Método para Embarcar Voz na Codificação de Vídeo Baseado em Transformadas Senoidais

Tiago Ramalho Melo, Sérgio Campos Sant'Ana, Bruno Macchiavello e Alexandre Zaghetto

***Resumo*— O presente projeto tem como objetivo principal acoplar um sinal de voz na codificação de um vídeo digital documental onde a voz é a única informação sonora que o compõe. Para atingir esse objetivo os coeficientes do sinal de voz calculados na fase de análise do vocoder LPC (*Linear Prediction Coding*), são inseridos em posições especificas dos coeficientes DCT (*Discrete Cosine Transform*) do vídeo codificado. No processo de decodificação do vídeo estes coeficientes LPC são recuperados e a fase de síntese do vocoder é realizada.**

***Palavras-Chave*— codificação de vídeo, codificação de voz, MPEG-2, vocoder LPC.**

***Abstract*— The main objective of this project is to use one single encoder to compress simultaneously digital speech and video, working on films that use only voice as a audio signal, for example, a video narrative. To achieve this goal, a method has been developed to insert the LPC (Linear Prediction Coding) voice coefficients into the video signal. More specifically, the LPC coefficients of the speech signal, are inserted into specific positions of the DCT (Discrete Consine Transform) coefficients of the encoded video. At the decoder these coefficients are recovered and the speech is synthesized.**

***Keywords*— video coding, speech coding, MPEG-2, LPC vocoder.**

I. Introdução

Em transmissões multimídias de vídeo e áudio os pacotes de cada um dos sinais são transmitidos em sessões separadas. Por exemplo, temos o protocolo RTP (*Real-time Transport Protocol*) [1] que especifica um formato para transmissão de dados em tempo real, utilizado em vídeo conferências, e que faz uso desta divisão dos sinais. A sincronização entre áudio e vídeo pode se feita por meio de *flags*, juntamente com a utilização de protocolos específicos, como o NTP (*Network Time Protocol*) [2]. A separação dos sinais também ocorre quando estes são armazenados.

Entretanto, existem formas alternativas onde a divisão dos sinais não é necessária, como a técnica descrita em [3], a qual mistura os sinais de áudio e vídeo de forma síncrona, e o método mostrado em [4] que descreve uma forma de acoplar o áudio no domínio da freqüência do vídeo quando este é comprimido. Em ambos os métodos referenciados existe a preocupação no sincronismo dos sinais, ou seja, o áudio é codificado conjuntamente com o quadro (*frame*) do vídeo a qual ele pertence.

Este artigo tem por objetivo propor um sistema de codificação conjunta do sinal sonoro com o sinal de vídeo para filmes que possuem apenas a voz como sinal de áudio, condição comum em documentários narrados, vídeo conferências ou telejornais. Logo, o objetivo não é codificar conjuntamente qualquer sinal de áudio, mas especificamente um sinal de voz. O foco será concentrado na preservação da qualidade dos sinais, não levando em consideração a questão da sincronia, pois espera-se que a sincronização seja realizada pelo decodificador.

Utilizou-se como padrão de codificação de vídeo o MPEG-2 [5], baseado na DCT (*Discrete Cosine Transform*) [6], [7], e que opera removendo redundâncias espacias, aquelas que ocorrem no quadro do vídeo, e redundâncias temporais, que acontecem entre os quadros da seqüência. Para a análise e síntese de voz foi usado o vocoder LPC (*Linear Prediction Coding*) [8], [9] que é um codificador paramétrico, ou seja, ele usa, por meio da modelagem do trato vocal, características do sinal a ser codificado para a posterior síntese da voz.

O restante do artigo está organizado da seguinte forma: a Seção II descreve o método elaborado para ocultar o sinal de voz na codificação do vídeo e são descritos os algoritmos e testes feitos para que o sistema pudesse ser construído, na Seção III são detalhados os testes feitos para comprovar a viabilidade da técnica proposta, a Seção IV apresenta as conclusões e a Seção V mostra as perspectivas para trabalhos futuros.

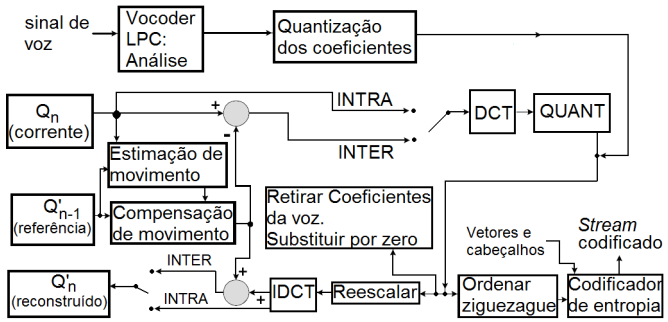
II. Arquitetura Proposta

Esta seção irá explicitar quais os processos, técnicas e ferramentas foram utilizadas no decorrer do projeto, descrevendo o codec do vídeo e o vocoder LPC utilizados, suas modificações em relação ao sistema original e como se dá o acoplamento dos sinais.

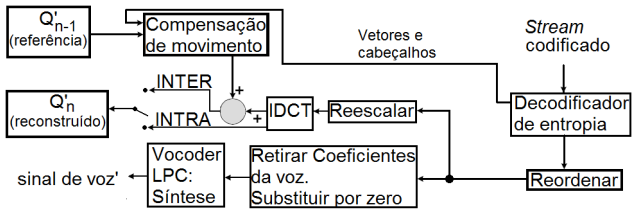
Os dados necessários para síntese de voz (coeficientes LPC, o ganho, a classificação surdo/sonoro e a freqüência fundamental) podem ser organizados em uma matriz, com cada linha representando as informações de cada quadro do sinal de voz. Já para a codificação do vídeo, temos que os passos antes do ordenamento em *stream* é a DCT e a quantização, ou seja, neste ponto todos os processos de transformações das imagens que compõe o vídeo já foram feitos, restando apenas a codificação entrópica, que não altera a qualidade da imagem. Logo, é possível "esconder" a matriz de coeficientes do sinal de voz entre os coeficientes DCT do vídeo, e deixar que o processo de codificação siga normalmente, com a codificação entrópica compactando os quadros do vídeo conjuntamente com os coeficientes do sinal de voz. Para a descompactação basta retirar dos coeficientes DCT os coeficientes da voz, e

Tiago Ramalho Melo, Sérgio Campos Sant'Ana, Bruno Macchiavello e Alexandre Zaghetto, Departamento de Ciência da Computação, Universidade de Brasília, Brasil. E-mails: tiargos@gmail.com, scsbsb@gmail.com, bruno@cic.unb.br, alexandre@cic.unb.br

organizá-los de forma que o sintetizador LPC recupere o sinal de voz. Este modelo pode ser visualizado na Figura 1.



(a) Codificador MPEG-2 com coeficientes da voz



(b) Decodificador MPEG-2 com coeficientes da voz

Fig. 1: Modelo para embarcar voz na codificação de vídeo

Existem diversas formas de ocultar os coeficientes da voz na DCT do vídeo. Para descobrir o melhor método de ocultação, nós realizamos vários testes com diferentes algoritmos. Os testes foram divididos em termos de posição no bloco DCT e componente YUV do vídeo. É importante salientar que os coeficientes LPC e o ganho do sinal de voz são originalmente dados em ponto flutuante, que é um formato incompatível com os coeficientes DCT, que são representados por inteiros. Logo, foi necessário quantizar as informações do sinal de voz para que elas pudessem ser incorporadas na DCT e posteriormente pudessem ser compactadas pelo codificador de entropia.

Em sistemas digitais os coeficientes LPC são normalmente quantizados para gerar um representação binária dos mesmos. A quantização proposta neste trabalho é multiplicar o coeficiente LPC em ponto flutuante por 10^n e truncar o resultado para um número inteiro, cada casa decimal do número quantizado será então oculta na DCT do sinal de vídeo. Foram utilizados nos testes de posição e componente quatro casas decimais para cada coeficiente LPC, ou seja, os coeficientes da voz em ponto flutuante foram multiplicados por 1000 e truncados.

A. *Posição*

Foram desenvolvidos cinco algoritmos para saber qual é a melhor posição onde se deve colocar o coeficiente de voz no bloco de coeficientes DCT. Entende-se por melhor posição o método de inserção que produz a menor degradação da qualidade da imagem e maior taxa de compressão. Pouca degradação ocorre quando o vídeo que embarcou em sua compactação os coeficientes do sinal de voz for comparado qualitativamente com o vídeo que foi compactado normalmente, e as suas qualidades resultantes são as mais próximas possíveis. Para mensurar esta qualidade foi utilizada a medida objetiva PSNR (*Peak Signal to Noise Ratio*).

Os cinco métodos desenvolvidos são:

1) O primeiro algoritmo coloca o coeficiente da voz sempre no último AC do bloco, ou seja, na posição (8,8), pois esta posição é a mais distante do coeficiente DC. Como em muitos casos o coeficiente substituído tem o valor zero, na recuperação do coeficiente de voz o valor colocado em seu lugar é sempre zero. Este método pode ser observado na Figura 2.

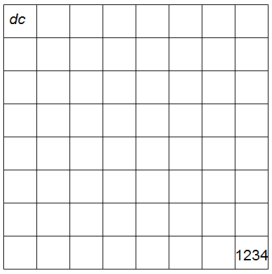


Fig. 2: Primeiro modelo de inserção dos coeficientes do vocoder LPC na DCT. Exemplo: inserindo um coeficiente com o valor 1234.

2) No segundo algoritmo cada um dos quatro dígitos do coeficiente da voz é colocado sempre em uma das quatro últimas posições do bloco. Esta forma pode ser melhor visualizada na Figura 3. Na retirada do coeficiente as posições alteradas recebem o valor zero, pois ,mesmo que este não seja seu valor original, o valor zero é o que menos degrada o bloco.

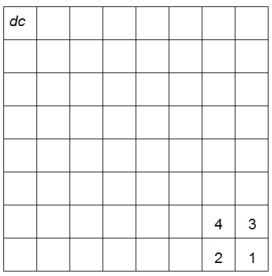


Fig. 3: Segundo modelo de inserção dos coeficientes do vocoder LPC na DCT. Exemplo: inserindo um coeficiente com o valor 1234.

3) Diferentemente dos algoritmos anteriores, o terceiro utiliza posições variadas para a inclusão. É feita uma busca em todas as posições do bloco, exceto a posição DC, para achar o melhor local de inserção para cada dígito do coeficiente da voz. É escolhida a posição que gerou o menor erro quadrático entre o dígito do coeficiente da voz e o coeficiente DCT. A busca não é feita pelo valor total, mas por cada dígito, pois o valor do coeficiente da voz é alto após a multiplicação por 1000 e os valores dos coeficientes DCT são em sua maioria pequenos, o que ocasiona um erro quadrático muito grande.

Por utilizar pontos variados de inserção, este método necessita de um "vetor de posições" que associa uma posição (x,y) para cada dígito de cada coeficiente do sinal de voz, o que ocasiona um aumento na taxa de bits de compactação quando comparado com os métodos anteriores.
Quando são retirados os coeficientes de voz não é feita a substituição por zero, pois teoricamente as posições em que se encontram os dígitos possuem os valores mais próximos dos valores originais, devida a busca feita inicialmente. Um exemplo deste algoritmo está na Figura 4.

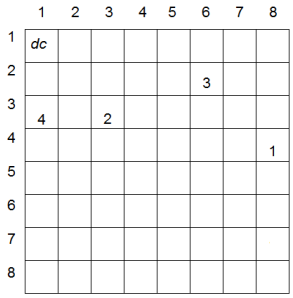


Fig. 4: Terceiro modelo de inserção dos coeficientes do vocoder LPC na DCT. Exemplo: inserindo um coeficiente com o valor 1234, que gera o vetor de posições 4,8,3,3,2,6,3,1.

4) O quarto algoritmo implementado utiliza a ideia do primeiro método, com a diferença que em cada bloco é colocado um dígito do coeficiente da voz na posição (8,8). Assim, todo coeficiente precisa de quatro blocos. Na recuperação, as posições de onde são retirados os dígitos recebem o valor zero. Processo que pode ser visualizado na Figura 5, onde os dígitos são colocados nos quatro primeiros blocos.

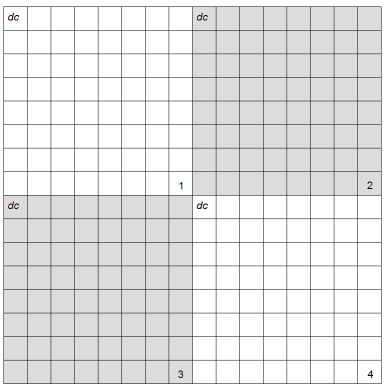


Fig. 5: Quarto modelo de inserção dos coeficientes do vocoder LPC na DCT. Exemplo: inserindo um coeficiente com o valor 1234, com cada dígito em um bloco.

5) O quinto método também utiliza a ideia de um dígito por bloco, mas a posição de inserção não é fixa. Assim como no terceiro algoritmo, é feita uma busca no bloco para encontrar a posição onde está o coeficiente DCT que gera o menor erro quadrático em relação ao dígito que será inserido. Logo, será necessário também a utilização do vetor posições e na recuperação dos coeficientes do sinal de voz as posições de onde são retirados os dígitos não recebem o valor zero.

Utilizando os cinco algoritmos descritos anteriormente na codificação da seqüência de teste "foreman.yuv", no formato CIF, chegamos aos gráficos das Figuras 6 e 7. No padrão MPEG-2 a quantização é feita por um passo de quantização QP e um fator de escala S, dada pela equação 1. Assim, os testes foram feitos com o QP fixo igual a 16, variando a escala S com incremento 4, a partir do valor 4 até 60. Para medir a taxa foi realmente calculada a entropia do sinal.

$$CoefDCTQuant = round(\frac{8 * coefDCT}{S * QP}) \qquad (1)$$

Observando o gráfico resultante dos testes fica claro que o melhor método de inserção é dado pelo primeiro algoritmo, que praticamente tem os mesmos valores da codificação normal. Este método tem a vantagem de pouco alterar o bloco, mudando o valor apenas de uma posição, diferentemente dos algoritmos que separam os coeficientes em dígitos, o que quadruplica as alterações. Outra vantagem é a utilização de uma posição fixa para inserções, dispensando o uso do vetor de posição que aumenta o tamanho das informações a serem compactadas. Uma terceira vantagem é a sua baixa complexidade computacional, já que buscas e construções de outras estruturas não são necessárias. É importante notar que para os métodos 3 e 5 a entropia inclui o sinal de vídeo e o vetor de posições.

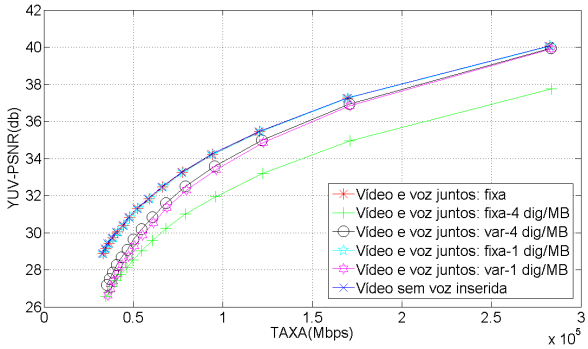


Fig. 6: Gráfico comparativo entre os 5 métodos de inserção, utilizando o vídeo "foreman.yuv"

B. Componente

Sabendo qual é o melhor posicionamento para o acoplamento dos coeficientes LPC na DCT, temos agora que verificar em qual componente do sistema de cores YUV a inserção dos coeficientes do vocoder LPC causa menor degradação. No presente trabalho utilizou-se o formato YUV 4:2:0, no qual as componentes de cor U e V possuem $\frac{1}{4}$ da resolução da componente de luminância. Logo, na luminância existe um maior espaço para guardarmos as informações da voz. Os testes foram aplicados ao vídeo "foreman.yuv", com a resolução 352x288 (CIF), e para efeitos de comparação também utilizamos a medida PSNR dos vídeos codificados.

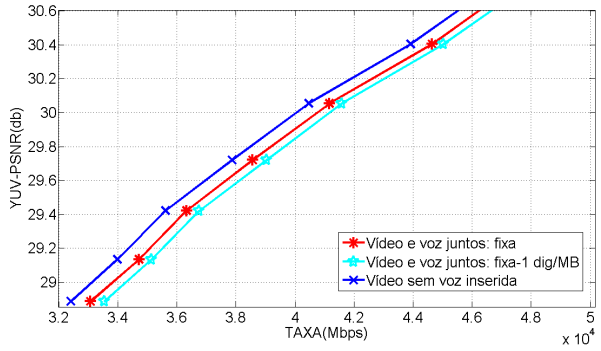


Fig. 7: Resultado comparativo entre os métodos de inserção mais bem-sucedidos, utilizando o vídeo "foreman.yuv"

Antes dos testes era esperado que o melhor resultado fosse o da inserção na crominância, já que, como dito anteriormente, o olho humano é mais sensível ao brilho. Mas como pode ser observado no gráfico da Figura 8, o melhor resultado se dá inserindo-se na componente de luminância. Isto deve-se ao fato do sinal de crominância ser muito suave, então a entropia aumenta ao inserir os coeficientes do vocoder LPC.

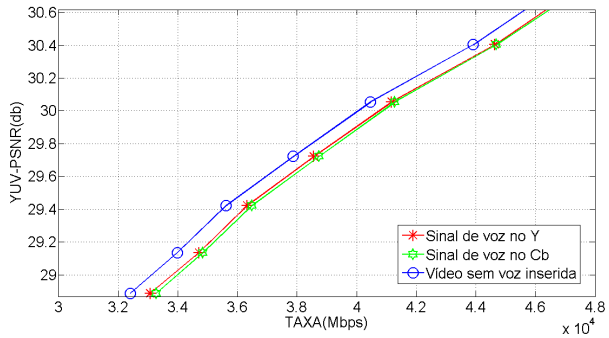


Fig. 8: Gráfico comparativo entre os métodos de inserção nas componentes, utilizando o vídeo "foreman.yuv"

A partir dos resultados dos testes de posição e componente, o método desenvolvido, o qual pode ser observado na Figura 1, consiste em na codificação inserir cada coeficiente do vocoder LPC na posição (8,8) de um bloco da componente de luminância. Na decodificação o valor desta posição é substituído por zero.

III. Resultados Experimentais

Para examinar o desempenho do método proposto foram feitos vários testes, onde foram utilizados sete vídeos padrões no formato *raw*, com resoluções QCIF, CIF e 4CIF. Como estes vídeos não possuem áudios próprios, e o projeto é restrito a sinal de voz, gravamos nossos próprios arquivos de voz para que estes pudessem ser codificados conjuntamente com os sinais de vídeo. Foram concebidos dois arquivos, um com voz masculina e outro com voz feminina, ambos amostrados à uma taxa de 8 KHz e possuindo apenas um canal (mono). O tempo de duração do arquivo de voz coincide com o tempo do vídeo no qual ele será inserido, por exemplo, para um vídeo com 300 quadros (10s), o sinal de voz a ser inserido dura 10 segundos.

Nos testes três parâmetros foram controlados. O GOP (*Group of Pictures*), com as opções IPPPP (um quadro I a cada quatro quadros P), todos os quadros I, IPIP (intercalar quadros I e P) e todos P (à exceção do primeiro quadro que é do tipo I). O segundo parâmetro é o fator de multiplicação dos coeficientes da voz, que determina quantas casas decimais eles terão. Foram usados os fatores 1000, 10000 e 100000, que produzem coeficientes com 4 casas, 5 casas e 6 casas respectivamente. Por último, foram usados vários níveis de quantização para o sinal de vídeo por meio da variação do fator de escala S ou do passo de quantização QP da equação 1. Foram determinadas 15 níveis de quantizações diferentes. Assim, para cada GOP foi variado os fatores do coeficiente da voz, e para cada fator de quantização da voz foi variada a quantização do vídeo, logo, foram feitas 180 (4x3x15) iterações para cada seqüência de teste. Escolhidas as seqüências testes, as avaliações foram feitas gerando gráficos como os das Figuras 9 e 10.

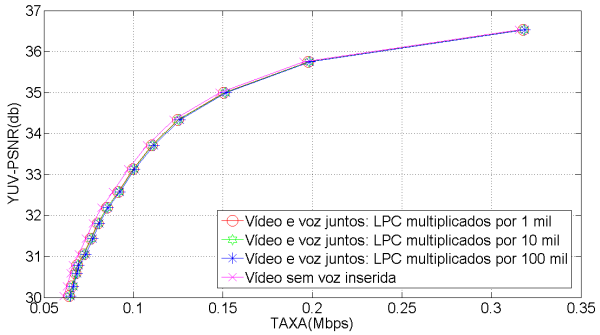


Fig. 9: Akiyo: GOP IPPPP

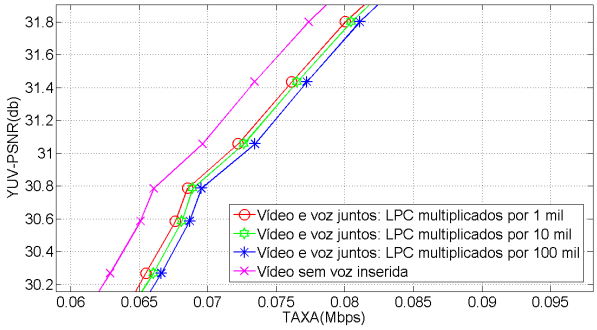


Fig. 10: Akiyo: GOP IPPPP - *zoom*

IV. Conclusões

Os resultados obtidos demonstram que existe um comportamento padrão, mesmo variando a resolução do vídeo e a GOP da codificação. Em todos os casos, a entropia do vídeo, codificado conjuntamente com o sinal de voz, aumenta à medida em que o nível de quantização dos coeficientes do vocoder LPC diminui. Resultado que já era esperado, pois, mais bits serão utilizados pelos coeficientes inseridos, aumentando o arquivo codificado.

Outro ponto importante, é o fato da qualidade do vídeo codificado em conjunto com o sinal de voz ser muito próxima à qualidade do vídeo codificado isoladamente. Em vários casos as PSNR's são as mesmas para ambos os processos. Observando os resultados dos menores fatores de quantização *S* do vídeo, os quais geram maiores divergências pelo fato do coeficiente DCT a ser substituído ter uma menor probabilidade de ser zero, é possível verificar que ainda assim as PSNR's são extremamente próximas ou iguais, ou seja, para quantizações menores na DCT a probabilidade de ser zero na posição (8,8) diminui, mas os valores ali encontrados não são muito distantes de zero, tornando o erro pequeno.

Quando a DCT sofre um processo alto de quantização, seus coeficientes tendem a ter valores próximos a zero, principalmente os coeficientes AC mais distantes do DC. Desta forma, é muito provável que o valor encontrado na posição onde será colocado o coeficiente LPC seja igual a zero, fazendo com que a substituição não ocasione problemas, já que na decodificação o coeficiente será substituído por zero. Variando o passo de quantização *QP* na quantização do vídeo, foram obtidos resultados semelhantes aqueles anteriormente encontrados variado o fator escala *S*.

As tabelas de quantização do MPEG-2 normalmente utilizam um passo de quantização muito maior para os coeficientes de mais alta freqüência. Por isso, o coeficiente onde o LPC é inserido é normalmente é igual ou próximo a zero. Em outro padrões de codificação como o H.264/AVC é possível preservar melhor a informação de alta freqüência, ou inclusive, codificar sem perdas, nesse caso a inclusão dos coeficientes LPC possivelmente deve gerar uma maior distorção no sinal do vídeo.

Percebeu-se que a GOP tem uma influência mínima na codificação de vídeo com inserção de voz. À medida em que o número de quadros do tipo P aumenta na GOP, a diferença entre as qualidades dos vídeos codificados com e sem ocultação de voz também cresce. Entretanto, estas divergências são ínfimas, girando em torno da média $1.85\text{x}10^{-3}$.

A codificação conjunta dos sinais de vídeo e voz não consegue obter as mesmas taxas utilizadas pelos sinais codificados separadamente, ou seja, somando-se taxas individuais da codificação de cada sinal consegue-se uma taxa de bits resultante menor do que a taxa proveniente da codificação acoplada. Dentre os motivos que ocasionam este fato, estão a quebra de padrões que otimizam os resultados do codificador de entropia, ocasionada pela inserção dos coeficientes da voz, e o uso de codificadores entrópicos que levam em consideração as características específicas de cada sinal.

Em relação ao sinal de voz, os dados revelam que a voz sintetizada pelos coeficientes retirados da DCT é idêntica à voz sintetizada pelo método o qual os coeficientes LPC não são inseridos na codificação do vídeo. Esta igualdade acontece em virtude do processo de ocultação na DCT não modificar em nenhum aspecto os coeficientes gerados no analisador LPC. Os únicos processos aos quais os coeficientes do vocoder LPC são submetidos, após serem gerados, é a quantização e truncamento, com a finalidade de transformá-los em valores inteiros. Estes processos fazem parte dos dois sistemas de sintetização, com ocultação na DCT e sem ocultação. Portanto, é possível afirmar que a qualidade da voz sintetizada não é prejudicada pelo procedimento de inserção na codificação do vídeo. Para medir a qualidade do sinal de voz foi utilizado o PESQ (*Perceptual Evaluation of Speech Quality*).

V. Perspectiva para Trabalhos Futuros

Uma extensão deste trabalho é o projeto para imagens impressas, onde o sinal de voz pode ser ocultado na DCT da imagem, que posteriormente é impressa. Assim, teríamos uma imagem em papel na qual é ocultado um sinal de voz. Sendo que para recuperá-lo bastaria escanear a imagem, retirar os coeficientes da voz que estão na DCT, substituindo-os por zero, e passar os coeficientes para o vocoder LPC para que a síntese do sinal seja feita.

Porém, alguns problemas foram encontrados nos testes preliminarmente executados, apesar destas avaliações demonstrarem a viabilidade da proposta. Assim, para trabalhos futuros fica a perspectiva de soluções para esses problemas. Dentre estas soluções está encontrar um método de inserção dos coeficientes da voz que não degrade, ou danifique minimamente, a imagem impressa, e que este processo possibilite recuperá-los de forma eficiente, gerando um sinal de voz sintético com qualidade aceitável. Na realização do trabalho sugere-se aperfeiçoar a modelagem do problema e testar realmente os processos de impressão e escaneamento, para verificar os reais danos que estes procedimentos físicos ocasionam na imagem, e elaborar métodos que os elimine.

Em relação ao tema principal do projeto, propõem-se elaborar e avaliar outros algoritmos de ocultação dos coeficientes do vocoder LPC, em termos de posição na DCT e componente YUV. Testar este sistema, e porventura outros elaborados, em diferentes padrões de codificação de vídeo, principalmente no H.264, e utilizar também outros processos de análise e síntese de voz. Posteriormente, verificar a possibilidade da ocultação de um sinal de áudio, e não apenas voz, na codificação do vídeo.

Referências

[1] P. Sanjoy, *Multicasting on the Internet and its Applications*. Kluwer Academic Publishers, 1998.

[2] D. L. Mills, Disponível em: `http://www.ntp.org/`. Acesso em: janeiro de 2011.

[3] H. Chen, Y. Zhao e L. Qi, *Inserted Audio-Video Mixed Signal Synchronous Coding Technique*. Chinese patent CN1599464, 2005.

[4] H. Chen, Y. Zhao, e L. Qi, *Audio-Embedded Video Frequency in Audio-Video Mixed Signal Synchronous Compression and Method of Extraction*. Chinese patent CN1655616, 2005.

[5] P. Sommer e M. Orzessek, *ATM and MPEG-2: Integrating Digital Video Into Broad*. Prentice Hall, 1998.

[6] I. Richardson, *H.264 and MPEG-4 Video Compression*. John Wiley and Sons Ltd, 2004.

[7] R. C. Gonzalez e R. E. Woods, *Digital Image Processing*. Prentice Hall, 2006.

[8] X. Huang, A. Acero e H. Hon, *Spoken Language Processing: A guide to Theory, Algorithm, and System Development*. Prentice Hall, 2001.

[9] L. Rabiner e R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall.