7

ÁUDIO PARTE 7
CODIFICAÇÃO E TRANSMISSÃO

Fabio Montoro
Revisado em 12-3-2015

7.1 Introdução

Vamos abordar aqui a codificação digital do sinal analógico, ou seja, a transformação do sinal analógico em um sinal digital.

O processo de digitalização de um sinal analógico possui três etapas:

- **Amostragem**: primeiramente se executa a amostragem do sinal analógico em uma frequência que seja pelo menos o dobro da frequência superior da faixa que se deseja capturar e reproduzir posteriormente (teorema de Nyquist-Shannon).

Harry Nyquist

Claude Shannon

- **Quantização:** cada valor de tensão amostrado é convertido para a forma digital para depois ser serializado segundo algum protocolo.

- **Codificação:** esta etapa é a codificação propriamente dita, na qual o sinal sofre um processamento, podendo ou não ser aplicada alguma compressão.

Para sinais de áudio, confinados em uma faixa de frequência de até 20 kHz, teoricamente, a frequência de amostragem deve ser, no mínimo, 40 kHz.

A máxima relação sinal-ruído da codificação é função da quantidade de bits da conversão AD. Quanto mais bits, melhor. Considerando um fator de pico de 4,8 dB, temos:

$$SNR_{\max(CR=4,8)} = 20 \cdot \log(2^N)$$

O mínimo aceitável, pelos padrões de qualidade atuais, é o padrão CD, com amostragem em 16 bits e frequência de 44,1 kHz:

$$SNR_{CD(CR=4,8)} \geq 20 \cdot \log(2^{20}) \cong 96\ dB$$

Se a amostragem for a 20 bits:

$$SNR_{20(CR=4,8)} \geq 20 \cdot \log(2^{20}) \cong 120\ dB$$

Alguns exemplos relação sinal-ruído máxima:

Normalmente os equipamentos profissionais digitalizam em 24 bits:

$$SNR_{profissional(CR=4,8)} \geq 20 \cdot \log(2^{24}) \cong 144\ dB \qquad [7.3]$$

- Fases
  - Amostragem
  - Quantização
  - Codificação
    - sem compressão
    - com compressão
      - sem perda (*lossyless*)
      - com perda (*lossy)*

A codificação sem compressão teoricamente mantém as características originais do sinal analógico mas, matematicamente, ainda possui redundâncias que podem ser eliminadas por algum processo de compressão sem perda.

Compressão sem perda

Os processos de compressão sem perda não conseguem reduzir tanto o tamanho do arquivo pois devem manter as informações essenciais do sinal original.

Compressão com perda

Os processos de compressão com perda não possuem compromisso com a manutenção das informações essenciais do sinal original e focam principalmente em reduzir o tamanho do arquivo.

7.2 Teorema da amostragem

O teorema da amostragem, um dos conceitos mais importantes em toda a teoria das comunicações, determina que a menor frequência em que se pode amostrar um sinal *s(t)*, de banda limitada, preservando todas as informações que esse sinal carrega, é o dobro da maior componente de frequência do espectro de *s(t)*.

A amostragem não precisa necessariamente ser periódica mas, no limite, ou seja, se quisermos amostrar com a menor frequência possível, sim.

Vamos supor que a amostragem possui um período constante (é o que se faz na prática) e igual a "T_a". Uma vez amostrado, o sinal *s(t)* se transforma em um sinal discreto no tempo. Vamos chamar de $s_a(t)$.

O sinal discreto só existe em instantes espaçados de "T_a" segundos e tem o valor exato do sinal *s(t)* naquele instante, ou seja, possui amostras do sinal *s(t)*, tomadas a cada " T_a " segundos. Então:

$$s_a(t) = \begin{cases} s(t) & para \quad t = nT_a \\ \\ 0 & para \quad t \neq nT_a \end{cases}$$

Colocado de forma mais rigorosa, o teorema teria a seguinte redação:

"Se um sinal contínuo no tempo, *s(t)*, possui banda limitada (transformada de Fourier, *X(jω)*, limitada em frequência, ou seja: *X(jω) = 0* quando $\omega \geq \omega_M$ onde ω_M é a maior componente do espectro) então *x(t)* poderá ser representado se forma única, por amostras igualmente espaçadas no tempo, desde que

$$T_a < \frac{1}{2 \cdot f_M} \qquad \text{ou} \qquad f_a > 2 \cdot f_M$$

onde T_a é o período de amostragem, $\omega_M = 2 \cdot \pi \cdot f_M$ e $f_a = 1/T_a$

A figura 7.1 mostra o espectro de um sinal *x(t)*, limitado em banda.

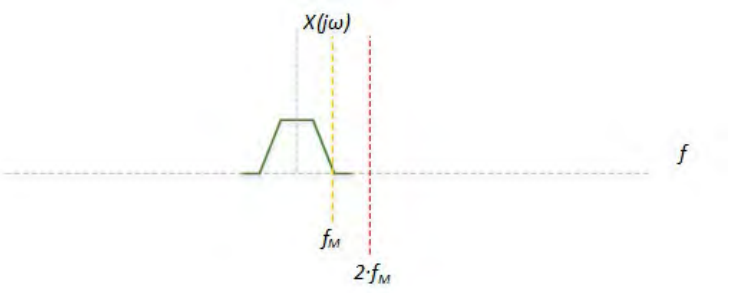

Fig. 7.1: Espectro de um sinal de banda limitada

A frequência máxima do espectro é f_M. A figura 7.1 mostra também o dobro desta frequência, que é o limite mínimo da frequência de amostragem.

A figura 7.2 mostra que o espectro do sinal amostrado $x_a(t)$, para $f_a > 2f_M$ é uma série de réplicas do espectro original, deslocado para $\pm n \cdot f_a$.

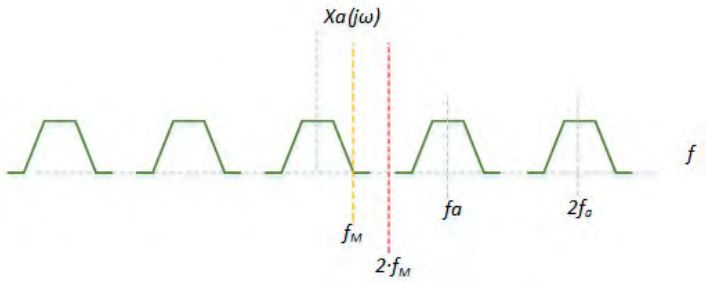

Fig. 7.2: Espectro de um sinal amostrado com $f_a > 2f_M$

Pela figura 7.2, pode-se ver que o sinal original pode ser obtido pela filtragem do espectro do sinal amostrado, com um filtro passa-baixa cuja frequência de inflexão esteja entre f_M e $2f_M$.

A figura 7.3 mostra o espectro do sinal amostrado $x_a(t)$, para $f_a < 2f_M$. Neste caso, as réplicas se sobrepõem, tornando impossível a reconstrução do sinal original.

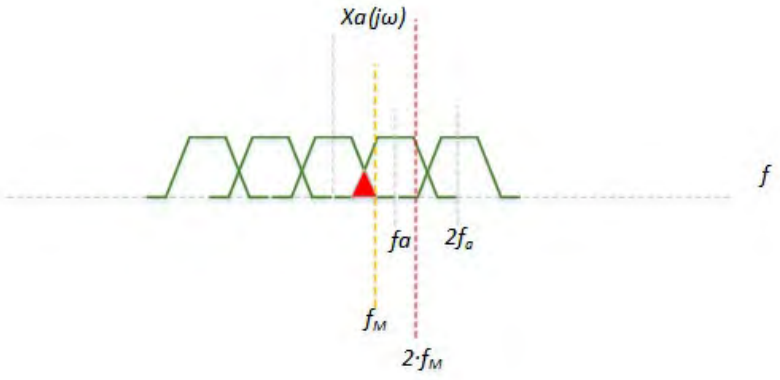

Fig. 7.3: Espectro de um sinal amostrado com $f_a < 2f_M$

7.3 Voz e música

Com relação ao conteúdo do sinal, há dois grandes grupos, com características intrínsecas bem diferentes, além da faixa de frequência ocupada por cada um deles. Essa distinção é importante pois o tipo de codificação pode ser personalizado para aplicações

específicas. Por exemplo: transmissão de voz via sistema de telefonia celular, gravação de um discurso, música com alta compressão, música de alta qualidade, etc.

Os dois grandes grupos e as respectivas faixas ocupadas em frequência, são:

Aplicação da codificação	Voz.....	300 a 3.400 Hz
	Música.....	20 a 20.000 Hz

7.4 Quantização do sinal de voz

Para transmitir ou armazenar o sinal amostrado em equipamentos digitais, é preciso transformar cada amostra $s_a(nT_a)$, que é o valor da voltagem no instante "nT_a", em um número binário, ou seja, é preciso transformar cada amostra, de analógico para digital. Essa técnica é chamada de PCM ("*Pulse Code Modulation*").

No processo de conversão para digital, escolhe-se uma quantidade "N" de bits a ser utilizada para codificar a amostra. O valor analógico será aproximado para o valor digital mais próximo, sendo que há 2^N = M possibilidades. A diferença entre a amplitude do sinal original amostrado, ou seja, a amplitude de $s_a(nT_a)$, e o sinal representado em binário, é o erro de quantização.

Naturalmente, deseja-se que o erro seja tão pequeno ao ponto de não prejudicar a qualidade do sinal quando ele for reconvertido para o formato analógico e reproduzido. Ao mesmo tempo, deseja-se que a quantidade de bits utilizada em cada amostra seja a menor possível a fim de reduzir o tempo de transmissão e o espaço de memória para armazenar a informação e, ainda, deseja-se que tudo isso seja implementado com o menor custo possível.

Ora, não dá pra querer o melhor dos três mundos!

Será preciso estabelecer um compromisso entre eles.

Esse compromisso estará diretamente ligado à aplicação e ao objetivo de marketing do fabricante do equipamento.

Ainda há uma quarta dimensão nesse universo compromissado: a demanda do mercado por equipamentos compatíveis, que se comunicam independentemente da marca, e que a gravação possa ser ouvida 20 anos depois. Os fabricantes não podem desconsiderar essa questão.

Então, os aspectos a serem compromissados pelo sinal digitalizado são:

- ✓ menos bits
- ✓ mais qualidade
- ✓ menor custo
- ✓ mais compatibilidade

7.4.1 Quantização linear

O sinal de voz, limitado entre 300 e 3400 Hz, deve ser amostrado, teoricamente, em pelo menos a 6800 Hz para manter suas características. Na prática utiliza-se a frequência de 8000 Hz, para garantir as variações de filtros e outros circuitos envolvidos.

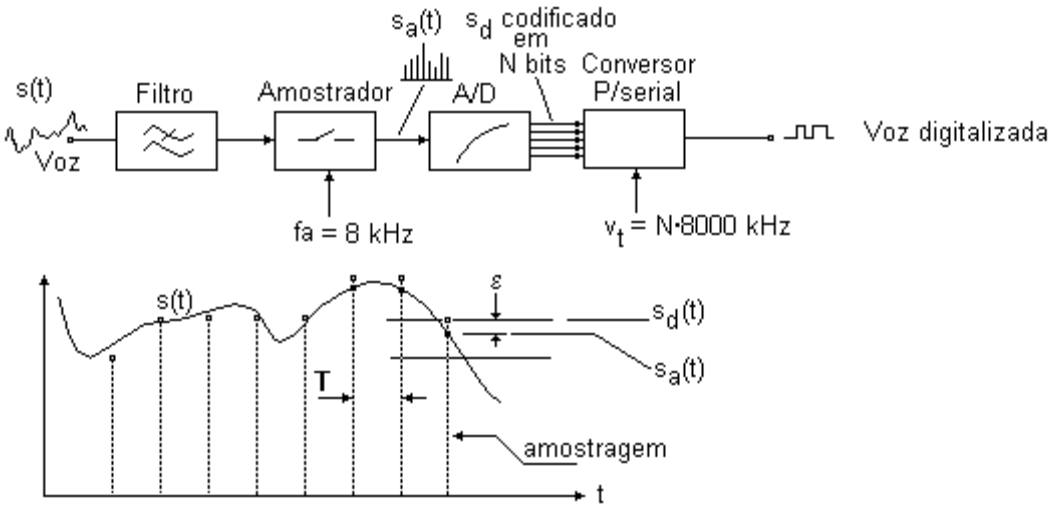

Fig. 7.4: Processo de digitalização

Veja a figura 7.4:

- O sinal de voz passa primeiro por um filtro passa-baixa, para cortar as componentes de frequência acima do permitido.
- O sinal filtrado é amostrado pelo circuito amostrador, na frequência de 8 kHz.
- O sinal amostrado $s_a(t)$ é convertido para o formato digital pelo conversor Analógico/Digital (A/D) utilizando "N" bits para cada amostra. O sinal resultante é $s_d(t)$. Este processo é chamado de "quantização". Inicialmente consideraremos que a quantização é feita de forma **linear**: o acréscimo de um bit no sinal quantizado sempre representa o mesmo acréscimo na tensão elétrica analógica.
- As amostras digitais são convertidas de paralelo para serial, a fim de serem transmitidas. A voz digitalizada, portanto, corresponde a um canal de dados com a taxa v_t = N.8000.

A relação entre a quantidade de bits e a qualidade é direta.
Vamos calcular a relação sinal-ruído gerada pelo processo de quantização. Seja:

N = quantidade de bits utilizada para codificar
$M = 2^N$ = quantidade total de níveis possíveis
$V_{máx}$ = valor de pico (máximo) que o sinal de entrada pode ter sem saturar

$v_a = 2 \cdot V_{máx}/M$ = intervalo entre dois valores quantizados

O erro de cada amostra, em relação à amplitude do sinal original, será a diferença entre o valor do sinal amostrado, $s_a(nT_a)$, e o sinal $s_d(nT_a)$, já quantizado.

O sinal $s_d(nT_a)$ somente pode assumir um dos 2^N valores discretos possíveis, múltiplos de v_a.

Considerando que a amplitude do sinal $s_a(nT_a)$ pode estar em qualquer posição entre dois níveis de $s_d(nT_a)$, com a mesma probabilidade, então podemos dizer que a função densidade de probabilidade do erro, $p(\varepsilon)$, é constante entre os valores que ele pode assumir ($-v_a/2$ e $+v_a/2$) e igual a $1/v_a$ pois a área dessa função deve ser igual a 1.

A variância do erro, ou seja, seu desvio médio quadrático, ou ainda, a potência do sinal "erro", é dada pela integral abaixo:

$$P(\varepsilon^2) = \int_{-va/2}^{+va/2} p(\varepsilon)\varepsilon^2 d\varepsilon \qquad [7.4]$$

Como $p(\varepsilon)$ é constante (e igual a $1/v_a$) entre $-v_a/2$ e $+v_a/2$, então:

$$P\left(\varepsilon^2\right) = \int_{-va/2}^{+va/2} p(\varepsilon)\varepsilon^2 d\varepsilon = \frac{1}{v_a}\left[\frac{\varepsilon^3}{3}\right]_{-va/2}^{+va/2} = \frac{v_a^2}{12} = \frac{(2 \cdot V_{máx}/M)^2}{12} = \frac{V_{máx}^2}{3 \cdot 2^{2N}} \qquad [7.5]$$

Vamos considerar que o sinal de entrada, $s(t)$, é aleatório e tenha uma potência de V^2_{RMS} (RMS = valor médio quadrático $=\sigma^2$). Registre-se que este valor deve ser menor que $V^2_{Máx}$ para que o sinal fique dentro da faixa do digitalizador, sem saturar, acomodando o fator de pico do sinal. A relação sinal-ruído, portanto, será:

$$SNR = \frac{V_{RMS}^2}{P\left(\varepsilon^2\right)} = \frac{3 \cdot 2^{2N} \cdot V_{RMS}^2}{V_{máx}^2} \qquad [7.6]$$

$$SNR[dB] = 10 \cdot \log\left\{\frac{3 \cdot 2^{2N} \cdot V_{RMS}^2}{V_{máx}^2}\right\} = 10 \cdot \log\left(3 \cdot 2^N\right) + 10 \cdot \log\left(\frac{V_{RMS}^2}{V_{máx}^2}\right) \qquad [7.7]$$

$$SNR[dB] = 10 \cdot \log(3) + 10 \cdot \log\left(2^{2N}\right) + 20 \cdot \log\left(\frac{V_{RMS}}{V_{máx}}\right)$$

$$\boxed{SNR[dB] \cong 4{,}8 + 6 \cdot N - CF[dB]} \qquad [7.8]$$

A equação 7.8 mostra que, para uma determinada amplitude relativa do sinal de entrada, a relação sinal-ruído vai melhorando linearmente conforme aumenta a quantidade "N" de bits da digitalização. A figura 7.5 ilustra esse efeito para uma relação de -12 dB entre o sinal de entrada e o nível máximo permitido.

A equação 7.8 nos mostra também que, para um determinado número de bits "N", a relação sinal-ruído varia linearmente com a amplitude do sinal de entrada em relação à amplitude máxima permitida. A figura 7.6 mostra esse efeito para o caso de N=8.

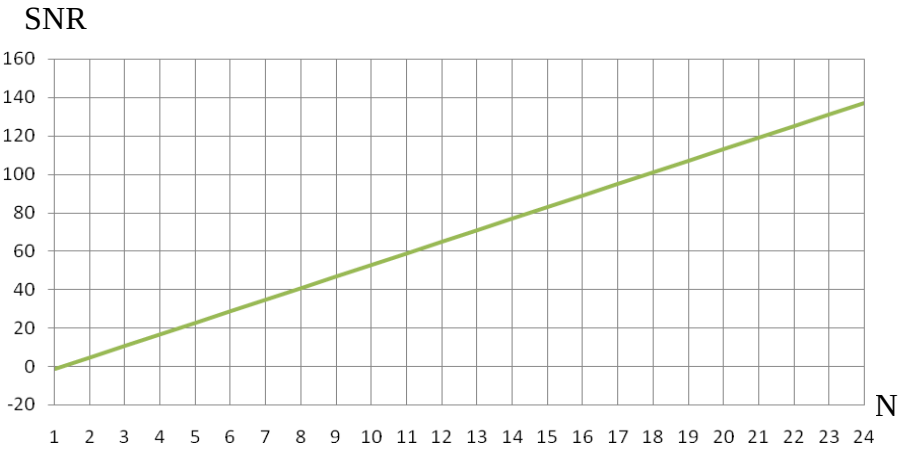

Fig. 7.5: Variação da SNR em função da quantidade de bits "N"

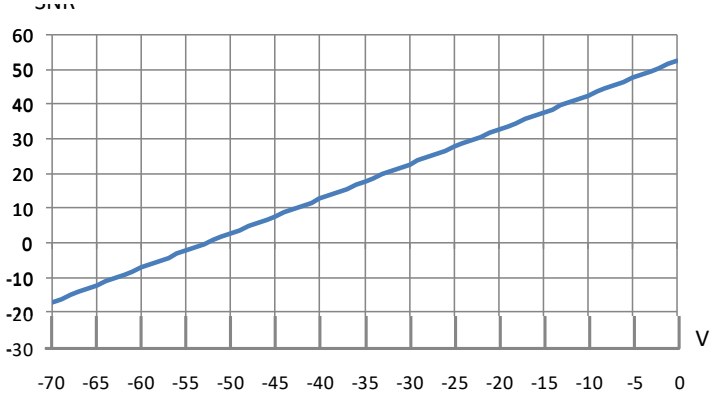

Fig. 7.6: Variação da SNR em função da amplitude do sinal de entrada

Com este cenário (sinal codificado com 8 bits), para se conseguir uma voz de "boa qualidade" (SNR = 40 dB), é preciso que:

- O sinal de entrada tenha um fator de crista de, no máximo, 12 dB
- Utilize toda a dinâmica disponível (oferecida pelo quantizador)

A transmissão desse sinal de voz digital terá uma taxa de 8·8000 = 64 kbps, sem contar os bits de overhead do protocolo que eventualmente seja utilizado.

7.4.2 Quantização logarítmica

A quantização logarítmica tem a finalidade de melhorar a relação sinal-ruído para níveis de recepção mais baixos que $V_{máx}$, como pode ser visto na figura 7.8 para o caso de 8 bits e na figura 7.9 para o caso de 12 bits.

Para melhorar essa situação, quando o sinal de entrada é fraco, utiliza-se uma quantização logarítmica que, na verdade, executa uma compressão do sinal de entrada.

A função de transferência de $s(nT_a)$ para $s_a(nT_a)$, específica da quantização chamada "lei μ" (μ law), com μ = 255, é dada pela equação 7.9.

Pela figura 7.7 podemos comparar as funções de transferência das quantizações linear (em azul) e logarítmica (em vermelho).

$$s_a(t) = \frac{\ln[1+\mu \cdot s(t)]}{\ln[1+\mu]} \qquad [7.9]$$

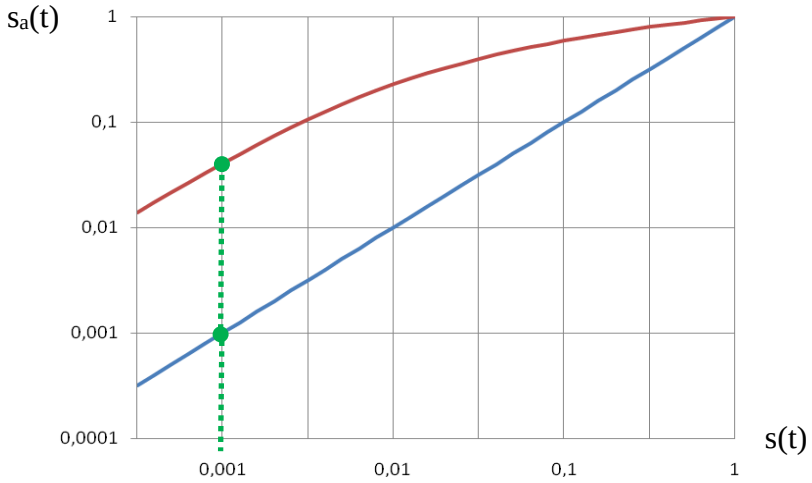

Fig. 7.7: Codificações linear (azul) e logarítmica (vermelho)

Uma amostragem de 1 mV (0,001 volt) é mapeada em 1 mV na linear e em 40 mV na logarítmica e assim por diante.

A relação sinal-ruído da compressão μ-law, considerando uma distribuição Gaussiana do sinal de entrada é dada pela equação 7.10.

$$SNR[dB] = 10 \cdot \log\left\{ \frac{2^{2N}}{10{,}25} \cdot \frac{1}{1 + \frac{1{,}6}{\mu \cdot \frac{V_{RMS}}{V_{máx}}} + \frac{1}{\left(\mu \cdot \frac{V_{RMS}}{V_{máx}}\right)^2}} \right\} \qquad [7.10]$$

A figura 7.8 ilustra a relação sinal-ruído para uma quantização linear (azul) e uma logarítmica (vermelho), com 8 bits por amostra. Para baixos níveis do sinal de entrada podemos notar que a quantização logarítmica é muito melhor que a linear: com o sinal de entrada em - 45 $dBV_{máx}$, por exemplo, a linear oferece uma SNR de 7,8 dB enquanto a logarítmica está em 34 dB. A partir desse ponto, a quantização logarítmica tem uma SNR praticamente constante até 0 $dBV_{máx}$.

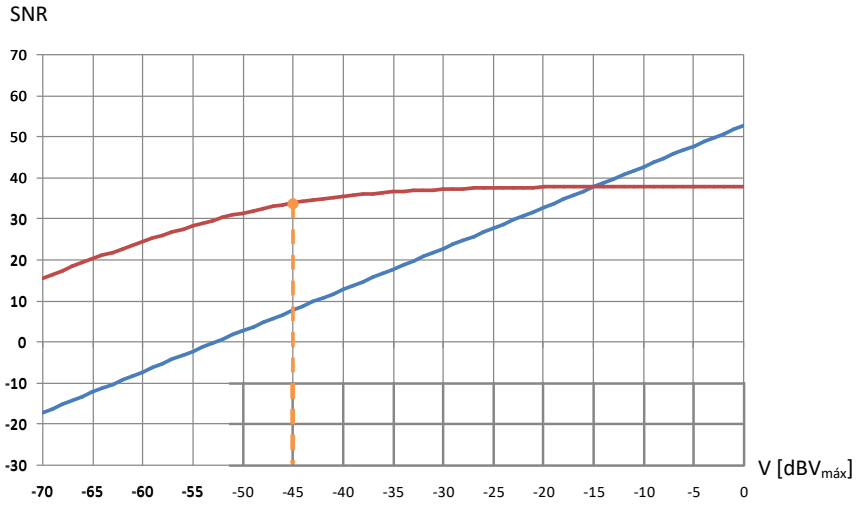

Fig. 7.8: Codificações linear (azul) e logarítmica (vermelho) em 8 bits

No caso da codificação de voz, da qual estamos tratando agora, as experiências mostraram que, para se obter uma qualidade aceitável para o sistema telefônico, precisamos ter uma relação sinal-ruído melhor que 30 dB na quantização, quando o sinal de entrada estiver acima de -45 $dBV_{máx}$.

Pela figura 7.8 vemos que, com 8 bits, somente a quantização logarítmica consegue atender a esse requisito.

Pela equação 7.8 podemos constatar que para atender a esses requisitos com a quantização linear seriam precisos 12 bits:

$$6 \cdot N = SNR[dB] - 4{,}8 - V[dBV_{máx}] = 30 - 4{,}8 + 45 \Rightarrow N = 11{,}66$$

A figura mostra 7.9 mostra o desempenho das duas quantizações para 12 bits.

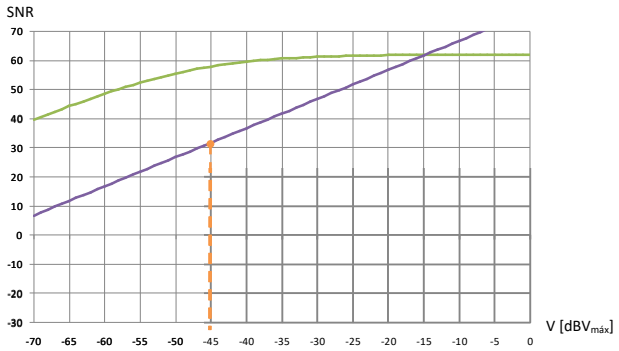

Fig. 7.9: Codificações linear (azul) e logarítmica (vermelho) em 12 bits

O quantizador logarítmico também é chamado de compansor pois a quantização corresponde a uma compressão do sinal de entrada. Na recepção o sinal deve sofrer uma expansão.

7.4.3 Codificação DPCM

A terceira fase corresponde à codificação propriamente dita.

O primeiro codificador de voz a se popularizar foi o DPCM - Differential Pulse Code Modulation, ou PCM diferencial.

Na figura 7.10 o quantizador está representado por um único bloco, que contempla o amostrador e o conversor A/D. O sinal de voz *s(t)*, na entrada, é convertido para um sinal digital q(n), com N bits por amostra.

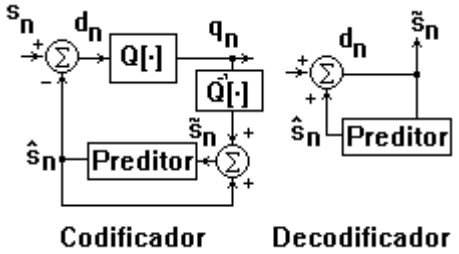

Fig. 7.10: Codificador DPCM

O sinal de voz possui uma componente periódica, envolta em um envelope que varia lentamente com o tempo, cuja forma é determinada pelo tubo do trato vocal humano, e essa propriedade permite estimar a próxima amostra com boa precisão, devido à correlação existente entre uma amostra e a próxima.

A codificação DPCM explora essa propriedade.

Veja, pela figura 7.10, que o codificador DPCM é formado por um quantizador linear e um preditor que estima o valor da próxima amostra. O preditor é um filtro digital. A saída do preditor é subtraida da entrada a fim de se obter somente a diferença, que será codificada e transmitida. O quantizador codifica somente a diferença, que, somada à estimativa anterior, entra no preditor. A decodificação DPCM é a operação inversa.

7.4.4 Codificação ADPCM

Quando o codificador diferencial utiliza uma técnica adaptativa para determinar o valor do degrau de quantização v_a, é chamado de ADPCM ("Adaptive Differential PCM"). A recomendação CCITT G.721 especifica um algoritmo para implementar esse tipo de codificador, para digitalizar o sinal de voz em 32 kbps, sendo 8000 amostras de 4 bits por segundo. Nesse caso, a saída do quantizador é um sinal de quatro bits. Ao diagrama do DPCM acrescenta-se um circuito de adaptação de v_a.

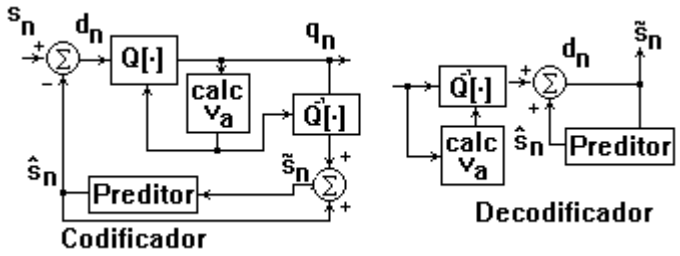

Fig. 7.11: Codificador ADPCM

7.4.5 A mudança de paradigma

O sinal de voz analógico, uma vez digitalizado, passa a ser um sinal digital serial e pode trafegar em qualquer rede digital, como os demais sinais de dados, desde que encapsulado adequadamente nos protocolos de rede Ethernet e IP.

Historicamente, o constante aumento na velocidade de transmissão dos modems e a melhoria dos codificadores de voz, fez mudar o paradigma analógico-digital, a partir do momento em que um sinal de voz digitalizada passou a ocupar menos banda (em bps) do que a velocidade máxima conseguida por um modem em um canal de voz.

A digitalização do sinal de voz começou com 64 kbps e caiu até 2400 bps.

Os modems começaram com 300 bps e chegaram a 33.600 bps.

Um modem ao transmitir 33.600 bps em um canal de voz ultrapassou a taxa da voz digitalizada, que foi reduzida a 2.400 bps. Essa inversão de situação ocorreu em meados de 1980.

A figura 7.12 mostra a mudança de paradigma.

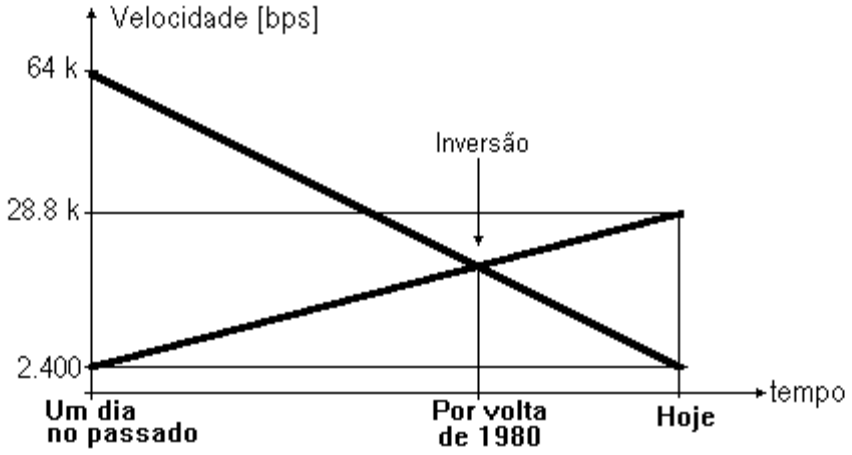

Fig. 7.12: Inversão da estratégia

A relação entre as velocidades da transmissão pela linha telefônica e da voz digitalizada cresceu 3000 vezes! O modem, padrão V.34, passou a ser capaz de transmitir 10 canais de voz com qualidade aceitável ou 5 canais de voz com "boa qualidade".

A integração voz-dados, que era feita no domínio da frequência (FDM), passou a ser possível e interessante sob os pontos de vista econômico e técnico, que fosse feita no domínio do tempo, na forma de multiplexação digital (TDM).

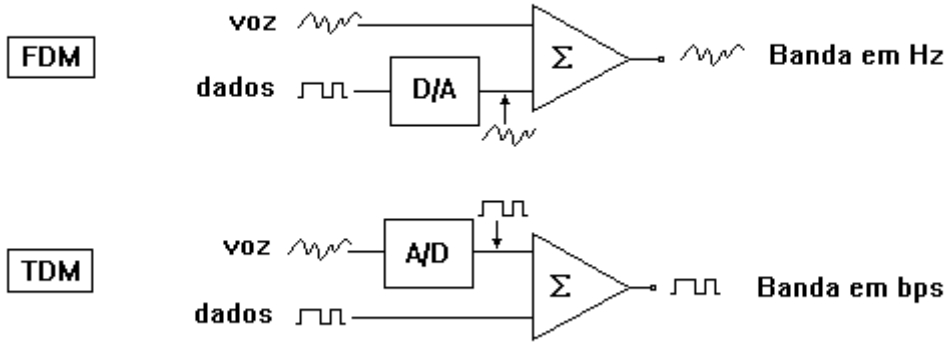

Fig. 7.13: Mudança de paradigma

7.4.6 Qualidade da voz

Há quatro métodos para se medir a qualidade de um sinal de voz, contaminado por interferência ou ruídos, que podemos considerar de maior importância:

- **Medida da relação sinal-ruído**: não traduz bem as distorções inclusive aquelas causadas pela digitalização.

- **IA, índice de articulação**: mede a percentagem de palavras e sílabas entendidas corretamente por uma platéia que anota o que ouve do sistema sob teste. Mede mais a influência do ruído e não é satisfatório quando há reverberação.

- **DTR, "Diagnostic Rhyme Test"**: mede a inteligibilidade da voz, pela capacidade de se distinguir palavras com fonemas parecidos.

- **MOS, "Mean Opinion Score"**: um método de avaliação subjetiva da qualidade do sinal de voz, baseado na opinião de uma platéia.

Como vimos, o sinal de voz digitalizado apresentará uma contaminação de ruído intrínseca do processo de digitalização que ele sofreu. Esses testes ajudam a avaliar qual codificação apresenta a melhor relação benefício-custo. Os mais relevantes são:

DRT

Este teste concentra seu objetivo em levantar a inteligibilidade do sinal de voz. Uma contagem dos fonemas interpretados corretamente "C", de um total de fonemas ditados, certos e errados, "C+E", determina o escore de inteligibilidade: DRT = [C-E]/[C+E]. O resultado é definido como:

- Excelente se DRT $>$ 0,96
- Muito bom se 0,91 $<$ DRT $\leq$ 0,96
- Bom se 0,87 $<$ DRT $\leq$ 0,91
- Moderado se 0,83 $<$ DRT $\leq$ 0,87
- Fraco se 0,79 $<$ DRT $\leq$ 0,83
- Ruim se 0,75 $<$ DRT $\leq$ 0,79

MOS

Este método de avaliação se mostrou bastante eficaz, principalmente para voz digitalizada em baixa velocidade, e é bastante utilizado, apesar de ser subjetivo. Um sinal de voz é digitalizado utilizando o equipamento a ser avaliado. Esse sinal é novamente transformado em sua forma analógica original e reproduzido para a platéia de ouvintes. Cada pessoa deve indicar sua opinião, baseada em uma escala com 5 notas:

5 - Excelente qualidade. Não há como notar qualquer distorção.

4 - Alta qualidade. Dificilmente se nota alguma distorção.

3 - Média qualidade. Há distorções perceptíveis no sinal, mas são bem aceitáveis.

2 - Baixa qualidade. Sinal distorcido, mas aceitável para comunicação de serviço.

1 - Má qualidade. Sinal de voz muito degradado, prejudicando a inteligibilidade.

O resultado do teste é a média das notas dadas pelas pessoas da platéia. O escore 5 representa a qualidade perfeita e é raramente atingido. Entre 4 e 5 a qualidade é considerada excelente e o escore 4 corresponde a um sinal de alta qualidade. Um escore MOS igual ou maior a 4 indica que a voz reproduzida tem a inteligibilidade do sinal original, praticamente isento de distorções, e, nesse caso, diz-se que temos uma "toll quality".

Para se ter uma idéia, a voz natural possui um escore MOS de 4,5 e o PCM μLaw a 64 kbps possui um escore de 4,3. Um sinal de voz, codificado a 16 kbps, pelo algoritmo ATC, possui um escore equivalente ao PCM, ou seja, alta qualidade.

Escores entre 3 e 4 são considerados de qualidade aceitável para a comunicação, significando que a inteligibilidade ainda é muito boa e se há distorções, não são totalmente perceptíveis.

7.4.7 Codificação do sinal de voz em baixa velocidade

A digitalização da voz em velocidades abaixo dos 32 kbps do ADPCM, visto anteriormente, é chamada de voz em baixa velocidade.

Os três parâmetros básicos na digitalização do sinal de voz são:

- Qualidade da voz
- Velocidade de digitalização
- Retardo do circuito

Veja abaixo uma relação de alguns algoritmos utilizados na codificação da voz em equipamentos comercialmente disponíveis no mercado, com as respectivas taxas para as quais foram desenvolvidos.

Algoritmo	Descrição	taxa [bps]
PCM μLaw	Pulse-Code Modulation, CCITT G.711	64.000
ADPCM	Adaptive Differential PCM, CCITT G.721	32.000
LD-CELP	Low Delay Code-Excited Linear Prediction	16.000
TDHS	Time Domain Harmonic Scaling	9.600
ATC	Adaptive sub-band Transform Coding	8.000
V-SELP	Vector Sum Excitation Linear Prediction	8.000
CS-CELP	Conjugate Structure Code-Excited Linear Prediction	8.000
CELP	Code-Excited Linear Prediction	4.800
ACELP	Algebraic Code-Excited Linear Prediction	4.800
IMBE	Multiband Excitation	2.400
LPC10E	Linear Prediction Coding	2.400

A figura 7.14 mostra a variação do escore MOS de alguns algoritmos em função da taxa de digitalização.

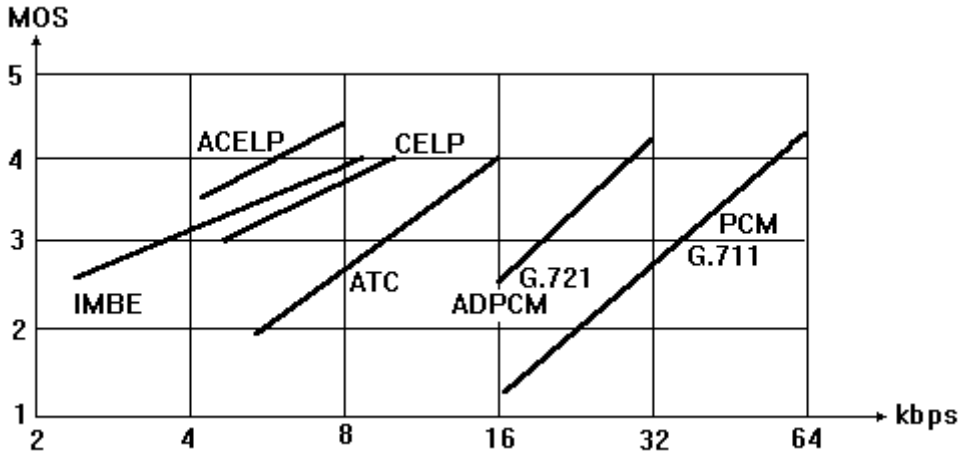

Fig. 7.14: Taxa e qualidade MOS dos algoritmos

7.4.8 Taxa de digitalização

A taxa de digitalização do sinal de voz é um parâmetro, antes de mais nada, que reflete o aproveitamento da banda disponível. Quanto menor, melhor. Porém há que se manter uma qualidade mínima a fim de atender aos requisitos da aplicação. Nas redes corporativas se usa voz ATC ou CELP a 8000 bps.

A mais nova opção de voz, ACELP a 4800 bps, com alta qualidade, foi apresentada ao mercado no inicio de 1995. Para aplicações especiais é possível utilizar voz IMBE a 2400 bps.

7.4.9 Retardo do circuito

Com relação à capacidade de processamento disponível no codificador, quanto maior o retardo do circuito melhor a qualidade da voz pois a estimativa dos parâmetros do sinal ficam mais exatas.

Um retardo maior que 150 ms, porém, já prejudica a qualidade geral do sinal digitalizado. Em equipamentos disponíveis comercialmente, o algoritmo ATC provoca um retardo em torno de 120 ms, enquanto o IMBE provoca um de 90 ms.

O algoritmo ACELP possui um retardo da ordem de 30 ms.

7.4.10 Padrões ITU-T

Os padrões atualmente utilizados para codificação de voz seguem as normas do ITU-T[1] e as principais estão relacionadas na tabela a seguir.

Norma ITU-T	Título	Data
G.711	Pulse Code Modulation (PCM) of Voice Frequencies	1972
G.722	7 kHz Audio Coding Within 64 kbps	1988
G.723	Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbps	1996
G.728	Coding of Speech at 16 kbps Using Low-Delay Code Excited Linear Prediction	1995
G.729	Coding of Speech at 8 kbps Using Conjugate Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP)	1996

7.4.11 Padrão ITU-T G.729

A primeira edição dessa norma foi em 1996, como revisão 1.

A última revisão, revisão 3, foi publicada em junho de 2012.

O codificador inicialmente amostra linearmente o sinal em 8 kHz com 16 bits por amostra, o que dá uma taxa de 8000 x 16 = 128 kbps.

[1] International Telecomunicatio Union, Genebra

Em seguida recolhe, a cada 10 ms, os 1280 bits de informação do sinal de voz digitalizada, que correspondem a 80 amostragens do sinal, ou seja, 80 x 16 = 1280 bits.

O codificador analisa a informação colhida e extrai os parâmetros necessários conforme definido no algoritmo. Esses parâmetros são codificados em 80 bits e transmitidos a cada 10 ms.

A taxa de transmissão do codificador, portanto, será:

$$v = \frac{80\ bits}{10\ ms} = \frac{80}{0{,}01} = 8000\ bps \qquad [7.11]$$

7.4.12 Padrões utilizados em telefonia celular

A primeira geração de telegonia celular (1G) utilizava voz no formato analógico, mas a partir da segunda geração (2G, 3G e 4G) todas passaram a utilizar codificação digital para o sinal de voz.

Está fora do escopo deste curso entrar em detalhes sobre as codificações utilizadas no sistema celular mas, a título de exemplo, a especificação ETSI 06.10 para sistemas GSM define um codificador com amostragem linear em 8 kHz e 13 bits por amostra, que utiliza o algoritmo chamado de RPE (*Regular Pulse Excited*) que separa o sinal de voz em pacotes de 20 ms para análise e codificação.

É importante que o codificador para telefone celular seja simples, para reduzir custos e espaço no equipamento, como é o caso do codificador definido pela ETSI, que comprometeu um pouco a qualidade.

7.5 A codificação padrão CD

O CD-DA, ou Compact Disk para armazenamento de Áudio, foi lançado em 1982, pelas empresas N.V.Philips e Sony Corporation, sendo capaz de armazenar 74 minutos de dois canais de áudio (estéreo) codificados em PCM (Pulse Code Modulation) com taxa de amostragem de 44,1 kHz e 16 bits por amostra.

A taxa de bits de áudio é dada por:

$$Taxa\ de\ bits = (2\ canais)\cdot(44100\ Hz)(16\ bits) \cong 1{,}41\ Mbps$$

A gravação na bolacha plástica de 12 cm utiliza uma codificação em que cada byte (8 bits) de áudio se transforma em 17 bits e os dados são agrupados em pacotes de 24 bytes de dados, ou 24 x 17 = 408 bits de áudio, juntamente com mais 180 bits de controle, ou seja, cada pacote gravado na bolacha possui 588 bits.

A máxima relação sinal ruído da codificação do CD (CF = 4,8 dB) é dada por:

$$SNR_{CD} = 20\cdot\log(2^{16}) \cong 96\ dB$$

Como parâmetro comparativo, o disco de vinyl possui SNR entre 50 e 60 dB.

A taxa de erro do CD gira em torno de 10 ppm (10 partes por milhão).

7.6 A codificação padrão DVD - Digital Versatile Disk

Este padrão, lançado em 1994, veio para superar as limitações do CD em capacidade de armazenamento e velocidade de transferência de dados.

O padrão prevê a utilização da mídia para dados, áudio e vídeo, definindo as categorias de produto "*DVD-Vídeo*" e "*DVD-Áudio*", entre outras.

Capacidade [Gbyte]	Tempo aproximado de programa [minutos]
4,7	133
8,5	241
9,4	266
17,0	482

O áudio da categoria "*DVD-vídeo*" possui as seguintes alternativas de codificação:

- LPCM fs = 48 ou 96 kHz 16, 20 ou 24 bits
  - até 8 canais máx = 8 canais
  - 6,144 Mbps
- Dolby AC3 fs = 48 kHz. até 6 canais (suround 5.1), máx = 448 kbps
- MPEG-1 AAC fs = 48 kHz até 8 canais (suround 7.1), máx = 384 kbps
- MPEG-2 AAC fs = 48 kHz até 8 canais (suround 7.1), máx = 912 kbps

A categoria "*DVD-Áudio*" expande as especificações do "*DVD-vídeo*" para aplicações específicas de áudio, permitindo, por exemplo, taxa se amostragem de 192 kHz.

7.7 WAVE

Não é uma codificação propriamente dita. É um formato de arquivo que possui áudio codificado (normalmente em PCM).

É um dos tipos de arquivo definidos na especificação "*Multimedia Programming Interface and Data Specifications*", publicada pela Microsoft.

A especificação da Microsoft define um formato genérico de arquivo de multimidia, denominado "*Resource Interchange File Format*", o qual pode assumir diversas formas. Uma delas é o Wave (terminação .wav).

O arquivo possui um cabeçalho que carrega diversas informações sobre o conteúdo, tais como: taxa de amostragem, quantidade de canais (um ou dois), a taxa de transferência média com que os dados devem ser transferidos (bps), o formato (PCM ou outro formato registrado), quantidade de bits por amostra e outros.

7.8 BWF

É também um formato de arquivo. Na verdade uma extensão do Wav, inclusive o arquivo possui a mesma terminação (.wav). Consta de norma publicada pela EBU (*European Broadcasting Union*) em 1997.

A idéia central deste formato foi acrescentar metadados, ou seja, informações adicionais de texto.

Pelo nome do arquivo não dá saber se é Wave ou Bwf. É preciso olhar as propriedades do arquivo.

7.9 A codificação MP3

MP3 é o nome popular para a codificação MPEG-1-Layer3, para áudio e vídeo, mas com esta sigla estamos nos referindo apenas ao áudio.

É uma codificação com perda de conteúdo, desenvolvida pelo Motion Picture Experts Group (MPEG), que explora as características cognitivas do ouvido humano com o objetivo de gerar um conteúdo cuja perda seja imperceptível ao ouvido humano. Dessa forma, o algoritmo procura remover apenas informações supostas irrelevantes com relação à percepção auditiva.

Foi adotada como padrão ISO em 1992.

A taxa de amostragem pode ser 32, 44,1 ou 48 kHz.

Suporta sequências com um ou dois canais de áudio, que podem ter taxas de transmissão de 32 a 320 kbps.

A codificação MP3 consegue fator de compressão variando de 1:2,7 até 1:24 sobre o arquivo de dados gerados pela amostragem.

Por exemplo, uma amostragem de 44,1 kHz a 16 bits (CD) gera uma taxa de 44100 x 16 = 705,6 kbps por canal. Com uma compressão de 1:24 chega-se a uma taxa de aproximada de 29,4 kbps.

É preciso observar a taxa de transmissão da codificação MP3. A codificação defaut do iTunes, por exemplo, é 160 kBps. Acima dessa taxa tem 192, 224, 256 e 320 kbps. Entretanto, poucas pessoas conseguem identificar perda de conteúdo em uma codificação a 192 kbps comparada a uma a 320 kbps.

O arquivo de áudio MP3 pode conter, além do áudio propriamente dito, informações adicionais, como taxa de amostragem, nome da música, etc.

7.9.1 Codificador MP3

Inicialmente o sinal digitalizado, no formato PCM, entra no codificador e passa pelo estágio de Análise, que consiste em um banco de 32 filtros e um circuito de transformada discreta de cosseno modificada (MDCT).Ao mesmo tempo o sinal passa por um circuito de transformada de Fourier e vai a um bloco de modelagem psico-acústica. Em seguida é feita a codificação e serialização da sequência.

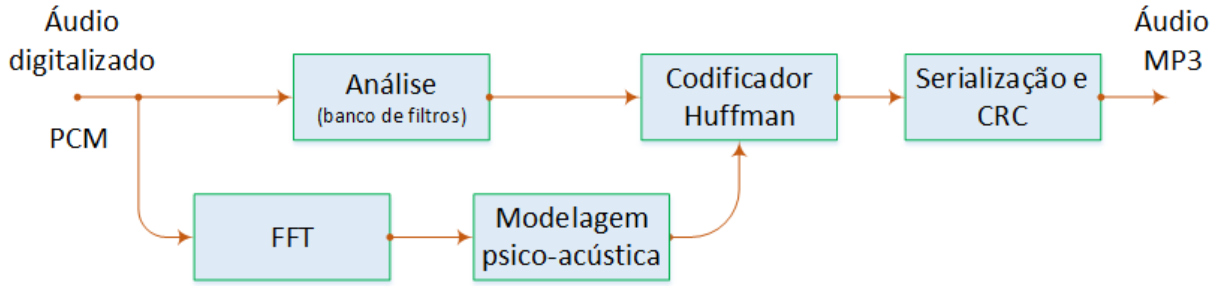

Fig. 7.15: Codificador MP3

7.10 A codificação FLAC

FLAC - *Free Lossless Audio Codec* - é um padrão de codificação desenvolvido por Josh Coalson, que utiliza um algoritmo de compressão que reduz o arquivo para cerca de 50% de seu tamanho original, sem perdas, gerando pacotes de dados auto suficientes (que permitem a sua decodificação sem depender de informação de outros pacotes).

É um padrão aberto bem documentado, não proprietário, não está vinculado a patentes e não exige pagamento de royalties.

Utiliza um preditor linear, em um esquema semelhante ao que foi abordado na codificação da voz, e é assimétrico a favor do decodificador, ou seja, permite uma decodificação com baixo retardo, o que o torna interessante para transmissões de áudio.

7.11 Comparativo entre as codificações sem perda

	ALAC	FLAC	WMAL
Velocidade de codificação	😐	☺	😐
Velocidade de decodificação	😐	☺	☹
Compressão	☺	☺	☺
Detecção de erro	☹	☺	☺
Streaming	☺	☺	☺
Suporte em hardware	😐	☺	☹
Suporte em software	😐	☺	😐
Código aberto	☺	☺	☹
Tag	iTunes	Vorbis	ASF

7.12 Qual é a qualidade é aceitável?

Depende.

Não é só uma pura questão de qualidade para consumo nos dias atuais. Também deve-se pensar no futuro: o padrão de digitalização será suportado pelos diversos softwares aplicativos daqui a 20 anos?

Para a captura e preservação de áudio com conteúdo importante é recomendável que se use a taxa de amostragem de 96 kHz com 24 bits.

O espectro principal do sinal original é um parâmetro importante para definir a profundidade da captura e preservação. O padrão de CD (44,1 kHz com 16 bits) é aceitável para gravações de voz se não houver requisito adicional. Pelo menos o padrão de DVD (48 kHz com 16 bits) deve ser utilizado para áudios cujo conteúdo se estende mais em frequência, como ruídos da natureza, de animais, ou máquinas.

Os formatos de arquivo ".aiff" e ".wav" devem ser escolhidos, preferencialmente, por três motivos:

- São padrões cuja codificação não executa compressão
- Estão bem documentados
- São públicos.

O formato ".alac", codificação ALAC (*Apple Lossless Audio Codec*), sem compressão, proprietária da Apple Computer, é uma opção aceitável.

A formato ".mp3", codificação MP3, possui perda mas sua documentação é pública e o padrão é amplamente utilizado, podendo ser uma opção na categoria "com perda".

Os seguintes formatos devem ser evitados: ".aac", padrão AAC (*Advanced Audio Coding*), ou MPEG-2, que é proprietário e possui perda; ".ra", real audio e ".wma", (*Windows Media Audio*).

Além da questão da preservação do patrimônio gravado, a qualidade aceitável depende do nível de exigência de quem vai ouvir, do ambiente em que o som será reproduzido e do conteúdo sonoro.

Nem sempre se consegue distinguir entre uma gravação em alta resolução (24 bits) e sua versão MP3. O sistema auditivo e a experiência de quem está ouvindo influencia consideravelmente.

A figura 7.16 ilustra a evolução da gravação de música para uso popular.

Fig. 7.16: a) Walkman da Sony
Analógico

b) iPod Nano da Apple
44.1 kHz, 16 bits
MP3

c) Pono da Pono Music
192 kHz, 24 bits
FLAC

7.13 Tempo de gravação versus espaço de memória

O espaço de memória ocupado para armazenar uma gravação depende da codificação utilizada.

A figura 7.17 foi extraída do manual do gravador Tascam modelo DR-44WL, a fim de exemplificar a relação entre codificação e espaço de memória.

A tabela considera que as gravações são feitas no modo estéreo (dois canais). Se a gravação for feita em mono (um canal) os tempos de gravação dobram.

File format (recording setting)			SD/SDHC/SDXC cards capacity			
			1GB	4GB	8GB	32GB
WAV/BWF 16 bit (STEREO)		44.1kHz	1 hour 41 minutes	6 hour 44 minutes	13 hour 28 minutes	53 hour 52 minutes
		48kHz	1 hour 33 minutes	6 hour 12 minutes	12 hour 24 minutes	49 hour 36 minutes
		96kHz	46 minutes	3 hour 06 minutes	6 hour 12 minutes	24 hour 48 minutes
WAV/BWF 24 bit (STEREO)		44.1kHz	1 hour 07 minutes	4 hour 30 minutes	9 hour 00 minutes	35 hour 44 minutes
		48kHz	1 hour 02 minutes	4 hour 08 minutes	8 hour 16 minutes	33 hour 04 minutes
		96kHz	31 minutes	2 hour 04 minutes	4 hour 08 minutes	16 hour 32 minutes
MP3 (STEREO/MONO)	32 kbps	44.1kHz/48kHz	74 hour 32 minutes	298 hour 08 minutes	596 hour 16 minutes	-
	64 kbps	44.1kHz/48kHz	37 hour 16 minutes	149 hour 04 minutes	298 hour 08 minutes	
	96 kbps	44.1kHz/48kHz	24 hour 50 minutes	99 hour 20 minutes	198 hour 40 minutes	
	128 kbps	44.1kHz/48kHz	18 hour 38 minutes	74 hour 32 minutes	149 hour 04 minutes	
	192 kbps	44.1kHz/48kHz	12 hour 25 minutes	49 hour 40 minutes	99 hour 20 minutes	
	256 kbps	44.1kHz/48kHz	9 hour 19 minutes	37 hour 16 minutes	74 hour 32 minutes	
	320 kbps	44.1kHz/48kHz	7 hour 27 minutes	29 hour 48 minutes	59 hour 36 minutes	

• The recording times shown above are estimates, They might differ depending on the SD, SDHC, and SDXC cards in use.

Fig. 7.17: Tempo de gravação x memória

7.14 Testes de inteligibilidade em ambiente interno

7.14.1 AL-CONS

"Articulation Loss of Consonants". Este teste consegue avaliar a relação entre o som direto e o reverberante, além da relação sinal-ruído.

Realizado com um analisador, quanto menor o valor melhor a inteligibilidade. Normalmente a nota máxima aceitável é 10%.

7.14.2 RASTI

RASTI, "Rapid Speed Transmission Index", mede a perda de modulação de um sinal gerado no ''ambiente" devido à presença de ruído e reverberação. Seu valor vai de 0 a 1.

Quanto maior, melhor. Acima de 0,75 é considerado excelente. De 0,6 a 0,75 é bom. De 0,45 a 0,60 é médio. De 0,30 a 0,45 é fraco e abaixo de 0,30 é ruím.

7.14.3 C50 e C80

Os testes C50 e C80, "Clarity 50 e Clarity 80 ", são utilizados para avaliar a inteligibilidade da voz e clareza da música, respectivamente.

Consiste em determinar a relação entre a potência do som direto nos primeiros 50 ms (ou 80 ms) e a potência do som reverberante juntamente com o ruído. O mínimo considera aceitável é 0 dB (potências iguais). O valor considerado bom é +4 dB ou superior.

Para um resultado mais confiável é recomendável posicionar o emissor em pelo menos duas posições dentro do ambiente e instalar microfones em pelo menos 10 posições onde estarão ouvintes.

7.15 Codificação para transmissão

O sinal de áudio, após ser codificado, segundo um dos métodos descritos anteriormente, passa a ser a informação principal a ser armazenada ou transmitida.

Caso seja necessário transmitir o sinal digital em um enlace de par trançado de cobre (canal de comunicação), que é o meio de comunicação mais comum em uma rede interna, é preciso fazer uma codificação adicional chamada de "codificação de linha", a fim de reduzir a banda em frequência do sinal transmitido e, portanto, reduzir a distorção e contaminação de ruído, levando a uma menor taxa de erro.

Neste curso vamos ver apenas duas codificações de linha, tendo em vista que são as mais utilizadas na Rhox atualmente: a primeira é a codificação de linha do protocolo Ethernet 100 Mbps sobre cobre. A outra é a codificação utilizada pela Rane em seus dispositivos remotos de áudio (RAD):

- Ethernet 100baseTX $\Rightarrow$ 4B5B com MLT-3
- Rane RAD $\Rightarrow$ Bifase-marca (AES-3)

7.15.1 Ethernet 100baseTX

A codificação utilizada pelo protocolo Ethernet 100baseTX possui duas fases:

i. primeiro o sinal é segmentado em grupos de 4 bits (16 combinações possíveis) que são mapeados em símbolos de 5 bits (32 combinações possíveis). Em seguida esses bits são serializados novamente. Então, uma sequência de dados a 100 Mbps se transforma em uma sequência de 125 Mbps. Obtém-

se aí uma redundância que pode ser utilizada para evitar padrões indesejáveis e transmitir controles.

ii. em seguida o sinal serial de 125 Mbps é codificado em MLT-3, que possui três níveis de linha e muda de nível sempre que há uma transição.

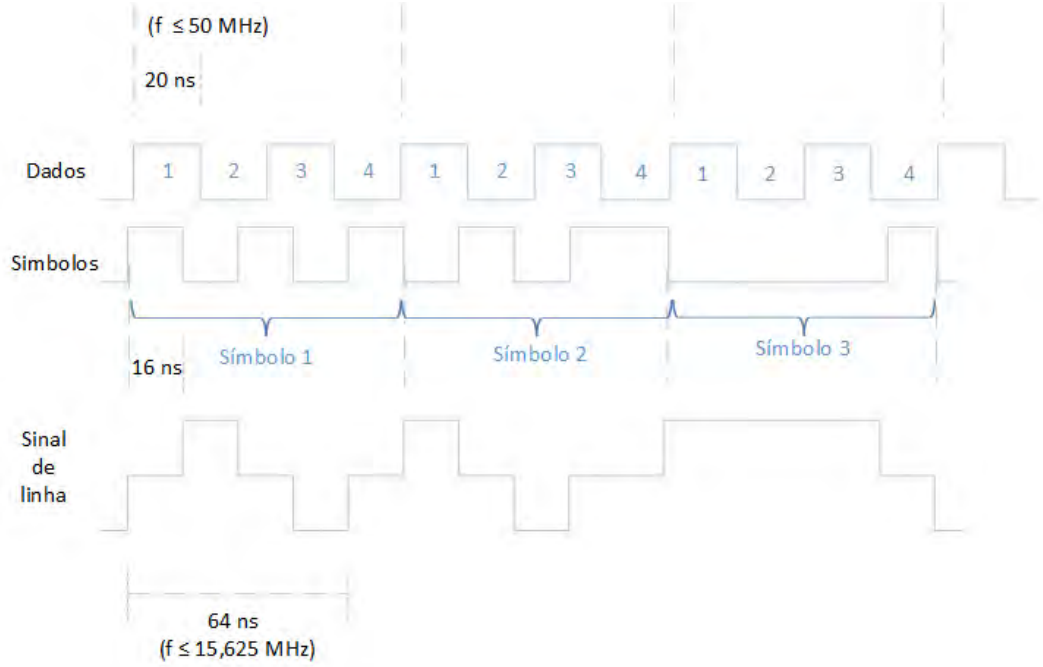

Fig. 7.18: Ethernet 100baseTX - Codificação de linha

A figura 7.18 mostra que a cada 4 bits de dados, de uma sequência a 100 Mbps, é gerado um símbolo de 5 bits. Supondo que os bits de dados fiquem sempre se alternando entre "0" e "1", a frequência fundamental máxima dessa sequência de dados seria de 50 MHz.

Entretanto, uma codificação MLT-3 é aplicada a fim de reduzir esse alcance.

Cada grupo de quatro bits de dados (um símbolo) é mapeado em 5 bits, gerando uma redundância.

Esses 5 bits de dados são codificados em MLT-3 que é um sinal de linha com três níveis que segue a seguinte regra: quando há uma transição da sequência de bits o sinal muda de nível subindo até o máximo e depois descendo até o mínimo, circularmente. Desta forma, se os bits de símbolos se alternarem sempre, o sinal de linha alcançará a frequência fundamental máxima de 15,625 MHz (período de 4 x 16 ns = 64 ns).

Essa codificação gera um espectro de frequência que alcança duas vezes a fundamental, ou seja, de 31,25 MHz.

O mérito da codificação de linha foi reduzir o espectro do sinal transmitido de 100 MHz (se transmitisse diretamente os bits na linha) para 31,25 MHz.

Tendo em vista que os protocolos de codificação de áudio, de camada 2, como o Cobranet, são encapsulados no pacote Ethernet, é justificável a sua abordagem neste curso.

7.15.2 Rane RAD

A codificação utilizada pela Rane para transmitir os dados entre o roteador Mongoose ou o processador HAL e os dispositivos remotos (RAD = *Remote Audio Device*) é a mesma da norma AES-3, ou seja, é a codificação bifase-marca, que funciona da seguinte maneira:

- Cada bit a ser transmitido corresponde a um símbolo de dois estados.
- O primeiro estado de um símbolo é sempre diferente do segundo do símbolo anterior.
- O segundo estado será igual ao primeiro se o bit for "0" e será diferente se o bit for "1".

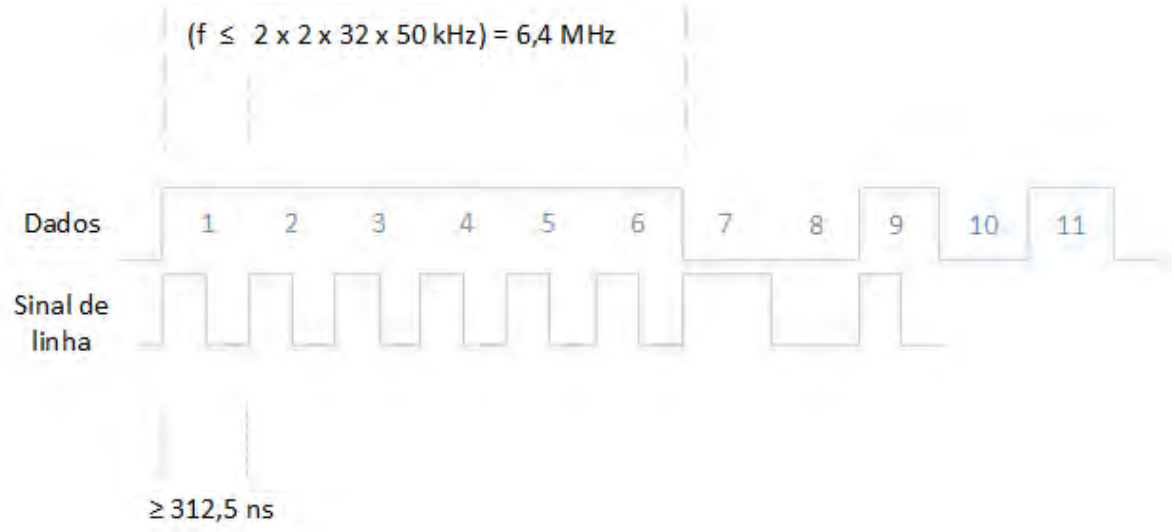

Fig. 7.19: Rane - Codificação de linha

A norma AES-3 estabelece que devem ser montados pacotes (ou quadros) com dois sub-pacotes de 32 bits. A taxa máxima de pacotes é de 50 kHz.

A norma também estabelece que a os pacotes consigam transmitir amostras com até 24 bits.

A figura 7.19 mostra que o pior caso para essa codificação é quando todos os bits são iguais a "1", gerando uma frequência de linha máxima 6,4 MHz.

A parte principal do espectro ocupado pela codificação bifase especificada na norma AES-3 e utilizada pela Rane chega a aproximadamente 5,5 MHz.

A norma prevê a transmissão balanceada em lances de par trançado de cobre, blindado, 100 Ω, com até 100 metros, uma taxa de transmissão de pacotes de 50 kHz e taxas de amostragem de 32kHz, 44,1 kHz, 48 kHz e 96 kHz.

7.15.3 Codificação para transmissão em banda-base

As transmissões que não fazem uso de modulação do sinal de dados por uma portadora, ou seja, o espectro original dos dados não é deslocado por multiplicação por uma portadora, e utilizam exclusivamente recursos de codificação dos bits, são chamadas de transmissão em banda-base, como é o caso dos exemplos citados.

A codificação visa alcançar os seguintes objetivos:

1. Reduzir ao máximo a componente DC
2. Reduzir ao máximo o espectro do sinal transmitido
3. O sinal codificado deve conter boa informação de sincronismo, a fim de facilitar sua recuperação na recepção
4. O sinal codificado deve ter boa imunidade a ruído

A figura 7.20 ilustra algumas codificações de linha: Bipolar, HDB-3, Bifase, Bifase Diferencial, Miller e 2B1Q.

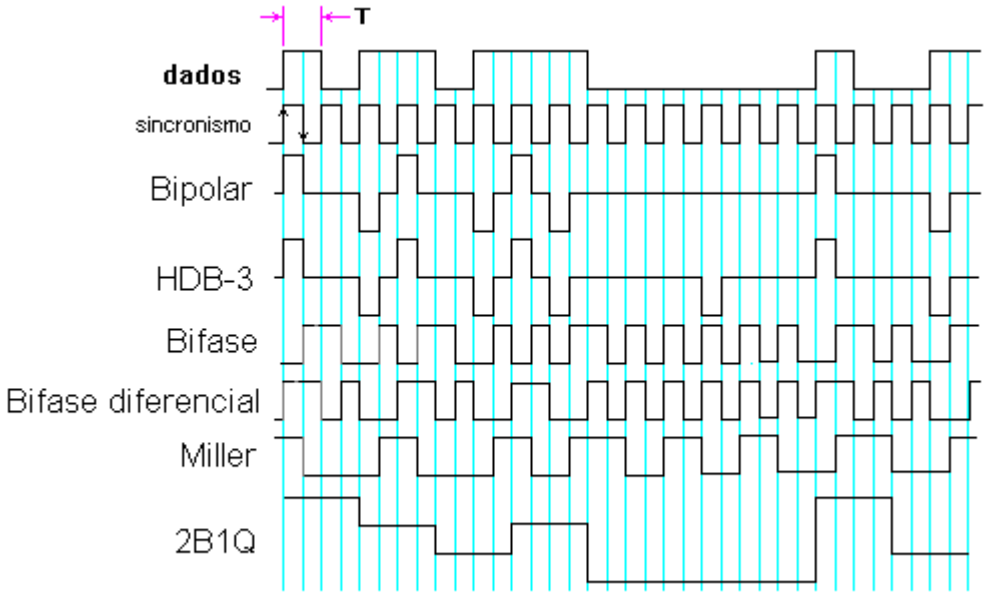

Fig. 7.20: Codificações em banda-base

A figura 7.21 ilustra os espectros resultantes das transmissões utilizando essas codificações de linha, considerando que a sequência de dados é aleatória.

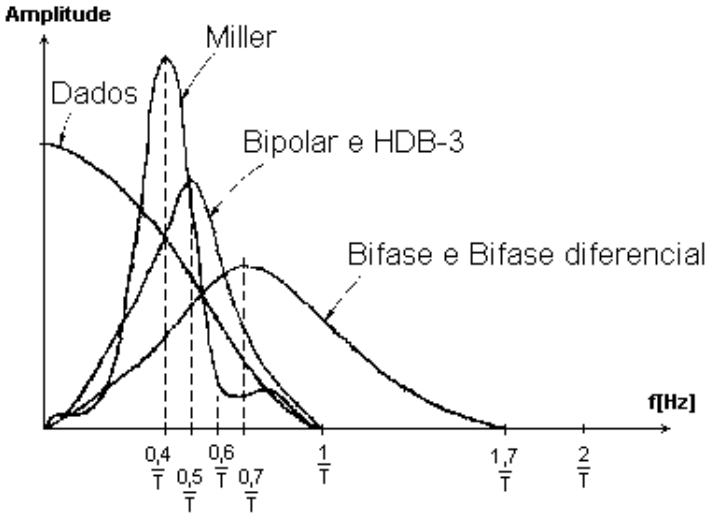

Fig. 7.21: Espectros de algumas codificações de linha

7.16 Encapsulamento e transmissão em rede

O modelo OSI de sete camadas define uma estrutura para organizar a comunicação de dados de forma hierárquica.

Na verdade, o número da camada é de pouca importância. O que interessa é saber quem vai encapsulado em quem.

Entretanto, as denominações e aplicações típicas das camadas são:

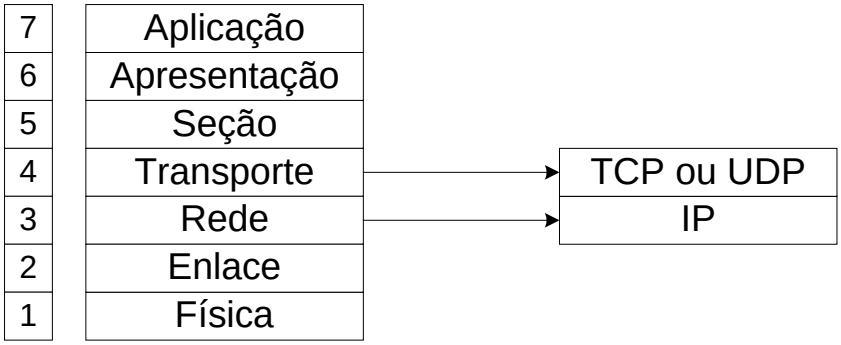

Fig. 7.22: Modelo OSI ("*Open Systems Interconnection*")[2]

O pacote de uma cada camada, contendo um cabeçalho com controles e o conteúdo (dados) é encapsulado no pacote da camada inferior, até chegar à camada 1, responsável pela transmissão. Então, o pacote da camada 1 carrega todos os demais.

A idéia do modelo OSI, ou o conceito das 7 camadas, se tornou uma referência de estrutura para os protocolos de comunicação, sendo amplamente utilizada, principalmente no que se refere ao nome e número das camadas. No entanto, a estrutura de uma comunicação não necessariamente utiliza as sete camadas. O TCP-IP, por exemplo, pode ser

[2] Em 1983 a ISO publicou o trabalho intitulado "The Basic Reference Model for Open Systems Interconnetion"

visto como estruturado em 4 ou 5 camadas, onde o aplicativo (camada 7), como o HTTP, SMTP, TELNET, DHCP, etc, fica diretamente sobre a camada de transporte (camada 4), UDP ou TCP.

A figura 7.23 mostra a relação entre as camadas do modelo OSI e a estrutura do TCP-IP, em dois exemplos: uma rede local utilizando Ethernet e uma conexão de longa distância utilizando o protocolo Frame Relay.

O protocolo Ethernet abrange as camadas 1 e 2, pois define aspectos das interfaces física e elétrica, bem como de transmissão e conexão de enlace.

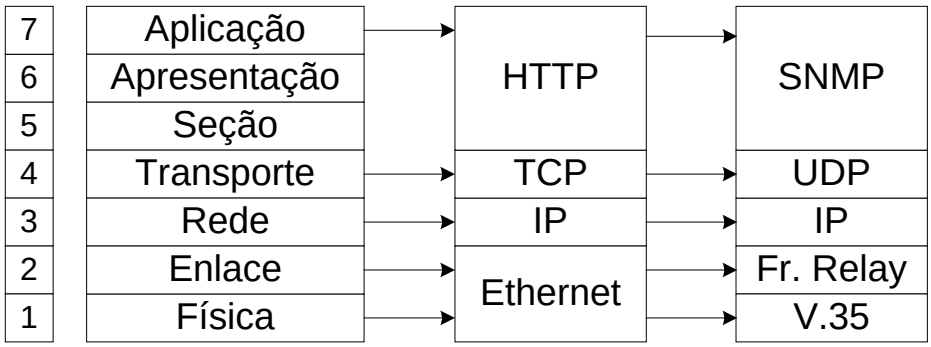

Fig. 7.23: TCP-IP e o modelo OSI

Como o pacote do protocolo da camada superior é encapsulado no campo de dados do pacote da camada imediatamente inferior, todas as informações do protocolo da camada superior, incluindo seus dados e informações de controle, são transportadas de forma transparente pelo protocolo da camada inferior.

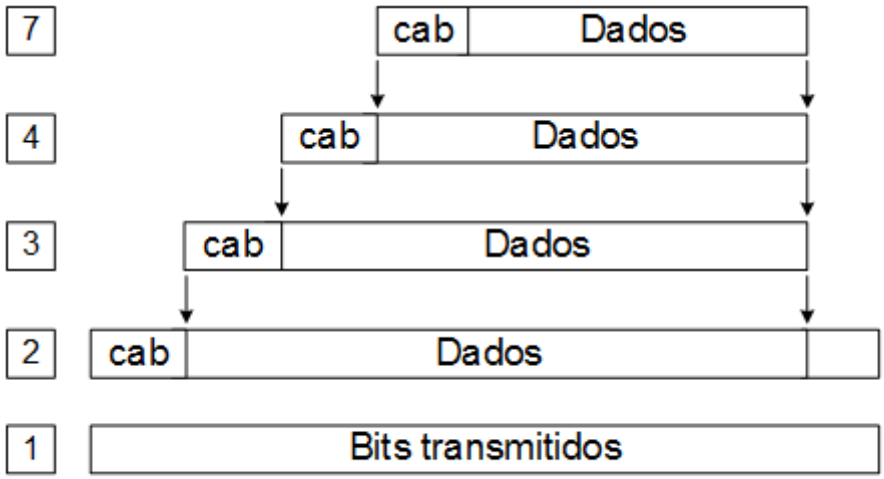

Fig. 7.24: Encapsulamento de pacotes

A figura 7.25 mostra a estrutura dos protocolos TCP-IP e como eles são encapsulados nas camadas 2 (Enlace) e 1 (Física).

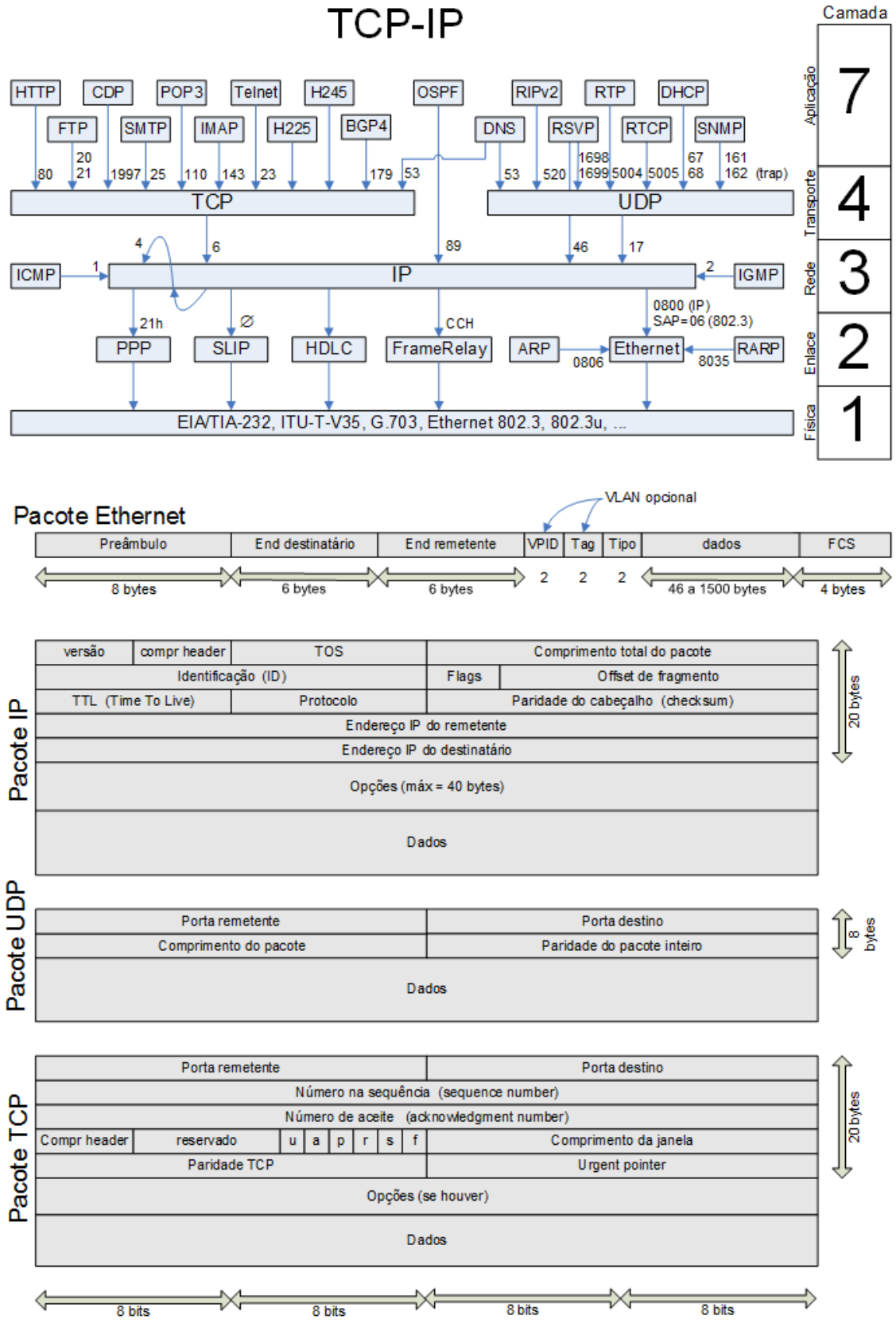

Fig. 7.25: Estrutura dos protocolos TCP-IP

Neste curso, nosso foco é nos protocolos relacionados com o transporte de áudio. A tabela mostra os principais protocolos relacionados com o transporte de áudio, ressaltando que eles podem ser de camada 2 ou 3.

Protocolo	OSI	Resolução	Amostragem suportada	Latência	Qtde de canais	Quando surgiu	Proprietário
Cobranet	2	≤ 24 bits	≤ 96 kHz	≤ 5,33 ms	≤ 32 in ≤ 32 out	1997	Cirrus Logic
Ether Sound	2	≤ 24 bits	≤ 192 kHz	Typ 1,5 ms	≤ 128	2001	Digigram
Dante	3	≤ 32 bits	≤ 192 kHz	Typ 1 ms	512	2006	Audinate
Q-LAN	3	≤ 24 bits	≤ 96 kHz	Typ 1 ms	512	2009	QSC
Ravenna	3	≤ 32 bits	≤ 384 kHz	Typ 1 ms	64	2010	ALC NetworX
AVB	2	≤ 32 bits	≤ 192 kHz	≥ 0,25 ms	400	2013	- - -

Os protocolos de camada 2 são encapsulados diretamente em um pacote camada 2 e, no caso do Ethernet, identificados pelo campo "Tipo". Por exemplo:

Campo "Tipo" do pacote Ethernet	Protocolo
0800	IP V4
0806	ARP
8035	RARP
8819	Cobranet (áudio)
8896	Ethersound (áudio)
88B5	AVB (áudio e vídeo)
88F7	PTP - Precision Time Protocol

Os protocolos de camada 4 são encapsulados em um pacote camada 3, no caso, um pacote IP.

Normalmente a camada 4 (transporte, roteamento) é utilizada conforme previsto na suite TCP-IP, com o protocolo UDP, que possui menor latência que o TCP. O protocolo de aplicação, normalmente o RTP (protocolo de transporte para aplicações em tempo real), definido pelas RFC 3050 (especificação) e 3051 (profiles), com os dados de amostragem do áudio e as informações de profile, é encapsulado no UDP.

Os protocolos que transportam áudio não operam em redes wireless (sem fio) da sére 802.11a/b/g/n, devido às limitações de desempenho das mesmas.

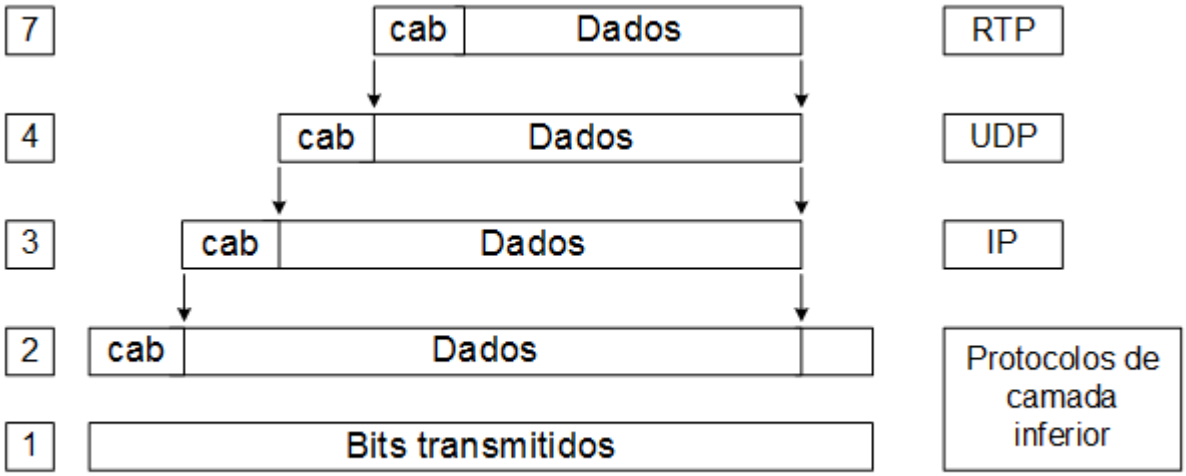

Fig. 7.26: Encapsulamento RTP

7.17 Protocolo Cobranet

Interface e protocolo, desenvolvidos pela Peak Audio, uma divisão da Cirrus Logic, camada 2, encapsulado em pacotes Ethernet. Opera em enlaces com taxa de 100 Mbps ou maior. Deve operar com retardo entre dois pontos Cobranet, menor que 3800 -0 + 250 µs.

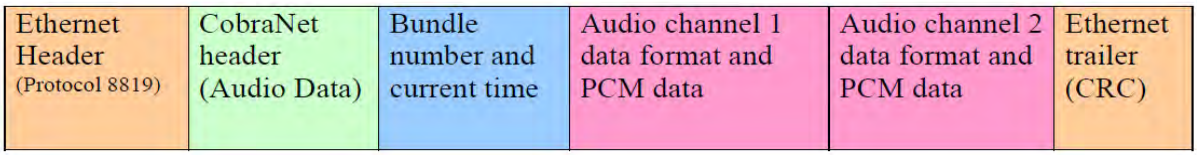

Fig. 7.27: Pacote Cobranet

A transmissão Cobranet se dá em feixes (bundle), numerados de 1 a 65279, que podem levar até 8 canais de áudio, com taxa de amostragem de 48 kHz, com resolução de 16, 20 ou 24 bits, em pacotes Ethernet que são transmitidos a cada ciclo isócrono.

Há dois tipos de feixes:

- Multicast: feixes 1 a 255
- Unicast: feixes 255 a 65279

O fabricante recomenda observar as latências dos switches ao desenhar a solução (150 µs para 100 Mbps e 15 µs para 1 Gbps) e os retardos dos enlaces envolvidos (20 µs para um enlace de fibra de 2 km, por exemplo).

Cada interface Cobranet pode ter 64 canais, organizados em feixes: 4 transmissores (32 canais) e 4 receptores (32 canais).

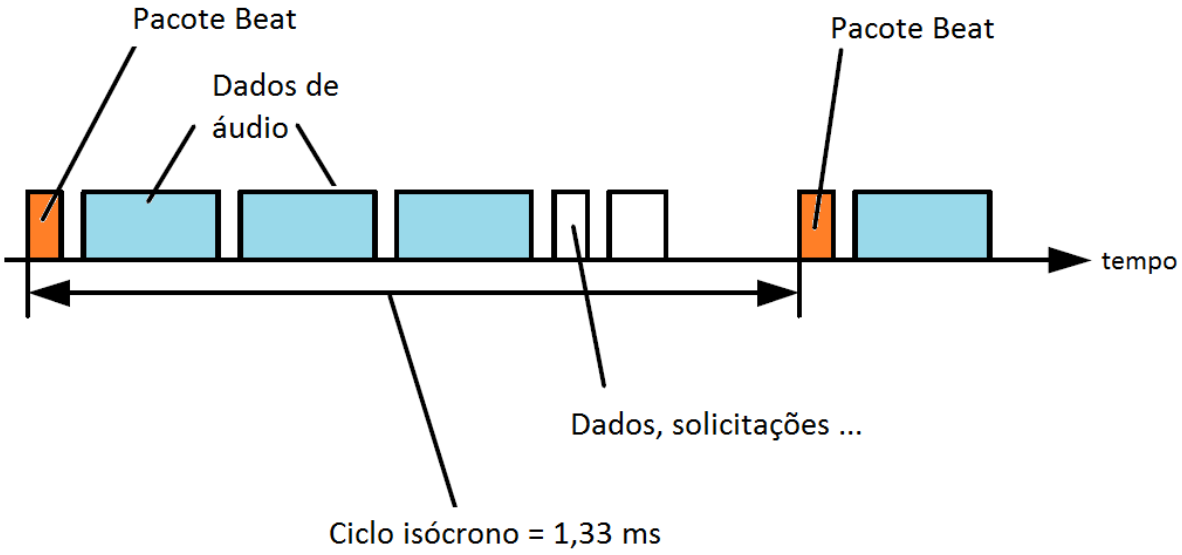

Fig. 7.28: Temporização da transmissão Cobranet

O sistema CobraNet possui latência configurável para 1,33 ms, 2,66 m, ou 5,33 ms, que deve ser aplicada a todo o sistema.

Pelo exposto acima, a menor latência de um enlace Cobranet será da ordem de 1,6 ms (latência de 1,33 ms mais os retardos de switch e transmissão)

Qualquer latência adicional, oriunda de conversões A-D ou processamento de sinal (DSP), soman-se às latências de transporte, entretanto, uma latência de até 10 ms é perfeitamente aceitável na maioria das aplicações.

Como os demais protocolos de áudio em rede, não deve passar por enlaces WiFi. Na verdade a variação do retardo no trajeto dos pacotes não deve exceder 0,25 ms sob pena de perda de sincronismo.

7.18 AES-67: streaming audio-over-IP

Abstract

*High-performance media networks support professional quality audio (**16 bit, 44,1 kHz and higher**) with low latencies (**less than 10 milliseconds**) compatible with live sound reinforcement. The level of network performance required to meet these requirements is available on local-area networks and is achievable on enterprise-scale networks. A number of networked audio systems have been developed to support high performance media networking but until now there were no recommendations for operating these systems in an interoperable manner. This standard provides comprehensive interoperability recommendations in the areas of synchronization, media clock identification, network transport, encoding and streaming, session description an connection management.*

A norma, publicada em 2013, define a interoperabilidade de áudio de alto desempenho (amostragem de 44,1 kHz ou maior, PCM linear com resolução de 16 bits ou maior) sobre redes IP, considerando baixa latência (menor que 10 ms).

Os principais parâmetros aceitos pela norma são:

Resolução

- 16 bits (RFC 3551)
- 24 bits (RFC 3190)

Amostragem

- 44,1 kHz (16 bits)
- 48 kHz (16 e 24 bits)
- 96 kHz (24 bits)

A norma ressalta o sistema deve possuir um único sinal de sincronismo (clock), seguindo a norma IEEE 1588:2008 Precision Time Protocol - PTP, e sinais amostrados na mesma taxa, o que permitirá a combinação de canais no receptor, como um mixer, por exemplo.

O protocolo RTP deve ser utilizado com um payload máximo de 1440 bytes, usando também o RTCP, O transporte deve ser feito com o protocolo UDP.

A norma recomenda que os equipamentos suportem "Packet Time", que é a duração do áudio digitalizado que vai codificado no pacote, conforme a tabela a seguir,

"Packet Time"	Taxa de amostragem [kHz]		
	48	44,1	96
125 µs	6	6	12
250 µs	12	12	24
333 µs	16	16	32
1 ms	48	48	96
4 ms	192	192	Não especificado

7.19 Protocolo Dante

Protocolo de transporte de áudio, baseado na camada 3 do modelo OSI, que roda encapsulado em UDP.

Na rede local o fabricante recomenda utilizar switches gerenciáveis gigabit Ethernet, principalmente se tiver 32 ou mais canais de áudio, com QoS de 4 ou mais filas, Diffserv.

O tráfego Dante pode ocorrer nos mesmos enlaces que os demais tráfegos, observadas as boas práticas de uma rede Ethernet.

O fabricante alerta para desabilitar o recurso EEE (Energy Efficient Ethernet ou 'Ethernet verde') se o switch tiver essa possibilidade.

7.20 Protocolo AVB

AVB ou "Audio Video Bridging" é uma interface com protocolo para transporte de dados com temporização crítica, como áudio e video, sobre Ethernet, publicado pelo IEEE como um conjunto de normas:

- IEEE 1722 Layer 2 Transport Protocol
- 802.1AS – Timing and Synchronization for Time-Sensitive Applications
- 802.1Qav – Forwarding and Queuing for Time-Sensitive Streams
- 802.1Qat – Stream Reservation Protocol (SRP)

O protocolo especifica latência máxima de 2 ms entre dois nós, permitindo até 7 switches com capacidade AVB, no trajeto.

O protocolo define "falantes" e "ouvintes" de áudio. Os falantes geram fluxos (streams) de áudio com um ou mais canais em pacotes de 125 µs e os ouvintes se inscrevem para receber os fluxos que desejarem. Um reprodutor de CD ou um microfone podem ser falantes e uma caixa de som ou um microcomputador podem ser ouvintes.

Como exemplo de sistemas, podemos ter:

- um reprodutor de CD gerando cinco canais de áudio para cinco caixas de som
- vários microfones gerando áudio em uma conferência
- vários microfone e instrumentos gerando áudio para uma mesa de mixagem de som

Não há regras sobre a dimensão do sistema AVB, apenas limites impostos pelos equipamentos de transmissão, falantes e ouvintes.

Um enlace Ethernet com 100 Mbps consegue transportar 9 fluxos de áudio estéreo (18 canais) ou um único fluxo com 45 canais.

7.21 Qual é o protocolo mais usado?

Esta pergunta não poderia deixar de ser respondida, a fim de dar uma idéia do que está acontecendo no mercado.

O site da **RH Consulting - Intelligent Audio Advice**[3] informa que realizou uma medida no mercado e obteve um total de 796 equipamentos utilizando protocolo de áudio em rede, com a seguinte posição para 2014:

Protocolo	Qtde
CobraNet	377
Dante	326
EtherSound	92
AVB	58
Ravenna	26
	879

[3] http://www.rhconsulting.eu/blog/files/every-networked-product-2014.html

Cobranet continua sendo o mais ofertado nos produtos, mas pela próxima figura, também extraída desse site, parece que está perdendo a força para o Dante. Isso é o que o mercado está dizendo, mas comparar protocolos de camadas diferentes envolve vários aspectos técnicos e estratégicos.

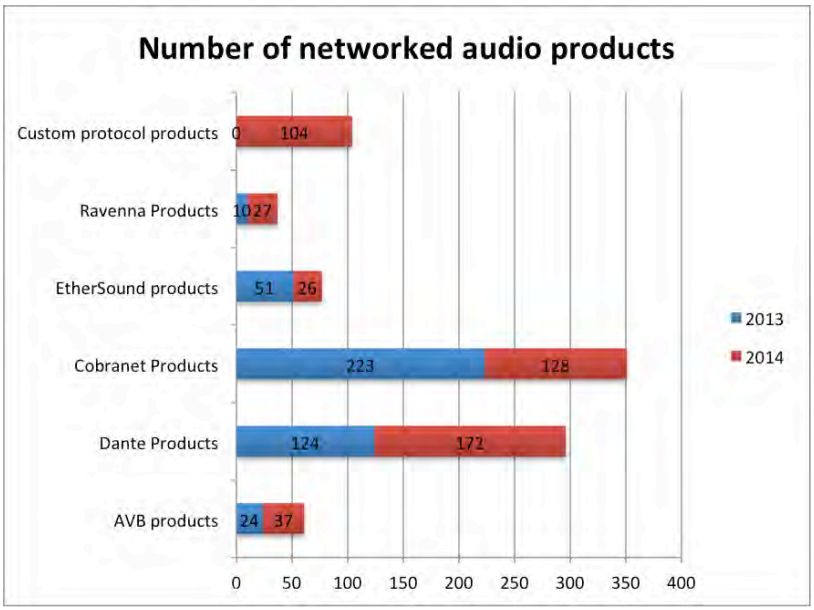

Fig. 7.29: Quantidade de produtos equipados com protocolos de áudio

Outra informação fornecida nesse site, é a quantidade de fabricantes ofertando cada protocolo, onde o protocolo Dante já ultrapassou o Cobranet:

Protocolo	Qtde
CobraNet	35
Dante	57
EtherSound	12
AVB	8
Ravenna	2
	186

7.22 Rane Mongoose - Cobranet

O Mongoose da Rane é um equipamento interessante.

É um roteador de áudio, de rotas fixas, apropriado para a transmissão de poucos canais (até 4) por um cabo de rede categoria 5e, sendo no máximo dois em cada sentido, entre dispositivos remotos denominados **RAD** (*Remote Access Device*), instalados em pontos de terminação de rede.

O mesmo cabo de rede cat 5e pode carregar sinais de microfone e de linha. A Rane disponibiliza vários modelos de RAD, que permitem escolher a configuração adequada.

A transmissão digital preserva o espectro original do sinal, sem distorções, tanto nas baixas quanto nas altas frequências, além de garantir uma excelente rejeição ao ruído.

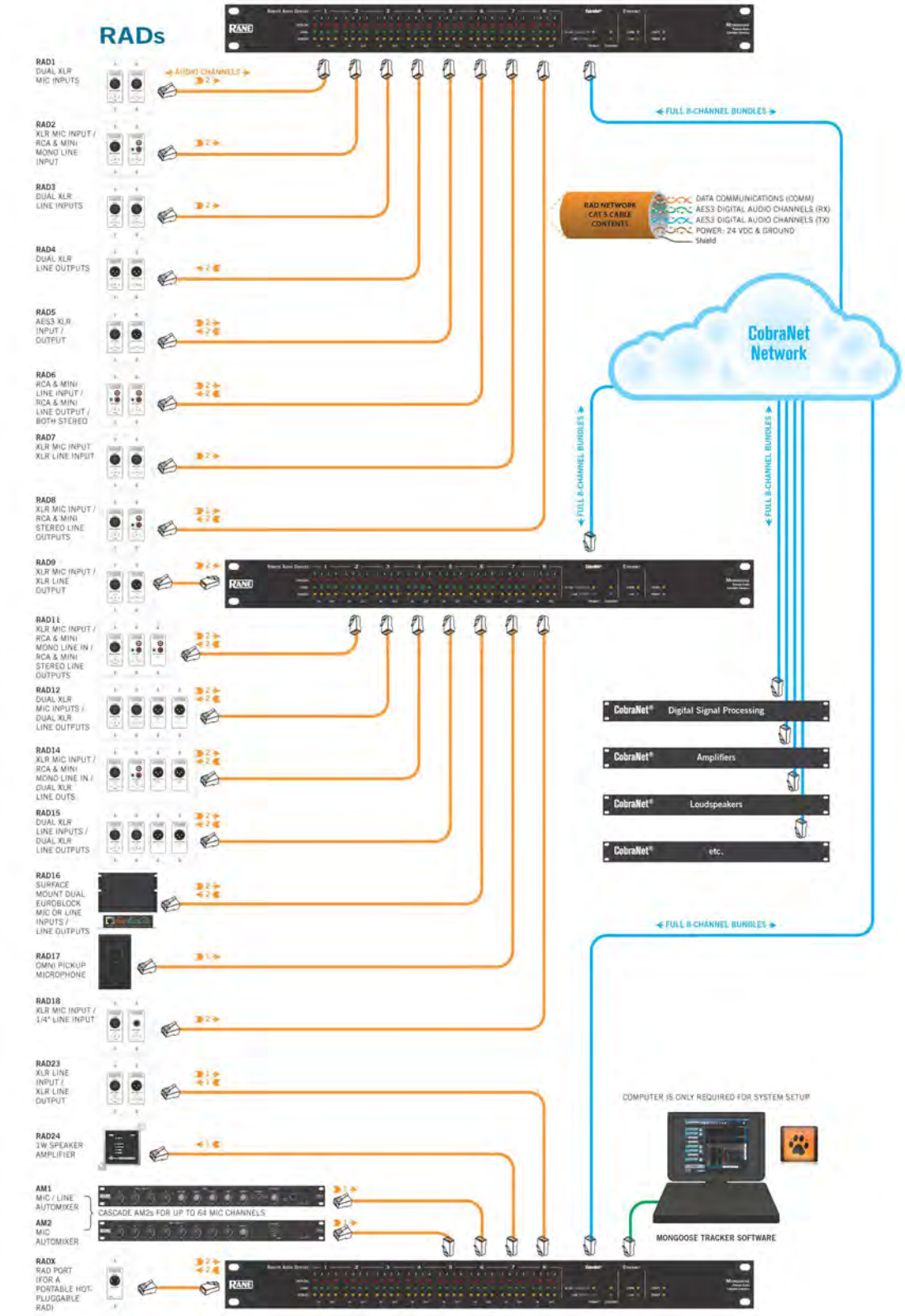

Fig. 7.30: Rane Mongoose - potencial de aplicações

São vantagens usar o Mongoose: transmissões com mais qualidade e fidelidade, flexibilidade para alterar rotas, flexibilidade para alterar o tipo de terminação e redução da quantidade de cabos da instalação. É uma excelente solução para abrir bundles Cobranet e distribuir os canais de áudio em posições com poucos canais (1 a 4).

A figura 7.31 mostra um esquema geral, que ilustra os possíveis dispositivos remotos do Mongoose e seu potencial de conexão de áudio em rede.

Características do Mongoose:

- 64 canais de áudio (32x32). São 32 com RAD e 32 com CobraNet.
- Para cada canal com RAD há 3 LEDs canal no painel frontal: "habilitado", "com sinal" e "sobrecarga"
- 32 LEDs no painel traseiro para indicar a situação de cada enlace RAD
- 8 portas RJ-45 para RAD, com alimentação pelo cabo
- Cada porta pode trafegar até 4 canais de áudio: 2 IN + 2 OUT
- 2 portas RJ-45 Cobranet para conexão em rede Ethernet. Uma primária e uma secundária (mesmo endereço MAC), para uso em solução CobraNet redundante. Estas portas não possuem auto-MDIX - cabo direto para ligar a um switch e cabo cross para ligar a um equipamento CobraNet.
- 2 bundles Cobranet com 8 canais cada
- 1 porta Ethernet para gerenciamento e configuração

Fig. 7.31: Rane Mongoose - Matriz de roteamento

As rotas são estabelecidas pelo software de gerenciamento. É possível criar rotas entre RADs e canais CobraNet dentro de um dos dois bundles. É possível mandar o mesmo sinal de entrada para diversos canais de saída (distribuição).

7.22.1 Aplicação Mongoose: RAD - RAD

A figura 7.32a mostra aplicação mais básica: um Mongoose fazendo o roteamento de canais em um RAD para canais em outro RAD. A figura 7.32b mostra um exemplo da matriz de configuração fazendo uma distribuição do canal de entrada para dois canais de saída, em RADs distintos.

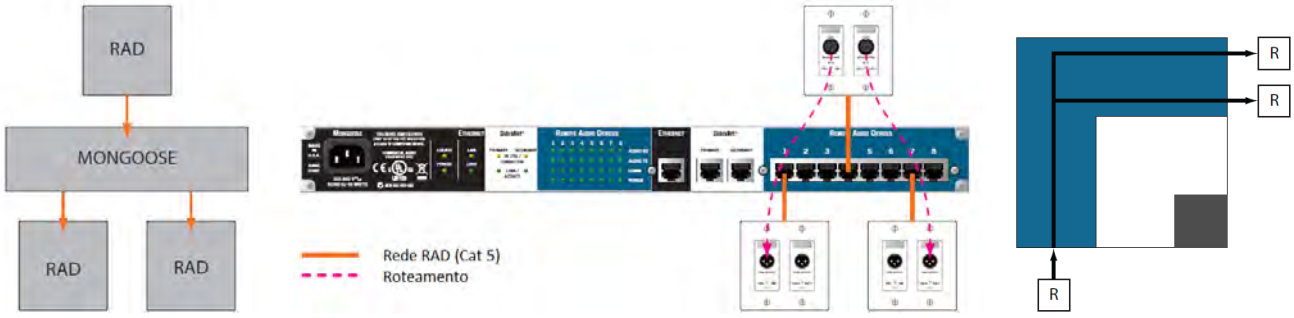

Fig. 7.32: Aplicação do Mongoose: RAD - RAD

7.22.2 Aplicação Mongoose: RAD - CobraNet

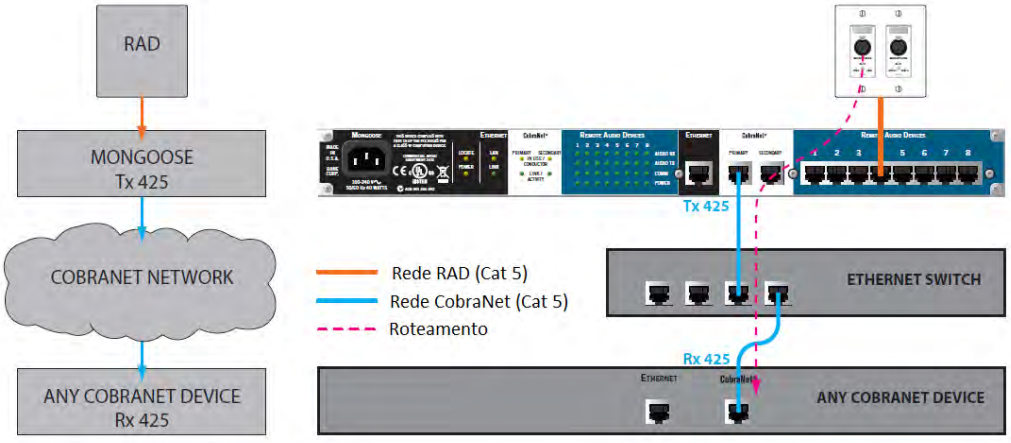

Fig. 7.33: Aplicação do Mongoose: RAD - CobraNet

Nesta aplicação, um canal de entrada via RAD é encaminhado utilizando o bandle 425 para um dispositivo CobraNet como Tx425. O dispositivo CobraNet recebe como Rx425.

7.22.3 Aplicação Mongoose: RAD - CobraNet - RAD

Nesta aplicação, um canal de entrada via RAD é encaminhado para outro RAD, passando por uma rede CobraNet, que utiliza um switch de rede Ethernet.

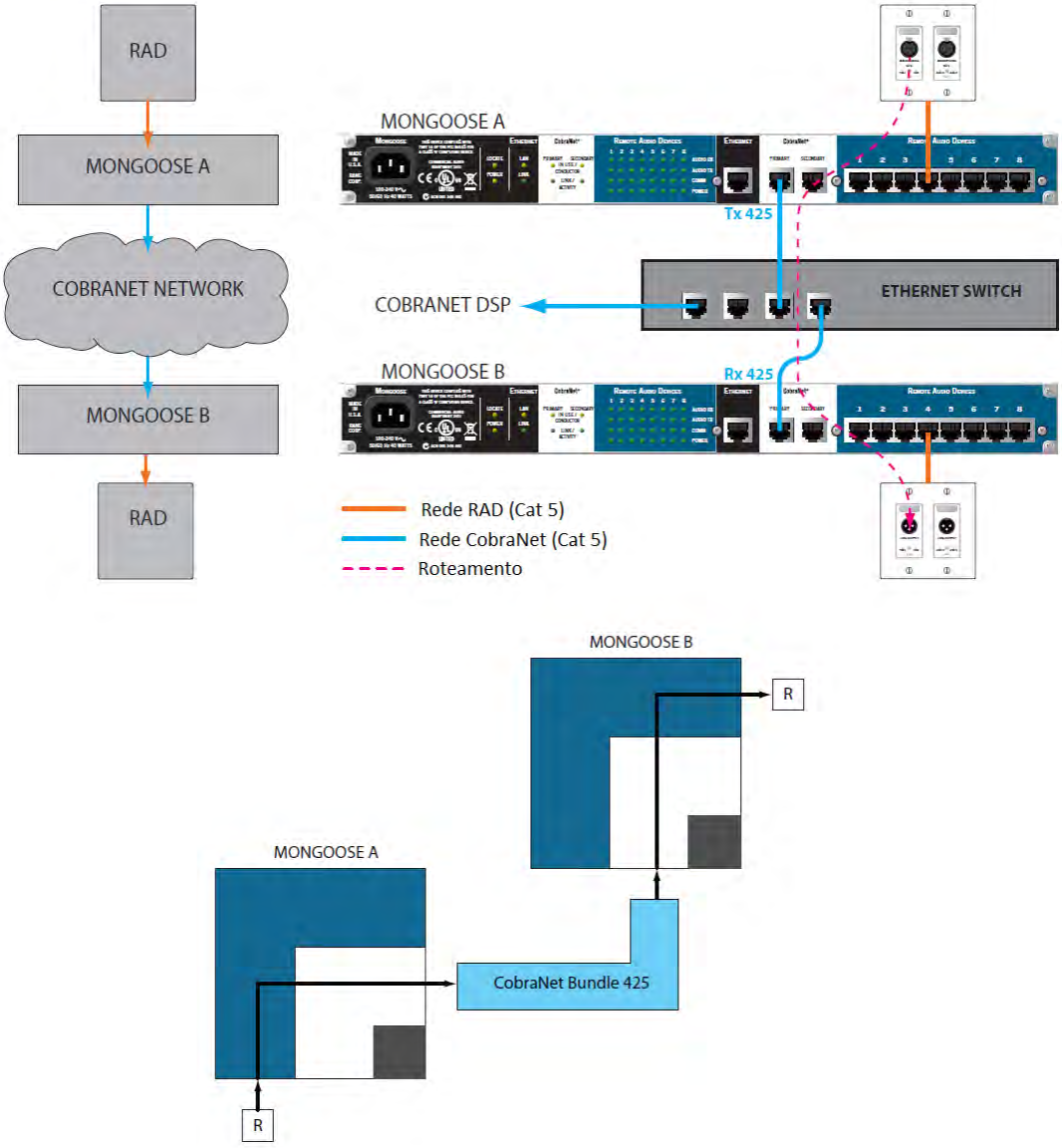

Fig. 7.34: Aplicação do Mongoose: RAD - CobraNet - RAD

7.22.4 Aplicação Mongoose: n-Mongoose-RAD - CobraNet

Esta aplicação mostra como os canais oriundos de dois Mongoose diferentes, podem ser combinados e encaminhados a um dispositivo CobraNet.

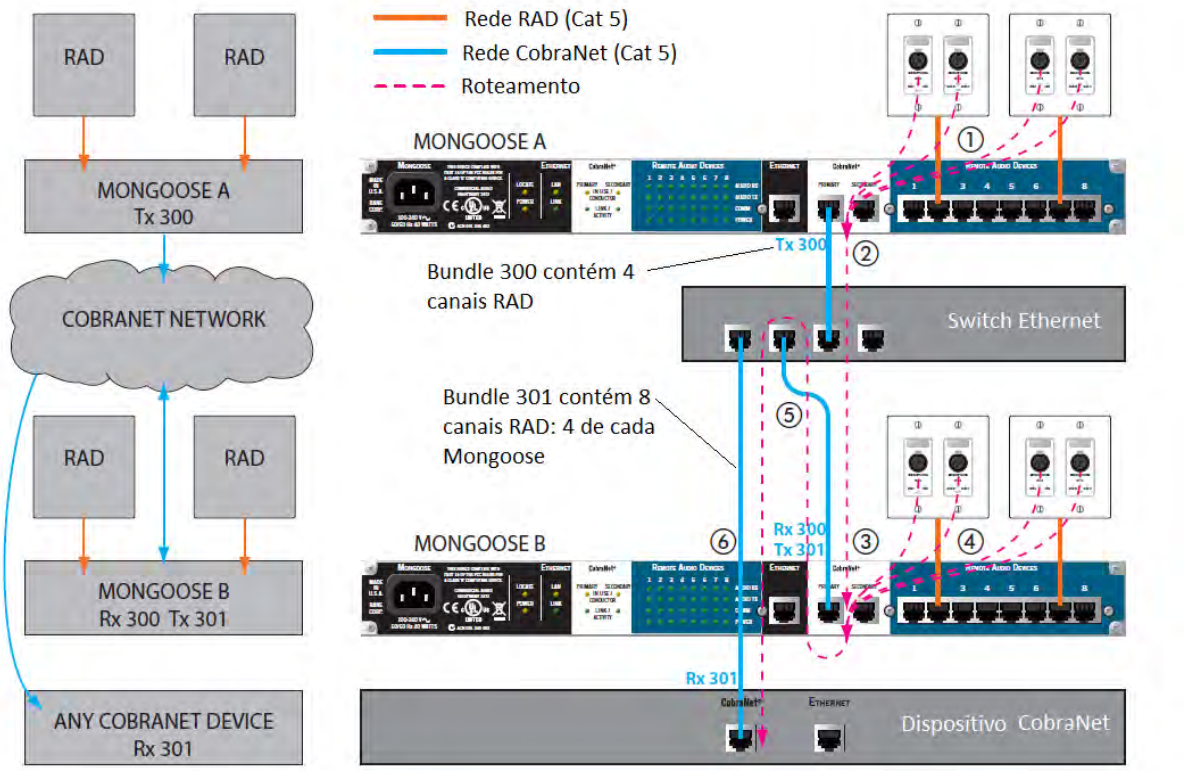

Fig. 7.35a: Aplicação do Mongoose: n-Mongoose-RAD - CobraNet

A sequência de eventos fica:

1) Mongoose **A** recebe 4 canais de RADs
2) Mongoose **A** manda 4 canais para o Mongoose **B** via CobraNet bundle 300.
3) Mongoose **B** recebe 4 canais de RADs
4) Mongoose **B** também recebe 4 canais pelo bundle 300
5) Mongoose **B** manda os 8 canais para o Mongoose **B** via CobraNet bundle 301.
6) O dispositivo CobraNet recebe os 8 canais no bundle 301

A matriz de configuração fica com o formato da figura 7.35b.

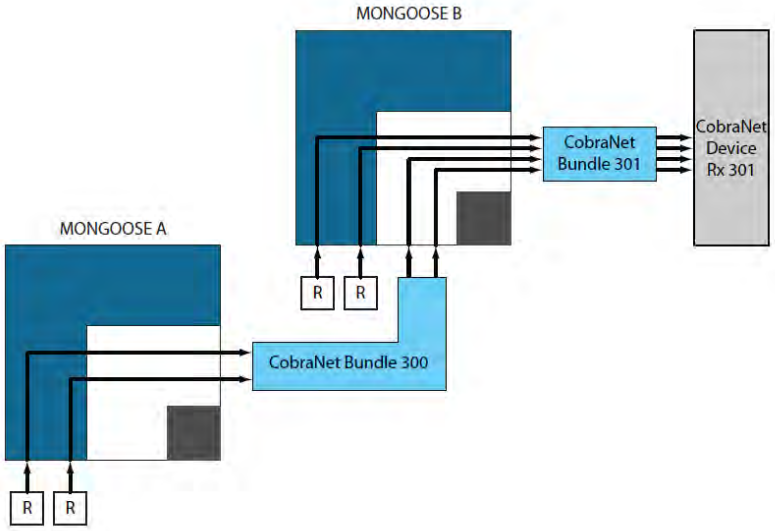

Fig. 7.35b: Aplicação do Mongoose: n-Mongoose-RAD - CobraNet

7.23 Terminação de rede Cobranet

Terminações de rede falando diretamente o protocolo Cobranet sobre Ethernet, com PoE, podem ser necessárias, dependendo do projeto. Como exemplo temos a terminação Atterotech INBOX X2, que possui duas entradas analógicas XLR/TRS (linha ou microfone) e uma interface Cobranet Ethernet PoE.

Fig. 7.36: Terminação de rede CobraNet

7.24 Processador Cobranet de pequeno porte

Processadores CobraNet de pequeno porte podem ser interessantes em algumas soluções pequenas e médias. O equipamento Atterotech Voice 4IO é um processador Cobranet com 4 entradas e quatro saídas de linha, analógicas, e uma interface CobraNet com um bandle.

Fig. 7.37: Processador CobraNet de pequeno porte

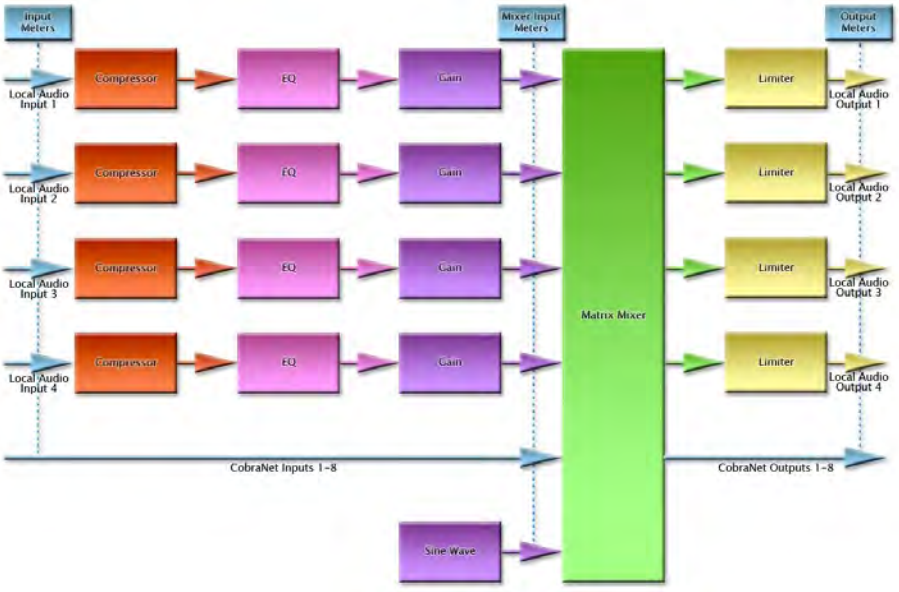

Fig. 7.38: Diagrama do Atterotech Voice 4IO

7.25 Rane Exp2x - Dante

O processador de sinais HAL1x, da Rane, possui um módulo de expansão com interface Dante, denominado **Exp2x**, com duas portas de rede que podem operar em modo redundante.

Cada Exp2x adiciona 32 canais de recepção de áudio e 32 de transmissão, seguindo o protocolo Dante, que suportam taxa de amostragem de até 96 kHz.

O equipamento executa conversão automática da taxa de amostragem para 48 kHz, podendo se conectar a equipamentos com taxas de 44,1 kHz, 48 kHz, 88,2 kHz e 96 kHz.

Como qualquer outro dispositivo Dante, as taxas de amostragem da entrada e da saída devem ser iguais.

Com a conexão do módulo, o software Halogen do processador HAL1x passa a ter acesso aos canais Dante.

Fig. 7.39: Painel traseiro do Rane Exp2x - Dante

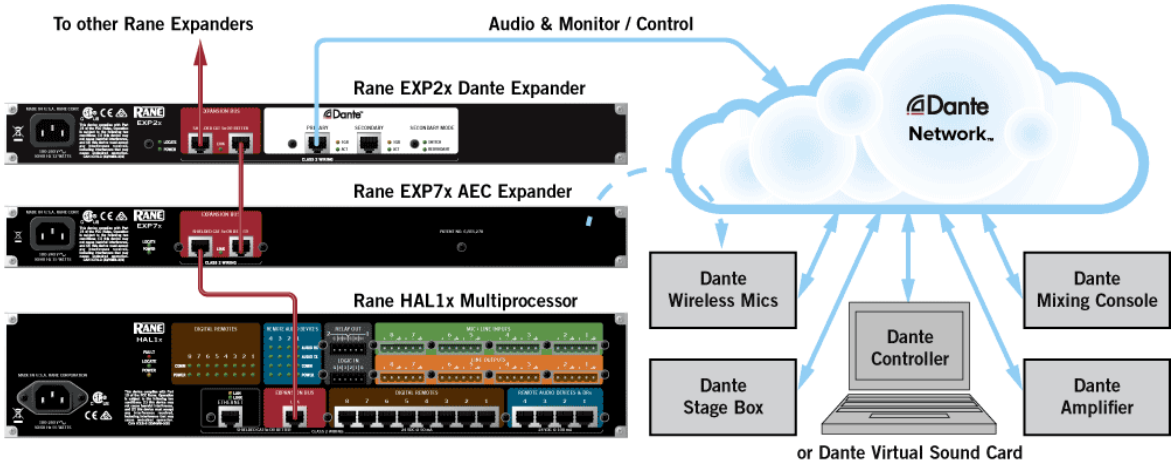

Fig. 7.40: Aplicação do Rane Exp2x - Dante

7.26 Processador de áudio com Dante (Extron)

Fig. 7.41: Matriz de áudio Extron DMP128

Este processador de áudio da Extron possui com 12 entradas (Mic/Line) e 8 saídas analógicas e uma matriz interna para mixagem dos canais. Possui modelos com AEC, Dante e entrada para telefone (POT)

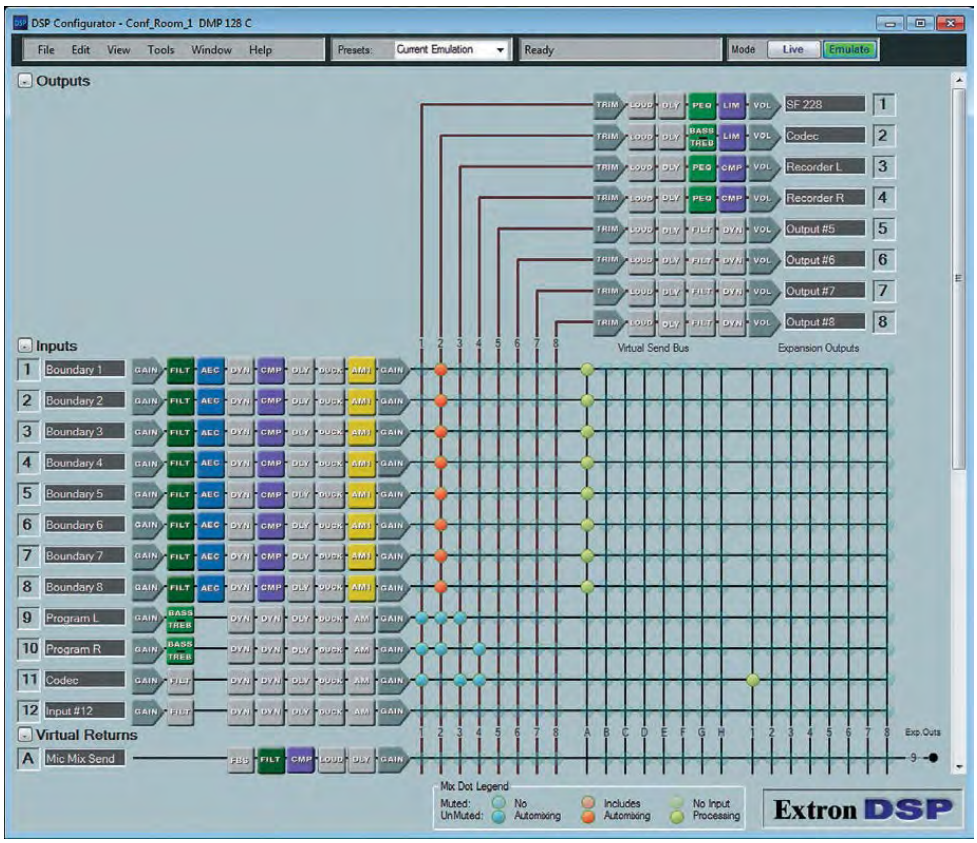

Fig. 7.42: Matriz de áudio Extron DMP128

Modelo	Característica	Código
DMP 128	12x8 ProDSP	60-1211-01
DMP 128 AT	12x8 ProDSP, Dante	60-1211-10
DMP 128 C	12x8 ProDSP, AEC	60-1178-01
DMP 128 C AT	12x8 ProDSP, AEC, Dante	60-1178-10
DMP 128 C P	12x8 ProDSP, AEC, POTS	60-1179-01
DMP 128 C P AT	12x8 ProDSP, AEC, POTS Dante	60-1179-10

7.27 Terminação de rede Dante (Extron)

Terminações de rede falando diretamente o protocolo Dante sobre Ethernet, com PoE, podem ser necessárias, dependendo do projeto. Como exemplo temos a terminação Extron AXI 22 AT 2, que possui duas entradas analógicas XLR/TRS (linha ou microfone), duas saídas analógicas de linha e uma interface Dante Ethernet PoE.

Fig. 7.43: Extron AXI 22 AT 2: Terminação de rede Dante

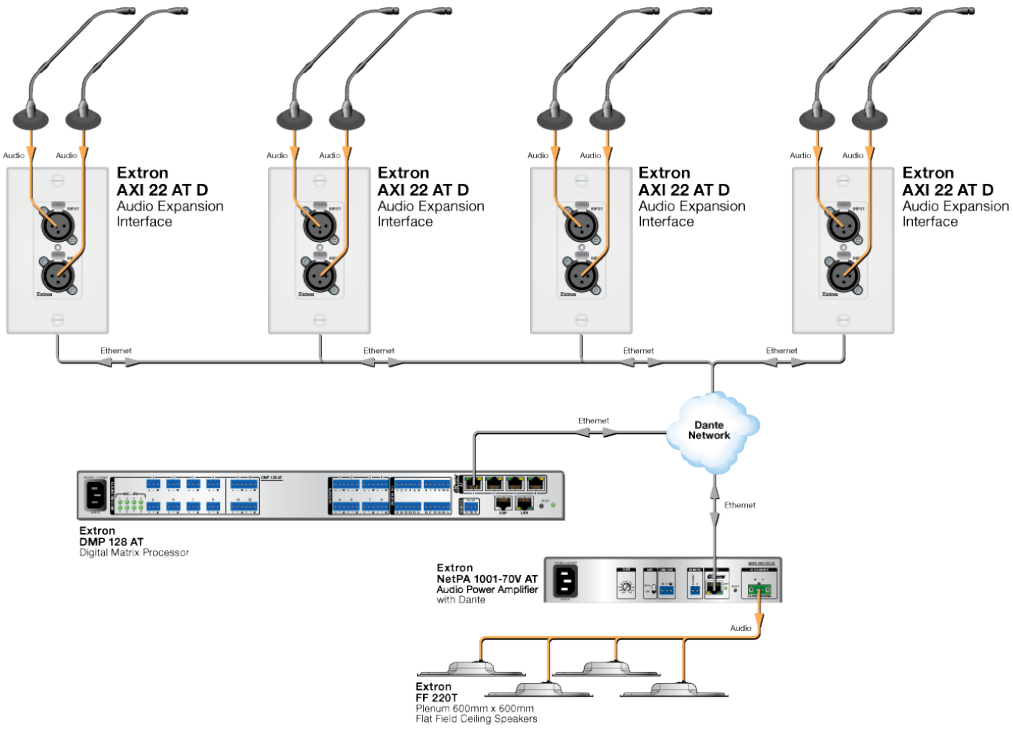

Fig. 7.44: Extron: Aplicação com terminações Dante remotas

7.28 Microfone Dante (Audio-Technica)

O microfone cardióide capacitivo ATND921 da Audio-Technica se conecta diretamente em rede, utilizando o protocolo Dante sobre Ethernet PoE.

Fig. 7.45: Audio-Technica: microfone Dante

7.29 Cognição auditiva: a decodificação cerebral do som

A decodificação do som pelo nosso sistema auditivo vai além da questão da percepção do som em seus parâmetros básicos: frequência e intensidade.

Todos os sons gerados em um ambiente são somados mecanicamente no espaço, ou seja, as ondas sonoras chegam misturadas em nosso ouvido.

Podemos segregar e identificar a voz de uma determinada pessoa falando ao mesmo tempo que outras, ou o solo de um específico instrumento musical tocando em uma banda.

O cérebro consegue separar uma sequência específica de som, que incorpora características além da frequência e amplitude, percebidas isoladamente, tais como a forma com que o espectro de frequência muda ao longo do tempo e a defasagem entre os sinais recebidos nos dois ouvidos, identificando-a como uma combinação de seu interesse naquele instante e eventualmente compará-la a um padrão já conhecido. Por exemplo, podemos estar interessados em apreciar o solo de um violino, executado no meio de uma música, juntamente com muitos outros instrumentos.

O que chega ao nosso ouvido é a soma global das pressões geradas por todas as fontes a cada instante, ou seja, a componente de 440 Hz de um piano se soma à componente de 440 Hz de um violão e de todos os instrumentos de uma orquestra, por exemplo. Como nosso ouvido identifica e separa, exatamente a intensidade de cada instrumento, permitindo a identificação a composição espectral de cada um ao longo do tempo?

É um problema complexo, convenhamos. O cérebro tem que identificar as sequências, conhecendo apenas a soma delas para cada frequência a cada instante, sem saber até mesmo quantas são e, à primeira vista, sem ter outra informação, e separá-las claramente ao longo do tempo. O trabalho do cérebro vai além da execução de uma análise espectral do sinal recebido, pois tal processo, sozinho, não é capaz de identificar quantas parcelas se

somaram, e a quem pertencem, para formar uma determinada componente de frequência em um determinado instante, do sinal global recebido.

Nosso sistema sensorial como um todo trabalha para estabelecer uma avaliação global do ambiente que nos cerca, seja por necessidade de sobrevivência, segurança, lazer ou arte.

Este assunto é matéria de pesquisa da psico-acústica, que estuda a percepção subjetiva do som.

Mais ligado, talvez, à neuro-ciência, a interpretação matemática do processo de reconhecimento das sequências sonoras ainda é matéria de muitas pesquisas.

O processo de análise da cena acústica pelo cérebro humano é conhecido como ASA, ou "*Auditory Scene Analysis*"

O processo foi estudado pelo psicólogo Bregman[4], que descreve a integração de sequências sonoras e a segregação frente a outras sequências, pelo nosso cérebro. Experiências mostram que o cérebro tende a segregar se os padrões repetitivos se afastam no domínio da frequência ou se aproximam no domínio do tempo, ou seja:

- A frequência de repetição dos padrões aumenta, ou
- Os espectros de frequência, dos padrões, de afastam.

7.29.1 Experiências

Estas experiências foram criadas no laboratório de pesquisa de cognição auditiva do Departamento de Psicologia da Universidade de McGill, no Canadá, pelo professor Albert Bregman, especialista em linguagem, percepção e cognição.

- Laptop
- Caixa amplificada
- Track 1: "Stream Segregation - 6 tones"

Demonstração da segregação de sequências em função da aproximação no domínio do tempo

1) Alta = 1600, 2000, 2500 Hz (intervalos ≈ 4 semitons)
2) Baixa = 350, 430, 550 Hz (intervalos ≈ 4 semitons)

- Laptop
- Caixa amplificada
- Track 3: "Loss of Rhythm"

Demonstração da segregação de sequências em função da aproximação no domínio do tempo e espectros afastados

1) Quando f_1 = 500 Hz e f_2 = 1400 Hz (≈ 18 semitons)

[4] Bregman, A.S., "*Auditory scene analysis*", *Encyclopedia of Neuroscience*, Academic Press, Oxford, UK, 2008

2) Quando f_1 = 1320 Hz e f_2 = 1400 Hz (≈ 1 semitom)

- Laptop
- Caixa amplificada
- Track 2: "Pattern Recognition"

Demonstração de como uma sequência adicional pode atrapalhar o processo cerebral de identificação de sequências: Princípio da camuflagem

- Laptop ♫ ♫
- Caixa amplificada
- Música: "Cononic Sonata" de Telemman

Exemplo de sequências na música clássica

- Laptop ♫ ♫
- Caixa amplificada
- Música: "Carinhoso " de Pixinguinha e , executada por Damien Groleau

Exemplo de sequências na música popular

https://www.youtube.com/watch?v=_EN1HNJhPJk

- Laptop ♫ ♫
- Caixa amplificada
- Música: "Apanhei-te cavaquinho" de Ernesto Nazareth, executada por Armandinho

Exemplo de sequências na música popular