PAPER

Técnicas de Codificação de Voz Aplicadas em Sistemas Móveis Celulares

João Luiz A. CARVALHO *(97/07867) e* Danilo DIAS *(97/07760)*

RESUMO Este trabalho aborda os conceitos básicos das principais técnicas de codificação de voz utilizadas em sistemas móveis celulares. Trata inicialmente das características do sinal de voz e das diferenças entre codificadores de forma de onda, paramétricos e híbridos. Na seqüência, são apresentados os codificadores LPC e CELP e os principais algoritmos efetivamente aplicados na telefonia móvel: RPE-LTP, VSELP, ACELP e QCELP.

palavras-chave: *codificação de sinais de voz, codificadores de forma de onda, codificadores paramétricos, codificadores híbridos, LPC, RPE-LTP, GSM, CELP, VSELP, IS-54, ACELP, IS-136, QCELP, IS-95.*

1. Introdução

O desenvolvimento de técnicas avançadas de codificação de voz tornou possível e viável a introdução dos sistemas digitais de telefonia móvel, diminuindo a largura de banda requerida por usuário. Assim, foi possível aumentar o número de usuários do sistema, tornando a telefonia celular um sistema mais barato e acessível à população.

No entanto, o processo de codificação deve ser simples e rápido o suficiente para que o sistema funcione em tempo real com processadores relativamente baratos e de baixo consumo. Além disso, a qualidade da voz codificada deve ser tal que permita não só a inteligibilidade do que é ouvido, mas também que se possa reconhecer o interlocutor e perceber outras informações como a entonação e a emoção.

As técnicas de codificação abordadas neste trabalho fornecem qualidade de voz suficiente para estes fins, e devido à possibilidade de uso de codificação de canal, além da própria robustez intrínseca de cada algoritmo e do sistema digital em si, conseguem qualidade de voz superior à dos sistemas analógicos frente a ruído no canal.

Este trabalho aborda inicialmente os conceitos básicos envolvidos na codificação de voz, tratando sobre as características do sinal de voz e as diferenças entre codificadores de forma de onda, paramétricos e híbridos. A seguir, idéias básicas sobre amostragem, quantização e codificação de linha são apresentadas, e introduzem-se os codificadores LPC e CELP nos quais são baseadas as demais técnicas abordadas. As principais técnicas aplicadas nos sistemas móveis atuais (Tabela 1) são então apresentadas e, em conclusão, faz-se uma análise comparativa sobre as técnicas discutidas.

Tabela 1 Sistemas móveis e técnicas de codificação aplicadas

Padrão	*Técnica*	*Taxa de bits*
GSM	RPE-LTP	13 kbit/s
IS-54B	VSELP	7,95 kbit/s
IS-136	VSELP/ACELP	7,95 kbit/s
IS-95	QCELP	1,2/2,4/4,8/9,6 kbit/s

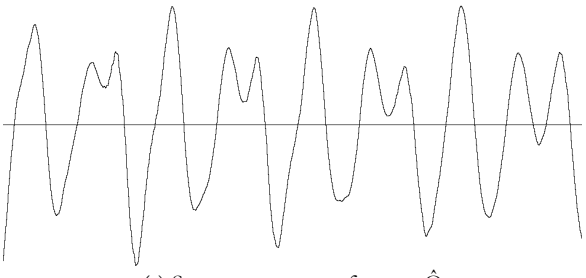
2. Características do Sinal de Voz

Para se desenvolver uma técnica de codificação eficiente, é necessário antes conhecer as características básicas do sinal de voz. O mecanismo de produção da voz apresenta uma resposta limitada em freqüência, com limite por volta de 10kHz. Como a maior parte da energia do sinal de voz está concentrada na faixa de freqüência entre 300 e 3400 Hz, pode se limitar o canal a esta banda com uma perda tolerável em qualidade.

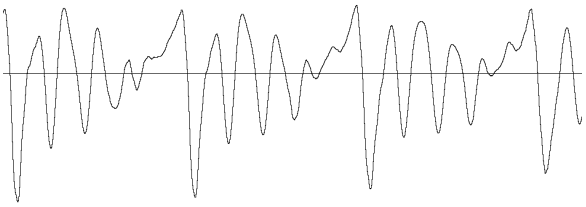
O sinal de voz se apresenta de forma sonora ou surda, conforme haja vibração ou não das cordas vocais. São classificados como sinais surdos na fala, fonemas com características de ruído, como o 'S' e o 'CH'. Já os sinais sonoros são as vogais e consoantes com características não ruidosas. Nestes sinais, a vibração das cordas vocais se dá a partir de uma freqüência fundamental, ou o *pitch*. As demais harmônicas definem o timbre, que é o que modela a forma de onda periódica, trazendo assim informações importantes, já que é essa forma de onda que permite o reconhecimento de um fonema e também do interlocutor. Portanto, o timbre é uma característica fundamental para que se possam distinguir vozes de mesma freqüência que sejam emitidas por diferentes pessoas. Uma outra característica importante do sinal de voz é a amplitude, que determina a intensidade do som, e é função da força ou potência com que a voz é produzida.

Todas as características citadas acima podem ser observadas nos segmentos de voz da Fig. 1, que estão na mesma escala de tempo e amplitude. Nos segmentos sonoros das Figs. 1(a) e 1(b) é fácil notar que as amplitudes são altas e que há uma periodicidade,

determinada pelo *pitch*. Já no segmento surdo da Fig. 1(c) os níveis de amplitude são relativamente baixos e não há periodicidade, fazendo com que o sinal lembre um ruído. Comparando ainda os dois segmentos sonoros, é possível perceber que além de terem formas de onda bastante diferentes, permitindo assim distinção dos diferentes fonemas, o segmento da Fig. 1(a) tem um período de *pitch* menor que o da Fig. 1(b), o que caracteriza uma voz mais aguda. As pequenas diferenças percebidas no sinal a cada período de *pitch* é um dos fatores que torna mais difícil a codificação.



(a) Segmento sonoro: fonema Ô



(b) Segmento sonoro: fonema A

(c) Segmento surdo: fonema S

Fig. 1 Comparação entre segmentos sonoros e surdos

Para codificação de sinais de voz, algumas propriedades do sinal são exploradas com mais ênfase:

- a distribuição de probabilidade da amplitude do sinal não é uniforme;
- a autocorrelação entre amostras sucessivas da voz é diferente de zero;
- o espectro de freqüência do sinal não é uniforme;
- é possível distinguir no sinal de voz segmentos sonoros e segmentos surdos;
- existe no sinal de voz uma quase periodicidade.

As propriedades citadas acima permitem uma quantização com baixo ruído e técnicas de codificação que tendem a diminuir a redundância do sinal de voz, que, por ser limitado em banda, pode ser discretizado no tempo a uma taxa finita por um processo de amostragem, e ser reconstruído completamente de suas amostras. A amostragem, a quantização e a codificação desses sinais serão discutidas nos itens a seguir.

3. Codificadores de Forma de Onda

Os codificadores de forma de onda, ou de linha, são esquemas que tentam aproximar o sinal gerado ao sinal de voz original. A forma básica de codificação aplicada no sinal de voz é a digitalização, uma vez que o sinal obtido é analógico ou contínuo no tempo. Para isso, são usados sistemas que amostram o sinal, quantizam as amostras obtidas e as convertem para o sistema binário.

Por implicarem em sinais digitais com altas taxas de bits, esses sistemas não são utilizados diretamente nos sistemas de telefonia móvel, mas são necessários para que formas de codificação mais eficientes sejam aplicadas, uma vez que todas essas técnicas utilizam processamento digital de sinais.

O sinal amostrado é definido com um sinal PAM (*Pulse Amplitude Modulation*), e após a quantização e a codificação, o sinal digital é definido como um sinal PCM, ou *Pulse Code Modulation*. Cada um desses passos da digitalização do sinal de voz é discutido a seguir.

3.1 Amostragem

Apesar de ser limitado em banda, nem toda a informação inicial do sinal analógico é amostrada. Como já mencionado, consegue-se uma boa qualidade de voz considerando-se apenas componentes de freqüência compreendidas na faixa entre 300 e 3400 Hz. Explorando essa característica, o sinal original - na faixa de 10kHz - é filtrado, para limitá-lo dentro de faixa de 4kHz, e a seguir amostrado, segundo o Teorema de Nyquist, a uma taxa de 8kHz, ou seja, 8000 amostras por segundo.

Se for utilizada uma taxa de amostragem inferior a 8kHz, é provável que ocorra superposição espectral (*aliasing*), resultando em distorções nas componentes mais altas de freqüência. Utilizando-se taxas mais elevadas, serão necessários amostradores mais caros, sem ganho na qualidade, e resultando em uma maior taxa de bits após a codificação.

3.2 Quantização

O sinal amostrado é discreto no tempo, mas tem valores contínuos de amplitude. A quantização é o processo que consiste em discretizar esses valores, explorando características da distribuição de probabilidade da amplitude do sinal, obtendo assim uma perda menos sensível de informações devido ao truncamento do sinal amostrado.

Essa perda de informação é chamada de ruído de quantização e pode ser minimizada utilizando-se níveis

de quantização com distância variada de acordo com a amplitude, uma vez que amostras de menor amplitude são mais prováveis, trazendo assim a maior parte da informação. Na Fig. 2 é possível observar como essa técnica melhora a quantização de sinais de baixa amplitude, uma vez que com a quantização uniforme a informação amostrada praticamente não descreve a original, enquanto que com a quantização não uniforme o sinal quantizado se aproxima mais da onda que deve ser codificada. Há, no entanto, uma maior distorção na quantização de amostras de maior amplitude, que por serem menos prováveis, influem menos na quantização do sinal como um todo.

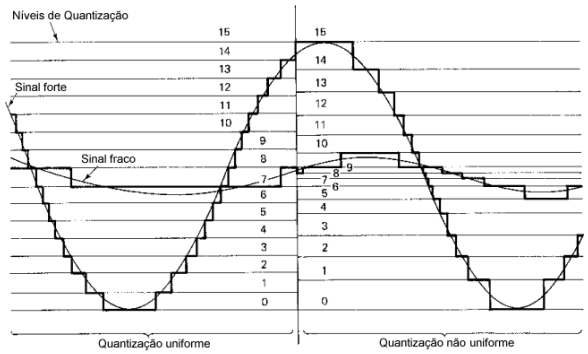


Fig. 2 Comparação: quantização uniforme e não uniforme[12]

As técnicas que exploram as características da distribuição de probabilidade da amplitude do sinal são chamadas de log-PCM, sendo que os esquemas mais utilizados são as duas leis de compressão recomendadas pelo antigo CCITT (atual ITU-T): a lei-μ e a lei-A. Esses sinais são obtidos através de uma compressão do sinal antes da quantização e uma expansão após a decodificação. Esse processo é conhecido como compansão e é utilizado no sistema de telefonia fixa.

3.3 Codificação

O sinal amostrado e quantizado, já com características discretas tanto no tempo como na amplitude, deve ser então convertido para um esquema que possa ser facilmente processado por microcomputadores e outros sistemas digitais e que possa permitir técnicas de codificação que minimizem sua redundância e técnicas de modulação que melhorem o desempenho na transmissão. Obviamente esse esquema é o binário.

Cada amostra do sinal é codificada como uma seqüência de bits que traz a informação do valor quantizado da amplitude da amostra em questão. Um maior número de bits permite um número maior de níveis de quantização, segundo a equação $L=2^n$, onde n é o número de bits e L é o número de níveis de quantização. Usando mais níveis de quantização, diminui-se o intervalo entre níveis e assim o ruído de quantização.

Utilizando o log-PCM, é possível conseguir uma boa qualidade de voz com apenas 256 níveis, ou 8 bits. Como a taxa de amostragem é de 8KHz, esse esquema implica em uma taxa de bits de 64kbit/s, usado na telefonia fixa. Essa taxa é considerada muito alta para um sistema de telefonia móvel, onde a banda utilizada por cada usuário deve ser a menor possível. Portanto, são necessárias técnicas de codificação adicionais que diminuam essa taxa de bits para valores mais adequados a esse tipo de sistema.

3.4 PCM Adaptativo e Diferencial

Algumas técnicas adicionais podem ser utilizadas para diminuir a taxa de bits nos codificadores de forma de onda. Entre elas convêm destacar o PCM Adaptativo (APCM), o PCM Diferencial (DPCM) e o PCM Adaptativo e Diferencial (ADPCM).

No APCM, o passo de quantização varia com o tempo, de modo a acompanhar as variações de amplitude do sinal de voz, baseando-se nas amostras passadas do mesmo. Assim reduz-se a faixa dinâmica do sinal e conseqüentemente a taxa final de transmissão.

Já o DPCM explora a significativa correlação entre amostras sucessivas do sinal de voz, uma vez que este é bastante redundante. Nesta técnica quantiza-se a diferença de amplitude entre amostras adjacentes, que por ser relativamente pequena, pode ser representada com menos bits. O sinal de entrada no quantizador é a diferença entre o sinal original e uma predição do mesmo, baseada nas amostras passadas, resultando em um sinal chamado de erro de predição, que por sua vez é codificado a uma taxa de 32 kbit/s.

Os codificadores ADPCM empregam quantização e/ou predição adaptativas. A predição adaptativa consiste no ajuste dinâmico do preditor de acordo com variações no sinal de voz. Assim codificadores ADPCM apresentam boa qualidade de voz para taxas entre 24 e 48kbit/s.

4. Codificadores Paramétricos e Híbridos

Codificadores paramétricos operam utilizando um modelo de como o sinal foi gerado, e tentam extrair do sinal os parâmetros desse modelo. São esses parâmetros que são enviados ao decodificador. Codificadores paramétricos para sinais de voz são chamados de vocoders, e como são baseados nas estatísticas do sinal de voz, não funcionam para outro tipo de sinais. Além disso, a qualidade é inferior a dos padrões telefônicos, e a voz reproduzida tem um aspecto sintético ou não

natural, mas esses sistemas fornecem inteligibilidade da voz com taxas abaixo de 4kbit/s.

Nos vocoders, o trato vocal é representado como um filtro variante no tempo e é excitado ou com uma fonte de ruído branco, para segmentos surdos do sinal de voz, ou com um trem de pulsos separados pelo período de *pitch*, para segmentos sonoros. Portanto, a informação que deve ser enviada ao decodificador é a especificação do filtro, um *flag* de segmento sonoro/surdo, o ganho aplicado ao sinal de excitação (G) e o período de *pitch,* para os segmentos sonoros. Essas informações são atualizadas a cada 10-20ms para seguir a natureza não-estacionária da fala.

Esses parâmetros do modelo podem ser determinados pelo codificador por diferentes métodos, usando técnicas no domínio do tempo ou no domínio da freqüência. Porém, um tipo de vocoder muito utilizado é o LPC, que extrai características perceptivelmente importantes do sinal de voz diretamente da forma de onda no tempo.

Com taxas de bits entre 4 e 16 kbit/s (um pouco maior que as dos vocoders), os codificadores híbridos, que exploram técnicas tanto dos codificadores paramétricos como dos de forma de onda, conseguem uma qualidade muito próxima à dos codificadores de linha, que requerem taxas acima de 16kbit/s. Os codificadores híbridos também são baseados nos modelos de produção da voz e utilizam uma excitação mais apurada para o sintetizador, que propicia uma melhora na qualidade da voz sintetizada, tornando-a mais inteligível que nos vocoders convencionais.

A qualidade da voz, em função da taxa de bits e do tipo de codificador, é apresentada no gráfico da Fig. 3.

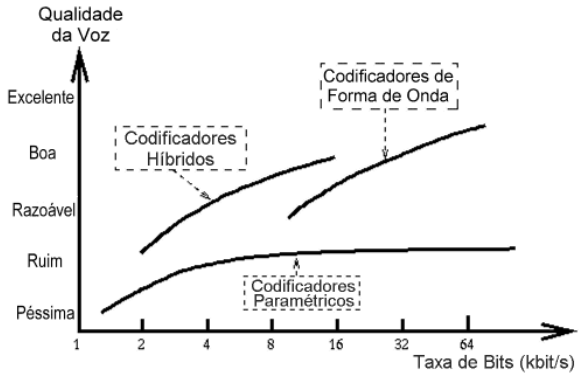


Fig. 3 Comparação entre classes de codificadores de voz [3]

As técnicas de codificação abordadas neste trabalho distinguem-se basicamente por alterações no processo de geração da excitação, já que a modelagem do trato vocal basicamente não sofre modificações em qualquer codificador de voz que utilize predição linear a curto termo. Por este motivo o vocoder LPC é apresentado a seguir.

5. Linear Prediction Coding: LPC

O LPC (Codificação Por Predição Linear) é um codificador paramétrico de sinais de voz muito utilizado e que, como já mencionado, extrai os parâmetros para o modelo do trato vocal diretamente da forma de onda no tempo, obtendo um resultado melhor que outros tipos de vocoders que obtêm seus parâmetros a partir do espectro de freqüência.

Fundamentalmente, um LPC analisa a forma de onda para produzir um filtro de síntese variante no tempo que modela o trato vocal, reproduzindo sua função de transferência. No receptor, cujo diagrama de blocos é apresentado na Fig. 4, um sintetizador recria o sinal de voz pela passagem de uma excitação específica pelo mesmo modelo matemático do trato vocal, que é atualizado periodicamente.

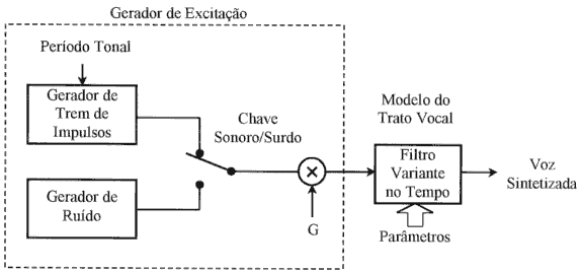


Fig. 4 Diagrama de blocos do decodificador LPC [3]

A excitação ideal a ser aplicada no filtro de síntese é o sinal residual obtido na saída do filtro inverso, quando na sua entrada é aplicado o próprio sinal de voz. Porém, no LPC essa excitação é modelada parametricamente, sem que se busque reproduzir sua forma de onda, mas somente as características mais marcantes do seu espectro de amplitude. É a modelagem "trem de impulsos/ruído", onde a excitação, que é amplificada com um ganho G relativo à amplitude do sinal de voz, pode ser uma fonte de ruído branco (para segmentos surdos) ou um trem de pulsos separados pelo período de *pitch* (para segmentos sonoros).

O modelo do trato vocal é conseguido por um filtro digital só de pólos e variante no tempo, do tipo:

$$H(z) = \frac{1}{A(z)} \tag{5.1}$$

onde,

$$A(z) = 1 - \sum_{i=1}^{10} \alpha_i z^{-i} \tag{5.2}$$

e os parâmetros α_i são os coeficientes preditores.

O filtro de síntese H(z) tem a função de introduzir no sinal sintetizado a correlação a curto termo encontrada no sinal original. É a análise STP, ou s*hort-*

term prediction, na qual o cálculo dos coeficientes preditores pode ser realizado da forma apresentada a seguir.

1) Para estimar a amostra presente, o preditor usa uma soma ponderada das últimas 10 amostras. Assim, pode-se reescrever a amostra sn como:

$$s_n = \sum_{i=1}^{10} \alpha_i s_{n-i} + e_n \quad (5.3)$$

onde, s_n: amostra atual
s_{n-i}: amostras anteriores
e_n: erro de predição.

2) Os coeficientes são calculados de modo a minimizar a energia média *E* do sinal e_n:

$$E = \sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N} \left(\sum_{i=0}^{10} \alpha_i s_{n-i} \right)^2 \quad (5.4)$$

onde, por (5.2), $\alpha_0 = -1$ e *N* é o número de amostras contidas na janela de tempo pela qual o erro de predição é computado. Valores típicos são uma janela de 10 ms, correspondendo a um valor de *N* = 80 amostras.

3) Para minimizar *E* a respeito de um coeficiente α_m, é necessário fazer sua derivada parcial igual a zero:

$$\frac{\partial E}{\partial \alpha_m} = \sum_{n=1}^{N} 2 s_{n-m} \sum_{i=0}^{10} \alpha_i s_{n-i} = 0 \quad (5.5)$$

$$= \sum_{i=0}^{10} \sum_{n=1}^{N} s_{n-m} s_{n-i} \alpha_i = 0 \quad (5.6)$$

onde m=1,2,...,10.

4) A soma mais interna pode ser vista como o coeficiente de correlação C_{im}, reescrevendo:

$$\sum_{i=1}^{10} C_{im} \alpha_i = C_{0m} \quad (5.7)$$

5) Após determinar os coeficientes de correlação C_{im}, a equação (5.7) pode ser usada para determinar matricialmente os coeficientes preditores, já que se trata de um sistema linear de 10 equações e 10 variáveis.

Os coeficientes preditores são atualizados a cada 20ms aproximadamente, e são enviados junto com os demais parâmetros do sistema:

- o *flag* de segmento sonoro/surdo;
- o período de *pitch*, no caso de segmento sonoro;
- o fator de ganho de excitação G, que determinará a amplitude do sinal de saída do filtro de síntese.

Após serem codificados de forma binária, os coeficientes do filtro e os parâmetros descritos acima são enviados a uma taxa de 2,4 kbit/s. O resultado é um sinal de voz de qualidade razoável, levando-se em conta a relativamente baixa taxa de bits.

6. Regular-Pulse Excited LPC with a Long-Term Predictor: RPE-LTP

O codificador utilizado no padrão GSM (*Global System for Mobile communications*), sistema de telefonia móvel digital largamente utilizado na Europa, é o RPE-LTP, que é um esquema híbrido baseado no LPC, no qual a excitação consiste em pulsos regularmente espaçados e de amplitude variada e que possui um preditor a longo termo.

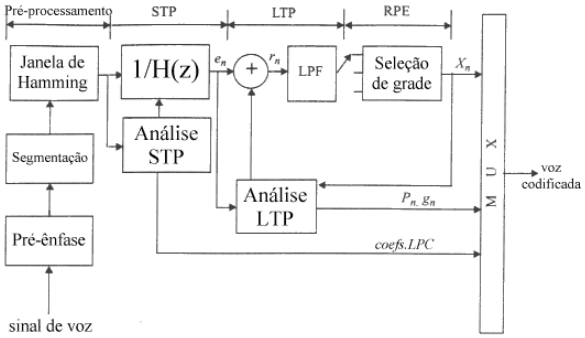
Este codificador tem uma taxa de bits de 13 kbps e combina vantagens do RELP (*Residual Excited Linear Predictive*) de banda básica, proposto pela França, e o MPE-LTP (*Multi-Pulse Excited – Long-Term Prediction*), proposto pela Alemanha, que não serão abordados neste trabalho. O que deve ser levado em consideração, no entanto, é que o RPE-LTP modifica o codificador RELP, incorporando algumas características do MPE-LTP, e com isso reduz a taxa de bits de 14,77 kbit/s para 13 kbit/s sem perda de qualidade. A principal dessas modificações é a adição de um esquema de predição a longo termo, a chamada análise LTP (*long-term prediction*). Além disso, deve-se considerar que no MPE-LTP, parte da informação transmitida é referente a posição dos pulsos usados na excitação, enquanto no RPE-LTP esses pulsos são regularmente espaçados dado um intervalo fixado, cabendo ao codificador determinar somente a posição do primeiro pulso e a amplitude de cada um deles, o que explica o termo *Regular Pulse Excited*.

O codificador do RPE-LTP, apresentado na Fig. 5(a), é composto de quatro blocos de processamento principais:

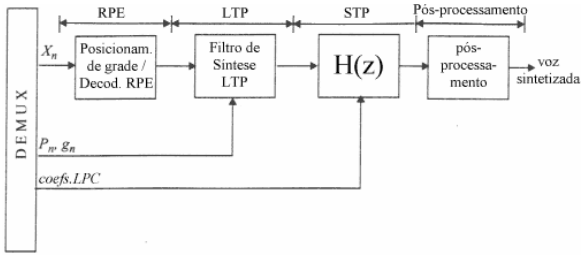
- 1ª Etapa: A seqüência de voz passa por uma equalização do tipo pré-ênfase, é ordenada em segmentos de 20 ms, e passa por um janelamento de Hamming;
- 2ª Etapa: Um filtro faz a análise STP do sinal, encontrando os coeficientes preditores do filtro de síntese. A seguir, esses parâmetros são utilizados para construir o filtro LPC inverso, que determinará o erro de predição e_n, que na verdade é a excitação que deveria ser utilizada no decodificador;
- 3ª Etapa: Porém, uma análise LTP do erro de predição encontra um período de *pitch* e um fator de ganho tais que minimizam esse sinal, maximizando a correlação cruzada de amostras sucessivas do mesmo. O erro de predição minimizado é chamado de LTP residual, ou r_n. Esse sinal é ponderado e a seguir decomposto em três possíveis seqüências de excitação;
- 4ª Etapa: A seqüência de maior energia é selecionada para representar o LTP residual, sendo

normalizada em relação ao pulso de maior amplitude, e transmitida a uma taxa de 9,6 kbit/s.

O receptor, mostrado na Fig. 5(b), realiza o processo inverso. A seqüência de pulsos de excitação X_n é decodificada e processada por um filtro de síntese LTP, que usa o *pitch* e o ganho para sintetizar o sinal *long-term*. Este passa por um filtro de síntese STP construído a partir dos coeficientes preditores, recriando assim o sinal original.



(a) Codificador



(b) Decodificador

Fig. 5 Diagrama de blocos do RPE-LTP [2]

Os blocos de 20 ms de voz contêm 260 bits, que são ordenados, dependendo de sua importância, em grupos de 50, 132 e 78 bits cada. Os do primeiro grupo são os que mais afetam a qualidade da voz, por isso têm um maior número de bits para correção e detecção de erros. Os do segundo grupo são de média importância e os do terceiro grupo, por serem menos significativos, não têm correção ou detecção de erros.

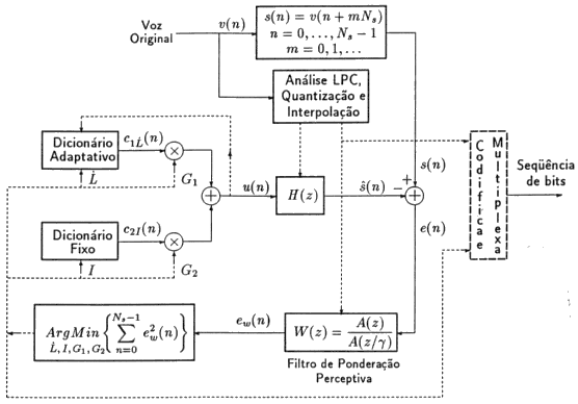
O RPE-LTP fornece boa qualidade de voz, mas consome bastante potência e requer uma alta taxa de bits se comparado com outros codificadores mais modernos. Sua maior vantagem é sua relativa simplicidade, uma vez que ele pode rodar em tempo real em um PC 486 de 66 MHz. Porém, com o barateamento dos DSP's (processadores de sinais digitais) e a demanda cada vez maior por banda nos sistemas móveis, o RPE-LTP já vem sendo substituído por algoritmos de codificação mais eficientes, como o ACELP.

7. Code Excited Linear Predictive: CELP

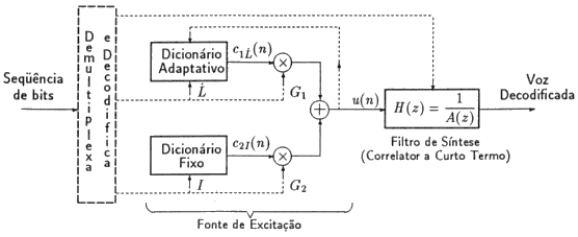
A codificação preditiva linear excitada por dicionário de códigos, ou CELP, é um algoritmo híbrido, que trabalha com um modelo matemático do trato vocal ao mesmo tempo em que tenta aproximar o sinal gerado ao sinal de voz original, conseguindo assim reproduções de boa qualidade a taxas de até 4 kbit/s, consideradas relativamente baixas.

A maioria das técnicas de codificação de voz aplicadas na telefonia celular derivam do CELP, se distinguindo deste basicamente por modificações nos dicionários de código que dão origem à excitação. Por isso, embora não seja utilizado diretamente nos sistemas móveis, o CELP é abordado neste item.

A Fig. 6(a) mostra um diagrama de blocos genérico do codificador CELP, que tem um modelo básico do tipo "fonte de excitação/filtro sintetizador", como nos codificadores paramétricos.



(a) Codificador



(b) Decodificador

Fig. 6 Diagrama de blocos de um sistema CELP [9]

A diferença básica entre o CELP e os vocoders está na fonte de excitação, que é não é do tipo trem de pulsos/ruído como no LPC, mas sim constituído por um esquema de quantização vetorial de dois estágios em paralelo.

O primeiro estágio possui um dicionário adaptativo que é constituído por K_1 seqüências retiradas da

excitação passada e é responsável pela reprodução da periodicidade dos segmentos sonoros do sinal de voz, substituindo o filtro correlator a longo termo empregado nos primeiros codificadores CELP. Valores típicos para K_1 são 128 e 256.

O segundo estágio possui um dicionário fixo constituído por K_2 seqüências estocásticas, determinísticas, ou obtidas por meio de um procedimento de treinamento. Valores usuais de K_2 são 128, 256, 512 e 1024.

Para uma freqüência de amostragem de 8 kHz e período de atualização do preditor de *pitch* de 5ms, estas seqüências devem ter 40 amostras cada. Os parâmetros do filtro síntese, que são em geral 10, são por sua vez atualizados a cada 20ms.

A seqüência de excitação com a qual se deseja reproduzir o segmento de voz é obtida como uma combinação linear de duas seqüências obtidas uma de cada dicionário, que são escolhidas por um procedimento onde são testadas diversas excitações possíveis para sintetizar o segmento de voz corrente, escolhendo-se aquela que minimiza a medida de erro ponderado na saída do filtro de ponderação. É a chamada análise-por-síntese.

Portanto, o que é enviado ao decodificador, apresentado na Fig. 6(b), não é a seqüência de excitação em si, mas os ganhos e índices que identificam as duas seqüências nos dicionários, que também existem no decodificador, implicando assim em uma boa qualidade de voz a uma baixa taxa de bits.

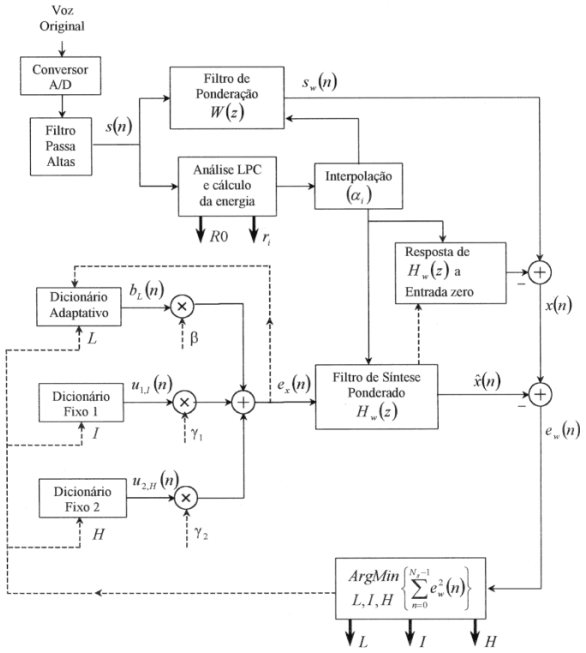
8. Vector Sum Excited Linear Predictive: VSELP

A codificação linear preditiva excitada por soma e vetores, ou VSELP, é o algoritmo de codificação de voz utilizado nos sistemas TDMA IS-54B e IS-136, de telefonia móvel celular. O IS-136 ainda pode usar outros algoritmos, como o ACELP, descrito no item 9.

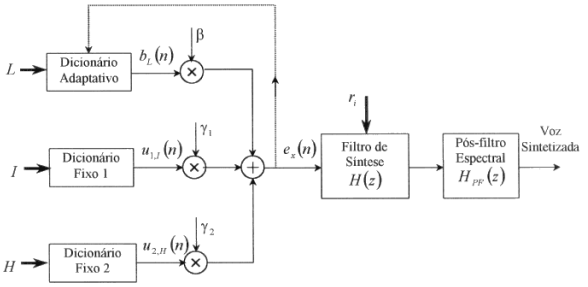
O codificador de voz VSELP é uma variação do algoritmo CELP, e foi projetado para suprir suas deficiências, ao atingir três objetivos: alta qualidade de voz, complexidade computacional modesta e robustez a erros de canal.

O VSELP conta com um dicionário adaptativo, como o CELP, mas o dicionário fixo é substituído por dois dicionários de código organizados com uma estrutura pré-definida, evitando a procura por "força-bruta". Isso reduz significativamente o tempo necessário para busca da palavra-código ótima, além de trazer alta qualidade de voz e uma maior robustez a erros de canal, mantendo baixa complexidade, a uma taxa de 7,95 kbit/s.

A Fig. 7(a) mostra o diagrama de blocos de um codificador VSELP, que tem sua seqüência de excitação formada pela combinação linear de 3 seqüências, obtidas uma de cada dicionário.



(a) Codificador



(b) Decodificador

Fig. 7 Diagrama de blocos do VSELP [3]

Como no CELP, o dicionário adaptativo é responsável pela predição a longo termo, introduzindo a periodicidade encontrada no sinal de voz (*pitch*). Cada um dos dois dicionários fixos do VSELP contém 128 vetores de 40 amostras, formados a partir de combinações lineares de 7 vetores base.

A busca nestes dicionários é realizada logo após a determinação do atraso ótimo do dicionário adaptativo (L) e é feita de modo seqüencial. Primeiro determina-se a melhor seqüência do primeiro dicionário estruturado, levando-se em conta a seqüência já escolhida do dicionário adaptativo. Em seguida, determina-se a melhor seqüência do segundo dicionário estruturado, considerando-se agora as duas seqüências previamente escolhidas.

Envia-se ao decodificador, mostrado na Fig. 7(b), o atraso do dicionário adaptativo (L), os índices dos vetores dos dicionários fixos (I e H), os respectivos ganhos (β, γ_1 e γ_2), e os parâmetros LPC do filtro de síntese. Assim o receptor gera a excitação fazendo uma combinação linear das três seqüências de excitação especificadas e a utiliza como sinal de entrada no filtro LPC, que introduz a correlação a curto termo, gerando a voz sintetizada.

Devido à estruturação dos dicionários utilizados no VSELP, um procedimento de busca sub-ótimo é conseguido, o que mantém a carga computacional baixa o suficiente para implementação em tempo real com os DSP's disponíveis. Aliam-se a isso os ganhos obtidos na qualidade de voz e na robustez a erros de canal, além da ainda relativamente baixa taxa de bits, que justificam a grande utilização deste codificador nos sistemas móveis atuais.

9. Algebraic Code Excited Linear Predictive: ACELP

Como já mencionado no item 8, o sistema TDMA IS-136 pode utilizar vários tipos de codificação de voz, visando manter compatibilidade com sistemas anteriores, enquanto oferece maior qualidade de voz aos consumidores. Com o desenvolvimento das tecnologias de compressão de voz, melhores codificadores se tornaram disponíveis. Por isso, além do VSELP, original do IS-54B, o IS-136 pode usar o codificador de voz do IS-641, que é um algoritmo EFR (*Enhanced Full Rate*) de codificação linear preditiva excitada por código algébrico, o ACELP [14].

Padronizado pelo ITU-T SG15 na recomendação G.723, o ACELP tem qualidade de voz igual, se não melhor que o ADPCM a 32 kbit/s, e transmite voz de com melhor qualidade e de forma menos sensível a ruído que o VSELP, à mesma taxa de 7,95 kbit/s.

A diferença básica entre o ACELP, cujo diagrama de blocos é apresentado na Fig. 8, e o CELP está no dicionário fixo, que é baseado em uma estrutura algébrica. Assim como no CELP, a seqüência de excitação é obtida como uma combinação linear de duas seqüências, uma obtida do dicionário fixo e outra do dicionário adaptativo. Porém, a análise-por-síntese é mais eficiente que no CELP.

A busca no dicionário adaptativo é realizada somente ao redor de uma região limitada por um período de *pitch*, estimado a cada 10ms por duas análises diferentes: *open-loop* e *closed-loop* (esses dois processos serão discutidos no item 10). Com esse parâmetro, seleciona-se a melhor seqüência com uma resolução de 1/3 para o atraso de *pitch*.

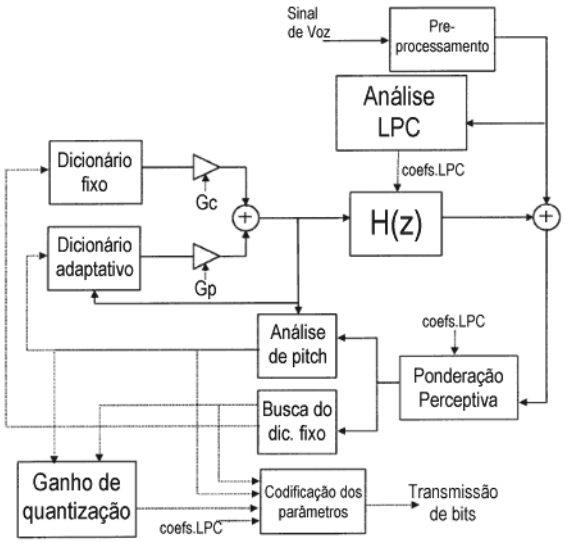


Fig. 8 Diagrama de blocos do codificador ACELP [6]

Como já mencionado, o dicionário fixo é baseado em uma estrutura algébrica. Neste dicionário, cada vetor de palavra código contém quatro pulsos diferentes de zero. Cada pulso pode ter uma amplitude de +1 ou –1 e assumir as posições dadas na Tabela 2. Como o dicionário é estruturado, é possível se realizar uma busca mais eficiente. No caso ela é realizada por quatro laços em série, e uma tática de pesquisa focalizada é utilizada para tornar mais simples o procedimento de busca.

Tabela 2 Estrutura do dicionário fixo [6]

Pulso	*Sinal*	*Posição*
i_0	s_0: ±1	m_0: 0, 5, 10, 15, 20, 25, 30, 35
i_1	s_1: ±1	m_1: 1, 6, 11, 16, 21, 26, 31, 36
i_2	s_2: ±1	m_2: 2, 7, 12, 17, 22, 27, 32, 37
i_3	s_3: ±1	m_3: 3, 8, 13, 18, 23, 28, 33, 38 4, 9, 14, 19, 24, 29, 34, 39

A estrutura algébrica do dicionário fixo também garante ao ACELP sensibilidade reduzida a erros no canal e voz codificada de melhor qualidade. A ref. [11] apresenta uma comparação da qualidade de sinais de voz codificados com ACELP e VSELP e também no AMPS (modulação analógica FM), a diferentes condições de relação sinal/ruído. Os exemplos fornecidos evidenciam a robustez do codificador ACELP, uma vez que este mantém sua inteligibilidade e apresenta qualidade de voz relativamente boa mesmo aos mais baixos níveis de C/N.

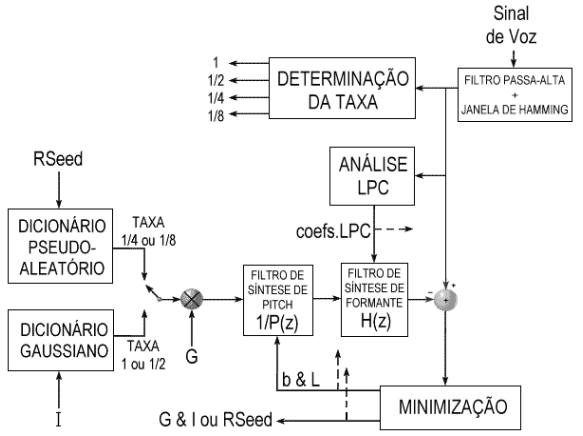
A ref. [6] apresenta ainda uma proposta de um algoritmo de taxa variável para o ACELP, conseguindo diminuir a taxa média de bits para 4,4 kbit/s, sem degradação da qualidade de voz, ao levar em consideração a atividade de voz.

10. Qualcomm Code Excited Linear Predictive: QCELP

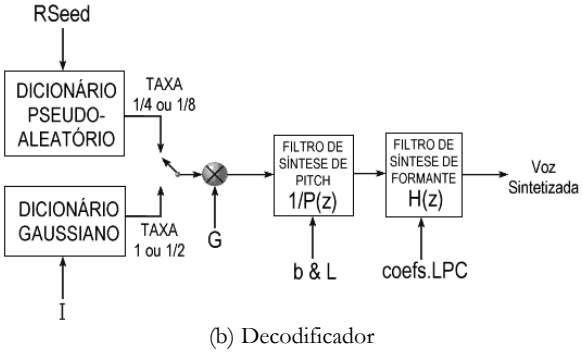
O sistema IS-95 de telefonia móvel CDMA utiliza o codificador de voz QCELP, que é um padrão em sistemas de espectro espalhado, podendo utilizar quatro taxas de bit diferentes - 9,6 / 4,8 / 2,4 / 1,2 kbit/s – de acordo com o segmento de sinal que está sendo codificado. A Qualcomm, empresa que desenvolveu o algoritmo QCELP, também produz equipamentos com os quais pode-se fixar a taxa de bits em 9,6 ou 4,8 kbit/s, o que pode ser desejado em função do tráfego no sistema CDMA.

Além da taxa variável, a diferença básica do QCELP para outros algoritmos baseados no CELP, está na forma como a correlação a longo termo é codificada. Já a predição a curto termo é realizada com um filtro LPC, como nos sistemas já abordados.

O primeiro passo para a codificação, cujo diagrama de blocos simplificado é apresentado na Fig. 9(a), é um pré-processamento, que consiste basicamente em um filtro passa-altas que retira a componente DC do sinal, seguido de um janelamento de Hamming que reduz o efeito da divisão do sinal em blocos.



(a) Codificador

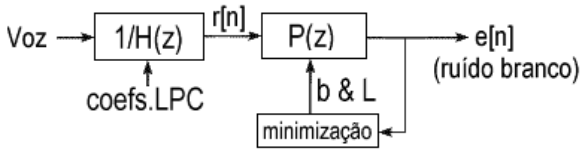


(b) Decodificador

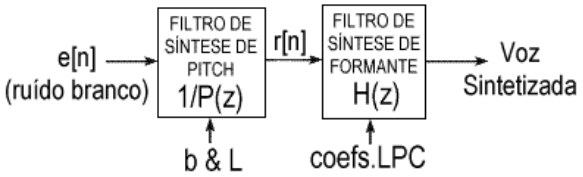
Fig. 9 Diagrama de blocos simplificado do QCELP

Segue-se a isso a análise LPC, que determina os parâmetros do filtro de síntese, e, em paralelo, um procedimento para determinação da taxa de dados, que analisa as características do quadro para decidir se este pode ser codificado a uma taxa reduzida sem afetar a qualidade de voz. Para sons surdos, utiliza-se ¼ da taxa (2,4 kbit/s) e para pausas ou ruído de fundo utiliza-se $^1/_8$ da taxa (1,2 kbit/s). Para segmentos sonoros, a taxa máxima (9,6 kbit/s) é usada em quadros transitórios, com periodicidade reduzida, ou que não são bem modelados, os quais requerem taxa máxima para que se consiga boa qualidade de voz. Para segmentos sonoros bem modelados, estacionários e periódicos é usada ½ da taxa (4,8 kbit/s).

A busca do período de *pitch*, que modela a correlação a longo termo, pode ser realizada através de dois modelos: *open-loop* e *closed-loop*. Em um modelo *open-loop*, como ilustrado na Fig. 10(a), retira-se a correlação a curto termo do sinal de voz ao passá-lo pelo filtro LPC inverso. A seguir, o sinal residual *r[n]* entra no filtro preditor de *pitch* P(z), que tenta retirar a correlação a longo termo, produzindo o *pitch* residual *e[n]*. O filtro P(z) tem dois parâmetros, ganho de *pitch* b e atraso de *pitch* L, que devem ser otimizados para que a energia média de *e[n]* seja minimizada, resultando em um sinal cujas características lembram as de um ruído branco.



(a) Processo *open-loop*



(b) Processo reverso

Fig. 10 Modelo básico de reconstrução da voz usado no algoritmo QCELP

O modelo básico de reconstrução da voz usado no QCELP é o *closed-loop*, que se baseia no processo reverso, mostrado na Fig. 10(b). Usando-se ruído branco como fonte de excitação, passando-a por um filtro de síntese de *pitch* 1/P(z) e na seqüência por um filtro de síntese de formante H(z), que introduzem respectivamente a correlação a longo e a curto termos, resultando em o que se espera ser, voz sintetizada. A busca pelos parâmetros b e L é feita de forma a minimizar a diferença entre o sinal original e o sinal

sintetizado obtido a partir de uma excitação que se aproxima de ruído branco.

Essa excitação é obtida a partir de um dicionário pseudo-aleatório ou de um dicionário gaussiano, para taxas de ¼ e 1/8 ou 1 e ½ respectivamente. O dicionário gaussiano tem 128 vetores de 128 amostras e dois parâmetros, índice I e ganho G. O dicionário pseudo-aleatório usado é chamado de dicionário circular com deslocamento unitário, onde o vetor de uma linha é o vetor da linha anterior descolado de uma amostra. O parâmetro que identifica o deslocamento é o RSeed, e não há parâmetro de ganho.

O decodificador, que é apresentado na Fig. 9(b), recebe então os coeficientes do filtro LPC, os parâmetros do filtro de síntese de *pitch* (b e L), e os parâmetros da excitação (RSeed ou G e I). Com essa informação ele monta os filtros de síntese e utiliza a excitação especificada para sintetizar o sinal de voz.

Portanto, o QCELP consegue sintetizar o sinal de voz com taxas médias de bits menores que os demais codificadores, sem requerer um dicionário adaptativo para modelar a correlação a longo termo. No entanto, a qualidade obtida é inferior à dos demais codificadores aqui abordados, mesmo à taxa máxima.

11. Conclusão

As técnicas abordadas neste trabalho são todas derivadas do codificador paramétrico LPC, tendo o modelo do trato vocal baseado em um filtro que introduz a correlação a curto termo presente no sinal de voz. A diferença básica está na forma como é gerada a excitação, responsável por introduzir a correlação a longo termo e descrever as pequenas diferenças percebidas entre períodos tonais sucessivos do sinal a ser codificado.

No RPE-LTP a informação da correlação a longo termo é obtida por um filtro LTP e as diferenças percebidas entre períodos tonais são descritas pela seqüência de excitação residual enviada ao decodificador. Nos codificadores VSELP e ACELP o filtro LTP é substituído por um dicionário adaptativo e a seqüência residual é obtida a partir de dicionários de código estruturados. No QCELP não há dicionário adaptativo, mas um filtro de síntese de *pitch*. A seqüência residual, no entanto, é obtida a partir de dicionários fixos como nos demais codificadores derivados do CELP.

Essas diferentes soluções para descrever a excitação a ser usada no decodificador implicam em diferentes taxas de bits e diferentes qualidades de voz, como mostra a Tabela 3. O índice MOS (*mean opinion score*) apresentado é obtido a partir de testes subjetivos onde várias pessoas ouvem determinadas seqüências de diálogo ou duplas de palavras rimadas, com os quais obtém-se uma avaliação em uma escala de 1 a 5 pontos para a inteligibilidade das amostras, onde 1 é muito ruim e 5 é excelente. Nota-se, portanto, que o RPE-LTP e o VSELP já não são as melhores soluções em termo de taxa de bits nem de qualidade de voz, e vêm sendo substituídos nos novos sistemas TDMA pelo codificador ACELP, que ainda é mais robusto a erros devido à estrutura algébrica do seu dicionário fixo. Nos sistemas de espectro espalhado se usa o QCELP, que apesar de ter uma qualidade de voz inferior ao ACELP, oferece uma taxa de bits, em média, menor.

Tabela 3 Comparação entre as técnicas abordadas

Técnica	*Padrão*	*Taxa de bits*	*Excitação*	*Modelo do Trato Vocal*	*Qualidade (MOS rating)*
RPE-LTP	GSM	13 kbit/s	pulsos regulares e análise LTP	filtro LPC (análise STP)	3,54
VSELP	IS-54B e IS-136	7,95 kbit/s	1 dicionário adaptativo 2 dicionários fixos estruturados	filtro LPC (análise STP)	3,8
ACELP	IS-136	7,95 kbit/s	1 dicionário adaptativo 1 dicionário com estrutura algébrica	filtro LPC (análise STP)	3,9
QCELP	IS-95	1,2/2,4 kbit/s	1 dicionário pseudo-aleatório	filtro LPC (análise STP)	3,45 (taxa máxima)
		4,8/9,6 kbit/s	1 dicionário gaussiano		

Reconhecimento

Este trabalho foi orientado pelo Prof. Paulo H. P. de Carvalho como projeto final para a disciplina Tópicos Especiais em Telecomunicações - Sistemas de Comunicações Móveis, ministrada no período 1º/00.

Além do Prof. Paulo, colaboraram também o Prof. Lúcio M. Silva e o Prof. Sebastião Nascimento, indicando e fornecendo referências bibliográficas e compartilhando um pouco do seu vasto conhecimento sobre o assunto em questão.

Referências

[1] ALENCAR, Marcelo S., "Telefonia Digital", 2ª Edição, Editora Érica Ltda., São Paulo-SP, 1999.

[2] RAPPAPORT, Theodore S.,"Wireless Communications: Principles & Practice", Prentice Hall, New Jersey, 1996.

[3] FIACADOR, Altair R., "Estudo e Simulação do Codificador de Voz VSELP do Padrão IS-136", Relatório de Projeto Final, Universidade de Brasília, Dezembro de 1999.

[4] UNIVERSITY OF SOUTHAMPTON, UK, "Commonly Used Speech Codecs".

http://www-mobile.ecs.soton.ac.uk/speech_codecs/common_classes.html

[5] MOTOROLA INC., "Vector Sum Excited Linear Prediction (VSELP) 13000 Bit Per Second Voice Coding Algorithm Including Error Control for Digital Cellular - Technical Description". Motorola Inc, 18 de outubro de 1989.

[6] CHUNG, Woosung e Sangwon KANG, "Design of a Variable Rate Algorithm for the CS-ACELP Coder", IEIC Trans. Inf. & Syst., Vol. E82-D, no.10, Outubro de 1999, pp. 1364-1371.

[7] WANG, Derek Q., "QCELP Vocoders in CDMA Systems Design". http://www.csdmag.com/main/1999/04/9904feat3.htm

[8] LINCOM CORPORATION, "RPE-LTP (13 Kbps) - Regular-Pulse Excited LPC with a Long-Term Predictor", 1998. http://www.lincom-asg.com/ssadto/rpe-ltp.html

[9] SILVA, Lúcio M., "Contribuições para a Melhoria da Codificação CELP a Baixas Taxas de Bits", Tese de Doutorado, Pontifícea Universidade Católica do Rio de Janeiro, 1996.

[10] RABINER, L. R., "Applications of Voice Processing to Telecommunications", Proc IEEE vol. 82, 1994, pp. 197-228.

[11] UNIVERSAL WIRELESS COMMUNICATIONS CONSORTIUM, "Comparison of ACELP, VSELP and AMPS Voice Quality Under Low Carrier-to-Noise RF Conditions". http://www.uwcc.org/edge/papers/acelp.html

[12] SKLAR, Bernard, "Digital Communications: Fundamentals and Applications", 1a Edição, Prentice Hall, New Jersey, 1988.

[13] CARNEGIE MELLON UNIVERSITY, "Speech at CMU Web Page". http://fife.speech.cs.cmu.edu/speech/

[14] HARTE, Lawrence J., Adrian D. SMITH e Charles A. JACOBS, "IS-136 TDMA Technology, Economics and Services", Artech House, Boston - London, 1998.

[15] HUGHES NETWORK SYSTEMS, "News Releases: HNS Offers Enhanced Quality Voice Via Commercially Available IS-136 TDMA". http://www.hns.com/news/pressrel/snd_pres/p030397.htm