Codificador CELP a 4 Kbps Utilizando Coeficientes Mel-cepstrais Generalizados

Ricardo José da Rocha Cirigliano e Fernando Gil Vianna Resende Jr.

***Resumo*— Este trabalho apresenta um codificador CELP a 4 Kbps utilizando parâmetros mel-cepstrais generalizados (MCG) para modelagem do trato vocal ao invés dos tradicionais coeficientes LPC. A forma de obtenção da excitação do codificador proposto é similar à técnica dos codificadores CELP tradicionais, através de análise-por-síntese. Testes subjetivos mostram que a qualidade dos sinais codificados pela técnica CELP-MCG é supeiror à dos sinais codificados pela técnica CELP tradicional, sem que haja aumento na taxa de bits.**

***Palavras-Chave*—Codificação, CELP, Mel-cepstrais Generalizados.**

***Abstract*— This article presents a CELP coder at 4 Kbps using mel-generalized-cepstral (MGC) coefficients to model the vocal tract instead of the traditional LP coefficients. In this coder, the excitation is obtained in the same way as in the traditional CELP coders, through analysis-by-synthesis. Subjective tests showed that the quality of signals coded by the CELP-MGC is superior to the quality of signals coded by a traditional CELP coder, without increase on the bit rate.**

***Keywords*— Coding, CELP, Mel-generalized-cepstral.**

I. Introdução

O processamento digital de voz tem avançado muito nas últimas décadas. Isso tem sido motivado em grande parte pelo desenvolvimento da tecnologia digital, possibilitando a realização de tarefas antes consideradas de alto custo, e pela crescente necessidade dos sistemas de telefonia móvel em oferecer melhores produtos e serviços. Para o caso de codificação, a intenção é realizar a transmissão ou armazenamento da voz com a menor ocupação possível dos meios físicos, atingindo um requisito de qualidade que irá depender da aplicação a que se destina o codificador.

Dentre todas as técnicas de codificação de voz digital, a predição linear com excitação por códigos de dicionários (*code-excited linear prediction*, CELP) surgiu como uma possível alternativa capaz de atender a estes requisitos [1]. Contudo, a representação do trato vocal pelos coeficientes de predição linear não modela o espectro de forma fiel, uma vez que o filtro de síntese é composto somente por pólos, fazendo com que os zeros não sejam modelados. Uma alternativa ao uso dos coeficientes de predição linear são os coeficientes mel-cepstrais generalizados (MCG). Em [2] é mostrado que estes coeficientes, quando ajustados corretamente, conseguem modelar o espectro da voz de forma mais completa que os coeficientes de predição linear.

Ricardo José da Rocha Cirigliano e Fernando Gil Vianna Resende Jr., Programa de Engenharia Elétrica, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, E-mails: rjcirig@lps.ufrj.br e gil@lps.ufrj.br.

Neste trabalho foi desenvolvido um codificar CELP a 4 Kbps utilizando coeficientes MCG. A estrutura deste codificador é similar à estrutura dos codificadores CELP tradicionais, pela extração e síntese de coeficientes MCG. Testes subjetivos mostraram que os sinais codificados por este codificador apresentam qualidade superior à dos sinais codificados pelo codificador CELP apresentado em [3].

Este trabalho está dividido da seguinte forma: na Seção II são apresentados os coeficientes MCG e a forma como eles são extraídos e sintetizados; a Seção III descreve o codificador proposto; a Seção IV mostra os resultados dos testes subjetivos e a Seção V apresentas as conclusões.

II. Codificação CELP-MCG

A Figura 1 apresenta o diagrama de blocos simplificado do codificador CELP-MCG [4]. A estrutura do codificador CELP-MCG é muito semelhante à estrutura do codificador CELP convencional. Contudo, a utilização da análise MCG ao invés da análise LP resulta em mudanças significativas no interior de cada um dos blocos da Figura 1.

A. Análise MCG

Na análise MCG o espectro de voz $H(e^{j\omega})$ é modelado por um conjunto de coeficientes $c(m)$ por

$$H(z) = K.S(z) \tag{1}$$

onde K é um ganho e

$$S(z) = \begin{cases} (1+\gamma\sum_{m=0}^{M} c(m)\tilde{z}^{-m})^{1/\gamma}, -1 \leq \gamma < 0, \\ exp\sum_{m=0}^{M} c(m)\tilde{z}^{-m}, \gamma = 0 \end{cases} \tag{2}$$

onde $\tilde{z}^{-1}$ é definido como

$$\tilde{z}^{-1} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}, |\alpha| < 1. \tag{3}$$

O parâmetro α controla a compressão na freqüência. Ao aproximar α de zero, é obtida uma escala linear. Para o codificador apresentado neste trabalho o parâmetro α foi selecionado através de testes subjetivos informais que apresentaram como melhor valor 0,31. O parâmetro γ controla a acurácia na representação dos pólos e zeros da função de transferência. A função $H(z)$ passa a ser somente pólos quando $\gamma = -1$ e $\alpha = 0$. De forma a simplificar o filtro de síntese, apresentado na próxima seção, o parâmetro γ foi ajustado para -0,5. A Figura 2 apresenta a envoltória espectral de um trecho de voz modelado por coeficientes LPC e coeficientes MCG. Pode-se observar que os coeficientes MCG modelam o espectro de forma mais uniforme em relação aos pólos e zeros do que os coeficientes LPC que modelam somente os pólos.

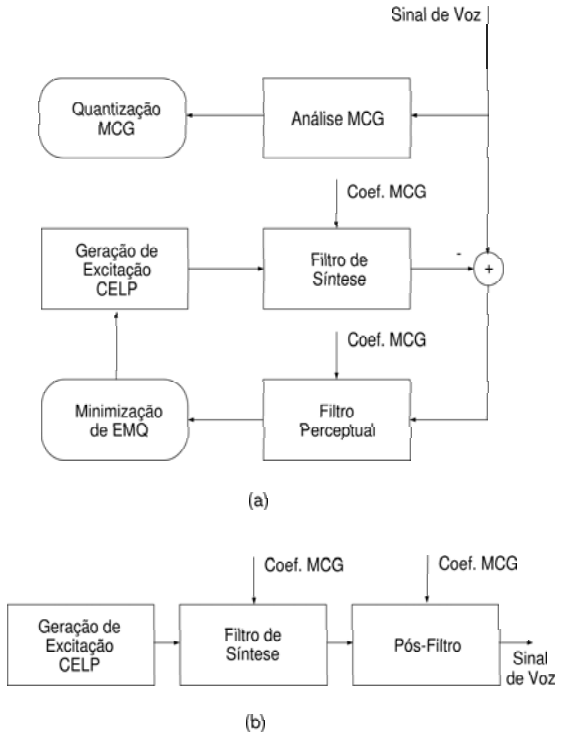

Fig. 1. Diagrama de blocos do codificador CELP-MCG.

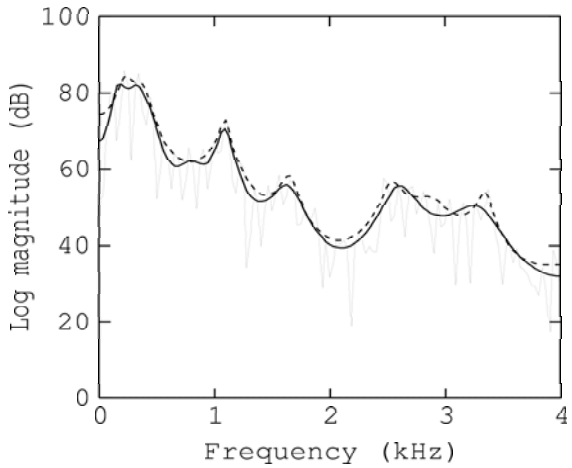

Fig. 2. Comparação entre a envoltória espectral dos coeficientes LPC (linha pontilhada) e dos coeficientes MCG (linha cheia).

B. Filtro de Síntese MCG

A utilização do parâmetro $\gamma = -0,5$ permite que o filtro de síntese apresente a seguinte função de transferência racional

$$S(z) = \frac{1}{(C(\tilde{z}))^2} \tag{4}$$

onde

$$C(\tilde{z}) = 1 + \gamma \sum_{m=0}^{M} c(m)\tilde{z}^{-1} \tag{5}$$

A equação (5) pode ser reescrita na forma

$$C(\tilde{z}) = 1 + \gamma \sum_{m=1}^{M} b(m)\Phi_m(z) \tag{6}$$

onde

$$\Phi_m(z) = \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\tilde{z}^{-(m-1)}, m \geq 1 \tag{7}$$

e os coeficientes $b(m)$ são obtidos por

$$b(m) = \begin{cases} c(M), m = M \\ c(m) - \alpha b(m+1), 0 \leq m < M. \end{cases} \tag{8}$$

C. Filtro Perceptual

O filtro perceptual é definido pelos coeficientes MCG como

$$S_p(z) = \frac{C(\tilde{z}/\beta_1)}{C(\tilde{z}/\beta_2)} \tag{9}$$

onde $\tilde{z}/\beta$ é uma expansão de banda no plano $\tilde{z}$. O filtro $C(\tilde{z}/\beta)$ possui a mesma estruta de $C(\tilde{z})$,

$$C(\tilde{z}/\beta) = 1 + \gamma \sum_{m=1}^{M} b_\beta(m)\Phi_m(z) \tag{10}$$

onde os coeficientes $b_\beta(m)$ são definidos em duas etapas. Na primeira são calculados os parâmetros $\hat{b}_\beta(m)$ através de

$$\hat{b}_\beta(m) = \begin{cases} \beta^M c(M), m = M, \\ \beta^m c(m) - \alpha\hat{b}(m+1), 0 \leq m < M. \end{cases} \tag{11}$$

Em seguida os coeficientes $b_\beta(m)$ são obtidos fazendo-se

$$b_\beta(m) = \frac{\hat{b}_\beta(m)}{1 + \gamma\hat{b}_\beta(0)}, 1 \leq m \leq M. \tag{12}$$

Testes subjetivos informais apresentaram melhores resultados fazendo-se $\beta_1 = 0,9$ e $\beta_2 = 0$. O valor de β_1 neste trabalho difere do valor sugerido em [4], que é 1.0.

D. Pós-Filtro

O pós-filtro do codificador CELP-MCG é composto de dois filtros em cascata: o $S_{sp}(z)$ e o $S_{tl}(z)$. O filtro $S_{sp}(z)$ é definido por

$$S_{sp}(z) = \frac{C(\tilde{z}/\beta_3)}{C(\tilde{z}/\beta_4)} \tag{13}$$

e é da mesma forma que o filtro perceptual apresentado em (9). O filtro $S_{tl}(z)$ é um filtro de compensação de *tilt* e possui a estrutura

$$S_{tl}(z) = (1 - \mu z^{-1})^n \tag{14}$$

onde o parâmetro μ controla o *tilt* espectral global. O valor de μ é obtido de forma que estes dois filtros em cascata façam $c(0) = 0$ [5]. Assim, μ é definido por

$$\mu = \frac{-\gamma(\beta_4 - \beta_3)c_1(1)}{-\alpha\gamma(\beta_4 - \beta_3)c_1(1) + n(1-\alpha^2)}. \tag{15}$$

Testes subjetivos informais apresentaram melhores resultados utilizando-se $\{\beta_3\ \beta_4\ n\} = \{0,8\ 0,9\ 2\}$.

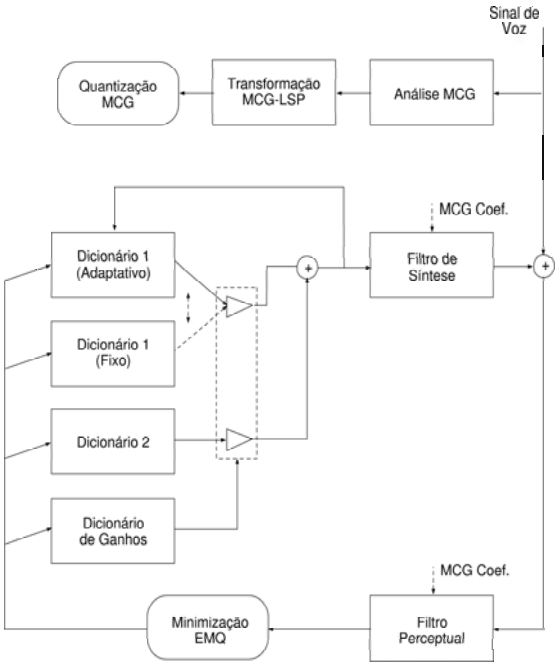

Fig. 3. Diagrama de blocos completo do codificador CELP-MCG.

III. Codificador CELP-MCG à 4 kbps

Esta seção descreve o codificador CELP-MCG à 4 kbps. O diagrama em blocos completo do codificador é apresentado na Figura 3. Cada bloco processado é composto por 240 amostras (30 ms) divididas em 4 sub-blocos de 60 amostras cada (7,5 ms).

A. Extração de Parâmetros MCG

A extração dos parâmetros MCG é realizada uma vez a cada bloco utilizando-se uma janela de Hamming de 200 amostras centralizada no último sub-bloco de cada bloco. São extraídos $M = 20$ coeficientes. Os coeficientes MCG são transformados em coeficientes MCG-LSP [7], pois estes apresentam melhores resultados de quantização e interpolação. O vetor de pesos utilizado na interpolação dos parâmetros MCG-LSP é dado por

$$\mathbf{p}_i = [0{,}1 \quad 0{,}6 \quad 0{,}8 \quad 1{,}0] \tag{16}$$

B. Quantização dos coeficientes MCG-LSP

Os coeficientes MCG-LSP são quantizados pela técnica SSVQ (*switched split vector quantization*) [8]. Primeiramente é calculada a distância entre o vetor de MCG-LSP [7] a ser quantizado e cada vetor de um dicionário direcional. O objetivo deste dicionário é indicar uma "direção" para a qual o vetor original aponta. A distância utilizada é a Euclidiana ponderada, onde os pesos [9] são dados por

$$p_q = \begin{cases} [|H(e^{j\omega_i})|]^{0.6}, 0 \leq i < 11, \\ 0.8[|H(e^{j\omega_i})|]^{0.6}, 11 \leq i < 17, \\ 0.4[|H(e^{j\omega_i})|]^{0.6}, 17 \leq i < 20, \end{cases} \tag{17}$$

Uma vez identificada a "direção" do vetor de MCG-LSP, este é quantizado através de dois dicionários utilizando-se a técnica *split*, onde os $M = 20$ coeficientes são divididos ao meio, sendo 10 quantizados pelo primeiro dicionário e 10 pelo segundo. Cada uma das direções possui um par de dicionários, o que torna a quantização mais robusta. O dicionário direcional é composto por 8 vetores e cada um dos dicionários *split* possui 64 vetores.

C. Parâmetros de Excitação

O codificador CELP-MCG apresentado neste trabalho possui dois dicionários de excitação: o primeiro dicionário é composto por um dicionário adaptativo e um dicionário fixo [6], e o segundo é um dicionário gaussiano.

1) Primeiro Dicionário de Excitação: O primeiro dicionário de excitação é composto por um dicionário adaptativo e um dicionário fixo, utilizando-se a técnica PSI-CELP [6]. O dicionário adaptativo apresenta somente atrasos inteiros, variando de 20 até 146 amostras. O dicionário fixo apresenta vetores aleatórios com distribuição gaussiana e média zero. A busca em ambos os dicionários é feita passando-se o vetor de excitação candidato $\mathbf{x}_{d1}$ pelo filtro de síntese e correlacionando-se a saída $\mathbf{y}_{d1}$ com o sinal de voz original $\mathbf{t}$ através de

$$corr = \sum_{n=0}^{N-1} y_{d1}(n)t(n) \tag{18}$$

onde N é o número de amostras das seqüências. Somente para as seqüências com $corr > 0$ é calculada a distorção. Se todas as seqüências candidatas apresentarem $corr \leq 0$, a excitação do sub-bloco será composta apenas por um vetor do segundo dicionário de excitação e o seu respectivo ganho. Ambos dicionários possuem 128 vetores.

2) Segundo Dicionário de Excitação: O segundo dicionário de excitação é composto por vetores gaussianos com média zero e variância unitária. Em cada vetor foi utilizado um limitador de forma que todas as amostras com valor abaixo de $1{,}645$ fossem alteradas para zero. Isto garantiu que cada vetor possuisse apenas cerca de 5 ou 6 pulsos, resultando em 90% de esparcidade no dicionário. Este dicionário é composto por 512 vetores.

D. Ganhos dos Dicionários

Os ganhos G_{d1} e G_{d2} calculados através de

$$G_{d1} = \frac{\mathbf{t}^T\mathbf{y}_{d1_{ot}} - \mathbf{y}_{d1_{ot}}^T\mathbf{y}_{d2_{ot}}\mathbf{t}^T\mathbf{y}_{d2_{ot}}}{\mathbf{y}_{d1_{ot}}^T\mathbf{y}_{d1_{ot}}} \tag{19}$$

$$G_{d2} = \frac{\mathbf{t}^T\mathbf{y}_{d2_{ot}} - \mathbf{y}_{d1_{ot}}^T\mathbf{y}_{d2_{ot}}\mathbf{t}^T\mathbf{y}_{d1_{ot}}}{\mathbf{y}_{d2_{ot}}^T\mathbf{y}_{d2_{ot}}} \tag{20}$$

onde $\mathbf{y}_{d1_{ot}}$ e $\mathbf{y}_{d2_{ot}}$ são os vetores ótimos respectivamente do primeiro e do segundo dicionário de excitação. Após o cálculo ambos são quantizados de forma escalar utilizando respectivamente 4 e 5 bits.

E. Taxa de Bits

A Tabela I apresenta a alocação de bits total do codificador CELP-MCG.

TABELA I
ALOCAÇÃO DE BITS.

	Sub-bloco	Bloco
MCG-LSP	-	21
Dicionário 1	7	7x4
Dicionário 2	9	9x4
G_{d1}	4	4x4
G_{d2}	5	5x4
Total	25	121

IV. RESULTADOS

Foram realizados dois testes subjetivos para avaliar o codificador: MOS (*mean opinion score*) e DMOS (*degradation mean opinion score*). Ambos os testes foram realizados também com o codificador CELP tradicional apresentado em [3] com o objetivo de comparar os resultados com o codificador apresentado neste trabalho.

A. Teste MOS

Neste teste foram apresentadas a cada ouvinte 6 frases codificadas por cada técnica, 3 faladas por locutores masculinos e 3 por locutores femininos, para as quais eles deveriam dar uma nota entre 1 e 5. O teste foi realizado com 6 ouvintes. Os resultados podem ser encontrados na Tabela II.

TABELA II
RESULTADOS DO TESTE MOS.

Sentença	CELP-MCG	CELP
Masc. 1	4.4	4.1
Masc. 2	3.4	3.5
Masc. 3	4.2	3.8
Fem. 1	3.8	3.9
Fem. 2	3.9	3.7
Fem. 3	3.9	3.5
Média	3.93	3.76

O resultado do teste subjetivo MOS indica que os sinais codificados pela técnica CELP-MCG apresentam qualidade superior aos sinais codificados pela técnica CELP tradicional. Durante os testes foi solicitado que cada ouvinte apontasse o artefato mais marcante de cada um dos dois sinais apresentados. De modo geral, o artefato mais sinalizado pelos ouvintes no sinal codificado pela técnica CELP tradicional foi a presença de altas freqüências que distorciam o sinal de voz, enquanto no sinal codificado pela técnica CELP-MCG o artefato mais sinalizado era a presença de ruído, principalmente nos trechos de silêncio. Nenhum ouvinte apontou distorções espectrais como artefato nos sinais codificados pela técnica CELP-MCG.

B. Teste DMOS

Neste teste, a cada ouvinte foi apresentado um sinal de referência (sinal original) ao qual estava associada a nota 5. Em seguida foi apresentado o mesmo sinal codificado pelas duas técnicas para os quais eles deveriam dar notas entre 1 e 5. Foram apresentadas 6 frases, sendo 3 faladas por locutores masculinos e 3 por locutores femininos. O teste foi realizado com 6 ouvintes, Os resultados são apresentados na Tabela III.

TABELA III
RESULTADOS DO TESTE DMOS.

Sentença	CELP-MCG	CELP
Masc. 1	4.1	4.0
Masc. 2	3.4	3.4
Masc. 3	3.8	3.6
Fem. 1	3.7	3.9
Fem. 2	3.8	3.7
Fem. 3	3.6	3.3
Média	3.73	3.65

O teste DMOS, assim como o teste MOS, mostra que o sinal codificado pela técnica CELP-MCG apresenta qualidade superior ao sinal codificado pela técnica CELP tradicional.

V. CONCLUSÕES

Neste trabalho foi apresentado um codificador de voz CELP a 4 kbps utilizando parâmetros mel-cepstrais generalizados (CELP-MCG) no lugar dos tradicionais parâmetros LPC para a modelagem do trato vocal. A forma de obtenção da excitação no codificador CELP-MCG é similar à técnica utilizada nos codificadores CELP tradicionais (análise-por-síntese). Dois testes subjetivos, MOS e DMOS, foram realizados para avaliar a qualidade do sinal codificado pela técnica CELP-MCG quando comparado ao sinal codificado pela técnica CELP tradicional. Os resultados mostram que os sinais codificados pela técnica CELP-MCG apresentam qualidade superior à dos sinais codificados pela técnica CELP tradicional, sem que haja aumento na taxa de bits.

REFERÊNCIAS

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates", Proc. of the IEEE ICASSP, p. 937-940, 1985.

[2] K. Koishida, G. Hirabayashi, K. Tokuda and T. Kobayashi, "A 16 kb/s Wideband CELP-Based Speech Coder Using Mel-Generalized Cepstral Analysis", IEICE Trans. Inf. & Syst., v. E83-D, n. 4, Abril, 2000.

[3] R. S. Maia, R. J. R. Cirigliano, D. Rojtenberg and F. G. V. Resende Jr., "CELP speech coding: a comparison in terms of quantization techniques for the synthesis filter parameters", Proc. ITS2002, Setembro, 2002.

[4] K. Tokuda, T. Kobayashi, T. Chiba, and S. Imai, "Spectral estimation of speech by mel-generalized cepstral analysis", IEICE Trans., v. J75-A, n. 7, p.1124-1134, Julho, 1992.

[5] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "CELP speech coding based on mel-generalized cepstral analysis", IEICE Trans., v. J81-A, n. 2, p. 252-260, Fevereiro, 1998.

[6] S. Miki, K. Mano, H. Ohmuro and T. Moriya, "Pitch synchronous Innovation CELP (PSI-CELP)", Proc. EUROSPEECH, p. 261-264, 1993.

[7] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "Spectral representation of speech based on mel-generalized cepstral coefficients and its properties", IEICE Trans., v. J80-A, n. 11, p. 1999-2006, Novembro, 1997.

[8] S. So and K. K. Paliwal, "Efficient Vector Quantisation of Line Spectral Frequencies Using the Switched Split Vector Quantiser", Proc. ICSLP-2004, Outubro, 2004.

[9] McCree, K. Truong, E. George, T. Barnwell, and V. Viswanathan, "A 2.4 kbits/s MELP coder candidate for the new U.S. federal standard", Proc. ICASSP, p. 200-203, 1996.