Modelagem e Síntese de Fricativos e Oclusivos em Codificadores Multibandas a 1,75 kb/s

Rodrigo C. de Lamare e Abraham Alcaim
CETUC - PUC-RIO, 22453-900, Rio de Janeiro - Brasil
e-mails: delamare@infolink.com.br e alcaim@cetuc.puc-rio.br

Resumo

Nesse artigo é investigado o uso de técnicas de modelagem e síntese de sons fricativos e oclusivos em codificadores que utilizam excitação mista sonoro-surdo em multibandas. Um codificador com excitação mista sonoro-surdo em multibandas operando a 1,75 kb/s é empregado como plataforma de comparação entre a técnica de excitação de sons fricativos e oclusivos e a abordagem de excitação ruidosa. A análise é realizada por meio de testes de comparação do tipo A/B. Os resultados mostram que o uso da modelagem e síntese de fricativos e oclusivos apresenta qualidade de voz superior ao da excitação ruidosa.

1 Introdução

Com o advento da telefonia celular digital e aplicações como redes de voz baseadas no protocolo IP (VOIP), algoritmos de codificação de voz a baixas taxas de bits tiveram um considerável aumento de importância. Uma grande parte dos algoritmos de codificação de voz a baixas taxas de bits como a excitação mista em multibandas (EMM) [1,2] e o codificador de predição linear com excitação mista (MELP) [3,4], são baseados em codificação de predição linear (LPC), onde um sinal de excitação é aplicado a um filtro de apenas pólos, que representa a informação da envoltória espectral da voz. Apesar desses algoritmos de compressão produzirem uma qualidade de voz decodificada bastante boa, eles não são adequados para codificar alguns sons não estacionários como os fricativos e oclusivos surdos. Por este motivo, neste trabalho são investigadas técnicas de modelagem e síntese que melhoram a codificação deste sons. Além disso, é realizada uma comparação desta abordagem com o método tradicional de excitação ruidosa, por meio de testes de avaliação subjetivos.

Esse artigo é organizado da seguinte forma. A Seção 2 descreve as técnicas de modelagem e síntese de sons fricativos e oclusivos. A Seção 3 apresenta o codificador multibandas que será usado como plataforma de comparação para os testes. A Seção 4 é dedicada aos testes de avaliação e à discussão dos resultados, enquanto na Seção 5 são apresentadas as conclusões deste trabalho

2 Codificação de Sons Fricativos e Oclusivos

A excitação mista permite aos codificadores multibandas (EMM) uma flexibilidade significativa na decisão sonoro ou surdo. No entanto, a excitação mista de ruído e de um trem de pulsos não é capaz de reproduzir sinais específicos como aqueles encontrados nos sons oclusivos e fricativos. Para produzir uma melhor qualidade de voz nas sentenças contendo sons fricativos e oclusivos, é usada uma estratégia baseada nos algoritmos introduzidos por Unno *et al.* [5] e Ehnert [6]. Estes algoritmos envolvem a detecção e a modelagem e síntese destes sinais.

2.1 Detecção de Sons Fricativos e Oclusivos

Para a detecção de sons oclusivos é usado o valor de pico do sinal residual LPC $r(n)$ e uma janela deslizante é empregada a fim de localizar a posição do quadro que maximiza o valor de pico [5]. O valor de pico com a janela deslizante é dado por

$$P = max_{i=-T_s}^{i=T_s} P_i \tag{1}$$

$$P_i = \frac{\frac{1}{N}\sum_{n=0}^{N-1} r(n+i)^2}{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} |r(n+i)|}} \tag{2}$$

onde N é o comprimento do quadro de voz e T_s é o valor máximo do deslizamento que também é usado em (1). Além da medida de pico, a energia do sinal passa-baixa é calculada e usada para distinguir os rápidos ataques de vogais dos sons oclusivos. Na abordagem adotada neste trabalho existem dois tipos de sinais oclusivos, já que duas entradas no dicionário de excitações são reservadas para estes sons. O primeiro corresponde aos sinais cujas amplitudes máximas são encontradas na primeira metade dos quadros de voz, enquanto o segundo é associado àqueles cujas amplitudes máximas são localizadas na segunda parte dos quadros.

A detecção de sons fricativos é baseada no uso de limiares apropriados para o número de cruzamentos por zero e a energia de cada quadro. Em geral, estes sinais de baixa energia apresentam entre 60 e 140 cruzamentos por zero, enquanto os quadros sonoros típicos não cruzam o eixo mais de 60 vezes por quadro [6]. Um limiar de energia também é empregado para distinguir os sons fricativos dos quadros em silêncio. Note que apenas os sinais fricativos e oclusivos surdos efetivamente necessitam deste modelo. O algoritmo de detecção do período fundamental separa os fricativos e oclusivos surdos dos sonoros.

2.2 Modelagem e Síntese de Fricativos e Oclusivos

Técnicas de excitação multipulso são capazes de produzir sons fricativos e oclusivos sintetizados com alta qualidade, ainda que requeiram uma taxa de bits relativamente alta. No entanto, para aplicações a baixas taxas de bits um modelo alternativo que permite uma qualidade de voz perceptivamente satisfatória para esses sons é mais adequado do que o ruído. Nesse modelo, os sinais fricativos e oclusivos $f|s(n)$ são produzidos através da filtragem LPC de protótipos pré-armazenados de sinais residuais $r(n)$ e dos coeficientes LPC transmitidos:

$$f|s(n) = Gr(n) + \sum_{i=1}^{p} a_i f|s(n-i) \qquad (3)$$

onde G é o ganho baseado na energia do sinal fricativo ou oclusivo original e $a_1, ..., a_p$ são os coeficientes LPC transmitidos ao decodificador. Os protótipos são cuidadosamente escolhidos para que consigam representar de forma adequada a grande maioria dos sons modelados. Foi usado um sinal residual como protótipo para sintetizar os sons fricativos, enquanto dois sinais residuais são empregados para reproduzir os oclusivos.

3 A Plataforma de Codificação Multibandas

Codificadores de voz a baixas taxas que seguem o modelo clássico de Atal e Hanauer [7] geralmente resultam em uma qualidade de voz sintética devido à uma deficiência que normalmente resulta em chiado. A excitação mista em multibandas (EMM) [1,2] ataca o problema do chiado diretamente, através da divisão do sinal de voz em diversas bandas de freqüências. Essas bandas de freqüências são avaliadas por um detector sonoro ou surdo individualmente, com uma excitação periódica ou uma excitação ruidosa sendo selecionada para cada subbanda do quadro de voz. O diagrama em blocos do codificador e do decodificador é mostrado na Figura 1.

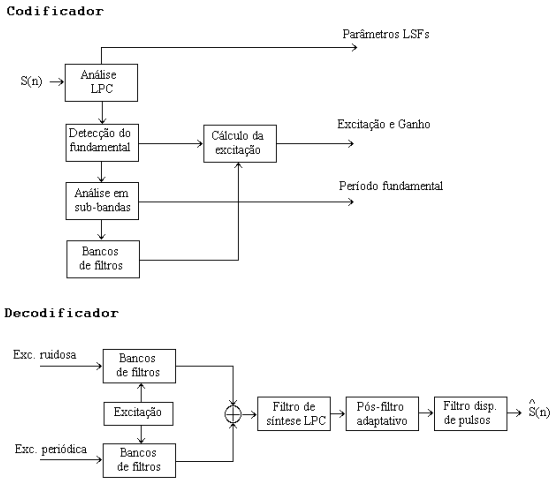

Figura 1: Diagrama em blocos do codificador e decodificador.

3.1 Codificador

De acordo com o esquema do codificador, na Figura 1, realiza-se uma análise de predição linear a cada quadro de 20 ms. Um algoritmo de detecção de período fundamental similar àquele empregado no codificador MELP [3,4] é usado para determinar alguma evidência de quadros sonoros. Os coeficientes LPC são transformados em parâmetros LSF e codificados com 21 bits por quadro por uma estrutura de quantização vetorial preditiva chaveada (QVPC)[8]. Neste esquema, emprega-se um QV preditivo e um QV sem memória. Uma busca de ambos esquemas

é realizada para cada quadro e o melhor candidato, no que diz respeito a um critério de distorção, é codificado e transmitido. No codificador empregado neste trabalho foram usados esquemas de QV multiestágios com busca em árvore com parâmetro de busca $M = 12$ e 4 estágios. O ganho é quantizado uniformemente com 5 bits por quadro e a excitação é codificada com 3 bits por quadro. Os quadro de voz classificados como sonoros são divididos em 3 bandas de freqüências, que são implementadas com bancos de filtros fixos, e uma análise em sub-bandas é realizada a fim de determinar se as bandas são sonoras ou surdas. Para quadros surdos, é empregada a técnica de modelagem e síntese de fricativos e oclusivos, descrita na Seção 2. A alocação de bits do codificador proposto é mostrada na Tabela 1.

Tabela 1: Alocação de Bits

Parâmetros	**Sonoro**	**Surdo**
LSFs	21	21
Ganho	5	5
Excitação	3	3
Fundamental	6	0
Outros	0	6
Total bits/20 ms	35	35
Bit rate	1,75 kb/s	1,75 kb/s

3.2 Decodificador

No decodificador, os sinais de excitação para os quadros sonoros são filtrados por um par de bancos de filtros. Para a reprodução destes quadros de voz, a excitação mista é gerada como a soma das excitações periódica e ruidosa filtradas. Para os quadros surdos, a excitação é declarada totalmente surda e o sinal de excitação não é aplicado ao banco de filtros. Em seguida, aplica-se o sinal de excitação ao filtro de síntese LPC com os coeficientes correspondentes às LSFs interpoladas e o ganho decodificado ao sinal de voz sintetizado. Para reduzir o ruído de codificação e melhorar a qualidade da voz decodificada, é empregado o conhecido filtro de melhoria espectral adaptável (MEA) [9], que possui a seguinte função de transferência:

$$H_{MEA} = \frac{A(z/\alpha)}{A(z/\beta)}(1 - \nu z^{-1}) \qquad (4)$$

onde $A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$ é o filtro de síntese inverso e a_i é o conjunto de coeficientes LPC. Os valores apropriados para α, β e ν a baixas taxas de bits são 0.5, 0.8 e $0.4k_1$, respectivamente, onde k_1 é o primeiro coeficiente de reflexão do modelo de predição linear [9]. Note que esse filtro é seguido de um filtro fixo de dispersão de pulso (DP), baseado em um pulso triangular com espectro aplainado, que espalha a energia da excitação dentro do perído fundamental, reduzindo a aspereza da voz sintética.

Para ilustrar a melhoria alcançada com a modelagem e síntese de oclusivos em codificadores a baixas taxas de bits, a Figura 2 mostra um exemplo da reprodução de sons oclusivos da frase "O Atabaque do Tito é coberto com pele de gato". O sinal original (a) contém alguns sinais oclusivos. No sinal sintetizado (b) uma excitação ruidosa é aplicada ao quadro de voz associado com o som oclusivo e degrada a qualidade da voz. Na Figura 2 (c) o sinal de voz sintetizado com a abordagem descrita na Seção 2 permite uma melhor reprodução dos sons oclusivos, melhorando a qualidade da voz.

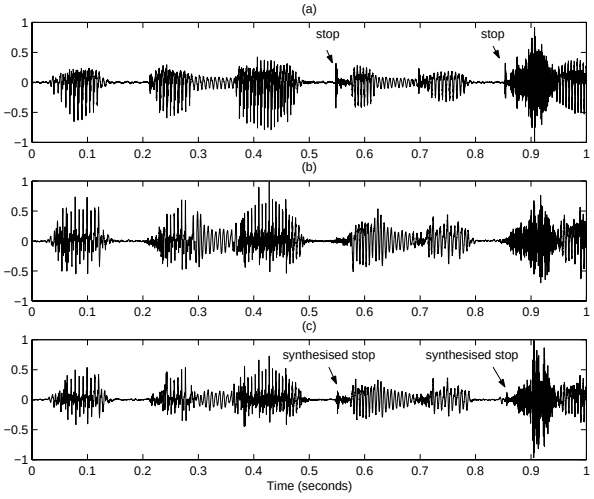

Figura 2: Reprodução de sinais oclusivos. (a) Sinal de voz original. (b) Sinal sintetizado com excitação ruidosa. (c) Sinal sintetizado com modelagem e síntese de oclusivos.

4 Resultados dos Testes de Avaliação Subjetiva

Para avaliar a qualidade da voz sintetizada, foi realizado um teste de comparação do tipo A/B com dez pares de sentenças, onde cada sentença foi gravada por um locutor diferente. Cinco locutores femininos e cinco masculinos foram usados no experimento. O material de teste incluiu apenas voz limpa e foi apresentado a 20 ouvintes, que escolhiam ou a primeira sentença (correspondente a um dos casos avaliados)

como de melhor qualidade, ou a segunda sentença, ou consideravam as duas como de qualidade comparável. Como cada par de sentenças foi também apresentado aos avaliadores com a ordem invertida, o teste inclui um total de 400 opiniões. O codificador descrito na Seção 3 foi usado como plataforma de testes para comparar a modelagem e síntese de sons oclusivos e fricativos com a excitação ruidosa empregada na maioria dos codificadores a baixas taxas. Os resultados mostrados na Tabela 2 revelam que 40% dos ouvintes não mostraram uma preferência clara, 36% deles preferiram a abordagem de fricativos e oclusivos, enquanto 24% acharam a excitação ruidosa superior. De fato, o método de modelagem e síntese de sons oclusivos e fricativos apresentou um desempenho superior ao da tradicional excitação ruidosa, aumentando a qualidade da voz decodificada.

Tabela 2: Comparação A/B.

Resultados	%
Mod. e Sínt. de Fricativos e Oclusivos	36
Qualidade Comparável)	40
Excitação Ruidosa	24

5 Conclusões

Neste trabalho foram investigadas técnicas de modelagem e síntese de sons fricativos e oclusivos. Um codificador multibandas operando a 1,75 kb/s foi descrito e usado como plataforma de análise das técnicas de excitação examinadas. Foi realizada uma análise comparativa da técnica de excitação de fricativos e oclusivos e da tradicional excitação ruidosa, usada em grande parte dos codificadores a baixas taxas de bits, por meio de testes de comparação do tipo A/B. Os resultados mostram que a abordagem proposta apresenta uma qualidade de voz decodificada superior à da excitação ruidosa, representando, portanto, uma abordagem promissora para codificação desses sons a baixas taxas de bits.

Agradecimentos

Gostaríamos de agradecer a Joaquim Pedro Cordeiro, pela aplicação dos testes de avaliação subjetiva.

Referências

[1] D. W. Griffin, J. S. Lim, "Multiband excitation vocoder", *IEEE Trans. on Acoustics Speech and Signal Processing*, pp. 1223-1235, August 1988.

[2] K. A. Teague, B. Leach and W. Andrews, "Development of a high-quality MBE based vocoder for implementation at 2400 bps", *Proc. IEEE Wichita Conf. Communications, Networking and Signal Processing*, pp. 129-133, April 1994.

[3] A. McCree, T. P. Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding", *IEEE Trans. Speech and Audio Processing*, pp. 242-250, 1995.

[4] L. M. Supplee, R. P. Cohn, J. S. Collura e A. V. McCree, "MELP: The New Federal Standard at 2400 bps", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1591-1594, 1997.

[5] T. Unno, T. P.Barnwell III, K. Truong, "An Improved Mixed Excitation Linear Prediction (MELP) Coder", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, USA, 1999

[6] W. Ehnert, "Variable-rate speech coding: coding unvoiced frames with 400bps", *Proc. EUSIPCO'98*, Rhodes, Greece, pp. 1437-1440, 1998.

[7] B. Atal, S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", *The Journal of the Acoustical Society of America*, 1971, vol. 50, no. 2, pp. 637-655.

[8] T. Eriksson, J. Lindén and J. Skoglund, "Exploiting Interframe Correlation in Spectral Quantization: A Study of Different Memory VQ Schemes", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996.

[9] J. Chen e A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech", *IEEE Trans. Speech and Audio Processing*, 1995, vol. 3, pp. 59-71.