COMPENSAÇÃO DE ESPARGIMENTO EM CODIFICADORES DE VOZ

Miguel Arjona Ramírez

Depto. de Engª de Sistemas Eletrônicos - Escola Politécnica
Universidade de São Paulo, São Paulo, SP
miguel@lps.usp.br

RESUMO

A excitação esparsa possibilita o emprego de algoritmos de busca eficientes para a obtenção de alta qualidade de reprodução da voz. Entretanto, conforme se diminui a taxa de transmissão, o espargimento começa a degradar a qualidade com o surgimento de efeitos auditivos que podem ser compensados em fase. Descreve-se um procedimento para treinar a resposta impulsiva de um filtro para compensação de fase. Os filtros de compensação de dois codificadores ACELP de quatro pulsos à taxa de 5,3 kbit/s são treinados com os sinais de uma extensa base de dados. Um dos codificadores usa a busca focalizada e outro emprega a busca conjunta de posição e amplitude, operacionalmente menos complexa. Efetuam-se testes subjetivos de audição envolvendo estes codificadores compensados e também versões sem compensação operando a 8 kbit/s. Os resultados indicam que a compensação treinada de espargimento eleva a qualidade do codificador de taxa baixa com busca conjunta a um nível próximo da qualidade dos codificadores de taxa mais alta. Por outro lado, a elevação da qualidade do codificador compensado com busca focalizada é bem menos significativa.

1. INTRODUÇÃO

A maior parte das amostras das excitações esparsas é nula e as posições das amostras não-nulas são atribuídas de forma estocástica ou determinística. O uso de dicionários fixos esparsos permitiu sua aplicação generalizada devido à existência de algoritmos de busca eficientes. Além disso, um efeito benéfico que também contribui para essa popularidade é a melhora da qualidade perceptual que a excitação esparsa proporciona [1]. Em particular, dentre os codificadores de voz com algoritmos ACELP e dicionários multipulso algébricos padronizados mais recentemente encontra-se o codificador G.729 da ITU-T [2], que opera à taxa de 8 kbit/s. Este tipo de excitação determinística esparsa é usada também em codificadores de voz com taxas mais baixas de operação como o G.723.1 da ITU-T [3] para a taxa de 5,3 kbit/s. Porém, este último codificador apresenta distorção mais audível em decorrência do espargimento mais acentuado da sua excitação. Por esta razão, foi escolhido para a aplicação de dois métodos de compensação do espargimento enquanto o primeiro codificador foi tomado como referência.

Como já foi mencionado, a motivação principal para o uso de excitação esparsa é a aplicação de algoritmos computacionalmente mais eficientes para a busca da excitação fixa. Assim, enquanto os codificadores G.729 e G.723.1 de referência usam o algoritmo de busca focalizada ("focused search") da excitação fixa, desenvolveu-se o método de busca em árvore "depth first tree search", DFTS, especialmente para um codificador multimídia para voz digital e dados simultâneos ("digital simultaneous voice and data" - DSVD), que veio a ser incorporado na recomendação G.729A da ITU-T [4].

Uma aplicação de larga escala de codificadores esparsos eficientes ocorre no sistema D-AMPS TDMA de telefonia celular IS-136, que incorpora o codec IS-641-A como versão melhorada para a taxa plena, denominado "enhanced full rate (EFR) codec" [5].

Os algoritmos de busca DFTS e EFR ganham eficiência sobre a busca focalizada de referência na determinação das posições dos quatro pulsos porque realizam buscas preliminares sobre subconjuntos de pares de posições [5, 6].

Um algoritmo de busca eficiente de excitação esparsa, denominado "busca conjunta de posição e amplitude" ("joint position and amplitude search" - JPAS) [7], preserva mais a informação de amplitude dos pulsos e tem mais precisão na determinação da posição do pulso dominante, apresentando um ganho de eficiência decorrente da redução do número de combinações de posições pesquisadas para os demais pulsos.

Os algoritmos acima foram anteriormente comparados em complexidade e em qualidade de reprodução [8] para a taxa de operação de 8 kbit/s. Para a medida de complexidade operacional de implementações com aritmética de ponto fixo usou-se a unidade WMOPS

("weighted million operations per second"), contabilizada no pior caso, e tomou-se a complexidade da busca focalizada, 6,31 WMOPS, como referência. Assim, a complexidade medida para a busca JPAS foi abaixo de 25% enquanto as complexidades medidas para as buscas DFTS e EFR foram de 30% e 54%, respectivamente. As qualidades de reprodução da voz pelos codificadores com os três algoritmos de busca eficiente excederam a qualidade percebida do codificador de referência com a busca focalizada.

Assim, para avaliar o impacto de um método de busca eficiente no processo de compensação de espargimento, implementou-se a busca JPAS num codificador com taxa de operação de 5,3 kbit/s.

2. COMPENSAÇÃO DE FASE

O espargimento da excitação fixa é mais notado auditivamente quando a excitação ideal do filtro de síntese apresenta uma distribuição mais uniforme de energia no tempo. Isto ocorre quando o segmento de voz em questão é surdo, solicitando uma contribuição relativamente maior do dicionário fixo na composição da excitação total, que não é periódica. Por outro lado, quando o segmento de voz é sonoro, o dicionário adaptativo fornece a maior contribuição relativa e a componente complementar do dicionário fixo é apropriada para a modelagem da excitação ideal, que é de natureza periódica impulsiva e, portanto, esparsa neste caso. De fato, a melhora perceptual ocasionada pela excitação esparsa, comentada na Seção 1, deve decorrer deste casamento de características entre o dicionário esparso e a excitação necessitada pelos segmentos quase-periódicos.

As observações acima permitem compreender que os efeitos causados pela excitação esparsa sobre o sinal de voz reconstruído ocorram predominantemente para os sons surdos, quando podem chegar a ser percebidos como uma componente quase-periódica estranha à natureza do sinal [9]. Sabe-se também que tais efeitos podem ser significativamente atenuados pela adição de uma componente aleatória ao espectro de fase nas altas freqüências [9]. Embora este procedimento reduza a periodicidade do sinal, ele não afeta significativamente a sensação de tom porque o "pitch" é percebido preponderantemente a partir dos harmônicos mais baixos da freqüência fundamental.

Essa alteração do espectro de fase é efetuada no nível da excitação porque neste ponto se pode aproveitar o filtro de síntese na redução de efeitos transitórios causados pelo tratamento do sinal em blocos. Esta localização da compensação de espargimento permite que ela seja implementada através de um filtro $F(z)$ aplicado à excitação fixa como está representado na Figura 2. Isto ocorre porque os espectros de fase da excitação e do filtro se somam nesta disposição.

Assim, o filtro de compensação de espargimento $F(z)$ é um passa-tudo não-recorrente (FIR) projetado para ter um espectro de fase nulo até 3 kHz e uniformemente distribuído entre $-\pi/2$ e $\pi/2$ na faixa entre 3 e 4 kHz. A resposta impulsiva resultante está apresentada na Figura 1 depois do truncamento dentro dos limites do sub-bloco $0 \leq n < 60$.

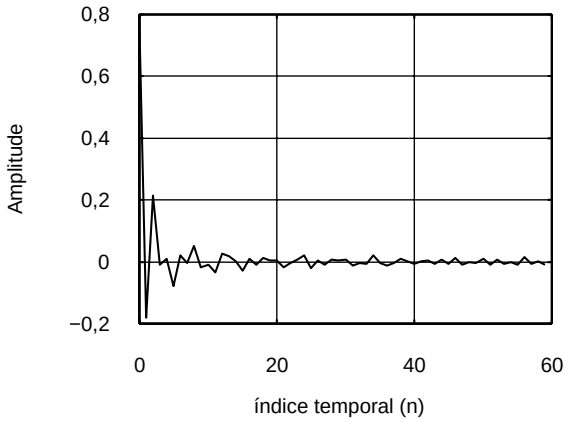

Figura 1: Resposta impulsiva projetada para compensar a fase da excitação esparsa.

3. TREINAMENTO DA COMPENSAÇÃO DE ESPARGIMENTO

A compensação de fase é um componente muito valioso num dispositivo destinado à compensação das conseqüências do espargimento da excitação. Entretanto,

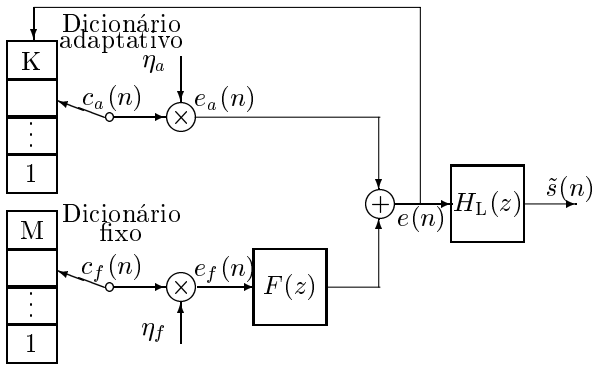

Figura 2: Síntese CELP com excitação compensada contra espargimento.

considerou-se que a eficácia da compensação poderia ser aumentada após um processo de treinamento. Portanto, elaborou-se um processo de treinamento que vai moldando gradativamento a resposta impulsiva do filtro de compensação, inicialmente identificada com a resposta impulsiva do filtro de compensação de fase projetado.

A excitação composta injetada no filtro de síntese do decodificador resulta da combinação

$$e(n) = \eta_f c_f(n) * f(n) + \eta_a c_a(n), \quad (1)$$

onde $c_f(n)$ e $c_a(n)$ representam os vetores-código fixo e adaptativo, respectivamente, cujos ganhos são η_f e η_a, e $f(n)$ é a resposta impulsiva do filtro de compensação.

No codificador, durante a fase de busca da excitação fixa, o vetor-alvo $u_f(n)$ é reconstruído como

$$\tilde{u}_f(n) = \eta_f c_f(n) * f(n) * h(n), \quad (2)$$

onde $h(n)$ é a resposta impulsiva do filtro de síntese ponderado.

Permutando-se a ordem das duas convoluções na Equação (2), obtém-se

$$\tilde{u}_f = \eta_f c_f(n) * h(n) * f(n), \quad (3)$$

que é mais conveniente de se usar na fase de treinamento, em que a resposta impulsiva $f(n)$ da compensação é incógnita. Além disso, notando-se que

$$q_u(n) = \eta_f c_f(n) * h(n) \quad (4)$$

é a reconstrução sem compensação do vetor-alvo, sua versão compensada na Equação (3) pode se expressar como

$$\tilde{u}_f = q_u(n) * f(n) \quad (5)$$

ou, com notação matricial, ela pode ser escrita como

$$\tilde{\mathbf{u}}_f = \mathbf{Q}\mathbf{f}. \quad (6)$$

A matriz $\mathbf{Q}$ na Equação (6) representa a operação de convolução com a reconstrução sem compensação $\mathbf{q}_u$. Portanto, ela tem uma estrutura Toeplitz triangular inferior com o vetor $\mathbf{q}_u$ na primeira coluna.

Assim, o vetor de erro ponderado no sub-bloco m é dado por

$$\begin{aligned} \mathbf{e}_m &= \mathbf{u}_{f,m} - \tilde{\mathbf{u}}_{f,m} \\ &= \mathbf{u}_{f,m} - \mathbf{Q}_m\mathbf{f} \end{aligned} \quad (7)$$

de forma tal que o erro quadrático total acumulado ao longo de toda a base de dados pode ser representado como

$$\begin{aligned} \sum_{m=1}^{N_S} \mathbf{e}_m^T \mathbf{e}_m &= \sum_{m=1}^{N_S} \left(\mathbf{u}_{f,m}^T - \mathbf{f}^T\mathbf{Q}_m^T\right)\left(\mathbf{u}_{f,m} - \mathbf{Q}_m\mathbf{f}\right) \\ &= \sum_{m=1}^{N_S} \left(\mathbf{u}_{f,m}^T\mathbf{u}_{f,m} - 2\mathbf{u}_f^T\mathbf{Q}_m\mathbf{f} + \mathbf{f}^T\mathbf{Q}_m^T\mathbf{Q}_m\mathbf{f}\right) \end{aligned} \quad (8)$$

Tabela 1: Desempenho objetivo de codificadores com compensação de fase projetada e treinada a 5,3 kbit/s com as buscas focalizada e conjunta da excitação fixa ACELP.

Iteração	Focalizada		Conjunta	
	SNRSEG (dB)	WSNRSEG (dB)	SNRSEG (dB)	WSNRSEG (dB)
Inicial	9,20	4,65	8,94	4,53
Final	9,26	4,73	8,99	4,59

onde N_S é o número total de sub-blocos contidos na base de dados.

A minimização do erro quadrático ocorre quando seu gradiente em relação à resposta impulsiva de compensação é nulo, isto é,

$$\begin{aligned} \mathbf{0} &= \nabla \sum_{m=1}^{N_S} \left(\mathbf{e}_m^T\mathbf{e}_m\right) = \sum_{m=1}^{N_S} \frac{\partial}{\partial \mathbf{f}} \left(\mathbf{e}_m^T\mathbf{e}_m\right) \\ &= -2\sum_{m=1}^{N_S} \mathbf{u}_f^T\mathbf{Q}_m + 2\sum_{m=1}^{N_S} \mathbf{Q}_m^T\mathbf{Q}_m\mathbf{f} \end{aligned} \quad (9)$$

Após uma reordenação da Equação (9), obtém-se o sistema de equações

$$\mathbf{A}\mathbf{f} = \mathbf{b}, \quad (10)$$

onde $\mathbf{A} = \sum_{m=1}^{N_S} \mathbf{Q}_m^T\mathbf{Q}_m$ e $\mathbf{b} = \sum_{m=1}^{N_S} \mathbf{u}_f^T\mathbf{Q}_m$. Este sistema acumulativo é resolvido na última etapa de cada iteração, fornecendo a nova resposta impulsiva de compensação.

Usou-se o codificador ACELP G.723.1 operando a 5,3 kbit/s [3] para treinar as respostas impulsivas de filtros de compensação de espargimento, tomando-se a resposta impulsiva de fase projetada, que está representada na Figura 1, como a primeira estimativa. Efetuaram-se dois treinamentos completos, usando-se num deles o algoritmo de busca focalizada [2] e no outro o algoritmo de busca conjunta de posição e amplitude [7, 8] para buscar a excitação no dicionário fixo ACELP. Todos os sinais contidos na partição de teste da base de dados TIMIT foram usados nos treinamentos, totalizando uma duração de 5187 s ou 691,6 mil sub-blocos. Os treinamentos foram executados com o critério de maximização da relação sinal-ruído segmentada (WSNRSEG) no nível do vetor-alvo, que é o procedimento seguido pelo codificador durante a busca da excitação. As medidas obtidas nas situações inicial e final estão na Tabela 1.

As respostas impulsivas das compensações de espargimento obtidas com o treinamento para a busca

focalizada e para a busca conjunta estão representadas nas Figuras 3 e 4, sendo que esta última é apresentada como um incremento sobre a primeira para ressaltar a diferença que existe entre ambas.

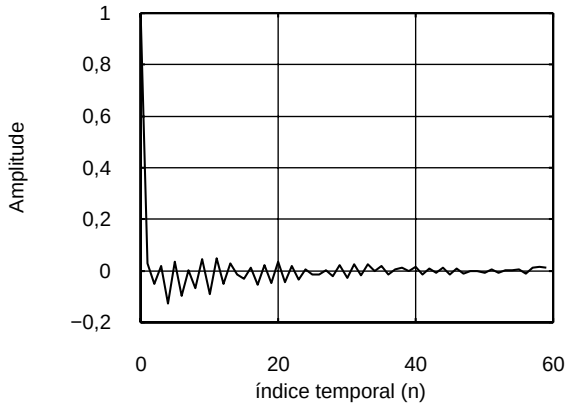

Figura 3: Resposta impulsiva da compensação treinada com a busca focalizada no codificador G.723.1 à taxa de 5,3 kbit/s.

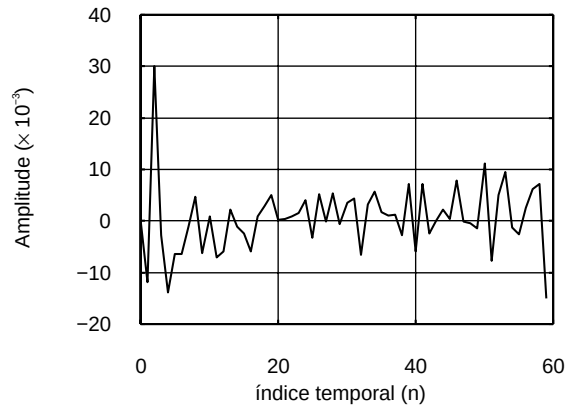

Figura 4: Resposta impulsiva da compensação treinada com a busca conjunta no codificador G.723.1 à taxa de 5,3 kbit/s, representada como um incremento sobre a resposta treinada com a busca focalizada.

4. COMPLEXIDADE E DESEMPENHO RESULTANTES

Executou-se um teste de avaliação auditiva subjetiva do codificador G.723.1 a 5,3 kbit/s com cada algoritmo de busca, focalizada e conjunta, usando-se cada um deles tanto na versão sem compensação quanto em duas versões compensadas, além de dois codificadores G.729 a 8 kbit/s usando cada um dos dois algoritmos de busca sem compensação. Ao todo, nove condições foram avaliadas por 12 ouvintes femininos e 12 masculinos. Usaram-se quatro pares de frases, que foram reproduzidas através de fones de ouvido durante as audições. Cada par de frases é composto por duas sentenças curtas proferidas em português por dois locutores femininos e dois masculinos, com duração aproximada de 3 s e com um intervalo entre elas de 0,5 s de silêncio. O teste subjetivo foi do tipo de categorização absoluta ("absolute category rating" - ACR) com uma escala de 5 pontos onde 5 é excelente e 1 é ruim [10]. Os índices médios de opinião (MOS) resultantes são mostrados na Tabela 2, evidenciando que o codificador com busca conjunta compensada a 5,3 kbit/s tem um desempenho muito próximo do codificador com busca focalizada a 8 kbit/s. Ao contrário, o codificador com busca focalizada compensada a 5,3 kbit/s não consegue atingir o nível de qualidade dos codificadores que operam a 8 kbit/s.

Tabela 2: Qualidade subjetiva de codificadores com buscas ACELP em aritmética de ponto fixo nas variedades desprovidas e providas de compensação de espargimento à taxa de 5,3 kbit/s e apenas nas variedades desprovidas de compensação à taxa de 8 kbit/s em testes subjetivos de categorização absoluta a partir de gravações feitas com ruído ambiente de escritório.

Condição	Ouvintes		
	Fem.	Masc.	Todos
Original	3,65	4,13	3,89
FOCS	3,52	3,58	3,55
FOCF	3,50	3,75	3,63
FOCT	3,52	3,63	3,57
JPAS	3,60	3,48	3,54
JPAF	3,54	3,60	3,57
JPAT	3,60	3,73	3,67
FOC8	3,60	3,79	3,70
JPA8	3,77	3,83	3,80

FOCS: Busca focalizada sem compensação a 5,3 kbit/s
FOCF: Busca focalizada compensada em fase
FOCT: Busca focalizada com compensação treinada
JPAS: Busca conjunta sem compensação a 5,3 kbit/s
JPAF: Busca conjunta compensada em fase
JPAT: Busca conjunta compensada por filtro treinado
FOC8: Busca focalizada sem compensação a 8 kbit/s
JPA8: Busca conjunta sem compensação a 8 kbit/s

Todos os codificadores utilizados foram implementados com aritmética de ponto fixo e suas complexidades operacionais foram medidas em milhões de operações ponderadas por segundo (WMOPS), tomadas nos piores casos. Conforme aparece na Tabela 3, a

Tabela 3: Medidas de complexidade operacional no pior caso de implementações em ponto fixo de dois algoritmos de busca ACELP nas versões sem compensação (normal) e compensada operando à taxa de 5,3 kbit/s com os sinais da partição de teste da base de dados TIMIT.

Busca Fixa	Função	Complexidade (WMOPS)	
		Normal	Compensada
Focalizada	Codif.	3,54	3,94
Conjunta	Codif.	1,24	1,59
Focalizada	Decod.	0,59	0,62
Conjunta	Decod.	0,55	0,59

complexidade do codificador pode aumentar mais do que um quarto para acomodar a compensação de espargimento. Mas parece haver condições de redução da complexidade porque as amplitudes das respostas impulsivas de compensação decaem rapidamente com o afastamento a partir da origem (Figuras 3 e 4), podendo muito possivelmente ser truncadas em analogia com as respostas mais breves usadas no filtro de conformação de pulsos em codificadores que operam por volta de 4 kbit/s [11].

5. CONCLUSÃO

Testaram-se vários codificadores de voz com dicionários fixos de quatro pulsos operando nas taxas de 5,3 e 8 kbit/s. Os codificadores à taxa mais baixa foram melhorados com dois tipos de filtros de compensação de espargimento: um projetado para tornar a fase mais aleatória nas altas freqüências e outro cuja resposta impulsiva foi refinada num processo de treinamento a partir da resposta impulsiva do primeiro. Ademais, dois algoritmos de busca da excitação fixa foram usados para cada condição: a busca focalizada de referência e a busca conjunta de posição e amplitude. Em testes subjetivos de audição, a compensação de espargimento treinada foi capaz de elevar a qualidade da busca conjunta à taxa de 5,3 kbit/s a um nível muito próximo ao da busca focalizada a 8 kbit/s ao passo que não pôde promover uma melhora significativa de qualidade da busca focalizada. Uma possível causa deste comportamento é o maior cuidado exercido pela busca conjunta na determinação dos sinais dos pulsos.

AGRADECIMENTOS

O autor agradece a colaboração de Fernando Kazuyoshi Takikawa nas medidas de complexidade e na preparação dos testes subjetivos.

REFERÊNCIAS

[1] W. B. Kleijn, D. J. Krasinski e R. H. Ketchum, "Fast methods for the CELP speech coding algorithm", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 8, pp. 1330-1342, Aug. 1990.

[2] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon e Y. Shoham, "Design and description of CS-ACELP: A toll quality 8 kb/s speech coder", *IEEE Trans. Speech Audio Processing*, vol. 6, no. 2, pp. 116-130, Mar. 1998.

[3] ITU-T, *Recommendation G.723.1 Dual rate speech coder for multimedia applications transmitting at 5.3 and 6.3 kbit/s,* ITU-T, Geneva, March 19, 1996.

[4] R, Salami, C. Laflamme, B. Bessette and J.-P. Adoul, "Description of ITU-T Recommendation G.729 Annex A: Reduced complexity 8 kbit/s CS-ACELP codec", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Munich, 1997, vol. 2, pp. 775-778.

[5] T. Honkanen, J. Vainio, K. Järvinen e P. Haavisto, "Enhanced full rate codec for IS-136 digital cellular system", in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Processing,* Munich, 1997, vol. 2, pp. 731-734.

[6] R. Salami, C. Laflamme, B. Bessette e J.-P. Adoul, "ITU-T G.729 Annex A: Reduced complexity 8 kb/s CS-ACELP codec for digital simultaneous voice and data", *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 56-63, Sept. 1997.

[7] M. Arjona Ramírez e M. Gerken, "A multistage search of algebraic CELP codebooks", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, 1999, vol. 1, pp. 17-20.

[8] M. Arjona Ramírez e M. Gerken, "Joint position and amplitude search of algebraic multipulses", *IEEE Trans. Speech Audio Processing*, vol. 8, no. 5, Sept. 2000 (a ser publicado).

[9] R. Hagen, E. Ekudden, B. Johansson e W. B. Kleijn, " Removal of sparse-excitation artifacts in CELP" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Seattle, 1998, vol. 1, pp. 145-148.

[10] P. Kroon, "Evaluation of speech coders", in W. B. Kleijn e K. K. Paliwal, Ed., *Speech Coding and Synthesis*, Amsterdam, Elsevier Science, 1995, pp. 467-494.

[11] T. Amada, K. Miseki e M. Akamine, "CELP speech coding based on an adaptive pulse position codebook", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, 1999, vol. 1, pp. 13-16.