Um Estudo sobre Codificadores de Voz a Baixas Taxas operando em Ambientes Ruidosos e Redes IP

Fred Berkowicz, Rodrigo C. de Lamare e Abraham Alcaim

***Resumo*— Este artigo analisa o problema da transmissão da informação da envoltória espectral sobre a qualidade da voz em codecs a baixas taxas de bits, usando diferentes esquemas de quantização vetorial (QV). Estes codecs operam em redes IP e em ambientes com diferentes tipos de ruído: branco, fábrica e falatório. É também investigado o impacto da utilização de um supressor de ruído baseado em wavelets (Wavelet Denoising) na qualidade da voz codificada. A avaliação da qualidade da voz é feita utilizando-se a medida PESQ (Perceptual Evaluation of Speech Quality) da recomendação ITU-T P. 862 e testes subjetivos de comparação A/B.**

***Palavras-Chave*— Quantizadores vetoriais multiestágio, redes IP, perda de quadros, avaliação perceptiva de qualidade da voz (PESQ), Wavelet Denoising.**

***Abstract*— This paper analyses the problem of transmitting the spectral envelope information over the voice quality of low bit rates codecs, using different vector quantisation schemes. These codecs operate in IP networks and enviroments with different types of noise: white, factory and babling. It is also investigated the impact of the use of a Wavelet Denoising scheme on the encoded speech quality. The speech quality analysis is carried out using the PESQ measurement of ITU-T P.862 recommendation and through A/B comparison subjective listening tests.**

***Keywords*— Multistage vector quantisation, IP networks, frame losses, PESQ, wavelet denoising.**

I. INTRODUÇÃO

Devido ao grande crescimento da Internet e dos sistemas de comunicações móveis celulares, as aplicações de processamento de voz nesses meios têm despertado um interesse crescente. Tanto no caso de redes IP, como de telefonia móvel celular, os problemas relacionados ao projeto de codificadores são acentuados pelas altas taxas de erro de bits e perdas de pacotes, fora outros problemas usuais na concepção destes sistemas, como o ruído ambiente.

A maioria dos algoritmos de codificação de voz a baixas taxas é baseada na técnica de análise de predição linear (ou análise LPC - "Linear Predictive Coding"), onde um sinal de excitação é aplicado a um filtro só de pólos, caracterizado pelos parâmetros LPC, que representa a informação da envoltória espectral do sinal de voz. Este artigo trata do problema da transmissão da informação da envoltória espectral da voz e seu impacto sobre a qualidade da voz em codecs a baixas taxas, usando esquemas de quantização vetorial (QV) em redes IP e em ambientes com diferentes tipos de ruído: branco, fábrica e falatório. Bolot [1], [2] estudou a distribuição de perda de pacotes na Internet e concluiu que ela poderia ser aproximada por um modelo de perda Markoviano, também conhecido como Modelo de Gilbert. Neste trabalho as freqüências em linhas espectrais ou parâmetros LSF foram escolhidas para representar os coeficientes LPC, uma vez que são mais adequadas para procedimentos de quantização e interpolação [3].

Na quantização vetorial sem memória (QVSM) [4], [5], cada vetor de LSFs é quantizado de maneira independente de qualquer outro conjunto de LSFs. Entretanto, esta não é a melhor forma de codificar os parâmetros LSF, uma vez que ganhos podem ser conseguidos ao explorar a correlação entre conjuntos de LSFs adjacentes. Um esquema eficiente de codificação das LSFs, onde emprega-se quantização vetorial preditiva chaveada (QVPC) foi apresentado em [6], [7]. Neste artigo, o QVSM e dois esquemas de QVPC são comparados em um cenário envolvendo os seguintes problemas: canais sujeitos a perdas de quadros (simulados de acordo com o Modelo de Gilbert) e em ambientes ruidosos. Ainda, são comparadas as qualidades das vozes codificadas nessas condições, com sinais que passaram por um processo de supressão de ruído baseado em transformadas wavelets (*wavelet denoising*) [8], [9] antes da codificação. A qualidade da voz é avaliada através da medida PESQ da recomendação ITU-T P.862 [10] e testes subjetivos de comparação A/B.

Este trabalho é organizado da seguinte forma. A Seção II apresenta uma breve descrição do codec e dos diferentes tipos de quantizadores vetorias utilizados. A Seção III descreve o modelo de perdas de pacotes em redes IP. A Seção IV faz uma breve introdução à teoria de wavelets aplicada à supressão de ruído (Wavelet Denoising). A Seção V mostra e discute os resultados de simulações e a Seção VI é dedicada às conclusões deste artigo.

II. OS QUANTIZADORES VETORIAIS E O CODEC A BAIXAS TAXAS

Uma forma simples de melhorar o desempenho das estruturas de QV sem memória descritas em [11] e de obter vantagens na exploração da memória de uma fonte é usar quantização vetorial preditiva (QVP) [4]. Entretanto, existem situações de rápidas mudanças na envoltória espectral da voz e, portanto, baixas correlações entre os conjuntos de LSF adjacentes. Essa observação motivou a combinação de técnicas de QVSM e QVP para codificar quadros de baixa correlação separadamente dos quadros com alta correlação [12]. Dentre os esquemas de quantização vetorial preditiva chaveada (QVPC), destacamos o proposto por McCree e De

Fred Berkowicz, Rodrigo C. de Lamare e Abraham Alcaim, CETUC, Pontifícia Universidade Católica do Rio de Janeiro, 22453-900 Rio de Janeiro, RJ. E-mails: fredberko@ig.com.br, delemare@infolink.com.br e alcaim@cetuc.puc-rio.br. Este trabalho foi parcialmente financiado pelo CNPq.

Martin [13] (QVPC2), que usa 2 estruturas multiestágios de QVP com busca em árvore, onde cada preditor é projetado para um banco de dados de treinamento específico e opera com 21 bits por quadro.

Um outro esquema de QVPC [6], [7], denominado QVPC4, combina QVSM e quantização vetorial preditiva (QVP) para codificar quadros com alta correlação temporal de forma separada de quadros com baixa correlação temporal. O QVPC4 usa uma estrutura composta por 3 QVPs e 1 QVSM. Neste artigo escolhemos 3 quantizadores vetoriais das LSF, todos operando a 21 bits por quadro. O QVPC4 e o QVSM foram escolhidos, tendo em vista que testes em [7] mostraram que estes são mais robustos em presença de ruído. O terceiro quantizador vetorial que será utilizado é o QVPCP2, para que se possa avaliar também, o desempenho de um sistema composto só por QVPs operando simultaneamente em canais com perda de quadros e em ambientes ruidosos.

O codec utilizado neste trabalho, proposto por de Lamare e Alcaim [14], é baseado em excitação mista multibandas e opera em uma taxa de bits média de 1,2kb/s. O codificador realiza uma análise de predição linear a cada quadro de 20 ms e emprega o quantizador vetorial preditivo chaveado QVPC4, para codificar 10 parâmetros LSF com 21 bits por quadro. Neste trabalho, além de utilizar QVPC4, utilizaremos também os esquemas QVSM e QVPCP2. O ganho é quantizado uniformemente com 5 bits por quadro e a excitação é codificada com 3 bits por quadro. Os quadros de voz classificados como sonoros são divididos em 3 bandas de freqüências, que são obtidas com bancos de filtros fixos. Em seguida, uma análise em sub-bandas é realizada a fim de classificar as sub-bandas em sonoras e surdas. A excitação mista consiste da soma de pulsos periódicos nas sub-bandas sonoras com ruído branco nas sub-bandas restantes. Para quadros surdos, é empregada uma técnica de modelagem e síntese de fricativos e oclusivos. Com o objetivo de reduzir o ruído e melhorar a fala codificada são adotados um pós-filtro de restauração da envoltória espectral combinado com redução de ruído e um esquema de supressão de ruído. A alocação de bits do codificador é mostrada na Tabela I.

TABELA I
ALOCAÇÃO DE BITS DO CODEC OPERANDO A 1,2KB/S

Parâmetros	Quadro sonoro	Quadro surdo
LSFs	21	0
Ganho	5	5
Excitação	3	3
Fundamental	6	0
Total bits/20 ms	35	8
Bit rate	1,75 kb/s	0,4 kb/s

III. MODELO DE PERDAS DE PACOTES EM REDES IP

O sistema considerado neste trabalho é uma rede IP onde no terminal de transmissão, um conjunto de parâmetros LSF é quantizado e codificado a cada quadro de voz em uma palavra-código de 21 bits (Figura 1). Cada seqüência de conjuntos de parâmetros LSF codificados está associada a uma seqüência de quadros, que por sua vez é caracterizada por uma seqüência de bits. Esses quadros são encapsulados em pacotes, de acordo com o mecanismo de transmissão da rede, e enviados pelo canal. No destinatário, o enquadramento é desfeito pelo decodificador e os parâmetros LSF quantizados são recebidos.

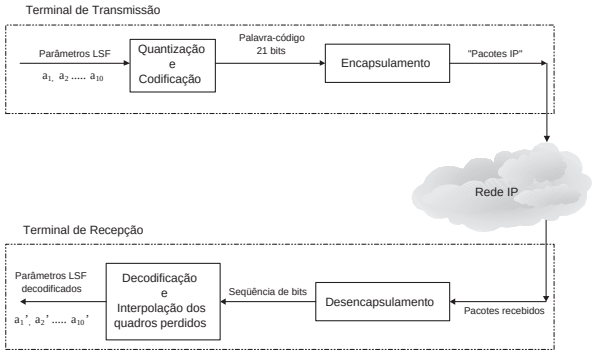

Fig. 1. Diagrama em blocos do sistema de transmissão e recepção de pacotes IP.

Em uma rede IP, havendo congestionamento, poderá ocorrer a situação de *buffer overflow* nos *switches* ou roteadores levando ao descarte ou perda de pacotes, caracterizando um canal com perda ou apagamento de quadro (*Frame Erasure*– FE). As perdas de pacotes em redes IP normalmente ocorrem em rajadas. Supõe-se, sem perda de generalidade, que um quadro do quantizador é encapsulado em um pacote. Para avaliar o desempenho dos quantizadores, adotamos um modelo Markoviano de dois estados, também conhecido como Modelo de Gilbert, para representar o canal com perdas. Os dois estados se referem aos eventos "pacote recebido" e "pacote perdido", respectivamente. Como mostrado na Figura 3, p é a probabilidade de transição do estado "pacote recebido" para o estado "pacote perdido", e q a probabilidade de transição do estado "pacote perdido" para "pacote recebido". A taxa de perda de quadro (TPQ), também conhecida como probabilidade de perda incondicional é dada por : $TPQp/(p+q)$. O comprimento médio da rajada (*B*) é dado por $B = 1/(1-ppc)$, onde *ppc* é a probabilidade de perda condicional, que é, a probabilidade de transição do estado "pacote perdido" para "pacote perdido" (*ppc = 1 - q*).

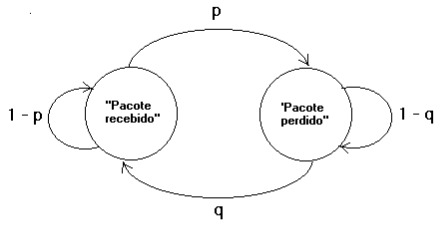

Fig. 2. Modelo de Gilbert.

Com a finalidade de combater os efeitos de perdas de quadros nos codificadores adotamos, neste artigo, a interpolação linear dos parâmetros LSF. Recebe-se, por exemplo, um conjunto de parâmetros LSF#1 e, por causa das imperfeições do canal, deixa-se de receber as LSF#2 e recebe-se as LSF#3. A interpolação permite que se obtenha uma

aproximação das LSF#2 às custas de um pequeno retardo adicional. Nota-se que é possível realizar a interpolação de mais de um conjunto de parâmetros LSFs às custas de um retardo cada vez maior [15], [16].

IV. Wavelet Denoising

Nas últimas décadas, as transformadas *wavelets* têm sido aplicadas a um número cada vez maior de problemas ligados direta ou indiretamente à área de processamento de sinais. Das várias aplicações pode-se citar a supressão de ruído (*denoising*) em sinais, bem como compressão, detecção e reconhecimento de padrões [17]. A capacidade de analisar sinais com espectro variante no tempo é um dos grandes atrativos da teoria de wavelets. Geralmente, os sinais são estudados como função do tempo, ou como função da freqüência. Entretanto, em muitas aplicações práticas, é de grande utilidade caracterizar o sinal tanto no domínio temporal, quanto no frequencial, como no caso de processamento de sinais de voz [18].

A idéia básica utilizada na supressão de ruído utilizando transformada *wavelet* é escolher quais coeficientes serão mantidos para preservar a informação do sinal, removendo assim os coeficientes associados à contribuição do ruído. Existem duas propriedades das transformadas *wavelet* que torna possível a supressão de ruído nos sinais. A primeira é que apenas alguns poucos coeficientes de decomposição serão não-nulos se as funções-base forem selecionadas adequadamente de acordo com as características do sinal analisado; assim se consegue uma alta concentração de energia nesses poucos coeficientes. A outra propriedade é que, se o sinal apresenta distribuição gaussiana, os coeficientes *wavelet* também apresentarão tal distribuição. Dessas propriedades, observa-se que os coeficientes da transformada *wavelet* de um sinal terão amplitude comparativamente superior aos coeficientes da transformada do ruído, esta diferença de amplitude torna possível uma operação de filtragem onde as componentes espectrais do sinal e do ruído podem estar superpostas em tempo e freqüência, o que não é possível com métodos baseados na transformada de Fourier [19].

Uma abordagem genérica para resolver o problema da supressão de ruído foi proposta originalmente por Donoho e Johnstone [19][20]. Ela consiste em eliminar os coeficientes menores que um certo limiar (*thresholding*), estabelecido de acordo com algum critério. O algoritmo genérico proposto por Donoho para realizar supressão de ruído, pode ser resumido nos seguintes passos: aplicar algum algoritmo *wavelet* rápido aos dados de entrada, dividindo o sinal analisado em componentes de alta escala (aproximações) e componentes de baixa escala (detalhes); aplicar uma função de *thresholding* (limiar) aos coeficientes de detalhe da transformada , usando um limiar especialmente estimado; e, por fim, inverter o algoritmo de *wavelet* para compor o sinal no domínio do tempo. O procedimento genérico de supressão de ruído depende principalmente da definição do limiar. Na supressão de ruído normalmente utilizam-se dois tipos de funções de limiar. O mais simples é o *Hard-Thresholding*, que substitui os coeficientes menor que um certo limiar por zero. O outro é o *Soft-Thresholding*, que apresenta propriedades matemáticas mais interessantes pois "encolhe"os coeficientes, evitando descontinuidades e instabilidade. Por essa razão, neste trabalho é usado o *Soft-Thresholding*. No *Soft-Thresholding* os coeficientes de detalhe da transformada do sinal limpo, são estimados por

$$\hat{\beta}_{jk} = \begin{cases} \text{sign}(Z_{jk})(|Z_{jk}| - t), & |Z_{jk}| \leq t \\ 0, & \text{caso contrário.} \end{cases} \quad (1)$$

onde t é o limiar e Z_{jk} é o k-ésimo coeficiente de detalhe da transformada *wavelet* de um sinal no nível j.

Para o cálculo do limiar, utilizou-se uma classe de regras que computa o limiar de forma adaptativa aos dados, conhecida como *SureShrink*. O estimador do limiar é expresso por [21]:

$$\hat{t}_j = \min_{t \geq 0} \sum_{k=1}^{N} (2\sigma_j^2 + t^2 - Z_{jk}^2) I\{|Z_{jk}| \geq t\}. \quad (2)$$

onde σ_j é estimado por $\hat{\sigma}_j = m_j/0,6745$, sendo que m é o desvio da mediana do valor absoluto calculado no nível j da transformada wavelet, e $I\{\cdot\}$ é uma função indicadora e definida como:

$$I\{x\} = \begin{cases} 1, & x \text{ verdadeiro;} \\ 0, & x \text{ falso.} \end{cases} \quad (3)$$

É importante ressaltar, ainda, que não é qualquer base de funções que pode ser usada como *wavelets* para representar um sinal. Esta base deve atender a duas condições: a primeira é que deve ser ortonormal, para assim poder realizar a reconstrução do sinal estimado; a segunda condição se refere à propriedade da decomposição *wavelet* do sinal possuir poucos coeficientes não nulos [8]. Uma transformada *wavelet* adequada deve concentrar mais de 90% da energia do sinal nos primeiros $N/2$ coeficientes [22]. Existem algumas famílias de funções, comumente apresentadas pela abreviação do pesquisador que as desenvolveram ("coif"para Coifman ou "db"para Daubechies") que podem ser usadas como funções *wavelet*. Como para sinais de voz a transformada Daubechies 10 ($db10$) satisfaz as condições, esta será usada nas simulações [22]. Além disso, testes de escuta informais, mostraram que a utilização de 5 níveis na transformada *wavelet* $db10$ era a melhor opção a ser utilizada.

V. Simulação e Discussão

Nessa seção são inicialmente apresentados os resultados obtidos utilizando os três esquemas de quantização vetorial descritos na Seção II (QVSM, QVPCP2 e QVPC4), em um codificador a baixa taxa de bits operando em ambientes ruidosos e redes IP. Posteriormente, são mostrados os resultados para vozes codificadas, quando se utiliza Wavelet Denoising, antes destas serem processadas.

O codec utilizado nas simulações e os diferentes esquemas de quantização testados foram apresentados na Seção II deste artigo. A base de dados usada para treinamento dos QVs foi produzida por 20 locutores masculinos e 20 femininos, onde cada locutor pronunciou 2 conjuntos de dez frases obtidas de listas foneticamente balanceadas para o português falado no Rio de Janeiro [23], resultando em um total de 800 frases. Os desempenhos foram aferidos utilizando-se um conjunto de 60 frases distintas, também foneticamente balanceadas [23],

produzidas por 2 locutores masculinos e 2 femininos, gerando uma coleção de 8922 quadros de LSFs.

Para avaliar a qualidade da voz em Codecs operando sobre redes IP utilizamos testes subjetivos de comparação A/B e a recomendação ITU-T P.862 de avaliação perceptiva de qualidade da voz (PESQ - *Perceptual evaluation of speech quality*), que consiste de uma técnica de medição objetiva para estimar a qualidade subjetiva que seria obtida em testes de escuta [10]. A saída do PESQ é a predição da qualidade percebida que seria dada a uma fala decodificada por um ouvinte em um teste de audição subjetivo como o MOS (*Mean Opinion Score*). A pontuação do PESQ é mapeada em uma escala tipo MOS, com um limite entre 1,0 e 4,5. Em situações de distorções extremamente altas, o resultado pode ficar abaixo de 1,0, mas isso é muito incomum.

O modelo de perda de quadros, descrito na Seção III, foi simulado para as condições de rede consideradas em um trabalho recente [1] e mostrada na Tabela II.

TABELA II
PARÂMETROS DO MODELO DE GILBERT USADO PARA SIMULAR AS CONDIÇÕES DA REDE COMO EM [1].

$TPQ(\%)$	ppc	B	p	q
0	–	–	0	0
0,6	0,147	1,17	0,005	0,853
9	0,330	1,49	0,066	0,670
28,6	0,500	2,00	0,200	0,500
38,5	0,600	2,50	0,250	0,400

A. Resultados com Vozes Ruidosas sem Wavelet Denoising

A Tabela III apresenta os resultados do teste de qualidade da voz codificada (utilizando a medida PESQ), com 3 diferentes tipos de ruído, processada por um codec operando em uma rede IP sem perda de quadros e utilizando os esquemas QVPC4, QVPCP2 e QVSM para quantizar os parâmetros LSF. Para RSR baixas, como $-5dB$ e $0dB$ a qualidade do sinal é bastante prejudicada em todos os ambientes ruidosos, independente do quantizador vetorial adotado. Como o codec opera a uma taxa de bits bastante baixa, o que já introduz um ruído de codificação, a simulação em ambientes ruidosos prejudica significativamente o seu desempenho. Entretanto, ao adotar o esquema QVPC4 proposto em [14], obteve-se um desempenho ligeiramente superior ao QVPCP2 em todos os ambientes ruidosos simulados e em toda faixa de RSR analisada. Já o QVSM forneceu um desempenho muito inferior em relação aos outros esquemas. Ainda é importante ressaltar que para a voz codificada, o ruído de falatório é o mais prejudicial, causando a maior queda de qualidade da voz codificada. Isso pode ser melhor observado na Figura 3, que apresenta o resultado PESQ para a voz codificada em presença dos 3 tipos de ruído, utilizando o QVPC4 e uma rede IP sem perda de quadros.

Em taxas de perda de quadros de até 9%, as simulações mostraram que ocorrem variações muito pequenas na qualidade da voz codificada em relação a TPQ = 0%. Quando ocorrem perda de quadros a taxas mais altas como 28,6% e 38,5% a qualidade da voz começa a cair consideravelmente.

TABELA III
RESULTADO PESQ PARA VOZES CODIFICADAS EM AMBIENTES RUIDOSOS E COM TPQ = 0%.

Ruído	$RSR(dB)$	-5	0	5	10
Falatório	QVPC4	1,833	1,966	2,132	2,275
	QVPCP2	1,808	1,944	2,112	2,224
	QVSM	1,727	1,845	2,001	2,098
Fábrica	QVPC4	1,965	2,112	2,271	2,415
	QVPCP2	1,953	2,090	2,260	2,402
	QVSM	1,824	1,946	2,086	2,224
Branco	QVPC4	2,032	2,155	2,327	2,475
	QVPCP2	2,022	2,145	2,311	2,461
	QVSM	1,988	2,032	2,148	2,264
	$RSR(dB)$	15	20	25	30
Falatório	QVPC4	2,389	2,502	2,590	2,670
	QVPCP2	2,345	2,466	2,561	2,658
	QVSM	2,217	2,293	2,347	2,399
Fábrica	QVPC4	2,536	2,598	2,649	2,696
	QVPCP2	2,509	2,586	2,638	2,661
	QVSM	2,306	2,356	2,389	2,392
Branco	QVPC4	2,597	2,659	2,698	2,733
	QVPCP2	2,586	2,646	2,680	2,710
	QVSM	2,353	2,387	2,423	2,430

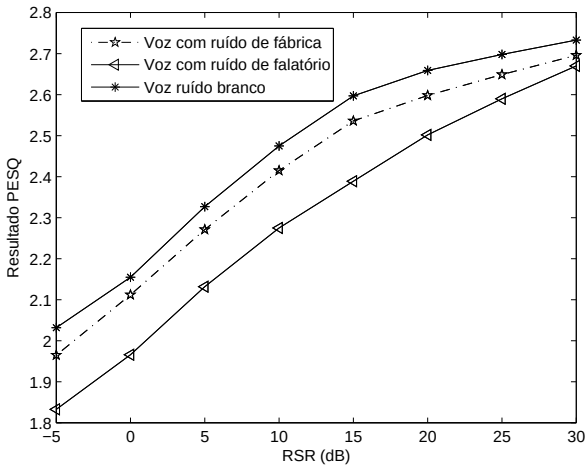

Fig. 3. Resultado do PESQ com vozes codificadas em ambientes ruidosos utilizando o QVPC4 e sem perda de quadros.

A uma TPQ = 28,6% a qualidade da voz começa a ficar bastante comprometida em RSR mais baixa que $10dB$, nos diversos ambientes ruidosos testados. Com uma TPQ 38,5%, o desempenho de todos os esquemas de QV fica bastante prejudicado. Entretanto, vale destacar que o desempenho do QVPC4 é ligeiramente melhor que o QVPCP2 e superior ao QVSM, mesmo em situações extremas de perda de quadros e RSR baixas.

Para melhor visualização e compreensão do impacto da perda de quadros e da presença de ruído ambiente na qualidade da voz, a Figura 4 mostra o desempenho em termos do

PESQ versus a razão sinal-ruído do ambiente, para as vozes codificadas em presença de ruído de fábrica, utilizando o QVPC4 em diversas taxas de perda de quadros. Vale informar que os comentários a seguir, também são válidos para os outros esquemas de quantização vetorial testados nos diversos tipos de ruídos. Verifica-se que com uma TPQ = 0,6% a voz codificada, tem qualidade similar à obtida em uma rede sem perda (TPQ = 0%). Já a uma TPQ = 9%, a qualidade tem uma pequena queda. A partir da TPQ = 28,6% a qualidade da voz começa a sofrer uma queda significativa. Finalmente, considerando o pior cenário em relação à perda de quadros simulado nesse artigo (38,5%), tem-se uma voz com qualidade bastante ruim. É importante destacar, que a queda de qualidade da voz codificada devido à perda de quadros de LSF, principalmente a uma TPQ = 38,5%, é muito agravada em situações de RSR baixas, chegando a um valor de PESQ inferior a 1,9.

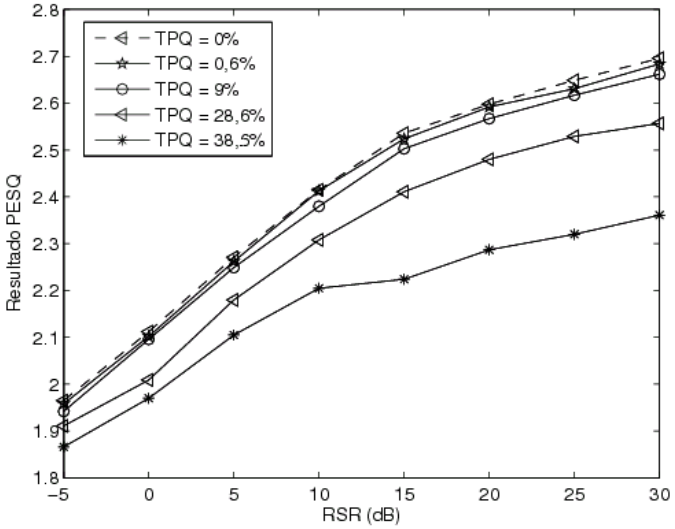

Fig. 4. Resultado do PESQ para vozes codificadas em presença de ruído de fábrica utilizando QVPC4 e com perda de quadros.

B. Resultados Obtidos com Vozes Ruidosas Utilizando Wavelet Denoising

O objetivo aqui, é verificar se a utilização de *wavelet denoising* nas vozes ruidosas, antes destas serem processadas pelo codec, melhora a qualidade da voz codificada. Para testar a qualidade da voz, utilizaremos dois métodos: a medida PESQ e testes subjetivos de comparação A/B. As simulações foram feitas utilizando o codec e as mesmas condições de rede IP utilizadas nas simulações anteriores. Para quantização das LSFs utilizou-se o QVPC4, por ter apresentado o melhor desempenho em relação aos outros dois esquemas testados.

A Tabela IV apresenta os resultados PESQ da voz codificada e sem perda de quadros. Pode-se verificar que o desempenho do codificador, de acordo com a medida PESQ, não melhorou com a utilização do algoritmo de *wavelet denoising*. Observa-se que para quase toda a faixa de RSR testada nos diferentes tipos de ruído, o codificador fornece melhores resultados sem a utilização do método de supressão de ruído. O desempenho do codec também foi avaliado com sinais processados ou não pelo *wavelet denoising*, a taxas de perda de quadros de 0,6%, 9%, 28,6% e 38,5%. Para todas as taxas não verificou-se melhora, segundo o critério de avaliação PESQ, com a utilização de supressão de ruído. Entretanto, testes informais de escuta sinalizam uma melhora de qualidade na voz codificada, pois a utilização do *denoising* ocasiona uma diminuição significativa do ruído de fundo. Por isso, decidimos usar uma medida subjetiva de qualidade - testes de comparação A/B - , além da medida objetiva, a fim de comparar os resultados obtidos com o PESQ.

TABELA IV

TABELA COMPARATIVA DO RESULTADO PESQ, PARA VOZES CODIFICADAS COM TPQ = 0%, EM AMBIENTES RUIDOSOS, COM E SEM A UTILIZAÇÃO DE WAVELET DENOISING.

$RSR(dB)$	Wavelet Denoising	Falatório	Fábrica	Branco
-5	Não	1,833	1,965	2,032
	Sim	1,899	1,880	1,376
0	Não	1,966	2,112	2,155
	Sim	2,049	2,034	1,825
5	Não	2,132	2,271	2,327
	Sim	2,145	2,204	2,266
10	Não	2,275	2,415	2,475
	Sim	2,276	2,347	2,420
15	Não	2,389	2,536	2,597
	Sim	2,345	2,439	2,559
20	Não	2,502	2,598	2,659
	Sim	2,500	2,570	2,641
25	Não	2,590	2,649	2,698
	Sim	2,570	2,637	2,694
30	Não	2,670	2,696	2,733
	Sim	2,643	2,670	2,709

Para realização dos testes subjetivos de comparação A/B, utilizou-se dois conjuntos de frases. Um com frases corrompidas por ruído de falatório e outro formado por frases com ruído branco. Cada conjunto foi formado com RSR variando entre $5dB$ e $15dB$. As frases foram codificadas com o codec operando a uma TPQ = 0%.

O teste A/B foi realizado com 9 pares de sentença para cada conjunto. O material de teste incluiu voz codificada corrompida por ruído de falatório, com e sem a utilização de *wavelet denoising*, em um conjunto, e voz codificada corrompida por ruído branco, com e sem a utilização de *wavelet denoising* em outro conjunto. As sentenças foram apresentadas a 30 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 incluiu um total de 540 opiniões para cada conjunto considerado. Os resultados mostrados na Tabela V, referentes ao conjunto de vozes com ruído de falatório, revelam que 42% dos ouvintes não mostraram uma preferência clara, 45,9% deles preferiram a voz codificada com a utilização de *wavelet denoising* e 12% preferiram a codificação sem a utilização de *wavelet denoising*. A Tabela VI, referente ao conjunto de vozes com ruído branco, mostra que 25,4% dos ouvintes não têm uma preferência clara, 52,4% deles preferiram a voz codificada com a utilização de *wavelet*

denoising e 22,2% preferiram a codificação sem a utilização de *wavelet denoising*. Observa-se, portanto, que os resultados com PESQ não foram corroborados com a medida subjetiva, concluindo-se, assim, que o emprego de *wavelet denoising* é vantajoso.

TABELA V

COMPARAÇÃO A/B PARA VOZES COM RUÍDO DE FALATÓRIO.

Resultados para vozes codificadas	%
Usando Wavelet Denoising	45,9
Qualidade Comparável	42,0
Sem Wavelet Denoising	12,0

TABELA VI

COMPARAÇÃO A/B PARA VOZES COM RUÍDO BRANCO.

Resultados para vozes codificadas	%
Usando Wavelet Denoising	52,4
Qualidade Comparável	25,4
Sem Wavelet Denoising	22,2

VI. CONCLUSÕES

Nesse trabalho, verificou-se que o esquema de quantização vetorial QVPC4, proposto por de Lamare e Alcaim [14], foi o que alcançou melhor desempenho em todos os ambientes ruidosos e em todas as taxas de perda de quadros quando comparados aos esquemas QVPCP2 e QVSM. O QVSM, foi o que forneceu o pior desempenho nas condições simuladas. Verificou-se, ainda, que até a uma taxa de perda de quadros de 9% o desempenho dos quantizadores não foi muito afetado, ficando próximo do desempenho conseguido em redes sem perdas. Os resultados mostraram que em RSR baixas, menores que 10dB, a qualidade da voz processada pelo codec a baixas taxas fica bastante comprometida, tendo um resultado PESQ muito baixo. Essa situação é agravada quando a rede opera com taxas de perdas de quadros altas, como 28,6% e 38,5%. De qualquer modo, o melhor desempenho, mesmo em ambientes muito hostis, é atingido pelo QVPC4.

Também foi avaliada a qualidade da voz codificada, com a utilização de *wavelet denoising* e em diferentes taxas de perda de quadros, através da medida PESQ. Nessa medida o desempenho do codec não melhorou com a utilização da técnica de supressão de ruído apresentada. No entanto, testes subjetivos de comparação A/B, mostraram que a maioria dos ouvintes preferem a voz codificada com a utilização de *wavelet denoising*. Ressalte-se, portanto, que a medida PESQ deve ser usada com cautela, pois em determinadas situações – como as aqui estudadas – são necessários testes de qualidade subjetivos para melhor avaliar a qualidade de voz percebida pelos ouvintes.

REFERÊNCIAS

[1] D. Quercia, L. Docio-Ferandez, C. Garcia-Mateo, L. Farinetti and J. C. De Martin, "Performance analysis of distributed speech recognition over IP networks on the AURORA database", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 3820-3823, 2002.

[2] J.-C. Bolot, "Characterizing end-to-end packet delay and loss in the Internet", *Proc. ACM SIGCOMM*, pp. 289-298, September 1993.

[3] F. K. Soong e B.H. Juang, "Line spectrum pair (LSP) and speech data compression", *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc.*, 1984.

[4] A. Gersho e R. M. Gray, *Vector quantization and signal compression*, Kluwer Academic Publishers, 1992.

[5] K. K. Paliwal e B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame", *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp. 3-14, 1993.

[6] R. C. de Lamare e A. Alcaim, "Analysis of LSF switched-predictive vector quantisers", *Proc. International Symposium on Signal Processing and its Applications*, Kuala-Lumpur, Malaysia, 2001.

[7] R. C. de Lamare e A. Alcaim,"Noisy channel perfomance of LSF switched-predictive vector quantisers", *Proc. IEEE International Conference on Information, Communications and Signal Processing*, Singapore, 2001.

[8] S. G. Mallat,"A theory for multiresolution signal decomposition : the wavelet representation.", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674-693, July 1989.

[9] D. L. Donoho,"De-noising by soft-thresholding", *IEEE Transactions on Information Theory*, vol. 41, pp. 613-627, May 1995.

[10] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,"Recommendation P.862, ITU-T, Fev. 2001.

[11] W.P. LeBlanc, B. Battacharya, S.A. Mahmoud and V. Cupperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding", *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 4, pp. 373-385, 1993.

[12] M. Yong, G. Davidsson and A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction", *Proc. ICASSP-88*, vol. 1, pp. 402-405, New York, USA, 1988.

[13] A. McCree and J.C. De Martin, "A 1.7 KB/S Melp Coder with Improved Analysis and Quantization", *Proc. ICASSP-98*, 1998.

[14] R. C. de Lamare and A. Alcaim, "Strategies to Improve the Performance of Very Low Bit Rate Speech Coders and Application to a 1.2 kb/s Codec", *IEE Proceedings on Vision, Image and Signal Processing*, vol. 152, no. 1, pp. 74-86, February 2005.

[15] E. Daniel e K. Teague, "Federal standard 2.4 kbps MELP over IP", *Proc. 43rd IEEE Midwest Symp. on Circuits and Systems*, 2000.

[16] J. Wang e J. Gibson, "Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001.

[17] M. Bahoura and J. Rouat, "Wavelets Speech Enhancement Based on the Teager Energy Operator", *Proc. IEEE Signal Processing Letters*, pp. 10-12, January 2001.

[18] S. Qian and D. Chen, "Joint time-frequency analysis: methods and applications", *Prentice Hall PTR*, EUA 1996.

[19] D. L. Donoho and I. M. Johnstone, "Ideal Spatial Adaptation via Wavelet Shrinkage", *Department of Statistics, Stanford University*, EUA 1992.

[20] D. L. Donoho and I. M. Johnstone, "Ideal denoising in an Orthonormal Basis Chosen from a Library of Bases", *Department of Statistics, Stanford University*, EUA 1994.

[21] C. A. Medina, "Realce de Voz Aplicada à Verificação Automática de locutor ", *Dissertação de Mestrado, Instituto Militar de Engenharia*, Rio de Janeiro, 2003.

[22] J. I. Agbinya, "Discrete Wavelet Transform Techniques in Speech Processing", *IEEE TENCON - Digital Signal Processing Applications*, pp. 514-519, 1996.

[23] A. Alcaim, J. A Solewicz, e J. A. Moraes, "Freqüência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro ", *Revista da Sociedade Brasileira de Telecomunicações*, vol. 7, pp. 23-41, 1992.