DISPERSÃO SELETIVA DE PULSOS PARA 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

Os codificadores de voz para taxas médias e baixas, como o CELP algébrico (ACELP) e o vocoder LPC com excitação mista (MELP), usam excitações esparsas, conciliando uma boa reconstrução dos segmentos sonoros do sinal de voz com a codificação compacta da excitação. Entretanto, os segmentos surdos tendem a apresentar efeitos tônicos, que são compensados com algum tipo de dispersão da excitação impulsiva. Propõe-se a aplicação de filtros de dispersão impulsiva apenas aos sub-blocos surdos do sinal num codificador ACELP com busca conjunta de posição e amplitude (JPAS). Apresentam-se dois procedimentos para a obtenção da resposta de dispersão. O primeiro é um projeto baseado na distribuição do espectro de fase da excitação e o segundo é um processo de treinamento. O treinamento parte sempre de uma resposta de dispersão projetada. Mostra-se que o ganho de desempenho resultante do treinamento pode se situar próximo aos 2 dB quando a dispersão projetada inicial é de qualidade inferior, reduzindo-se a valores por volta de 0,5 dB quando a dispersão inicial já é mais eficaz. Entretanto, em ambas as situações os resultados finais são equivalentes.

1. INTRODUÇÃO

O modelo CELP, em particular sua vertente algébrica denominada ACELP, tem tido sucesso nas últimas disputas para o estabelecimento de padrões de codificação de voz para a telefonia digital celular e fixa e para as comunicações multimídia às taxas de 8 kbit/s [1], de 7,4 kbit/s [2] e de 5,3 kbit/s [3]. Inclusive, mais recentemente um conjunto de melhoras do modelo CELP básico, renomeado CELP estendido (eX-CELP) [4], está servindo de base para o candidato definitivo [5] ao padrão de 4 kbit/s da União Internacional de Telecomunicações – Telecommunication Union - Telecommunication Standardization Sector (ITU-T).

A evolução tecnológica registrada recentemente com os algoritmos eficientes de busca de sinais de inovações [6, 2, 7] influenciou decisivamente nessa difusão dos codificadores ACELP . Os algoritmos para efetuar essas buscas eficientes estão disponíveis na literatura técnica para a "busca conjunta de posição e amplitude" ("joint position and amplitude search" - JPAS) [8] e na literatura normativa para a busca fixa do "enhanced full rate (EFR) codec" [9] e para a "depth first tree search" (DFTS) [10].

Conforme se diminui a taxa de transmissão abaixo de 8 kbit/s, vão aparecendo efeitos indesejáveis causados pela natureza esparsa da excitação ACELP [11], que podem ser compensados através do treinamento da resposta de um filtro para a excitação fixa [12]. É interessante que a resposta impulsiva deste filtro de compensação de espargimento pode ser truncada a aproximadamente um terço do comprimento do sub-bloco [13].

Há razões para se supor que diferentes dispersões impulsivas, selecionadas de acordo com a natureza do segmento do sinal de voz, sejam mais eficientes do que uma única resposta de dispersão [12, 14, 15]. Neste trabalho explora-se a seleção da resposta de dispersão de acordo com duas classes de segmentos de voz, sonoros ou surdos. No caso da primeira classe, a resposta de dispersão usada é puramente impulsiva, isto é, não há dispersão, enquanto para a segunda classe usam-se dispersões treinadas ou projetadas.

Há outros modelos de codificadores de voz que também empregam dispersões impulsivas, como a predição linear com excitação mista ("mixed excitation linear prediction") – MELP [16, 17], que inclui um filtro para dispersão dos pulsos, dentre vários outros processamentos para melhora do sinal de excitação, para atingir uma boa qualidade de voz a 2,4 kbit/s.

2. COMPENSAÇÃO DA EXCITAÇÃO ESPARSA

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. Decorre desta observação a motivação para o treinamento restrito aos segmentos surdos da compensação de espargimento.

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 [12] 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 [11]. Sabe-se também que tais efeitos podem ser significativamente atenuados pela adição de uma componente aleatória ao espectro de fase nas altas freqüências [11]. 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.

Assim, faz sentido que os pulsos da excitação esparsa sejam dispersados apenas quando o sinal de voz for surdo, sendo esta hipótese fundamental a base deste trabalho. Então, previamente à aplicação da dispersão impulsiva, é necessário determinar a natureza surda ou sonora do segmento de voz em questão. Neste aspecto, há métodos que se baseiam em parâmetros básicos do codificador para tomar esta decisão, como o ganho do dicionário adaptativo em codificadores CELP [11, 18]. Em seguida será vista outra alternativa.

3. CLASSIFICAÇÃO SONORO-SURDA DOS SEGMENTOS DO SINAL DE VOZ

No caso de um codificador CELP, a geração da excitação composta está ilustrada na Figura 1 após a determinação do atraso P ou índice do dicionário adaptativo. Desta forma a excitação adaptativa é dada por

$$e_a(n) = \eta_a \left(e_a(n-P) + e_f(n-P) \right) \tag{1}$$

em função da excitação fixa $e_f(n)$.

Na situação em que a excitação fixa ainda não foi determinada, assume-se que ela seja nula, isto é, $e_f(n) \equiv 0$, definindo-se a excitação adaptativa de forma recorrente como

$$e_a(n) = \eta_a e_a(n-P). \tag{2}$$

Neste ponto, pode-se estimar o ganho η_a através de uma predição no domínio da excitação como

$$e_a(n) = \eta_a e_a(n-P) + d(n), \tag{3}$$

em que foi introduzido o sinal de erro $d(n)$, cujo valor quadrático médio $\varepsilon = \mathrm{E}\left[e^2(n)\right]$ será minimizado.

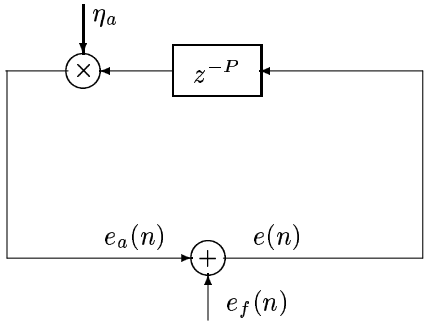

Figura 1: Geração da excitação composta $e(n)$ a partir das excitações fixa $e_f(n)$ e adaptativa $e_a(n)$.

O erro quadrático mínimo $\varepsilon_{\min}$ é obtido para

$$\eta_a = \frac{\mathrm{cov}\left(e_a(n), e_a(n-P)\right)}{\mathrm{var}\left(e_a(n-P)\right)} \tag{4}$$

com valor igual a

$$\varepsilon_{\min} = \left(1-\rho^2\right) \mathrm{var}\left(e_a(n)\right), \tag{5}$$

onde o coeficiente de correlação é

$$\rho = \frac{\mathrm{cov}\left(e_a(n), e_a(n-P)\right)}{\sigma\left(e_a(n)\right) \cdot \sigma\left(e_a(n-P)\right)}, \tag{6}$$

assumindo valores em $-1 \leq \rho \leq 1$.

Para a utilização do coeficiente de correlação da excitação adaptativa na classificação sonoro-surda do sinal de voz, deve-se notar que valores altos em módulo estão mais associados a segmentos sonoros, que têm periodicidade maior. Resta, então estabelecer um nível limiar para a decisão sonoro-surda. Como a Recomendação G.723 já inclui um classificador [3], que é usado na regeneração do sinal em caso de "apagamento de bloco" ("frame erasure"), pode-se usar o mesmo nível de limiar desse regenerador, que declara o bloco de voz sonoro quando

$$\rho^2 > \frac{1}{8}. \tag{7}$$

Um detalhe de implementação que deve ser levado em conta diz respeito à periodicidade da classificação sonoro-surda, que, enquanto é suficiente que ocorra a cada bloco para a regeneração, deve ocorrer a cada sub-bloco para a dispersão impulsiva adaptativa.

4. TREINAMENTO SELETIVO DA DISPERSÃO

Usou-se o codificador ACELP G.723.1 operando à taxa de 5,3 kbit/s [3] para treinar as dispersões impulsivas da excitação fixa nos sub-blocos surdos a partir da dispersão impulsiva projetada com fase distribuída entre 0 e $\pi/2$ radianos na

faixa de 1,5 a 4 kHz [13], que está representada na Figura 2. A classificação sonoro-surda dos sub-blocos de voz foi efetuada de acordo com o procedimento descrito na Seção 3. A busca conjunta de posição e amplitude (JPAS) [6], implementada através de um algoritmo eficiente [8], foi usada 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 para três comprimentos da dispersão impulsiva, que foram obtidos pelo simples truncamento da dispersão projetada da Figura 2. Esse truncamento é a melhor forma de obtenção dessas respostas mais curtas na ausência de treinamento [13].

Nota-se que o ganho de treinamento excede 1,8 dB nos três casos, que, comparado com valores por volta de 0,1 dB obtidos quando a dispersão impulsiva é aplicada de forma irrestrita [12], atestam a importância da classificação para a aplicação seletiva da dispersão. Também, o treinamento é mais eficiente em aproximadamente 0,1 dB no caso da dispersão mais curta com 10 amostras de comprimento. Portanto, embora uma tendência de melhor desempenho das dispersões mais curtas já houvesse sido observada nos treinamentos não-seletivos [13], o acréscimo obtido com o treinamento seletivo da dispersão impulsiva restrito aos sub-blocos surdos foi ampliado de duas a três vezes em relação ao treinamento irrestrito.

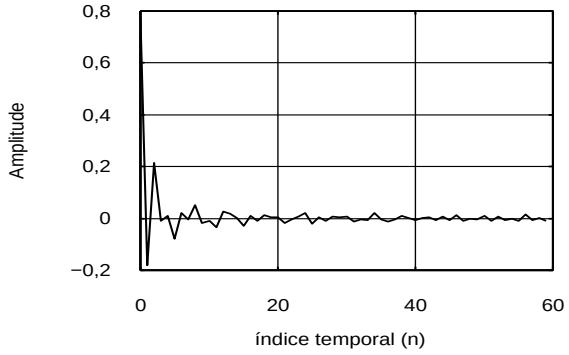

Figura 2: Dispersão impulsiva projetada com distribuição de fase entre 0 e $\pi/2$.

A dispersão impulsiva treinada que teve melhor desempenho foi a de comprimento 10, que se encontra ilustrada na Figura 3.

Para avaliar a sensibilidade em relação à condição inicial e o seu impacto sobre o ganho de treinamento, fez-se um novo treinamento a partir de uma dispersão impulsiva

Tabela 1: Desempenhos objetivos das dispersões impulsivas truncadas treinadas seletivamente a partir da distribuição de fase de 0 a $\pi/4$ sobre sub-blocos surdos em codificadores a 5,3 kbit/s com a busca conjunta do dicionário fixo ACELP.

Comprimento da Dispersão	Iteração	SNRSEG (dB)	WSNRSEG (dB)
10	Inicial	7,02	4,48
	Final	8,92	4,63
20	Inicial	6,99	4,45
	Final	8,84	4,62
60	Inicial	6,96	4,43
	Final	8,78	4,59

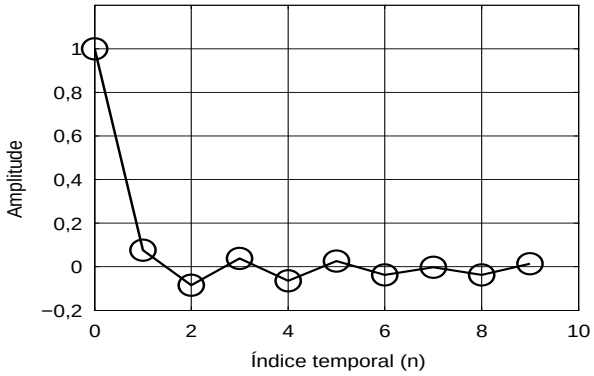

Figura 3: Dispersão impulsiva treinada a partir da distribuição com fase entre 0 e $\pi/2$.

projetada com distribuição uniforme de fase entre $-\pi$ e π radianos na faixa de freqüências de 3 a 4 kHz e com fase nula na faixa de freqüências inferior, que se encontra representada na Figura 4.

Nota-se da Tabela 2 que os desempenhos iniciais das dispersões projetadas são mais altos em aproximadamente 1,3 dB no caso destas dispersões com distribuição de fase de maior amplitude sobre uma faixa de freqüências mais alta. Entretanto, os ganhos de predição caíram ao nível de 0,5 dB, atingindo desempenhos objetivos praticamente equivalentes aos anteriores, podendo situar-se ligeiramente abaixo ou acima nas situações finais.

O codificador utilizado foi implementado em aritmética de ponto fixo e as complexidades operacionais de suas versões com diferentes dispersões impulsivas foram medidas em milhões de operações ponderadas por segundo, tomadas nos piores casos (WMOPS). Os resultados aparecem na Tabela 3, que também inclui as medidas realizadas com o codificador operando sem a seleção dos sub-blocos para aplicação da dispersão [13]. Nota-se que a aplicação da seleção sonoro-surda causa um acréscimo menor que 20% na

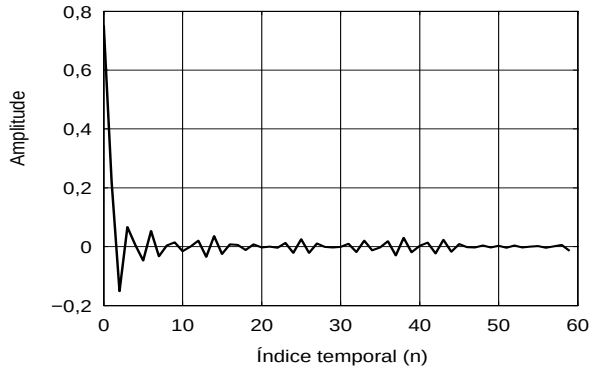

Figura 4: Dispersão impulsiva projetada com distribuição de fase entre $-\pi$ e π.

Tabela 2: Desempenhos objetivos das dispersões impulsivas truncadas treinadas seletivamente a partir da distribuição de fase de $-\pi$ a π sobre sub-blocos surdos em codificadores a 5,3 kbit/s com a busca conjunta do dicionário fixo ACELP.

Comprimento da Dispersão	Iteração	SNRSEG (dB)	WSNRSEG (dB)
10	Inicial	8,37	4,63
	Final	8,87	4,65
20	Inicial	8,32	4,59
	Final	8,77	4,62
60	Inicial	8,29	4,57
	Final	8,72	4,61

complexidade operacional do codificador.

5. CONCLUSÃO

Apresentaram-se as aplicações das excitações esparsas nos codificadores de voz mais recentes com considerações sobre suas deficiências na modelagem dos segmentos surdos do sinal de voz. Conseqüentemente, aplicaram-se dispersões impulsivas apenas aos sub-blocos surdos da excitação, selecionados com um classificador sonoro-surdo conveniente. Essas dispersões impulsivas foram projetadas com base na distribuição do espectro de fase da excitação ou obtidas por um processo de treinamento. O treinamento parte sempre de uma resposta de dispersão projetada. Constatou-se que o ganho de treinamento pode se aproximar de 2 dB quando a dispersão inicial tem desempenho inferior, superando por larga margem os ganhos próximos a 0,1 dB que são obtidos quando a dispersão é aplicada de forma irrestrita. Por outro lado, quando a dispersão inicial já é de alto desempenho, o ganho de treinamento reduz-se a valores por volta de 0,5 dB. Entretanto, os desempenhos obtidos na iteração final do treinamento são equivalentes em ambos os casos.

Tabela 3: Medidas de complexidade operacional no pior caso de implementações em ponto fixo do algoritmo de busca JPAS da excitação com dispersão impulsiva treinada e truncada operando à taxa de 5,3 kbit/s com os sinais da partição de teste da base de dados TIMIT. Fornecem-se as complexidades dos casos irrestrito e seletivo sobre sub-blocos surdos de dispersão.

Comprimento da dispersão	Complexidade com seleção (WMOPS)	Complexidade sem seleção (WMOPS)
10	1,68	1,42
20	1,74	1,48
60	1,85	1,59

6. REFERÊNCIAS

[1] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and 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.

[2] T. Honkanen, J. Vainio, K. Järvinen, and 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.

[3] "*Dual rate speech coder for multimedia applications transmitting at 5.3 and 6.3 kbit/s*," ITU-T Recommend. G.723.1, Mar. 1996.

[4] Y. Gao, A. Benyassine, J. Thyssen, H. Su, and E. Shlomot, "eX-CELP: A speech coding paradigm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, 2001, a ser publicado.

[5] J. Thyssen, Y. Gao, A. Benyassine, E. Shlomot, C. Murgia, H. Su, K. Mano, Y. Hiwasaki, H. Ehara, K. Tasunaga, C. Lamblin, B. Kovesi, J. Stegmann, and H. Kang, "A candidate for the ITU-T 4 kbit/s speech coding standard," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, 2001, a ser publicado.

[6] M. Arjona Ramírez, "A low-complexity search algorithm for speech coders with sparse excitation," *Revista da Sociedade Brasileira de Telecomunicações*, a ser publicado.

[7] 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.

[8] M. Arjona Ramírez and M. Gerken, "Joint position and amplitude search of algebraic multipulses," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 5, pp. 633–637, Sept. 2000.

[9] "*IS-641-A TDMA Cellular/PCS - Radio Interface Enhanced Full-Rate Voice Codec, Revision A*," TIA/EIA TR45, Sep. 1997.

[10] "*Reduced complexity 8 kbit/s using CS-ACELP speech codec*," ITU-T Recommend. G.729 Annex A, Nov. 1996.

[11] R. Hagen, E. Ekudden, B. Johansson, and 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.

[12] M. Arjona Ramírez, "Compensação de espargimento em codificadores de voz," in *Anais do Simpósio Brasileiro de Telecomunicações*, Gramado, 2000, CD-ROM.

[13] M. Arjona Ramírez, "Treinamento da compensação de excitações de voz esparsas," *Revista da Sociedade Brasileira de Telecomunicações*, a ser publicado.

[14] K. Ozawa, "4 kb/s multi-pulse based CELP speech coding using excitation switching," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, 1999, vol. 1, pp. 189–192.

[15] K. Yasunaga, H. Ehara, K. Yoshida, and T. Morii, "Dispersed-pulse codebook and its application to a 4 kb/s speech coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, 2000, vol. 3, pp. 1503–1506.

[16] A. McCree and T. P. Barnwell *III*, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 242–250, Jul. 1995.

[17] A. McCree, K. Truong, E. B. George, T. P. Barnwell, and V. Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new U. S. Federal Standard," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, 1996, vol. 1, pp. 200–203.

[18] T. Amada, K. Miseki, and 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.