Relatório Técnico

Sintese de Voz com Qualidade

Geraldo Lino de Campos

Evandro Bacci Gouveia

1 – Introdução

Este relatório apresenta a implementação de um sistema de síntese de voz a partir de fonemas utilizando a técnica de predição linear excitada por código. A idéia básica consiste em formar uma biblioteca e unidades básicas, fonemas, difones, trifones etc, e concatenar essas unidades para a produção da fala.

Neste relatório apresenta-se a teoria envolvida e os principais resultados a que foi possível chegar durante a implemetação do sistema.

2 – Fundamentos Teóricos

Na técnica de análise por *Predição Linear*, um sinal é modelado como uma combinação linear de seus valores passados e presente e dos valores passados de uma entrada hipotética a um sistema cuja saída é o sinal dado.

Os parâmetros desse modelo são então obtidos pela minimização do erro no sentido dos mínimos quadrados, isto é, minimiza-se a soma dos quadrados das diferenças entre as amostras reais e as preditas, e, com isso, obtém-se um conjunto único de coeficientes de predição.

Dado um sinal contínuo no tempo $s(t)$ este sinal é amostrado, obtendo o sinal $s(nT)$ discreto no tempo, onde n é uma variável inteira e T e o intervalo de amostragem. A freqüência de amostragem é, então, $f_s=1/T$. Abreviaremos $s(nT)$ por s_n, para facilitar a notação.

No modelo adotado, o sinal s_n será considerado como a saída de um sistema com entrada desconhecida u_n, de forma que se mantenha a relação

$$s_n = -\sum_{k=1}^{p} a_k \hat{s}_{n-k} + G\sum_{l=0}^{q} b_l u_{n-l} \qquad (1)$$

onde a_k, $1 \le k \le p$, b_l, $1 \le l \le q$ e o ganho G são os parâmetros do modelo. Sem perda de generalidade, consideraremos $b_0=1$.

Especificando a equação acima no domínio de freqüência, através da transformada *z*, obtemos a função de transferência$H(z)$ do sistema

$$H(z) = \frac{S(z)}{U(z)} = G \frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

onde

$$S(z) = \sum_{n=-\infty}^{\infty} s_n z^{-n} \tag{3}$$

é a transformada *z* de s_n e $U(z)$, analogamente, é a transformada *z* de u_n.

O uso deste modelo geral, com pólos e zeros na função de transferência, no entanto, torna a solução do problema, isto é, a determinação dos a_k e b_l bastante incômoda do ponto de vista computacional, já que implica na resolução de um sistema com muitas variáveis. Além disso, existem métodos heurísticos bastante eficientes para modelos mais simplificados. A função de transferência do trato vocálico, para sons produzidos com trem de pulsos e sem componentes nasais, não apresenta zeros, o que sugere o uso do modelo só com pólos. Para sons nasais e sons gerados por excitação por ruído, no entanto, podem aparecer zeros. O efeito de um zero numa função de transferência, porém, pode ser conseguido incluindo mais pólos, uma vez que

$$(1 - az^{-1}) \approx \frac{1}{1 + az^{-1} + a^2 z^{-2} + \ldots}$$

se $|a| < 1$, o que é o caso para zeros no círculo unitario.

Além disso, os zeros tem muito menor influência na percepção dos sons. Enquanto os pólos determinam as freqüências de ressonância, os zeros, em geral, apenas alteram a forma do espectro [Ata71].

Assim, o modelo adotado terá como função de transferência

$$H(z) = \frac{G}{A(z)} \tag{4}$$

onde

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{5}$$

é o chamado *filtro inverso* do sistema.

No domínio do tempo, as equações acima correspondem a assumir que o sinal s_n pode ser obtido pela combinação linear de valores passados desse sinal e alguma entrada u_n

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{6}$$

desde que se conheçam os*p* parâmetros a_k e o ganho *G*.

2.1– Parâmetros do Preditor Linear

Os par^ametros do preditor linear podem ser determinados atrav'es de diversas t'ecnicas, como se vera a seguir.

2.1.1 Método de Autocorrelações

Suponha que a entrada u_n do sistema seja totalmente desconhecida. Neste caso, o sinal s_n só pode ser estimado pela soma ponderada de amostras passadas.

Seja $\hat{s}_n$ essa aproximação, isto é

$$\hat{s}_n = -\sum_{k=1}^{p} a_k s_{n-k} \tag{7}$$

Assim, o erro entre o valor real s_n e o preditor $\hat{s}_n$ será

$$e_n = s_n - \hat{s}_n = s_n + \sum_{k=1}^{p} a_k s_{n-k} \tag{8}$$

Essa soma pode ser realizada sobre um intervalo de duração infinita, que leva ao método de resolução conhecido como *método de autocorrelações*, ou de duração finita, que leva ao *método de covariância*. Aqui, faremos uso do método de autocorrelações. O erro quadrático médio total será, então

$$E = \sum_{n=-\infty}^{\infty} e_n^2 = \sum_{n=-\infty}^{\infty} \left(s_n + \sum_{k=1}^{p} a_k s_{n-k}\right)^2 \tag{9}$$

A minimização será obtida fazendo-se

$$\frac{\partial E}{\partial a_i} = 0, \quad 1 \le i \le p$$

(10)

A partir de (9) e (10) obtém-se

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_{n-k} s_{n-i} = -\sum_{n=-\infty}^{\infty} s_n s_{n-i}, \quad 1 \le i \le p \tag{11}$$

A resolução do sistema de equações (11) fornece os valores a_k que minimizam E em (9).

Levando-se em conta (9) e (11), pode-se obter para o cálculo do erro quadrático médio total mínimo a expressão

$$E_p = \sum_{n=-\infty}^{\infty} s_n^2 + \sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_n s_{n-k} \tag{12}$$

Definindo-se a função de autocorrelação do sinal s_n como

$$R(i) = \sum_{n=-\infty}^{\infty} s_n s_{n+i} \tag{13}$$

podemos reduzir as expressões (11) e (12) a

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i), \quad 1 \le i \le p \tag{14}$$

$$E_p = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{15}$$

Note que a função $R(i)$ é par, isto é

$$R(-i) = R(i) \tag{16}$$

Com isso, é fácil verificar que, quando expandimos (14), obtemos a matriz

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ R(2) & R(1) & R(0) & \dots & R(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \\ a_p \end{bmatrix} = -\begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \dots \\ R(p) \end{bmatrix} \tag{17}$$

que é uma matriz simétrica e onde os elementos de cada diagonal são iguais. Tal tipo de matriz é chamada de matriz de Toeplitz. Da matriz também fica claro que se dividirmos todos os coeficientes $R(i)$ por uma constante a solução do sistema não se altera.

Escolhendo-se a constante $R(0)$ podemos trabalhar com o *coeficiente de autocorrelação normalizad***o**

$$r(i) = \frac{R(i)}{R(0)} \tag{18}$$

Na prática, costuma-se janelar o sinal s_n usando uma função de janelamento w_n obtendo a função s_n' dada por

$$s_n' = \begin{cases} s_n w_n & 0 \le n \le N-1 \\ 0 & \textit{no complemento.} \end{cases} \tag{19}$$

E importante a escolha de uma função adequada para se realizar o janelamento. No método utilizado, pode-se notar que, no início da janela, tentam-se predizer amostras não nulas a partir de amostras anteriores que foram arbitrariamente fixadas em zero, e, no fim da janela, tenta-se prever o valor zero como somatório de valores não nulos. Pode-se concluir que, sem uma função de janela mento conveniente, o erro obtido seria muito grande.Por isso, é importante que essa função de enquadramento tenha valores próximos a zero nas extremidades, como é o caso das funções de Hamming ou de Hanning. Uma função como a retangular não seria, por esse mesmo motivo, uma boa escolha.

Determinados os parâmetros $\hat{a}_k$ resta, na equação (6), a determinação do ganho G. Este pode ser obtido igualando-se a energia das amostras do sinal analisado com a energia das amostras preditas.

Mostra-se em [Mak75] que o ganho é dado por

$$G^2 = E_p = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{20}$$

2.1.2 Algoritmo de Durbin

O cálculo dos coeficientes $a_k, 1 \leq k \leq n$ como visto por (17), pode ser realizado pela resolução de um sistema de p equações a p incógnitas. Existem inúmeros métodos para a resolução de um sistema de tal tipo. Como nos restringimos ao método de autocorrelações, porém, podemos aproveitar a característica de a matriz resultante ser de Toeplitz para utilizarmos um algoritmo bem mais eficiente que os métodos mais gerais de resolução. Este algoritmo, de Durbin, consiste nas seguintes equações:

$$E_0 = R(0) \tag{21}$$

$$k_i = -\frac{R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{E_{i-1}} \tag{22}$$

$$a_i^{(i)} = k_i \tag{23}$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}, \quad 1 <= \text{ j } <= \text{ i-1} \tag{24}$$

$$E_i = (1-k_i^2) E_{i-1} \tag{25}$$

As equações (21) a (25) são resolvidas recursivamente para $i = 1,2, \ldots, p$ e a solução final é dada por:

$$a_j = a_j^{(p)}, \quad 1 \leq j \leq p \tag{26}$$

Observe-se que para se calcular os coeficientes de predição de ordem p é necessário calcularmos os coeficientes de predição de todos os preditores de ordem menor que p.

2.1.3 Algoritmo de Le Roux e Gueguen

As variáveis intermediárias k_i do método de Durbin, chamadas de *coeficientes PARCOR* ou *coeficientes de reflexão*, podem ser obtidas diretamente a partir do coeficiente de correlação normalizado $r(i)$. Os parâmetros de predição linear a_i e as variáveis k_i são equivalentes, no sentido de que caracterizam de modo único o preditor linear. A vantagem de se utilizar os coeficientes de reflexão é que estes possuem melhores características de quantização e interpolação que os coeficientes de predição linear.

O termo coeficiente de reflexão se deve a uma interpretação física que aparece quando se modela o trato vocálico por uma sucessão de tubos de diâmetros diferentes. k_m é o coeficiente de reflexão entre o tubo m e o tubo m+1.

Em [Rou77] deduz-se as seguintes equações:

$$e^{h+1_i} = e^{h_i} + k_{h+1} e_{h+1-i}^{h} \tag{27}$$

$$k_{h+1} = -\frac{1}{e^{h_{h+1}}} e^{h_0} \tag{28}$$

$$e_0^{h+1} = e^{h_0}(1-k_{h+1}^2) \tag{29}$$

que dão a solução recursiva para as variáveis k_m a partir dos valores de $e^{0_i} = r(i)$.

Os coeficientes de predição a_k podem ser obtidos a partir do coeficientes de reflexão através das equações recursivas:

$$a^{i}_{i} = k_i \tag{30}$$

$$a^{i}_{j} = a^{i-1}_{j} + k_i a^{i-1}_{i-j}, \quad 1 \le j \le i-1 \tag{31}$$

$$a_j = a^{p}_{j}, \quad 1 \le j \le p \tag{32}$$

Pode-se mostrar que o erro normalizado para o preditor de ordem *h* será

$$V_h = \prod_{h}^{i=1} (1-k_i^2) \tag{33}$$

Em [Mar73] prova-se que

$$|k_m| < 1 \tag{34}$$

e isso garante a estabilidade do filtro modelado. Em função desse resultado, vê-se também que o erro normalizado V_h decresce com a ordem *h* do filtro.

No método introduzido em [Rou77], todas as variáveis intermediárias e_h tem valores entre −1 e 1 desde que se utilizem os coeficientes de autocorrelação normalizados, o que o torna bastante conveniente para implementação em processadores de ponto fixo.

2.1.4 Pares de Linhas Espectrais

A análise por predição linear resulta num filtro $\frac{1}{A_p(z)}$ da forma

$$\frac{1}{1}A_p(z) = \frac{1}{1+a_1 z^{-1}+a_2 z^{-2}+ \ldots +a_p z^{-p}} \tag{35}$$

para o qual vale a relação

$$A_{p+1}(z) = A_{p(z)-k_{p+1}} z^{-(p+1)} A^{-1}_{p(z)} \quad n = 0,1, \ldots ,p-1 \tag{36}$$

e onde $A_{0(z)=1}$.

Consideremos dois casos artificiais extremos de condições de contorno, quais sejam, $k_{p+1}=1$ e $k_{p+1}=-1$. Esses casos correspondem, respectivamente, a um fechamento total e uma abertura total da glótis no modelo de tubos acústicos.

Nessas condições, obtemos, de (36), os polinômios

$$\begin{aligned} P(z) &= A_{p(z)} - z^{-(p+1)} A_p(z^{-1}) \\ &= 1+(a_1-a_p)\, z^{-1}+ \ldots \\ &\quad +(a_p-a_1)\, z^{-p}+z^{-(p+1)} \end{aligned} \tag{37}$$

para o caso $k_{p+1}=1$ e

$$\begin{aligned} Q(z) &= A_p(z)+z^{-(p+1)} A_p(z^{-1}) \\ &= 1+(a_1+a_p) z^{-1}+ \ldots \\ &\quad +(a_p+a_1) z^{-p}+z^{-(p+1)} \end{aligned} \tag{38}$$

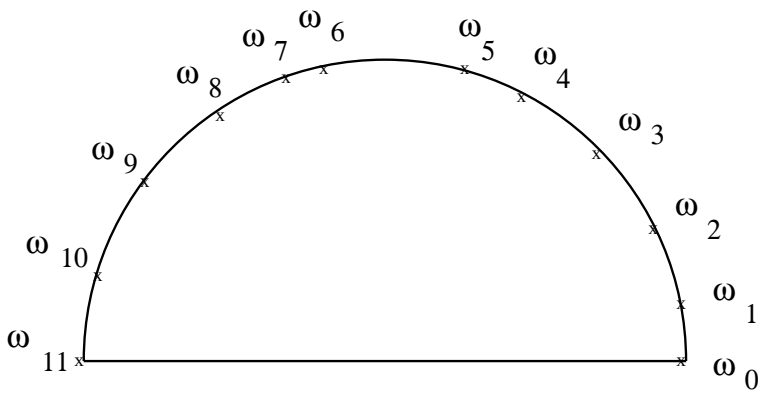

Figura 1 - Pares de linhas espectrais

para $k_{p+1}=-1$.

Por comodidade, vamos nos limitar ao caso de p par. Em [Sug88] mostra-se que os polinômios $P(z)$ e $Q(z)$ podem ser expressos por:

$$P(z) = (1-z^{-1}) \prod_{i=2,4,\ldots,p} (1-2z^{-1}\cos\omega_i + z^{-2}) \tag{39}$$

e

$$Q(z) = (1+z^{-1}) \prod_{i=1,3,\ldots,p-1} (1-2z^{-1}\cos\omega_i + z^{-2}) \tag{40}$$

E fácil verificar pelas equações acima que $e^{j\omega_i}$ $i=1,2,\ldots,p+1$ são raízes dos polinômios $P(z)$ e $Q(z)$. Os parâmetros ω_i $i=1,\ldots,p$ são definidos como os parâmetros de *pares de linhas espectrais* (parâmetros LSP). Observe-se que $\omega_0=0$ e $\omega_{p+1}=\pi$ são sempre raízes de $P(z)$ e $Q(z)$ e são excluídos dos parâmetros LSP. Assim, os parâmetros LSP podem ser interpretados como as freqüências de ressonância do trato vocálico sob as duas condições de contorno artificiais extremas, na glótis.

Os polinômios $P(z)$ e $Q(z)$ possuem algumas propriedades bastante interessantes, dentre as quais (figura 1):

- Todas as raízes de $P(z)$ e $Q(z)$ estão no círculo unitário.
- As raízes de $P(z)$ e $Q(z)$ estão ordenadas, e se alternam,

ou seja:

$$0=\omega_0<\omega_1<\omega_2<\ldots<\omega_{p-1}<\omega_p<\omega_{p+1}=\pi \tag{41}$$

As duas propriedades acima garantem a estabilidade do filtro, sendo, portanto, condições a serem respeitadas quando se faz a quantização destes parâmetros.

A potência da função de transferência $H(z)$ pode ser calculado usando parâmetros LSP como se segue.

observe-se inicialmente que

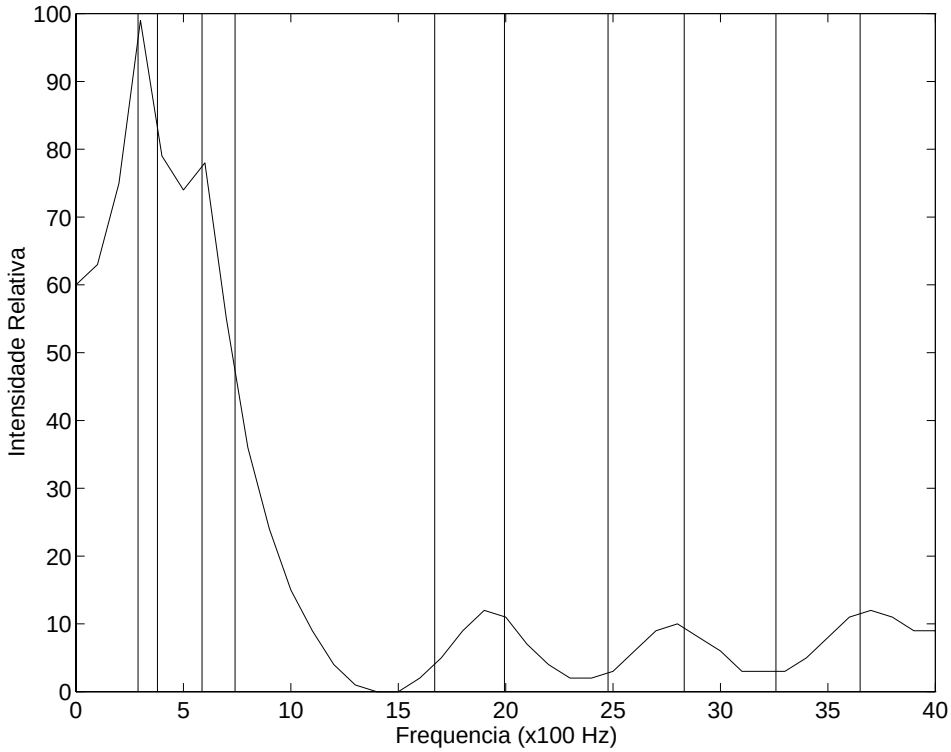

Figura 2 - Pares de linhas espectrais e espectro

$$A_{p(z)} = \frac{2}{P(z)+Q(z)} \tag{42}$$

Assim, temos

$$\begin{aligned} |H(e^{j\omega})|^2 &= \frac{1}{|A^{j\omega}_{p(e)}|^2} \\ &= \frac{4}{P(e^{j\omega})+Q(e^{j\omega})|^2} \end{aligned} \tag{43}$$

que pode ser calculado como

$$|H(e^{j\omega})|^2 = \frac{2^{-p}}{\sin^2\frac{\omega}{2}\prod_{i=2,4,\ldots,p}(\cos\omega - cos\omega_i^2) + \cos^2\frac{\omega}{2}\prod_{i=1,3,\ldots,p-1}(cos\omega - \cos\omega_i^2)} \tag{44}$$

A equação (44) implica que $|H(e^{j\omega})|^2$ tem uma ressonância bastante grande na freqüência ω quando pelo menos dois parâmetros LSP têm valores próximos de ω (figura 2). Assim, os parâmetros LSP representam um filtro que só tenha pólos pela densidade de distribuição de *p* freqüências discretas.

Existem outros conjuntos de parâmetros, além dos coeficientes de reflexão e dos parâmetros LSP, que trazem informações equivalentes a respeito do preditor, no sentido de que trazem a mesma informação espectral. Podemos citar, por exemplo, os coeficientes de predição, os pólos da função de transferência *H*(*z*), os coeficientes cepstrais, etc. Em [Rab78] apresentam-se as relações de transformação entre vários destes parâmetros.

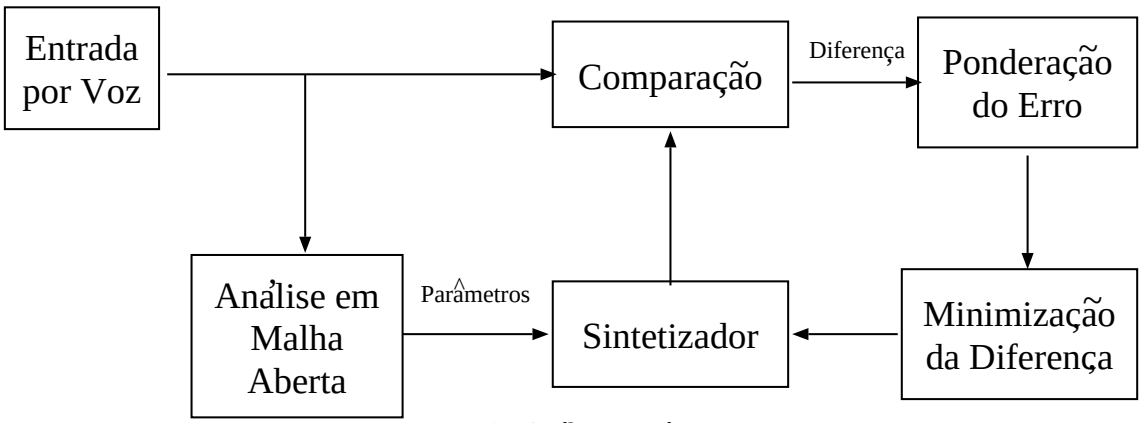

Figura 3 - Análise por síntese

Em [Vis75] apresenta-se um estudo comparativo entre vários conjuntos de parâmetros, com relação a propriedades de quantização. Entre esses conjuntos de parâmetros, incluem-se os coeficientes de reflexão e de predição linear, mas não os parâmetros LSP. Iniciam colocando duas propriedades que são importantes quando se faz a quantização dos parâmetros. A primeira delas é a estabilidade do filtro, que significa que os pólos da função de transferência $H(z)$ continuarão dentro do círculo unitário após a quantização. A segunda é a de ordenação natural dos parâmetros. Os parâmetros a_i, por exemplo, possuem uma ordenação intrínseca. Dos conjuntos estudados pelos autores, só os coeficientes de reflexão possuem as duas propriedades.

Analisando estudos sobre desvio espectral devido à quantização, sobre considerações computacionais, sobre estabilidade e ordenação natural, os autores citados concluem que os coeficentes de reflexão são o melhor conjunto de parâmetros para se utilizar. Além disso, são o único conjunto para o qual os valores já calculados podem ser utilizados quando se muda a ordem do preditor, isto é, k_i,i, não muda quando se muda p.

Em [Sug88] argumenta-se que os parâmetros obtidos usan-do o método de pares de linhas espectrais possuem propriedades de quantização e interpolação melhores que os coeficientes de predição ou os coeficientes de reflexão.

Em [Pap87] afirma-se que os parâmetros LSP como excelentes para a síntese, pois são os mais adequados para a interpolação requerida na síntese. A conclusão que se pode tirar é que o armazenamento e, para certas aplicações, a transmissão devem ser feitos por pares de linhas espectrais. O cálculo destes parâmetros, no entanto, exige uma capacidade computacional grande, e uma alternativa seria, então, o cálculo dos coeficientes de reflexão ou de predição, e a conversão destes para parâmetros LSP, para eventuais quantizações e/ou interpolações. A escolha entre o método de Durbin ou de LeRoux-Gueguen seria, então baseada na disponibilidade ou não de processador de ponto flutuante.

2.2– Estimação do Espectro

O espectro de potências $P(\omega)$ do sinal de voz s_n pode ser estimado a partir dos parâmetros de predição linear, já que estes, conforme foi colocado anteriormente, são uma estimativa do sinal.

Neste método, é feito um mapeamento de freqüências no círculo unitário no plano e ω_B, a $\omega=0$ corresponda $\theta=0$ e a $\omega=\omega_B$ corresponda $\theta=\pi$. A metade inferior do círculo unitário é completada de maneira que $P(\omega)=P(-\omega)$. Assim, sendo $S(z)$ a transformada z do sinal s_n

$$P(\omega) = |S(e^{j\omega})|^2 \tag{45}$$

Este espectro será modelado por $\hat{P}(\omega)$ usando a função de transferência $H(z)$ do modelo de Predição Linear, isto é

$$\hat{P}(\omega) = |H(e^{j\omega})|^2 = \frac{1}{G^2}|A(e^{j\omega})|^2 \tag{46}$$

onde $A(z)$ é o filtro inverso dado por(5).

Desta forma, o espectro de potências do sinal pode ser estimado através dos parâmetros de predição linear por

$$\hat{P}(\omega) = \frac{G^2}{|1+\sum_{k=1}^{p} a_k e^{-jk\omega}|^2} \tag{47}$$

A ordem do preditor está relacionada com a precisão com que esta estimativa do espectro será realizada, ou seja, quanto menor a ordem, mais suave será a curva de aproximação do espectro. Este método tende, portanto, a fornecer a envoltória do espectro.

2.3– Análise por Síntese

A idéia básica da técnica conhecida como {êm análise por síntese} é bastante simples. Num sistema qualquer, conhecendo-se o processo, ou a função de transferência entre a entrada e a saída, e a saída, basta testar todas as entradas possíveis, e a entrada que gerar a saída mais próxima daquela já conhecida é a entrada que se buscava.

Aplicado ao processamento de voz (figura 3), dada uma elocução, pode-se obter o filtro que modela aquela elocução, por exemplo, através de um modelo de predição linear. Dados o filtro e a elocução, basta procurar a função de excitação entre todas as funções possíveis. E claro que se não se fizer alguma limitação nessa busca, essa é uma técnica que não permite chegar a um resultado em tempo finito. A busca pode tornar-se viável desde que se limite essa busca a um conjunto finito de possibilidades, e que seja orientada pelos conhecimentos que se possuem sobre as características do sinal de excitação. A dificuldade vai estar relacionada apenas ao tamanho desse conjunto de funções de entrada que se quer buscar.

O algoritmo CELP (*Code Excited Linear Prediction*) limita essa busca a um conjunto finito de vetores, chamado livro de código (*codebook*). Os elementos desses vetores são amostras representativas de um sinal de excitação, ou sinais que guardam grande semelhança com o sinal a ser representado, como ruído Gaussiano, por exemplo. Utilizando o mesmo livro de código para a análise e para a posterior síntese, basta termos o índice do vetor que produziu o menor erro entre o sinal original e o reconstruído para conseguirmos reproduzir aquela elocução.

A grande dificuldade desse algoritmo é a busca pela entrada que melhor gera a saída, o que equivale à análise da elocução de entrada. A elocução de entrada é dividida em segmentos, que são analisados. Aumentando-se o tamanho desses segmentos, aumenta a demanda computacional do algoritmo.

Aumentando-se o tamanho do livro de código, e, portanto, o número de vetores com os quais se comparam esses segmentos, também aumenta essa demanda. Uma maneira de reduzir a complexidade computacional exigida é o uso de um ganho separado para ajustar a escala, enquanto a forma é dada pelo vetor de excitação.

2.3.1 Critério de Erro

Já que o objetivo do algoritmo é encontrar a excitação que gera o menor erro entre o sinal original e o sinal reconstruído, é necessária a escolha de um critério de erro adequado. Um critério bastante comum, graças sobretudo a sua simplicidade e desempenho, é o de mínimos quadrados. Contudo, aproveitando características de mascaramento do sistema auditivo, é possível o uso de critérios de erro que resultem num ruído menor no sinal sintetizado.

Por exemplo, em [Kro88] demonstra-se experimentalmente que o sistema auditivo tem uma capacidade limitada para detetar pequenos erros em faixas de freqüência em que o sinal de voz tem maior energia, como nas regiões das formantes. Assim, o ruído de quantização deve ser distribuído levando-se em conta a distribuição de energia do sinal original. Isso pode ser feito pela minimização do erro ponderado. Um filtro de ponderação do erro bastante adequado, segundo {Kro88], é:

$$W(z) = \frac{A(z)}{A(z/\gamma)} \tag{48}$$

ou seja

$$W(z) = \frac{1+\sum_{i=1}^{p} a_i z^{-i}}{1+\sum_{i=1}^{p} a_i \gamma^i z^{-i}} \tag{49}$$

O parâmetro γ é uma fração entre 0 e 1 que controla a energia do erro nas regiões de formantes.

2.3.2 Algoritmo de Busca

Consideremos uma busca que seja realizada em dois estágios. O filtro de predição linear não é suficiente para a representação exata da elocução analisada. O resíduo ponderado da predição linear, somado aos erros de codificação introduzidos nas buscas passadas, é o vetor-objetivo do primeiro estágio é. O objetivo do segundo estágio é o objetivo do primeiro menos a excitação encontrada no primeiro estágio.

Sejam os vetores $\boldsymbol{s}$, $\hat{\boldsymbol{s}}$ e $\boldsymbol{e}$ respectivamente o sinal de voz original, o sinal sintético e o erro ponderado. Seja $\boldsymbol{v}$ o vetor de excitação que se está buscando e $\boldsymbol{u}$ o vetor de excitação do estágio anterior. Lembrando, é uma busca em dois estágios. No primeiro estágio, o vetor $\boldsymbol{u}$ é o vetor nulo.

O vetor de excitação $\boldsymbol{v}$ é caracterizado por um índice i do livro de código e por um parâmetro de ganho g_i. Assim, este vetor pode ser escrito

$$\boldsymbol{v}^{(i)} = g_i \boldsymbol{x}^{(i)} \tag{50}$$

onde $\boldsymbol{x}^{(i)}$ é um vetor do livro de código que está sendo utilizado na busca.

Seja $\boldsymbol{H}$ uma matriz de Toeplitz $L \times L$ triangular inferior contendo em sua primeira coluna a resposta ao impulso $\{h_i\}$ *do filtro de predição linear truncada em amostras.*

$$\boldsymbol{H} = \begin{bmatrix} h_0 & 0 & \dots & 0 \\ h_1 & h_0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ h_{L-1} & \dots & \dots & h_0 \end{bmatrix} \tag{51}$$

e seja $\boldsymbol{W}$ a matriz análoga para o filtro de ponderação de erro

$$\boldsymbol{W} = \begin{bmatrix} w_0 & 0 & \dots & 0 \\ w_1 & w_0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ w_{L-1} & \dots & \dots & w_0 \end{bmatrix} \tag{52}$$

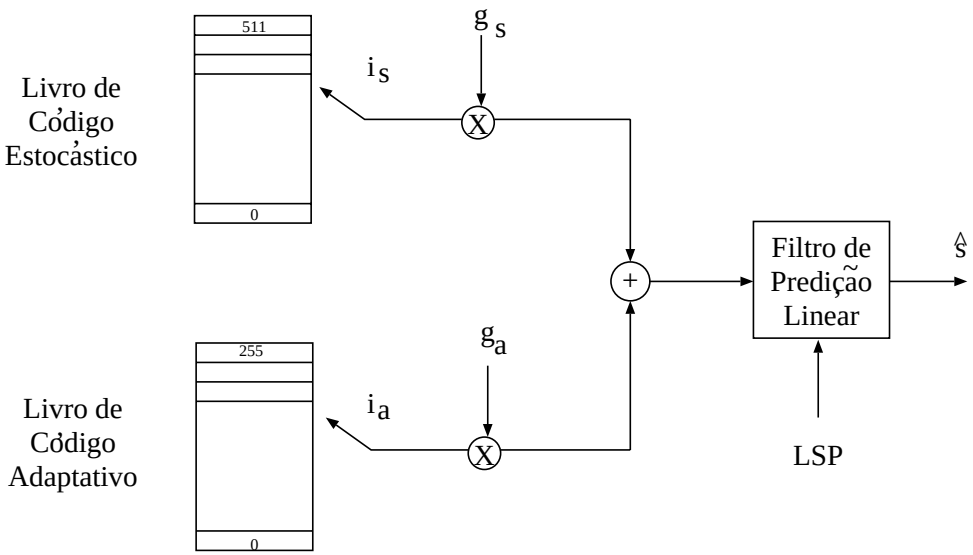

Figura 4 - Predição linear excitada por código

Sendo $\hat{\boldsymbol{s}}^{(0)}$ a resposta do filtro de predição linear com entrada nula, resultado das condições iniciais do filtro, o sinal sintético será dado por

$$\hat{\boldsymbol{s}}^{(i)} = H(\boldsymbol{u}+\boldsymbol{v}^{(i)})+\hat{\boldsymbol{s}}^{(0)} \tag{53}$$

O erro ponderado será, então

$$\boldsymbol{e}^{(i)} = \boldsymbol{W}(\boldsymbol{s}-\hat{\boldsymbol{s}}^{(i)}) \tag{54}$$

ou, levando-se em conta (53)

$$\mathrm{D}\boldsymbol{e}^{(i)} = \boldsymbol{e}^{(0)} - \boldsymbol{WHv}^{(i)} \tag{55}$$

onde o vetor-objetivo é dado por

$$\boldsymbol{e}^{(0)} = \boldsymbol{W}(\boldsymbol{s}-\hat{\boldsymbol{s}}^{(0)})-\boldsymbol{WHu} \tag{56}$$

Desta forma, o erro ponderado $\boldsymbol{e}^{(i)}$ para o vetor *i* do livro de código é

$$\boldsymbol{e}^{(i)} = \boldsymbol{e}^{(0)}-g_i\boldsymbol{y}^{(i)} \tag{57}$$

onde $\boldsymbol{y}^{(i)}$ é o vetor $\boldsymbol{x}^{(i)}$ filtrado, isto é

$$\boldsymbol{y}^{(i)} = \boldsymbol{WHx}^{(i)} \tag{58}$$

O objetivo é minimizar o erro quadrático

$$E_i = \boldsymbol{e}^{(i)^T}\boldsymbol{e}^{(i)} \tag{59}$$

onde o *T* denota transposição. Expandindo, temos

$$E_i = \boldsymbol{e}^{(0)T}\boldsymbol{e}^{(0)}-2g_i\boldsymbol{y}^{(i)T}\boldsymbol{e}^{(0)}+g_i^2\boldsymbol{y}^{(i)T}\boldsymbol{y}^{(i)} \tag{60}$$

Para um dado valor de i, o ganho g_i pode ser calculado derivando E_i e igualando a zero, ou seja

$$\frac{\partial E_i}{\partial g_i} = -2g_i \mathbf{y}^{(i)T}\mathbf{e}^{(0)} + g_i \mathbf{y}^{(i)T}\mathbf{y}^{(i)} = 0 \tag{61}$$

Assim, obtemos que o ganho para o erro quadrático médio mínimo é dado pela razão entre a correlação cruzada do vetor-objetivo e do vetor filtrado e a energia do vetor filtrado

$$g_i = \frac{\mathbf{y}^{(i)^T}\mathbf{e}^{(0)}}{\mathbf{y}^{(i)T}\mathbf{y}^{(i)}} \tag{62}$$

Minimizar o erro E_i em relação a i é equivalente a maximizar o oposto da soma dos últimos dois termos de (60), já que o primeiro termo independe do vetor $\mathbf{f}x_i$. Isso corresponde a maximizar

$$m_i = g_{i(2g_i}\mathbf{y}^{(i)^T}\mathbf{e}^{(0)} - g_i \mathbf{y} sup(i)^T \mathbf{y}^{(i)}) \tag{63}$$

Substituindo (62) em (63), vem

$$m_i = \frac{(\mathbf{y}^{(i)^T}\mathbf{e}^{(0)})^2}{\mathbf{y}^{(i)^T}\mathbf{y}^{(i)}} \tag{64}$$

Assim, buscamos o vetor $\mathbf{x_i}$ que maximiza m_i. Esse vetor terá então sua magnitude multiplicada pelo fator de ganho g_i, resultando no erro ponderado perceptivo mínimo.

2.3.3 Formação dos Livros de Código

Seguindo recomendação de [Fst91], são utilizados dois livros de código na busca citada (figura 4). No primeiro estágio, é utilizado um livro de código adaptativo, o qual modela a periodicidade do sinal, ou o {êm pitch}. No segundo estágio, é utilizado um livro de código estocástico, fixo, que modela o resíduo da última etapa.

2.3.4 Livro de Código Adaptativo

Este livro de código é atualizado com a excitação presente do codificador. Ele guarda, portanto, a história recente das excitações sintéticas. A busca neste livro encontra a seqüência, na história recente das excitações, que melhorrepresente o vetor-objetivo. Utilizando atrasos, por exemplo, entre 20 e 147 amostras, para uma taxa de amostragem de 8 kHz, podem-se obter portanto, freqüências fundamentais entre 54 e 400 Hz. A construção do livro de código adaptativo é descrita a seguir.

Seja $\mathbf{r}$ o livro de códigos presentemente armazenado, onde cada elemento é um vetor com amostras sobrepostas aos outros vetores. Como as amostras são sobrepostas, é possível representar esse livro de código como um único vetor. Suponhamos, para tornar a explicação mais concreta, que cada vetor possui 60 amostras, e que o livro de código possui 147 amostras:

$$\mathbf{r} = [r(-147), r(-146), \ldots, r(-1)] \tag{65}$$

onde o índice −1 indica o último elemento (e, portanto, o mais recente) que foi introduzido no livro de código. Os vetores 60 a 147 do livro de código são formados pelos elementos -60 a -1, -61 a -2, ~...~ até -147 a -88, respectivamente. Para vetores entre 20 e 59 (considerando livro de código com 128 vetores de atraso

por um número inteiro de amostras), o vetor é replicado até se chegar aos 60 elementos, ou seja, o vetor 20 será formado pelos elementos -20 a -1 replicados 3 vezes, e o vetor 59 será formado pelos elementos -59,~...~, -1, -59.

Seja ***r'*** um vetor candidato obtido do livro de código

$$\boldsymbol{r'} = [r'(0), r'(1), \ldots, r'(59)] \tag{66}$$

e seja ***r''*** a concatenação de ***r*** e ***r'***

$$\begin{aligned}\boldsymbol{r''} = [&r(-147), r(-146), \ldots, r(-1), r'(0), r'(1), \ldots, r'(59)] \\ [&r''(-147), r''(-146), \ldots, r''(-1), r''(0), r''(1), \ldots, r''(59)]\end{aligned} \tag{67}$$

Assim, podemos escrever

$$\boldsymbol{r'_M}(i) = \boldsymbol{r''_M} 9i0 = \boldsymbol{r''_M}(i-M) \quad \begin{cases} i = 0, 1, \ldots, 59 \\ M = 20, 21, \ldots, 147 \end{cases} \tag{68}$$

Após completada a busca por todo o livro de código, o livro de código adaptativo ***r*** é atuaalizado com o vetor de excitação escolhido, ***e***, soma dos vetores escolhidos em cada livro de código, adaptativo e estocástico, multiplicados pelo fator de ganho. Esta atualização descarta os elementos mais atrasados, deslocando o livro de código

$$\boldsymbol{r}(i) = \boldsymbol{r}(i+60) \qquad i = -147, -146, \ldots, -61 \tag{69}$$

$$\boldsymbol{r}(i) = \boldsymbol{e}(i+60) \qquad i = -60, -59, \ldots, -1 \tag{70}$$

No caso de atrasos por um número não inteiro de amostras, os vetores do livro de código são formados por interpolação, usando ponderação pelo vetor ***w***

$$w_{f(j)} = h(12(j+f)) \frac{sin((j+f)\pi)}{(j+f)\pi} \begin{cases} j = \frac{-N}{2}, \frac{-N}{2}+1, \ldots, \frac{N}{2}-1 \\ f = \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4} \end{cases} \tag{71}$$

onde h é a função de Hamming

$$h(k) = 0{,}54 + 0{,}!46 cos(\frac{k\pi}{6N}) \tag{72}$$

O número de pontos usados na interpolação é 40, segundo recomendação de [Fst91]. O atraso não inteiro vai ter uma parte inteira M e uma parte fracionária f. Esta parte fracionária vai ser um dos 5 valores apresentados em (71), e determinará qual o conjunto de funções de peso a ser utilizado. Para o cálculo do elemento do vetor ***r'*** pode ser usada a seguinte fórmula recursiva

$$r'_{M+f} = r''_{M+f} = \sum_{j=-N/2}^{N/2-1} w_{f(j)} r''_{M+f}(i-M+j) \begin{cases} i = 0, 1, \ldots, 59 \\ M = 20, 21, \ldots, 147 \end{cases} \tag{73}$$

2.3.5 Livro de Código Estocástico

O livro de código estocástico recomendado por [Fst91] é um livro estocástico com ceifagem central. Isto significa que ele é formado a partir de um livro estocástico gausssiano de média nula, ao qual foi aplicada a seguinte função

$$f_{i(n)} = \begin{cases} c_i(n) & |c_i(n)| > n_{cc} \\ 0 & \textit{no complemento} \end{cases} \tag{74}$$

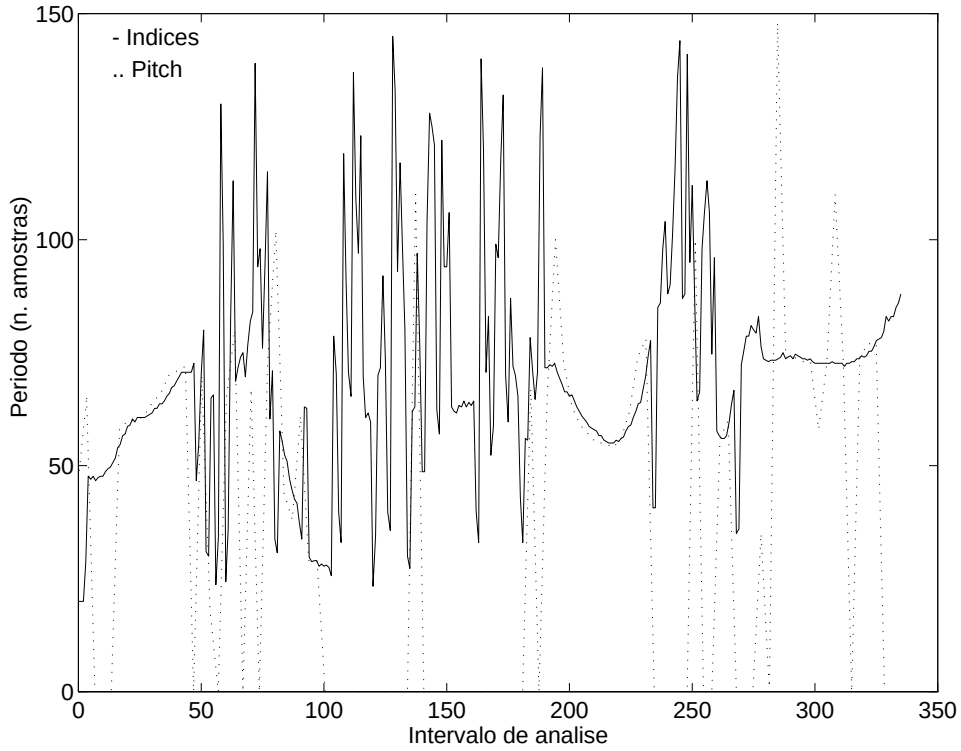

Figura 5 - Elocução "O Leo chegou? Sim, chegou"

onde c_i representa os elementos do livro de códigos estocástico gaussiano, ,$\boldsymbol{f}_i$, o livro com ceifagem central, e n_{cc}, o n'ivel de ceifagem. Segundo KLEIJN et al [], a grande vantagem de se fazer a ceifagem 'e a redu'c 'ai da exig^encia computacional do m'etodo. Como a busca se baseia em produtos de elementos de uma matriz pela coluna da outra, quando este elemento for nulo basta desprezar a coluna correspondente. A mesma refer^encia afirma que n~ao se observa perda de qualidade significativa.

Além da ceifagem central, pode-se trabalhar com livros de código ternários. Definindo a função

$$sign(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases} \tag{75}$$

o livro de código estocástico ternário com ceifagem central será obtido por

$$t_i(n) = sign(f_i(n)) \tag{76}$$

isto reduz ainda mais a demanda computacional do método, substituindo multiplicações por meras comparações. Contrariamente ao livro de código estocástico, ceifagem central do livro de código adaptativo reduz a qualidade do sinal sintético, e, portanto, não é recomendada. Em [Kro88] propoe-se o valor n_{cc}=1,2 como adequado para o nível de ceifagem central, dada a variância unitária, o que faz com que o livro de código tenha 76 % de esparsidade.

3 – Resultados

Para a implementação do sintetizador, foram necessárias três etapas de estudos. A primeira delas foi o entendimento da relação entre as características físicas do sinal de voz, tais como intensidade, pitch etc, e os parâmetros obtidos com a técnica CELP.

Ao analisarmos uma elocução qualquer com esta técnica, obtemos 14 parâmetros, que são os 10 parâmetros de pares de linhas espectrais (parâmetros LSP), e mais o índice e o ganho relativos ao livro de código adaptativo e o índice e ganho relativos ao livro de código estocástico (parâmetros CELP). Os parâmetros LSP estão relacionados à configuração do trato vocálico [Sug88]. Os demais parâmetros representam o sinal de excitação. Sua relação com as características físicas desse sinal foram o objeto de estudos dessa primeira fase da implementação. Ressalte-se que as características físicas importantes para a implementação do sintetizador são o pitch e o ganho do sinal de excitação.

Numa segunda fase, foi implementada uma biblioteca de fonemas. Para se realizar a síntese a partir de textos irrestritos, é conveniente que se utilizem unidades fonéticas de curta duração, tais como os difones. Esses difones são armazenados segundo alguma representação conveniente. No caso, essa representação é a citada acima, que engloba os parâmetros LSP e os parâmetros CELP. Esse conjunto de parâmetros forma a biblioteca que será consultada quando da síntese.

A síntese, em sua forma mais rudimentar, é obtida pela mera concatenação de difones. O resultado obtido, porém, não é muito satisfatório, pois os difones da biblioteca utilizada foram obtidos em contextos diferentes. E necessário, então, algum algoritmo que permita suavizar as transições entre difones, fazendo um tratamento dos parâmetros nas vizinhanças. Esta foi a terceira etapa de estudos para a implementação do sintetizador.

3.1– Relação entre Parâmetros e Características Físicas

Na técnica CELP, o livro de código adaptativo contém um histórico das excitações passadas. Nos primeiros intervalos da análise, portanto, este livro de código contém informações incorretas, que não devem ser utilizadas na síntese. Uma forma de contornar isto é, então, zerar o ganho do livro de código adaptativo nos intervalos iniciais. Anulando também o ganho do livro de código estocástico, obtém-se um sinal de excitação nulo, que irá zerar o conteúdo do livro de código adaptativo. A amostra mais antiga do livro de código adaptativo utilizado corresponde a um atraso de 147 amostras.

Atualizando este livro de código a cada 60 amostras, é necessário zerar o ganho do livro de código adaptativo nos três primeiros intervalos. Basta zerar o ganho do livro de código estocástico nos dois primeiros intervalo para que o livro adaptativo contenha apenas informações corretas sobre a elocução analisada.

3.1.1 Pitch

Suponha o caso ideal de excitação apenas por pulso, com pitch constante. No algoritmo CELP, o índice do livro de código adaptativo indica a excitação atrasada que melhor representa a excitação presente. Se a excitação é constante, esse atraso é exatamente o período dessa excitação. Assim, existe uma forte correlação entre o atraso indicado pelo índice do livro de código adaptativo e o período de pitch, conforme pode ser comprovado pela figura 55 , que traz uma comparação entre o atraso indicado pelo índice do livro de código adaptativo e o período de pitch obtido pela técnica descrita em [Gol69].

Uma das características que se pode salientar é a de que trechos em que ocorre variação bastante grande no atraso correspondem a fonemas surdos, enquanto que trechos onde essa variação é pequena correspondem a fonemas sonoros.

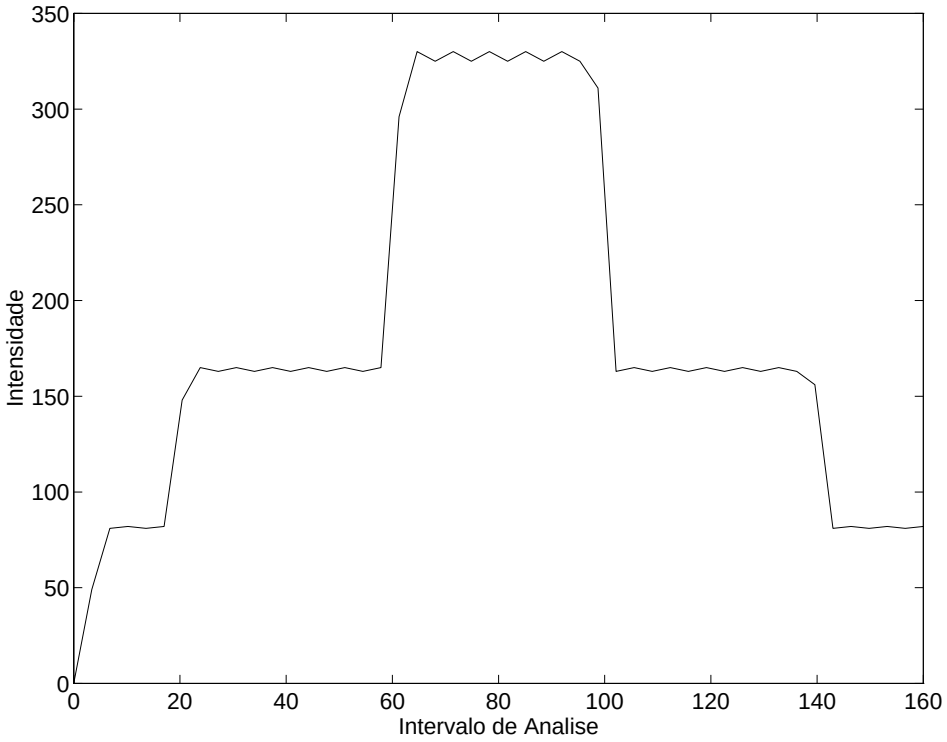

Figura 6 - Intensidade da elocução sintética "aaaaa"

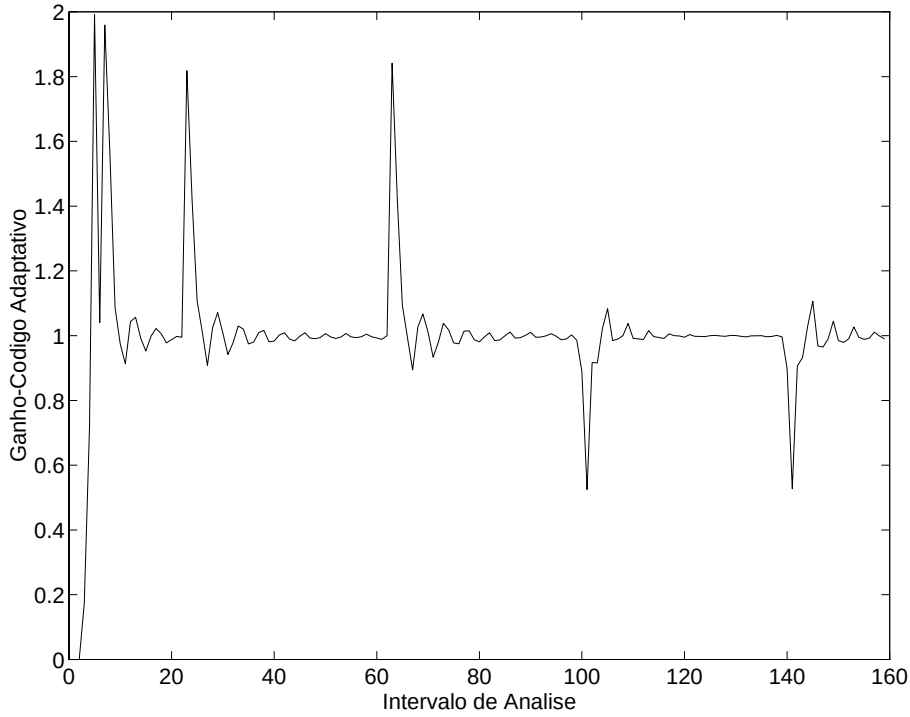

Figura 7 - Ganho da elocução sintética "aaaaa"

Estudos sobre o perfil de variação do pitch estão além do objetivo deste trabalho. Para a implementação do sintetizador, basta saber quais parâmetros devem ser variados para se obter uma variação do pitch. Do exposto, conclui-se que este parâmetro é o índice do livro de código adaptativo, e o atraso por este indicado está diretamente ligado à variação do pitch.

3.1.2 Ganho

Inicialmente, façamos distinção entre duas características relacionadas ao ganho: a intensidade do sinal e a variação dessa intensidade. A primeira característica corresponde ao que se entende geralmente por ‘‘volume’’, e a segunda está relacionada ‘a entonação.

Num instante qualquer i da síntese, a excitação $e(i)$ será dada por

$$e(i) = e(i–D) \times g_a + x \times g_s \tag{77}$$

onde D é o atraso correspondente ao índice do livro de código adaptativo, g_a e g_s são os ganhos dos dois livros de código e x é um elemento do livro de código estocástico.

Conforme já exposto, nos primeiros intervalos de análise os dois ganhos são nulos. Os primeiros $e(i)$ não nulos são dados por

$$e(i) = x \times g_s \tag{78}$$

Dados os parâmetros CELP correspondentes a elocução e multiplicando-se o ganho g_s em todos os intervalos de análise por um mesmo fator K, obtemos que os primeiros $e(i)$ não nulos serão

$$e(i) = K \times x \times G_s \tag{79}$$

e o sinal de excitação num instante i será

$$e(i) = e(i–D) \times g_a + K \times x \times g_s \tag{80}$$

Levando-se em conta (79), no entanto, vê-se que (80) traz, recursivamente, o fator K. Então, multiplicar o ganho g_s para toda a elocução por um fator K corresponde a aumentar por esse fator a intensidade da elocução, linearmente.

Na técnica CELP, o ganho do livro de código adaptativo g_a multiplica um vetor que representa uma excitação passada ao sistema.

Este produto é um dos dois componentes do sinal de excitação.

Supondo, num caso ideal, que o outro componente seja nulo, a excitação presente será a mesma que alguma excitação passada, a menos do ganho g_a. Este ganho tem a função, portanto, de manter, atenuar ou acentuar a excitação passada.

A figura 6 mostra a intensidade da elocução sintética /a/. A elocução dura alguns segundos, e a variação de intensidades ocorreu sem intervalos de silêncio. Foi utilizada elocução sintética para que se possa ter controle sobre as intensidades relativas do sinal ao longo do tempo.A síntese foi realizada com um sistema baseado em [Cam80]. A figura 7 traz o ganho g_a resultante da análise dessa elocução sintética. Da comparação das figuras, pode-se observar que um aumento na intensidade do sinal da ordem de 2 faz com que o ganho g_a assuma um valor próximo de 2. Uma diminuição na intensidade da ordem de 2 faz o ganho g_a assumir valor próximo de 0,5. Nos trechos em que não há variação de intensidade, o ganho g_a se mantém em torno de 1. Assim, o perfil de variação do ganho num sinal sintetizado a partir dos parâmetros CELP é dado pelo ganho do livro de código adaptativo g_a.

3.2– Biblioteca de Fonemas

Utilizando os meios descritos no capítulo anterior, foram gravadas elocuções de palavras que contivessem os difones de interesse, e os trechos correspondentes a esses difones foram selecionados.

Inicialmente, foram selecionadas palavras que contivessem os difones formados pela associação das vogais /a/, /ã/,/e/, /ɛ/, /i/, /o/, /\c o/ e /u/ com as consoantes /b/, /c/, /d/, /f/, /g/, /\v z/, /l/, /ʎ/, /m/, /n/,/\ñ/, /p/, /r/, /rr/, /s/, /t/, /v/, /\v, /s/, /e /z/. Cada vogal foi associada com cada consoante, sendo o difone sempre iniciado pela consoante (/ba/, /bã/, /be/,~...~,/mi/, /mo/,~...~,/zu/). Para cada par formado, foi feita a elocução de uma palavra. A escolha de palavras obedeceu aos seguintes critérios:

- a palavra deveria existir em Língua Portuguesa;
- a vogal do difone deveria ser a vogal tônica da palavra;
- o difone deveria estar no meio da palavra, em geral trissilábica;
- a consoante em cada par deveria ter vogais como fonemas vizinhos;
- a sílaba posterior deveria ser, preferencialmente, iniciada por plosivo.

Esses critérios foram adotados visando facilitar a posterior separação do difone na palavra. Optou-se por retirar os difones de palavras, ao invés de simplesmente fazer as elocuções /ba/, /ca/, etc, buscando naturalidade na forma como esses difones seriam pronunciados. Por esse mesmo motivo, não faria sentido utilizar palavras fictícias.

Optou-se pelo uso de vogais como tônicas por elas se apresentarem, nessa situação, como mais características. Numa palavra como {êm pede}, por exemplo, o primeiro {êm e}, tônico, é pronunciado nitidamente como /é/, enquanto que o segundo é pronunciado como um fonema não muito bem caracterizado, entre /e/ e /i/.

O início da amostragem é sempre com silêncio. Consoantes como /s/, por exemplo, são difíceis de distinguir de silêncio, examinando apenas intensidade ou forma de onda. Por essa razão, optou-se pela escolha de palavras que não se iniciassem pelo difone de interesse. Estando o difone de interesse no meio da palavra, seu início estaria bem caracterizado.

Para facilitar a delimitação do difone, foi escolhido que a sílaba posterior ao difone fosse plosiva. As plosivas têm um intervalo de silêncio no seu início. Assim, há uma distinção nítida entre o final do difone de interesse e o início do difone seguinte. Outras classes de fonemas causariam ruído na vogal do difone, dificultando o isolamento deste.

Para facilitar a caracterização de consoantes, optou-se também por escolher palavras em que essa consoante estivesse cercada por vogais, pois essas são bem caracterizadas, no espectrograma, e definem muito bem os limites da consoante que está no meio.

Utilizando basicamente os mesmos critérios, foram selecionadas palavras que contivessem as semivogais /l/ e /r/ seguindo consoantes, ou seja, os encontros /bl/, /cl, /gl/, /pl/, /tl,/br/, /cr/, /dr/, /gr/, /pr/ e /tr/. Além disso, certas consoantes aparecem no final de sílabas, e é necessário também acrescentá-las à biblioteca. São as consoantes /m/ (te<u>m</u>po), /n/(te<u>n</u>to}, /r/\po<u>r</u>ta e /s/(pa<u>s</u>ta). As vogais (/a/, /ã/,~...~, /u/), isoladamente, também fazem parte da biblioteca.

Tendo as elocuções das palavras que continham os difones expostos acima, foi feita a análise, e, com os parâmetros obtidos nessa análise, foram selecionados os trechos relativos aos difones. A referência básica para seleção desse trecho foi o espectrograma da elocução, e o controle de qualidade definitivo era a síntese do trecho. A duração dos trechos foi normalizada em 195 ms, o que corresponde a 26 intervalos de análise de 7,5 ms.

Os difones, porém, foram obtidos em contextos bastante diferentes, e pelo menos uma característica precisa ser normalizada antes de sua concatenação: a intensidade.

A intensidade de um difone (restringe-se a discussão a difones formadas por vogais, por simplicidade) depende muito mais da vogal do que da consoante. Assim, os difones formados por *consoante+vogal* deveriam, para uma dada vogal, ter intensidades próximas. Foram então obtidas as intensidades de todas as combinações *consoante+vogal*, e, para cada vogal, as intensidades foram normalizadas num valor médio. Essa normalização é bem fácil de ser realizada pela multiplicação do ganho do livro de código estocástico, em cada trecho relativo aos difones, por um fator conveniente. Esse fator conveniente é dado pela razão entre o valor médio da intensidade original e o valor em que se quer normalizar.

Feito isto, tem-se as intensidades ajustadas para cada vogal. A relação entre as intensidades das elocuções de difones com vogais diferentes é mais subjetiva. Para duas vogais, mantém-se o difone que contenha uma delas com intensidade constante, e varia-se a intensidade do outro difone. Obtém-se, então, a relação entre as intensidades dos difones dessas duas vogais que apresentam o melhor efeito subjetivo. Os difones que contém uma dada vogal já tinham intensidades compatíveis. Assim, o ajuste final de intensidades é imediato.

3.3– Interpolações nas Vizinhanças

Considerem-se inicialmente os difones da biblioteca iniciados por consoantes. Eles podem ser divididos em duas classes: os que se iniciam por consoantes oclusivas e os que se iniciam por outras consoantes. A diferença entre elas, basicamente, é que, para os difones da primeira classe, existe um intervalo de silêncio entre este difone e o anterior, numa elocução. Para os da segunda classe, ocorre uma interação maior com o difone anterior. Este intervalo de silêncio causa um isolamento entre os difones que atenua os efeitos da coarticulação.

Os difones da biblioteca que não se iniciam por consoantes são as vogais. A associação de uma vogal com o difone anterior define um ditongo, em que há sempre grande influência entre os difones.

Esse ditongo pode ser formado pela aproximação de duas vogais ou de uma vogal e uma semivogal (/l/, /r/).

Assim, há três possibilidades para os encontros entre difones:

- encontro envolvendo consoante oclusiva;
- não envolvendo consoante oclusiva;
- ditongo.

Seguindo o modelo de predição linear, a produção de voz ocorre pela excitação de um filtro por um sinal conveniente. Assim, a discussão que se segue procura mostrar como interagem, no encontro entre difones, tanto os parâmetros que representam o filtro (parâmetros LSP) quanto os que representam a excitação (parâmetros CELP).

Encontros que envolvam consoante oclusiva são os mais fáceis de tratar. Como a influência do efeito de coarticulação é bem atenuada, não é necessária interpolação entre parâmetros nas vizinhanças para se obter uma qualidade de síntese aceitável. Os difones dessa classe da biblioteca de fonemas caracterizam-se por terem um intervalo de silêncio seguido de uma ‘‘explosão’’, que é o início propriamente dito do difone. Estes difones foram selecionados mantendo-se um intervalo de silêncio da ordem de 30 ms. Na síntese, notou-se que este intervalo era muito curto. Introduz-se, na síntese, então, um intervalo entre os difones que dura cerca de 30 ms.

Inicialmente, o intervalo introduzido era se silêncio. Observou-se, no entanto, que a terminação brusca do difone anterior causava uma sensação subjetiva não muito agradável. Assim, ao invés de silêncio, foi testada a alternativa de interpolação linear dos parâmetros LSP, entre o último intervalo do primeiro difone e o primeiro do segundo, e prolongamento do sinal de excitação, atenuado. Para este prolongamento do sinal de excitação, foi mantido o *pitch* do difone anterior. Para se obter a atenuaçcão deste sinal, basta atribuir ao ganho do livro de código adaptativo um valor menor do que 1 (seção 3.1.2).

Encontros envolvendo difones que se iniciam por outras consoantes que não as oclusivas sofrem grande influência do efeito de coarticulação. A coarticulação, segundo [Sha87], tende a ser um fenômeno que precede o difone,

influenciando mais o difone anterior. Assim, para simular o efeito de coarticulação, fez-se interpolação dos parâmetros LSP num intervalo de cerca de 60 ms, iniciando nos últimos 45 ms do difone anterior, e terminando nos primeiros 15 ms do difone posterior. Foi utilizada interpolação linear dos parâmetros LSP. A interpolação dos parâmetros CELP mostrou-se desnecessária.

Os ditongos, que correspondem a transições suaves entre configurações estáveis do trato vocálico, podem ser modelados por interpolação linear entre parâmetros. Neste caso, porém, o intervalo em que é feita essa interpolação é maior, da ordem de 160 ms, e distribuído simetricamente, começando nos últimos 80 ms do difone anterior e terminando nos primeiros 80 ms do difone posterior, se o encontro for entre vogais, ou um intervalo total de cerca de 80 ms, quando entre semivogal e vogal.

Inicialmente, não se fez nenhuma interpolação dos parâmetros CELP para os encontros vocálicos. Quando se sintetizava uma elocução como /aaa/, ou seja, encontro vocálico envolvendo apenas uma vogal, repetida, o efeito obtido, ao invés de um /a/ longo, era de vários /a/ separados.

Analisando, por CELP, elocuções naturais em que estivessem presentes ditongos, observou-se que a excitação, nesses trechos, era muito parecida com a excitação para única vogal, ou seja, o índice do livro de código adaptativo (

No caso de uma única vogal, o ganho do livro de código estocástico (g_s) assume, nos dois primeiros intervalos não nulos, valores muito altos, da ordem de 1000, e depois se mantém entre 50 e 100, não levando em conta o sinal. Também com os ditongos, numa elocução natural, acontece isto.

Foram realizadas, então, essas três mudanças quanto à excitação, para os ditongos: repetição de i_a na transição, atribuição do valor 1 a g_a, e atenuação de g_s.

Observou-se que a síntese de elocuções como /aaa/ passaram a ser realizadas adequadamente. Ditongos em que houvesse vogais diferentes, porém, eram sintetizadas com intensidade subjetiva muito distante entre uma vogal e outra.

Analisando a dinâmica do processo de síntese por CELP, notamos que o ganho g_s fornece uma referência inicial, que é mantida pelo ganho g_a, já que a excitação inicial depende exclusivamente do livro de código estocástico. Essa referência inicial é diferente para cada vogal. Assim, para que tenhamos a mesma sensação subjetiva de intensidade, é necessário que as vogais sejam sintetizadas com intensidades diferentes, as quais dependerão do ganho do livro de código estocástico.

Na síntese do ditongo, com as mudanças citadas acima, a intensidade é mantida num patamar, pois g_a é mantido no valor unitário. Ditongos como /au/, por exemplo, soam como se o /u/ fosse um urro, enquanto que, no caso de /ua/, mal se ouve o /a/, pois a referência inicial para a segunda vogal está, em ambos os casos, totalmente errada.

Para se contornar isso, ao invés de, na transição, ser mantido $g_{a=1}$, corrige-se a diferença de intensidades, atribuindo-se a g_a a razão entre intensidades que forneça uma sensação subjetiva adequada.

4 – CONCLUSÕES

5 – BIBLIOGRAFIA

[Ata71] ATAL, B. S. and HANAUER, S. L. "Speech analysis and synthesis by linear prediction of the speech wave", Journal of the Acoustical Socitey of America **50**:2, 637-55, Aug 1971.

[Cam80] CAMPOS, G. L. "Síntese de voz para o idioma português", Tese de doutoramento, Escola Politécnica, Universidade de São Paulo, São Paulo, 1980.

[Fst91] Federal standard 1016, telecommunications: analog to digital conversion of radio voice by 4,800 bit/second code excited linear prediction (CELP)}, U. S. Office of Technology and Standards, Feb 1991.

[Gol69] GOLD, B. and RABINER, L. "Parallel processing techniques for estimating pitch periods of speech in the time domain", Journal of the Acoustical Socitey of America **46**:2, 442-8, Aug 1969.

[Kro88] KROON, P. and Ed DEPRETTERE, F. "A class of analysis by synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbits/s", IEEE Journal on Selected Areas in Communications **6**:2, 353-63, Feb 1988.

[Mak75] MAKHOUL, J. "Linear prediction: a tutorial review", Proceedings of the IEEE **63**:4, 561-80, Apr 1975

[Mar73] MARKEL, J. D. and GRAY, Jr., A. H. "On autocorrelation equations as applied to speech analysis", IEEE Transactions on Audio and Electroacoustics **21**:2, 69-79, Apr 1973

[pap87] PAPAMICHALIS, P. E. "Practical approaches to speech coding", Prentice-Hall, Englewood Cliffs, 1987

[Rab78] RABINER, L. R. and SCHAFER, R. W. "Digital processing of speech signals", Prentice Hall, Englewood Cliffs, 1978.

[Rou77] le ROUX, J. and GUEGUEN, C. "A fixed point computation of partial correlation coefficients", IEEE trans. on Acoustics, Speech and Signal Processing **25**:3, 257-9, Jun 1977.

[Sha87] O'SHAUGHNESSY. D. "Speech communication: human and machine", Addison-Wesley, Nova York, 1987.

[Sub88] SUGAMURA, N and FARVARDIN, N. "Quantizer design in LSP analysis-synthesis", IEEE Journal on Selected Areas in Communications **6**:2, 432-40, Feb 1988.

[Vis75] VISWANATHAN, R. and MAKHOUL, J. "Quantization properties of transmission parameters in linear predictive systems", IEEE Trans. on Acoustics, Speech and Signal Processing **23**:3, 309-21, Jun 1975.