p amostras depois do fim da janela, predizendo amostras com valor zero à custa de amostras com valor não zero. Os limites dos somatórios da equação 3.17 serão entre 0 e $N+p\text{-}1$ e os somatórios corresponderão à função de autocorrelação $R[i]$ do sinal depois de multiplicado pela janela. Para minimizar os erros no início e no fim da trama deverá ser usada uma janela que tenda para zero nos seus extremos, como por exemplo uma janela de Hamming,

$$w[n] = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right). \tag{3.29}$$

Esta janela minimiza também o efeito da convolução no domínio da frequência entre a transformada do sinal e a transformada da janela (equação 3.28), quando comparado com a utilização de uma janela rectangular.

Com estes considerandos, o sistema de equações (3.18) colocado sobre a forma matricial dá origem a uma matriz de Toeplitz [pxp], em que todos os coeficientes ao longo das diagonais são iguais:

$$\begin{bmatrix} R[0] & R[1] & R[2] & ..... & R[p\text{-}1] \\ R[1] & R[0] & R[1] & ..... & R[p\text{-}2] \\ R[2] & R[1] & R[0] & ..... & R[p\text{-}3] \\ ..... & ..... & ..... & ..... & ..... \\ R[p\text{-}1] & R[p\text{-}2] & R[p\text{-}3] & ..... & R[0] \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ ..... \\ a_p \end{bmatrix} = - \begin{bmatrix} R[1] \\ R[2] \\ R[3] \\ ..... \\ R[p] \end{bmatrix} \tag{3.30}$$

Uma das vantagens do método da autocorrelação é a de que o filtro resultante é sempre estável. Sendo a matriz de Toeplitz, é ainda possível resolver este sistema de equações de um modo recursivo e portanto computacionalmente eficiente. Levinson propôs um desses algoritmos, reformulado mais tarde por Robinson [Markel (74)]. Mas um dos métodos mais eficientes foi proposto por Durbin [Makhoul (75)]. Este algoritmo, que tem como entrada os valores da autocorrelação do sinal de fala até à ordem p, pode ser descrito do modo seguinte:

$$E^{(0)} = R[0] \tag{3.31a}$$

$$k_i = \frac{\left\{ R[i] - \sum_{j=1}^{i-1} a_j^{(i-1)} R[i-j] \right\}}{E^{(i-1)}} \qquad 1 \le i \le p \tag{3.31b}$$

$$a_i^{(i)} = k_i \tag{3.31c}$$

$$a_j = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \qquad 1 \le j \le i\text{-}1 \tag{3.31d}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \tag{3.31e}$$

As equações 3.31b) a 3.31e) são resolvidas recursivamente para i=1,2,...,p e a solução final é dada por,

$$a_j = a_j^{(p)} \qquad 1 \le j \le p. \tag{3.32}$$

De notar que utilizando esta recursão os coeficientes dos preditores para ordens inferiores a p são todas calculados. Os valores intermédios k_i, $1 \le i \le p$, representam os coeficientes de reflexão do tracto vocal quando modelado por p secções sem perdas. Estes coeficientes são também conhecidos por coeficientes de correlação parcial (PARCOR). É possível apenas com estes coeficientes determinar os coeficientes

LPC. Através da recursão inversa, é também possível recalcular os coeficientes de reflexão a partir dos coeficientes LPC.

3.1.8.2 Método da covariância

Uma alternativa para a minimização do erro E_e numa trama de dimensão N será delimitar o somatório das equações 3.17 entre 0 e N-1. Se para além desta imposição nada se supuser sobre o sinal fora dessa trama, os somatórios deixam de representar a correlação do sinal multiplicado pela janela, como no método da autocorrelação, para representarem a covariância entre dois sinais muito parecidos mas não iguais. A matriz resultante embora simétrica deixa de ser de Toeplitz, pelo que a sua solução é computacionalmente mais exigente e não é garantida a estabilidade do filtro resultante. Melhoramentos ao método da covariância tornam o filtro LPC estável. A matriz resultante sofre uma decomposição de Cholesky [Rabiner (78)], resultando numa matriz triangular inferior, que dá por sua vez origem a coeficientes de reflexão k_i. A estabilidade é garantida pela condição:

$$|k_i| < 1 . \tag{3.33}$$

O método da covariância é mais exacto que o método da autocorrelação uma vez que evita os erros no início e fim da trama, embora para tramas de maior dimensão (20 ms) os dois métodos se equiparem, pois o erro cometido pelo método da autocorrelação torna-se numa percentagem pouco significativa do erro total.

3.1.8.3 Método *lattice*

Os métodos da autocorrelação e da covariância são métodos que estimam os coeficientes de LPC em dois passos: o cálculo da matriz de correlação e a solução de um conjunto de equações lineares. O método *lattice* é uma formulação em que estes dois passos estão combinados

num método recursivo para calcular os parâmetros de predição linear. Uma explicação detalhada deste método pode ser encontrada por exemplo em [Rabiner (78)].

3.2 Coeficientes LSF

O preditor linear pode ser definido com base num conjunto alternativo de coeficientes denominados coeficientes LSF (*Line Spectrum Frequencies*), obtidos por transformação. Os coeficientes LSF foram introduzidos por Itakura em 1975 [Itakura (75)], tendo sido as suas propriedades estudadas mais tarde por Soong e Juang [Soong (84)]. Por definição os coeficientes LSF são as frequências correspondentes às raízes de dois polinómios de ordem $p+1$, $P(z)$ e $Q(z)$, derivados do filtro inverso de predição linear $A(z)$, de ordem p. $P(z)$ corresponde ao tracto vocal com a fonte glotal completamente fechada (coeficiente de reflexão k_{p+1}=-1) e $Q(z)$ representa o tracto vocal com a fonte glotal completamente aberta (coeficiente de reflexão k_{p+1}=1). Como se verá na secção 3.3.3, estes coeficientes têm informação sobre os formantes do tracto vocal. Através da recursão do cálculo dos coeficientes de predição linear para uma ordem superior,

$$A_{p+1}(z) = A_p(z) - k_p z^{-(p+1)} A(z^{-1}), \tag{3.34}$$

virá:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) = A(z)\left[1 + z^{-(p+1)}\,\frac{A(z^{-1})}{A(z)}\right] \tag{3.35a}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) = A(z)\left[1 - z^{-(p+1)}\,\frac{A(z^{-1})}{A(z)}\right], \tag{3.35b}$$

representando os coeficientes LSF as frequências correspondentes às raízes destes dois polinómios. O filtro inverso $A(z)$ é recuperado de $P(z)$ e $Q(z)$ através de:

$$A(z) = \frac{P(z) + Q(z)}{2}. \tag{3.36}$$

Definindo $H(z)$ como:

$$H(z) = z^{-(p+1)} \frac{A(z^{-1})}{A(z)}, \tag{3.37}$$

pode-se calcular as raízes de $P(z)$ impondo a condição de $H(z)$=-1 e as raízes de $Q(z)$ impondo $H(z)$=1. $H(z)$ tem a característica de um filtro passa-tudo com ganho unitário,

$$\left|H(z)\right| = 1 \text{ para } \left|z\right| = 1, \tag{3.38}$$

pelo que todos os coeficientes LSF se situam sobre o círculo unitário. Os polinómios $P(z)$ e $Q(z)$ têm respectivamente uma raiz em z=-1 e em z=1. Factorizando $H(z)$ [Soong (84)],

$$H(z) = z^{-1} \prod_{k=1}^{p} \frac{(z_k z - 1)}{z - z_k}, \tag{3.39}$$

em que $z_k = r_k e^{jw_k}$ k=1,...,p correspondem às raízes de $A(z)$. A fase de $H(w)$ pode ser expressa através de:

$$\phi(w) = -(p+1)w - \sum_{k=1}^{p} 2\tan^{-1} \frac{r_k \sin(w - w_k)}{1 - r_k \cos(w - w_k)}, \tag{3.40}$$

e o atraso de grupo, definido como o simétrico da derivada da fase, pode ser expresso por,

$$\tau(w) = \frac{1}{f_s}\left(1 + \sum_{k=1}^{p} \frac{1 - r_k^2}{1 + r_k^2 - 2r_k \cos(w - w_k)}\right), \tag{3.41}$$

sendo uma função sempre positiva desde que o filtro seja estável, ou seja, $r_k<1$, pelo que $\phi(w)$ é uma função monótona decrescente. Dado que o módulo de $H(w)$ é sempre igual a 1, os coeficientes LSF podem ainda ser calculados impondo a condição:

$$\phi(w) = -m\pi, \qquad m = 1..p, \tag{3.42}$$

estando os coeficientes LSF entrelaçados, ou seja, à medida que a frequência aumenta ocorre alternadamente uma raiz de $P(z)$ e outra de $Q(z)$. Esta condição é suficiente para manter a estabilidade do filtro de predição linear após uma eventual quantificação. A transmissão dos coeficientes LSF em vez dos coeficientes LPC, em sistemas de codificação, é aliás a principal utilização destes, já que a quantificação directa dos coeficientes LPC pode resultar num filtro instável.

3.3 Estimação de formantes

As frequências dos formantes e respectivas larguras de banda são alguns dos parâmetros espectrais que se podem extrair do sinal de fala. A sua importância deve-se à relação que têm com a posição dos articuladores na fase de produção de fala e com a inteligibilidade da fala, sendo largamente utilizados na codificação, síntese e reconhecimento de sinais de fala.

3.3.1 Método dos máximos do espectro

As frequências dos formantes podem ser estimadas através dos máximos locais do espectro de curta duração (figura 2.3), calculado por exemplo através da Transformada de Fourier localizada ou pela envolvente espectral produzida por predição linear. Este método de análise é contudo pouco preciso, principalmente utilizando a Transformada de Fourier localizada, pois a energia do espectro está

concentrada em múltiplos da frequência fundamental. Para vozes femininas e de crianças, em que a frequência fundamental é elevada, este problema agrava-se pois só por acaso a frequência do formante situa-se perto de um múltiplo de *F*0. Quando a largura de banda do formante é menor que a frequência fundamental, esta pode mesmo não incluir nenhuma harmónica.

3.3.2 Método das raízes do filtro de predição linear

Um outro processo de estimação dos formantes envolve a análise das raízes do denominador do filtro de predição linear, pois uma raiz pode ser identificada como formante se apresentar uma largura de banda suficientemente estreita.

Decompondo o polinómio do denominador do filtro LPC em sistemas ressonantes de segunda ordem, a função de transferência de cada *k*-ésimo sistema ressonante, correspondente à *k*-ésima raiz $r_k e^{j\omega k}$ e à respectiva raiz complexa conjugada, é descrita por,

$$T_k(z) = \frac{A_k}{(1 - r_k e^{j\omega_k} z^{-1})(1 - r_k e^{-j\omega_k} z^{-1})}, \tag{3.43}$$

O valor de A_k é calculado de modo a que o sistema tenha um ganho unitário para z=1 através de:

$$A_k = (1 - r_k e^{j\omega_k})(1 - r_k e^{-j\omega_k}), \tag{3.44}$$

As frequências de ressonância F_k e as larguras de banda B_k de cada sistema ressonante podem ser obtidas através de [Atal (71)]:

$$F_k = \frac{\omega_k}{2\pi T_s} \tag{3.45}$$

$$B_k = -\frac{\ln(r_k)}{\pi T_s}, \tag{3.46}$$

pelo que F_k corresponderá a um formante se B_k for suficientemente estreita, ou equivalentemente, se r_k apresentar um valor perto da unidade.

Repare-se que cada dois pólos do filtro LPC correspondem a um formante, pelo que a ordem de análise deverá ser de pelo menos 2 vezes o número máximo de formantes esperado na largura de banda considerada do sinal. Para sinais limitados a 4 kHz são esperados 3 a 4 formantes, pelo que a ordem de análise deverá ser no mínimo de 8. Em sistemas de codificação em que se modela o tracto vocal através do filtro LPC é geralmente utilizado um valor típico de 10, servindo os dois pólos adicionais para melhor modelar o declive espectral.

3.3.3 Método dos coeficientes LSF

A conversão de coeficientes de predição linear em coeficientes LSF converte uma raiz de $A(z)$ num par de raízes no círculo unitário. Uma das características dos coeficientes LSF é a de se aproximarem da respectiva raiz de $A(z)$. A tabela 3.1 ilustra esta aproximação. É simulado um filtro $A(z)$ a partir da definição arbitrária das suas raízes, sendo os respectivos coeficientes LSF determinados com uma resolução de 1 Hz, através da equação (3.42).

ordem	F_k	B_k	r_k	ω_k	P	Q	P^*	Q^*
						0		0
1	200	77,56	0,97	0,16	195	312	364	727
2	600	130,62	0,95	0,47	581	727	1091	1455
3	1600	568,23	0,80	1,26	1397	1794	1818	2181
4	2700	130,62	0,95	2,12	2500	2702	2545	2909
5	3400	908,27	0,70	2,67	2899	3422	3272	3636
					4000		4000	

Tabela 3.1
Relação entre as raízes de um filtro $A(z)$ e os coeficientes LSF, calculados com uma resolução de 1 Hz. P^* e Q^* representam os coeficientes LSF para um sinal com espectro plano (fs=8kHz, p=10).

A figura 3.7 ilustra, para a simulação descrita na tabela 3.1, a resposta de fase $\phi(f)$ e o atraso de grupo $\tau(f)$ correspondentes ao filtro $H(f)$, baseados respectivamente nas equações 3.40 e 3.41.

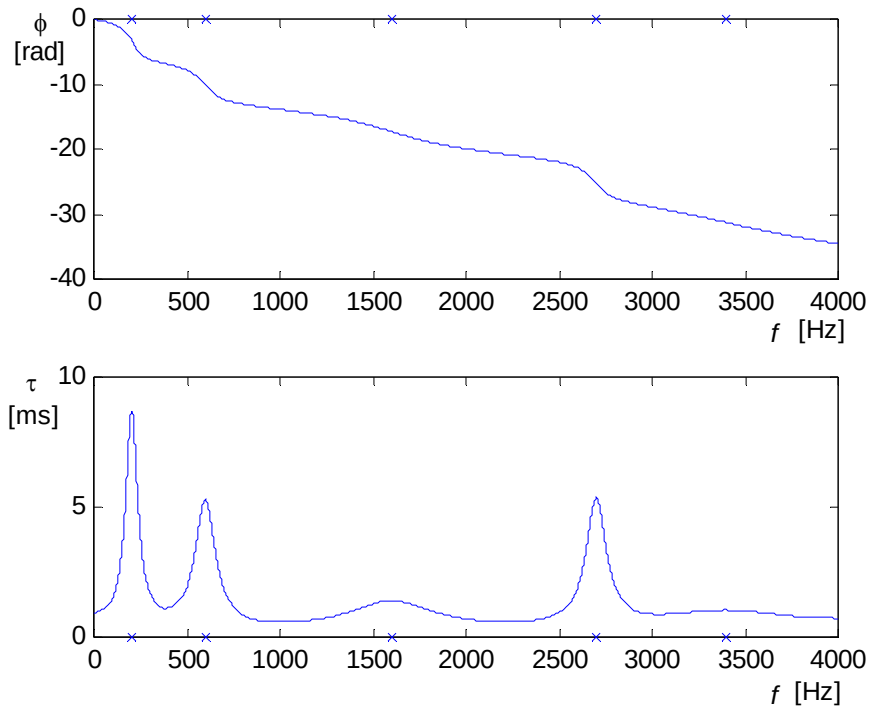

Figura 3.7
Resposta do filtro $H(f)$ da simulação da tabela 3.1
Em cima: resposta da fase. Em baixo: atraso de grupo.
Os × marcam as frequências das raízes de $A(z)$.

Se o espectro de entrada for plano, os coeficientes LSF estão separados uniformemente entre 0 e $fs/2$. Se uma raiz de $A(z)$ apresentar um valor do seu módulo perto da unidade, a largura de banda é estreita (equação (3.46)), sendo muito provável que este pólo corresponda a um formante. Nesta situação, como ilustrado na figura 3.8, também correspondente à simulação ilustrada na tabela 3.1, os coeficientes LSF aproximam-se da respectiva raiz de $A(z)$, uma vez que a variação da fase é muito grande nesta zona (o atraso de grupo é grande à volta de w_k) [Kang (85)].

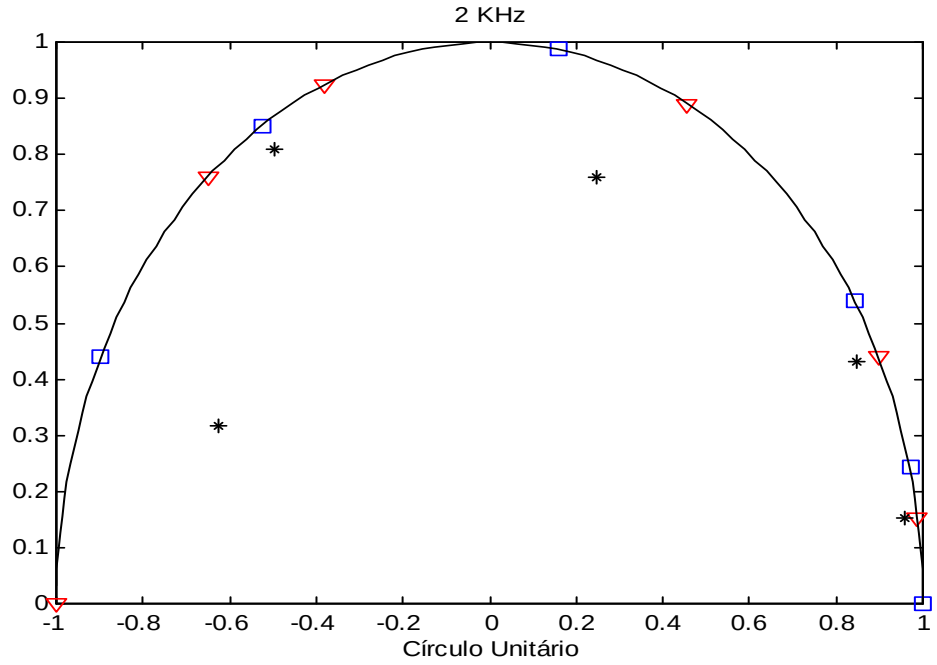

Figura 3.8
Coeficientes LPC e LSF.
Os * marcam as raízes de $A(z)$, correspondentes à simulação descrita na tabela 3.1; Os ∇ marcam as raízes de $P(z)$ e os □ às raízes de $Q(z)$. O semi-plano inferior corresponde à imagem do semi-plano superior.

No caso extremo de uma raiz ter módulo unitário, este e os coeficientes LSF respectivos coincidiriam. Pelo contrário, se uma raiz de $A(z)$ apresentar um valor baixo em módulo, será grande a correspondente largura de banda e a sua contribuição traduzir-se-á apenas na inclinação espectral, estando as raízes correspondentes de $P(z)$ e $Q(z)$ afastadas. Obviamente, poderá ainda existir interacção entre pólos consecutivos do filtro de predição linear, que contribua também para a definição de formantes ou para a inclinação espectral.

3.4 Análise *cepstral*

O *cepstrum* de um sinal é definido como a transformada inversa de Fourier[2] do logaritmo do espectro do sinal. Para um espectro S(w)

[2] Originalmente definida como a T.F. directa embora a inversa seja hoje a mais utilizada.

(raiz quadrada da função espectral de potência), a representação em série de Fourier de logS(*w*) é expressa através de,

$$log\ S(w) = \sum_{n=-\infty}^{\infty} c_n e^{-jwn}\ , \tag{3.47a}$$

sendo

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(S(w)) e^{jwn} dw\,, \tag{3.47b}$$

em que $c_n = c_{-n}$ são reais, definidos como coeficientes *cepstra* reais ou simplesmente coeficientes *cepstra*. Note-se que,

$$c_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} log|S(w)|dw\,. \tag{3.48}$$

Uma das vantagens da utilização dos coeficientes *cepstra* é a facilidade com que se pode separar a excitação do tracto vocal deste último, devido à operação logarítmica que transforma o produto no domínio da frequência na soma das duas componentes. Assim, os coeficientes *cepstra* podem ser decompostos por,

$$c_n = c_n^e + c_n^v\,, \tag{3.49}$$

em que c^e e c^v correspondem respectivamente aos coeficientes *cepstra* da excitação e da resposta do filtro que modela o tracto vocal, correspondendo os primeiros aos valores de ordem mais elevada e os segundos aos valores de ordem mais baixa de c_n.

Quando o sinal é modulado através de um filtro de fase mínima só de pólos, correspondente ao filtro preditor linear, os coeficientes *cepstra*, que neste caso são denominados de coeficientes *cepstra* de LPC, são obtidos pelas equações:

$$c_n = -a_n - \frac{1}{n}\sum_{k=1}^{n-1} k c_k a_{n-k} \qquad p \geq n > 0 \tag{3.50a}$$

$$c_0 = \log\left(G^2\right) \tag{3.50b}$$

$$c_n = -\frac{1}{n}\sum_{k=1}^{n-1} k c_k a_{n-k} \quad n > p \tag{3.50c}$$

em que a_0=1, G é o ganho do filtro LPC e p é a ordem do filtro.

Para um par de espectros $S(w)$ e $S'(w)$, é possível aplicando o teorema de Parseval relacionar a distância euclidiana *cepstral* (erro quadrático) com a distância *rms* do logaritmo espectral através de:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\log S(w) - \log S'(w)\right|^2 dw = \sum_{n=1}^{\infty} \left(c_n - c'_n\right)^2 , \tag{3.51}$$

sendo c_n os coeficientes *cepstra* correspondentes a $S(w)$ e c'_n os coeficientes correspondentes a $S'(w)$. Repare-se que não foi considerado o termo c_0 correspondente à energia. Os coeficientes *cepstra* são utilizados com sucesso no reconhecimento. De modo a associar diferentes pesos a cada coeficiente, o termo do lado direito da equação 3.51 é alterado para,

$$\sum_{n=0}^{\infty} W(n)\left(c_n - c'_n\right)^2 , \tag{3.52}$$

sendo $W(n)$ a janela que pesa os diferentes coeficientes *cepstrais*. Uma das janelas mais utilizadas, exemplificada na figura 3.9, é a da meia arcada sinusoidal,

$$W(n) = 1 + \frac{L}{2} sin\left(\frac{n\pi}{L}\right) \qquad 0 < n \leq L , \tag{3.53}$$

correspondendo L à ordem de truncatura no cálculo dos coeficientes *cepstra*, tipicamente entre 10 e 16.

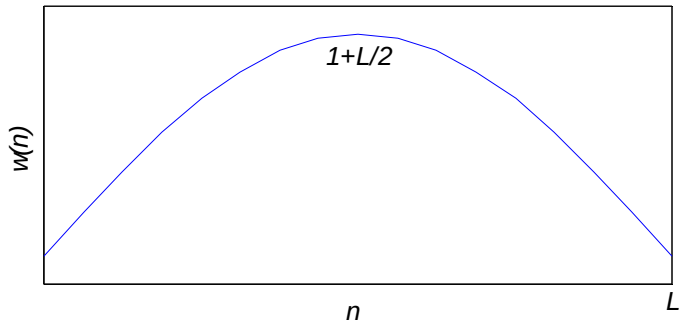

Figura 3.9
Exemplo de janela de pesos dos coeficientes *cepstra* utilizada no reconhecimento de fala.

3.5 Estimação da Frequência Fundamental

Os limites possíveis de vibração das cordas vocais situam-se aproximadamente entre os 50 e os 500 Hz. As frequências mais baixas são típicas das vozes graves masculinas e as frequências mais altas são típicas das crianças e de algumas vozes femininas, mais agudas. Existem vários métodos descritos na literatura para a detecção de vozeamento e estimação da frequência fundamental numa trama, mas basicamente ou utilizam métodos no domínio do tempo, no domínio da frequência, ou mistos. Seguidamente descrevem-se sucintamente alguns desses métodos, podendo-se encontrar uma descrição e sua comparação mais pormenorizada por exemplo em Rabiner [Rabiner (76)] ou no capítulo 14 [Talking (95)] do livro [Kleijn (95)].

3.5.1 Método da Autocorrelação

Como mostra a figura 3.10, a periodicidade da forma de onda encontra-se também presente na função de autocorrelação. Se o primeiro valor máximo da função de autocorrelação, normalizado pela energia do sinal e procurado dentro dos limites possíveis de vibração das cordas vocais, apresentar um valor razoável (*e.g.* maior que 0,3),

então a trama é considerada vozeada e o valor da frequência fundamental estimado como o valor dessa periodicidade.

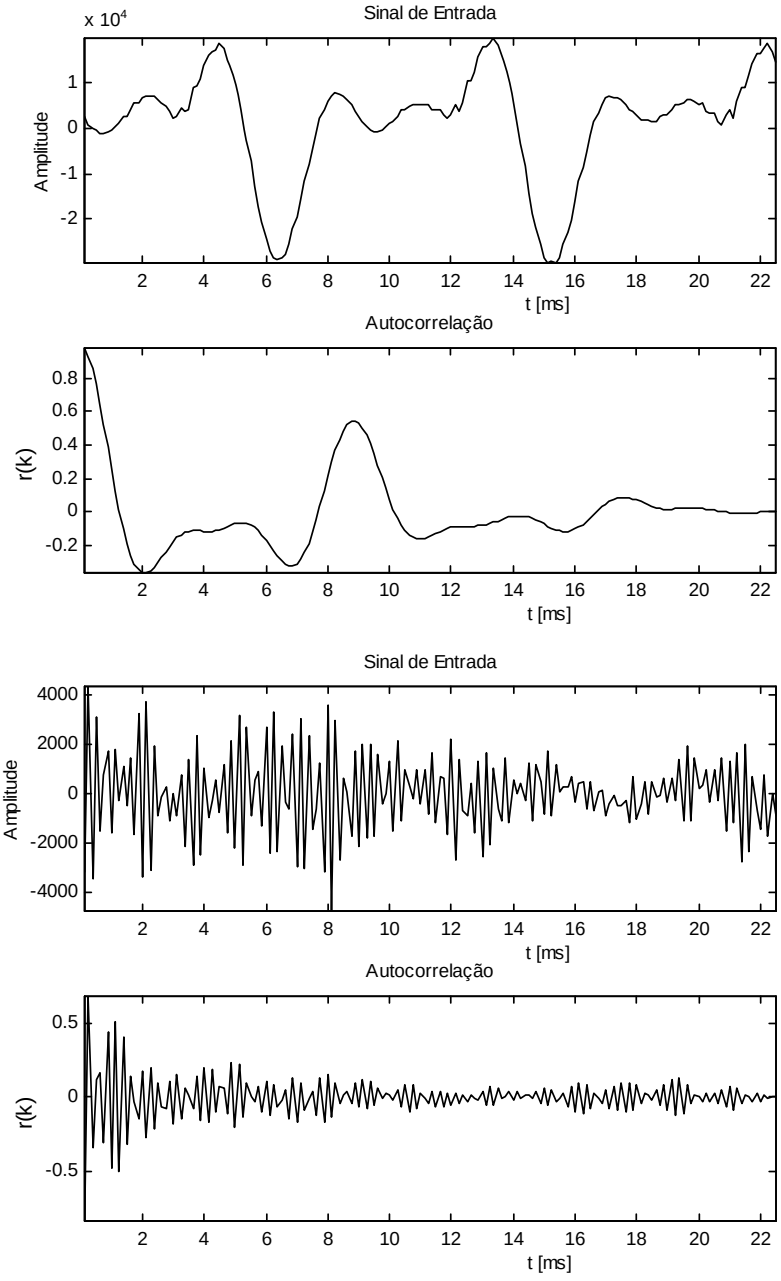

Figura 3.10
Em cima: Zona Vozeada e respectiva função de autocorrelação normalizada.
Em baixo: Zona não vozeada e respectiva função de autocorrelação normalizada.

Uma vez que a frequência fundamental tem um valor baixo e pode existir uma componente não vozeada forte nas frequências mais

elevadas, o sinal deve ser previamente filtrado por um filtro passa-baixo com frequência de corte a cerca de 900 Hz [Rabiner (77)]. Embora os sinais de fala tenham geralmente média nula, é possível encontrar uma componente DC localizada na trama de análise, que deve também ser retirada antes do cálculo da autocorrelação.

Embora este método seja razoavelmente robusto na presença de ruído, tem como principal desvantagem a necessidade de utilizar janelas relativamente grandes (2 vezes o máximo período possível), que introduzem maior atraso e tornam o método sensível às variações entre períodos, nomeadamente às variações de amplitude. Uma segunda desvantagem prende-se com a diminuição do intervalo de cálculo, à medida que se aumenta o período candidato, que faz variar a imunidade ao ruído e às variações entre períodos e enfatiza as harmónicas coincidentes com o primeiro formante do tracto vocal. O efeito dos formantes pode ser atenuado calculando a função de autocorrelação sobre o resíduo de predição de LPC em vez do próprio sinal, que mantém a mesma periodicidade (figura 3.3) do sinal mas em que é retirada a contribuição dos formantes dada pelo filtro LPC.

A decisão de vozeamento pode tomar em consideração outras características, nomeadamente o número de passagens por zero. Se este valor for grande (*e.g.* maior que 2 passagens por milisegundo [Markel (74)]) a trama é considerada não vozeada. Outras características [Ribeiro (91-b)] a ter em conta são: a relação entre a energia de baixas e altas frequências, já que para tramas vozeadas o espectro apresenta o típico declive espectral passa-baixo proveniente da vibração da glote (figura 3.6); e a grande correlação com as tramas anteriores, havendo a tendência para se mudar a decisão apenas quando há fortes indícios nesse sentido.

3.5.2 Método da função de correlação cruzada normalizada

A função de correlação cruzada normalizada *(*NCCF - *normalised cross-correlation function*) ultrapassa as dificuldades da função de autocorrelação na estimação da frequência fundamental [Talkin (95)]. Definindo *m* como o primeiro elemento da janela de análise de dimensão *N*, centrada na *i*-ésima trama e *k* o intervalo temporal em número de amostras para a qual se está a calcular a NCCF, esta é definida como:

$$\Phi_i[k] = \frac{\sum_{j=m}^{m+N-1} s[j]s[j+k]}{\sqrt{e_m e_{m+k}}}, \qquad k \leq N \qquad (3.54)$$

sendo e_j dado por,

$$e_j = \sum_{l=j}^{j+N-1} s[l]^2 . \qquad (3.55)$$

A dimensão da janela de análise é no mínimo igual ao máximo período possível da frequência fundamental e não ao dobro desse valor, como no caso da utilização da função de autocorrelação. Os valores da NCCF estão limitados entre -1 e 1, correspondendo os valores perto da unidade a múltiplos do período fundamental, independentemente das variações de amplitude entre períodos, e o primeiro máximo local, desde que suficientemente representativo, ao período fundamental.

Para testar valores do máximo da NCCF que não estejam próximos da unidade, mas que mesmo assim sejam razoavelmente significativos, o procedimento de decisão de vozeamento pode socorrer-se, como para o método da autocorrelação, de outras características do sinal de fala para ajudar à decisão.

3.5.3 Métodos no domínio da frequência

Devido à periodicidade nas zonas vozeadas, o espectro é essencialmente um conjunto de réplicas do espectro da janela de análise espaçadas da frequência fundamental. A energia das baixas frequências nas zonas vozeadas é também elevada, enquanto nas zonas não vozeadas a energia se espalha de um modo mais uniforme pelo espectro. Estas características podem ser utilizadas para a decisão de vozeamento e estimação do período fundamental, usando por exemplo o algoritmo *harmonic sieve* proposto por Sluyter em 1982 [Sluyter (82)].

Neste método, a decisão de vozeamento depende da intensidade espectral dada pelo máximo do módulo da Transformada de Fourier Localizada (TFL) no intervalo 200 a 800 Hz. O algoritmo de decisão é relativamente complexo e tem em conta não só este valor mas também o valor da intensidade espectral e a decisão nas 5 últimas tramas. Após a decisão da trama como vozeada, a estimação do período fundamental é feita por tentativas, verificando qual dos candidatos tem o melhor alinhamento das suas harmónicas.

3.5.4 *Pitch* Fraccionário

Como a vibração das cordas vocais não coincide em princípio com um múltiplo da frequência de amostragem, o período fundamental não é um número inteiro mas sim um número fraccionário (*pitch fraccionário*) [Marques (89)]. Os valores intermédios entre amostras são obtidos por interpolação do tipo $sen(x)/x$, modelando a reconstrução do sinal por um filtro ideal passa-baixo com frequência de corte de metade da frequência de amostragem (figura 3.11). Este método, embora aumente em muito os requisitos de cálculo e sejam necessários mais bits na codificação, faz diminuir o erro de predição.

Um procedimento alternativo é modelar este efeito com um preditor de dois ou três elementos de atraso, escolhidos à volta do período real.

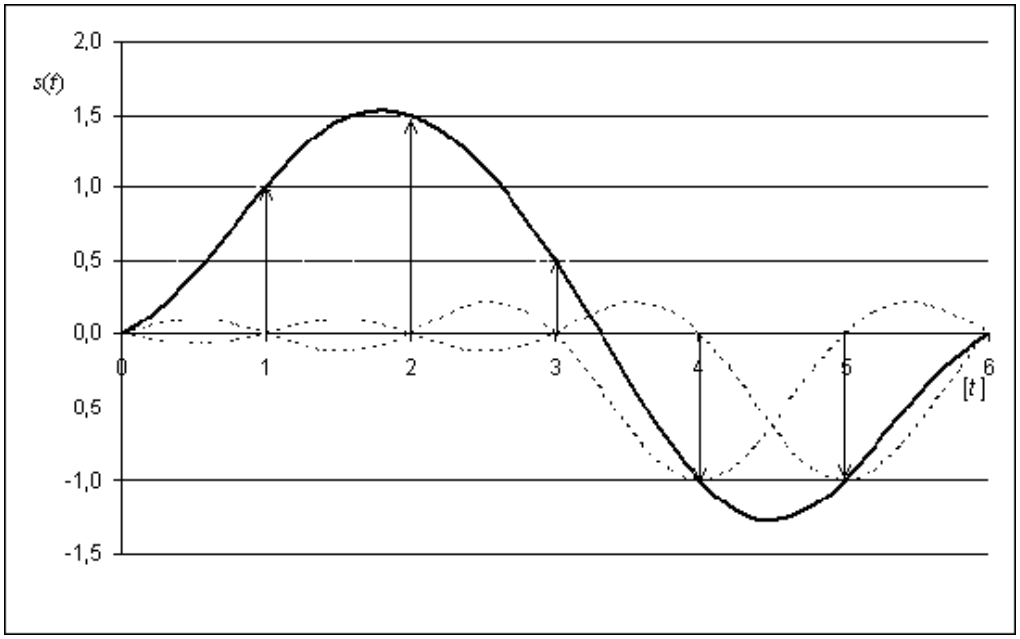

Figura 3.11
Reconstrução utilizando interpolação seno(x)/x, correspondente à resposta impulsiva de um filtro passa-baixo ideal com frequência de corte de fs/2.

3.6 Identificação do género do orador

O género do orador é, talvez, a primeira característica da fala que um ouvinte humano é capaz de identificar. A dimensão do tracto vocal e a espessura e comprimento das cordas vocais são as causas principais desta distinção tão clara, que divide o espaço dos sinais de fala. A dimensão do tracto vocal influencia a localização dos formantes e a espessura e comprimento das cordas vocais influenciam a frequência fundamental, pelo que os oradores do género feminino têm, tipicamente, formantes e frequências fundamentais mais elevados que os oradores do género masculino. Uma pré-identificação automática do género pode, assim, assistir com sucesso a algumas aplicações do processamento de fala, tais como o reconhecimento de fala independente do orador e a identificação ou a verificação do orador.

Apenas baseado no valor da frequência fundamental é possível implementar m método de baixa complexidade para identificação do género do orador [Meneses (99-a)], situando-se o ponto de distinção à volta dos 150 a 170 Hz. Como exemplo apresenta-se na figura 3.12 um gráfico das frequências de ocorrências dos valores médios da frequência fundamental em segmentos fonéticos, para 8 oradores (4 de cada género), correspondendo a 32 minutos, separadamente para oradores do género feminino e masculino, verificando-se uma razoável separação (157 Hz) entre as duas classes. A taxa de sucesso para 48 oradores (24 de cada género) corresponde a 90%.

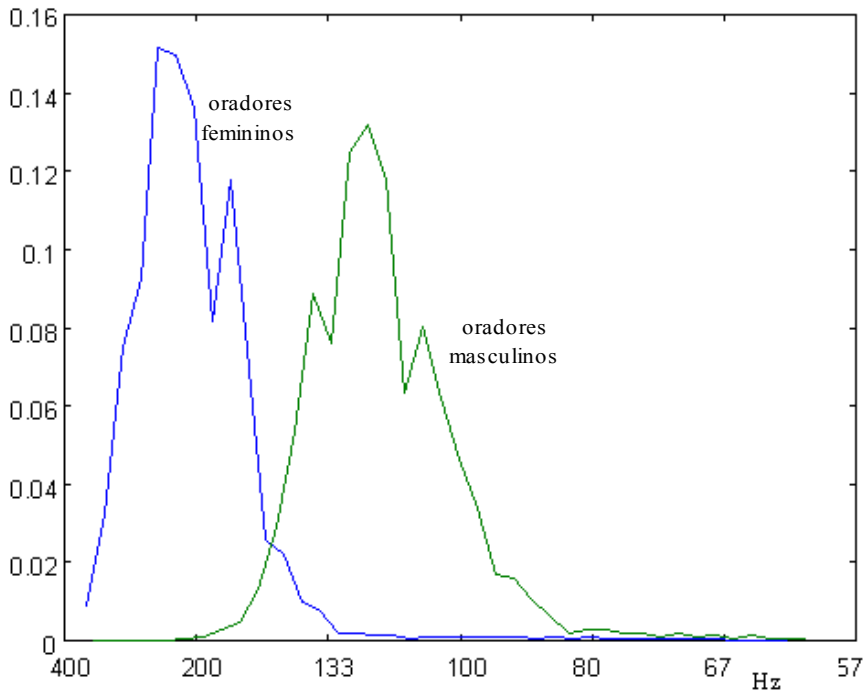

Figura 3.12
Frequências de ocorrência do valor médio da frequência fundamental, calculadas por segmentos fonéticos.

A identificação utilizando o valor obtido com base numa trama é, no entanto, pouco robusta, já que as variações prosódicas aumentam ou diminuem localmente o valor da frequência fundamental. De modo a

tornar mais robusta a decisão, deve-se manter uma estimativa do seu valor médio, por exemplo através de uma filtragem passa-baixo do tipo:

$$\overline{FO}'(k) = (1-\alpha)\overline{FO}'(k-1) + \alpha FO(k) \quad 0 < \alpha < 1, \tag{3.56}$$

em que α corresponde ao coeficiente de filtragem.

3.7 Detecção de actividade de voz

Em algumas aplicações é necessário detectar a presença de actividade de voz (VAD – *Voice activity detection*). Esta detecção é razoavelmente simples caso se esteja num ambiente silencioso, mas em zonas de mais baixa relação sinal-ruído ou com música de fundo esta tarefa é dificultada. As aplicações da VAD são: (1) reconhecimento de palavras isoladas ou fala contínua; (2) transmissão por pacotes de sinais de fala em que as zonas de ausência de actividade de voz podem ser codificadas com um débito binário menor, não sobrecarregando a rede de transmissão (uma conversa telefónica típica é constituída por apenas cerca de 50% de actividade de fala em cada sentido); (3) algoritmos de cancelamento de ruído e cancelamento acústico de eco em que é necessária uma boa estimativa do ruído, obtida nas zonas de ausência de fala.

3.7.1 Detecção de palavras isoladas

A detecção de palavras isoladas toma também a designação de detecção de extremos, ou seja do início e fim da palavra. A aplicação óbvia é o reconhecimento de palavras isoladas. Um método clássico com vista ao reconhecimento, proposto por Lamel [Lamel (81)] assume que cada palavra compreende uma sequência de um ou mais pulsos de energia. O problema a solucionar passa por encontrar estes pulsos e determinar quais pertencem a cada palavra.

A detecção dos pulsos de energia é um procedimento esquerda-direita, ou seja, uma função monótona do tempo. O sinal é dividido em tramas (Lamel propõe tramas de 15 ms), sendo estimadas as respectivas energias $R_l[0]$, sendo l o índice de cada trama. Como mostrado na figura 3.13, quando $R_l[0]$ ultrapassa o limiar $k1$, a trama respectiva ($A1$) é considerada o início de um pulso de energia se mais tarde a energia exceder o limiar $k2$ (na trama $A2$), a menos que a duração entre $A2$ e $A1$ seja demasiado longa e neste caso o início do pulso é considerado em $A2$. O fim de pulso é detectado de maneira similar, utilizando os limiares de decisão $k2$ e $k3$, sendo detectado como fim do pulso a trama $A4$, a menos que a duração entre $A4$ e $A3$ seja demasiado longa (tipicamente indica respiração no fim da palavra) em que neste caso é considerada a trama $A3$. São ainda efectuados mais dois testes: (1) A energia máxima é calculada e se o seu valor for inferior ao limiar $k4$ o pulso é rejeitado; (2) a duração total é calculada e se for inferior a 75 ms o pulso é rejeitado (bater de porta, pancada numa mesa etc.).

Figura 3.13
Detecção de pulsos de energia [Lamel (81)]

A estimação da energia $R_l(0)$ deve ter em conta o ruído de fundo. A energia do ruído é estimada nas zonas de ausência de voz, pressupondo que a sua energia se mantém durante a sua presença. O valor da energia do ruído pode assim ser "descontado" na estimação

global da energia da trama, uma vez que o ruído e o sinal não são correlacionados.

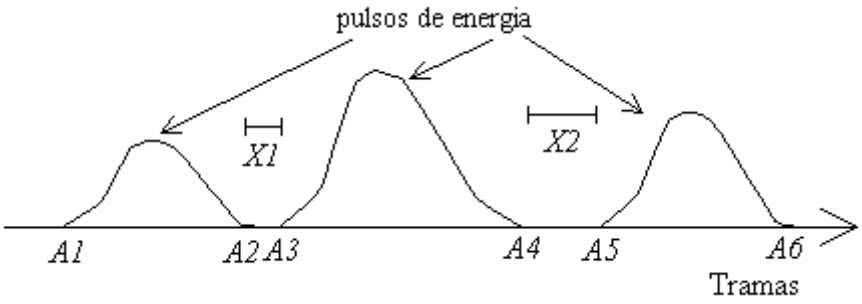

Figura 3.14
Detecção de palavras com base em pulsos de energia [Lamel (81)]

Dado que uma palavra é constituída por um conjunto de um ou mais pulsos, estes são considerados como possivelmente pertencentes à mesma palavra caso a distância que os separa seja inferior a, por exemplo como proposto por Lamel, 75 ms. Caso contrário considera-se como pertencentes a palavras diferentes. A palavra contudo não deverá ter uma dimensão menor que 150 ms e, função da aplicação, uma dimensão máxima. No exemplo da figura 3.14, assumindo que as distâncias entre pulsos *X*1 e *X*2 sejam inferiores a 75 ms, é considerada a hipótese mais provável de extremos de palavra as tramas *A*1 e *A*6. Esta hipótese é testada por um reconhecedor de palavras isoladas e assumida como válida se a distância para a palavra reconhecida for suficientemente pequena. Caso contrário, de um modo recursivo, são testadas as hipóteses (*A*1, *A*4) (caso *X*2 seja maior que *X*1), (*A*3, *A*6) e (*A*3, *A*4). A escolha final é assim escolhida de um modo integrado entre o reconhecedor e hipóteses cada vez menos prováveis de extremos de palavras. Caso *X*2 seja inferior a *X*1 a sequência testada é (A1, A6), (A3, A6), (*A*1, *A*4) e (*A*3, *A*4).

3.7.2 Detecção de fala contínua

A detecção de fala contínua utiliza tipicamente não só a energia mas também parâmetros típicos dos sinais de fala, tais como o a periodicidade das zonas vozeadas, a grande energia de baixas frequências e a não estacionaridade, esta última característica de alguns tipos de ruído. A recomendação de codificação do ITU-T G.729 Anexo B (96), por exemplo, utiliza as diferenças com estimativas dos respectivos valores médios da: energia e energia da banda de baixa frequência; número de passagens por zero; coeficientes espectrais (LSF). Outro exemplo, a norma de codificação GSM AMR [Vahatalo (99)] estima o somatório da relação sinal-ruído por (10) bandas;