OTIMIZAÇÃO DOS CODIFICADORES VSELP E EFR POR REFINAMENTO DA MODELAGEM AUTOREGRESSIVA

IRENE HELEONORA S. P. FANTINI E LUÍS GERALDO P. MELONI

Laboratório de Processamento Digital de Fala - LPDF
Departamento de Comunicações
Faculdade de Engenharia Elétrica e de Computação – UNICAMP
P.O. Box 6101, CEP: 13083-970 – Campinas – SP – BRASIL
{irene,meloni}@decom.fee.unicamp.br

ABSTRACT

This paper extends the use of the refined autoregressive modeling to the EFR coder. The refined redesign is based on the observation that a better formant estimation is obtained by the use of a frame length multiple of pitch period and a synchronous frame position to glottal opening and closure. Simulations using telephone speech signals have shown that it is possible to obtain coding gains up to 3.5 dB with the segmental signal-to-noise ratio. The decoder does not require any modification and yet offers a better speech quality. The paper also compares the VSELP and EFR performances for both original and refined coders using segmental SNR and PSQM measures.

Palavras-chaves – *Codificação de Fala, EFR, VSELP, ACELP, TDMA, PSQM.*

1. INTRODUÇÃO

Nos últimos anos houve um avanço significativo na área de codificação de fala, oferecendo boa qualidade para taxas de transmissão de bits em torno de 8 kbit/s. Os codificadores que operam nesta taxa são geralmente baseados em modelos de produção da fala que representam o trato vocal e a fonte de excitação. Este trabalho se dedica ao estudo de dois codificadores padronizados pela *Telecommunications Industry Association* (TIA): o VSELP - *Vector-Sum Excited Linear Predictive Coding* [1] e o EFR - *Enhanced Full Rate* [2]. Estes codificadores são utilizados na telefonia celular pela técnica de múltiplo acesso TDMA.

Este artigo compara a qualidade dos dois codificadores e emprega um método de refinamento na modelagem do trato vocal (modelagem autoregressiva - AR)[3] ao codificador EFR. Sabe-se que a modelagem AR usando um quadro de análise de tamanho múltiplo do período fundamental (*pitch*) e com posição síncrona à abertura ou fechamento da glote melhora a estimação dos formantes [4][5]. Como os quadros de análise podem ser de curta duração (da ordem de 5 ms), o método da covariância na modelagem AR é mais apropriado. Para garantir a estabilidade dos modelos AR's, utiliza-se o método da covariância modificada [6]. Embora existam algoritmos para se determinar os instantes de abertura e fechamento da glote [7], neste trabalho, a busca da posição do quadro dentro do quadro padrão dos codificadores é feita de forma exaustiva. Não é intenção a aplicação em tempo real, mas a verificação da melhoria dos codificadores através de simulações por computador. Recentemente foi publicada pelos autores a aplicação deste método para o codificador VSELP [3]. Este trabalho estende os resultados para o codificador EFR e apresenta resultados por uma medida perceptiva, a *Perceptual Speech Quality Measure* (PSQM) [8]. Esta é uma medida objetiva que tem forte correlação com a medida subjetiva MOS. Este trabalho também apresenta resultados comparativos dos codificadores originais VSELP e EFR. O codificador EFR apresenta um conjunto de melhorias em relação ao primeiro. Estas melhorias refletem na qualidade da SNR segmentar média de 3,4 dB acima da do VSELP.

O trabalho envolve dois tipos de simulações :

- a primeira busca a maior SNR variando a duração e a posição do quadro de análise dentro do quadro do padrão;
- a segunda utiliza um detector de *pitch* [9] [10] para os quadros sonoros como estimativa da duração ótima, e busca apenas a posição do quadro de análise a fim de obter a maior SNR.

A modificação mantém a mesma distribuição de bit do codificador original. Desta forma não é necessária nenhuma modificação no decodificador, existindo uma compatibilidade total entre os codificadores original e refinado.

2. REVISÃO DO CODIFICADOR EFR

Esta seção apresenta uma breve revisão do codificador EFR. Um diagrama de blocos simplificado do codificador EFR é mostrado na Fig. 1.

O sinal de fala é filtrado por um filtro passa-alta com freqüência de corte de 80 Hz que além de eliminar as freqüências indesejáveis realiza uma redução da faixa dinâmica do sinal. A modelagem AR é feita pelo método da autocorrelação através do algoritmo de Levinson-Durbin [11]. A matriz de autocorrelação é obtida a partir do sinal filtrado e é multiplicada por uma janela exponencial para expandir as bandas de formantes [2]. Os coeficientes AR são transformados em pares de freqüências espectrais (LSF) [12]. Estas freqüências espectrais estão relacionadas com as raízes dos polinômios

$$F_{1,2}(z) = A(z) \pm z^{-11} A(z^{-1}) . \qquad (1)$$

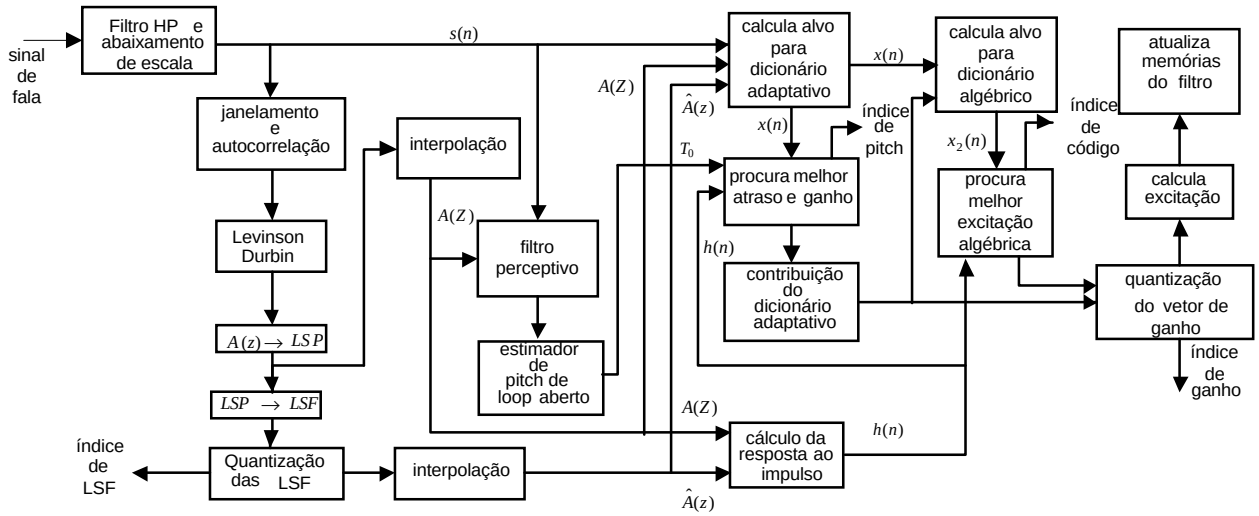

Fig. 1 - Diagrama em blocos do codificador de fala IS-641 EFR.

Os coeficientes LSF quantizados e não quantizados são interpolados linearmente para obtenção dos coeficientes de quatro subquadros.

O sinal *s(n)* é filtrado por um filtro perceptivo. Este filtro atua dinamicamente de forma a mascarar o sinal de erro em zonas espectrais ao redor dos formantes.

Após a modelagem AR do sinal, é realizada uma estimativa do período de *pitch*, que ocorre em duas fases. Primeiro, estima-se um valor de *pitch* inteiro em malha aberta; e então, um valor mais preciso é calculado em malha fechada. Este valor é quantizado pelo dicionário adaptativo. O terceiro dado a ser transmitido é o dicionário fixo, que representa a excitação do sinal. Este dicionário é do tipo algébrico, portanto o codificador é da família ACELP - *Algebraic Coder Excited Linear Prediction*.

Os coeficientes LSF são codificados a cada 20 ms (duração do quadro). Os parâmetros de estimação de *pitch* e excitação são calculados a cada subquadro (5 ms). Os ganhos da excitação e de *pitch* são quantizados vetorialmente em conjunto. Os quadros de análise de 240 amostras de fala são superpostos em 80 amostras. Estas melhorias no codificador oferecem um sinal de fala de alta qualidade.

3. A MEDIDA PSQM

A PSQM [8] é uma medida objetiva da qualidade subjetiva que tem forte correlação com o MOS.

A PSQM é uma medida relativa que compara o sinal fonte com o sinal codificado. O método busca imitar a percepção do som em situações reais e abstrair as diferenças imperceptíveis. Em particular, se a entrada e a saída forem idênticas, a PSQM irá predizer uma qualidade perfeita, independente da natureza do sinal de entrada.

Na PSQM os sinais físicos de voz (fonte e codificado) são mapeados em representações psicofísicas que coincidem com as representações internas dos sinais de fala. Estas representações fazem uso de equivalentes psicofísicos de freqüência e de intensidade. O mascaramento dos sons é modelado de maneira simples: apenas quando dois componentes tempo-freqüência coincidem em ambos os domínios, o mascaramento é levado em conta.

A qualidade do sinal de fala processado é julgada com base em diferenças da representação interna. Estas diferenças são usadas para o cálculo da perturbação de ruído como uma função do tempo e da freqüência. Na PSQM a perturbação média de ruído é diretamente relacionada com a qualidade da fala codificada.

A transformação do domínio físico (externo) para o domínio psicofísico (interno) é realizada por três operações :

- mapeamento tempo-freqüência;
- conversão não-linear da escala de freqüências; e,
- conversão não-linear da escala de intensidades (compressão).

No cálculo da perturbação do ruído, é feita uma modelagem cognitiva com o intuito de obter uma alta correlação entre medidas objetivas e subjetivas.

Na Fig. 2 tem-se um gráfico da relação MOS x PSQM para a língua inglesa [13]. Note que para valores altos da MOS, tem-se pequenos valores da PSQM. Portanto a qualidade do sinal será tanto melhor quanto menor for o valor da PSQM. Por meio deste gráfico de conversão, o valor máximo que pode ser obtido na escala MOS é de 4. Embora não exista uma regra estrita de conversão da medida PSQM para a MOS, emprega-se neste trabalho esta conversão apenas para referência à medida MOS, por ser esta última mais antiga.

4. REFINAMENTO DO MODELO

O método de refinamento da modelagem autoregressiva foi apresentado em detalhes em [3]. São realizados dois tipos de simulações. A primeira realiza a busca exaustiva da posição e do tamanho do quadro de análise dentro do quadro do padrão. O menor período de *pitch* adotado é de 30. Para o VSELP [14] o tamanho do quadro é de 170 amostras, assim o intervalo de

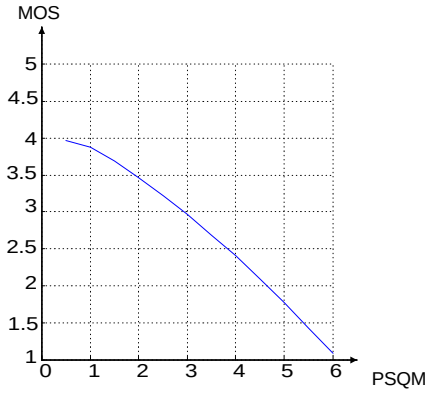

Fig. 2 - Gráfico da relação MOS x PSQM

variação do *pitch* é de 30 a 170 amostras. Como a posição ótima é procurada a cada quadro, o número total de variações de tamanho e posição é 9.730 para cada 20 ms de sinal. Para o EFR o tamanho do quadro é de 240 amostras e a variação de *pitch* é de 30 a 240. Neste caso o número total de variações de tamanho e posição de quadro é de 21.945. Esta simulação é chamada de *busca exaustiva*.

Adicionando-se um detetor de *pitch*, o esforço computacional diminui drasticamente. De posse do valor estimado *P*, a busca da posição ótima é realizada de 1 a 170-*P* para o VSELP, e de 1 a 240-*P* para o EFR. Este método é chamado de *busca simplificada.*

5. SIMULAÇÕES

Os sinais de fala usados nas simulações são retirados do "*Telephone Network Acoustic-Phonetic Continuous Speech Corpus"*- NTIMIT [15]. Esta base de sinais emprega linhas telefônicas reais e as elocuções são do Inglês americano.

Foram escolhidas seis frases (três masculinas e três femininas) as quais são amostradas a 16 kHz e listadas na Tabela 1. As frases foram filtradas e decimadas para a freqüência de amostragem de 8kHz conforme filtro definido em [3]. Ambos os codificadores trabalham com esta freqüência de amostragem.

Tabela 1 - Sinais de fala do NTIMIT usados nas simulações.

No.	Locutor	Sexo	Duração
1	DR5\MHIT0\SX263	Masculino	2,0 s
2	DR3\MADC0\SX17	Masculino	1,7 s
3	DR8\MBSB0\SX363	Masculino	2,0 s
4	DR2\FAEM0\SX402	Feminino	2,8 s
5	DR5\FBMH0\SX56	Feminino	2,5 s
6	DR1\FCJF0\SX307	Feminino	1,6 s

Uma das medidas de desempenho utilizada é a SNR segmentar. Esta é calculada a partir do sinal após o pré-filtro e anterior ao pós-filtro. O pós-filtro tem como função melhorar a qualidade perceptiva do sinal de fala. A medida PSQM é realizada com o sinal de fala original e o sinal decodificado após o pós-filtro.

Como é assumido que a melhoria da qualidade do modelagem AR ocorre para quando o mesmo é síncrono ao período de *pitch*, os resultados são analisados apenas para quadros sonoros. Para os quadros surdos utiliza-se o tamanho original do quadro. Observa-se que a SNR para os quadros surdos são menores que para quadros sonoros nos codificadores originais.

Um detector sonoro/surdo é utilizado com o propósito de calcular a SNR média nos trechos sonoros. O detector sonoro/surdo é baseado nos seguintes parâmetros: o valor da correlação cruzada do detector de *pitch*, o valor RMS do quadro, a taxa de cruzamento por zero e o coeficiente de autocorrelação de curto termo normalizado para atraso unitário [3]. O desempenho deste detector não é crítico, e qualquer outro detector pode ser usado sem mudanças significativas nos resultados.

A PSQM recomendada pela ITU-T trabalha com sinais com freqüência de amostragem de 16 kHz. Para o emprego desta medida, faz-se a interpolação dos sinais codificados a 8 kHz.

Os resultados das simulações para o codificador VSELP estão nas Tabelas 2 e 3. A Tabela 2 mostra o comportamento da SNR segmentar para os métodos utilizados. Para a busca exaustiva temos o maior valor médio de SNR segmentar. Em relação ao codificador original temos a diferença média de 2,35 dB. Para a busca simplificada o ganho médio é de 1,45 dB. O melhor resultado foi obtido para a frase 1 com 3,23 dB de ganho para o método exaustivo e 1,9 dB de ganho para o método simplificado.

Tabela 2 - SNR segmentar para o codificador VSELP.

No.	Codificador Original	Método Exaustivo	Método Simplificado
1	10,71	13,94	12,61
2	10,60	12,42	11,82
3	9,97	12,65	11,68
4	10,58	13,23	12,26
5	11,25	13,72	12,90
6	9,98	11,23	10,56
Média	10,52	12,87	11,97

Na Tabela 3, tem-se os resultados da medida PSQM para o VSELP. Note que para o método simplificado, mesmo havendo ganhos para a medida SNR segmentar, não há ganhos para a medida PSQM, e consequentemente para a MOS. Observa-se uma melhoria subjetiva do sinal codificado para o método exaustivo, com um ganho de 22% obtido pelo método exaustivo. A correspondência entre a PSQM e o MOS é feita a partir da Fig. 2.

Tabela 3 - Valores PSQM obtidos para o codificador VSELP e o valor MOS médio para cada método.

No.	Codificador Original	Método Exaustivo	Método Simplificado
1	2,12	1,66	2,29
2	2,32	1,35	2,37
3	1,60	1,41	1,65
4	2,33	1,81	2,23
5	1,63	1,42	1,79
6	1,09	0,98	1,08
Média	1,85	1,44	1,90
MOS	3,5	3,7	3,5

Os resultados das simulações para o EFR estão nas Tabelas 4 e 5. Observa-se na Tabela 4 que a SNR segmentar média do codificador original está 3,47 dB acima daquela do VSELP original.

Tabela 4 - SNR segmentar para o codificador EFR.

No.	Codificador Original	Método Exaustivo	Método Simplificado
1	13,88	16,57	15,85
2	13,81	16,00	15,70
3	13,36	16,89	15,71
4	14,39	16,45	16,00
5	14,63	16,75	16,33
6	13,88	15,62	15,24
Média	13,99	16,55	15,81

Observa-se também que a qualidade subjetiva do codificador EFR é maior que a do VSELP. No EFR o valor MOS médio é 8% maior que o MOS médio do VSELP. Isto comprova o melhor desempenho do codificador EFR.

Da Tabela 4, tem-se que a SNR segmentar média para o método exaustivo é 2,56 dB maior que o codificador original. O maior ganho foi obtido para a frase 3, de 3,53 dB. Para o método simplificado o ganho é de 1,82.

Tabela 5 - Valores PSQM obtidos para o codificador EFR e o valor MOS médio para cada método.

No.	Codificador Original	Método Exaustivo	Método Simplificado
1	1,26	1,20	1,18
2	1,47	0,72	1,32
3	1,04	0,96	1,02
4	1,41	1,34	1,38
5	1,06	1,04	1,02
6	0,82	0,42	0,82
Média	1,18	0,95	1,12
MOS	3,78	3,95	3,85

Da Tabela 5, observa-se um ganho de 19,5% obtido pelo método exaustivo em relação ao codificador original para o valor PSQM médio. Tal como ocorre para o VSELP, a melhoria é menor no método simplificado. O ganho da PSQM é de 5%.

6. SUMÁRIO

Este artigo apresenta simulações que comprovam um melhor desempenho dos codificadores VSELP e EFR por meio de um refinamento da modelagem AR. Dois tipos de simulações são realizados: uma baseada em busca exaustiva do tamanho e posição do quadro de análise e outra que se baseia na estimação do período de *pitch*.

A qualidade subjetiva do sinal codificado foi medida. Para o método simplificado praticamente não há ganho em relação ao codificador original. Com relação ao método exaustivo, há ganho de 22% para o VSELP e 19,5% para o EFR na PSQM. Portanto, o método de fato introduz melhorias refletidas tanto por medidas objetivas quanto subjetivas. O uso de um algoritmo de estimação de *pitch* reduz bastante o esforço computacional, porém compromete os ganhos de qualidade.

7. REFERÊNCIAS

[1] TIA/EIA/IS-136.2, "800 MHz TDMA Cellular – Radio Interface – Mobile Station – Base Station Compatibility – Traffic Channels and FSK Control Channel", Telecomm. Industry Association, Dez. 1994.

[2] TIA/EIA/IS-641, "TDMA Cellular/PCS – Radio Interface – Enhance Full-Rate Speech Codec", Telecomm. Industry Association, Maio 1996.

[3] Fantini, I.H.S.P. and Meloni, L.G.P. "Enhanced VSELP coding by a refined autoregressive modeling", XVII Simpósio Brasileiro de Telecomunicações, Vila Velha, ES, pp. 126-129, Set. 1999.

[4] Markel J. D. e Gray A. H. *"Linear Prediction of Speech"*, Springer-Verlag, 1976.

[5] Rabiner L.R. e Atal B.S. "LPC prediction error - analysis of its variation with the position of the frame analysis". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, no. 5, pp. 434-442, Out. 1977.

[6] Dickinson B.W. "Autoregressive estimation using energy ratios", *IEEE Trans. on Information Theory*, vol. 24, no. 4, pp. 503-506, Jul. 1978.

[7] Yegnanarayana B. e Ananthapadmanabha T.V. "Epoch extraction from linear prediction residual for identification of closed glottis interval", *IEEE Trans. Acoust., speech, Signal Processing*, vol. 27, no. 4, pp. 309-319, Ago. 1979.

[8] ITU-T Recommendation P.861, "Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs", Ago. 1996

[9] Hess W. "*Pitch determination of Speech signals"*, Springer-Verlag, 1983.

[10] Medan Y., Yair E. e Chazan D., "Super resolution pitch determination of speech signals", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, no. 1, pp. 40-48, Jan. 1991.

[11] Hayes M.H. *"Statistical Digital Signal Processing and Modeling"*, John Wiley & Sons, 1996.

[12] Honkanen T., Vanio J., Järvinen K. e Haavisto P., "Enhanced full rate speech codec for IS-136 digital cellular system", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Munich, Germany, pp. 731-734, 1997.

[13] Celso S. Kurashima, "*Implementação de um Pós-Filtro Adaptativo para a Melhoria de Qualidade Perceptual de Sinais de Voz com Ruído*", Dissertação de Mestrado, Escola Politécnica da USP, 1999.

[14] Gerson I.A. e Jasuik M., "Vector sum excited linear prediciton (VSELP) speech coding at 8 kb/s", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Albuquerque, pp. 461-464, Abril, 1990.

[15] Jankowski C., Kalyanswamy A., Basson S. e Spitz J. "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Albuquerque, Abril, 1990.