Capítulo 3

Codificação de Voz

Francisco Madeiro, Waslon Terllizzie Araújo Lopes

3.1 Introdução

A compressão de sinais, cujo objetivo fundamental é reduzir o número de bits necessários para representar adequadamente os sinais (voz, imagem, áudio, vídeo), desempenha um papel importante em aplicações que necessitam minimização dos requisitos de largura de faixa e/ou de capacidade de armazenamento, tais como: sistemas multimídia, redes digitais de serviços integrados, videoconferência, sistemas de resposta vocal, correio de voz (*voicemail*), difusão de música, facsímile de alta resolução, televisão de alta definição (HDTV, *high definition television*), telefonia móvel, sistemas de armazenamento de imagens médicas e de impressões digitais e transmissão de imagens de sensoriamento remoto obtidas por satélites.

Apesar de alguns sistemas não apresentarem grandes limitações de largura de faixa, como é o caso das redes de comunicações por fibra ótica, e embora a evolução tecnológica esteja continuamente contribuindo para o surgimento de memórias com grande capacidade de armazenamento, a compressão de sinais desempenha um papel importante, devido a uma série de fatores, tais como (Jayant and Noll, 1984; Gersho and Gray, 1992):

- a grande utilização dos sistemas multimídia tem levado ao aumento da demanda no tocante ao armazenamento de voz, música, imagens, vídeo e dados em forma comprimida;
- um maior número de canais de comunicação pode ser multiplexado em sistemas de faixa larga, por meio do uso de técnicas de compressão para reduzir os requisitos de largura de faixa de cada sinal a ser multiplexado;
- nos sistemas de reconhecimento de fala e nos sistemas de resposta vocal, vocabulários maiores podem ser armazenados por meio da redução dos requisitos de memória necessários para cada padrão de voz;
- nas redes digitais de serviços integrados (ISDN, *integrated services digital networks*), as técnicas de compressão permitem uma integração eficiente de sinais e dados;
- em telefonia móvel celular, a largura de faixa é severamente limitada, o que tem motivado muitos estudos em compressão de voz.

A Figura 3.1 ilustra o princípio básico da compressão de sinais. A informação dos sinais é composta das seguintes partes (Jayant and Noll, 1984):

- componente relevante;
- componente irrelevante, que corresponde à parcela supérflua da informação, de que o destinatário não necessita;

- componente não redundante;
- componente redundante, que não necessita ser transmitida, uma vez que o receptor tem condições de reconstituí-la automaticamente.

O objetivo é reduzir as componentes redundante e irrelevante da informação, transmitindo apenas o que é imprescindível para o receptor, ou seja, a parte essencial da informação.

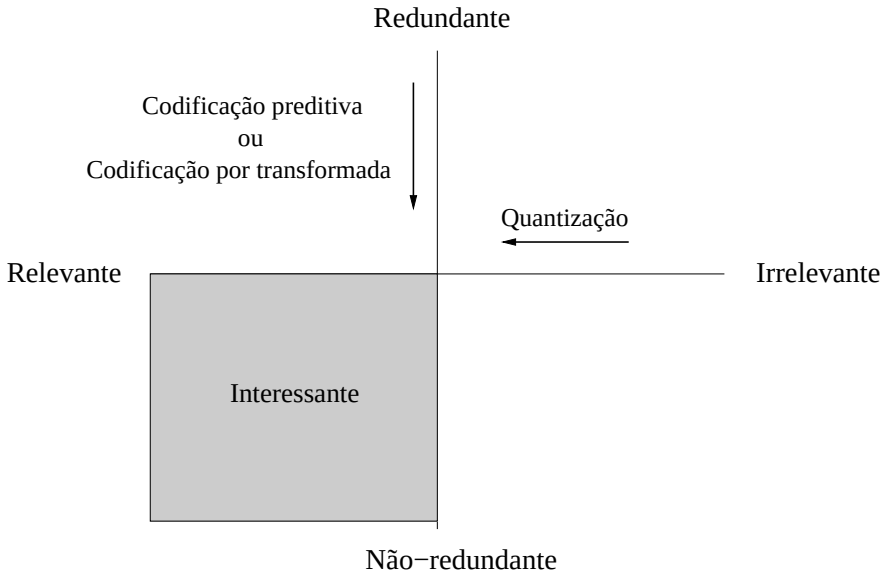


Figura 3.1: Princípio da compressão de sinais.

A parcela irrelevante da informação é reduzida por meio do processo de quantização, que introduz perda de caráter irreversível. A parcela redundante da informação é reduzida por meio de técnicas de predição e de transformação do sinal, que apresentam caráter reversível.

3.2 Codificação de Sinal – Preliminares

Codificação de sinal é o processo de representar um sinal de informação de tal maneira a atingir um objetivo de comunicação desejado, tal como conversão analógico-digital, transmissão a baixas taxas de bits ou criptografia de mensagens. Na literatura, os termos codificação de fonte, codificação digital, compressão de dados, compressão de largura de faixa e compressão de sinal são todos utilizados para referir-se a técnicas usadas para obter uma representação digital compacta de um sinal. Uma questão importante em se tratando de compressão de sinais é a preocupação com o receptor humano ao final do processo de comunicação (Figura 3.2) (Jayant, 1992; Jayant et al., 1993).

A Figura 3.3 apresenta um sistema de comunicação digital. Enquanto o codificador de fonte visa minimizar a taxa de bits necessária para representação adequada de um sinal de entrada, o *mo*dulador-*dem*odulador (*modem*) procura maximizar a taxa de bits que pode ser sustentada em um dado canal ou meio de armazenamento sem causar um nível inaceitável de probabilidade de erro de bit. Na codificação de fonte, a taxa de bits é medida em bits por amostra ou bits por segundo (usualmente denotada por b/s ou ainda por bps). Na modulação, é medida em bits por segundo por hertz ($b/s/\mathrm{Hz}$). Os blocos de codificação de canal adicionam redundância à seqüência de bits visando proteção contra erros. Nos sistemas de modulação codificada, as operações de codificação de canal e modulação são integradas para o propósito de maior eficiência geral. Os processos de codificação de fonte e de codificação de canal também podem ser integrados.

Ao longo deste capítulo, o termo codificação de voz será exaustivamente utilizado, referindo-se especificamente à codificação de fonte.

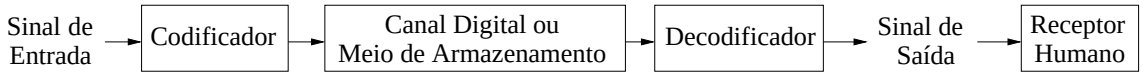


Figura 3.2: Codificação digital para compressão de sinais.

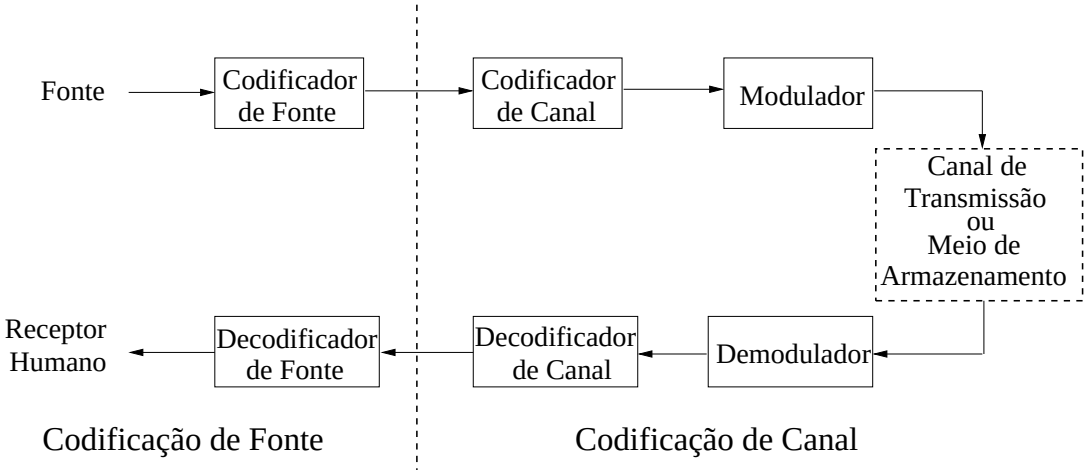


Figura 3.3: Diagrama de blocos de um sistema de comunicação digital.

3.3 Amostragem

No processo de amostragem um sinal contínuo no tempo é transformado em um sinal discreto no tempo. A taxa de amostragem deve ser suficientemente alta para que o sinal contínuo no tempo possa ser reconstruído a partir de suas amostras com precisão adequada. O Teorema da Amostragem, de Nyquist, que é a base para determinar a taxa de amostragem adequada para um dado sinal, é abordado a seguir.

3.3.1 Teorema da Amostragem

Um sinal $f(t)$ limitado em faixa em B Hz (ou seja, $F(w) = 0$ para $|w| > 2\pi B$) pode ser reconstruído exatamente (sem erro) a partir de suas amostras tomadas a uma taxa $R_S > 2B$ Hz (amostras por segundo). Em outras palavras, a freqüência de amostragem mínima é $f_S = 2B$ Hz (Lathi, 1988).

Considerações

Seja um sinal $f(t)$ (Figura 3.4) cujo espectro é limitado em faixa em B Hz. A amostragem do sinal $f(t)$ pode ser realizada multiplicando $f(t)$ por um trem de impulsos, $\delta_{T_S}(t)$, que se repetem a cada T_S segundos, em que $T_S = 1/f_S$. Isto leva ao sinal amostrado $\hat{f}(t)$ apresentado na Figura 3.4, dado por

$$\hat{f}(t) = f(t)\delta_{T_S}(t) = \sum_n f(nT_S)\delta(t - nT_S). \tag{3.1}$$

Como o trem de impulsos $\delta_{T_S}(t)$ é um sinal periódico de período T_S, ele pode ser expresso por meio de série de Fourier, da forma

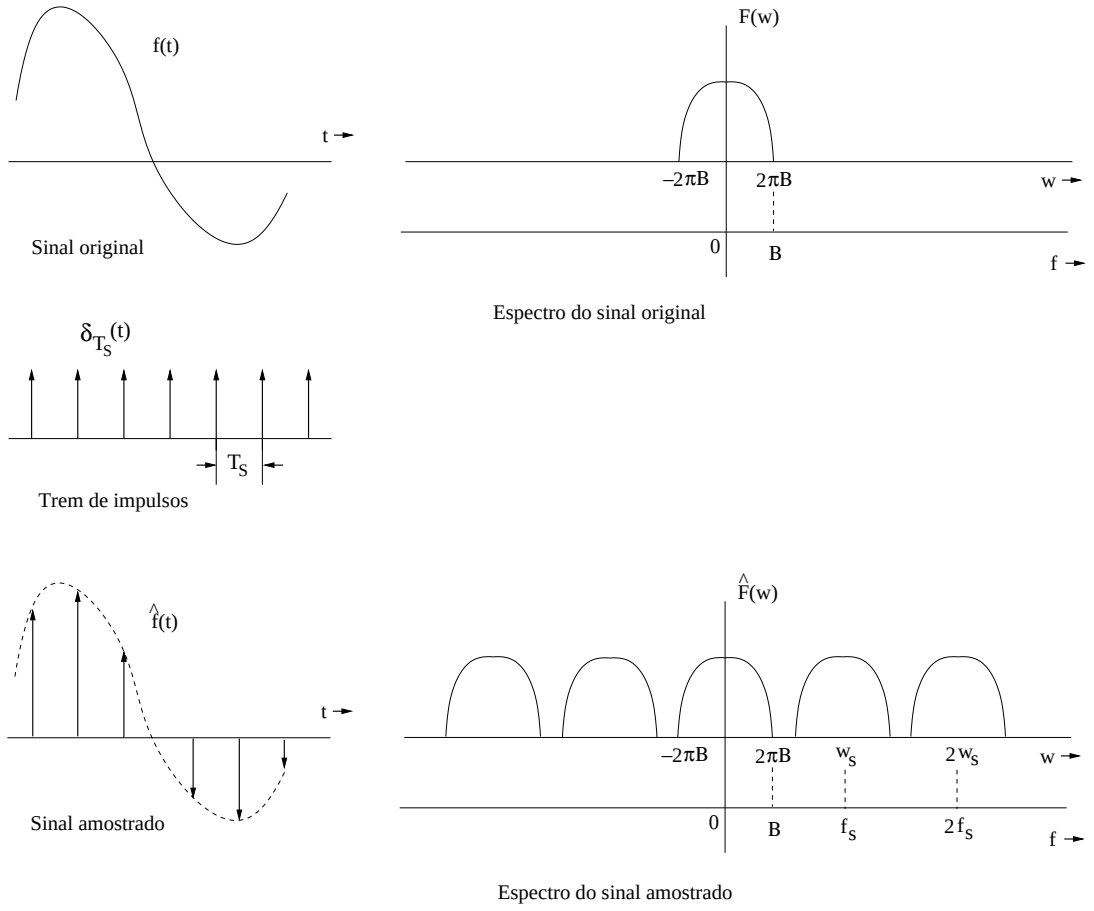


Figura 3.4: Sinais original e reconstruído, com os espectros correspondentes.

$$\delta_{T_S}(t) = \frac{1}{T_S}[1 + 2\cos w_S t + 2\cos 2w_S t + 2\cos 3w_S t + \cdots], \tag{3.2}$$

com $w_S = \frac{2\pi}{T_S} = 2\pi f_S$.

Portanto,

$$\hat{f}(t) = \frac{1}{T_S}[f(t) + 2f(t)\cos w_S t + 2f(t)\cos 2w_S t + 2f(t)\cos 3w_S t + \cdots]. \tag{3.3}$$

Desta forma, $\hat{F}(w)$, a transformada de Fourier de $\hat{f}(t)$, é dada por

$$\hat{F}(w) = \frac{1}{T_S}\sum_{n=-\infty}^{\infty} F(w - nw_S). \tag{3.4}$$

O espectro $\hat{F}(w)$ consiste de $F(w)$ repetindo-se periodicamente com período $w_S = 2\pi/T_S$ rad/s, ou $f_S = 1/T_S$ Hz, como mostra a Figura 3.4.

Se desejarmos recuperar $f(t)$ a partir de $\hat{f}(t)$, devemos ser capazes de recuperar $F(w)$ a partir de $\hat{F}(w)$. Isso é possível se não houver sobreposição entre ciclos sucessivos de $\hat{F}(w)$. A Figura 3.4 mostra que a condição

para isso é

$$f_S > 2B. \tag{3.5}$$

O intervalo de amostragem é $T_S = 1/f_S$. Portanto,

$$T_S < \frac{1}{2B}. \tag{3.6}$$

Desta forma, com uma freqüência de amostragem f_S maior que duas vezes a largura de faixa B (em hertz), $\hat{F}(w)$ consistirá de repetições de $F(w)$ que não se sobrepõem. Assim, $f(t)$ pode ser recuperado a partir de suas amostras passando o sinal amostrado $\hat{f}(t)$ por um filtro passa-baixa ideal com largura de faixa B Hz. A taxa de amostragem mínima, $f_S = 2B$, requerida para recuperar $f(t)$ a partir de suas amostras é denominada taxa de Nyquist ou freqüência de Nyquist para $f(t)$, e o intervalo de amostragem correspondente, $T_S = 1/2B$, é chamado intervalo de Nyquist para $f(t)$ (Lathi, 1988).

Para aplicações em telefonia, a freqüência de amostragem adotada internacionalmente é $f_S = 8$ mil amostras por segundo, abreviada para 8 k amostras/s. O sinal de voz é então quantizado, como será visto posteriormente, para 256 níveis distintos. Cada nível corresponde a um código de 8 bits ($2^8 = 256$). Após a codificação, o sinal é transmitido a uma taxa de 8 k amostras/s$\times$8 bits/amostra = 64 kbits/s e ocupará uma banda passante de aproximadamente 64 kHz.

Se a freqüência de amostragem w_S for menor que $2\pi B$ haverá superposição dos espectros e, portanto, perda de informação. À medida que w_S se torna menor que $2\pi B$, a taxa de amostragem se torna mais baixa, havendo perda parcial de informação. Portanto, a freqüência ótima de amostragem é $w_S = 2\pi B$, conhecida como freqüência de amostragem de Nyquist.

Caso a freqüência de amostragem seja inferior à freqüência de Nyquist, o sinal não poderá ser recuperado completamente, pois haverá superposição espectral, resultando em distorção nas freqüências mais altas. Esse fenômeno é conhecido como *aliasing.* Por outro lado, um aumento na freqüência de amostragem, além da freqüência de Nyquist, implicará separação entre os espectros superior à necessária para a recuperação do sinal.

3.4 O Desempenho de um Sistema de Compressão de Sinais

O problema geral em compressão de sinal é minimizar a taxa de bits na representação digital do sinal, mantendo os níveis requeridos de qualidade do sinal, complexidade de implementação e retardo de comunicação. Cada um desses aspectos é abordado a seguir.

3.4.1 Qualidade de Sinais Reconstruídos

Um dos grandes desafios em codificação digital de sinais é a concepção e o desenvolvimento de metodologias de avaliação de qualidade de sinais reconstruídos (obtidos com a aplicação de técnicas de compressão). De forma geral, as medidas utilizadas para avaliação da qualidade de sinais enquadram-se em duas classes: medidas de qualidade subjetivas e medidas de qualidade objetivas. As primeiras baseiam-se em comparações (realizadas por meio de testes de escuta ou de visualização), entre o sinal original e o sinal processado, realizadas por um grupo de pessoas, que subjetivamente classificam a qualidade do sinal processado segundo uma escala pré-determinada. As medidas objetivas, por sua vez, baseiam-se numa comparação matemática direta entre os sinais original e processado (Deller Jr. et al., 1993).

Para serem úteis, as medidas de qualidade objetivas devem apresentar no mínimo duas características. Primeiramente, devem ter significado subjetivo, no sentido de que pequenas e grandes variações das medidas objetivas devem significar pequenas e grandes variações da qualidade subjetiva dos sinais reconstruídos, respectivamente – devem apresentar, portanto, uma correlação forte, positiva ou negativa, com resultados de avaliação subjetiva. Segundo, devem ser matematicamente tratáveis e facilmente implementáveis.

Medidas de qualidade subjetivas são utilizadas para avaliar de forma plena, definitiva, a qualidade de algoritmos/técnicas de codificação de sinais. Os testes subjetivos, contudo, são de difícil realização, uma vez que requerem a disponibilidade de um grande número de avaliadores (entre leigos, especialistas e possíveis usuários do sistema de codificação de sinais), envolvem grande volume de sinal processado e necessitam da disponibilidade de laboratórios com condições adequadas ao desenvolvimento das avaliações subjetivas, enfim, são bastante dispendiosos de tempo, implicando custo elevado de realização. Devido a esses problemas, as medidas de qualidade objetivas, por serem pouco dispendiosas de tempo, desempenham um papel importante no processo de avaliação de qualidade de sinais reconstruídos e constituem ferramenta valiosa no processo de ajuste de parâmetros de algoritmos/técnicas de compressão.

A avaliação de qualidade de sinais de voz e imagem tem sido objeto de estudo de diversas pesquisas, tendo sido abordada suscinta ou profundamente em diversos trabalhos (*e.g.* (Eskicioglu and Fischer, 1995; Dimolitsas, 1991)).

A seguir serão abordadas algumas medidas de qualidade utilizadas para a avaliação de técnicas de compressão de sinais de voz e imagem. Serão descritos o escore médio de opinião (MOS, *mean opinion score*) e os testes de preferência, que constituem alternativas para avaliação de qualidade subjetiva. Também serão apresentadas a relação sinal-ruído (SNR, *signal-to-noise ratio*), a relação sinal-ruído segmental (SNRseg, *segmental signal-to-noise ratio*) e a distorção espectral (SD, *spectral distortion*), que são medidas de qualidade objetivas utilizadas para avaliação de qualidade dos sinais de voz. Por fim, será apresentada a relação sinal-ruído de pico (PSNR, *peak signal-to-noise ratio*), que constitui uma das medidas objetivas mais utilizadas para avaliação da qualidade de imagens reconstruídas.

Escore Médio de Opinião

Uma medida subjetiva bastante utilizada para avaliação de desempenho de sistemas de compressão de voz e imagem denomina-se escore médio de opinião (MOS). No teste MOS, cada avaliador atribui um escore de qualidade ao sinal reconstruído, segundo a escala graduada apresentada na Tabela 3.1 (Jayant and Noll, 1984; Deller Jr. et al., 1993). É calculada a média aritmética dos escores obtidos e determinado o valor final da avaliação, ou seja,

$$MOS = \frac{1}{L}\sum_{l=1}^{L} s_l, \tag{3.7}$$

em que L é o número de avaliadores utilizados no teste e s_l é o escore atribuído pelo l-ésimo avaliador.

Tabela 3.1: Escala para o teste MOS.

Escore (s)	Qualidade
5	Excelente
4	Boa
3	Razoável
2	Pobre
1	Ruim

Testes de Preferência

Os testes de preferência são realizados por comparação entre pares de sinais.

Uma forma de realização desse tipo de avaliação subjetiva consiste em atribuir um conceito de acordo com três possíveis resultados de comparação, ou seja (Aguiar Neto, 1995):

- Conceito A - A qualidade do primeiro sinal é melhor do que a do segundo;
- Conceito B - A qualidade do segundo sinal é melhor do que a do primeiro;
- Conceito C - A qualidade de ambos sinais não se distingue.

Outra forma de realização de testes de preferência, muito comum para a avaliação de imagens, consiste em proceder a comparação com relação a uma imagem de referência, utilizando uma escala de graduação com valores que variam de um a cinco, na qual cada valor correspondente a um conceito obtido do processo de comparação (Aguiar Neto, 1995):

- Conceito 5 - A imagem sob teste tem qualidade muito superior à apresentada pela imagem de referência;
- Conceito 4 - A imagem sob teste tem qualidade um pouco superior à apresentada pela imagem de referência;
- Conceito 3 - A imagem sob teste tem a mesma qualidade da apresentada pela imagem de referência.
- Conceito 2 - A imagem sob teste tem qualidade um pouco inferior à apresentada pela imagem de referência;
- Conceito 1 - A imagem sob teste tem qualidade muito inferior à apresentada pela imagem de referência.

Em se tratando de codificação de voz, avaliações por meio de MOS são bem aceitas e algumas vezes complementadas com medidas de inteligibilidade, como por exemplo MRT (*modified rhyme test*) e DRT (*diagnostic rhyme test*) (Deller Jr. et al., 1993). Existem outras medidas de qualidade subjetiva de voz, como IAJ (*isometric absolute judgment*), QUART (*quality acceptance rating test*) e DAM (*dianostic acceptability measure*) (Deller Jr. et al., 1993).

O diagnóstico chamado teste de verso (DRT) (Bellamy, 1991) é uma maneira de determinar a inteligibilidade de um sistema de voz por meio de uma porcentagem de palavras corretas, reconhecidas de uma lista de pares de palavras padronizadas. O teste DRT com interferência mede a porcentagem de reconhecimento de palavras quando um ruído de acompanhamento é adicionado ao teste de palavras antes da codificação. O diagnóstico de medição de aceitabilidade (DAM) (Bellamy, 1991) considera a inteligibilidade e a aceitabilidade subjetiva, usando procedimentos que eliminam grande parte da dependência da preferência pessoal dos ouvintes. As contagens são normalizadas de 0 a 100.

Relação Sinal-Ruído (SNR)

Sejam $x(n)$ o sinal original, $y(n)$ o sinal processado e $e(n) = x(n) - y(n)$ o sinal erro no instante de tempo n.

A energia contida no sinal original é

$$E_x = \sum_n x^2(n). \tag{3.8}$$

A energia contida no sinal erro é

$$E_e = \sum_n e^2(n) = \sum_n [x(n) - y(n)]^2. \tag{3.9}$$

A medida SNR resultante, expressa em dB, é dada por

$$\text{SNR} = 10 \log_{10} \frac{E_x}{E_e} = 10 \log_{10} \frac{\sum\limits_n x^2(n)}{\sum\limits_n [x(n) - y(n)]^2}. \tag{3.10}$$

Relação Sinal-Ruído Segmental (SNRseg)

Apesar da simplicidade matemática, a medida SNR apresenta uma limitação incômoda: a igual ponderação de todos os erros no domínio do tempo. Por essa razão, um indesejável valor elevado de SNR pode ser obtido se uma seqüência de fala apresentar alta concentração de segmentos vocais, sonoros, de alta energia, uma vez que o efeito do ruído é maior nos segmentos de baixa energia, como por exemplo os sons fricativos surdos.

Uma medida de qualidade mais refinada pode ser obtida se for tomada a média da relação sinal-ruído medida em curtos intervalos de tempo. É definida, então, a relação sinal-ruído segmental (ou razão sinal-ruído segmentar):

$$\text{SNRseg} = E[\text{SNR}(j)], \tag{3.11}$$

em que $\text{SNR}(j)$ denota a relação sinal-ruído (SNR) convencional para o j-ésimo segmento (janela de tempo) do sinal.

A medida SNRseg é formulada como

$$\text{SNRseg} = \frac{1}{J}\sum_{j=0}^{J-1} 10\log_{10}\left[\frac{\sum\limits_{n=m_j-N_A-1}^{m_j} x^2(n)}{\sum\limits_{n=m_j-N_A-1}^{m_j}[x(n)-y(n)]^2}\right], \tag{3.12}$$

em que $m_0, m_1, ..., m_{J-1}$ são os instantes finais para as J janelas de tempo, de N_A amostras, de comprimento típico de 15 a 25 ms (Deller Jr. et al., 1993; Aguiar Neto, 1995).

Segundo Voran (Voran, 1999), as medidas SNR e SNRseg podem fornecer uma indicação de qualidade subjetiva em alguns codificadores de voz que têm por objetivo a representação da forma de onda. Contudo, quando utilizadas em sistemas de codificação e transmissão mais gerais, SNR e SNRseg apresentam pouca correlação com resultados de avaliação subjetiva da qualidade de voz. Voran (Voran, 1999) atribui a popularidade de SNR e SNRseg a razões históricas (tratam-se de medidas clássicas, utilizadas há bastante tempo), à simplicidade dessas medidas, e ao fato de inexistirem medidas objetivas que tenham sido amplamente testadas e aceitas, de tal maneira que possam substituir SNR e SNRseg.

Como outras medidas de avaliação de qualidade objetiva de voz, podem ser citadas a LAR (*log-area ratio*), a medida *log-likelihood* de Itakura (Deller Jr. et al., 1993) e a distorção espectral, descrita a seguir.

Distorção Espectral

Para aplicações que requisitam codificação de voz a baixas taxas (como é o caso de sistemas de comunicações móveis), é imprescindível quantizar precisa e adequadamente os parâmetros LSF (*line spectral frequencies*), utilizando o menor número de bits possível. O desenvolvimento de métodos de codificação LSF tem sido objeto de interesse de muitos pesquisadores, e. g. (Paliwal and Atal, 1993; LeBlanc et al., 1993; Eriksson et al., 1999).

A qualidade da quantização dos parâmetros LSF é avaliada por meio da distorção espectral (SD), definida como

$$\text{SD} = \left[\frac{1}{F_S}\int_0^{F_S}[10\log_{10} S(f) - 10\log_{10}\hat{S}(f)]^2 df\right]^{1/2}, \tag{3.13}$$

em que $S(f)$ e $\hat{S}(f)$ denotam, respectivamente, a envoltória espectral original e quantizada.

Relação Sinal-Ruído de Pico (PSNR)

Dentre as diversas medidas objetivas utilizadas para avaliação de qualidade de imagens (Eskicioglu and Fischer, 1995), a relação sinal-ruído de pico (ou razão pico-ruído) apresenta-se como a opção mais utilizada, apesar de serem freqüentemente registradas críticas, como relata por exemplo (Kubrick and Ellis, 1990), no que diz respeito à correlação de PSNR com resultados de avaliações subjetivas.

A relação sinal-ruído de pico é definida como 10 vezes o logaritmo na base 10 da razão entre o quadrado do valor de pico da amplitude do sinal de entrada, v_p^2, e o erro médio quadrático (MSE, *mean square error*):

$$\text{PSNR} = 10 \log_{10} \left[\frac{v_p^2}{\text{MSE}} \right]. \tag{3.14}$$

Para o caso de uma imagem original codificada a 8,0 bpp (256 níveis de cinza),

$$\text{PSNR} = 10 \log_{10} \left[\frac{255^2}{\text{MSE}} \right], \tag{3.15}$$

em que o erro médio quadrático entre as imagens original e reconstruída é definido como

$$\text{MSE} = \frac{1}{L \cdot C} \sum_{l=1}^{L} \sum_{c=1}^{C} [F(l,c) - \widehat{F}(l,c)]^2, \tag{3.16}$$

em que $F(l,c)$ e $\widehat{F}(l,c)$ representam os valores de pixels das imagens original e reconstruída, l designa a l-ésima linha e c denota a c-ésima coluna de uma imagem (matriz) $L \times C$.

3.4.2 Taxa de Bits

A taxa de bits de uma representação digital pode ser medida em bits por amostra, bits por pixel (bpp), bits por segundo, dependendo do contexto. A taxa de bits por segundo é simplesmente o produto da taxa de amostragem e o número de bits por amostra. A taxa de amostragem é, geralmente, ligeiramente superior a duas vezes a largura de faixa do sinal, conforme estabelecido pelo teorema da amostragem de Nyquist (Jayant and Noll, 1984; Lathi, 1988).

A Tabela 3.2 ilustra alguns formatos comumente utilizados para áudio (Jayant, 1992; Jayant et al., 1993). São utilizadas taxas de amostragem típicas de 8 kHz para voz telefônica, 16 kHz para áudio AM, 32 kHz para áudio FM, e 44,1 kHz ou 48 kHz para CD (*compact-disk*) ou áudio DAT (*digital audio tape*), sendo ambos sinais de largura de faixa 20 kHz. Observe que as larguras de faixa respectivas são menores que metade da taxa de amostragem correspondente, seguindo o princípio da amostragem de Nyquist. Cumpre mencionar que um sinal PCM tem sua banda limitada por um filtro passa-baixa. A banda de telefonia normalmente transmitida é de 300 Hz a 3400 Hz na Europa e América Latina e de 200 Hz a 3400 Hz nos Estados Unidos e Japão.

3.4.3 Complexidade

A complexidade de um algoritmo de codificação está relacionada ao esforço computacional requerido para implementar os processos de codificação e decodificação. Diz respeito, portanto, à capacidade aritmética e aos requisitos de memória. A complexidade é comumente medida em MIPS (milhões de instruções por segundo) (Aguiar Neto, 1995). Outras medidas relacionadas à complexidade são o tamanho físico do codificador, decodificador ou *codec* (*co*dificador mais *dec*odificador), seu custo e o consumo de potência (medida, por exemplo, em milliwatt, mW), sendo este um critério particularmente importante para sistemas portáteis (Jayant, 1992; Jayant et al., 1993)

Tabela 3.2: Formatos de áudio digital.

Formato	Taxa de Amostragem (kHz)	Largura de Faixa (kHz)	Faixa de Freqüência
Telefonia	8	3,2	200-3400 Hz
Teleconferência	16	7	50-7000 Hz
Compact Disk (CD)	44,1	20	20-20000 Hz
Digital Audio Tape (DAT)	48	20	20-20000 Hz

3.4.4 Retardo de Comunicação

O aumento de complexidade em um algoritmo de codificação é geralmente associado a um aumento de atraso de processamento no codificador e decodificador. A importância do retardo em um sistema de comunicação depende da aplicação. Dependendo do ambiente de comunicação, o atraso total tem que ser mantido em um limite severo, como no caso da utilização em redes telefônicas (Jayant, 1992; Jayant et al., 1993). Assim, o retardo produzido por um *codec* impõe certas restrições práticas quanto à utilização em sistemas de comunicações tendo em vista que o retardo não deve ultrapassar um determinado limite. No entanto, o retardo de comunicação pode ser visto como irrelevante em aplicações que envolvem comunicação unidirecional, a exemplo de sistemas de difusão de TV, ou armazenamento e envio de mensagens, como é o caso do correio de voz (Aguiar Neto, 1995).

3.4.5 Codificação e Comunicação Digital

A Figura 3.5 (Jayant, 1992; Jayant et al., 1993) descreve os critérios de desempenho em comunicação digital, que se aplicam não só à codificação de fonte, mas também à codificação de canal e à modulação, nas quais as unidades de qualidade e taxa de bits são diferentes. Na Figura 3.5, no eixo de qualidade do sinal, MOS diz respeito à codificação de fonte, ao passo que p_e, probabilidade de erro, diz respeito à codificação de canal e modulação. No eixo de eficiência, bps diz respeito à codificação de fonte, e bps/Hz diz respeito à codificação de canal e modulação. As unidades para atraso e complexidade são as mesmas, embora esses parâmetros estejam em contextos diferentes em se tratando de codificação de fonte e em codificação de canal. O atraso de processamento é usado em codificação de fonte para remover a redundância do sinal. Em codificação de canal, pode ser usado por exemplo para adicionar bits de proteção de erros e para procedimentos com propósito de assegurar "aleatoriedade" nos erros em surto, como é o caso do entrelaçamento (*interleaving*).

O desempenho de um sistema de codificação pode ser avaliado levando em consideração os quatro parâmetros previamente mencionados. Se ignorarmos por um momento a complexidade e o retardo de comunicação, os ganhos de desempenho de um codificador de fonte podem ser avaliados de duas formas: medindo os ganhos de qualidade do sinal reconstruído a uma taxa de bits especificada ou atingindo um nível de qualidade de sinal reconstruído a uma taxa de bits mais baixa. Dependendo da aplicação, uma dessas abordagens é mais relevante que a outra. Por exemplo, em problemas de codificação de voz telefônica a 16 kbps (quilobits por segundo) e HDTV a 15-20 Mbps, as taxas de bits são definidas pelas aplicações ou pelos padrões e o alvo da pesquisa em codificação é melhorar a qualidade do sinal a essas taxas. Por outro lado, no campo de difusão de áudio digital, em que a qualidade do sinal deve ser transparente (elevada) com relação ao algoritmo de codificação, o alvo é obter a qualidade requerida a taxas cada vez mais baixas.

A seção a seguir apresenta algumas características dos sinais de voz. Vale salientar que o estudo da produção e da percepção de voz pelos seres humanos trouxe grandes benefícios para o desenvolvimento de técnicas de compressão de voz.

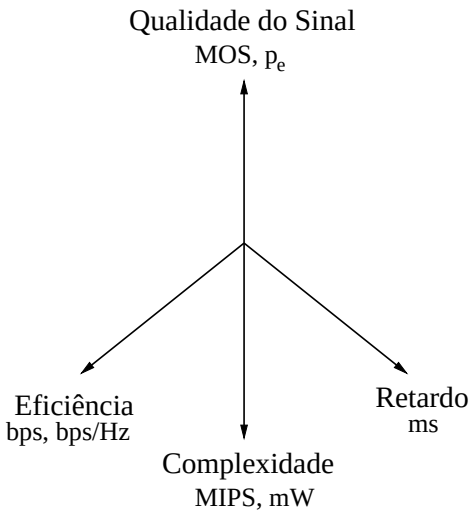


Figura 3.5: Parâmetros de desempenho de um codificador.

3.5 Caracterização dos Sinais de Voz

O conhecimento das características dos sinais de voz tem sido utilizado eficientemente em técnicas de codificação de voz, de síntese de voz, bem com em sistemas de reconhecimento de fala e de locutor.

O mecanismo de produção de voz humana, como ocorre em qualquer sistema físico, apresenta resposta em freqüência limitada. O limite varia de pessoa para pessoa, situando-se, em média, em torno de 10 kHz (Aguiar Neto, 1995). Nos sistemas telefônicos, entretanto, limita-se o sinal de voz na faixa de 200-3400 Hz, sem grandes prejuízos em termos de qualidade. Em outras aplicações, como teleconferência, limita-se o sinal de voz em 7 kHz, o que caracteriza a transmissão de voz em faixa larga.

Os sons da voz humana podem ser classificados em três classes distintas, como sonoros (tais como /a/ e /i/), surdos (como /sh/) ou explosivos (como /p/, /t/ e /k/) (Rabiner and Schafer, 1978; Aguiar Neto, 1995). Os sinais de voz classificados como sonoros são quase-periódicos no domínio do tempo e harmonicamente estruturados no domínio da freqüência, enquanto os sons surdos têm uma natureza aleatória e uma faixa larga. Além disso, a energia dos segmentos sonoros em geral é maior que a energia dos segmentos surdos. Os sons sonoros, portanto, dizem respeito a ondas de pressão quase-periódicas excitando o trato vocal, que, atuando como um ressonador, produz freqüências de ressonância denominadas formantes, que caracterizam os diferentes sons sonoros. Portanto, o envelope espectral que se ajusta ao espectro de curto prazo da voz sonora é caracterizado por um conjunto de picos, denominados formantes. Em geral há de três a cinco formantes abaixo de 5 kHz. As amplitudes e as localizações dos primeiros três formantes são muito importantes para a síntese e a percepção de voz (Spanias, 1994). A freqüência fundamental dos sons sonoros fica entre 80 Hz (para homens) e 350 Hz (para crianças), sendo 240 Hz um valor típico para mulheres (Aguiar Neto, 1995). Na geração dos sons explosivos, o ar é totalmente dirigido à boca, estando esta completamente fechada. Com o aumento da pressão, a oclusão é rompida bruscamente. Há ainda os sons com excitação mista – como os sons fricativos sonoros (como /j/, /v/ e /z/), que são produzidos combinando a vibração das cordas vocais e a excitação turbulenta, e os sons oclusivos (ou explosivos) sonoros (como /d/ e /b/).

Os sinais de voz são não-estacionários. Entretanto, podem ser considerados como quase-estacionários em curtos segmentos (curtos intervalos de tempo), tipicamente 5-20 ms (Spanias, 1994). Os sinais de voz, assim, podem ser considerados ergódicos (médias estatísticas e médias temporais são idênticas) (Aguiar Neto, 1995). As propriedades estatísticas e espectrais da voz são definidas, portanto, em curtos intervalos de tempo.

A descrição mais simples de alguma forma de onda é feita por meio do gráfico amplitude versus tempo, como ilustrado na Figura 3.6. Na forma de onda, podem ser identificados segmentos de voz de alta energia,

de baixa energia e segmentos de pausas intersilábicas. Estes últimos e os segmentos de pausa entre palavras corrrespondem a cerca de 50 a 60% do tempo do sinal (Aguiar Neto, 1995).

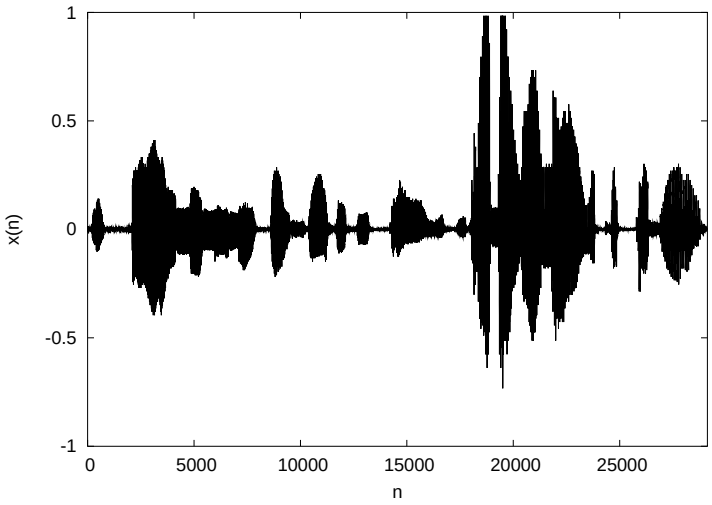


Figura 3.6: Forma de onda do sinal de voz correspondente à sentença "*O sol ilumina a fachada de tarde. Trabalhou mais do que podia*.", correspondente a 29.120 amostras, que equivalem a 3,64 s, tendo sido o sinal amostrado a 8 kHz.

Os sinais de voz apresentam uma variação bastante grande de amplitudes, como mostrado na Figura 3.7. A variação de amplitude do sinal, que corresponde a cerca de 50 dB, é denominada faixa dinâmica do sinal. Vale mencionar que a aquisição (resolução 8,0 bit/amostra e taxa de amostragem 8 kHz) dos sinais de voz utilizados nas Figuras 3.6 e 3.7, bem como nas Figuras 3.8 e 3.9, foi realizada usando uma estação de trabalho Sun®, instalada com utilitários de processamento de áudio.

No que diz respeito à função densidade de probabilidades (fdp) para as amplitudes dos sinais de voz, podem ser utilizadas aproximações seguindo os modelos exponencial *two-sided* ou laplaciano, Gamma e gaussiano (Aguiar Neto, 1995).

Quanto à energia do sinal de voz, concentra-se na região de freqüências mais baixas do espectro, notadamente na faixa de 500 a 800 Hz. No entanto, mesmo contendo baixos valores de energia, as componentes de freqüências mais altas são importantes pois determinam, em grande parte, a inteligibilidade da voz. O espectro decai cerca de 8-10 dB por oitava. As freqüências abaixo de 500 Hz contribuem muito pouco para a compreensão da fala, mas têm um papel importante no tocante à naturalidade da voz reproduzida (Aguiar Neto, 1995).

A Figura 3.8 representa, no espaço de padrões euclidiano bidimensional, um sinal de voz (cujo histograma é apresentado na Figura 3.9), correspondente a 10 frases foneticamente balanceadas (extraídas de (Alcaim et al., 1992) e pronunciadas por 10 locutores diferentes, sendo 5 masculinos e 5 femininos). Algumas características típicas dos sinais de voz podem ser observadas nas Figuras 3.8 e 3.9, tais como predominância de amostras de baixa amplitude e correlação entre amostras consecutivas (os vetores de voz concentram-se próximo à direção correspondente à componente principal do sinal de voz, em que $x_2 = x_1$.).

As seções a seguir abordam o processo de quantização. São enfocadas a quantização escalar e a quantização vetorial.

3.6 Quantização Escalar

O processo de quantização pode ser visualizado como o mapeamento do sinal, a partir do domínio contínuo, para um número contável de possíveis níveis de saída. A necessidade de representar sinais com um número

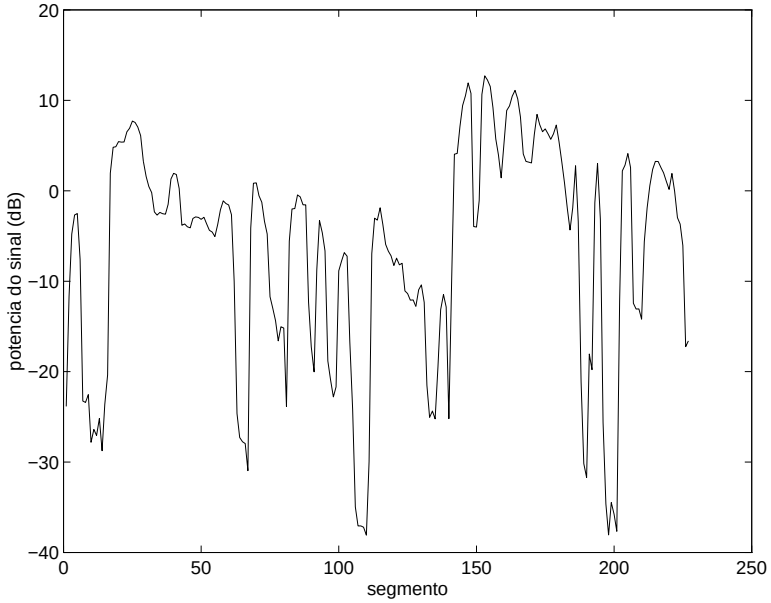


Figura 3.7: Faixa dinâmica do sinal de voz "*O sol ilumina a fachada de tarde. Trabalhou mais do que podia*.", correspondente a 29.120 amostras, que equivalem a 3,64 s, tendo sido o sinal amostrado a 8 kHz (Madeiro, 1998).

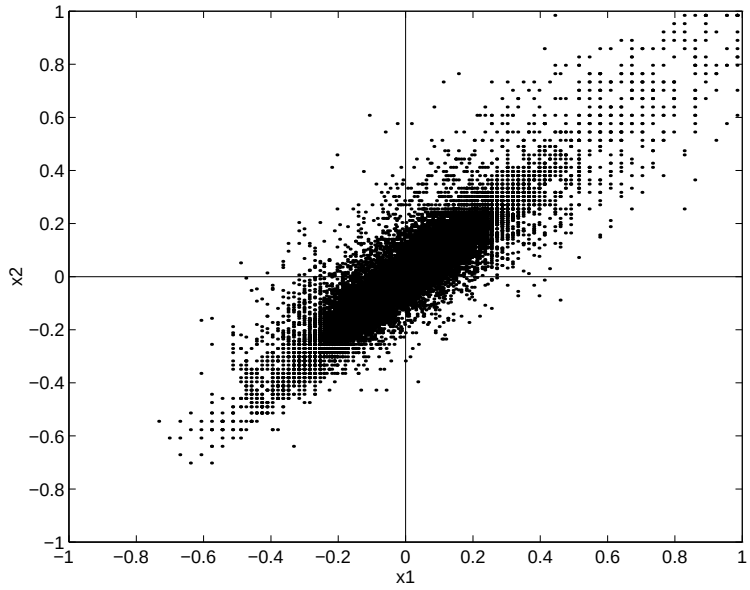


Figura 3.8: Sinal de voz consistindo de 10 frases foneticamente balanceadas (18,76s, 75.040 vetores). As coordenadas x_1 e x_2 representam a primeira e a segunda componentes dos vetores de voz $\boldsymbol{x} \in R^2$, respectivamente.

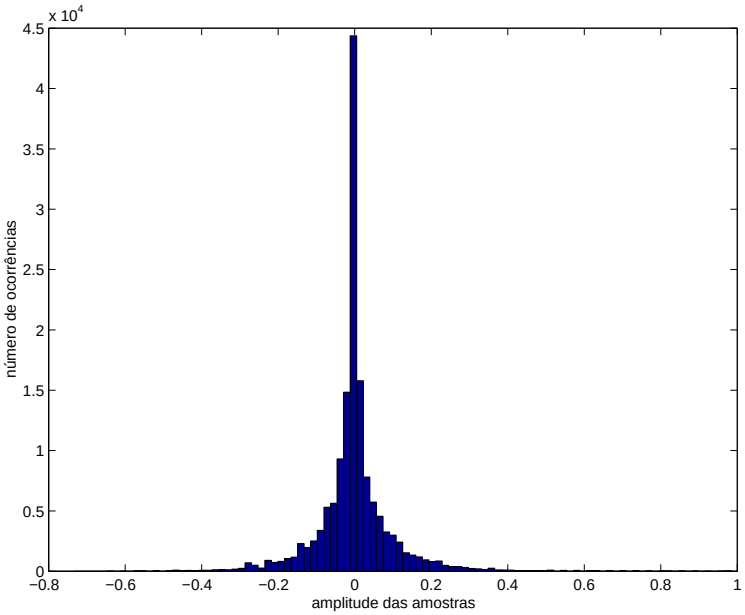


Figura 3.9: Histograma de um sinal de voz.

finito de bits faz com que o ruído de quantização esteja presente em quase todos os sistemas de processamento digital de sinais. Isto implica que o erro de quantização é intrínseco ao processo de conversão analógico-digital (Sripad and Snyder, 1977).

Apesar de bastante simples em sua descrição e construção, o quantizador uniforme tem provado ser surpreendentemente difícil de analisar, precisamente em função de sua inerente não-linearidade. Como estimar o espectro do ruído de quantização e recuperar o máximo de sinal, dado que esse erro está presente, é um dos objetivos da pesquisa na área.

A Figura 3.10 ilustra o esquema de quantização para um quantizador uniforme com passo de quantização d. É válido mencionar que a relação sinal-ruído de quantização cai com o inverso da amplitude do sinal. Em suma, valores menores de amplitude sofrem mais com o ruído de quantização.

3.6.1 O Ruído de Quantização

O erro, ou ruído, de quantização consiste na diferença entre o sinal na entrada do quantizador e o sinal na saída, $n = x - y$, no qual $y = q(x)$ e $q(\cdot)$ representa a função de quantização. O desempenho de sistemas de codificação ou processamento de sinais está limitado pelo nível do ruído de quantização. A própria capacidade do canal se limita em função desse ruído. Por conseguinte, a figura de mérito mais utilizada em análises comparativas é a SQNR (*signal to quantization noise ratio* – SQNR)[1](Paez and Glisson, 1972). O erro médio quadrático, para um quantizador uniforme, é dado aproximadamente por $d^2/12$ (Bennett, 1948), supondo-se uma distribuição uniforme para o ruído de quantização, em que d representa o passo de quantização. Esse resultado é mostrado a seguir.

Assumindo uma distribuição de probabilidades do tipo uniforme para o ruído, no intervalo $[-d/2, d/2]$,

$$p_N(n) = \frac{1}{d}, \tag{3.17}$$

[1]Ao longo do capítulo também se utiliza a notação SNR para denotar a relação sinal-ruído de quantização.

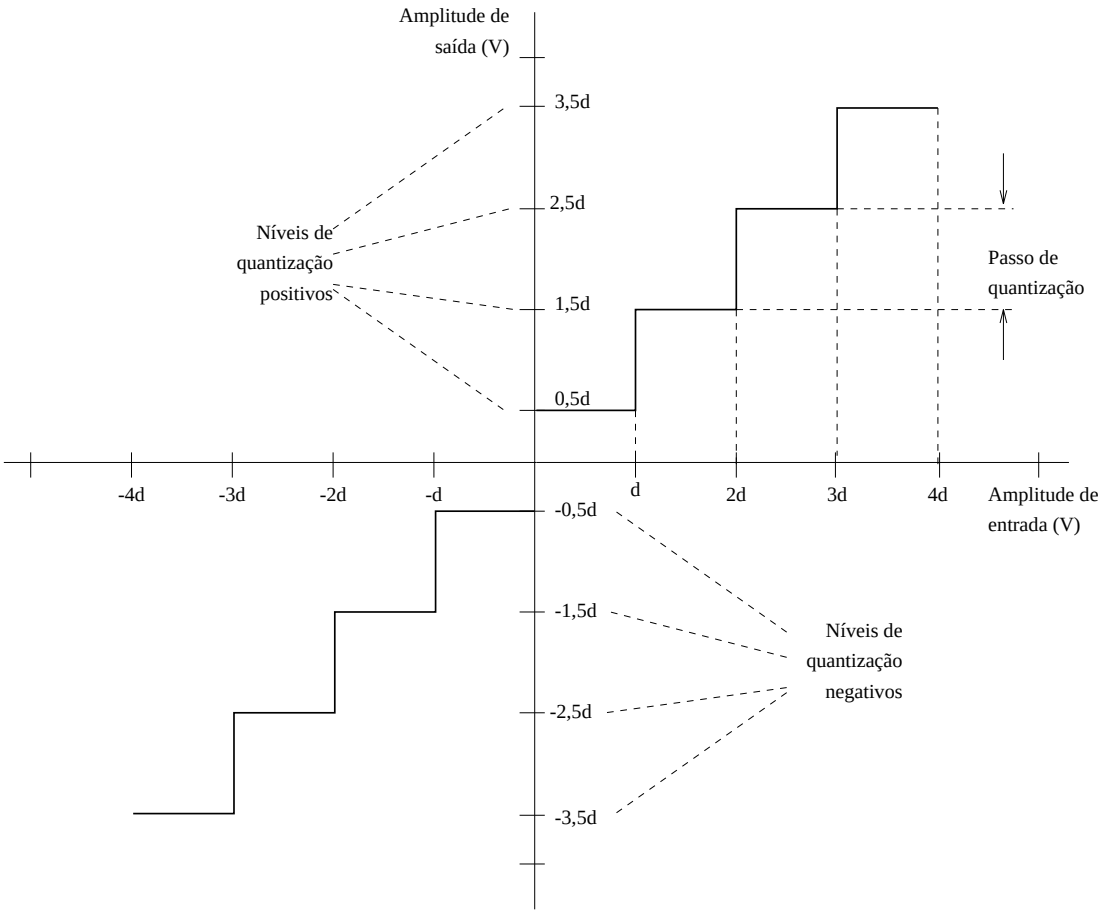


Figura 3.10: Esquema de quantização.

a potência do ruído de quantização será calculada pela fórmula

$$P_N = \int_{-\infty}^{\infty} n^2 p_N(n) dn. \tag{3.18}$$

Substituindo a distribuição assumida, obtém-se

$$P_N = \frac{1}{d}\int_{-d/2}^{d/2} n^2 dn = \frac{1}{d}\frac{d^3}{12} = \frac{d^2}{12}. \tag{3.19}$$

Este resultado foi primeiramente desenvolvido por Claude E. Shannon, em 1948 (Shannon, 1948c).

Com esse valor para a potência do ruído de quantização, pode-se calcular o efeito de aumento na SQNR como sendo de 6 dB para cada bit adicional, conforme mostrado a seguir.

A SQNR é dada por

$$\mathrm{SQNR} = 10 \log \frac{P_X}{P_N}, \tag{3.20}$$

em que P_X representa a potência do sinal a ser quantizado e $P_N = d^2/12$ a potência do ruído de quantização.

Para um sinal com faixa dinâmica, *i.e*, intervalo de variação das amplitudes do sinal, igual a $2\sqrt{P_X}$ e $N = 2^m$ níveis de quantização, em que m representa o número de bits de codificação, o passo de quantização

será dado por

$$d = \frac{2\sqrt{P_X}}{N} = \frac{2\sqrt{P_X}}{2^m}. \tag{3.21}$$

Por conseguinte, a potência do ruído de quantização é dada por

$$P_N = \frac{d^2}{12} = \frac{4P_X}{12N^2} = \frac{P_X}{3 \cdot 4^m}. \tag{3.22}$$

Substituindo a Equação 3.22 em 3.20, vem

$$\text{SQNR} = 10 \log 3 \cdot 4^m = 10 \log 3 + 10m \log 4 \approx 5 + 6m \text{ dB}. \tag{3.23}$$

Portanto, a partir da Equação 3.23 percebe-se que a SQNR aumenta 6 dB para cada bit adicional do código utilizado, considerando um quantizador uniforme.

3.6.2 Quantização Não Uniforme

Na quantização uniforme, a relação sinal-ruído de quantização é dada pela Equação 3.20, com a potência do ruído de quantização dependendo apenas da largura d dos degraus do quantizador, conforme mostra a Equação 3.19. A SQNR, portanto, depende diretamente da potência do sinal de entrada. Como a faixa dinâmica dos sinais de voz é de cerca de 50 dB, a SQNR varia bastante, decrescendo com o decréscimo da potência do sinal de entrada. Isso significa que a SQNR pode variar amplamente. A qualidade do sinal pode se deteriorar, por exemplo, nos intervalos de tempo em que uma pessoa conversa muito brandamente. Estatisticamente, observa-se uma predominância das amostras de baixas amplitudes nos sinais de voz, o que significa que a SQNR será baixa na maior parte do tempo. Idealmente, no entanto, deseja-se obter uma SQNR que possa ser a mais constante (mesma qualidade) possível para todos os valores de potência do sinal mensagem (sinal de entrada). O uso de um quantizador não uniforme procura assegurar essa constância de SQNR: neste caso, o passo de quantização não é constante, mas é função do valor da amplitude do sinal. Para níveis mais baixos do sinal mensagem, o passo ou degrau de quantização é menor. Aumenta-se logaritmicamente a largura do degrau de quantização à medida que o nível do sinal mensagem aumenta.

O método de codificação de forma de onda PCM está definido na CCITT G.711, e AT&T 43801. Basicamente, o sinal é amostrado a uma taxa de 8000 vezes por segundo. Em se tratando de quantização não uniforme, a expressão compansão foi cunhada para designar os processos de compressão e expansão do sinal a ser codificado. A compressão é necessária para elevar os níveis mais débeis do sinal, em comparação com os níveis elevados, e com isso tornar o codificador mais robusto. A expansão é realizada no receptor, como uma função inversa da compressão.

As duas leis de compressão recomendadas pelo antigo CCITT (atual ITU-T) são a Lei μ e a Lei A, dadas pelas equações seguintes

Lei μ

$$y = C(x) = \frac{V \ln(1 + \mu x/V)}{\ln(1 + \mu)}, \text{ se } x > 0 \tag{3.24}$$

ou

$$y = -C(-x), \text{ se } x \leq 0 \tag{3.25}$$

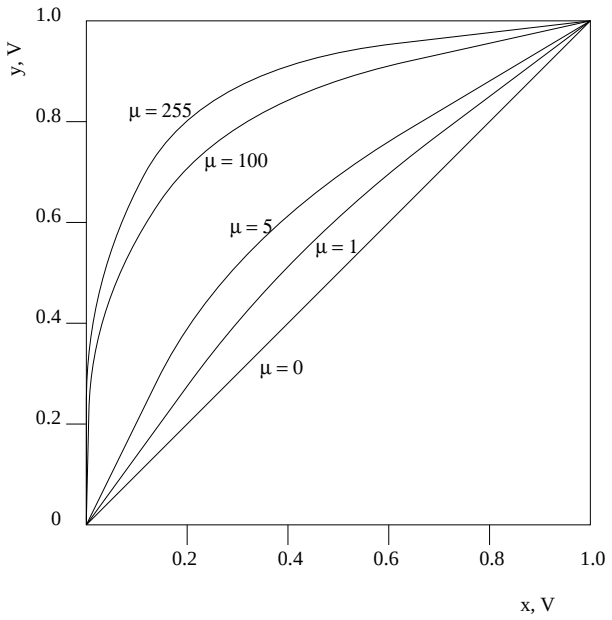


Figura 3.11: Curvas de compressão para a Lei μ.

Lei A

$$y = C(x) = \frac{Ax}{1 + \ln A},\ 0 \leq x \leq V/A \tag{3.26}$$

e

$$y = C(x) = \frac{V\,(1 + \ln(Ax/V))}{1 + \ln A},\ V/A \leq x \leq V \tag{3.27}$$

ou ainda

$$y = -C(-x),\ \text{se}\ x \leq 0 \tag{3.28}$$

As curvas de compressão correspondentes são ilustradas nas respectivas Figuras 3.11 e 3.12. A Lei A foi adotada inicialmente na Europa e também no Brasil. A Lei μ foi originalmente desenvolvida nos Estados Unidos e depois adotada como padrão pelo CCITT.

As curvas ilustram o efeito causado pela variação dos parâmetros μ e A. Um aumento no respectivo parâmetro implica aumentar a não-linearidade da curva para os dois quantizadores. Para $A = 1$ e $\mu = 0$, as curvas de compressão respectivas ficam lineares. Os valores normalmente utilizados, obtidos a partir de testes subjetivos de voz (*Mean Opinion Square* – MOS), são: $\mu = 255$ e A = 87,6. O padrão G.711 PCM Lei-A ou Lei-μ tem um MOS em torno de 4,2.

A curva de relação sinal-ruído de quantização, em função do inverso da amplitude do sinal, para quantizadores não uniformes fica mais plana. Dessa forma, os níveis mais baixos do sinal são preservados.

3.7 Quantização Vetorial

A quantização vetorial (Gersho and Gray, 1992; Gray, 1984), que pode ser vista como uma extensão da quantização escalar em um espaço multidimensional, encontra-se fundamentada na Teoria da Distorção Versus

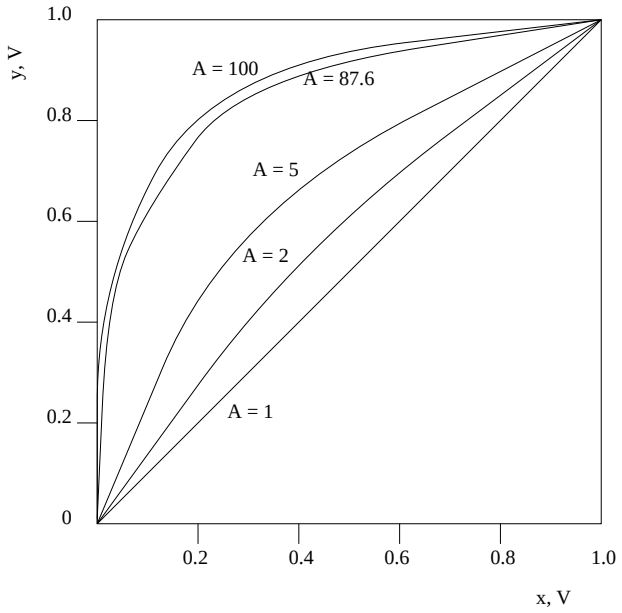


Figura 3.12: Curvas de compressão para a Lei A.

Taxa, formulada por Shannon, segundo a qual um melhor desempenho é obtido codificando blocos de amostras (isto é, vetores) ao invés de amostras individuais (isto é, escalares). Em outras palavras, essa teoria ressalta a superioridade da quantização vetorial sobre a quantização escalar. Matematicamente, a quantização vetorial pode ser definida como um mapeamento Q de um vetor $\boldsymbol{x}$ pertencente ao espaço euclidiano K-dimensional, $\mathbb{R}^K$, em um vetor pertencente a um subconjunto finito W de $\mathbb{R}^K$, ou seja,

$$Q : \mathbb{R}^K \rightarrow W. \tag{3.29}$$

O dicionário $W = \{\boldsymbol{w}_i;\ i = 1,\ 2,\ \ldots,\ N\}$ é o conjunto de vetores-código K-dimensionais, também denominados vetores de reconstrução. O índice associado ao vetor-código $\boldsymbol{w}_i$ será denotado por $\boldsymbol{i}$. Assim, cada índice $\boldsymbol{i} \in \{0, 1\}^b$ representa uma palavra-binária de b bits. A taxa de codificação do quantizador vetorial, que mede o número de bits por componente do vetor, é $R = \frac{1}{K} \log_2 N = \frac{b}{K}$. Em codificação de forma de onda de voz, R é expressa em bits/amostra. Em se tratando de codificação de imagens, R é expressa em bits por pixel (bpp).

Em um sistema de codificação de sinais baseado em quantização vetorial, conforme apresentado na Figura 3.13, o codificador e o decodificador funcionam como descrito a seguir. Dado um vetor $\boldsymbol{x} \in \mathbb{R}^K$ do sinal a ser codificado, o codificador determina a distorção $d(\boldsymbol{x}, \boldsymbol{w}_i)$ entre esse vetor e cada vetor-código $\boldsymbol{w}_i$, $i = 1,\ 2,\ \ldots,\ N$ do dicionário W. A regra ótima de codificação é a regra do vizinho mais próximo, na qual a palavra-binária $\boldsymbol{i}$ é transmitida ao decodificador se o vetor-código $\boldsymbol{w}_i$ corresponder à distorção mínima, isto é, se $\boldsymbol{w}_i$ for o vetor que apresenta a maior similaridade com $\boldsymbol{x}$ dentre todos os vetores-código do dicionário. Em outras palavras, o codificador usa a regra de codificação $C(\boldsymbol{x}) = \boldsymbol{i}$ se $d(\boldsymbol{x}, \boldsymbol{w}_i) < d(\boldsymbol{x}, \boldsymbol{w}_j), \forall j \neq i$. A tarefa do decodificador é muito simples: ao receber o índice $\boldsymbol{i}$ de b bits, o decodificador simplesmente procura o vetor $\boldsymbol{w}_i$ em sua cópia do dicionário W e produz $\boldsymbol{w}_i$ como a reprodução (reconstrução) de $\boldsymbol{x}$. Ele segue, portanto, a regra de decodificação $D(\boldsymbol{i}) = \boldsymbol{w}_i$. O mapeamento de $\boldsymbol{x}$ em $\boldsymbol{w}_i$ é geralmente expresso como $\boldsymbol{w}_i = Q(\boldsymbol{x})$.

No cenário de codificação digital de sinais, a quantização vetorial, portanto, constitui uma técnica de compressão com perdas, visto que o sinal reconstruído é uma versão degradada do sinal original. O erro de quantização, introduzido ao se representar o sinal de entrada por sua versão quantizada, é chamado distorção do quantizador. Uma das questões principais no projeto de quantizadores vetoriais é o compromisso entre taxa e distorção. O alvo a ser perseguido é a obtenção de um dicionário ótimo, que minimize, para uma determinada

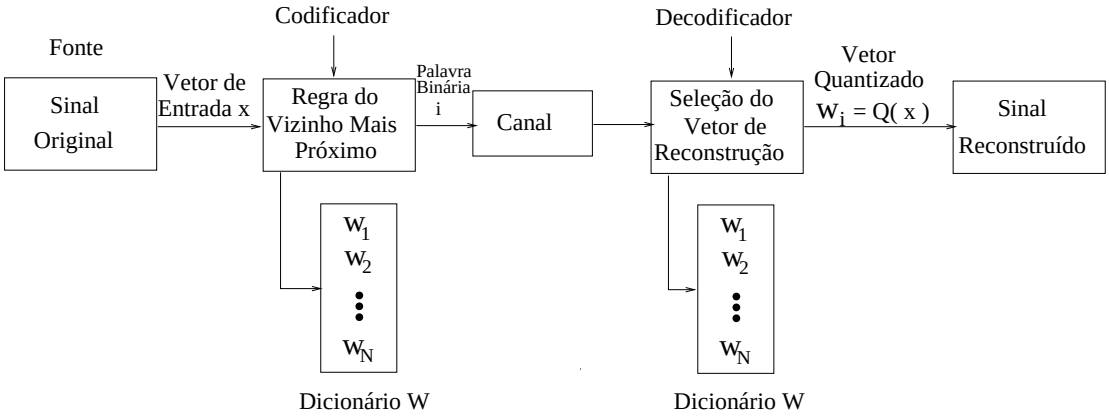


Figura 3.13: Sistema de codificação baseado em quantização vetorial.

taxa de codificação, a distorção média introduzida pela aproximação dos vetores de entrada por seus correspondentes vetores-código. Dentre as diversas técnicas para projeto de dicionários, o algoritmo LBG (Linde-Buzo-Gray) (Linde et al., 1980) destaca-se por sua ampla utilização. Outras abordagens têm sido utilizadas em projeto de dicionários, como por exemplo: algoritmo de Kohonen (Kohonen, 1990) e outros algoritmos não-supervisionados (Krishnamurthy et al., 1990; Chen et al., 1994); relaxação estocástica (Zeger et al., 1992); algoritmos *fuzzy* (Karayiannis and Pai, 1995) e algoritmo genético (Pan et al., 1995).

O projeto de dicionários tem um papel importante para o bom desempenho de sistemas de processamento de sinais baseados em quantização vetorial (QV) (Gray, 1984; Gersho and Gray, 1992). Em codificação de voz e imagem baseada em QV (e.g. (Abut et al., 1982; Ramamurthi and Gersho, 1986)), a qualidade dos sinais reconstruídos depende dos dicionários projetados. Em sistemas de identificação de locutor que utilizam QV paramétrica (e.g. (Soong et al., 1987; Fechine, 2000)), as taxas de identificação dependem dos dicionários de padrões acústicos de referência projetados para cada locutor cadastrado pelo sistema.

O mapeamento Q leva a um particionamento de $\mathbb{R}^K$ em N subespaços (células, denominadas regiões de Voronoi) $S_i, i = 1,\ 2,\ \ldots,\ N$, para os quais

$$\bigcup_{i=1}^{N} S_i = \mathbb{R}^K \text{ e } S_i \cap S_j = \emptyset \text{ se } i \neq j, \tag{3.30}$$

em que cada célula ou região S_i é definida como

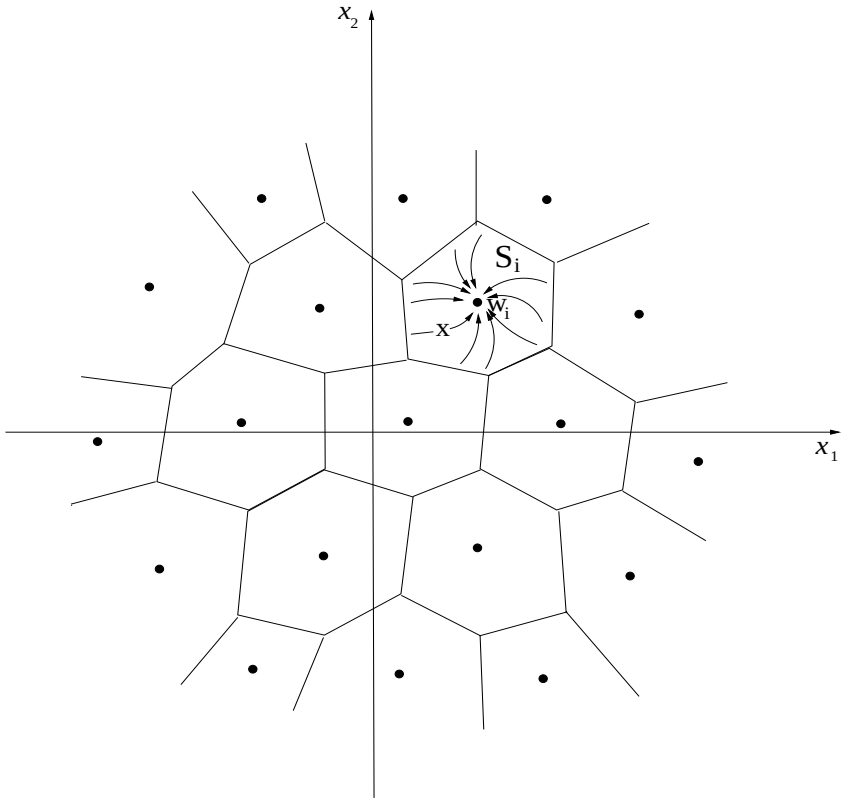
$$S_i = \{\boldsymbol{x} :\ Q(\boldsymbol{x}) = \boldsymbol{w}_i\} = \{\boldsymbol{x} :\ C(\boldsymbol{x}) = \boldsymbol{i}\}. \tag{3.31}$$

O vetor-código $\boldsymbol{w}_i$ constitui o vetor representativo de todos os vetores de entrada pertencentes à célula S_i, conforme ilustra a Figura 3.14. Como a quantização vetorial realiza um mapeamento de padrões de entrada (vetores de entrada $\boldsymbol{x}$) semelhantes em padrões de saída (vetores-código $\boldsymbol{w}_i$) semelhantes, ela pode ser vista como uma forma de reconhecimento de padrões, em que um padrão de entrada é "aproximado" por um padrão de referência, pertencente a um conjunto predeterminado (dicionário) de padrões (vetores-código) de referência (Gersho and Gray, 1992; Kosko, 1992).

3.7.1 Algoritmo LBG

Seja a iteração do algoritmo LBG denotada por n. Dados K, N e um limiar de distorção $\epsilon \geq 0$, o algoritmo LBG (Linde et al., 1980) consiste da seguinte seqüência de passos:

- *Passo 1*) inicialização: dado um dicionário inicial W_0 e um conjunto de treino $\boldsymbol{X} = \{\boldsymbol{x}_m;\ m = 1,\ 2,\ \ldots,\ M\}$, faça $n = 0$ e $D_{-1} = \infty$;



.

Figura 3.14: Partição do espaço euclidiano bidimensional, $\mathbb{R}^2$, introduzido pelo mapeamento dos vetores de entrada $\boldsymbol{x}$ nos vetores-código $\boldsymbol{w}_i$. As coordenadas x_1 e x_2 representam a primeira e a segunda componentes do vetor $\boldsymbol{x} \in \mathbb{R}^2$, respectivamente.

- *Passo 2*) particionamento: dado W_n (dicionário na n-ésima iteração), aloque cada vetor de treino (vetor de entrada) na respectiva classe (célula de Voronoi) segundo o critério do vetor-código mais próximo; calcule a distorção

$$D_n = \sum_{i=1}^{N} \sum_{\boldsymbol{x}_m \in S_i} d(\boldsymbol{x}_m, \boldsymbol{w}_i); \tag{3.32}$$

- *Passo 3*) teste de convergência (critério de parada): se $(D_{n-1} - D_n)/D_n \leq \epsilon$ pare, com W_n representando o dicionário final (dicionário projetado); caso contrário, continue;

- *Passo 4*) atualização do dicionário: compute os novos vetores-código como os centróides das classes de vetores; faça $W_{n+1} \leftarrow W_n$; faça $n \leftarrow n + 1$ e retorne ao *Passo 2*.

No algoritmo LBG a função distorção decresce monotonicamente, uma vez que o dicionário é iterativamente atualizado visando satisfazer as condições de centróide e de vizinho mais próximo. No algoritmo LBG, a distorção introduzida ao se representarem os vetores do conjunto de treinamento pelos correspondentes vetores-código (centróides) é monitorada a cada iteração. A regra de parada (teste de convergência) do algoritmo baseia-se nessa distorção monitorada – o treinamento do dicionário é encerrado quando $(D_{n-1} - D_n)/D_n \leq \epsilon$. Existem alguns problemas apresentados pelo algoritmo LBG, comumente relatados: alguns vetores-código podem ser sub-utilizados e, em casos extremos, até mesmo nunca serem utilizados, ou seja, o projeto do dicionário pode resultar em células de Voronoi pequenas ou até mesmo vazias; a velocidade de convergência e o desempenho do dicionário final dependem do dicionário inicial.

3.8 Parâmetros LPC

A maioria dos algoritmos de codificação de voz atuais faz uso do modelo fonte-filtro da produção de voz humana, no qual a voz é modelada como a resposta de um filtro de síntese linar variante no tempo a um sinal de entrada chamado excitação. Exemplos de codificadores de voz baseados nesse modelo incluem as numerosas variações do *vocoder* LPC (*linear predictive coding*) e a ampla família de codificadores de análise por síntese baseados em predição linear (LPAS, *linear-prediction-based analysis-by-synthesis*) – incluindo o CELP (*code-excited linar prediction*). O filtro de síntese determina o envelope espectral de curto prazo (*short-term*) da voz sintetizada e é caracterizado pelos coeficientes de predição linear, obtidos a partir da análise LP (*linear prediction*) no sinal de voz de entrada. Esses coeficientes são comumente denominados *coeficientes LPC*, que podem referir-se genericamente a qualquer um dos vários conjuntos diferentes (porém, equivalentes) de parâmetros que especificam o filtro de síntese.

Ao longo dos anos, têm sido intensas as atividades de pesquisa em quantização LPC (e.g. (Paliwal and Atal, 1993; Kleijn and Paliwal, 1995; Atal et al., 1993)). O principal objetivo da quantização LPC para codificação de voz é evitar a introdução de qualquer distorção perceptível na voz codificada, mantendo a codificação dos parâmetros à menor taxa possível. Se esse objetivo for alcançado, diz-se atingir a chamada *qualidade transparente* ou *quantização transparente.* Devido ao custo e à dificuldade de realizar testes de qualidade subjetivos, os pesquisadores têm utilizado medidas objetivas para avaliar a distorção no envelope espectral devido à quantização LPC. Especificamente, a distorção espectral ou distância espectral tem sido um critério de desempenho padrão. Associado a essa medida, há um critério objetivo para quantização LPC transparente, proposto por Paliwal e Atal (Paliwal and Atal, 1993), baseado no papel dos *outliers*.

O primeiro estudo de quantização vetorial aplicada a parâmetros LPC se deve a Buzo *et al.* em 1980 (Buzo et al., 1980). Recentemente, vários artigos envolvendo esquemas sofisticados de quantização vetorial têm sido apresentados. O trabalho de Paliwal e Atal (Paliwal and Atal, 1993) empregou eficientemente a técnica de split VQ (*split vector quantization*) para obter qualidade transparente à taxa de 24 bits por quadro. Esse trabalho tem sido uma referência freqüentemente utilizada como um *benchmark* para comparar outros resultados. Diversos outros pesquisadores têm obtido resultados similares ou melhores.

3.8.1 Quantização LPC

Em um codificador de voz baseado no modelo fonte-filtro, os coeficientes LPC $\{a_i\}$ são obtidos realizando análise preditiva linear (Markel and Gray, 1976) em cada quadro de voz. Esses coeficientes são usados para formar um filtro de síntese dado por $H(z) = 1/A(z)$, em que $A(z)$ é o filtro inverso, expresso por

$$A(z) = 1 + a_1 z^{-1} + \cdots + a_M z^{-M} \tag{3.33}$$

e M é tipicamente um número entre 10 e 16 denominado ordem do preditor. Devido à natureza quase-estacionária da voz, esse filtro é atualizado a cada quadro de voz, sendo 20 ms um tamanho típico de quadro, o que resulta em uma taxa de 50 quadros por segundo. Um descrição desse filtro de síntese deve ser comunicada ao receptor a cada quadro. O processo de quantização dos filtros a um número finito de de bits por quadro é conhecido como quantização do espectro LPC.

O objetivo da quantização LPC é codificar eficientemente os parâmetros LPC sem introduzir distorção audível na voz codificada. Conforme mencionado anteriormente, as dificuldades associadas aos testes de avaliação subjetiva levaram os pesquisadores a avaliar o desempenho de seus esquemas de quantização utilizando a distorção espectral, SD, que é expressa em dB e calculada para um quadro de voz de acordo com

$$\text{SD}^2 = \frac{2}{F_S} \int_0^{F_S} \{20 \log_{10} |H(e^{j\pi f/F_S})| - 20 \log_{10} |\hat{H}(e^{j\pi f/F_S})|\}^2 df, \tag{3.34}$$

em que F_S é a freqüência de amostragem e $\hat{H}(z)$ é a função de transferência do filtro de síntese quantizada. Portanto, SD^2 é o erro médio quadrático entre as log-magnitudes das respostas em freqüência do filtro de síntese

não quantizada e quantizada, considerando a média tomada na freqüência. Para avaliar o desempenho de um esquema de quantização LPC, é determinada a distorção espectral média (valor médio de SD considerando todos os quadros de voz) e é avaliada a percentagem de *outliers*, que são quadros cuja SD excede um determinado valor limiar. Se uma quantização muito ruim ocorre em um quadro particular, ela causará uma distorção muito incômoda, que causará uma correspondente percepção duradoura no ouvido, e a qualidade perceptual do sinal será prejudicada por um longo período de tempo, mesmo se os parâmetros LPC dos quadros subseqüentes forem quantizados adequadamente. Em (Paliwal and Atal, 1993) foram estabelecidas três condições baseadas em SD para quantização transparente, descritas a seguir:

1. A distorção espectral média é inferior a 1 dB;
2. Não há *outliers* com SD superior a 4 dB;
3. A percentagem de *outliers* com SD na faixa 2-4 dB é abaixo de 2%.

Para quantização dos parâmetros LPC, os coeficientes de predição $\{a_i\}$ são mapeados em uma representação equivalente, que tem boas propriedades de quantização em termos de distribuição, estabilidade e sensibilidade espectral. Representações como *log-area ratios*, *arcsines of reflection coefficients* e *line spectrum frequencies* (também denominados *line spectrum pairs*, LSP's) foram estudadas para esse propósito, fornecendo melhores eficiência de quantização e propriedades de estabilidade que os coeficientes LPC propriamente ditos. Desde a década de 1980, a representação LSF tem sido a forma dominante para o propósito de quantização do espectro LPC. Segundo (Soong and Juang, 1993), o resultado formal de uma avaliação DRT de um *vocoder* LSP a 800 bps é 87, o que é apenas 1,4 pontos inferior ao apresentado pelo *vocoder* LPC padrão 2400 bps. A 4800 bps, o escore DRT do *vocoder* LSP é apenas 0,7 ponto pior que o *vocoder* LPC *residual excited* a 9600 bps.

É válido mencionar que em algoritmos de codificação preditiva linear de voz a transmissão dos parâmetros LPC, em geral transformados para a representação LSF, consome grande parte da taxa total de bits do codificador.

Para definir os parâmetros LSF, o polinômio do filtro inverso é usado para construir dois polinômios:

$$P(z) = A(z) + z^{-(M+1)}A(z^{-1}) \tag{3.35}$$

e

$$Q(z) = A(z) - z^{-(M+1)}A(z^{-1}). \tag{3.36}$$

As raízes dos polinômios $P(z)$ e $Q(z)$ são chamadas LSF's. Os polinômios $P(z)$ e $Q(z)$ têm as seguintes propriedades: 1) todos os zeros de $P(z)$ e $Q(z)$ estão no círculo unitário e 2) zeros de $P(z)$ e $Q(z)$ encontram-se entrelaçados, isto é, os LSF's w_i encontram-se em ordem ascendente em $(0, \pi)$, da forma (Kim and Oh, 1999),

$$0 < w_1 < w_2 < \ldots < w_p < \pi. \tag{3.37}$$

Uma questão importante na codificação dos parâmetros LSF é que a relação de ordenamento é requerida para assegurar a estabilidade do filtro de síntese (Sugamura and Farvardin, 1988). As propriedades 1) e 2) previamente apresentadas ajudam na determinação dos parâmetros LSF a partir de $P(z)$ e $Q(z)$.

Até cerca de 1990, quase todos os esquemas de codificação utilizavam quantização escalar de alguma maneira (Eriksson et al., 1999). Questões de complexidade (número de operações realizadas para comparar um vetor de entrada com cada vetor-código do dicionário) e de requisitos de memória limitavam o uso da quantização vetorial. Tanto a complexidade quanto os requisitos de memória aumentam com o tamanho do dicionário. O primeiro trabalho a incorporar QV foi descrito em (Buzo et al., 1980), mas um desempenho longe do aceitável foi obtido com QV a 10 bits por quadro. Assim, tendo em vista o tamanho proibitivo do conjunto de treino, o elevado custo computacional do treinamento de um dicionário e os requisitos computacionais proibitivos para

quantização vetorial convencional (*full search vector quantization*), têm sido investigados esquemas híbridos de quantização (escalar e vetorial) (e.g. (Grass and Kabal, 1991; Laroia et al., 1991)), bem como esquemas para reduzir a complexidade da QV às expensas de uma perda de desempenho, como por exemplo o esquema proposto por Paliwal e Atal em (Paliwal and Atal, 1993), no qual o vetor de parâmetros LSF é dividido em dois vetores, sendo cada um quantizado por um dicionário distinto (esse procedimento é conhecido como *split VQ*), bem como o esquema de quantização vetorial multi-estágio (MSVQ) (LeBlanc et al., 1993).

Na quantização sem memória, cada vetor de parâmetros LSF é quantizado independentemente de vetores LSF passados. Esse, contudo, não é o modo mais eficiente de codificar vetores LSF, os quais apresentam uma significativa correlação *interquadro* (correlação entre quadros sucessivos). Conseqüentemente, é possível obter ganhos de desempenho de codificação com a exploração da correlação interquadros, o que tem sido feito em alguns trabalhos, dentre os quais citam-se as técnicas de quantização vetorial preditiva (Shoham, 1987) e FSVQ (*finite-state vector quantization*).

Além do tipo de quantização (escalar, vetorial, híbrida) e a inclusão ou exclusão de memória no processo de quantização, diversos aspectos, em geral interrelacionados, afetam o desempenho de um quantizador LSF, tais como (Ramachandran et al., 1995): a medida de distorção usada no projeto do quantizador, o projeto do dicionário, a complexidade da busca, o número de bits, os requisitos de memória para armazenar o dicionário e a robustez aos erros de canal.

CELP (*code excited linear prediction*) (Schoroeder and Atal, 1985) é uma classe de codificadores de voz que apresenta uma boa estratégia para transmissão digital de voz com alta qualidade a baixas taxas. Uma técnica importante dentro dessa classe de codificadores é o VSELP (*vector sum excited linear prediction*) (Electronic Industries Association (EIA), 1989).

3.9 Visão Geral da Codificação de Voz

A compressão de voz na largura de faixa de telefonia tem sido uma intensa área de pesquisa há décadas. Ainda assim, nos últimos anos tem havido um crescimento de interesse e atividade nessa área, com numerosas aplicações em telecomunicações e armazenamento, e muitos padrões estabelecidos.

Praticamente todos os trabalhos em compressão de voz envolvem compressão com perdas, em que a representação numérica das amostras do sinal nunca é recuperada exatamente após a decodificação. Existe uma ampla faixa de compromissos entre a taxa de bits e a qualidade do sinal reconstruído que são de interesse prático na codificação de voz telefônica, em que os usuários estão acostumados a tolerar níveis variados de degradação do sinal.

Os algoritmos de codificação de voz podem ser divididos em duas categorias principais (Gersho, 1994): *codificadores de forma de onda* e *vocoders*. O termo *vocoder* historicamente originou-se da contração de *voice* e *coder*. Em codificadores de forma de onda, os dados transmitidos do codificador para o decodificador especificam uma representação da voz original como uma forma de onda de amplitude versus tempo, de modo que o sinal reproduzido aproxima-se da forma de onda original e, conseqüentemente, fornece uma recriação aproximada do som original. Por outro lado, *vocoders* não produzem uma aproximação da forma de onda original: ao contrário, parâmetros que caracterizam segmentos de som individuais são especificados e transmitidos para o decodificador, que reconstrói uma forma de onda nova e diferente, com um som similar. *Vocoders* são também chamados de codificadores paramétricos por motivos óbvios. Freqüentemente esses parâmetros caracterizam o espectro de curto prazo de um som. Alternativamente, os parâmetros especificam o modelo matemático da produção de voz humana adequado para um som particular. Em todo caso, os parâmetros não fornecem informação suficiente para gerar uma boa aproximação da forma de onda original, mas a informação é suficiente para o decodificador sintetizar um som perceptualmente similar à voz. *Vocoders* operam a taxas de bits inferiores às dos codificadores de forma de onda, mas a qualidade da voz reconstruída, apesar de inteligível, sofre por uma perda de naturalidade e algumas características únicas que identificam um determinado locutor podem ser comprometidas. Portanto, nos codificadores de forma de onda a quantização se realiza dire-

tamente na forma de onda do sinal, ao passo que nos codificadores paramétricos a quantização é realizada nos parâmetros do modelo em consideração. Os codificadores híbridos baseiam-se nos modelos de produção de voz, mas utilizam uma excitação para o sintetizador mais apurada que a excitação utilizada nos codificadores paramétricos (surdo/sonoro) (Aguiar Neto, 1995).

Numerosos trabalhos em codificação de voz são baseados em voz em largura de faixa de telefonia, nominalmente limitada a 3,2 kHz (correspondente à faixa de 200 Hz a 3,4 kHz), amostrada à taxa de 8 kHz. Codificação de voz em banda larga tem tido um interesse crescente e se refere a sinais de 7 kHz, amostrados a 16 kHz.

Grande parte dos trabalhos em codificação de forma de onda de voz diz respeito a modificações e melhoramentos de métodos gerais bem estabelecidos.

Uma técnica de codificação notável e muito popular é o CELP. Como métodos de codificação, podem ser citados: ADM (*adaptive delta modulation*), ADPCM (*adaptive differential pulse code modulation*), APC (*adaptive predictive coding*), MP-LPC (*multipulse linear predictive coding*) e RPE (*regular pulse excitation*). Os codificadores MP-LPC, RPE e CELP pertencem à família dos algoritmos de análise por síntese, que podem ser vistos como codificadores híbridos, por combinarem algumas características de *vocoders* e de codificadores de forma de onda. O codificador de voz de análise por síntese mais usado e conhecido é provavelmente o GSM de taxa completa RPE-LTP 13 kbps, padronizado pelo ETSI em 1988 para o sistema móvel digital celular (Hersent et al., 2002).

Muito embora muitos *vocoders* sejam estudados há décadas, o sobrevivente mais importante é o *vocoder* LPC (*linear predictive coding*), muito usado em voz telefônica segura, constituindo também o ponto de partida de pesquisas em andamento. Outra abordagem de *vocoder* que não pode deixar de ser mencionada é a codificação senoidal (*sinusoidal coding*), cujas versões STC (*sinusoidal transform coding*) e MBE (*multiband excitation*) têm sido estudadas.

Nas seções a seguir são abordadas algumas técnicas utilizadas nos codificadores de forma de onda e nos codificadores paramétricos e híbridos.

3.10 Codificação de Forma de Onda

Os codificadores de forma de onda fazem a quantização diretamente sobre a forma de onda do sinal de voz. Têm como objetivo reproduzir a forma de onda, amostra por amostra, da maneira mais eficiente possível. São codificadores de pequena complexidade de implementação. Geralmente possuem um baixo retardo de voz e são muito apropriados para utilização em redes telefônicas em taxas de bits $R \geq 16$ kbit/s. Em relação à complexidade, geralmente podem ser classificados como de baixa, média e até alta complexidade. Os codificadores de forma de onda de baixa complexidade têm boa qualidade de voz em taxas de, no mínimo, 56 kbit/s. Taxas menores, de até 32 kbit/s, são obtidas com codificadores de média complexidade e taxas menores, de até 16 kbit/s, são obtidas com os codificadores de alta complexidade.

3.10.1 PCM Diferencial

A codificação diferencial explora o fato de que o sinal de voz apresenta correlação significativa entre amostras sucessivas. Isto implica que o sinal de voz é muito redundante. O objetivo da técnica DPCM (*differential pulse code modulation*) é a redução na redundância do sinal de voz. Isso é obtido quantizando a diferença de amplitude entre amostras adjacentes. Como essa diferença apresenta uma menor variação que a apresentada pelo sinal original, pode-se utilizar um menor número de bits para representá-la.

Com o sistema DPCM é possível reduzir a taxa de transmissão para 56 kbps com a mesma qualidade de um PCM. Isto significa uma economia de 1 bit por amostra codificada (Aguiar Neto, 1995).

3.10.2 PCM Diferencial Adaptativo

Os codificadores DPCM podem ter um ganho de desempenho se o processo de quantização e/ou predição for realizado de forma adaptativa. Neste caso o codificador é denominado ADPCM (*adaptive differential pulse code modulation*). A quantização adaptativa consiste em regular a largura dos degraus de quantização em função do nível do sinal: é realizada uma quantização não-uniforme na qual as larguras dos degraus de quantização não são pré-estabelecidas (Aguiar Neto, 1995). A predição adaptativa consiste no ajuste dinâmico dos coeficientes do preditor, de acordo com as variações do sinal de voz. Codificadores ADPCM apresentam boa qualidade de voz para taxas entre 24 e 48 kbit/s.

3.10.3 Modulação Delta

A modulação Delta é um caso especial do DPCM, no qual a variação de amplitude de amostra a amostra é quantizada, usando-se apenas dois níveis de quantização. A saída do quantizador de dois níveis é relacionada com a entrada pela expressão $y = 2du(x) - d$, em que $u(\cdot)$ é a função degrau e d representa o passo de quantização (Alencar, 1998).

Os resultados obtidos indicam que, para o quantizador utilizado no modulador Delta, o sinal de saída retém 64% da potência total de entrada. O ruído de quantização contribui com aproximadamente 36% da potência do sinal quantizado. Para um sinal gaussiano isto implica, para o pior caso, uma SQNR de 2,44 dB (Alencar and Tong, 1993).

As vantagens práticas de utilizar um pequeno número de níveis de quantização, incluindo a quantização com apenas um bit, têm sido indicadas pela popularidade de esquemas de codificação do tipo *sigma-delta* ($\Sigma - \Delta$). Esses esquemas têm ampla aceitação em virtude de serem robustos contra imperfeições dos circuitos e apropriados para implementação em VLSI (*very large scale integration*) (Gray, 1987; Zamir and Feder, 1995).

O ruído de quantização, provocado pelo estágio de quantização do modulador $\Sigma - \Delta$, tem sido analisado por diversas técnicas (Gray, 1989; Galton, 1993). A utilização do ruído pseudo-aleatório (*dither*) e a amostragem a taxas elevadas têm sido objeto de estudos recentes, que demonstram a viabilidade de recuperar um sinal quantizado com dois níveis (Chou and Gray, 1991; Shamai, 1994).

3.11 Codificação Paramétrica e Híbrida

Os codificadores paramétricos, ou *vocoders*, são baseados no modelo de produção da voz. Esse modelo é representado por um conjunto de parâmetros que tem atualização periódica. Para a determinação desses parâmetros, o sinal é segmentado a intervalos periódicos chamados quadros. Os parâmetros são usualmente atualizados a cada quadro. A taxa requerida pelos *vocoders* é baixa (menor que 4,8 kbit/s), mas o atraso e a complexidade são elevados e a voz soa sintética.

Nesse tipo de codificador, ao contrário dos codificadores de forma de onda, o processo de quantização é realizado sobre os parâmetros do modelo de produção de voz que serão utilizados para a sintetização do sinal de voz. Os parâmetros do modelo de voz são determinados em curtos intervalos de tempo, nos quais o sinal de voz pode ser considerado estacionário, e são então transmitidos para o sintetizador no receptor. Esses codificadores não fornecem a qualidade de voz requerida para a rede telefônica. São mais utilizados em aplicações com fins militares.

Codificadores híbridos combinam a qualidade dos codificadores de forma de onda com a eficiência dos codificadores paramétricos. Os codificadores híbridos, mais elaborados que os *vocoders*, são baseados nos modelos de produção de voz e utilizam uma excitação mais apurada para o sintetizador. A melhoria da excitação é responsável pelo melhoramento da qualidade da voz sintetizada, tornando-a mais inteligível que nos *vocoders* convencionais. Esse melhoramento é devido à quantização e codificação dos parâmetros que definem a excitação e aos parâmetros do filtro de síntese. Um processo conhecido como análise por síntese é responsável pela obtenção desses parâmetros utilizados em curtos intervalos de tempo. Esses codificadores são geralmente

complexos e fornecem, para taxas de 4 a 16 kbit/s, uma qualidade superior à obtida com os codificadores de forma de onda, com taxas mais elevadas (Alencar, 1998).

3.11.1 *Vocoder* de Canal

Os *vocoders* de canal foram desenvolvidos em 1928, por Homer Dudley (Dudley, 1993). A implementação original de Dudley comprimia as formas de onda em sinais analógicos com uma banda passante total de 300 Hz. Baseados no conceito original, *vocoders* de canal digital têm sido desenvolvidos operando em uma taxa de 1 a 2 kbit/s. A maior parte dos processos de codificação dos codificadores de canal envolve a determinação do espectro do sinal de voz amostrado como uma função do tempo. No *vocoder* de canal, um banco de filtros passa-faixas é usado para separar a energia do sinal de voz em sub-bandas, que são completamente retificadas e filtradas para determinar os níveis de potência relativos. Os níveis de potência individuais são codificados e transmitidos para o devido destino.

Em adição à medição do espectro do sinal, os *vocoders* de canal modernos também determinam a natureza da excitação da voz e contam as freqüências do som da voz. As medidas de excitação são usadas para sintetizar o sinal de voz no decodificador e são feitas pela passagem de uma fonte apropriada de sinal de um extremo a outro no domínio de freqüência da função de transferência do canal. A excitação da voz é simulada por um gerador de pulso usando uma taxa de repetição igual à taxa de medição da excitação. Devido à natureza sintetizada da excitação, essa forma de *vocoder* é denominada *vocoder* excitado por pulsos.

Muitas variações do *vocoder* de canal básico têm sido desenvolvidas, envolvendo a natureza da excitação do sinal de voz e os meios de codificação dos níveis de potência do sinal a ser codificado. Recentes avanços na tecnologia digital têm introduzido o uso de processamento digital de sinais, para determinar o espectro de entrada por meio de transformada de Fourier. Todas as formas de *vocoders* de canal que medem a densidade espectral de potência são referidas como *vocoders* de canal espectral para destingui-los dos *vocoders* de canal no domínio do tempo, tão bons quanto o *vocoder* preditivo linear, discutido posteriormente.

A maior parte das dificuldades encontradas na grande maioria das implementações dos *vocoders* está na determinação dos harmônicos da voz, pois certos harmônicos não são claramente classificáveis como puramente vocais. Deste modo, uma ampliação mais desejável do *vocoder* básico envolve uma caracterização mais acurada da excitação. Destituído da informação de excitação acurada, a qualidade da saída do codificador é pobre e muitas vezes depende da pessoa que está falando e dos sons particulares a cada voz. Alguns dos mais avançados *vocoders* de canal têm produzido alta inteligibilidade, apesar de produzirem na saída um som sintético, em conversas a uma taxa de 2,4 kbit/s (Bayless et al., 1973).

3.11.2 *Vocoder* de Formante

Como se sabe, os pulsos da densidade espectral da voz são distribuídos espaçadamente sobre toda a faixa de voz (300 Hz a 3400 Hz). Desse modo, a energia da voz tende a se concentrar em três ou quatro picos chamados *formantes*. Um *vocoder* de formante determina a localização e a amplitude desses picos e transmite essa informação em vez de todo o envelope espectral. Portanto, o *vocoder* de formante produz baixa taxa de bits somente pela codificação dos pulsos mais importantes do espectro da voz (Alencar, 1998).

O requisito mais importante para obter na saída de um *vocoder* de formante envolve um rastreamento acurado dos formantes. Uma vez que este requisito é satisfeito, um *vocoder* de formante pode prover inteligibilidade de voz a uma taxa menor que 1 kbit/s (Flanagan et al., 1979).

3.11.3 Codificador Preditivo Linear

O codificador preditivo linear (LPC) é um *vocoder* muito utilizado – extrai características perceptivelmente importantes do sinal de voz diretamente da forma de onda no tempo. O efeito é melhor que o obtido a partir do espectro de freqüência, como no caso do *vocoder* de canal e do *vocoder* de formante. Fundamentalmente, um LPC analisa a forma de onda da voz para produzir um modelo do trato vocal variante no tempo e para produzir

a função de transferência do modelo vocal. Um sintetizador, no terminal de recepção, recria o sinal de voz pela passagem da excitação especificada por um modelo matemático do trato vocal. Pela atualização periódica dos parâmetros do modelo e das especificações de excitação, o sintetizador se adapta às mudanças feitas. A Figura 3.15 apresenta o modelo básico de geração de voz da codificação preditiva linear. Essa figura é também um modelo de um codificador/sintetizador LPC.

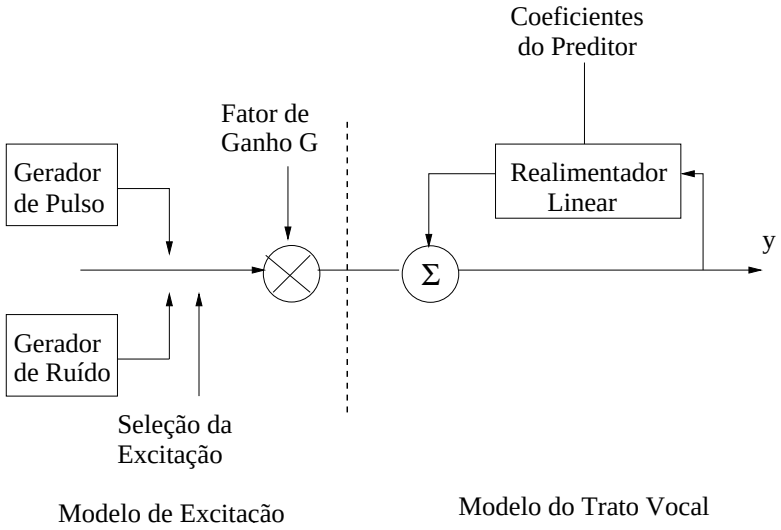


Figura 3.15: Modelo de geração de voz da codificação preditiva linear.

A equação do modelo do trato vocal é definida abaixo:

$$y(n) = \sum_{k=1}^{p} a_k y(n-k) + Gx(n), \tag{3.38}$$

em que $y(n) = n$-ésima amostra de saída, $a_k = k$-ésimo coeficiente do preditor, $G =$ fator de ganho, $x(n) =$ entrada amostrada em um tempo n e $p =$ ordem do modelo.

A saída de voz na Equação 3.38 é representada como a entrada atual do sistema, somada a uma combinação linear da saída predita do trato vocal. O modelo é adaptativo e o codificador determina, periodicamente, o novo conjunto de parâmetros correspondentes aos segmentos sucessivos de voz. Um LPC básico não faz a medição e codificação da diferença de formas de onda. Ao invés disso, os sinais de erro são minimizados em uma média quadrática, quando os coeficientes do preditor são determinados (Alencar, 1998).

A informação que o codificador/analisador LPC determina e transmite para o decodificador/sintetizador consiste em:

1. Natureza da excitação, se a partir de voz ou não;
2. Contagem do período para a excitação da voz;
3. Fator de ganho;
4. Coeficientes do preditor, ou seja, parâmetros do modelo do trato vocal.

A natureza da excitação (segundo a qual é utilizado um gerador de pulso ou um gerador de ruído) é determinada pela verificação de componentes mais intensos da forma de onda. A contagem desses componentes é determinada pela medição de sua periodicidade, quando ela existe. Em adição a essa medição, feita com técnicas similares às usadas por outros *vocoders*, um codificador/analisador LPC tem propriedades particulares que auxiliam na determinação dos componentes mais fortes da forma de onda (Rabiner and Schafer, 1978).

Os coeficientes do preditor podem ser determinados por vários procedimentos computacionais existentes. Todos os procedimentos usam amostras da forma de onda atual, como a saída desejada do sintetizador. Usando esses valores amostrados, um sistema linear de p equações e p incógnitas é produzido. Deste modo os coeficientes são obtidos pela inversão de uma matriz $p \times p$. Desde que a ordem da matriz pode variar de 6 a 12, dependendo da qualidade desejada da voz, a inversão da matriz pode levar a um grande dispêndio computacional. Dependendo das considerações específicas feitas para o modelo do trato vocal, as matrizes podem adquirir propriedades especiais que podem simplificar muito a solução das equações (Rabiner and Schafer, 1978).

Embora os codificadores preditivos lineares preservem a representação do sinal de voz no domínio do tempo, sua operação é tão boa que eles provêem boas estimativas dos picos do espectro do sinal de voz (Alencar, 1998). Além do mais, um codificador LPC é capaz de realizar, efetivamente, mudanças graduais na envoltória espectral. O resultado final é que os LPCs produzem uma representação mais natural da voz que os *vocoders* baseados puramente no domínio da freqüência (Bayless et al., 1973). A maior parte dos LPCs têm concentrado a codificação de voz no intervalo de 1,4 a 2,4 kbit/s.

3.11.4 Codificação Preditiva Linear com Excitação Aumentada

O algoritmo LPC básico sintetiza a voz no decodificador usando um modelo de excitação muito simples, que requer somente 10% da taxa de dados agregada (Alencar, 1998). A simplicidade do modelo, inevitavelmente, produz uma sonorização sintética da voz. Para superar essa falha, várias técnicas têm sido desenvolvidas para aumentar a excitação. Três dos algoritmos de excitação mais utilizados são: o LPC por excitação multipulso (MLPC), LPC por excitação residual (RELP) e LPC excitado por código (CELP).

Codificação Multipulso

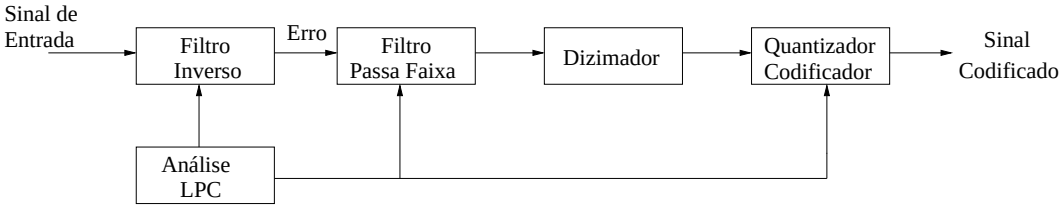
A codificação por excitação multipulso é uma extensão conceitualmente simples do LPC convencional, que usa erro de predição para determinar o período de repetição do sinal de voz. Um MLPC usa a predição residual para determinar uma seqüência de pulsos de excitação que forneça o menor erro possível, obtido pela comparação do sinal original com um sinal sintético no codificador utilizando essa seqüência. O caso mais simples de MLPC usa um número fixo de pulsos em um trem de pulsos e determina somente a fase inicial do trem de pulsos e as amplitudes e polaridades de cada pulso em relação à posição e amplitude dos pulsos. Em geral, a procura pelas posições e amplitudes dos pulsos é realizada a cada 5 ms. Uma característica favorável ao MLPC é que ele não precisa identificar onde um segmento de fala é iniciado; ele determina somente o período de repetição do sinal de voz (*pitch*). Em cada segmento de análise, a codificação multipulso se adapta, automaticamente, à natureza do sinal de excitação atual (Alencar, 1998).

Codificação por Excitação Residual Preditiva Linear

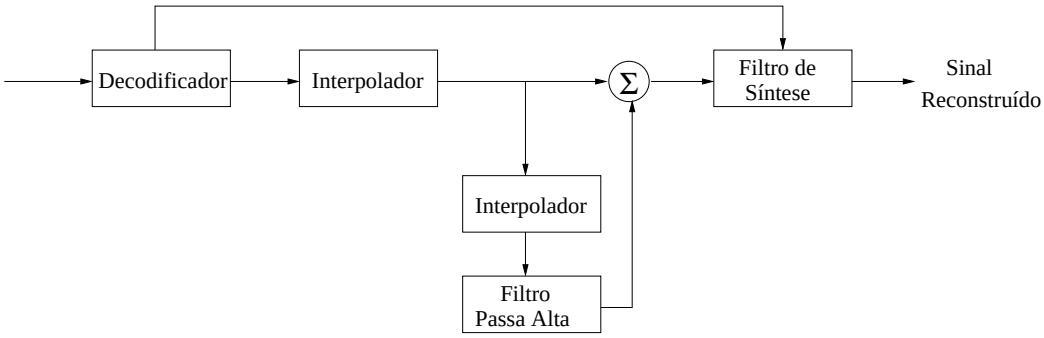
O termo codificador por excitação residual ou RELP se refere a sua estrutura que é idêntica à de um codificador preditivo adaptativo, mas que difere na maneira como o erro de predição é codificado. Um RELP não codifica o erro de predição diretamente, mas processa esse erro em uma predição que permita uma baixa taxa de dados. A propriedade fundamental desse processamento é que o erro de predição consiste em uma componente fundamental e múltiplos harmônicos. Deste modo, um RELP codifica e decodifica somente as componentes fundamentais. O decodificador reconstrói o erro de predição pela decodificação da componente fundamental e pela adição dos harmônicos.

A codificação por excitação residual preditiva linear baseia-se na produção da voz por excitação do filtro de síntese, presente no decodificador, por meio de um sinal chamado resíduo do filtro de análise. O princípio básico de um codificador RELP é mostrado na Figura 3.16.

O sinal resíduo é obtido por análise LPC do sinal original e representa a saída do filtro de predição como definido pela Equação 3.39.



(a) Codificador



(b) Decodificador

Figura 3.16: Sistema de codificação e decodificação RELP.

$$P(z) = 1 - H(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}, \tag{3.39}$$

em que $H(z)$ representa a função de transferência, no domínio da transformada Z, de um filtro inverso para o sistema representado por $P(z)$.

Tanto no filtro de análise quanto no filtro de síntese, os coeficientes são calculados a cada 20 ms. Antes de ser quantizado e codificado para ser transmitido, o sinal de resíduo ou erro de predição é primeiramente filtrado de maneira a limitar o espectro do sinal em 1 kHz, pois nessa faixa de freqüência estão as componentes mais importantes do sinal. É feita em seguida uma subamostragem pelo processo de decimação com o intuito de minimizar a redundância (Alencar, 1998).

No decodificador de um sistema RELP o sinal recebido passa por meio de um interpolador, seguido por um retificador e por um filtro passa-altas para regenerar as componentes de alta freqüência do sinal de erro de predição. A saída do interpolador é então somada à saída do filtro passa-altas para fazer a reconstituição do sinal resíduo que é utilizado como excitação do filtro de síntese.

3.11.5 Codificação por Excitação Linear Preditiva – CELP

Os codificadores CELP são parecidos com os codificadores de multipulso. O filtro de síntese e os preditores são os mesmos. A diferença entre eles é que nos codificadores CELP a seqüência de pulsos de excitação é selecionada de um conjunto de vetores aleatórios, previamente armazenados, formando uma espécie de dicionário (*codebook*), conforme mostra a Figura 3.17. Os vetores armazenados no dicionário possuem distribuição Gaussiana e média zero, com o intuito de aproximar as características estatísticas desses vetores com as características do sinal de voz de curto prazo (Aguiar Neto, 1995).

A escolha da seqüência ideal a ser utilizada como excitação do filtro de síntese no decodificador é feita por um processo de busca em um dicionário, utilizando a técnica de análise por síntese. De forma parecida com

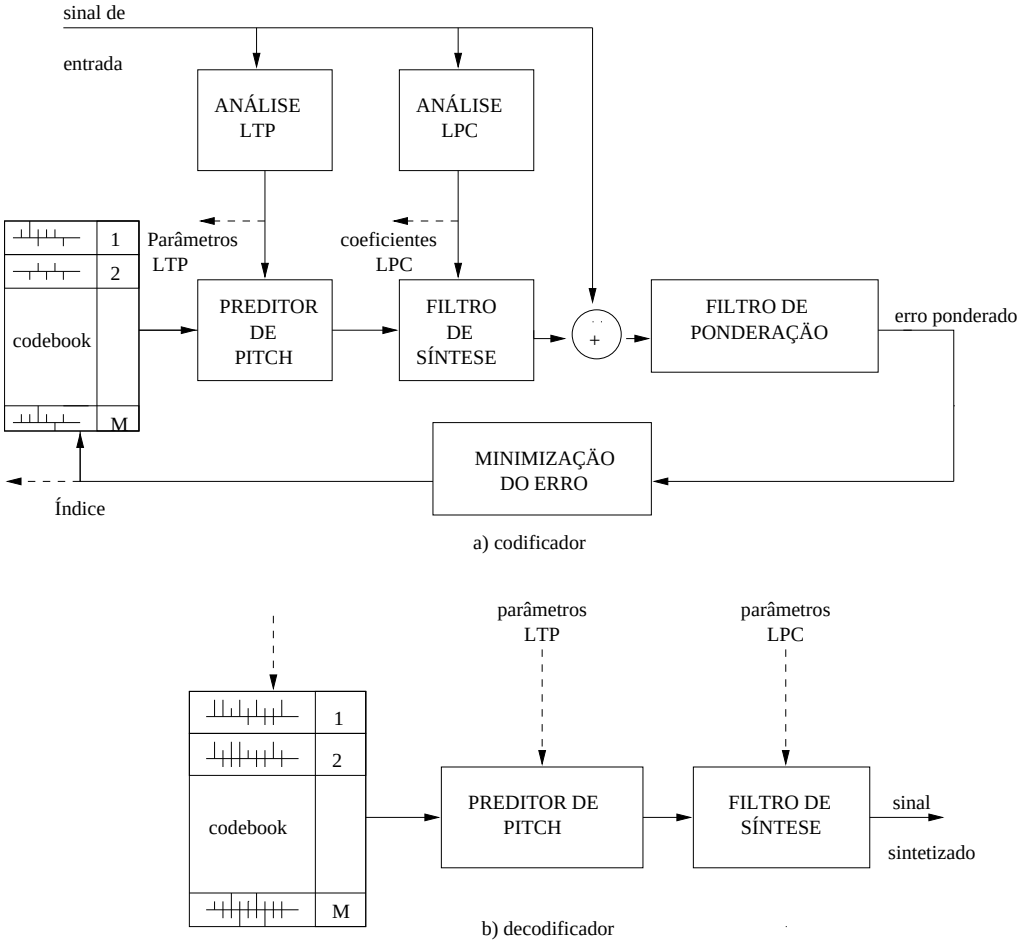


Figura 3.17: Sistema CELP de codificação.

o sistema MLPC, é feita uma análise de correlação do sinal de voz de curto prazao (LPC) e uma análise de correlação de longo prazo que geram, respectivamente, a envoltória espectral (formantes) e a periodicidade de *pitch* do sinal de voz. Com esses parâmetros, o sinal de voz é sintetizado no codificador de forma a compará-lo com o sinal original correspondente. A seqüência extraída do dicionário que proporcionar o menor valor de energia do sinal de erro do processo de comparação é então a seqüência escolhida como a ideal. O endereço dessa seqüência no dicionário é transmitido para o decodificador, que, utilizando um dicionário idêntico ao do decodificador, fornece a seqüência escolhida para a excitação do filtro de síntese do decodificador (Keiser and Strange, 1995).

O filtro de síntese possui de 10 a 12 parâmetros que são atualizados a cada 20 ms e os coeficientes do preditor de *pitch* são atualizados a cada 5 ms. No mesmo intervalo de tempo, é selecionada a seqüência de excitação ideal. Supondo uma freqüência de amostragem de 8 kHz, tem-se para cada intervalo de 5 ms um total de 40 amostras do sinal. Isto sugere vetores no dicionário com 40 amostras cada. Um valor razoável para o dicionário que produz uma boa qualidade de voz é de 1024 vetores, que requer palavras-código de 10 bits para a codificação de cada vetor do dicionário (Aguiar Neto, 1995).

Os codificadores CELP fornecem uma boa qualidade de voz para taxas de 4,8 kbit/s a 16 kbit/s (possuem excelência nesta faixa), apresentando um desempenho melhor em relação à qualidade versus taxa de bits, quando comparados com outros codificadores, como mostra a Figura 3.18. A codificação CELP, no entanto, demanda um esforço computacional muito grande. Durante muito tempo, isto constituiu um obstáculo à implementação prática em tempo real. Simplificações introduzidas em sua estrutura básica (métodos de busca eficientes em

dicionários ou uso de dicionários algébricos) e o surgimento de DSPs modernos, com o aumento de MIPS, operacionalizaram o método de codificação. Vale mencionar que o decodificador é muito mais simples que o codificador (tendo em vista que não há procedimento de busca de análise por síntese), podendo incluir uma filtragem posterior opcional (Hersent et al., 2002).

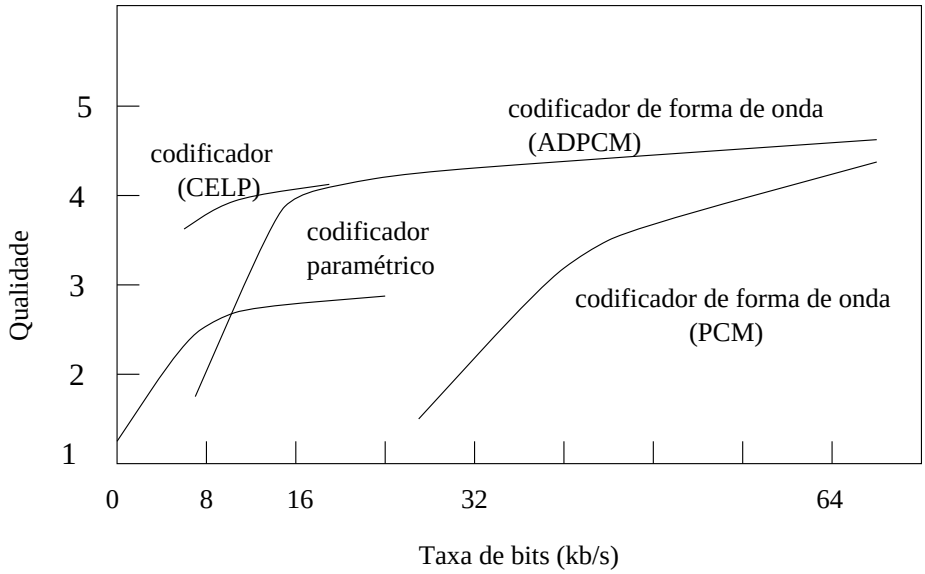


Figura 3.18: Comparação entre codificadores com relação à taxa de bits e qualidade da voz.

Uma modificação realizada no codificador CELP, de forma a reduzir o atraso de cerca de 20 a 40 ms para algo em torno de 2 ms, deu origem ao algoritmo conhecido como LD-CELP (CELP com baixo atraso). Nesse codificador, os vetores de excitação têm um tamanho de apenas cinco amostras, que correspondem a uma seqüência de excitação de 0,625 ms de duração a uma freqüência de amostragem de 8 kHz.

Muitos dos padrões internacionais na faixa de 4,8 kbps a 16 kbps são codificadores de voz CELP ou derivados do CELP, tais como: ITU-T G.729, padrão multimídia ITU-T G.723.1, codificador de voz avançado com taxa completa GSM ETSI e codificador de voz com meia taxa GSM.

3.12 Padrões em Codificação de Voz

Nas últimas duas décadas, o grande número de padrões de codificação de voz que têm sido estabelecidos reflete a maturidade da tecnologia de codificação de voz e a necessidade de satisfazer a crescente demanda por novas tecnologias de comunicação de voz. Esta seção apresenta alguns padrões de codificação de voz.

Uma diversidade de organizações de padronização tem sido responsável pela definição de novos padrões. A ITU (*International Telecommunications Union*) é uma parte da UNESCO (*United Nations Economic, Scientific and Cultural Organization*) e é responsável por estabelecer padrões globais de telecomunicações. Originalmente, o ITU era composto pelo CCITT e pelo CCIR. O CCITT estabelecia padrões de telecomunicações, incluindo padrões de codificação de voz, e o CCIR estabelecia padrões de rádio. Em 1993 o ITU foi reorganizado e as organizações CCITT tornaram-se parte do ITU-T (ITU *Telecommunications Standards Sector*). No ITU-T, o SG15 (*Study Group* 15) é responsável por formular padrões de codificação de voz. O SG12 é responsável por avaliar e caracterizar seu desempenho na rede e trabalha com o SG15 para testar os codificadores. Outros grupos de estudo no âmbito do ITU-T e do ITU-R (ITU *Radio Standarzation Sector*), que era originalmente o CCIR, estabelecem requisitos para que os padrões de codificação de voz atendam suas aplicações (Cox, 1995).

Padrões de telefonia celular digital são estabelecidos por organizações de padronização regionais. Essas organizações são responsáveis pela definição do sistema celular completo, do qual o codificador de voz é uma parte vital, contudo, pequena. Na Europa, o TCH-HS, que é parte do ETSI (*European Telecommunications Standards Institute* ou Instituto Europeu de Padronizações em Telecomunicações), é responsável por estabelecer padrões para o celular digital. Na América do Norte, essa função é desempenhada pela TIA (*Telecommunications Industry Association* ou Associação das Indústrias de Telecomunicações). No Japão, o RCR (*Research and Development Center for Radio Systems*) organiza esses padrões. Os diversos países em outras áreas do mundo têm padronizado sua tecnologia de acordo com os padrões estabelecidos por uma dessas organizações regionais (Cox, 1995).

Outras organizações também podem estabelecer padrões para aplicações específicas. Por exemplo, a Inmarsat (*International Maritime Satellite Corporation*), que regula a comunicação satélite geossíncrona para o mundo, tem estabelecido uma série de padrões para várias aplicações de telefonia baseadas em satélite. Além disso, alguns governos podem criar padrões dentro de seus próprios países. O governo dos Estados Unidos, por exemplo, padronizou a codificação de voz para telefonia segura. Outras organizações internacionais também podem formular padrões para seus países membros, a exemplo da NATO, que também criou padrões para telefonia segura (Cox, 1995).

A seguir, são discutidos alguns atributos dos codificadores de voz, importantes no cenário da padronização dos codificadores de voz.

3.12.1 Atributos do Codificador de Voz

Os codificadores de voz têm atributos que podem ser dispostos em quatro grupos: taxa de bits, qualidade, complexidade e atraso. Para uma determinada aplicação, alguns atributos são pré-determinados, sendo permitidos compromissos entre os demais. Por exemplo, o canal de comunicação pode impor um limite na taxa de bits; considerações de custo podem limitar a complexidade. A qualidade geralmente pode ser aumentada com o aumento da taxa de bits ou da complexidade, e às vezes com o aumento do atraso. Devem ser estabelecidos requisitos para todos esses atributos. Para o propósito da padronização, dois atributos adicionais são: o método de especificação e validação da conformidade; o cronograma no qual o plano será executado (Cox, 1995).

Taxa de bits

Os sinais de voz na largura de faixa de telefonia têm uma largura de faixa de aproximadamente 300-3400 Hz. Os codificadores de voz que têm sido padronizados nos últimos anos têm uma taxa de bits de 800 bps a 16 kbps (Cox, 1995). Alguns desses codificadores, notadamente os que foram padronizados para telefonia celular, também têm um codificador de canal a eles associado. Nesse caso, a taxa de bits aumenta (até, por exemplo, 22,8 kbps). As taxas mais baixas estão associadas principalmente à telefonia segura. Há algum tempo tem se observado uma sobreposição nessa área com aplicações de telefonia baseadas em satélite. As taxas de codificação de voz em telefonia celular cobrem a faixa de 3,3 a 13 kbps.

Sinais de voz de banda larga têm uma largura de faixa de 50-7000 Hz, sendo amostrados a 16 kHz. A largura de faixa mais ampla melhora a naturalidade e a inteligibilidade da voz e também contribui para diminuir a fadiga do ouvinte durante conversações longas. Dentre os padrões para telefonia em banda larga (7 kHz), pode ser citado o G.722, com taxas de 64, 56 e 48 kbps.

Atraso

Codificadores de voz de baixa taxa podem ser considerados codificadores de bloco. Eles codificam um bloco de voz, também denominado de um quadro, por vez. Dependendo da aplicação, o atraso total do sistema referente à codificação de voz é um múltiplo do tamanho do quadro. O atraso mínimo do sistema está geralmente entre três a quatro vezes o tamanho do quadro (Cox, 1995). Por exemplo, muitos codificadores de

baixa taxa têm tamanho de quadro de 20 ms, resultando em um atraso de 60 a 80 ms. Para os propósitos de padronização, o atraso é um fator importante em sistemas conversacionais em tempo real.

Complexidade

A maioria dos codificadores de voz é implementada primeiramente em *chips* DSP (*digital signal processor*) e pode posteriormente ser implementada em dispositivos VLSI de propósito especial. Velocidade e uso de RAM (*randon access memory*) são os fatores que mais contribuem para a complexidade. Quanto mais rápido o chip ou quanto maior o tamanho do chip, maior o custo. Esses mesmos atributos também influenciam no consumo de potência, que constitui um atributo crítico em aplicações portáteis. A complexidade, portanto, é uma questão importante para o custo e para o consumo de potência.

Sob a perspectiva dos padrões de codificação de voz, a complexidade é determinada pela aplicação. Ao se criar um padrão, o grupo de especialistas deve determinar qual o nível máximo de complexidade aceitável.

Velocidade é comumente medida pelo número de instruções por segundo (MIPS, *millions of instructions per second*) necessárias para implementação em tempo real do algoritmo de codificação de voz. A complexidade é geralmente especificada em termos de MIPS e de número de palavras de RAM necessários para uma implementação bem como ROM (*read-only memory*). Codificadores de voz requerendo menos que 15 MIPS são considerados de baixa complexidade. Aqueles que requerem 30 MIPS ou mais são considerados de alta complexidade (Rao et al., 2002). Do ponto de vista do projetista de sistema, mais complexidade resulta em custos mais elevados e mais gasto de potência. Para aplicações portáteis, mais dispêndio de potência significa tempo reduzido entre recargas de bateria ou utilização de baterias maiores, o que implica maiores custo e peso.

Qualidade

O SQEG (*Speech Quality Experts Group*) do ITU tem uma visão muito precisa do que constitui qualidade de voz para telefonia (*network toll quality*). O primeiro padrão de codificação de voz digital foi a recomendação CCITT G.711 para voz PCM a 64 kbps. A distorção introduzida por um *codec* G.711 é considerada uma QDU (*quantization distortion unit*). O SQEG usa a QDU para os propósitos de planejamento de rede.

O segundo padrão de codificação de voz digital foi o G.721, ADPCM a 32 kbps. Esse codificador foi padronizado para ser usado em combinação com um *codec* G.711 em sua entrada e saída. Considera-se que essa combinação tem uma distorção de 3,5 QDU (Cox, 1995). As diretrizes para o planejamento de rede do SQEG apontam para um máximo de 14 QDU para uma conexão internacional e de menos de 4 QDU para uma conexão doméstica (Cox, 1995). Considera-se que o codificador G.728 LD-CELP 16 kbps tem a mesma QDU que o G.721. Esses são codificadores ITU que apresentam *toll quality*.

Codificadores de voz a taxas mais baixas podem levar a uma voz inteligível, mas levam a uma distorção suficiente para causar fadiga dos ouvintes durante uma conversação prolongada. A qualidade de voz mais baixa para aplicações a baixas taxas também pode ser medida pelos resultados de testes subjetivos.

O desempenho dos codificadores de voz a baixas taxas pode ser comprometido quando a entrada contém outros sinais além de voz, tais como ruído ambiental ou música. Codificadores de voz a baixas taxas alcançam sua compressão utilizando modelos de produção de voz. Esses modelos tipicamente não se adequam ao modelamento de voz com um outro sinal. Como conseqüência, o codificador de voz transforma o outro sinal em algo que soa “artificial”. Podem ser ouvidos rangidos, grasnidos, estampidos ou outros efeitos estranhos. Esses artefatos podem não ter sérias conseqüências para o codificador, mas podem em alguns casos inviabilizar o uso do codificador.

Há outras vertentes da qualidade. A sensibilidade aos erros de canal também pode ser considerada como um aspecto da qualidade. No caso dos padrões de celular digital, são usados bits adicionais para codificação de canal, para proteger os bits de informação. Nem todos os bits de um codificador de voz têm a mesma sensibilidade aos erros de canal. É comum ter duas ou três classes de bits mais sensíveis, às quais se dá uma

maior proteção contra erros de canal, enquanto que às classes de bits menos sensíveis aos erros de canal não se dá proteção.

Para os propósitos de preservação de largura de faixa, detetores de atividade de voz são às vezes utilizados com os codificadores de voz. Durante os intervalos sem voz, há uma descontinuidade do *bit stream*. No receptor, é adicionado um "ruído de conforto" para simular o ruído acústico ambiental no codificador. Esse método é utilizado por alguns sistemas celulares e também em sistemas DSI (*digital speech interpolation*) para aumentar o número efetivo de canais ou circuitos (Cox, 1995). Muitas chamadas internacionais que utilizam cabos submarinos ou satélites utilizam sistemas DSI.

3.12.2 Organismos de Padronização

Esta seção apresenta uma breve descrição de vários organismos de padronização: ITU-T, TIA e T1A1 na América do Norte, ETSI na Europa e RCR no Japão.

ITU

O ITU-T (*Telecommunications Standardization Sector of the International Telecommunications Union*) é responsável por criar padrões de codificação de voz para telefonia de rede. Isto inclui as redes com e sem fio. No ITU os padrões são referidos formalmente como recomendações. O ITU-T é organizado em grupos de estudo (*Study Groups*) relacionados a diferentes áreas. O SG15 (*Study Group* 15) tem a responsabilidade de criar recomendações de codificação de voz. No entanto, outros grupos de estudo do ITU ou organizações nacionais podem requisitar que uma recomendação seja criada para uma aplicação particular (Cox, 1995). Por exemplo, o *Study Group* 14, que tem a responsabilidade pelas recomendações de *modem*, já requisitou que uma recomendação fosse criada para um codificador de voz para uso em aplicações com voz e dados simultâneos. O Grupo de Estudos 15 também pode se propor a criar um recomendação por si só.

No ITU, primeiramente um documento formal é criado, denominado "Termos de Referência" (ToR, *Terms of Reference*). Esse documento lista as aplicações pretendidas para a nova recomendação. Essas aplicações, por sua vez, determinam os requisitos para o codificador. Além dos requisitos também pode haver objetivos. Tratam-se de objetivos desejáveis mas que não mandatórios. Finalmente, o ToR inclui um plano de trabalho com uma agenda. Com base nos requisitos e objetivos, um plano de teste pode ser formulado. Esse plano geralmente contém os testes subjetivos para medir a qualidade da voz e as medidas objetivas para outros atributos (Cox, 1995).

Enquanto o SG15 tem a responsabilidade geral pela criação do ToR, o SQEG (*Speech Quality Experts Group*) do SG12 ajuda a estabelecer os alvos concernentes à qualidade de voz e é responsável por construir e conduzir o programa de testes que determina se os codificadores candidatos atendem os requisitos. Busca-se sempre realizar testes em pelo menos três línguas diferentes para que haja variabilidade suficiente no material de teste. Tanto a criação do plano de teste quanto o teste subseqüente são feitos à base de voluntariado. Assim, leva-se algum tempo até encontrar um número suficiente de laboratórios que prestem seus serviços voluntários.

Organizações de Padronização Norte-Americanas

O ANSI (*American National Standards Institute*) tem a responsabilidade norte-americana por todos os padrões. No âmbito do ANSI estão duas organizações que promulgam padrões de codificação de voz para aplicações específicas: a T1A1, que faz padrões para as telecomunicações norte-americanas, e a TIA (*Telecommunications Industry Association*), que faz padrões para telefonia celular e outras aplicações.

A TIA organiza seus comitês por aplicação. O Comitê TR-45 é responsável pelos padrões da telefonia digital celular norte-americana. O Subcomitê TR-45.3 é responsável pelos padrões de TDMA (*time division multiple access*), ao passo que o Subcomitê TR-45.5 é responsável pelos padrões de CDMA (*code division multiple access*). O Comitê TR-30 é responsável por padrões de *modems* e voz e dados simultâneos.

ETSI

O ETSI (*European Telecommunications Institute*) é o equivalente do ANSI, com a ressalva de que seus associados incluem países e companhias européias. O ETSI é um desdobramento da CEPT (*Conference of European Posts and Telephones*). O ETSI é fundamentalmente uma organização de fabricantes de equipamentos, enquanto a CEPT é atualmente apenas uma organização de provedores de serviços. O ETSI patrocina organizações de padronização específica como também organizações de pesquisa para as futuras necessidades de telecomunicações.

O principal exemplo de uma organização de padronização dentro da ETSI que teve um grande impacto na codificação de voz é o TCH-HS. Esse era originalmente conhecido como GSM (*Groupe Speciale Mobile*) e criou o padrão celular digital pan-europeu TDMA em 1987. Vale mencionar que o GSM foi formado antes que o ETSI surgiu da CEPT.

O ETSI é organizado por aplicação. Os documentos do ETSI não são considerados públicos até e a menos que o ETSI os libere. Membros do ETSI não são livres para compartilhar os documentos do ETSI com não-membros.

RCR

Os padrões de celular digital do Japão são criados pelo RCR (*Researh and Development Center for Radio Systems*), que constitui um corpo muito fechado.

3.12.3 Padrões de Codificação

Nesta seção são apresentados padrões de codificação de voz, sendo apresentadas algumas características dos codificadores de voz destacados.

G.711 64 kbps PCM

O CCITT padronizou o codificador PCM (com compansão) a 64 kbps. Na realidade, o CCITT padronizou dois codificadores. No Japão e Estados Unidos, o PCM com lei μ é utilizado. No resto do mundo, o PCM com lei A é utilizado. Ambos codificadores utilizam 8 bits para representar o sinal e têm uma relação sinal-ruído efetiva de 35 dB.

G.721, G.723, G.726 e G.727 ADPCM

O CCITT padronizou inicialmente o G.711 32 kbps ADPCM em 1984, tendo dois propósitos. Primeiro, foi pretendido para uso em DCME (*digital circuit multiplication equipment* ou equipamento de multiplexação de circuitos digitais). Isso representava um aumento de capacidade de $2:1$ em tais sistemas. Quando combinado com DSI (*digital signal interpolation*), que assegura uma vantagem de $2,5:1$, podia ser usado para aumentar a capacidade de cabos submarinos e *links* satélite por um fator de 5. A segunda razão para sua criação foi que esses *links* freqüentemente encontravam o problema de ter o PCM lei μ em um ponto e lei A em outro. O G.721 foi criado para aceitar tanto PCM com lei μ quanto PCM com lei A como entradas. É interessante o fato de que não tenha sido criado para aceitar PCM linear como entrada. O G.721 teve que ser repadronizado em 1986 em virtude de pequenos problemas da versão inicial. O G.721 32 kbps ADPCM foi selecionado para uso em padrões DECT (*digital european cordless telephone*) e CT2 (*cordless telephone* II), que utilizam formas de TDMA como esquemas de acesso.

O G.723 foi padronizado em 1988. Foi pretendido apenas para aplicações DCME e é mais uma padronização de ADPCM, para duas taxas de bit adicionais: 24 e 40 kbps.

O G.726 representa uma unificação do G.721 e do G.723. Trata-se do codificador com base em forma de onda padronizado mais comum, excetuando o PCM Lei-A ou Lei-μ ITU-T G.711. O codificador e o decodificador G.726 executam por volta de 4 MIPS em um DSP de ponto fixo de 16 bits (Hersent et al., 2002). Uma taxa de bits adicional, de 16 kbps, foi incluída, tendo mais uma vez DCME como principal aplicação. Os codificadores de 16 e 24 kbps não são de *toll quality*. No entanto, se eles são usados apenas de forma esparsa durante curtos períodos de sobrecarga, a qualidade geral da voz DCME ainda é *toll quality*. Com a adoção do G.726, as recomendações G.721 e G.723 foram tecnicamente removidas da lista dos padrões ITU correntes.

Uma característica interessante do codificador ADPCM é sua relativa insensibilidade a erros de bit quando comparado ao PCM (Hersent et al., 2002).

O G.727 inclui as mesmas taxas do G.726, mas todos os quantizadores têm um número par de níveis.

G.728 16 kbps LD-CELP

O programa de trabalho do G.728 foi iniciado pelo CCITT em 1998. Pretendia-se na ocasião criar um padrão universal de codificação de voz a 16 kbps com *toll quality*. Entendia-se por *toll quality* que o padrão deveria atingir ou ultrapassar o desempenho do G.721 ADPCM 32 kbps. Entendia-se por universal que deveria ser usado em qualquer lugar. Mais tarde o CCITT restringiu seu uso na rede telefônica com cobrança (*toll network*). A primeira aplicação do G.728 foram os videofones H.320 de baixa taxa de bit.

O desempenho do G.728 foi amplamente avaliado pelo SQEG. Observou-se que o padrão fornece um desempenho robusto a sinais com ruído de fundo ou música. O padrão é muito robusto a erros de bit aleatórios, mais que os padrões CCITT antecedentes G.711 e G.721.

O sistema G.728 tem um MOS em torno de 4. É usado no sistema H.320 de videoconferência, para substituir o G.711 64 kbps e possibilitar uma videoconferência que utiliza um só canal ISDN tipo B (64 kbps), deixando praticamente 48 kbps para vídeo (Hersent et al., 2002).

Padrões ITU-T de Codificação de Voz para Multimídia

O ITU-T padronizou três codificadores de voz que se aplicam a comunicações multimídia a baixas taxas (Rao et al., 2002). A recomendação ITU G.729 8 kbps CS-ACELP (*conjugate-structure* ACELP, ou ACELP de estrutura conjugada) foi originalmente projetada para aplicações sem fio, mas também se aplica a comunicações multimídia. O termo "estrutura conjugada" deve-se ao fato de que, para a codificação, o resíduo de predição linear é quantizado por meio de quantização vetorial de dois estágios (estrutura conjugada) (Hersent et al., 2002). O Anexo A da recomendação G.729 é uma versão de complexidade reduzida (10 MIPS para o codificador, inferior aos 18 MIPS do G.729) do codificador CS-ACELP. Foi projetada explicitamente para aplicações de voz e dados simultâneos (DSVD, *Digital Simultaneous Voice and Data*) que prevalecem em comunicações multimídia à baixa taxa de bits. A recomendação ITU G.723.1 com codificador de voz a 6,3 e 5,3 kbps (cuja complexidade em ponto fixo está em torno de 16 MIPS) para comunicações multimídia foi projetada originalmente para videofones à baixa taxa de bits. O G.723.1 foi selecionado para se tornar o codificador de voz padrão para voz sobre IP pela IMTC. O G.723.1 utiliza detecção de atividade de voz (VAD, *voice activity detection*), transmissão descontínua (DTX, *discontinuous transmission*) e geração de ruído confortável (CNG, *comfort noise generator*) (Hersent et al., 2002).

A seguir são apresentados alguns requisitos para o G.729. Qualidade (sem erros de bit): não inferior à apresentada pelo G.726 32 kbps. Erros de bit aleatórios (taxa de erro de bit menor que 10^{-3}): desempenho não inferior ao apresentado pelo G.726. Dependência do falante: desempenho não inferior ao apresentada pelo G.726 32 kbps. Quanto aos objetivos, exemplificam-se os seguintes. Capacidade de transmitir música: nenhum efeito incômodo gerado. Capacidade de transmitir tons de sinalização/informação: distorção a menor possível.

Quanto ao requisito, relatado no parágrafo anterior, de qualidade não inferior à apresentada pelo G.726, testes separados realizados com os codificadores G.729, G.723.1 e a versão DSVD do G.729 mostraram que esses três codificadores satisfazem esse requisito. Em se tratando de testes DCR (*degradation category rating*),

o G.729 recebeu escores menores que G.726. Entretanto, o MOS do G.729 nunca foi significativamente inferior ao do G.726, sendo às vezes até melhor. Todos os três codificadores codificam sinais de música, mas a qualidade da música é pobre. O desempenho global dos três codificadores foi semelhante. Os codificadores G.723.1 e G.729A (versão DSVD do G.729) parecem ser levemente menos robustos a ruído de fundo (Rao et al., 2002).

A especificaç ao G.729 também é recomendada para uso em voz sobre um sistema *frame relay*. A versão de baixa complexidade, G.729A (8 kbps, 10 MIPS), às vezes é usada em sistema VoIP (*voice over IP* ou voz sobre IP) (Hersent et al., 2002).

Codificação de Voz de Banda Larga

O codificador de voz em banda larga CCITT G.722 7 kHz é usado principalmente em teleconferência e videoteleconferência (Cox, 1995), em conjunto com a recomendação H.320. As pessoas tendem a achar a largura de faixa 50-7000 Hz menos cansativa que a largura de faixa de telefonia, 200-3400 Hz. A largura de faixa mais ampla aumenta a inteligibilidade da voz, especificamente para sons fricativos como o /f/ e o /s/, que são difíceis de distinguir na largura de faixa de telefonia. O codificador G.722, que data de 1988, utiliza codificação de sub-bandas, com taxas de bit de 64, 56 e 48 kbps, tem um atraso de 1,5 ms, complexidade de 10 MIPS e 1K RAM.

O G.722, assim como qualquer outro codificador ADPCM, é pouco sensível a erros de transmissão de bits e mais robusto que as seqüências PCM. Sua maior vantagem em relação aos codificadores modernos existentes é seu baixo atraso (Hersent et al., 2002).

O codificador de voz 7 kHz ITU-T teve como aplicações pretendidas: telefonia em banda larga ISDN, videotelefonia ISDN e videoconferência em taxas de acesso básicas, aplicações de pacotes em redes de faixa larga ISDN ou ATM, multiplex de circuito digital, aplicações PSTN (*public switched telephone network*) via *modem*, e envio de mensagens (*messaging*). Esse padrão data de 1998, tem dois modos (A e B, este último constituindo uma opção de menor complexidade) e operando a taxas de 24 kbps e 16 kbps tem, respectivamente, o desempenho do G.722 56 kbps e do G.722 48 kbps.

Codificadores de Voz de Telefonia Digital Celular na Europa

O codificador GSM 13 kbps RPE-LTP (*Regular Pulse Excitation with Long-Term Prediction*) foi padronizado pelo *Groupe Special Mobile* (GSM) do CEPT em 1987 para a telefonia digital celular pan-européia. Esse sistema de codificação usa um pulso de excitação regular (RPE, *Regular Pulse Excitation*) com uma predição de longo prazo (LTP, *Long-Term Prediction*). O uso de RPE para codificar o sinal residual leva a uma implementação de baixa complexidade comparada a alguns sistemas de busca multipulso. O RPE-LTP é apropriado para sistemas de comunicações móveis em virtude de sua elevada robustez a erros de transmissão (Hersent et al., 2002). A taxa total é de 22,8 kbps: os 9,8 kbps adicionais são utilizados para codificação de canal, para proteger o codificador dos erros de bit no canal rádio. Além da telefonia celular, esse codificador tem sido usado para outras aplicações, como envio de mensagens, em virtude de sua baixa complexidade, de 4,5 MIPS e 1K RAM. O MOS do padrão RPE-LTP é em torno de 3,8.

Tendo em vista que o padrão GSM RPE-LTP é dirigido principalmente a aplicações móveis, utilizam-se detecção de atividade de voz (VAD), transmissão descontínua (DTX) e geração de ruído confortável (CNG) para economizar energia no terminal (Hersent et al., 2002).

O codificador de voz GSM de meia-taxa foi padronizado pelo TCH-HS (o sucessor do GSM) com o objetivo de duplicar a capacidade do sistema celular GSM. Trata-se de um codificador VSELP 5,6 kbps. Um maior percentual de bits é utilizado para proteção uma vez que o canal de meia-taxa tem menos diversidade que a usada no sistema com taxa-total. O desempenho geral é similar ao do RPE-LTP, exceto para alguns sinais com ruído de fundo (Cox, 1995). Data de 1994 e tem uma complexidade de 30 MIPS e 4K RAM.

Codificadores de Voz de Celular Digital Norte-Americanos

O codificador IS54 7,95 kbps VSELP foi padronizado pela TIA para telefonia celular digital TDMA. Datando de 1989, esse codificador, submetido pela Motorola, é uma parte do padrão IS54 (*Interim Standard 54*), tem uma complexidade de 20 MIPS e 2K RAM.

O codificador IS96 8,5 kbps QCELP foi padronizado pela TIA para telefonia celular digital CDMA. É uma parte do IS96 e é usado no sistema especificado pelo IS95. A capacidade do sistema CDMA é sua característica mais atrativa. Para alcançar essa capacidade, quando não há voz a taxa dos canais é reduzida. O IS96 QCELP é um codificador de taxa variável que usa interpolação de sinal digital (DSI) para alcançar essa redução de taxa. Vale mencionar que o uso de DSI em celular diminui a qualidade de voz, especialmente em condições de elevado ruído de fundo. As taxas correspondem a 8,5, 4, 2 e 0,8 kbps. Na maior parte dos períodos com voz, o codificador opera a 8,5 kbps. Quando não há voz no canal, a taxa cai para 0,8 kbps. A essa taxa, o codificador está apenas fornecendo estatísticas a respeito do ruído de fundo. Essas taxas são as mais freqüentemente utilizadas durante a operação do IS96. Esse codificador data de 1993, tem uma qualidade inferior à do IS54, com uma complexidade de 20 MIPS e 2K RAM.

Codificadores de Voz de Telefonia Celular Digital no Japão

Abordemos inicialmente o padrão celular digital japonês VSELP. Esse codificador foi padronizado pelo RCR para o serviço de telefonia celular digital TDMA no Japão como padrão RCR STD-27B. O codificador foi submetido pela Motorola e é semelhante ao IS54 VSELP. Data de 1990, opera a 6,7 kbps e tem uma complexidade de 20 MIPS e 2K RAM.

O codificador JDC de meia-taxa 3,45 kbps PSI-CELP foi padronizado pelo RCR para duplicar a capacidade do sistema PDC (*Personal Digital Cellular*) TDMA japonês. Esse codificador, de meia-taxa, PSI-CELP (*Pitch Synchronous Innovation* CELP), foi padronizado em 1993. Tem mesma qualidade que o RCR PDC de taxa total. Sua complexidade, no entanto, é maior.

Codificador Inmarsat 4,15 kbps IMBE

Esse codificador foi padronizado pela *International Maritime Satellite Corporation* (Inmarsat) em 1990 para comunicações mundiais navio-costa via satélite (Inmarsat-M). É também utilizado para telefonia móvel via satélite. A taxa total de bits é 6,4 kbps, em virtude da codificação de canal. Esse codificador IMBE (*Improved Multi-band Excitation*) é um codificador paramétrico ou codificador baseado em modelo: baseia-se na classificação das faixas de freqüência em vozeadas (sonoras, *voiced*) ou não-vozeadas (surdas, *unvoiced*). Vale mencionar que quando o modelo no qual a IMBE se baseia não mais adere ao sinal de entrada, a qualidade de voz resultante é bastante reduzida. Em particular, isto ocorre quando há música ou ruído misturado com o sinal de voz. A qualidade desse padrão é inferior à do GSM. A complexidade é de 7 MIPS.

3.13 Exercícios

1. Fale sobre compressão de sinais, destacando sua importância e seus objetivos e apontando suas principais aplicações.

2. Apresente uma breve abordagem de um sistema de comunicação digital, destacando a função da codificação de fonte, da codificação de canal e da modulação.

3. Enuncie o Teorema da Amostragem, apresentando uma breve análise matemática do processo de amostragem. Em seguida comente: se a freqüência de amostragem for inferior à freqüência de Nyquist, o sinal não poderá ser recuperado completamente.

4. Discuta os aspectos utilizados para avaliar o desempenho de um sistema de compressão de sinais.

5. Apresente as principais características dos sinais de voz. Algumas dessas características são exploradas eficientemente nos sistemas de codificação de voz? Explique.

6. A modulação por codificação de pulsos (PCM) transforma um sinal analógico em uma série de pulsos binários. Qual a mínima freqüência de amostragem para um sinal com 3,4 kHz de banda?

7. O erro médio quadrático de um quantizador escalar uniforme é dado aproximadamente por $d^2/12$. Quais as justificativas usadas para se chegar a esse resultado?

8. O erro, ou ruído, de quantização escalar provoca perda na qualidade do sinal. Qual a melhoria, em termos de relação sinal/ruído de quantização (SQNR), que se obtém com a utilização de 2 bits adicionais no processo de codificação? Explique.

9. Defina matematicamente a quantização vetorial. Discuta sua importância no cenário da codificação de sinais.

10. Em se tratando de quantização LPC, o que se entende por quantização transparente?

11. Como você compararia os codificadores de forma de onda de voz com os codificadores paramétricos e híbridos sob o ponto de vista de qualidade do sinal reconstruído versus taxa de bits?

12. A modulação por codificação de pulsos (PCM) transforma um sinal analógico em uma série de pulsos binários. Explique o processo e mostre por que as etapas de amostragem e codificação não introduzem distorção apreciável no sinal.

13. Analise o funcionamento do ADPCM e explique por que esse esquema permite uma redução substancial na taxa de transmissão da fonte.

14. Fale sobre modulação Delta.

15. Fale, resumidamente, sobre os codificadores paramétricos. Qual a diferença básica entre esses e os codificadores de forma de onda?

16. Explique por que os sistemas de telefonia móvel celular necessitam de codificadores mais eficientes que os sistemas fixos.

17. Discuta os seguintes atributos dos codificadores de voz: taxa de bits, atraso, complexidade e qualidade.

18. Fale sobre as principais organizações para padronização dos codificadores de voz.

19. Quais os padrões de codificação de voz que você conhece? Apresente as principais características de cada um.