Avaliação Subjetiva de Versão Acelerada do Codificador de Voz MELP

Marcelo M. Ventura e Sergio L. Netto

***Resumo*— Este artigo apresenta os resultados de testes subjetivos realizados com o objetivo de avaliar a qualidade de sinais de voz submetidos a um vocoder MELP modificado. As modificações foram feitas na rotina de cálculo do pitch, particularmente na primeira estimativa desse parâmetro, e visavam a aceleração do codificador. Os resultados obtidos mostraram que as alterações propostas reduziram o tempo de execução em aproximadamente 13%, sem perda perceptível na qualidade do sinal.**

***Palavras-Chave*— codificador de voz, melp, avaliação subjetiva, aceleração.**

***Abstract*— This article presents the results obtained with subjective test evaluation of a modified MELP vocoder. The modification was aimed at acceleration, modifying the pitch calculation, more specifically in the first estimate of this parameter. The results have shown that the proposed modification reduced the execution time, in about 13%, without perceptible sound quality reduction.**

***Keywords*— vocoder, melp, subjective evaluation, acceleration.**

I. Introdução

O codificador MELP (*mixed excitation linear prediction*) surgiu na década de 90, tendo como embrião as idéias de Allan McCree. Posteriormente, em 1999, os EUA normatizaram o codificador na MIL-STD-3005. Em 2001, a OTAN estabeleceu a norma STANAG 4591 (Standardization Agreement), baseada no padrão DoD MIL-STD 3005, utilizando as mesmas técnicas de quantização e alocação de bits, e portanto, totalmente compatível com a MIL-STD 3005 [1][2][3].

Essencialmente, o MELP é um codificador de voz paramétrico, que trabalha a uma taxa de 2400 bps. Sua versão mais moderna oferece também as opções de 1200 e 600 bps, obtidas basicamente por meio de quantização vetorial de blocos de parâmetros. O grande mérito do MELP foi superar as limitações do antigo codificador de baixa taxa LPC10 (estabelecido na norma americana *Federal Standard* 1015), que também opera a 2400 bps, conferindo melhor qualidade de audio e maior robustez ao ruído ambiente [4][5]. Essa melhoria de qualidade foi possível graças ao uso de um modelo mais sofisticado de produção da voz, ao custo de uma maior complexidade computacional.

Este artigo trata da avaliação subjetiva de uma versão modificada do codificador de voz MELP, em que se buscou a aceleração do algoritmo com um mínimo de impacto na qualidade do sinal reconstruído. Este trabalho foi baseado em uma versão do MELP de 1996, obtida de [6], e por questão de disponibilidade, toma essa versão como parâmetro comparativo para os melhoramentos propostos neste estudo.

Marcelo M. Ventura e Sergio L. Netto¸ PEE-COPPE, Universidade Federal do Rio de Janeiro, RJ, 21941-972, Brasil, E-mails: ventura.mm@gmail.com, sergioln@lps.ufrj.br.

Os testes subjetivos realizados se basearam na recomendação ITU-T P.800 [7], mais especificamente nos tipos de testes *Absolute Category Rating (ACR)* e *Comparison Category Rating* (CCR). O primeiro deles objetiva estabelecer uma nota absoluta para a qualidade do audio, de acordo com uma escala de referência. Já o segundo fornece uma medição relativa por meio da comparação direta entre a versão original e a modificada.

O presente artigo está elaborado da seguinte forma: na Seção II o esquema geral do codificador MELP é visto de forma breve, com ênfase na etapa de cálculo de pitch. Em seguida, a Seção III descreve algumas modificações propostas para o MELP que resultaram em uma família de versões alteradas e aceleradas do codificador. Por fim, a Seção IV apresenta o processo de seleção de uma versão acelerada específica, cuja avaliação subjetiva é descrita na Seção V.

II. Descrição geral do algoritmo MELP

A figura 1 apresenta o diagrama em blocos do codificador de voz MELP. Desta figura, pode ser observado que o MELP mantém alguma similaridade com o LPC10, mas introduz algumas características novas. Essas características, representadas pelos blocos em cinza na figura 1, permitem obter uma voz sintetizada mais natural.

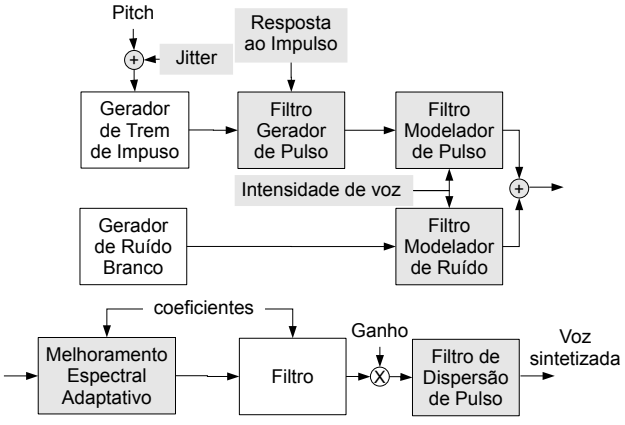

Fig. 1. Diagrama de blocos do decodificador MELP

Neste sentido, os principais melhoramentos introduzidos pelo MELP foram:

- mistura de excitações: o codificador permite combinar excitações periódicas e ruidosas, associadas respectivamente às classificações vozeadas e não vozeadas do LPC;
- pulsos aperiódicos: essa melhoria permite a geração de trens de impulso aperiódicos, de forma a modelar as

transições (vozeado / não vozeado) e variações de pitch, caracterizando um terceiro tipo de excitação: vozeado com *jitter*;
- modelamento pela "magnitude"de Fourier: tendo em vista que as excitações reais não são trens de impulso ideais e suas formas reais possuem informações relevantes sobre o sinal de voz, é possível realizar uma filtragem que resulta em um sinal de excitação mais próximo do original. Esse procedimento de modelagem de pulso é baseado no cálculo das magnitudes de Fourier do sinal de erro de predição;
- melhoramento espectral adaptativo: aumenta a qualidade perceptual do sinal sintetizado, ressaltando as características espectrais originais baseadas nos coeficientes de predição linear; e
- dispersão de pulso: torna o sinal de voz sintetizado mais natural.

A. Cálculo do pitch

Um diagrama em blocos do cálculo do pitch é apresentado na figura 2 e explicado a seguir.

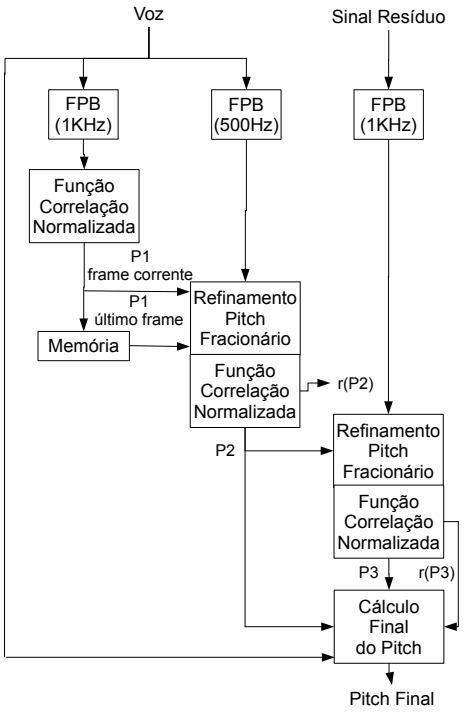

Fig. 2. Diagrama de blocos do decodificador MELP

O primeiro passo corresponde à passagem do sinal de voz digitalizado por um filtro passa-baixas com frequência de corte de 1 kHz, gerando o sinal s_k. A função de autocorrelação normalizada de s_k é então determinada por:

$$r(\tau) = \frac{c_\tau(0,\tau)}{\sqrt{c_\tau(0,0)c_\tau(\tau,\tau)}} \quad (1)$$

com

$$c_\tau(m,n) = c[m,n,\tau] = \sum_{k=-\lfloor \tau/2 \rfloor -80}^{-\lfloor \tau/2 \rfloor +79} s_{k+m}s_{k+n} \quad (2)$$

para $40 \leq \tau \leq 160$, onde $\lfloor \tau/2 \rfloor$ corresponde a um valor inteiro obtido por truncamento. A normalização utilizada na equação (1) visa compensar a variação de energia ao longo do tempo [8]. Neste processo, uma primeira estimativa P1 para o pitch é escolhida como o valor de τ que maximiza a função $r(\tau)$.

Posteriormente, uma etapa de refinamento do pitch é realizada, visando melhorar a estimativa inicial. Neste sentido, o procedimento verifica se o valor de máxima correlação está localizado no intervalo [T-1,T] ou [T,T+1]. Essa verificação ocorre avaliando se $c_T(0,T-1) > c_T(0,T+1)$, pois tal situação indica que o valor está em [T-1,T] e, nesse caso, $T = (T-1)$. Caso contrário, a etapa seguinte é executada diretamente, consistindo no uso das equações (1) e (2), com um novo sinal $\bar{s}_k$, saída de um filtro passa-baixas com frequência de corte de 500 Hz. Essa nova busca, contudo, é realizada somente para 5 amostras em torno do candidato a pitch P1 calculado no quadro corrente e no quadro anterior, ou seja, nos intervalos:

$$[P1_{quadro\ corrente} - 5, P1_{quadro\ corrente} + 5] \text{ e}$$
$$[P1_{quadro\ anterior} - 5, P1_{quadro\ anterior} + 5].$$

Os dois valores obtidos, que maximizam a função de autocorrelação normalizada nos intervalos considerados, são usados para determinar dois valores para a variável

$$\eta = \frac{C1 - C2}{C3 + C4}, \quad (3)$$

onde

$$C1 = c_T(0,T+1)c_T(T,T), \quad (4)$$
$$C2 = c_T(0,T)c_T(T,T+1), \quad (5)$$
$$C3 = c_T(0,T+1)(c_T(T,T) - c_T(T,T+1)), \quad (6)$$
$$C4 = c_T(0,T)(c_T(T+1,T+1) - c_T(T,T+1)). \quad (7)$$

A variável η, por fim, é utilizada no cálculo da correlação fracionária

$$r(\tau+\eta) = \frac{(1-\eta)c_T(0,T) + \eta c_T(0,T+1)}{\sqrt{c_\tau(0,0)C5}} \quad (8)$$

onde

$$C5 = (1-\eta)^2 c_T(T,T) + 2\eta(1-\eta)c_T(T,T+1) + \eta^2 c_T(T+1,T+1), \quad (9)$$

Desse forma, dentre os dois valores $(\tau+\eta)$, aquele que resulta no maior valor para (8) é escolhido como um candidato a pitch fracionário.

Na equação (3), os valores podem eventualmente ficar fora do intervalo [0,1], e por isso, são limitados entre [-1,2]. O pitch fracionário está limitado entre 20 e 160 [9].

O valor do pitch fracionário obtido acima é denominado P2 e está indicado na figura 2. O valor de P2 é também utilizado posteriormente na determinação da intensidade de voz, no cálculo do pitch final e no cálculo do ganho.

A etapa seguinte utiliza novamente a equação (1) para realizar uma busca levando em conta 5 amostras em torno do candidato P2, $[P2-5, P2+5]$, seguido de um refinamento do pitch para o valor que maximiza a função de autocorrelação normalizada no intervalo considerado. O sinal utilizado desta vez é o sinal resíduo obtido por um filtro passa-baixas com

frequência de corte de 1 kHz. O procedimento resulta no candidato a pitch P3.

O cálculo final do pitch é então realizado de acordo com o algoritmo apresentado na figura 3. Na etapa de dupla verificação do pitch, um procedimento analisa e corrige caso estejam sendo considerados múltiplos do pitch atual. O valor de Dth é um limiar utilizado nessa etapa.

Como pode ser percebido pela figura 3, o procedimento final eventualmente pode produzir novos valores para P3 e r(P3).

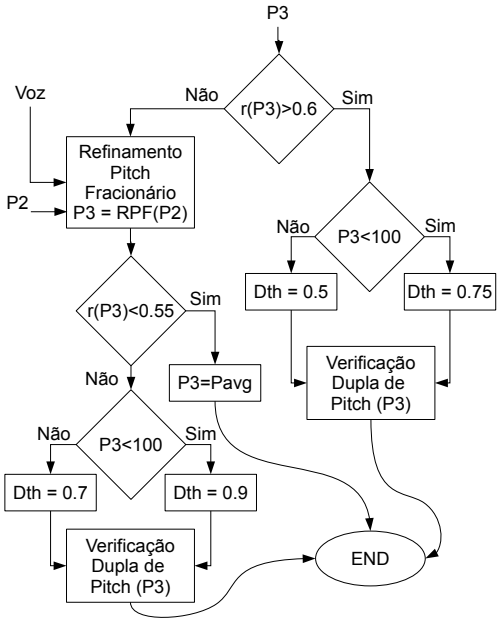

Fig. 3. Algoritmo para determinação do pitch final.

III. Modificações propostas

Esta seção apresenta as modificações que resultaram na versão acelerada do algoritmo, realizadas na etapa do cálculo da primeira estimativa inteira do pitch, ou seja, no cálculo de P1. Inicialmente é apresentada a forma "padrão", seguida da explanação da versão modificada.

A. Computação Padrão

Na prática, o cálculo para obtenção de P1 por meio da equação (1) pode ser realizado para $r(\tau)^2$, evitando assim a realização de uma raiz quadrada em cada etapa do cálculo:

$$r(\tau)^2 = \frac{c_\tau(0,\tau)c_\tau(0,\tau)}{c_\tau(0,0)c_\tau(\tau,\tau)}. \quad (10)$$

Dessa forma, a raiz quadrada pode ser efetuada apenas uma vez, após determinado o valor de τ que maximiza $r(\tau)^2$ no intervalo [40,160]. Sendo assim, para a busca do primeiro τ são realizadas 319 operações matemáticas (160 multiplicações e 159 adições) para cada um dos três diferentes $c_\tau[.]$ em (10), resultando em 960 operações para determinar o máximo valor de $r(\tau)^2$.

Além disso, podem ser utilizadas algumas relações de recorrência, permitindo o uso de valores prévios de $c_\tau(0,0)$ ou de $c_\tau(\tau,\tau)$ para determinar $c_{\tau-1}(0,0)$ ou $c_{\tau-1}(\tau-1,\tau-1)$, respectivamente, assumindo que os cálculos estão sendo realizados do maior valor de τ para o menor, ou seja, $\tau = 160, 159, \ldots, 41, 40$. Considerando valores ímpares de τ, temos:

$$c_\tau(0,0) = c_{\tau+1}(0,0) + s^2_{-\lfloor\tau/2\rfloor+79} - s^2_{-\lfloor\tau/2\rfloor-81}, \quad (11)$$

$$c_{\tau-1}(0,0) = c_\tau(0,0), \quad (12)$$

e

$$c_\tau(\tau,\tau) = c_{\tau+1}(\tau+1,\tau+1), \quad (13)$$

$$c_{\tau-1}(\tau-1,\tau-1) = c_\tau(\tau,\tau) + s^2_{-\lfloor\tau/2\rfloor+79+\tau} - s^2_{-\lfloor\tau/2\rfloor-81+\tau}. \quad (14)$$

Explorando essas recorrências, realizam-se 319 operações para calcular $c[0,\tau,\tau]$, 4 operações para corrigir $c[0,0,\tau-1]$ ou $c[\tau-1,\tau-1,\tau-1]$ (note que, para um dado τ, a correção é necessária apenas para um dos $c[.]$ do denominador de (10)) e 3 operações para calcular $r(\tau)^2$, correspondendo a um total de 326 operações.

Dessa forma, para encontrar P1 em um dado quadro, o algoritmo com uso de recorrência realiza $960 + (160-40) * 326 = 40.080$ operações, contra as $961 * 121 = 116.281$ operações efetuadas diretamente pela equação (1).

B. Parâmetro de Decimação

Considerando que o cálculo do pitch possui diversas etapas de refinamento até chegar ao seu valor final, a realização de decimação em k (na equação (2)) e/ou em τ (no intervalo [40,160]) na busca de P1 poderia reduzir substancialmente a complexidade computacional, seguindo a estratégia empregada em [10] para o Codec ITU-T G.729. A amplitude das decimações levariam a uma relação de custo benefício entre tempo de processamento e qualidade da voz reconstruída.

Sendo D_k e D_τ os fatores de decimação de k e τ, respectivamente, a equação (15) fornece a quantidade aproximada do número de operações requeridas na etapa de estimação do parâmetro P1:

$$N(D_k,D_\tau) = 2(320-D_k) + \lfloor\frac{121}{D_\tau}\rfloor(2\lfloor\frac{160}{D_k}\rfloor+2) + (\lfloor\frac{121}{D_\tau}\rfloor-1)(4D_t + (D_k-1)\lceil\frac{D_\tau}{2}\rceil)) \quad (15)$$

IV. Avaliação objetiva e seleção de amostra

Os fatores de decimação introduzidos na seção anterior foram implementados na versão de referência do MELP, variando cada fator de 1 a 10, resultando em 100 diferentes configurações. Cada configuração corresponde a uma complexidade computacional distinta, estimada pela equação (15) e uma qualidade de voz particular, estimada nessa etapa pelo algoritmo PESQ (*Perceptual Evaluation of Speech Quality*) [11].

Para avaliar a relação entre a complexidade computacional e a qualidade de voz, quando usados os fatores de decimação D_k e D_τ, foi utilizada uma base de dados consistindo em 40 sinais de falantes da Língua Inglesa americana (20 homens e 20 mulheres), tendo cada amostra uma duração média de 4,15 s, sendo realizada a codificação/decodificação por meio de cada uma das 100 diferentes versões do MELP.

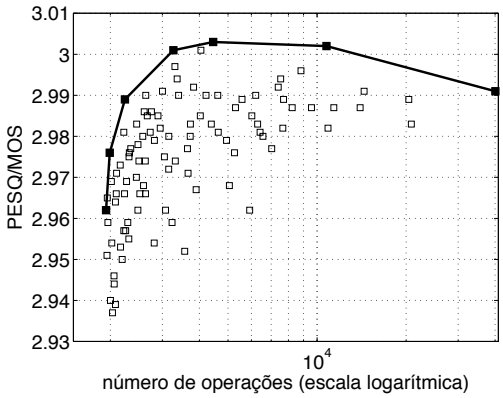

Fig. 4. Número de operações × valores PESQ-MOS das versões do MELP modificadas.

O resultado desse procedimento pode ser visto na figura 4, onde são apresentados os valores PESQ-MOS e suas correspondentes complexidades computacionais estimadas. A figura sugere que, para um dado nível de decimação, é possível obter um sinal com o mesmo nível de qualidade da voz do original, mas com o número de operações significativamente reduzido. Os pontos preenchidos do gráfico definem um feixe côncavo que representa a melhor relação custo-benefício entre boa qualidade de voz x complexidade.

A redução da complexidade computacional pode ser traduzida, em termos mais práticos, em menor tempo de processamento. Nesse contexto, foi realizada uma medição no tempo de processamento (*time profiling*) por meio da ferramenta *DTrace*, disponível nos sistemas operacionais baseados em Unix (Solaris, FreeBSD e MAC OS). Esse procedimento, realizado nos pontos constituintes do feixe côncavo da figura 4, resultou nos pontos apresentados na tabela I e na figura 5.

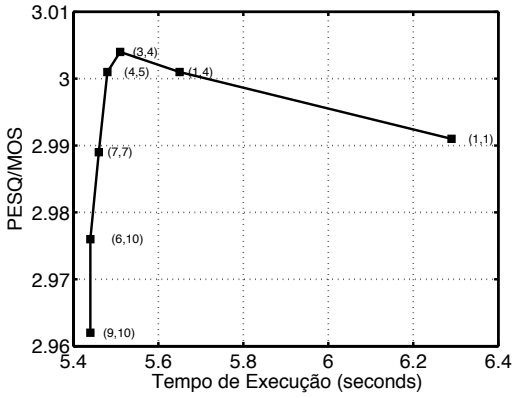

Fig. 5. Tempo de execução × PESQ-MOS para o feixe côncavo da Fig. 4.

Esses resultados sugerem que o uso da decimação dos parâmetros na etapa de estimação de P1 resultou em um significativo impacto no tempo total de processamento, permitindo uma redução de 13% e mantendo-se uma qualidade de voz aceitável.

TABELA I

Tempo de execução × PESQ-MOS para o feixe côncavo da Fig. 4.

D_k	D_τ	Tempo (s)	PESQ-MOS	% redução
9	10	5.44 ± 0.02	2.962	13.5
6	10	5.46 ± 0.02	2.976	13.2
7	7	5.45 ± 0.03	2.989	13.4
4	5	5.48 ± 0.02	3.001	12.9
3	4	5.51 ± 0.02	3.003	12.4
1	4	5.65 ± 0.03	3.002	10.1
1	1	6.29 ± 0.02	2.991	0.0

A tabela mostra que a partir de 13% o ganho no tempo de execução é pouco expressivo (valores dentro do intervalo de incerteza).

V. Avaliação subjetiva

Para realizar a avaliação subjetiva e confirmar a hipótese levantada na seção anterior, foram considerados dois tipos de testes, o *Absolute Category Rating* (ACR) e o *Comparison Category Rating* (CCR), descritos na recomendação P.800.

Em ambos os testes foram utilizadas uma versão MELP original e sua versão decimada com $D_k = 4$ e $D_\tau = 5$.

Durante os testes objetivos, o procedimento de analise e estabelecimento dos parâmetros foi realizado com uma base de vozes de falantes de língua inglesa falada nos EUA, com igual proporção de falas de homens e mulheres. Tendo em vista que os testes subjetivos devem ser realizados com uma base de vozes na mesma língua dos ouvintes, no caso, o português do Brasil, foram utilizadas frases nesse idioma. O mesmo banco de vozes foi utilizado nos dois testes subjetivos.

A. Absolute Category Rating - ACR

Neste teste, 32 sinais de fala, usando a Língua Portuguesa do Brasil, foram codificados e decodificados, por meio da versão original do MELP e de sua versão modificada. Estes sinais tinham duração entre 2 e 3 segundos e contemplavam diferentes falantes masculinos e femininos. Cada um dos sinais foi avaliado por 20 ouvintes "destreinados", que atribuíam uma nota MOS (*mean opinion score*) de acordo com a escala númerica apresentada na Tabela II.

TABELA II

Escala MOS

Escala MOS	Significado
5	Excelente
4	Bom
3	Regular
2	Ruim
1	Pobre

A média das notas para cada um dos 32 sinais é vista na Tabela III. O valor oficial esperado para o MOS do MELP é em torno de 3,2. Os resultados da Tabela III mostram uma pequena polarização, que tem sua explicação mais provável na dificuldade de se estabelecer uma referência por meio de valores de MOS conhecidos. Tal polarização, até mesmo mais

acentuada, pode ser observada também em outros trabalhos semelhantes [12].

TABELA III
MOS

MELP original	3,51 ± 0,31
MELP modificado	3,50 ± 0,32

B. Comparison Category Rating - CCR

Nesta avaliação os ouvintes realizam dois julgamentos por meio da resposta às perguntas: "Qual amostra apresenta melhor qualidade?" e "Quão melhor?". As respostas são quantificadas por meio da escala apresentada na Tabela IV. Os resultados obtidos para este teste estão apresentados no histograma da Fig. 6, onde as notas positivas favorecem o codificador modificado.

TABELA IV
Escala de teste CCR

Escala CCR	Significado
3	Muito melhor
2	Melhor
1	Pouco melhor
0	Igual
−1	Pouco pior
−2	Pior
−3	Muito pior

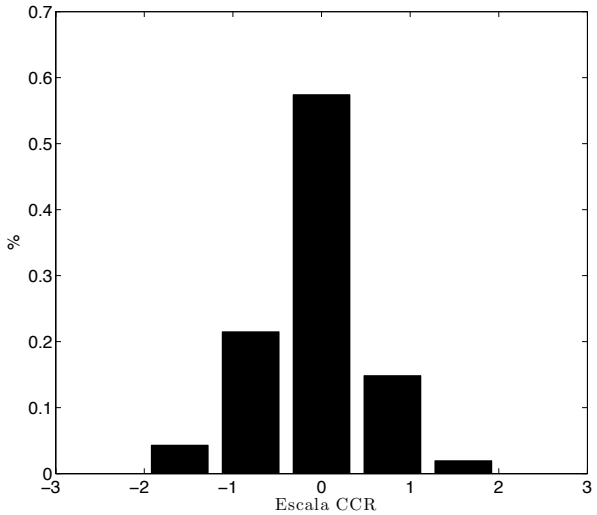

Fig. 6. Histograma do teste CCR.

De modo geral, os resultados de ambos os testes descritos nesta seção mostram uma certa equivalência na qualidade da voz resultante gerada pelas versões original e acelerada do algoritmo MELP. Deste fato, pode-se concluir que a aceleração obtida para o algoritmo de codificação, como descrito na Seção III, não provocou uma queda significativa de qualidade do sinal reconstruído.

VI. Conclusões

Os testes subjetivos apresentados por este trabalho fornecem subsídios que permitem concluir que o esquema de aceleração proposto permitiu reduzir o tempo de execução em aproximadamente 13%, sem afetar significativamente a qualidade da voz reconstruída.

Em termos práticos, os testes subjetivos do tipo ACR resultaram em valores essencialmente equivalentes. Essa conclusão foi complementada e ainda reforçada pelos resultados dos testes tipo CCR, que indicaram que os ouvintes em geral perceberam os sinais originais e modificados como iguais.

Ressalta-se a importância da ferramenta PESQ como guia para seleção de uma versão para realização dos testes subjetivos, tendo em vista que seria impraticável a realização de testes subjetivos com as 100 versões de MELP obtidas.

Percebe-se que considerar mais valores de k e/ou τ não necessariamente resulta em uma melhor qualidade de voz, sendo possível obter a mesma qualidade com menos esforço computacional, valendo-se do fato de que refinamentos adicionais do pitch são considerados pelo algoritmo do MELP em etapas posteriores ao cálculo de P1.

Os próximos passos do estudo visam estabelecer outros parâmetros e subrotinas passíveis de melhoramentos.

Agradecimentos

Os autores agradecem a Thiago de M. Prego por ter cedido o *setup* para testes subjetivos utilizado neste estudo, bem como a todas as pessoas que doaram parte do seu tempo para a realização destes mesmos testes.

Referências

[1] Supplee, L. and Cohn, R. and Collura, J. and McCree, A., "MELP: The new Federal Standard at 2400 bps", *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1591–1594, vol. 2, 1997.

[2] Tremain, T.E. and Kohler, M.A. and Champion, T.G., "A new mixed excitation LPC vocoder", *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1137–1140, vol. 2, 1996.

[3] Benesty, Jacob and Sondhi, M. M. and Yiteng Huang, *Springer Handbook of Speech Processing*, Berlin, Heidelberg, Springer Verlag, 2008.

[4] McCree, A.V. and Barnwell, T.P., III, "A 2400 bps mixed excitation LPC vocoder", Proc. IEEE Military Communications Conference, pp. 381–384, vol. 1, 1992.

[5] McCree, A.V. and Barnwell, T.P., III, "A new mixed excitation LPC vocoder", *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 593–596, vol. 1, 1991.

[6] Página educacional do Dr. Noam Amir, "Mixed Excitation Linear Prediction Speech Coder", Online: http://health.tau.ac.il/Communication%20Disorders/noam/speech/melp/Download/Download.htm, [13/03/2011].

[7] "Methods for Subjective Determination of Transmission Quality", ITU-T Rec. P.800, 1996.

[8] Chu, Wai C., *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, John Wiley & Sons, Inc., New York, USA, 2003.

[9] "Analog-to-digital Conversion of Voice by 2,400 bit/second Mixed Excitation Linear Prediction (MELP)", US DoD MIL-STD-3005, 1999 (Cancelada em 2008).

[10] Prego, T. de M., and Netto, S.L., "Efficient search in the adaptive codebook for ITU-T G.729 codec", *IEEE Signal Processing Letters*, pp. 881–884, vol. 16, Outubro 2009.

[11] "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs", ITU-T Rec. P.862, 2001.

[12] E. C. Tan and T. T. Teo, "Real-time Implementation of MELP Vocoder", *Journal of The Institution of Engineers*, Singapore. Vol. 44 Issue 3 2004.