Capítulo 4

Síntese de Fala

As primeiras tentativas de construir máquinas de produção de fala, embora sintetizando apenas 5 vogais, remontam a 1779, por C. G. Kratzenstein. Poucos anos mais tarde, em 1791, W. R. von Kempelen demonstrou uma máquina muito mais sofisticada e capaz de produzir fala contínua, provando que o sistema humano de produção de fala podia ser modelado artificialmente. No mesmo ano publicava um livro descrevendo os seus estudos sobre produção de fala e as experiências de duas décadas até chegar a esta máquina.

Em 1835, Wheatstone demonstrou, na *Dublin Association for the Advances of Sciences*, uma máquina construída com base nos princípios descritos no livro de von Kempelen. Esta máquina, representada na figura 4.1, usava um fole para fornecer ar a um ressoador feito em pele, sendo a sua secção alterada pela mão de um

operador. A outra mão manipulava quatro comandos que geravam constrições de modo a produzir consoantes.

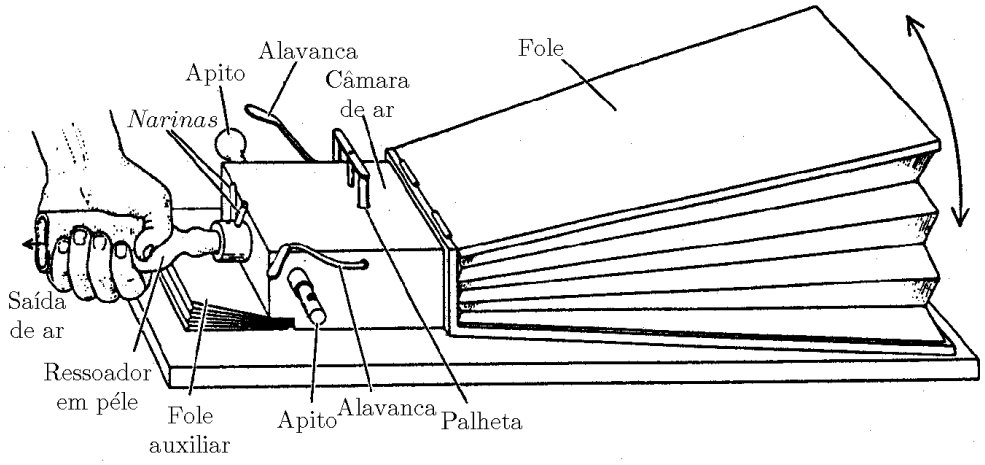

Figura 4.1
Versão de Wheatstone da máquina falante de von Kempelen.

O primeiro sintetizador de fala completamente eléctrico, conhecido por *Pedro the voder* (*vo*ice *de*monstrate*r*), foi inventado por Homer Dudley e apresentado em 1939 na feira mundial de Nova York. O nome Pedro provêm de Dom Pedro, Imperador do Brasil, presente em 1876 na primeira Exibição Centenária em Filadélfia, quando da demonstração do telefone por Alexander Bell. Um episódio que ficou célebre envolvendo D. Pedro foi quando este, ao ouvir uma voz proveniente do telefone, exclamou: *Meu Deus, fala*.

A figura 4.2 apresenta uma fotografia do *voder* a ser demonstrado por uma operadora, a Senhora Harper, tirada na feira de Nova York. Este era manipulado através de 14 teclas que controlavam a estrutura que modelava o tracto vocal, por uma barra que escolhia o tipo de excitação (ruído nas zonas vozeadas ou um oscilador simulando a frequência fundamental nas zonas não vozeadas) e um pedal que permitia a variação da frequência do oscilador.

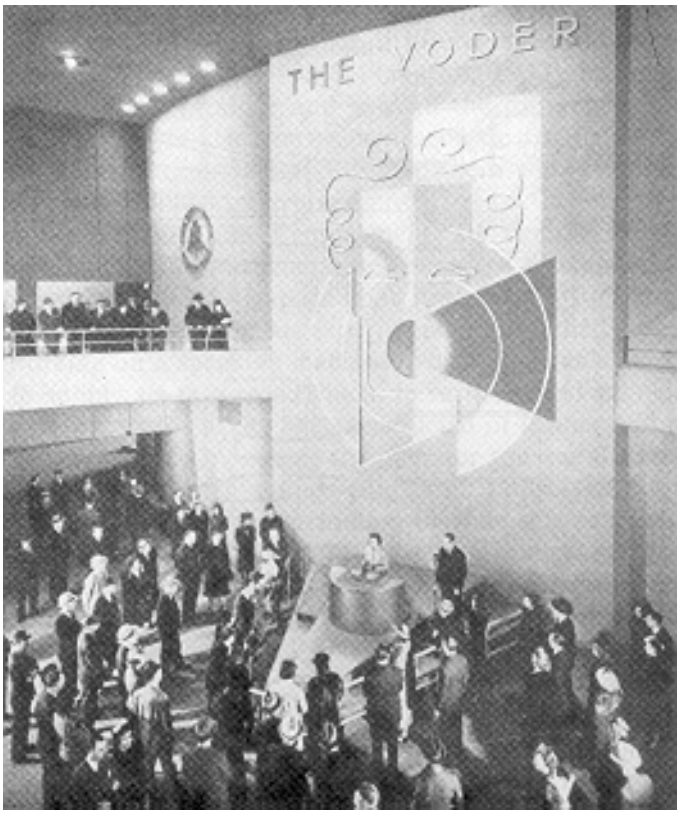

Figura 4.2
O *voder* em demonstração na feira de Nova York em 1939.

Ainda em 1939, Dudley propôs os *vocoders* (*vo*ice-*coders*) de canal [Dudley (39)], que representam os sinais de fala no domínio da frequência. O sinal é modelado através da energia de um conjunto de filtros passa-banda contíguos, que dividem a banda total num número fixo de canais. Este conjunto de filtros é excitado por ruído nas zonas não vozeadas e por pulsos periódicos nas zonas vozeadas modelando a abertura e fecho da glote. A maioria dos modelos de síntese mantém ainda esta estrutura fonte-filtro na modelação dos sinais de fala.

O aparecimento dos computadores produziu grandes desenvolvimentos no processamento de fala no resto do século XX, sendo os modelos principais de síntese apresentados resumidamente neste capítulo.

4.1 Síntese por predição linear

A síntese de sinais de fala baseada em predição linear tem como base o receptor do esquema de blocos apresentado na figura 3.1. Nesta figura o bloco preditor corresponde obviamente à predição linear, sendo a excitação do filtro uma representação do respectivo resíduo de predição *e*[*n*]. De modo a modelar com alguma precisão as variações lentas do tracto vocal durante a produção da fala, os coeficientes de predição devem ser actualizados num máximo de 30 ms e estimados com um dos métodos descritos na secção 3.1.8.

4.1.1 Predição de longa duração

Nas zonas vozeadas, quer o sinal de fala quer o respectivo resíduo de predição são quase periódicos. Esta quase periodicidade pode ser explorada para minimizar ainda mais o resíduo de predição através de um preditor do tipo:

$$e'_p[n] = \alpha e'[n - N0], \tag{4.1}$$

sendo $N0$ o valor do período de vibração das cordas vocais ou período fundamental, medido em múltiplos do período de amostragem e α o coeficiente de predição de longa duração. Este coeficiente tem tipicamente valores da ordem da unidade. Valores ligeiramente superiores ou inferiores são obtidos nas zonas de transição, respectivamente com aumento ou diminuição de energia. Nas zonas não vozeadas este preditor não se aplica. O valor óptimo de α, tendo como critério a minimização da energia do resíduo, é obtido substituindo na equação (3.12) o atraso de uma amostra pelo atraso $N0$,

$$\alpha = \frac{R_e[N0]}{E_e} = r_e(N0). \tag{4.2}$$

O resíduo *e*''[*n*] gerado pela cascata destes dois preditores, cujo esquema de blocos se apresenta na figura 4.3, denomina-se de resíduo de dupla predição.

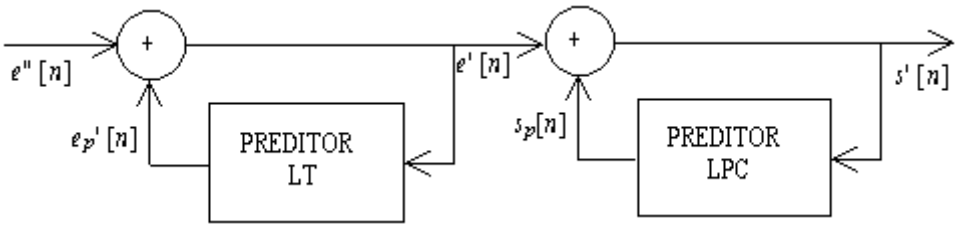

Figura 4.3
Síntese por dupla predição.

Ao contrário do preditor LPC que explora a correlação entre amostras consecutivas, este preditor explora a correlação entre períodos glotais, que têm uma duração maior, pelo que se denomina de preditor de longa duração (LT - *Long Term*).

Na figura 4.4 é apresentado o resíduo de dupla predição (LPC de ordem 10 e preditor de longa duração com $N0=71$) e comparado com o resíduo de predição de ordem 10. O sinal original é mostrado na figura 3.3. Como se pode verificar, a periodicidade que se encontrava ainda no resíduo de LPC foi praticamente eliminada. O resíduo de dupla predição aproxima-se de ruído gaussiano, com uma gama dinâmica ainda menor que o resíduo de LPC. Este preditor só não tem um melhor desempenho porque se está a restringir que o atraso $N0$ seja um número inteiro de períodos de amostragem.

A utilização de dupla predição na síntese de sinais de fala é um dos métodos mais utilizados em codificação de sinais de fala, tendo para isso sido desenvolvidos métodos eficazes de representação do resíduo de dupla predição e de quantificação dos coeficientes LPC, que serão descrito no capítulo sobre 6 sobre quantificação.

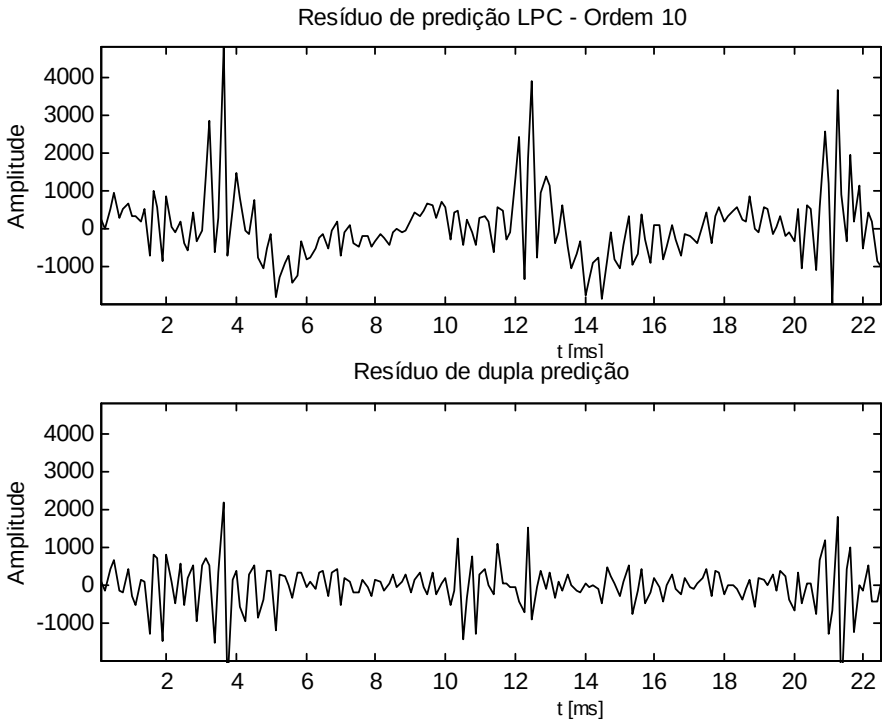

Figura 4.4
Em cima: Resíduo de predição LPC - Ordem 10.
Em baixo: Resíduo de dupla predição (LPC+LT)

4.1.2 *Vocoder* LPC

Os *vocoders* (*vo*ice *coders*) utilizam um modelo simplificado da produção da fala e não tentam reproduzir fielmente o sinal de entrada, mas apenas representá-lo de modo a manter as suas características perceptualmente mais importantes, tais como a envolvente espectral, a estrutura fina do espectro e a energia global.

Um exemplo de um *vocoder* bem sucedido é o mostrado na figura 4.5, em que para modelar o tracto vocal é utilizado um filtro de predição linear (*vocoder* LPC), excitado com ruído branco para produzir sons não vozeados, ou com um trem de pulsos com período igual ao da vibração das cordas vocais, para sons vozeados. Este sintetizador, proposto por Atal e Hanauer em 1971 [Atal (71)], torna-se completamente paramétrico, mas a qualidade do sinal sintetizado é menor do que se fosse conhecido o resíduo de predição.

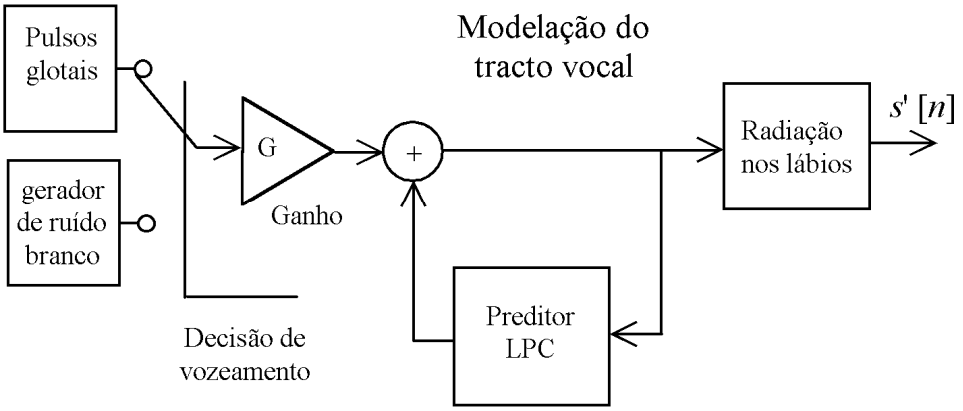

Figura 4.5
Síntese com o modelo *vocoder* LPC

De modo a modelar as variações do tracto vocal e da periodicidade dos pulsos glotais, os coeficientes do modelo deverão ser alterados a intervalos entre os 5 e os 30 ms.

Nas zonas não vozeadas, o ganho, calculado através da equação 3.19 assumindo à entrada um impulso, deverá ser ajustado de modo a que se adeque a uma entrada do tipo ruído gaussiano com variância unitária, para que se mantenha a mesma energia, ou seja,

$$G' = G \times \sqrt{\frac{1}{DimensãoTrama}}\,. \tag{4.3}$$

4.2 Modelo do pulso glotal

Os pulsos glotais podem ser modelados através de um trem de impulsos quase periódicos, representando as variações lentas da frequência fundamental. Contudo, para além de possuir uma envolvente espectral plana, este trem de impulsos imporia a mesma fase na origem para todas as harmónicas, o que produziria um sinal sintetizado com um pico maior que o da fala natural. Por outro lado, se a análise LPC for produzida com pré-ênfase (secção 3.1.7), deverá ser colocado um

filtro de de-ênfase após a predição, constituído por um filtro passa-baixo de primeira ordem, que imporá um declive espectral de -6 dB por oitava.

Para evitar o pico no sinal sintetizado, pode-se utilizar uma forma de onda com envolvente espectral plana, mas em que a variação de fase ao longo das harmónicas produz um sinal mais natural. Alternativamente, pode-se utilizar uma forma de onda mais parecida com a do resíduo de predição, aproximadamente equivalente à de um filtro passa-baixo de segunda ordem, modelado por Rosenberg [Rosenberg (71)] através do polinómio seguinte:

$$u_g(t) = \begin{cases} k_0 + k_1 t + k_2 t^2 + k_3 t^3 & \text{para } 0 \le t \le T_{op} \\ 0 & \text{para } T_{op} \le t \le T0 \end{cases}, \qquad (4.4)$$

em que T_{op} representa a duração da fase de abertura da glote, assumida como uma percentagem fixa do período fundamental $T0=1/F0$.

É normal incorporar no modelo da fonte glotal a característica de radiação nos lábios, essencialmente uma característica passa-alto de primeira ordem, derivando a equação 4.4. O declive espectral total corresponde então a -6 dB por oitava (-12 dB/oitava do modelo da fonte glotal e +6 dB/oitava da radiação nos lábios), o que é equivalente aos impostos pelo filtro de de-ênfase.

As constantes k_0, k_1, k_2 e k_3 da equação 4.4 podem ser calculadas [Oliveira (93)] impondo algumas restrições: a derivada do fluxo glotal é zero na origem e o integral num período fundamental da derivada do fluxo glotal é zero, de modo a evitar a introdução de uma componente DC. A derivada do fluxo glotal, ilustrada na figura 4.6, é então representada no domínio discreto por,

$$u_g'[n] = \begin{cases} \dfrac{(2N_{op} - 1)n - 3n^2}{N_{op}^2 - 3N_{op} + 2} & para\ 0 \le n < N_{op} \\ 0 & para\ N_{op} \le n < N0 \end{cases}, \qquad (4.5)$$

em que $N0$ representa o número de amostras do período fundamental e N_{op} o número de amostras da fase de abertura da glote.

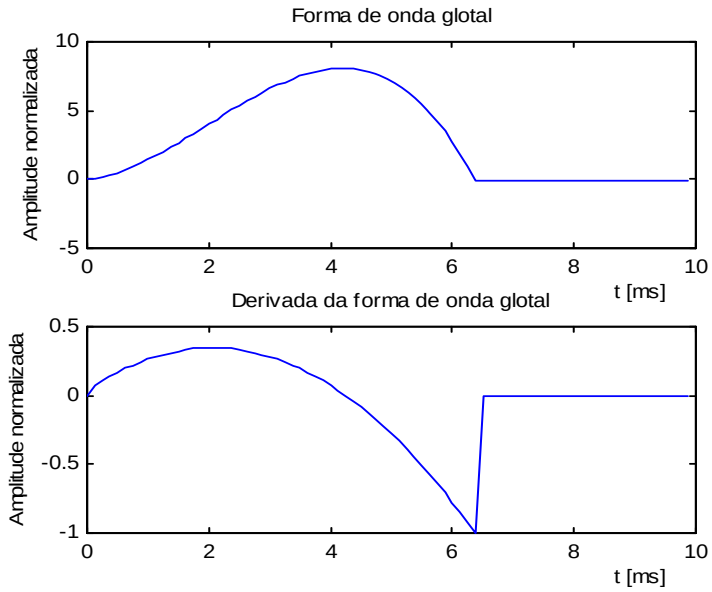

Figura 4.6
Modelo da forma de onda do fluxo glotal (em cima) e respectiva derivada do fluxo glotal (em baixo) (Adaptado de [Oliveira (93)]). O período fundamental ilustrado é de 10 ms (80 amostras), com uma duração da fase de abertura da glote de 66%.

O ganho, calculado através da equação 3.19 assumindo á entrada um impulso, deverá ser ajustado de modo a que se mantenha a mesma energia, ou seja,

$$G' = G \times \sqrt{\frac{PeriodoGlotal}{DimensãoTrama}}\,. \qquad (4.6)$$

4.3 Síntese sinusoidal

Os sinais vozeados são produzidos quando as cordas vocais vibram, produzindo um sinal quase-periódico. Esta quase periodicidade tem implicações no espectro de curta duração, tornando-o discreto se o sinal for absolutamente periódico (harmónicas da frequência fundamental). Por outro lado, o sinal torna-se não harmónico quando da diminuição da periodicidade ou nas zonas não vozeadas. Uma trama do sinal pode no entanto ser sempre descrita por uma soma de sinusóides. Sendo conhecidas as frequências e as respectivas amplitudes e fases (ou coeficientes complexos) de cada uma das sinusóides, é possível sintetizar o sinal.

Esta síntese pode ser implementada por transformada inversa mas, de forma a manter a continuidade do sinal nas fronteiras das tramas, o sinal deve ser sintetizado no domínio do tempo, o que facilita a variação das frequências e das amplitudes, através da sobreposição de sinusóides [Almeida (84-a)] [McAulay (84)]:

$$s'(t) = \sum_{k=1}^{K(t)} A_k(t) \sin\left(\varphi_k(t)\right), \tag{4.7}$$

em que $A_k(t)$ e $\varphi_k(t)$ representam, respectivamente, as amplitudes e fases da k-ésima sinusóide, cuja frequência é dada pela derivada da fase $\varphi_k(t)$, sendo $K(t)$ o número de sinusóides na banda considerada. As amplitudes, frequências e fases são calculadas nas fronteiras das tramas, sendo os valores intermédios obtidos por interpolação, utilizando interpolação linear das amplitudes e cúbica das fases (quadrática das frequências) [Almeida (84-a)].

Nas zonas vozeadas o sinal pode ser considerado periódico, restringindo as frequências a serem harmónicas da frequência

fundamental *F*0, sendo este modelo designado por *modelo harmónico*. A fase pode ainda ser restringida a ser contínua trama-a-trama (*vocoder* harmónico), sendo este modelo descrito por [McAulay (90)],

$$s'(n) = \sum_{k=1}^{K(n)} A_k(n) \sin\left(\frac{n 2\pi k F0(n)}{fs} + \theta_k(n) \right), \qquad (4.8)$$

em que *fs* é a frequência de amostragem e $\theta_k(n)$ controla a periodicidade do sinal. Os valores de *F*0*(n)* são também calculados nas fronteiras das tramas e interpolados linearmente amostra-a-amostra (interpolação quadrática da fase). Nas zonas vozeadas $\theta_k(n)$ é nulo, mantendo-se a continuidade de fase. Nas zonas não vozeadas a periodicidade é cortada pela soma à fase de cada harmónica de uma componente aleatória $\theta_k(n)$ de média nula.

Repare-se que o modelo harmónico com restrições na fase e em que as amplitudes são modeladas pelos coeficientes LPC tem os mesmos parâmetros de controlo do *vocoder* LPC. No entanto, principalmente para vozes de oradores do género masculino, o sinal pode soar demasiado tonal devido ao elevado número de harmónicas. A utilização de modelos harmónicos mais ruído [Stylianou (98-a)] ou a utilização de uma probabilidade de vozeamento que define uma frequência a partir da qual o sinal é considerado não vozeado [McAulay (95)], atenua o efeito tonal e modela melhor as zonas de transição de vozeamento.

4.4 Síntese de formantes

Decompondo o polinómio do denominador do filtro de predição linear (equação 3.21) numa cascata de *p*/2 sistemas ressonantes de segunda ordem, através das raízes índice *k*, $r_k e^{j\omega_k}$ e das respectivas raízes conjugadas, virá:

$$\frac{G}{1+\sum_{i=1}^{p} a_i z^{-i}} = G\prod_{k=1}^{p/2}\frac{1}{(1-r_k e^{j\omega k} z^{-1})(1-r_k e^{-j\omega k} z^{-1})}. \qquad (4.9)$$

Conhecendo as frequências de ressonância F_k dos formantes e correspondentes larguras de banda B_k, os módulos e fases das respectivas raízes são calculados a partir de:

$$w_k = 2\pi T_s F_k \qquad (4.10a)$$

$$r_k = e^{-\pi T_S B_k}, \qquad (4.10b)$$

sendo T_s o período de amostragem. Repare-se contudo que apenas os sistemas ressonantes que tenham raízes com um módulo perto da unidade correspondam a formantes. Para produzir fala o tracto vocal pode ser modulado através desta cascata de filtros, tal como mostrado na figura 4.7, utilizando a mesma excitação que a excitação do modelo do *vocoder* LPC, transformando-se num *vocoder* de formantes.

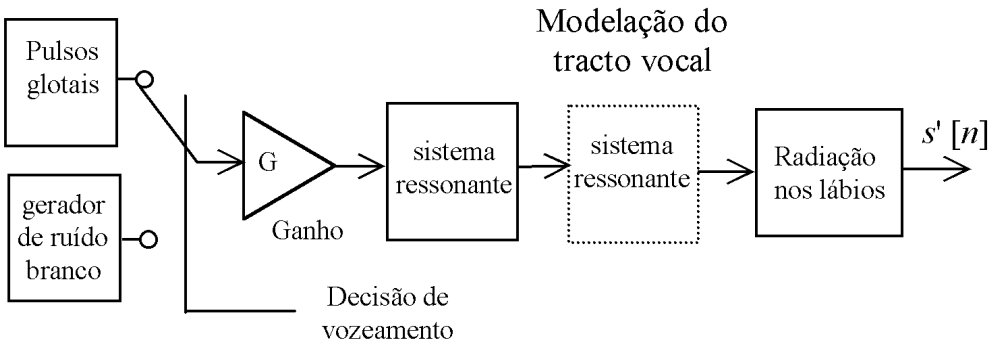

Figura 4.7
Vocoder de formantes.

Uma configuração alternativa à cascata de sistemas ressonantes é a configuração em paralelo, ilustrada na figura 4.8. Ao contrário da configuração em cascata, cada sistema tem um ganho individual.

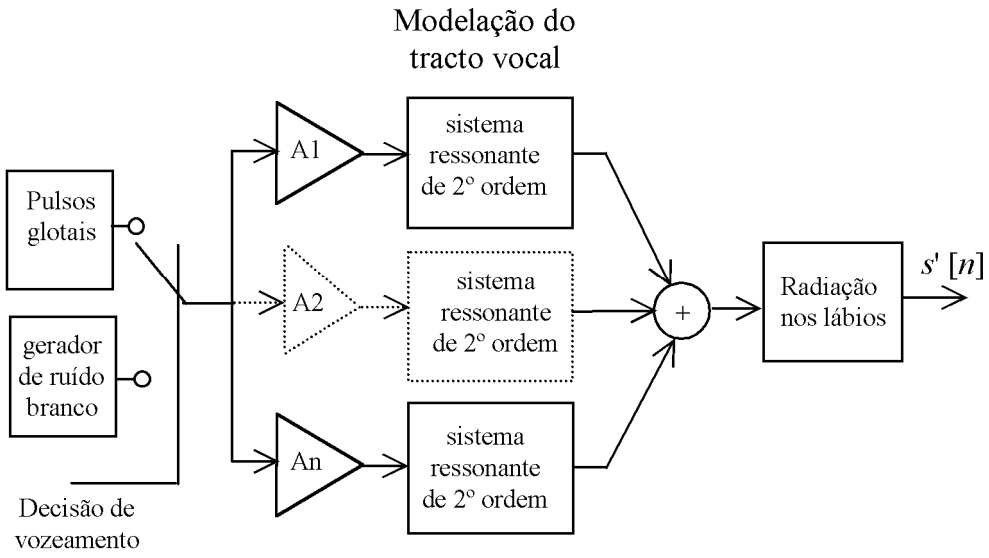

Figura 4.8
Sintetizador paralelo de formantes.

Tal como no *vocoder* LPC, quer na configuração em série quer na paralela, os parâmetros do sintetizador deverão ser alterados em intervalos entre os 5 e os 30 ms.

Um dos sintetizadores de formantes mais bem sucedidos, nomeadamente em aplicações de síntese a partir de texto, é o sintetizador de formantes introduzido por Klatt [Klatt (80)] em 1980.

4.5 Síntese por concatenação

Em muitos sistemas, como por exemplo nos sistemas de síntese de fala a partir de texto, é necessário produzir fala a partir de uma sequência fonética e respectiva informação prosódica (energia, duração e frequência fundamental), existindo um primeiro módulo que converte a sequência ortográfica nestes parâmetros. Para sintetizar fala, segmentos fonéticos pré-gravados são concatenados, num modelo a que se dá o nome de síntese por concatenação.

Uma das aproximações no desenvolvimento destes sistemas é o armazenamento da forma de onda de cada segmento fonético, mas as dificuldades na alteração da informação prosódica e na modelação do efeito da coarticulação torna estes sistemas de difícil implementação. Uma outra aproximação, cujo esquema de blocos é apresentado na figura 4.9, utiliza um modelo paramétrico de síntese trama-a-trama, como por exemplo a síntese por predição linear ou síntese de formantes. A informação espectral (coeficientes LPC ou formantes) de cada trama do segmento fonético é armazenada num livro de código e obtida através do índice do segmento respectivo. O livro de código deverá ter várias instâncias de cada segmento, de modo a poder modelar quer os efeitos da coarticulação (*e.g.*, *trifones* - segmentos em diferentes contextos à esquerda e à direita) quer variações na prosódia. A duração deve ser reajustada e os coeficientes interpolados entre segmentos. A energia e a frequência fundamental são também parâmetros de controlo destes modelos.

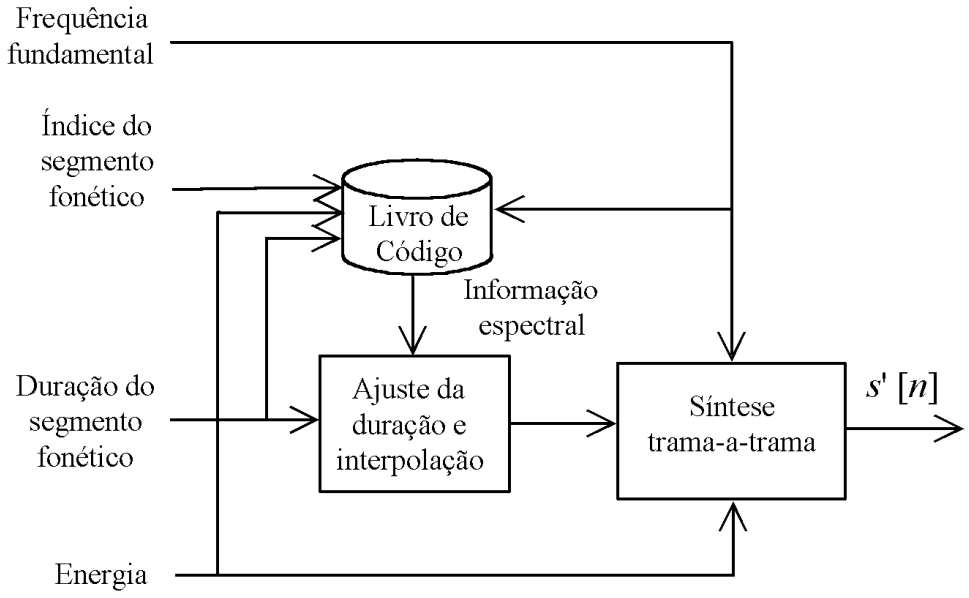

Figura 4.9
Sintetizador por concatenação de segmentos fonéticos.

Mesmo que a entrada do sistema seja uma sequência de segmentos fonéticos, não é obrigatório que a síntese do sinal seja processada com estas unidades. Os *difones*, definidos como o segmento entre os centros de dois segmentos fonéticos consecutivos, são dos segmentos mais utilizados em sistemas de síntese de fala a partir de texto, sendo a sua principal vantagem a concatenação em zonas estáveis, embora se deva arranjar critérios de minimizar as variações espectrais no ponto de concatenação [Sagisaka (95)] [Stylianou (98-b)].

Alternativamente à síntese em que os parâmetros sobre a envolvente espectral estão armazenados num livro de código, estes podem ser gerados por regras. Embora mais eficiente sobre o ponto de vista do armazenamento, esta técnica tem no entanto dificuldades na modelação da dinâmica dos parâmetros de controlo.

4.6 Disfarce da informação

A utilização de codificadores de fala em que o tráfego é efectuado através do protocolo IP (*Internet Protocol*). Uma das desvantagens deste protocolo é não ser garantido que todos os pacotes sejam entregues em tempo real, sendo necessário "disfarçar" a informação dos pacotes não entregues.

A recomendação do ITU-T G.711 (64 kbits/s *companding* amostra-a-amostra) foi recentemente utilizada em transmissão por pacotes, sendo em 1999 adoptado o Apêndice 1 que estabelece recomendações para disfarce de grande qualidade e baixa complexidade de pacotes perdidos. As suas principais características são:

- Pacotes (tramas) de 10 ms (80 amostras);
- Predição de longa duração com estimação da frequência fundamental através do método da co-variância, numa janela de 160 (20 ms) amostras e um intervalo de procura entre as 40 amostras (200 Hz) e as 120 amostras (66,7 Hz);
- Sobreposição com janelas triangulares de ¼ de período fundamental na zona anterior ao primeiro pacote perdido, de modo a minimizar descontinuidades perceptualmente audíveis na zona de *colagem*, impondo um atraso de 3,75 ms (120/4=30 amostras);
- Atenuação do sinal a partir do segundo pacote consecutivo perdido, de modo a que o sinal desvaneça completamente a partir do sexto pacote consecutivo perdido;
- Utilização de dois períodos fundamentais a partir do segundo pacote consecutivo perdido e de três períodos partir do terceiro, de modo à zona regenerada não soar demasiado tonal;
- Sobreposição final de ¼ período fundamental mais 4 ms por cada pacote perdido a partir do segundo, de modo a tornar mais suave a transição no fim da recuperação;

Como exemplo, é apresentado na figura 4.10 o disfarce do sinal para três tramas consecutivas perdidas, sendo mostradas a zona de sobreposição inicial e final. Na figura 4.11 é apresentado um troço do sinal com 3 pacotes recuperados e comparado com o sinal original. Apesar de diferenças significativas, tal como num sinal de fala real não existem descontinuidades e as características espectrais do sinal variam suavemente.

Em codificadores que utilizem predição linear os coeficientes das tramas perdidas são uma repetição dos valores das últimas tramas, normalmente atenuadas em amplitude e com aumento da largura de banda dos formantes, de modo a tender o sinal para ruído branco.

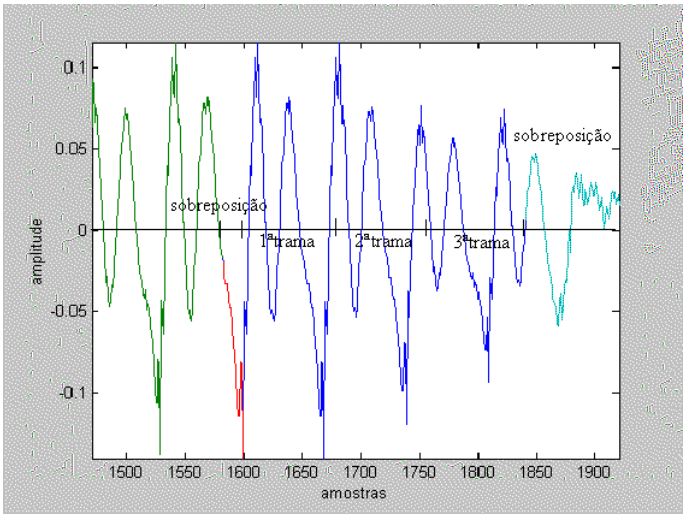

Figura 4.10
Exemplo de reconstrução de três pacotes consecutivos perdidos, sendo mostradas as zonas de sobreposição inicial e final e os três pacotes perdidos.

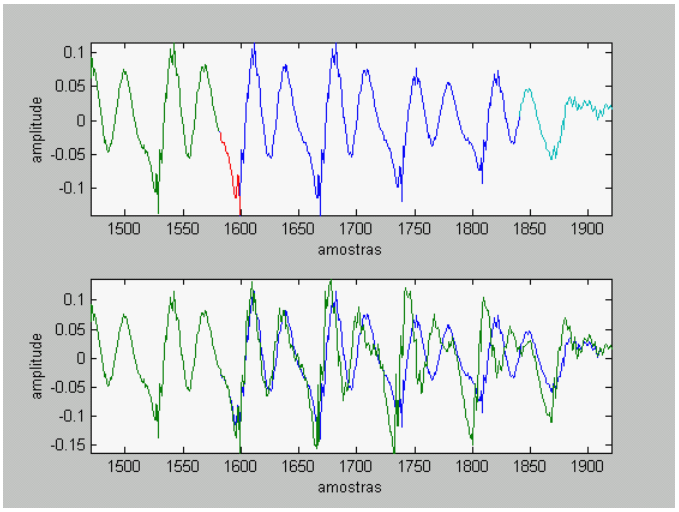

Figura 4.11
Em cima, exemplo já mostrado na figura 4.10, comparado na janela de baixo com o sinal original.