DICIONÁRIO ESTOCÁSTICO MULTIPULSO PARA CODIFICADORES CELP

Lúcio Martins da Silva[1] **e Abraham Alcaim**[2]
[1]Departamento de Engenharia Elétrica
Universidade de Brasília
[2]CETUC/PUC/Rio
alcaim@cetuc.puc-rio.br

Resumo - Este artigo apresenta uma nova técnica de projeto de dicionários de vetores esparsos para codificadores CELP (Code Excited Linear Prediction), baseada no esquema de excitação multipulso e no algoritmo LBG (Linde-Buzo-Gray) de treinamento. O dicionário resultante – denominado *multipulso* ou, simplesmente, MP — possui as vantagens inerentes aos dicionários esparsos convencionais. Seu grande mérito é ter ainda a vantagem de poder ser otimizado através de um procedimento de treinamento. Através de simulações, o dicionário MP foi comparado com os dicionários gaussiano e "vector-sum" (VS). Os resultados mostraram que, dentre os três tipos de dicionário testados, o dicionário MP é o que tem melhor desempenho em termos da qualidade da voz codificada. Neste artigo, são apresentadas também comparações quanto à complexidade associada aos três dicionários.

Abstract - This paper presents a new algorithm to design codebooks of sparse vectors for code excited linear prediction (CELP) speech encoders. The algorithm is based on multipulse excitations and on the LBG codebook training algorithm. The codebooks, which are referred to as multipulse (MP) codebooks, have the inherent advantages of sparse codebooks. Moreover, they can be optimized by a closed-loop training procedure. We present a comparative analysis of the MP codebook with vector-sum (VS) and Gaussian codebooks. Simulation results show that the MP codebook yields the best performance in terms of speech quality. Complexities of these codebooks are also compared.

Palavras chaves - Codificação de Voz, CELP, Otimização de Dicionário.

1. INTRODUÇÃO

A codificação CELP ("Code Excited Linear Prediction") [19] é atualmente uma das técnicas mais eficazes para compressão digital de voz com alta qualidade a taxas de bits na faixa entre 4 e 16 kb/s. Ela pertence à classe de algoritmos que têm por base a predição linear e codifica a excitação usando um procedimento de análise-por-síntese. Um diagrama de uma estrutura CELP usual é mostrado na Fig. 1. A voz reconstruída é obtida passando o sinal de excitação por um filtro de síntese baseado nos coeficientes de predição de curto termo. Os parâmetros deste filtro são determinados a partir do sinal de voz original e são adaptados a intervalos regulares. O sinal de excitação, $u(n)$, é formado pela soma ponderada de contribuições provenientes de dois dicionários :

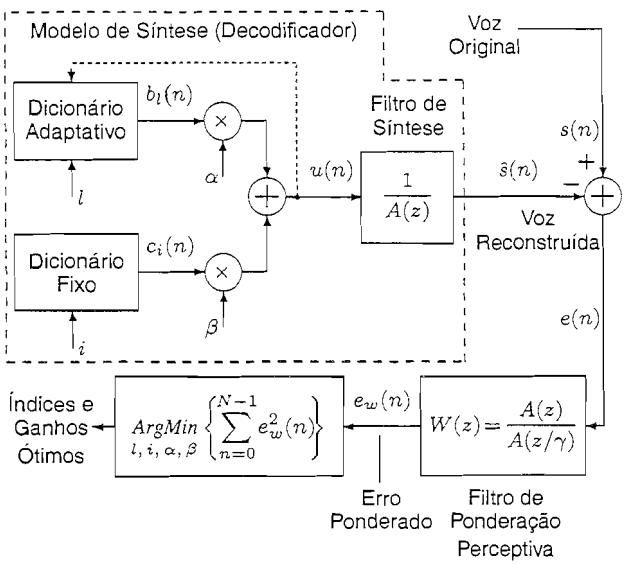

Figura 1. Diagrama de bloco ilustrativo da análise-por-síntese CELP para codificação da excitação.

um dicionário adaptativo e outro fixo. O dicionário adaptativo [14] é responsável pela periodicidade dos segmentos sonoros da voz e realiza uma função similar à de um preditor tonal (ou de longo termo). Para cada bloco de N amostras do sinal de voz original $s(n)$, um procedimento de análise-por-síntese determina os valores ótimos para os índices, l e i, e para os ganhos, α e β. Valores usuais de N correspodem a um comprimento em torno de 5 ms. A Fig. 1 ilustra a identificação da melhor seqüência de excitação para o bloco corrente do sinal de voz, que, sem perda de generalidade, está se assumindo ser o bloco $\{s(n) : n=0,1,\ldots,N-1\}$. Após a análise (ou síntese, no caso do decodificador) de cada bloco do sinal de voz, o dicionário adaptativo é atualizado, por isso a realimentação com $u(n)$ indicada com linha tracejada na Fig. 1. O filtro de ponderação $W(z)$ tem o propósito de prover uma medida de erro objetiva mais correlata com avaliações subjetivas. O seu uso é uma forma de explorar o fenômeno do mascaramento auditivo [2].

Na proposta original do algoritmo CELP [3], [19], o dicionário fixo é formado por seqüências (ou vetores-código) obtidas de um processo gaussiano i.i.d. (variáveis independentes e identicamente distribuídas). Contudo, a busca nesse tipo de dicionário, sem qualquer estruturação, despende um esforço computacional muito grande, além de requerer muita

memória para o seu armazenamento. Outros tipos de dicionários com restrições estruturais têm sido, então, propostos, visando conseguir uma ou mais das seguintes características : busca rápida, espaço de armazenamento reduzido, menor sensibilidade a erros no canal e voz de melhor qualidade. Um tipo de dicionário largamente usado é aquele formado por vetores esparsos, isto é, vetores cuja maioria dos elementos tem valor zero. Ele foi proposto inicialmente para reduzir a complexidade da busca e a memória de armazenamento, mas posteriormente se constatou que ele pode também propiciar voz de qualidade melhor do que a conseguida com um dicionário gaussiano i.i.d. [8], [15], [14]. Este tipo de dicionário é usualmente obtido através de um procedimento de ceifagem central dos vetores de um dicionário gaussiano i.i.d. Em geral se usa dicionários com vetores que tiveram 90 a 95% dos seus elementos ceifados, ou seja, cujas amplitudes foram reduzidas para zero.

O fato de um dicionário esparso propiciar uma voz reconstruída de qualidade melhor do que aquela conseguida com um dicionário gaussiano i.i.d. tem, segundo Kleijn *et al.* [13], a seguinte explicação : a seqüência de excitação (após remoção da correlação de longo-termo ou subtração da contribuição do dicionário adaptativo) ainda possui alguma estrutura relevante que é melhor aproximada após a ceifagem central das seqüências gaussianas i.i.d. É provável, na opinião dos autores deste artigo, que em muitos casos um dado vetor do dicionário seja eleito o melhor por conta de seus elementos de maior magnitude, embora alguns ou muitos dos seus elementos de pequena magnitude contribuam negativamente para a reconstrução do sinal de voz. Com a ceifagem central estes elementos de pequena magnitude são eliminados, o que teria, então, um efeito benéfico. Isto deve ocorrer particularmente na reconstrução dos ataques (inícios) dos sons sonoros. Nestes casos o dicionário adaptativo não é capaz de gerar os pulsos proeminentes observados na forma de onda do resíduo de predição linear, ficando para o dicionário fixo a tarefa de tentar gerá-los. Assim, o vetor-código que contiver um pulso proeminente na posição adequada será um forte candidato, mesmo que o restante de seus elementos não sejam tão apropriados. A desvantagem da ceifagem central como técnica de geração de dicionários esparsos é que não existe, ou pelo menos não é do conhecimento dos autores, um procedimento para treinar (otimizar) o dicionário. O treinamento dos dicionários propicia melhoria significativa à qualidade da voz decodificada, como comprovaram os projetistas do VSELP-8 kb/s [10] e do LD-CELP-16 kb/s [6]. A taxas de bits mais baixas, poder otimizar os dicionários representa uma vantagem ainda maior, pois, como os recursos são escassos, é preciso tirar o máximo deles.

Este artigo propõe uma nova técnica de projeto de dicionários de vetores esparsos baseada no esquema de excitação multipulso [1], [4], e na técnica LBG de treinamento [16]. O treinamento é executado em malha fechada, de modo análogo ao método de treinamento de dicionário estocástico não estruturado descrito em [7]. *Dicionário multipulso*, ou simplesmente dicionário MP, é o nome dado ao dicionário projetado com a técnica aqui proposta. Todos os vetores de um dicionário MP têm o mesmo número N_{nz} de elementos não-nulos, sendo que N_{nz} é escolhido muito menor do que N (dimensão dos vetores-código). O dicionário MP possui as vantagens dos dicionários esparsos convencionais, ou seja, ocupa pouco espaço de armazenamento, permite uma busca rápida e propicia voz de melhor qualidade comparado a um dicionário gaussiano i.i.d. Seu grande mérito é ter ainda a significativa vantagem de poder ser otimizado através de um procedimento de treinamento.

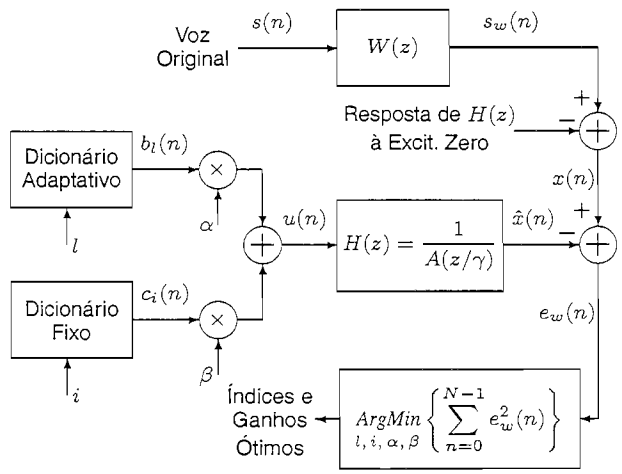

Figura 2. Versão do sistema de análise-por-síntese mostrado na Fig. 1, modificado de modo a torná-lo mais eficiente. O filtro $H(z)$ se encontra no estado zero (memórias zeradas) no instante $n=0$.

A Seção II do artigo apresenta a estrutura de análise-por-síntese usualmente empregada em codificadores CELP para escolha dos parâmetros da excitação. Na Seção III é descrito o método de projeto de um dicionário MP otimizado para a estrutura CELP. A Seção IV apresenta uma análise comparativa do dicionário MP com os dicionários VS ("vector-sum") [10] e GAU (Gaussiano). Finalmente, a Seção V contém as principais conclusões do trabalho.

2. ESTRUTURA DE ANÁLISE-POR-SÍNTESE EM CODIFICADORES CELP

O procedimento de análise-por-síntese CELP representado pela Fig. 1 é, na prática, implementado segundo o diagrama da Fig. 2.. As duas estruturas são equivalentes, mas aquela mostrada na Fig. 2. leva a algoritmos mais eficientes. Estas figuras ilustram a obtenção da seqüência de excitação para o bloco corrente do sinal de voz, que, sem perda de generalidade, está se assumindo ser o bloco $\{s(n) : n = 0, 1, \ldots, N-1\}$. Para tornar a notação mais compacta será adotada a representação vetorial para as seqüências; assim, uma seqüência qualquer $\{r(n) : n=0, 1, \ldots, N-1\}$ será representada pelo vetor $\mathbf{r} = (r(0)\, r(1) \ldots r(N-1))^T$, onde T indica a operação de transposição.

O procedimento de análise-por-síntese pode ser enunciado como se segue : dados o dicionário adaptativo $\mathcal{B} = \{\mathbf{b}_l : l = 1, \ldots, J\}$ e o dicionário fixo $\mathcal{C} = \{\mathbf{c}_i : i = 1, \ldots, K\}$, deseja-se identificar o par de vetores-código $(\mathbf{b}_L, \mathbf{c}_I)$ que resulta no vetor sintético $\hat{\mathbf{x}}$ (ver Fig. 2.) que melhor aproxima o vetor-alvo $\mathbf{x}$, ou seja, que minize a norma do vetor-erro $\mathbf{e}_w = \mathbf{x} - \hat{\mathbf{x}}$.

($\|\cdot\|$ denotará o operador norma, ou seja, $\|\mathbf{r}\| = \sqrt{\mathbf{r}^T\mathbf{r}}$.)

Como se pode deduzir da Fig. 2., o vetor sintético $\hat{\mathbf{x}}$, obtido com o par de vetores-código $(\mathbf{b}_l, \mathbf{c}_i)$, é dado por

$$\begin{aligned}\hat{x}(n) &= h(n) * u(n) \\ &= \alpha h(n) * b_l(n) + \beta h(n) * c_i(n), \end{aligned} \quad (1)$$

onde $n = 0, 1, \ldots, N-1$, e $h(n)$ é a resposta impulsional do filtro $H(z) = 1/A(z/\gamma)$. Usando notação matricial, (1) pode ser reescrita na forma :

$$\hat{\mathbf{x}} = \alpha\mathbf{H}\mathbf{b}_l + \beta\mathbf{H}\mathbf{c}_i, \quad (2)$$

onde $\mathbf{H}$ é a matriz Toeplitz triangular inferior assim definida :

$$\mathbf{H} = \begin{pmatrix} h(0) & 0 & 0 & \cdots & 0 \\ h(1) & h(0) & 0 & \cdots & 0 \\ h(2) & h(1) & h(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ h(N-1) & h(N-2) & h(N-3) & \cdots & h(0) \end{pmatrix} \quad (3)$$

Usualmente, a busca pelos vetores ótimos $(\mathbf{b}_L, \mathbf{c}_I)$ é feita seqüencialmente. Primeiro identifica-se o vetor-código do dicionário adaptativo $\mathcal{B} = \{\mathbf{b}_l : l = 1, 2, \ldots, J\}$ que melhor aproxima o vetor-alvo $\mathbf{x}$, ou seja, que minimiza o erro $\|\mathbf{x} - \alpha\mathbf{H}\mathbf{b}_l\|^2$. Pode ser mostrado que este vetor-código tem seu índice dado por

$$L = \underset{l=1,\ldots,J}{ArgMax} \left\{ \frac{\left(\mathbf{x}^T\mathbf{H}\mathbf{c}_l\right)^2}{\|\mathbf{H}\mathbf{c}_l\|^2} \right\} \quad (4)$$

e o ganho α deve ter o seguinte valor :

$$\alpha = \frac{\mathbf{x}^T\mathbf{H}\mathbf{c}_L}{\|\mathbf{H}\mathbf{c}_L\|^2}. \quad (5)$$

Selecionado $\mathbf{b}_L$ e calculado o valor ótimo para α, um novo vetor-alvo para busca no dicionário fixo é definido como

$$\mathbf{y} = \mathbf{x} - \alpha\mathbf{H}\mathbf{b}_L. \quad (6)$$

Em seguida, identifica-se, então, qual vetor-código do dicionário fixo minimiza o erro $\|\mathbf{y} - \beta\mathbf{H}\mathbf{c}_i\|^2$. Pode ser mostrado que este vetor-código tem seu índice dado por

$$I = \underset{i=1,\ldots,K}{ArgMax} \left\{ \frac{\left(\mathbf{y}^T\mathbf{H}\mathbf{c}_i\right)^2}{\|\mathbf{H}\mathbf{c}_i\|^2} \right\} \quad (7)$$

e o ganho β deve ter o seguinte valor :

$$\beta = \frac{\mathbf{y}^T\mathbf{H}\mathbf{c}_I}{\|\mathbf{H}\mathbf{c}_I\|^2} \quad (8)$$

3. PROJETO DE UM DICIONÁRIO MULTIPULSO OTIMIZADO

O projeto de um dicionário MP, com K vetores, se inicia com a construção de um dicionário inicial, que pode ser feita escolhendo-se aleatoriamente as posições e amplitudes dos N_{nz} elementos não nulos de cada um dos K vetores. Este dicionário é colocado no codificador CELP e então otimizado iterativamente de modo a minimizar a distorção entre a voz original e sua versão reconstruída. Este processo de otimização deve ser realizado com um sinal de voz de treinamento que contenha uma grande diversidade de falas e locutores.

A i-ésima iteração do processo de otimização se inicia com a análise-por-síntese e reconstrução de todo o sinal de voz de treinamento. Nesta operação o sinal de treinamento é dividido em M blocos justapostos de comprimento N amostras, aqui representados pelos vetores $\{\mathbf{s}_m : m = 1, \ldots, M\}$. Um dos resultados da análise é a separação (partição) destes blocos em K conjuntos ou células, $\{\mathcal{S}_k : k = 1, \ldots, K\}$. A célula $\mathcal{S}_k$ desta partição contém todos os blocos do sinal de treinamento que "escolheram" o k-ésimo vetor do dicionário MP (obtido na iteração anterior). O problema de otimização na i-ésima iteração pode, então, ser assim definido : dada a partição $\{\mathcal{S}_k : k = 1, \ldots, K\}$ quer se encontrar o dicionário $\mathcal{C} = \{\mathbf{c}_k : k = 1, \ldots, K\}$, onde $\mathbf{c}_k$ é um vetor com $N - N_{nz}$ elementos nulos, que minimiza a distorção global

$$D = \sum_{m=1}^{M} d_w(\mathbf{s}_m, \hat{\mathbf{s}}_m), \quad (9)$$

onde $\hat{\mathbf{s}}_m$ é a aproximação CELP para o vetor $\mathbf{s}_m$ e d_w é a mesma medida de distorção usada na análise-por-síntese CELP. Ou seja, de acordo com as Figuras 1 e 2, tem-se que

$$d_w(\mathbf{s}_m, \hat{\mathbf{s}}_m) = \|\mathbf{e}_{wm}\|^2 = \|\mathbf{x}_m - \hat{\mathbf{x}}_m\|^2, \quad (10)$$

ou, usando as equações (2) e (6),

$$d_w(\mathbf{s}_m, \hat{\mathbf{s}}_m) = \|\mathbf{y}_m - \beta_m\mathbf{H}_m\mathbf{c}_{i(m)}\|^2, \quad (11)$$

onde o subscrito m foi acrescentado às notações para indicar a dependência das grandezas com o bloco a que estão associadas.

Substituindo (11) em (9) e dividindo o somatório em K termos, um para cada célula $\mathcal{S}_k$, $k = 1, \ldots, K$, tem-se que

$$D = D_1 + \ldots + D_k + \ldots + D_K, \quad (12)$$

onde

$$D_k = \sum_{m \in \mathcal{S}_k} \|\mathbf{y}_m - \beta_m\mathbf{H}_m\mathbf{c}_k\|^2. \quad (13)$$

Para simplificar o processo de otimização assume-se que minimizar D com relação ao dicionário $\mathcal{C}$ é equivalente a minimizar cada termo D_k, $k = 1, \ldots, K$, em (12) com relação ao vetor-código correspondente $\mathbf{c}_k$. Em outras palavras, assume-se que os vetores-alvos $\{\mathbf{y}_m : m = 1, \ldots, M\}$ e os ganhos $\{\beta_m : m = 1, \ldots, M\}$ não dependem do dicionário que se está otimizando. O problema a ser resolvido passa a ser, então, o de encontrar o vetor $\mathbf{c}_k$, $k = 1, \ldots, K$, que minimiza o somatório em (13), onde $\{(\mathbf{y}_m, \beta_m) : m \in \mathcal{S}_k\}$ é independente de $\mathbf{c}_k$, com a restrição de que apenas N_{nz} dos elementos de $\mathbf{c}_k$ podem ter valores diferentes de zero.

A matriz $\mathbf{H}_m$, que aparece em (13), é definida conforme a equação (3) e pode ser escrita na forma

$$\mathbf{H}_m = [\mathbf{h}_{m,0}\ \mathbf{h}_{m,1}\ \ldots\ \mathbf{h}_{m,j}\ \ldots\ \mathbf{h}_{m,N-1}], \quad (14)$$

onde

$$\mathbf{h}_{m,j} = \left[\underbrace{0 \ \ldots \ 0}_{j \text{ zeros}} \ h_m(0) \ h_m(1) \ \ldots \ h_m(N-1-j) \right]^T . \quad (15)$$

Substituindo (14) em (13), tem-se que

$$D_k = \sum_{m \in \mathcal{S}_k} \left\| \mathbf{y}_m - \beta_m \sum_{j=1}^{N_{nz}} A_j \mathbf{h}_{m,l_j} \right\|^2 , \quad (16)$$

onde $\{l_1, l_2, \ldots, l_{N_{nz}}\}$ e $\{A_1, A_2, \ldots, A_{N_{nz}}\}$ são as posições e amplitudes, respectivamente, dos elementos não nulos de $\mathbf{c}_k$. O problema estará, então, solucionado se forem determinados os conjuntos $\{l_1, l_2, \ldots, l_{N_{nz}}\}$ e $\{A_1, A_2, \ldots, A_{N_{nz}}\}$ que minimizam a distorção D_k. Determinar simultaneamente todos os N_{nz} elementos não nulos é muito despendioso computacionalmente, por isso optou-se por um cálculo seqüencial sub-ótimo. Primeiro determina-se l_1 e A_1 supondo que c_{kl_1} é o único elemento não nulo de $\mathbf{c}_k$. Em seguida determina-se l_2 e A_2 supondo que c_{kl_1} e c_{kl_2} são os únicos elementos não nulos de $\mathbf{c}_k$ e que c_{kl_1} é fixo. E assim sucessivamente, até completar N_{nz} elementos não nulos. Portanto, para se calcular o q-ésimo elemento não nulo de $\mathbf{c}_k$, primeiro subtrai-se dos vetores-alvos $\{\mathbf{y}_m : m \in \mathcal{S}_k\}$ a contribuição dos $q-1$ elementos já determinados. Isto é, primeiro calcula-se novos vetores-alvos :

$$\hat{\mathbf{y}}_m = \mathbf{y}_m - \beta_m \sum_{j=1}^{q-1} A_j \mathbf{h}_{m,l_j}, \quad m \in \mathcal{S}_k. \quad (17)$$

Em seguida, determina-se os valores de l_q e A_q que minimizam a distorção $\hat{D}_{k,q}$, assim definida :

$$\hat{D}_{k,q} = \sum_{m \in \mathcal{S}_k} \| \hat{\mathbf{y}}_m - \beta_m A_q \mathbf{h}_{m,l_q} \|^2 . \quad (18)$$

O valor de A_q é obtido solucionando a equação $\partial \hat{D}_{k,q} / \partial A_q = 0$. Substituindo (18) nesta igualdade resulta :

$$A_q = \frac{\sum_{m \in \mathcal{S}_k} \beta_m \mathbf{h}^T_{m,l_q} \hat{\mathbf{y}}_m}{\sum_{m \in \mathcal{S}_k} \beta_m^2 \| \mathbf{h}_{m,l_q} \|^2} \quad (19)$$

Substituindo esta expressão em (18) chega-se ao seguinte resultado : a melhor posição l_q para o q-ésimo elemento não nulo é

$$l_q = \underset{l=0,\ldots,N-1}{ArgMax} \left\{ \frac{\left(\sum_{m \in \mathcal{S}_k} \beta_m \mathbf{h}^T_{m,l} \hat{\mathbf{y}}_m \right)^2}{\sum_{m \in \mathcal{S}_k} \beta_m^2 \| \mathbf{h}_{m,l} \|^2} \right\} \quad (20)$$

Para reduzir a subotimalidade, antes de determinar o $(q+1)$-ésimo elemento não nulo de $\mathbf{c}_k$, pode-se recalcular simultaneamente as amplitudes $\{A_1, A_2, \ldots, A_q\}$. Para isto, a distorção intermediária $\hat{D}_{k,q}$, definida por (18), é rescrita na seguinte forma :

$$\hat{D}_{k,q} = \sum_{m \in \mathcal{S}_k} \left\| \mathbf{y}_m - \beta_m \sum_{j=1}^{q} A_j \mathbf{h}_{m,l_j} \right\|^2 , \quad (21)$$

que é obtida substituindo (17) em (18). Os novos valores para $\{A_1, A_2, \ldots, A_q\}$ são, então, aqueles que minimizam esta distorção intermediária $\hat{D}_{k,q}$. Isto é, lhes são atribuídos os valores que tornam verdadeiras as igualdades $\partial \hat{D}_{k,q} / \partial A_i = 0$, $i = 1, 2, \ldots, q$, ou, equivalentemente, os valores obtidos com a solução de

$$\Phi \mathbf{A} = \rho, \quad (22)$$

onde

$$\mathbf{A} = [A_1 \ A_2 \ \ldots \ A_q]^T, \quad (23)$$

Φ é uma matriz $q \times q$, cujos elementos são dados por

$$\phi(i,j) = \sum_{m \in \mathcal{S}_k} \beta_m^2 \mathbf{h}^T_{m,l_i} \mathbf{h}_{m,l_j}, \quad 1 \le i,j \le q, \quad (24)$$

e ρ é um vetor-coluna formado pelos seguintes elementos :

$$\rho(i) = \sum_{m \in \mathcal{S}_k} \beta_m \mathbf{y}^T_m \mathbf{h}_{m,l_i}, \quad 1 \le i \le q. \quad (25)$$

Este procedimento de otimização das amplitudes é similar ao procedimento que é geralmente usado no cálculo da excitação em esquemas MP-LPC ("multipulse linear predictive coding") [4].

Uma vez que se tenha determinado l_q, usando (20), e recalculado $\{A_1, A_2, \ldots, A_q\}$, resolvendo a equação (22), se $q = N_{nz}$, está determinado o vetor-código multipulso $\mathbf{c}_k$. Em caso contrário, passa-se a calcular o $(q+1)$-ésimo elemento. Todo este processo tem que ser realizado para cada valor de $k \in \{1, \ldots, K\}$, onde K é o número de vetores no dicionário multipulso. Novas iterações são, então, realizadas enquanto a distorção global D, dada por (9), não for suficientemente pequena ou o seu valor não decrescer significativemente de uma iteração para outra.

Este método de otimização do dicionário MP não garante que a distorção total D irá decrescer monotonicamente à medida que as iterações são executadas, embora o método seja baseado no algoritmo LBG de treinamento de dicionários, que garante o decrescimento monotônico do erro. Contrariando a assunção que foi feita, os vetores-alvos $\{\mathbf{y}_m\}$ e os ganhos $\{\beta_m\}$ dependem do dicionário que se está otimizando. Isto é, de uma iteração para outra do treinamento, $\{\mathbf{y}_m\}$ e $\{\beta_m\}$ se modificam e isto faz com que não se tenha um decrescimento monotônico do erro. Na prática, entretanto, em geral o erro converge para um mínimo, embora não monotonicamente, e o método se mostra eficaz.

4. DESEMPENHO DO DICIONÁRIO MULTIPULSO

Para avaliar a qualidade do dicionário MP foi realizado um teste comparativo com os dicionários VS ("vector-sum") [10]

e gaussiano (GAU) i.i.d. O dicionário VS é um importante e efetivo dicionário estruturado. Os codificadores CELP's que usam este tipo de dicionário são denominados "vector-sum excited linear prediction (VSELP)" e foram adotados em padrões dos sistemas de telefonia celular digital TDMA da América do Norte e do Japão. Um dicionário VS com 2^B vetores é construído combinando-se um conjunto de B vetores estocásticos $\{\mathbf{b}_k^{(vs)} : k = 1, \ldots, B\}$ da seguinte forma

$$\mathbf{c}_i^{(vs)} = \sum_{k=1}^{B} \theta_{ik} \mathbf{b}_k^{(vs)}, \qquad i = 1, 2, \ldots, 2^B, \tag{26}$$

onde $\theta_{ik} = +1$ se o k-ésimo bit do índice binário i é igual a 1 ou, em caso contrário, $\theta_{ik} = -1$. Esta estruturação do dicionário reduz drasticamente a quantidade de memória para armazenamento e o esforço computacional despendido na busca [10]. Além disso, como a palavra binária i enviada ao receptor especifica diretamente a polaridade da combinação linear dos vetores de base, se o canal causar o erro de um único bit desta palavra, apenas um termo da soma em (26) é afetado, causando somente uma mudança moderada no vetor de excitação decodificado. Contudo, a dependência entre os vetores torna o dicionário menos eficiente.

Os vetores-base do dicionário VS são otimizados com uma base de dados (voz) de treinamento [10], de modo análogo ao proposto para um dicionário MP. O critério de otimização é também a minimização do erro ponderado total ou, equivalentemente, da distorção global D, dada por (9). Esta distorção pode ser expressa como uma função de cada uma das amostras dos B vetores-base, dados $x(n)$, L, I, $b_L(n)$, α, β e $h(n)$ (ver Fig. 2) para cada um dos blocos de codificação (N amostras) do sinal de voz de treinamento. Os vetores-base ótimos são, então, calculados através da solução de BN equações simultâneas que são obtidas tomando-se as derivadas parciais da distorção global em relação a cada amostra dos vetores-base e igualando-as a zero. Assim como na otimização de um dicionário MP, este procedimento deve ser repetido até que a distorção global D se torne suficientemente pequena ou o seu valor não decresça significativemente de uma iteração para outra.

4.1. MEDIDAS DE DISTORÇÃO USADAS PARA AVALIAR A QUALIDADE DA VOZ

O desempenho dos dicionários testados foi avaliado através da qualidade do sinal de voz decodificado, avaliada por três medidas : razão sinal-ruído-ponderado segmentar (*RSRpond-seg*), distância cepestral ($\mathcal{DC}$) e distância espectral ($\mathcal{DE}$). Estas medidas têm, segundo vários estudos [5], [11], [12], [21], uma boa correlação com medidas subjetivas de avaliação, quando aplicadas a codificadores de voz a baixas taxas de bits. A seguir são definidas cada uma destas medidas.

Sejam $s(n)$ e $\hat{s}(n)$ o sinal de voz original e degradado (decodificado), respectivamente, e seja $e_w(n)$ o erro ponderado, obtido passando-se o sinal-erro $e(n) = s(n) - \hat{s}(n)$ pelo filtro de ponderação $W(z)$ (veja Fig. 1). A *RSRpond-seg* é dada por

$$RSRpond\text{-}seg\ (dB) = \frac{1}{M} \sum_{m=0}^{M-1} 10 \log \frac{\sum_{n=0}^{Q-1} s^2(n + mQ)}{\sum_{n=0}^{Q-1} e_w^2(n + mQ)} \tag{27}$$

Isto é, para o cálculo da *RSRpond-seg*, o sinal de voz original e o erro ponderado são divididos em M segmentos de Q amostras ; em seguida é calculada a razão sinal-ruído-ponderado, em *dB*, associada a cada segmento : a média aritmética dos M valores obtidos é a *RSRpond-seg*, em *dB*. Neste trabalho, utilizou-se $Q = 80$, o que equivale a um intervalo de 10 *ms* (a taxa de amostragem é de 8 kHz). A *RSRpond-seg* tem, em geral, boa correlação com os testes subjetivos e a razão está na ponderação que dá pesos diferentes para os componentes do erro : pesos maiores para aqueles componentes localizados nas bandas de freqüência em que o ruído é mais perceptível auditivamente e pesos menores para aqueles localizados nas bandas em que o ruído é menos perceptível. Esta diferença na percepção auditiva dos componentes freqüenciais do ruído é devida ao fenômeno do mascaramento auditivo [2].

A distância cepestral ($\mathcal{DC}$) é definida como [12] :

$$\mathcal{DC}\ (dB) = \frac{10}{\ln 10} \sqrt{2 \sum_{k=1}^{P} [c_s(k) - c_{\hat{s}}(k)]^2} \tag{28}$$

onde $c_s(\cdot)$ e $c_{\hat{s}}(\cdot)$ são os coeficientes cepestrais LPC [18, p. 442] do sinal de voz original e degradado, respectivamente. A distância $\mathcal{DC}$ é calculada a intervalos de Q amostras e o seu valor médio usado como um indicador da qualidade do sinal de voz degradado. Também neste caso se utilizou $Q = 80$. As análises LPC's foram de ordem dezesseis e utilizaram janelas de 24 *ms*. O número de coeficientes cepestrais usado na medida foi $P = 32$.

A distância espectral ($\mathcal{DE}$), com compensação de amplitude, é definida por

$$\mathcal{DE}\ (dB) = 10 \sqrt{\frac{1}{L} \sum_{l=0}^{L-1} \left[\log \left(\frac{|S(l)|^2}{|\hat{S}(l)|^2} \right) - \log \left(\frac{\sigma_s^2}{\sigma_{\hat{s}}^2} \right) \right]^2} \tag{29}$$

onde $S(\cdot)$ e $\hat{S}(\cdot)$ são as transformadas de Fourier discretas do sinal de voz original e degradado, respectivamente, e σ_s^2 e $\sigma_{\hat{s}}^2$ são as energias destes sinais. A distância $\mathcal{DE}$ é também calculada a intervalos de Q amostras e o seu valor médio usado como uma medida da qualidade do sinal de voz degradado. As medidas foram realizadas com $Q = 80$ e $L = 256$. A distância espectral leva em conta o espectro de potência *total* do sinal, enquanto que a distância cepestral considera apenas a envoltória deste espectro.

4.2. TESTES E RESULTADOS

A plataforma CELP usada nos testes é como aquela mostrada na Fig. 1 e suas particularidades estão apresentadas na

Tabela 1 Dados da plataforma CELP usada na comparação dos dicionários MP, VS e Gaussiano. (†"Line Spectrum Frequency", ‡Quantizador Vetorial Híbrido.)

Freqüência de Amostragem	8 kHz
Análise e Codificação LPC	
Método	Autocorrelação
Ordem	10
Janela de análise (Hamming)	192 amostras (24 ms)
Freqüência das análises	Uma a cada 30 ms
Parâmetros quantizados	LSF†
Tipo do quantizador	QV híbrido (QVH‡) [20]
Taxa de bits	25 bits/vetor-LSF
Codificação da Excitação	
Comprimento dos vetores	$N = 48$ amostras (6 ms)
Faixa de atrasos do dicionário adaptativo	20-146 amostras
Ganhos α e β	Não quantizados
Parâmetro de $W(z)$	$\gamma = 0,8$

Tabela 2 Desempenho dos dicionários MP, VS e Gaussiano (não ceifado) em termos da qualidade objetiva da voz reconstruída. K denota o tamanho efetivo do dicionário; para os dicionários VS's, $K = 2^B/2$.

Dicionário			RSRpond-seg	$\mathcal{DC}$	$\mathcal{DE}$
Tipo	Treinamento	K	(dB)	(dB)	(dB)
MP	não	64	13,49	3,35	8,17
MP	sim	64	13,81	3,23	8,05
MP	não	128	13,84	3,32	8,12
MP	sim	128	14,06	3,26	8,01
MP	não	256	14,17	3,28	8,06
MP	sim	256	14,33	3,14	7,92
VS	não	64	12,97	3,47	8,44
VS	sim	64	13,33	3,45	8,41
VS	não	128	13,23	3,36	8,31
VS	sim	128	13,51	3,45	8,31
VS	não	256	13,46	3,37	8,23
VS	sim	256	13,76	3,41	8,29
GAU	não	64	13,31	3,40	8,27
GAU	não	128	13,66	3,30	8,16
GAU	não	256	13,98	3,29	8,11

Tabela 1. Para os dicionários MP's, escolheu-se $N_{nz} = 6$, para $N = 48$ (comprimento dos vetores-código). Esta escolha foi baseada em experimentos com diferentes valores para N_{nz}. Trabalhos sobre dicionários de vetores esparsos obtidos através de ceifagem central sugerem a manutencão de 5 a 10% das amostras de cada vetor-código [8], [15], [14]. O valor escolhido para N_{nz} corresponde a uma percentagem de 12,5%.

O sinais de voz usados nas simulações sofreram uma filtragem passa-faixa com banda 120-3400 Hz, foram amostrados a uma taxa de 8000 amostras/s e digitalizados com 12 bits/amostra. Na fase de testes foram usados sinais de voz de quatro locutores : dois homens e duas mulheres. Cada locutor pronunciou um par de frases, resultando um sinal com duração total de 18,83 segundos. Para o treinamento dos dicionários MP e VS foram usados sinais de voz de dezoito locutores (9 homens e 9 mulheres), que não participaram do conjunto de teste. A duração total do sinal de voz de treinamento foi de 84,25 segundos.

As simulações foram realizadas em um plataforma Pentium-200 MHz. Para treinar um dicionário MP com 128 vetores é despendido um tempo de máquina em torno de 3,5 horas, enquanto que o treinamento de um dicionário VS com 8 vetores-base ($B = 8$) despende em torno de 15 horas. Neste último, a solução das BN equações simultâneas despende um tempo muito grande. Para esta solução usou-se o método de Cholesky.

Os resultados obtidos nas simulações estão mostrados na Tabela 2. Foram testados dicionários com 64, 128 e 256 vetores. No caso dos dicionários VS o negativo-simétrico de um vetor-código é também um vetor-código e o ganho aplicado, β (veja Fig. 1), é sempre positivo, enquanto que com os dicionários MP e gaussiano, o ganho β pode ser positivo ou negativo. Portanto, o ganho de um dicionário VS requer um bit a menos para a sua codificação ou, em outras palavras, o tamanho efetivo de um dicionário VS é $2^B/2$, onde B é o *número de vetores da base VS*. Por isso, na Tabela 2 o tamanho de um dicionário VS equivale a $2^B/2$ e não a 2^B. (Os vetores-base VS não treinados foram obtidos de um processo gaussiano i.i.d.)

Um dado que deve ser observado nos resultados apresentados na Tabela 2 é a melhoria que se consegue quando o tamanho de um dicionário é duplicado. Notar que as variações nas medidas objetivas são relativamente pequenas. Quando um dicionário não treinado é duplicado em tamanho, o que subjetivamente pode ter um efeito significativo, o incremento da RSRpond-seg, por exemplo, varia entre 0,23 e 0,35 dB. Os decrementos das medidas de DC e DE são, em termos absolutos, ainda menores. Estes dados são importantes para a avaliação da melhoria introduzida com o treinamento dos dicionários.

Com o treinamento, os valores de RSRpond-seg obtidos com os dicionários MP's contendo 64, 128 e 256 vetores tiveram um aumento de 0,32, 0,22 dB e 0,16 dB, repectivamente. Portanto, pode-se concluir que o treinamento teve, nestes casos, um efeito equivalente ao de um aumento maior do que 50% no tamanho dos dicionários. No caso dos dicionários VS's, os aumentos dos valores de RSRpond-seg foram ainda um pouco maiores : 0,36, 0,28 e 0,30, respectivamente. Isto significa que, em termos da RSRpond-seg, o treinamento teve um efeito equivalente ao de um aumento maior do que 100% no tamanho dos dicionários VS's. Contudo, as medidas de distorção DC e DE não indicaram melhoria de qualidade do sinal de voz em decorrência do treinamento dos dicionários VS's. No caso dos dicionários MP's, ao contrário, o treinamento propiciou, em termos da DC e DE, melhorias (decrementos dos valores medidos) que são, em geral, superiores àquelas conseguidas com a duplicação do tamanho dos dicionários.

Com relação à comparação dos diferentes tipos de dicionários, notar, em primeiro lugar, que as medidas realizadas indicam que os dicionários MP's são superiores, em desempenho, aos *dicionários gaussianos*. Isto está de acordo com os resultados obtidos por Davidson et al. [8], Lin [15] e Kleijn et al. [14], segundo os quais um dicionário esparso pode propiciar voz de melhor qualidade do que um dicionário gaussiano i.i.d.

Quando se comparam os dicionários MP e VS, com base

Tabela 3 Complexidade dos dicionários MP, VS e Gaussiano em termos do tempo despendido na busca e memória requerida para o armazenamento do dicionário.

Dicionário		Tempo relativo	Memória
Tipo	K	despendido na busca	(bytes)
VS	64	1,00	1.344
	128	1,17	1.536
	256	1,53	1.728
MP	64	1,44	1.920
	128	2,84	3.840
	256	5,62	7.680
GAU	64	5,64	12.288
	128	11,19	24.576
	256	22,24	49.152

nos resultados apresentados na Tabela 2, a superioridade do dicionário MP é ainda maior. Notar, especialmente, que o dicionário MP, treinado, contendo 64 vetores, tem desempenho superior ao do dicionário VS, treinado, contendo 256 vetores. Para confirmar este resultado foi realizado um teste subjetivo (de escuta) informal. O sinal de voz de teste tinha, como já foi dito, oito frases, pronunciadas por quatro locutores (dois homens e duas mulheres). Para cada uma destas frases foi constituído um par de versões sintéticas : uma delas obtida com o dicionário MP de 64 vetores e a outra com o dicionário VS de 256 vetores. A ordem das versões foi alterada nos oito pares. Solicitou-se, então, a dez avaliadores que escutassem estes pares de sinais de voz e que, para cada dos pares, emitissem uma dentre as seguintes opiniões : a versão A é melhor, a versão B é melhor ou as duas versões são equivalentes. O resultado do teste foi : em $19,4\%$ das avaliações, a versão MP foi escolhida como a melhor, em $16,7\%$ a escolhida foi a versão VS e em $63,9\%$ elas foram consideradas equivalentes. Portanto, em termos da qualidade do sinal de voz codificado, os dicionários MP e VS com 64 e 256 vetores, respectivamente, podem ser considerados equivalentes.

O quesito qualidade da voz não pode, neste caso, ser considerado independentemente do quesito complexidade dos sistemas. Este último foi avaliado, neste trabalho, em dois itens : o esforço computacional despendido com a busca do melhor vetor-código e memória requerida para armazenamento do dicionário. Como estimativa para o primeiro item, foi escolhido o tempo despendido com a busca por uma plataforma Pentium-200 MHz. Para cada dicionário, foi medido o tempo total gasto em 3.200 buscas, realizadas para codificar o sinal de voz de teste. A Tabela 3 mostra os valores medidos normalizados com relação ao tempo despendido na busca no dicionário VS com 64 vetores.

A memória requerida pelos dicionários VS e gaussiano foi calculada como $4BN$ e $4KN$ (bytes), respectivamente — o fator 4 representa o número de bytes para cada número real armazenado. No caso dos dicionários MP's, é preciso armazenar KN_{nz} números reais, que ocupam quatro bytes cada um, e KN_{nz} números que ocupam um byte cada um e dão as posições das amostras não nulas dos vetores-código MP's. Ou seja, a memória requerida por um dicionário MP é $5KN_{nz}$ (bytes). A quantidade de memória requerida por cada um dos dicionários que fizeram parte das simulações é apresentada na Tabela 3.

Os dados da Tabela 3 mostram que, conforme já foi afirmado, o dicionário MP é superior ao dicionário gaussiano também no quesito complexidade. Por outro lado, a complexidade associada a um dicionário VS é significativamente menor do que a que está associada a um dicionário MP, quando ambos têm o mesmo número de vetores-código. Contudo, as complexidades associadas aos dicionários MP e VS com 64 e 256 vetores, respectivamente, são equiparáveis, o tempo despendido na busca MP é, inclusive, menor. Em termos da qualidade do sinal de voz, como já foi destacado, estes dicionários são equivalentes. Porém, com o dicionário VS, a taxa de bits despendida é 333,3 bits/s (2 bits a cada 6 ms) maior do que aquela despendida quando se usa o dicionário MP.

Outro aspecto relevante a ser considerado na comparação dos dicionários é a robustez a erros de transmissão. Os dicionários VS's são intrinsicamente robustos a erros de transmissão. Mas esta robustez decorre justamente da correlação (redundância) existente entre seus vetores, que faz com que um dicionário VS tenha pior desempenho, em canal sem erro, do que um dicionário MP de mesmo tamanho. No caso dos dicionários MP's, a robustez a erros de transmissão pode lhes ser atribuída através de uma associação conveniente entre os índices binários e os vetores-código [9], [17]. Isto é, vetores-código que são "parecidos" uns com os outros devem ser identificados por índices binários que diferem no menor número de bits possível. Desta forma, os índices recebidos pelo decodificador, ainda que tenham sido contaminados por erros de transmissão, levarão a vetores-código que se assemelham aos vetores-códigos "corretos". A vantagem desta estratégia, chamada de codificação com redundância zero, é que ela não afeta o desempenho do dicionário em canal sem erro, o mesmo não acontecendo com a estratégia usada no projeto dos dicionários VS's. É interessante salientar ainda que, tendo os dois dicionários, MP e VS, o mesmo tamanho, o dicionário MP tem a vantagem de ter um desempenho significativamente melhor em canal sem erro e por isso ele só terá pior desempenho em canal com erro se a codificação com redundância zero for muito menos eficiente do que a codificação com redundância executada intrinsicamente na construção de um dicionário VS.

5. CONCLUSÕES

Neste artigo foi proposto um dicionário estocástico esparso para codificadores CELP, denominado dicionário multipulso (MP), que pode ser otimizado através de um procedimento de treinamento. Estudos anteriores já haviam comprovado que os dicionários esparsos podem propiciar voz de melhor qualidade do que os dicionários gaussianos não esparsos, além de propiciar economia significativa na memória de armazenamento e no esforço computacional da busca do vetor ótimo. Entretanto, as estruturas esparsas que têm sido usadas não permitem otimização do dicionário através de treinamento. A nova técnica de projeto apresentada neste artigo vem exatamente permitir a otimização desse tipo atraente de estrutura de dicionário para codificadores CELP.

Comparando com dicionários do tipo "vector-sum" (VS), constiuídos de vetores cheios, os dicionários MP apresentam desempenho significativamente melhor em canal sem erro.

Já os dicionários VS são intrinsecamente robustos a erros de transmissão, mas a correlação existente entre seus vetores deteriora seu desempenho em canal sem erro. O emprego de código com redundância zero pode vir a superar deficiências do dicionário MP em presença de erros no canal, sem afetar seu desempenho em canal sem erro — o mesmo não acontecendo com a estratégia usada no projeto dos dicionários VS. Uma análise detalhada desses aspectos através de simulações constitui um tema de interesse para continuidade deste trabalho.

REFERÊNCIAS

[1] Atal, B. S., e J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low-bit rate," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Paris, França, pp. 614–617, 1982.

[2] Atal, B. S., e M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 247–254, Junho de 1979.

[3] Atal, B. S., e M. R. Schroeder, "Stochastic coding of speech signals at very low bit rates," *Proc. Int. Conf. Commun.*, pp. 1610–1613, 1984.

[4] Berouti, M., H. Garten, P. Kabal, e P. Mermelstein, "Efficient computation and encoding of the multipulse excitation for LPC," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, pp. 10.1.1–10.1.4, 1984.

[5] Campos Neto, S. F., *Metodologias de Avaliação de Algoritmos de Codificação de Voz*, Dissertação de Mestrado, UNICAMP, Abril de 1993.

[6] Chen, J., R. V. Cox, Y. Lin, N. Jayant, e M. J. Melchner, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 830–849, Junho de 1992.

[7] Cuperman, V., e A. Gersho, "Vector predictive coding of speech at 16 kb/s," *IEEE Trans. Commun.*, vol. COM-33, pp. 685–696, Julho de 1985.

[8] Davidson, G., e A. Gersho, "Complexity reduction methods for vector excitation coding," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Tóquio, Japão, pp. 3055–3058, Abril de 1986.

[9] Farvardin, N., "A study of vector quantization for noisy channels," *IEEE Trans. Inform. Theory*, vol. 36, pp. 799–809, Julho de 1990.

[10] Gerson, I. A., e M. A. Jasiuk, "Vector sum excited linear prediction (VSELP) speech coding at 8 kbps," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 461–464, 1990.

[11] Kitawaki, N., K. Itoh, M. Honda, e K. Kakehi, "Comparison of objective speech quality measures for voiceband codecs," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Paris, pp. 1000–1003, 1982.

[12] Kitawaki, N., H. Nagabuchi, e K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE J. Selected Areas in Commun.*, vol. SAC-6, pp. 242–248, Fevereiro de 1988.

[13] Kleijn, W. B., D. J. Krasinski, e R. H. Ketchum, "Fast methods for CELP speech coding algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-38, pp. 1330–1342, Agosto de 1990.

[14] Kleijn, W. B., D. J. Krasinski, e R. H. Ketchum, "Improved speech quality and efficient vector quantization in SELP," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 155–158, 1988.

[15] Lin, D., "New approaches to stochastic coding of speech sources at very low bit rates," in *Signal Processing III : Theories and Applications*, (*et al.*, I. T. Y., ed.), pp. 445–447, Amsterdam : The Netherlands : Elsevier, North-Holland, 1986.

[16] Linde, Y., A. Buzo, e R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Janeiro de 1980.

[17] Marca, J. R. B., N. Farvardin, N. S. Jayant, e Y. Shoham, "Robust vector quantization for noisy channels," *Proc. of Mobile Satellite Conference*, Pasadena, USA, pp. 515–520, Maio de 1988.

[18] Rabiner, L. R., e R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ : Prentice Hall, 1978.

[19] Schroeder, M. R., e B. S. Atal, "Code-excited linear prediction (CELP) : high quality speech at very low bit rates," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 937–940, 1985.

[20] Silva, L. M., e A. Alcaim, "Sub-optimal quantization of line spectral frequencies," *Intl. Telecomm. Symp.*, Acapulco, México, pp. 35–38, Outubro de 1996.

[21] Wang, S., A. Sekey, e A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Selected Areas in Commun.*, vol. 10, pp. 819–829, Junho de 1992.

Lúcio Martins da Silva - Graduado em Engenharia Eletrônica e de Telecomunicações pela PUC-MG em 1981, Mestre em Engenharia Elétrica pela Universidade de Brasília (UnB) em 1989 e Doutor em Ciências em Engenharia Elétrica pela PUC-Rio em 1996. Em 1982 e 1983 foi professor da PUC-MG e desde 1985 é professor da UnB. Suas áreas de interesse em pesquisa são : codificação de voz, reconhecimento de voz e reconhecimento de locutor.

Abraham Alcaim recebeu o diploma de Engenheiro Eletricista e o título de Mestre em Ciências em Engenharia Elétrica pela Pontifícia Universidade Católica do Rio de Janeiro (PUC/Rio) em 1975 e 1977, respectivamente, e os títulos de D.I.C. e Ph.D. em Engenharia Elétrica pelo Imperial College of Science and Technology, University of London, em 1981.

Desde 1976 ele é professor do Centro de Estudos em Telecomunicações da Universidade Católica (CETUC), tendo atualmente o cargo de Professor Associado.

O Dr. Alcaim trabalha há mais de 20 anos nas áreas de codificação digital e transmissão de sinais e processamento

digital de voz e imagem. Ele é autor de diversos artigos publicados em congressos e revistas nacionais e internacionais.

Em 1984 ele esteve por um período curto com o Centre National d'Etudes des Télécommunications (CNET), em Lannion, França, onde trabalhou em medidas de qualidade objetivas e subjetivas para codificadores de voz. De Dezembro de 1991 a Setembro de 1993 ele foi Cientista Visistante no Centro Científico Rio da IBM Brasil, onde trabalhou no projeto de novos codificadores de imagens, com aplicação especial para imagens obtidas por satélites de sensoreamento remoto.

O Dr. Alcaim foi o Technical Program Chairman dos simpósios internacionais SBT/IEEE International Telecommunications Symposiums de 1990 e 1994 – ITS'90 e ITS'94 – realizados no Rio de Janeiro em Setembro de 1990 e em Agosto de 1994, respectivamente. Ele é atualmente membro do comitê organizador do IEEE Global Telecommunications Conference (GLOBECOM'99) a ser realizado no Rio de Janeiro em Dezembro de 1999. O Dr. Alcaim é correspondente regional do IEEE Global Communications Newsletter.

Desde Março de 1996 ele é membro do Conselho Deliberativo da Sociedade Brasileira de Telecomunicações.