UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA
DEPARTAMENTO DE COMUNICAÇÕES

METODOLOGIAS DE AVALIAÇÃO DE ALGORITMOS DE CODIFICAÇÃO DE VOZ

SIMÃO FERRAZ DE CAMPOS NETO

Orientador: Prof.Dr. FÁBIO VIOLARO

Banca Examinadora:
Fábio Violaro (UNICAMP)
Abraham Alcaim (CETUC)
João Marcos Travassos Romano (UNICAMP)
Leonardo Mendes (UNICAMP)

Tese apresentada à Faculdade de Engenharia Elétrica da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de MESTRE EM ENGENHARIA ELÉTRICA.

Campinas, Abril de 1993

METODOLOGIAS DE AVALIAÇÃO DE ALGORITMOS DE CODIFICAÇÃO DE VOZ

SIMÃO FERRAZ DE CAMPOS NETO
Orientador: Prof.Dr. FÁBIO VIOLARO

Campinas, Abril de 1993

Campos Neto, Simão Ferraz de
Metodologias de Avaliação de Algoritmos de Codificação de Voz / Simão Ferraz de Campos Neto, 1993.
159 páginas.
Tese (Mestrado) - Universidade Estadual de Campinas, 1993.
1. Processamento e Análise de Voz 2.codecs 3. Análise Estatística. I. Título
621.38043

Este documento foi editorado com o sistema LaTeX e impresso numa impressora laser LPS20 (DEC). Versão de 13 de maio de 1993. Publicado pela Unicamp. Cópias desta tese podem ser obtidas:

Biblioteca da Faculdade de Engenharia Elétrica
(Prédio da Biblioteca Central)
Universidade Estadual de Campinas – Unicamp
13100-000 Campinas SP

Resumo

Neste trabalho são apresentados diversos aspectos relacionados à avaliação da qualidade subjetiva e objetiva de algoritmos de codificação de voz, como metodologia de testes, infra-estrutura, descrição de algoritmos de referência e de medidas objetivas. Este trabalho é importante por fornecer subsídios para a implementação de metodologias efetivas que garantam a qualidade de codificadores do sinal de voz quando utilizados na rede telefônica. Após a parte tutorial deste trabalho, analisam-se os resultados de um dos testes subjetivos para a língua portuguesa realizados durante a padronização da hoje Recomendação CCITT G.728 e os resultados de medidas objetivas de qualidade, bem como a sua capacidade de estimar a qualidade subjetiva.

Abstract

Many topics related to the subjective and objective assessment of speech quality for speech coding algorithms are presented in this work, such as: test methodologies, laboratorial facilities, description of reference algorithms, and objective measures. This work is important because it gives the basement for the development of effective methodologies that assure the quality of speech coders for use in the telephone network. After the tutorial part, it is presented the results of one of the subjective tests for the Portuguese language, made during the standardization of the present CCITT Recommendation G.728, and the results of some objective measures of quality, as well as their capacity of estimating the subjective quality.

À minha filha Íris e
à minha esposa Yoshiko.

Agradecimentos

Este trabalho não poderia ter sido realizado sem a ocorrência de diversos fatos. Primeiro, a decisão estratégica do CPqD/Telebrás, em 1989, de participar dos testes subjetivos para o codificador CCITT a 16 kbit/s como um mecanismo de formação de recursos humanos relacionados à avaliação de qualidade de algoritmos de codificação de voz. Fui escolhido para acompanhar esta atividade e devo deixar meu agradecimento ao Eng.José Sindi Yamamoto pela sua confiança e pelas discussões sempre produtivas. Também não houvesse a disposição em enveredar por novos caminhos, não teria contado com a orientação fiel do Prof.Dr.Fábio Violaro, que dedicadamente permitiu evoluir este trabalho, sempre incentivando minha liberdade criativa (mas sempre me pedindo para não deixar o trabalho crescer muito...). Também contei com a colaboração dos colegas, que gentilmente cederam suas vozes para a geração do material de voz para os testes subjetivos, ou seus ouvidos e julgamento para as sessões de testes subjetivos. Para esta última atividade, contei também com o apoio de amigos e parentes, que muitas vezes vieram ao Centro em feriados e finais de semana. A estes, meus sinceros agradecimentos. Aos colegas de trabalho, preciso agradecer a compreensão e desculpar-me pelos maus humores, sempre freqüentes ao longo desta jornada. E, em casa, agradecer o carinho constante de minha esposa e filha, sem cujo apoio este trabalho por certo malograria.

Preface

In the late 1980's, various methods, available on a regional level, were proposed for the coding of speech at 16 kbit/s. Not only was the proliferation of regional and incompatible algorithms detrimental to global interworking, but the modest quality of these algorithms threatened to impose undue constraints on overall network transmission planning. In this context, a question arose within the International Telephone and Telegraph Consultative Committee (CCITT) as to whether it was possible to define a new "universal" 16 kbit/s algorithm that could be used worldwide for all potential applications and still achieve network transparency. This question led to a set of highly challenging requirements which spurred speech researchers to investigate innovative ways to meet the challenge.

From the beginning, this CCITT effort required close coordination between experts in speech coding, and experts on transmission performance who produced a set of requirements and methodologies for subjective and objective testing of voice codecs. The subjective test plan called for extensive multilingual tests and required the collaboration of laboratories in many countries, while the objective test plan called for non-voice signaling performance assessments as well as objective measurements using voice signals.

In May 1922, following a four-year effort, CCITT Recommendation G.728 on Low-Delay Code-Excited Linear Prediction (LD-CELP) was finally approved for the "telephone-quality" coding of speech at 16 kbit/s. It was through the definition and implementation of extensive and detailed test methodologies, both subjective and objective, that the determination of the LD-CELP algorithm's ability to meet the specified performance requirements was possible.

It is noted that this was a pioneering experiment whereby standardization stimulated and led, rather than followed or consolidated, innovative developments in speech coding and testing. It is also noted that the test methodologies developed in this process are expected to be a model for years to come, as evidenced by their planned use in CCITT's 8 kbit/s speech coding standardization effort.

It is with this preface in mind that is my pleasure to introduce this present thesis, which I consider to provide a significant insight into the selection of speech coding technologies for the "network of the future".

Spiros Dimolitsas.
Chairman of the CCITT Ad Hoc Group
of Experts on 16 kbit/s Speech Coding.
Comsat Laboratories, April of 1993.

Conteúdo

Capítulo 1

Introdução

O ciclo de implementação de algoritmos de codificação de voz apresenta em geral três etapas. Inicialmente, versões de um algoritmo são simuladas em linguagem de alto nível (e.g. Fortran ou C), utilizando representação em ponto flutuante, muitas vezes com precisão dupla; esta é a fase *exploratória* do algoritmo. Uma vez adquirida a confiança na implementabilidade e qualidade do algoritmo, parte-se para a fase de *implementação*. Quando o objetivo é uma implementação com aritmética de ponto fixo, a partir da versão exploratória implementa-se uma simulação empregando somente operações aritméticas em ponto fixo, usando ainda uma linguagem de alto nível. Quando esta implementação tiver atingido o grau de qualidade necessário, ela será a referência para a *implementação em hardware* do algoritmo, ambas devendo se comportar de maneira idêntica e apresentarem a mesma qualidade. Em geral, esta etapa final da implementação se resume à re-escrita da implementação em ponto fixo na linguagem assembly de um DSP (*Digital Signal Processors*, microprocessadores com arquitetura otimizada para processamento digital de sinais [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]). Em outras palavras, a última fase resume-se ao desenvolvimento de um firmware. Para o caso de algoritmos com aritmética em ponto flutuante, a segunda etapa perde o sentido, passando-se então direto da fase exploratória para o desenvolvimento do firmware.

Nessa descrição um termo foi utilizado, mas não claramente definido: qualidade. O que é a qualidade de um algoritmo codificador de voz? Como se avaliar essa qualidade? Quais as ferramentas e infra-estrutura necessárias?

Desde o princípio das atividades de codificação de voz dentro do CPqD *(Centro de Pesquisas e Desenvolvimento)* da Telebrás, no começo de 1987, com o embrionário grupamento de processamento digital de voz, atenção foi dada à necessidade de se identificar ferramentas e metodologias para se avaliar a qualidade dos algoritmos implementados.

Esta dissertação de mestrado busca mostrar e documentar o trabalho realizado dentro da atividade de avaliação de qualidade de sinais codificados de voz dentro do CPqD, a infra-estrutura montada e as ferramentas disponíveis. Um subsídio muito importante a esta atividade veio do acompanhamento e participação intensa do autor em dois grupos de trabalho dentro do CCITT (Conselho Consultivo Internacional para Telegrafia e Telefonia), com sede em Genebra, Suíça, de Março de 1989 até hoje. Esses grupos, um lidando com os aspectos de codificação (Comissão de Estudos XV, Grupo de Trabalho 2, ou SG XV, WP XV/2), outro com os aspectos de qualidade (Comissão de Estudos XII, SG XII), produziram um trabalho muito

importante neste período de tempo que foi a geração do padrão CCITT de codificação de voz a 16 kbit/s, LD-CELP. O CPqD patrocinou a realização dos testes subjetivos (definidos pelo SGXII) deste codec para a língua portuguesa, produzindo uma significativa especialização na área, bem como inúmeras *Contribuições* a ambos os grupos do CCITT. Adicionalmente, junto ao SGXV, foi criado um Grupo de Usuários de Ferramentas de Software (UGST), com o propósito de definir e implementar uma biblioteca de software do CCITT para auxiliar a geração de padrões de codificação de voz, do qual o autor foi coordenador de Julho de 1990 até o fim de 1992, sendo responsável pela implementação de diversos softwares da biblioteca hoje publicada pelo CCITT, conhecida como STL92 [11], bem como de uma recomendação associada, G.191 [12].

No Capítulo 2 são descritos vários esquemas de codificação de voz padronizados pelo CCITT, com ênfase aos algoritmos utilizados durante os testes subjetivos do codificador LD-CELP, hoje o padrão CCITT G.728.

No Capítulo 3 é descrita a infra-estrutura básica de hardware e software necessária para a avaliação de algoritmos de codificação de voz.

No Capítulo 4 são apresentados os conceitos básicos e a fundamentação para a realização da avaliação subjetiva de codificadores de voz. Este é com certeza o capítulo mais denso deste trabalho.

No Capítulo 5 é feita uma apresentação de vários métodos de avaliação objetiva da qualidade de codificadores de voz, com maior ênfase a métodos tradicionais. Também são descritos alguns dos métodos ora em estudo pelo CCITT [13, 14].

No Capítulo 6 são apresentados resultados de avaliações subjetivas de alguns algoritmos de codificação de voz. São também apresentados resultados comparativos com algumas das técnicas objetivas de avaliação de qualidade descritas, bem como a avaliação do uso de diversas medidas objetivas como estimadores da qualidade subjetiva.

Finalmente no Capítulo 7 as conclusões fazem um balanço crítico dos resultados alcançados neste trabalho, bem como se apontam direções para melhoria da infra-estrutura e metodologia implementadas.

Ainda nesta introdução, é apresentado um glossário onde os termos mais usados nesta dissertação são definidos e explicados. Em relação à terminologia, sempre surgem dúvidas sobre como melhor utilizar os termos técnicos definidos em língua inglesa. Alguns possuem uma tradução fiel (*envelope* por "envoltória"), outros o uso consagrou uma versão não tão castiça (*ratio* por "relação"). Há ainda o caso das palavras cujo termo em português fica sendo o termo em inglês, por não haver uma palavra *única* com o sentido exato em nossa língua (por exemplo, *codebook* ao invés de "dicionário de símbolos"), ou aquelas que não constam de dicionários mas cuja forma, por provirem do latim, podem ser utilizadas, dada sua precisa definição, diretamente em português. A regra geral seguida nesta dissertação privilegia o jargão paulista, mais especificamente aquele utilizado pela comunidade de especialistas na Unicamp e CPqD, procurando sempre minimizar a agressão à língua-mãe.

1.1 Glossário

ACR:
Absolute Category Rating

ADPCM:
Adaptive Pulse Code Modulation. Algoritmo de codificação de voz padronizado pelo CCITT na Recomendação G.726.

ANOVA:
Analysis of Variance

ANSI:
American National Standards Institute

BetaMax, Sony:
Termo utilizado na literatura sobre testes subjetivos para designar um dispositivo de armazenamento de sinais de voz da Sony. Constituído de um digitalizador cujo sinal de saída é um sinal na faixa de vídeo, normalmente conectado a um gravador de vídeo cassete Betamax. Utilizado em testes subjetivos para o intercâmbio de sinais de voz no passado pelo CCITT, encontra-se em desuso.

Banda parcial:
Classe genérica de sinais que ocupam apenas uma pequena parte do espectro permitido, e.g. tons, sinalização e sinais de dados modulados pelos mais diversos tipos de modems. Termo utilizado na Rec.CCITT G.726.

CCIF:
Comitê Consultivo Internacional para Telefonia, findo em 1956

CCITT:
Comitê Consultivo Internacional para Telegrafia e Telefonia, criado em 1956 e findo em 1993.

Condição:
Num teste subjetivo, é ou um codec sob testes com uma condição de contorno (por exemplo, taxa de erros de 10^{-3} e 1 transcodificação em cascata) ou uma configuração de referência (por exemplo, MNRU com Q=5dB). As condições escolhidas para um teste devem representar um subconjunto das configurações possíveis que plenamente descrevam a situação de circuito real.

DAT:
Digital Audio Tape, Fita de áudio digital. Fita menor que a cassete normal, porém com armazenamento do sinal de áudio na forma digital, a uma taxa de amostragem de 48kHz. Utilizada em testes subjetivos para o intercâmbio de sinais de voz.

dBA:
o nível absoluto (medido com a ponderação "A" de freqüência) de um sinal com um medidor de índice sonoro que segue a Recomendação CCITT P.54 (Blue Book), ou, equivalentemente, a Norma IEC–179.

dBm:
o nível absoluto de potência em dB, referido à potência de 1 mW (i.e., um sinal com 1 mW de potência é dito estar a 0 dBm).

dBmp:
o nível absoluto de potência ponderada por um filtro psofométrico (CCITT Rec.P.53), expressa em dB, referido a 1mW.

dBm0:
o nível absoluto de potência em dB referido a 1mW, medido em um ponto de nível relativo zero no circuito.

dBm0p:
o nível absoluto de potência psofométrica em dB, referido a potência de 1 mW, medido em um ponto de nível relativo zero no circuito.

dBov:
Nível em dB relativo ao ponto de saturação *(overload)* de um dado sistema. Por exemplo, um sistema A/D de 16 bits tem seu ponto de 0dBov para a amplitude de 32768, i.e.,

$$0\text{dBov} = 20\log_{10}(32768).$$

dBPa:
Nível de pressão de um sinal acústico em relação a 1 μPascal.

dBr:
a potência relativa, em dB.

dBrnC:
o nível de ruído acima de -90 dBm (limiar de audição para ruído), ponderado pelo filtro C-MESSAGE (ANSI/IEEE Std.743-1984). Usado nos EUA, se relaciona ao dBmp pela relação dBmp=dBrnC-90;

dBSPL:
Abreviação geral para o Nível de Pressão Sonora (Sound Pressure Level), que na literatura de Comunicações se refere a -94dBPa (0dBSPL=-94dBPa).

dBV:
$20 \times log_{10}(volt)$, sendo 0dB relativo a 1V.

DCR:
Degradation Category Rating

EID:
Error Insertion Device, Dispositivo de inserção de erros. Ferramenta de software desenvolvida pelo UGST e presente na STL92.

ERP:
Ear Reference Point, Ponto de Referência do Ouvido. Um ponto localizado à entrada do ouvido externo do ouvinte.

ETSI:
European Telecommunications Standards Institute

Fônon:
O nível em dB de um tom de 1000Hz que seja julgado ter a mesma sonoridade de um outro sinal, potencialmente com diferente espectro de freqüência.

G.711:
Recomendação CCITT para a Codificação log-PCM a 64 kbit/s.

G.721:
Recomendação CCITT para a Codificação ADPCM a 32 kbit/s. Substituída pela G.726.

G.726:
Recomendação CCITT para a Codificação ADPCM a 40, 32, 24 e 16 kbit/s.

G.728:
Recomendação CCITT para a Codificação a 16 kbit/s usando o LD-CELP.

GSM:
Groupe Speciale Mobile

IRS:
Intermediate Reference System (CCITT Rec. P.48)

LD-CELP:
Low Delay, Code Excited Linear Prediction. Algoritmo de codificação de voz a 16 kbit/s padronizado pelo CCITT na Recomendação G.728.

Medidas Espectrais de Distorção:
São medidas de distorção objetiva baseadas na análise da representação espectral do sinal, como a distância cepstral.

Medidas Psicoacústicas de Distorção:
São medidas de distorção objetiva que levam em conta o processo de percepção da fala, modelando o sistema auditivo humano. As medidas, portanto, refletem objetivamente o sinal que é enviado ao córtex cerebral, após o pré-processamento pelo ouvido.

Medidas Temporais de Distorção:
São medidas de distorção objetiva baseadas na análise temporal da forma de onda de um sinal, como a relação sinal-ruído.

MNRU:
Modulated Noise Reference Unit (CCITT Rec. P.71)

MOS:
Mean Opinion Score

PCM:
Pulse Code Modulation (CCITT Rec. G.711)

PLL:
Preferred Listening Level, geralmente assumido como sendo -15dBPa no ponto de referência do ouvido *(Ear Reference Point, ERP)*.

Q:
Signal to Subjective Quantization Noise Equivalent Ratio

QDU:
Quantizing Distortion Unit, Unidade de Distorção de Quantização. Definida na Recomendação CCITT G.113, 1 *qdu* equivalente à distorção *média* (sic) de 1 conversão A/D–D/A segundo a lei de codificação na Recomendação G.711.

SINAD:
Sigla para "SIgnal, Noise, And Distortion", Sinal, Ruído e Distorção. É o recíproco da medida de distorção e é dado pela relação, em dB, do valor RMS do sinal adicionado a ruído mais distorção com o valor RMS só do ruído mais a distorção.

SNR, SNR a longo prazo:
No contexto de codificação digital de sinais de voz, é a razão entre a potência de um sinal e da potência da diferença entre esse sinal e uma versão sua distorcida, calculado *globalmente* para o sinal, isto é, para o sinal como um todo.

SNR segmentada:
No contexto de codificação digital de sinais de voz, é a razão entre a potência de um sinal e potência da diferença entre esse sinal e uma versão sua distorcida, calculada *localmente* para o sinal, isto é, para segmentos do sinal, tipicamente com 20ms de duração. Após o cálculo da SNR segmentada para todos os segmentos do sinal, uma figura de mérito é composta pela média aritmética das SNR calculadas para todos os segmentos.

Sônon:
Por definição, 1 *sônon* é o aumento de potência que faz dobrar a sonoridade percebida pelo usuário.

SQEG:
Speech Quality Experts Group, Grupo de Especialistas em Qualidade de Voz da Comissão de Estudos XII, responsável pela definição e revisão de Recomendações da Série P e do Manual de Telefonometria, além da elaboração e condução de testes subjetivos para processos de definição de padrões do CCITT.

STL, STL92:
Software Tool Library, Biblioteca de Ferramentas de Software do CCITT, que é referida como STL92 em sua versão publicada no ano de 1992.

SV6:
Voltímetro de voz implementado em hardware, desenvolvido pela British Telecom e comercializado pela Malden Electronics. Segue a recomendação CCITT P.56.

UGST:
Users Group on Software Tools, Grupo de Usuários de Ferramentas de Software; grupo criado dentro da Comissão de Estudos XV do CCITT para definir ferramentas de software a serem utilizadas no processo de geração de padrões (Speech Voltmeter, MNRU, etc).

Capítulo 2

Algoritmos de Codificação de Voz

2.1 Introdução

A introdução de técnicas de codificação digital de voz foi possível somente com o advento de circuitos integrados digitais de baixo custo e visava principalmente aumentar a capacidade de transmissão da rede telefônica então instalada.

No final da década de 60 concluíram-se os trabalhos do PCM logarítmico (log-PCM) a 64 kbit/s. Este foi o primeiro padrão internacional de codificação digital de voz do CCITT, sendo a Recomendação G.711 publicada em sua primeira versão no final de 1972.

Em 1984 foi publicada pelo CCITT a primeira versão do segundo padrão internacional de codificação digital de voz, o ADPCM a 32 kbit/s, que permitiu duplicar a capacidade de dos canais de comunicação. Ela já utilizava um bom nível de processamento digital e sua implementação foi viabilizada pelo avanço das técnicas digitais em silício.

Desde então, mais e mais técnicas digitais vêm sendo introduzidas visando diminuir ainda mais as taxas de bit dos sinais de voz. Adicionalmente, a digitalização de sinais vem sendo introduzida gradualmente tanto nas redes públicas como nas redes privadas (estas não restritas a algoritmos padronizados pelo CCITT), objetivando atingir, no futuro, uma rede completamente digitalizada.

Nas seções a seguir são descritos em seus aspectos básicos diversos esquemas de codificação de voz, padronizados pelo CCITT.

2.2 G.711: O algoritmo log-PCM do CCITT a 64 kbit/s

A codificação de sinais analógicos para transmissão em sistemas digitais foi introduzida a nível regional no começo dos anos 60. A taxa de amostragem escolhida foi de 8kHz (para abranger a faixa de telefonia, de 300 a 3400Hz) e, com uma codificação de 8 bits por amostra, resulta numa taxa de 64 kbit/s. Para a transmissão de um sinal digital através de símbolos binários, é necessária uma faixa mínima pelo menos 8 vezes maior que a faixa necessária para a transmissão do sinal sob a forma analógica (i.e., 32kHz contra 4kHz). O ímpeto para essa mudança de tecnologia foi a promessa de uma redução nos custos dos equipamentos

de transmissão e comutação muito maiores que os aumentos decorrentes do uso de uma faixa maior. De fato, a economia resultante, adicionada à integração entre comunicações e computação (o que resultou numa maior eficiência nas rotinas de Operação e Manutenção) foram tão grandes que, praticamente, novos equipamentos analógicos não têm mais sido instalados [15, Cap.6].

Em 1972, o CCITT publicou a Recomendação G.711, que constitui a padrão de referência para praticamente todos os sistemas de transmissão digital [16].

O princípio básico desse algoritmo é o de amostrar o sinal de voz mantendo a faixa de telefonia (300–3400 Hz). Para isso, usa uma taxa de amostragem de 8 kHz e codifica as amostras com 8 bits. Com essa combinação de número de bits e freqüência de amostragem, são formados feixes de bits de 64 kbit/s por canal de voz.

Descrição do algoritmo

O esquema mais natural de quantização é a quantização linear. No entanto, uma característica básica desta abordagem é que a relação sinal-ruído (SNR) varia com a amplitude dos sinais de entrada: quanto menor a amplitude dos sinais, menor a SNR [17]. Portanto, para um sinal não estacionário com grande variância, a SNR variará significantemente ao longo do tempo[1]. Por outro lado, utilizando-se quantização logarítmica (i.e., associação de compressão logarítmica e de quantização linear), pode-se obter uma relação sinal-ruído independente do nível do sinal. Do ponto de vista da qualidade percebida pelo usuário, um sistema utilizando quantização linear teria uma qualidade (SNR) que variaria significantemente com a amplitude do sinal de entrada, enquanto que outro com quantização logarítmica apresentaria uma qualidade (SNR) praticamente independente do nível do sinal de entrada, sendo, portanto, mais "estável".

Com isso em mente, diversos estudos foram conduzidos no final da década de 60 para se escolher um bom algoritmo de quantização logarítmica. Isto levou à definição de dois esquemas de compressão, um usando a lei de compressão μ:

$$c(x) = x_{max}\frac{\ln(1+\mu|x|/x_{max})}{\ln(1+\mu)}sgn(x)$$

e outro usando a lei A:

$$c(x) = \begin{cases} \dfrac{A|x|}{1+\ln(A)}sgn(x), & \text{para } 0 \le \dfrac{|x|}{x_{max}} \le \dfrac{1}{A} \\ x_{max}\dfrac{1+\ln(A|x|/x_{max})}{1+\ln(A)}sgn(x), & \text{para } \dfrac{1}{A} \le \dfrac{|x|}{x_{max}} \le 1 \end{cases}$$

Ambas as leis se comportam como a quantização linear para sinais de pequena amplitude, mas são verdadeiramente logarítmicas para sinais de grande amplitude. De fato, para sinais de grande amplitude, a SNR é dada por [17]:

$$SNR_{\mu} \approx 6.02B + 4.77 - 20\log_{10}(\ln(1+\mu))$$

[1] Isto acontece com o sinal de voz, que é um sinal não estacionário para períodos de tempo acima de algumas dezenas de milisegundos, sua potência variando em função dos fonemas falados, da distância entre o locutor e o transdutor, do volume da voz deste locutor, etc.

e

$$SNR_A \approx 6.02B + 4.77 - 20\log_{10}(1 + \ln A)$$

onde B é o número de bits usado para a quantização.

O CCITT escolheu os valores $A = 87.56$ e $\mu = 255$ para o padrão G.711, junto com $B = 8$. Isso simplifica as duas equações acima para:

$$SNR_\mu = 6.02B - 9.99 = 38.17dB$$

e

$$SNR_A = 6.02B - 10.1 = 38.06dB$$

De fato, a G.711 não especifica as leis como definido acima, mas antes usa uma aproximação linear-por-partes da característica de compressão. Isto se deve a uma maior facilidade de implementação em hardware, bem como a outras propriedades (veja [17, p.229]).

Esta aproximação usa 1 bit para o sinal (1 para positivo, 0 para negativo), usa os bits 2 a 4 para indicar o segmento (ou expoente) e os bits 5 a 8 para o nível (ou mantissa)[2]. Dentro de cada segmento, a quantização é linear (4 bits, ou 16 níveis, por segmento), havendo 15 segmentos de inclinações distintas para a lei μ e 13 para a lei A.

A lei A apresenta uma característica de quantização do tipo *mid-riser* (i.e., não existe o valor codificado zero), enquanto que a da lei μ é do tipo *mid-tread* (apresentando o valor decodificado 0^+ e 0^-) [18, pp.181-182]. Quanto à faixa das amplitudes de entrada, a lei A trabalha com sinais na faixa de –4096 a 4096, implicando numa representação com 13 bits. Para a lei μ, sinais no formato linear são aceitos na faixa de –8159 a 8159, o que pode ser representado por 14 bits. Apesar disto, a faixa dinâmica[3] para as leis A e μ equivale, respectivamente, à faixa dinâmica resultante de uma conversão linear de 12 e 13 bits [17, pg.234].

Um detalhe para a lei A é que os bits pares são invertidos. Isso foi introduzido para resolver problemas de recuperação de relógio decorrentes de longas seqüências de zeros, antes da introdução do código de linha HDB3. A razão disso vem do fato da distribuição de amplitudes para o sinal de voz ser aproximadamente Laplaciana [18] (i.e., fortemente concentrada em sinais de pequena amplitude), combinado com o fato da lei A usar palavras com a maioria de bits em '0' para codificar sinais de pequena amplitude (ao contrário da lei μ, que utiliza bits a sua maioria em '1' para pequenos sinais), como pode ser visto na coluna 6 das tabelas 1(a) e 1(b) da G.711. Como resultado, a população de bits '0' para a lei A é alta, bem como a probabilidade de ocorrência de longas seqüências de '0'. Com a inversão de bits, diminui a probabilidade de ocorrência de longas seqüências de bits '0', facilitando assim a recuperação de relógio com o código então usado (AMI).

A regra de conversão entre os formatos lei A ou μ e PCM linear é descrita em termos de tabelas na G.711. Dois modos são possíveis para sua implementação: algorítmico ou busca-a-tabela. Para implementação em silício, a segunda é mais vantajosa por ser de mais fácil implementação, ao custo de uma maior área. Para outras aplicações, como implementações

[2] Note que a numeração dos bits dentro do byte está na ordem reversa da notação normalmente utilizada em computação, sendo que o bit 1 da G.711 corresponde ao bit 7 (mais significativo) da representação usual e o bit 8 da G.711 ao bit 0 (menos significativo) da representação usual.

[3] Aqui definida como sendo a máxima amplitude dividida pelo menor passo de quantização [17, pg.234].

embutidas em DSPs ou implementações em software, busca-a-tabela ocuparia muita memória; portanto, nestes casos a implementação algorítmica é preferida.

A implementação (algorítmica) feita para o UGST (publicada pelo CCITT) é descrita em [11, Cap.5].

2.3 G.721: O algoritmo ADPCM a 32 kbit/s do CCITT

No meio de 1982, formou-se um grupo dentro da Comissão de Estudos XVIII do CCITT para estudar a padronização de uma técnica de codificação de voz que pudesse reduzir à metade a taxa de codificação dos sinais de voz, então em 64 kbit/s por canal (Rec.CCITT G.711), porém sem degradação da qualidade.

Concluiu-se que a técnica ADPCM poderia ser utilizada para se definir um codificador de boa qualidade, após muitas contribuições recebidas de diversas organizações. Este processo tomou 18 meses de desenvolvimentos e de testes objetivos (dados e sinalização) e subjetivos (voz), culminando com a geração de uma recomendação pelo CCITT, publicada em Outubro de 1984 na Série Vermelha como Rec.G.721 [19, 20].

Entretanto, no Período de Estudos de 1985–1988 descobriram-se problemas com dados FSK, o que implicou em mudanças no algoritmo. Estas mudanças foram aprovadas e publicadas em 1989 na Série Azul de Recomendações do CCITT, com mesmo número, mas substituindo a anterior (são incompatíveis) [21]. Ainda no mesmo Período de Estudos, identificou-se a necessidade de operação em outras taxas de bit para uso em sistemas de transmissão que exploram a estatística de atividade dos canais de voz (e.g. *Digital Circuit Multiplication Equipment, DCME* [22]). Assim, ao final do mesmo Período, publicou-se também a G.723 [23], com extensão do ADPCM para as taxas de 24 e 40 kbit/s.

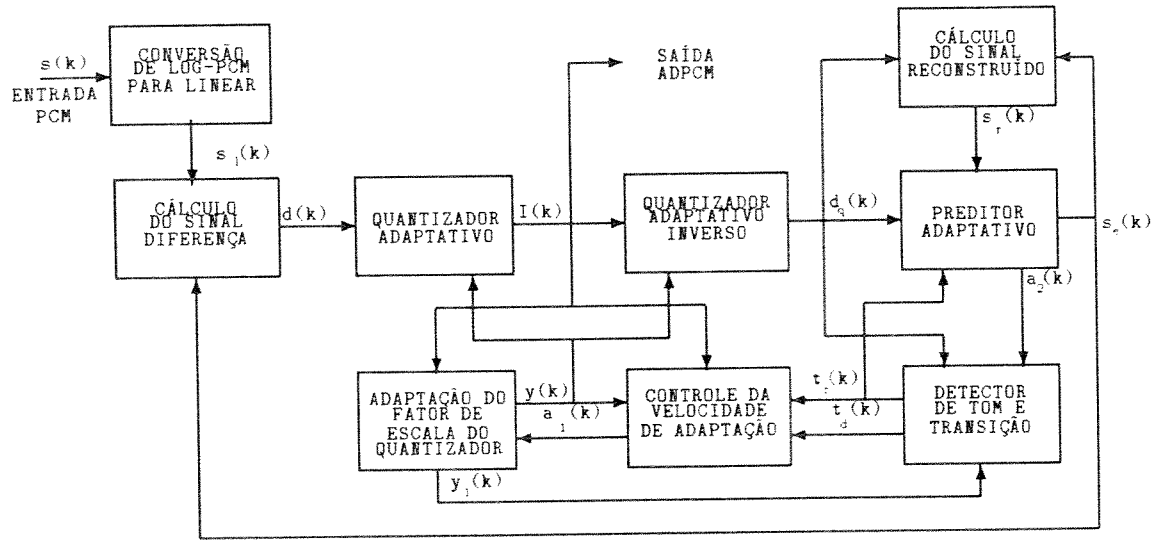
No último Período de Estudos (1989–1992), estas duas últimas recomendações foram fundidas em uma só (entretanto mantendo plena compatibilidade com os algoritmos anteriores), porém adicionando a taxa de 16 kbit/s para o ADPCM. Essa nova recomendação foi batizada de G.726 [24], que substitui as G.721 e G.723.

Apesar dessa mudança de numeração, os especialistas continuam a chamar de G.721 o algoritmo ADPCM do CCITT para codificação de voz a 32 kbit/s, apesar da referência oficial ser "G.726 operando a 32 kbit/s". Por simplicidade, neste trabalho continuaremos a nos referir ao ADPCM a 32 kbit/s como G.721.

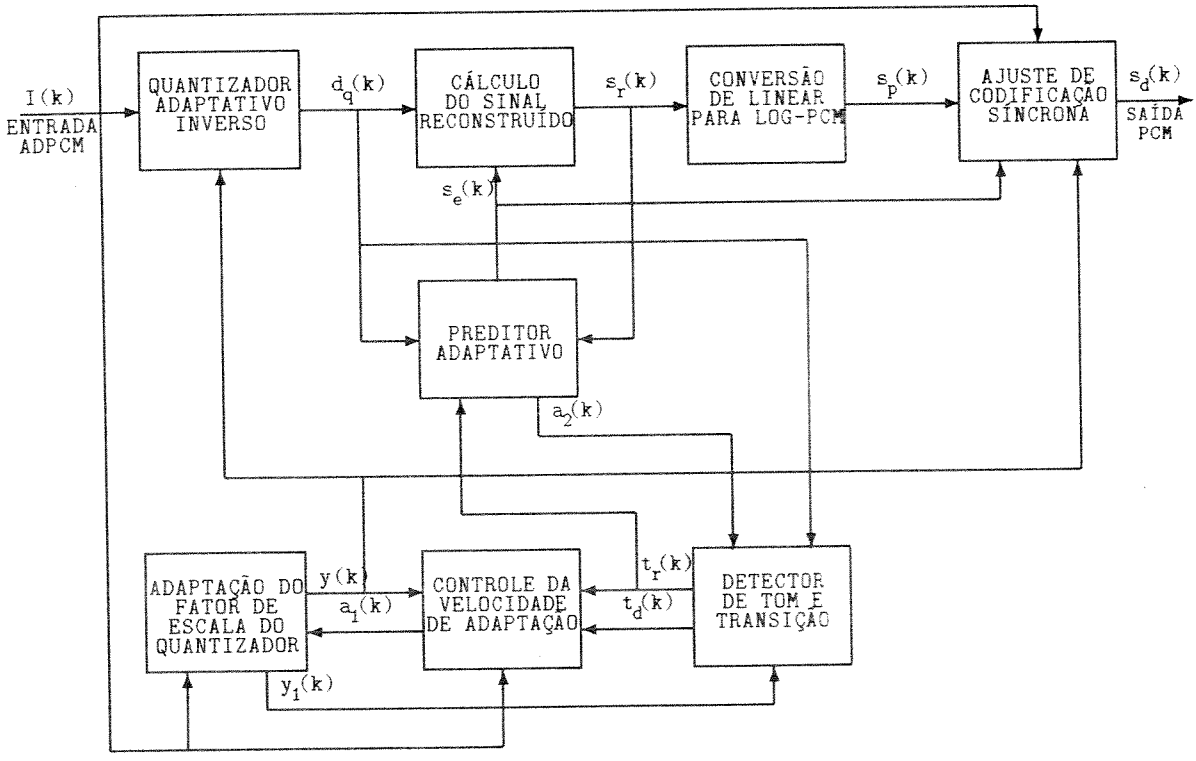
O algoritmo G.721 é descrito em [24] e seu diagrama em blocos está na figura 2.1. Outras análises do algoritmo podem também ser encontradas em diversos estudos [25, 20], apesar de se referirem à versão da Série Vermelha.

Visão geral do algoritmo

O codificador G.721 recebe sinais na faixa de voz já codificados no formato log-PCM da Recomendação G.711 (quer lei A ou μ) e gera em sua saída um sinal codificado com 4 bits por amostra. O decodificador faz o caminho contrário, retornando a 8 bits log-PCM as amostras ADPCM de 4 bits.



(a) Diagrama em blocos do codificador.



(b) Diagrama em blocos do decodificador.

Figura 2.1: Diagrama em blocos simplificado da G.721

Como uma descrição grosseira, para explorar a predizibilidade dos sinais de voz, um preditor adaptativo é usado para computar o sinal diferença $d(k)$ (baseado nas amostras $s_l(k)$), que é quantizado por um quantizador adaptativo utilizando 4 bits. Estes bits, enviados ao decodificador, são processados por um quantizador inverso e o sinal diferença é usado para se calcular o sinal reconstruído, $s_r(k)$, que é por sua vez comprimido (usando lei A ou μ) e exteriorizado pelo decodificador ($s_d(k)$).

Desta descrição, podem-se levantar alguns problemas:

- Se somente o sinal-diferença quantizado é transmitido, como pode o decodificador reconstruir o sinal?
- Como se pode assegurar a estabilidade do preditor?
- A redução em taxa não degrada muito a qualidade?

Estes e outros aspectos foram considerados no desenvolvimento do algoritmo, e muitos dos blocos do codificador foram adotados para garantir seu comportamento adequado. Por exemplo, o uso da técnica de adaptação retroativa *(backward)* do preditor e quantizador[4] permite que somente o uso do sinal-diferença quantizado garanta um "sincronismo" entre o codificador e o decodificador. Também, foram definidos fatores de fuga *(leak factors)* para garantir que o algoritmo sempre irá convergir, independentemente do estado inicial ou da ocorrência de erros de transmissão. Para instabilidades, alguns parâmetros tiveram sua excursão restringida. Assim, todos os pontos que possam parecer obscuros (para aqueles não familiarizados com o algoritmo) têm uma razão de ser, o que pode ser demonstrado por estudos cuidadosos [26]. A seguir é dada uma visão geral dos blocos que compõem o algoritmo.

Conversão do formato log-PCM para linear

O sinal de entrada $s(k)$, codificado em lei A ou μ, deve ser convertido em amostras lineares. Esta *expansão* é feita utilizando o algoritmo da G.711 [16], porém convertendo a representação das amostras de sinal-magnitude para complemento-de-dois com 14 bits.

Cálculo do sinal diferença

Este bloco simplesmente calcula a diferença entre o sinal de entrada (em formato linear) e o sinal estimado:

$$d(k) = s_l(k) - s_e(k)$$

Quantizador Adaptativo

Um quantizador adaptativo não-uniforme de 15 níveis é usado para quantizar o sinal diferença. Antes da quantização, esse sinal é convertido para uma representação logarítmica[5] e escalo-

[4] A técnica de adaptação retroativa *(backward)* tem como característica o uso do sinal quantizado (que é a saída do codificador) para a adaptação dos preditores e/ou quantizadores, ao invés do uso do próprio sinal. Isso implica que o sinal codificado e quantizado, além de ser transmitido para o decodificador, terá que ser realimentado para uma réplica do decodificador presente no próprio codificador e então utilizado para a adaptação dos parâmetros.

[5] Note que multiplicar amostras lineares equivale a somar seus logaritmos. Usando-se algoritmos eficientes para conversão entre os domínios logarítmico e linear (como implementado pela G.721), essa conversão torna-se vantajosa em termos de implementação.

nado pelo fator $y(k)$, que é computado no bloco "Adaptatação do Fator de Escala" (ver sua descrição mais à frente).

A saída deste bloco é $I(k)$, que tem duas funções: é a amostra ADPCM quantizada e é a entrada da parte retroativa *(backward)* do algoritmo ADPCM, que provê informação para a quantização das próximas amostras. Um ponto importante é que, como toda a computação na parte retroativa do algoritmo baseia-se na amostra quantizada, a saída do decodificador será idêntica ao sinal reconstruído calculado por esta parte do codificador (na ausência de erros de transmissão). É por isso que somente as amostras quantizadas são necessárias no decodificador, dispensando-se o envio de informação lateral *(side information)* para a adaptação do quantizador inverso.

Quantizador Adaptativo Inverso

O quantizador adaptativo inverso toma o sinal $I(K)$ e o converte de volta ao domínio linear, gerando uma *versão quantizada* do sinal diferença, $d_q(k)$. Esta é usada como entrada ao preditor adaptativo, de tal modo que o sinal estimado seja baseado numa versão quantizada do sinal diferença, ao invés de utilizar o sinal não-quantizado (original).

Adaptatação do Fator de Escala

Este bloco computa $y(k)$, o fator usado no quantizador adaptativo e no quantizador adaptativo inverso. Como entrada, este bloco precisa de $I(k)$, bem como de $a_l(k)$, que é o parâmetro de Controle da Velocidade de Adaptação (ver sub-seção abaixo). A razão deste último é que o algoritmo de escalamento tem dois modos de operação *(Adaptação Bimodal)*, lento e rápido. Isto foi feito para se acomodar diferentes tipos de sinal de entrada $s(k)$ que podem produzir sinais-diferença com flutuações amplas (e.g. voz) ou pequenas (e.g. tons ou dados na faixa de voz), considerados como sinais lentos e rápidos, respectivamente.

Este bloco produz dois fatores de escala (rápido, $y_u(k)$, e lento, $y_l(k)$), baseados em $I(k)$. De sua combinação usando $a_l(k)$ surge $y(k)$.

Controle da Velocidade de Adaptação

Este bloco calcula o parâmetro $a_l(k)$, que pode ser visto como uma indicação proporcional da velocidade do sinal de entrada e que assume valores entre 0 e 1. Se for 0, o sinal é considerado como tendo *variação lenta*; se 1, como tendo *variação rápida*.

Para se realizar isso, duas medidas na magnitude média de $I(k)$ são calculadas ($d_{ms}(k)$ e $d_{ml}(k)$). Estas, em conjunto com as sinalizações de Detecção Atrasada de Tom $(t_d(k))$ e Detecção de Transição $(t_r(k))$ (ver *Detecção de Transição de Tons*, a seguir), são usadas para calcular o parâmetro intermediário $a_p(k)$. O parâmetro de controle será $a_l(k) = \min(a_p(k-1), 1)$. O parâmetro $a_l(k)$ é naturalmente não-negativo, mas poderia assumir valores maiores que 1 se seu valor não fosse limitado a 1. Essa limitação provoca o atraso da transição do modo rápido para o lento até que a magnitude média fique constante por algum tempo. Com isso, evitam-se transições prematuras para sinais pulsados, como dados na faixa de voz com portadora comutada.

Uma análise de $a_p(k)$ fornece informações sobre a natureza do sinal de entrada: se seu valor estiver em torno de 2, a magnitude média de $I(k)$ está mudando, um tom foi detectado ou o canal está desocupado; por outro lado, se próximo a 0, a magnitude média de $I(k)$ estará praticamente constante.

Preditor Adaptativo e Cálculo do Sinal Reconstruído

O preditor adaptativo tem como função principal o cálculo do sinal estimado $s_e(k)$, baseado no sinal diferença quantizado, $d_q(k)$. O preditor tem dois pólos e seis zeros, estrutura que atende satisfatoriamente a toda a classe de sinais para os quais o algoritmo foi especificado. Com esses coeficientes e os valores passados de $d_q(k)$ e $s_e(k)$, calcula-se o valor do sinal estimado.

A atualização dos dois conjuntos de coeficientes ($a_i(k), i = 1..2$, para os pólos e $b_i(k), i = 1..6$, para os zeros) é feita com o algoritmo do gradiente simplificado (algoritmo do sinal). A faixa dinâmica dos coeficientes é limitada neste ponto para evitar instabilidades. Adicionalmente, se uma transição de sinais com banda parcial[6] é indicada (por $t_r(k)$), os coeficientes do preditor são zerados até que t_r volte a 0. Note que quando isso acontece, o quantizador é forçado ao modo rápido de adaptação.

O sinal reconstruído $s_r(k)$ é calculado usando o sinal estimado $s_e(k)$ e o sinal diferença quantizado, $d_q(k)$.

Detecção de Transição de Tons

Este bloco é uma das mudanças em relação à versão anterior (Livro Vermelho, 1985) da G.721. Ele foi adicionado para melhorar o desempenho do algoritmo para sinais gerados por modems FSK operando no modo-caracter e é implementado em dois passos. Inicialmente, verifica se o sinal tem banda parcial (e.g. tom) pelo exame do coeficiente $a_2(k)$ do preditor, definindo o sinal $t_d(k)$. Em seguida, a indicação de transição de sinal de banda parcial $t_r(k)$ é colocada em 1, de modo que os coeficientes do preditor possam ser colocados em 0 e o quantizador forçado para o modo rápido de adaptação.

Conversão do Formato PCM de Saída

Este bloco é exclusivo do decodificador. Sua única função é comprimir o sinal reconstruído $s_r(k)$, que está representado no formato PCM linear, usando a lei A ou μ. É complementar ao bloco de conversão do formato PCM de entrada, já descrito.

Ajuste de Codificação Síncrona

Este bloco também é exclusivo do decodificador. Ele foi projetado para impedir a acumulação de distorções que poderiam ocorrer em conexões tandem síncronas (por exemplo, uma conexão ADPCM–PCM–ADPCM, em que não ocorrem conversões intermediárias para o domínio analógico), desde que:

[6] O termo "sinais de banda parcial" *(partial band signals)*, usado na G.726, se refere a sinais que possuem freqüências somente em uma região específica dentro do espectro de 4kHz para telefonia. [27]. Assim, este termo genérico abrange tanto tons (e.g. sinalização) como sinais de dados modulados pelos mais diversos tipos de modens.

- as transmissão do ADPCM e PCM intermediários sejam livres de erro;
- os sinais ADPCM e PCM intermediários não sejam perturbados por outros dispositivos digitais de processamento de sinais.

A implementação em software da G.721 que foi adotada pelo UGST (publicada pelo CCITT) é descrita no Capítulo 6 da referência [11, Cap.6].

2.4 G.728: o algoritmo LD-CELP a 16 kbit/s do CCITT

Na evolução da padronização de algoritmos de codificação de voz, o CCITT estabeleceu em 1985 atividades de padronização de um algoritmo de codificação de voz a 16 kbit/s. Devido ao amplo espectro de suas potenciais aplicações, foram estabelecidos requisitos extremamente rígidos para seu desempenho [28], necessitando apresentar qualidade no mínimo igual à da G.721 para uma transcodificação, bem como atraso unidirecional inferior a 5 ms (preferencialmente menor que 2 ms), para evitar o uso de canceladores de eco.

O processo de identificação de Requisitos e Objetivos (Termos de Referência) para o codificador a 16 kbit/s começou em 85 e se prolongou até meados de 1990. Já o processo de identificação de algoritmos e de testes, após a formação de um grupo de especialistas no assunto em junho de 1988, começou em março de 1989, com dois candidatos potenciais: a BNR do Canadá [29] e a AT&T Americana [30]. Em 1989 a BNR desisitiu [31], porém aparecendo um outro candidato, do consórcio entre a Voicecraft, Universidade da Califórnia (UCSB) e a Universidade Simon Fraser (Canadá), que também desistiria logo depois [32, 33, 34, 35]. Assim, a AT&T prosseguiu como sendo a única proponente de algoritmo de codificação de voz a 16 kbit/s no processo de padronização conduzido pelo CCITT.

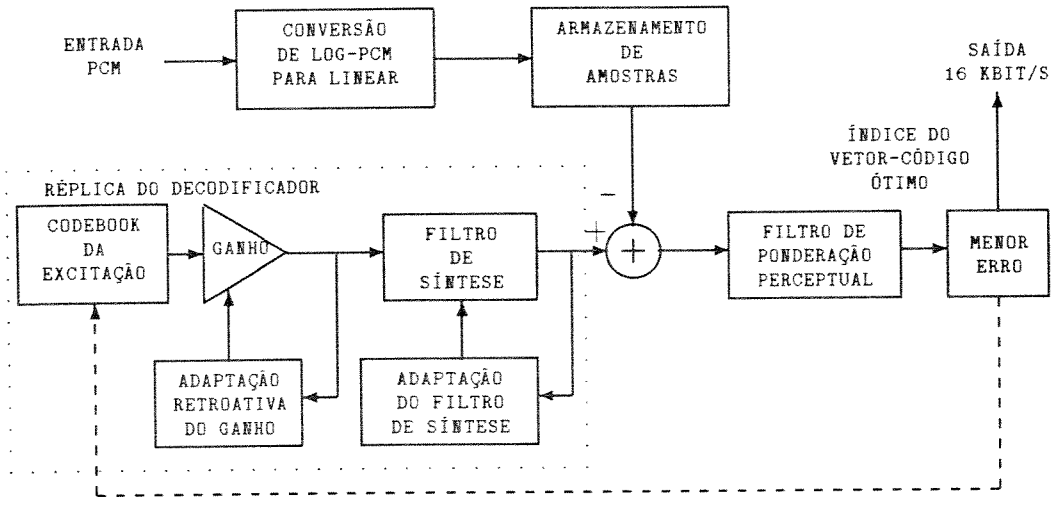
Seu algoritmo, o LD-CELP *(Low-Delay Code-Excited Linear Prediction)*, passou por duas fases de testes entre 1989 e 1992, sendo que oito organizações contribuíram para os trabalhos, inclusive o Brasil [36]. Após todos os testes e alguns refinamentos realizados entre as duas fases de testes para atender os requisitos de qualidade, o LD-CELP foi publicado em 1992 pelo CCITT como a Recomendação G.728 [37]. Seu diagrama em blocos simplificado encontra-se na figura 2.2.

2.4.1 Visão Geral do Algoritmo

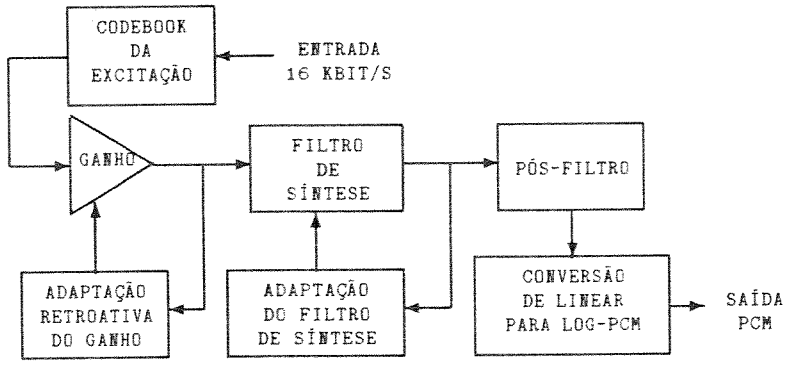
Características Gerais: O LD-CELP modificou a maioria das estruturas presentes em codificadores do tipo CELP [38] para atender ao requisitos de desempenho para o algoritmo, bem como para viabilizar sua implementação em tempo real. No entanto, manteve a estrutura básica dos codificadores CELP, que é a busca em um dicionário de códigos, ou *codebook*[7], usando análise-por-síntese *(analysis-by-synthesis)*.

Como tradicionalmente um sinal analógico é digitalizado em um feixe de 64 kbit/s com 8 bits por amostra (Rec.CCITT G.711), um feixe de dados a 16 kbit/s implica no uso de 2 bits por amostra. Como o LD-CELP utiliza um quadro básico de 5 amostras, isso implica que o *codebook* de quantização vetorial da excitação tenha 1024 vetores (representados por índices de 10 bits). Cada vetor-código é o produto de um ganho escalar de 3 bits e de um vetor de

[7] Manteremos aqui a palavra inglesa por ser o jargão de uso mais comum.



(a) Diagrama em blocos do codificador.



(b) Diagrama em blocos do decodificador.

Figura 2.2: Diagrama em blocos simplificado do LD-CELP

envoltória *(shape vector)* de dimensão 5 (representados por 7 bits). O ganho escalar possui 1 bit de sinal e 2 de magnitude, sendo simétrico em relação à origem; isto permite dobrar a faixa de amplitudes representada pelo *codebook* de envoltória sem contudo duplicar a complexidade da busca do vetor-código ótimo.

O *codebook* foi treinado com a mesma ponderação *perceptual*[8] implementada no codificador, o que leva em conta o efeito da adaptação do preditor e do ganho da excitação, apresentando um desempenho melhor do que a técnica de povoar o *codebook* com números aleatórios gaussianos, esta última sendo tradicionalmente usada em codificadores CELP [38].

Após a escolha do conteúdo do *codebook*, resta assinalar índices a esses valores (isto é, organizar os vetores-código dentro do *codebook*). Para garantir uma melhor SNR no decodificador para canais ruidosos, foi utilizada uma pseudo-codificação de Gray. Com isso, garante-se que um erro único desloque a palavra código no decodificador para uma outra bem próxima ao vetor-código originalmente transmitido, ao contrário do que aconteceria se os índices fossem distribuídos aleatoriamente.

Tipo de Especificação do Algoritmo: A especificação de um algoritmo pode ser feita em um dos três modos [39]: exata-em-bits *(bit-exact specification)*, de feixe de bits *(bitstream specification)* e exata do algoritmo *(algorithm exact specification)*. A especificação exata-em-bits implica que todas as variáveis e operações dentro do algoritmo têm seu comprimento em bits e tipo de representação completamente definidas; este é o tipo da especificação historicamente usado dentro do CCITT (G.721, G.722, G.723, G.726 e G.727). Uma outra abordagem é a de especificar somente o formato do feixe de bits entre codificador e decodificador; adicionalmente, provê-se um codificador e um decodificador de referência, mas que podem ser modificados para simplificar a implementação ou para melhorar o desempenho. Este tipo de especificação foi utilizado no passado em padrões de codificação segura de voz e mais recentemente nos padrões celulares norte-americano (VSELP) e japonês, bem como para a codificação de áudio e vídeo do padrão multi-meios do MPEG. Finalmente, a especificação exata do algoritmo implica em descrever detalhadamente todas as partes do algoritmo, sem entretanto especificar exatamente o comprimento das variáveis. Esta pode ainda ser definida em termos de uma aritmética de ponto fixo (somente variáveis inteiras) ou de ponto flutuante. Um exemplo desta abordagem é a G.728.

O algoritmo da G.728 é especificado em termos de operações em ponto flutuante. Por isso, é uma especificação que não é exata-em-bits *(bit-exact)*, mas que descreve precisamente a implementação [39]. Apesar de haver uma atividade dentro do CCITT para se especificar uma versão em ponto fixo (porém não *bit exact*), esta terá que manter plena interoperabilidade[9] com a versão já publicada. Uma implicação desta abordagem é a necessidade de se definir procedimentos de verificação de implementação e interoperabilidade bem mais sofisticados que os utilizados para os algoritmos com descrição exata-em-bits.

[8] O termo *perceptual*, apesar de não constar do vernáculo, constitui-se num jargão em amplo uso para designar os aspectos relacionados à maneira como um dado sinal é percebido pelo sistema auditivo humano. Por ser cotidiano, manteremos seu uso aqui.

[9] O conceito de interoperabilidade de duas implementações de um algoritmo surge em decorrência das especificações do tipo "*exata do algoritmo*". Ela se refere à necessidade de duas implementações diferentes terem de conversar entre si, apesar dos dispositivos utilizados serem diferentes. Por exemplo, um DSP em ponto flutuante e um DSP em ponto fixo, ou dois DSPs em ponto flutuante de diferentes fabricantes, em que a representação dos números apresente diferentes precisões. Essa pequenas diferenças podem provocar o acúmulo de erros que, ao longo do tempo, poderiam levar as duas implementações a não conversarem entre si (i.e., divergirem); elas, então, não seriam interoperáveis [39].

Historicamente, a escolha por uma implementação inicial em ponto flutuante foi permitida por haver DSPs disponíveis no mercado implementando operações em ponto flutuante. Porém, o fator determinante foi não haver certeza de se poder implementar, dentro do cronograma então definido e com a qualidade necessária, um algoritmo em ponto fixo (potencialmente exato-em-bits). Partiu-se então para esta abordagem de implementação, apesar de no final do processo ter-se admitido que a abordagem em ponto-fixo teria sido viável.

Atraso: O LD-CELP apresenta um atraso algorítmico[10] de 625 μs (quadro de 5 amostras) e um atraso (total) uni-direcional de menos de 2 ms. Para obter esse atraso, porém mantendo a qualidade necessária, optou-se por se utilizar a técnica de adaptação retroativa *(backward)* do preditor.

Adaptação Retroativa: Na técnica CELP tradicional, transmitem-se os parâmetros do preditor, o ganho e a excitação, enquanto que no LD-CELP, com a adaptação retroativa, apenas a excitação é transmitida. Para que isto funcione, a análise LPC é feita usando-se o sinal previamente quantizado; o ganho de excitação é atualizado a partir do valor do ganho embutido na amostra previamente quantizada.

No codificador do algoritmo LD-CELP existem três estruturas adaptadas retroativamente: o filtro de síntese LPC, o filtro de ponderação perceptual e a unidade de ganho da excitação. No decodificador se repetem estas três estruturas e se acrescenta a adaptação retroativa das estruturas do pós-filtro.

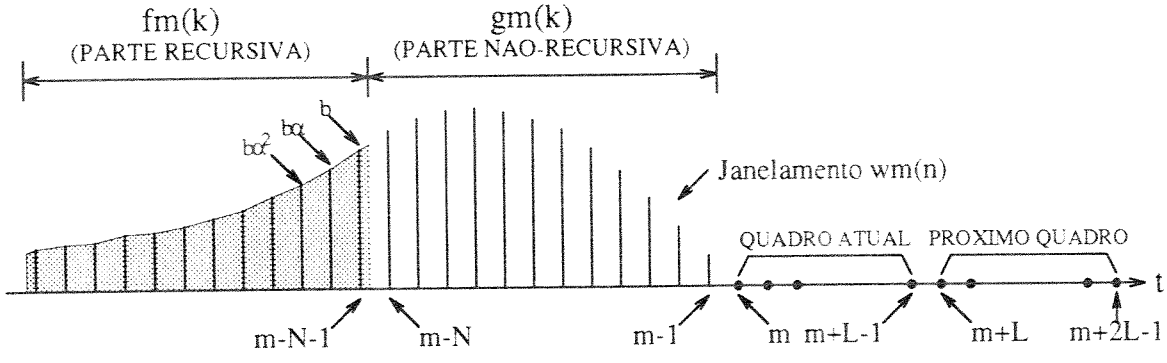


Figura 2.3: Janelamento Híbrido do LD-CELP

Janelamento usado nas adaptações : No LD-CELP, todas as estruturas adaptadas retroativamente (exceto o preditor de longo prazo no pós-filtro do decodificador) utilizam a análise LPC. O método utilizado para o cálculo dos coeficientes do preditor LPC é o da autocorrelação

[10] O atraso total introduzido por um codificador pode ser definido como o atraso intrínseco ao algoritmo de codificação (e.g. o número de amostras que ele necessita armazenar *antes* de poder começar a processar as amostras) somado ao tempo necessário para que o hardware que implementa o algoritmo consiga processar a amostra ou o quadro de amostras. No caso do ADPCM da G.721, o atraso algorítmico é zero, porém o tempo necessário ao processamento da amostra é não nulo e inferior a 125 μs; isto resulta num atraso total de 1 amostra, ou 125 μs. Obviamente, o tempo de processamento é altamente dependente da implementação, não podendo ser determinado a priori.

[18]. Neste método, faz-se necessário o uso de janelamento sobre as amostras para aumentar o ganho (e portanto, a eficiência) de predição.

Normalmente se utiliza o janelamento de Hamming, mas este não é adequado à estrutura do LD-CELP. O bloco de análise do LD-CELP é de apenas 5 amostras, em contraste com blocos de 160 a 256 amostras usados pelo janelamento de Hamming. Assim, usar o janelamento de Hamming implicaria numa sobreposição significativa de janelas e numa alta complexidade computacional [40]. Percebeu-se que no contexto retroativo do LD-CELP, a técnica de janelamento recursivo de Barnwell [41] resultava em ganhos de predição maiores que com o janelamento de Hamming, bem como numa maior qualidade subjetiva da voz processada. Por isso, a versão do LD-CELP testada na Fase 1 utilizou uma versão modificada do janelamento recursivo [42, 43]. Assim, conseguiu-se melhor qualidade, uma carga computacional mais balanceada, menor consumo de memória e menor complexidade computacional para atualizações freqüentes do preditor [44].

Porém, visando uma futura implementação em ponto fixo com mínimas mudanças com relação à versão em ponto flutuante (posto que necessitam ser *interoperáveis*), desenvolveu-se uma técnica de janelamento híbrido [40, 37], com resultados equivalentes à da versão de Barnwell[11] e que foi implementada na versão testada na Fase 2 de testes subjetivos e hoje permanece na Recomendação G.728 (ver figura 2.3). Com ela conseguiu-se manter a mesma qualidade (pois o formato da janela híbrida é o mesmo que da recursiva) e diminuir a complexidade computacional entre 20% e 30%. O objetivo foi reduzir a complexidade através da mistura de uma parte recursiva, como a de Barnwell, com uma não-recursiva, porém o conjunto resultando numa curva similar à de Barnwell. Foi então implementado um janelamento em que as m amostras passadas mais recentes são sobrepostas de maneira não-recursiva. Já as amostras anteriores à m-ésima amostra são incluídas no janelamento por um algoritmo recursivo. A característica dessa curva é ser um seno na parte não recursiva e uma exponencial decrescente na porção recursiva, de modo que amostras passadas mais antigas têm cada vez menor influência.

Em termos de equações [37, 42], a janela híbrida $w_m(k)$ é definida por:

$$w_m(k) = \begin{cases} f_m(k), & k < m - N \\ g_m(k), & m - N \leq k < m \\ 0, & k \geq m \end{cases}$$

onde

$$\begin{aligned} f_m(k) &= b a^{-[k-(m-N-1)]} \\ g_m(k) &= -\sin[c(k-m)] \end{aligned}$$

e $0 < a < 1$, $0 < b < 1$ e c são constantes definidas distintamente para cada uma das estruturas adaptadas retroativamente no algoritmo.

No método da autocorrelação, o i-ésimo coeficiente de autocorrelação $R_m(i)$ do sinal de

[11] A perda no ganho de predição ficou abaixo de 0.1 dB, um valor muito pequeno.

entrada $s(n)$ é calculado como:

$$\begin{aligned} R_m(i) &= \sum_{k=-\infty}^{\infty} s(k)w_m(k)s(k-i)w_m(k-i) \\ &= \sum_{k=-\infty}^{m-1} s(k)w_m(k)s(k-i)w_m(k-i) \\ &= r_m^{\mathcal{R}}(i) + r_m^{\mathcal{N}}(i) \end{aligned}$$

Onde $r_m^{\mathcal{R}}(i)$ e $r_m^{\mathcal{N}}(i)$ são respectivamente as componentes recursiva e não-recursiva do i-ésimo coeficiente de autocorrelação, descritos por:

$$r_m^{\mathcal{R}}(i) = \sum_{k=-\infty}^{m-N-1} s(k)s(k-i)f_m(k)f_m(k-i)$$

e

$$r_m^{\mathcal{N}}(i) = \sum_{k=m-N}^{m-1} s(k)s(k-i)g_m(k)g_m(k-i)$$

Consideremos agora um preditor de ordem M com um ciclo de adaptação (atualização) dos coeficientes de L amostras. Então, o i-ésimo coeficiente de autocorrelação do próximo ciclo de adaptação será:

$$R_{m+L}(i) = r_{m+L}^{\mathcal{R}}(i) + r_{m+L}^{\mathcal{N}}(i)$$

com

$$r_{m+L}^{\mathcal{R}}(i) = \alpha^{2L} r_m^{\mathcal{R}}(i) + \sum_{k=m-N}^{m+L-N-1} s(k)s(k-i)f_{m+L}(k)f_{m+L}(k-i)$$

e

$$r_{m+L}^{\mathcal{N}}(i) = \sum_{k=m+L-N}^{m+L-1} s(k)s(k-i)g_{m+L}(k)g_{m+L}(k-i)$$

Portanto, vê-se que $r_{m+L}^{\mathcal{R}}(i)$ é calculado *recursivamente* a partir de seu valor $r_m^{\mathcal{R}}(i)$ do ciclo de adaptação anterior e que os coeficientes de autocorrelação têm sempre uma componente recursiva e uma não-recursiva.

Correção de Ruído Branco: O uso do método da autocorrelação para o cálculo dos coeficientes de predição traz embutido nele a inversão da matriz de autocorrelação, apesar de isso não ser feito de maneira explícita em métodos como o de Levinson-Durbin, onde os coeficientes de predição (e coeficientes de reflexão) são obtidos iterativamente.

Entretanto, são operações equivalentes e, se a matriz de autocorrelação for mal-condicionada[12], os coeficientes LPC calculados pelo método de Levinson-Durbin poderão levar a um filtro instável. Este efeito será ainda mais pronunciado quando filtros LPC de alta ordem são utilizados, como será descrito mais à frente. Uma técnica simples para diminuir o mal-condicionamento da matriz de autocorrelação consiste em adicionar ruído branco ao sinal sobre o qual realizar-se-á a predição, pois isso preencheria os vales do espectro LPC com ruído

[12] A matriz de autocorrelação pode vir a ser mal-condicionada porque a representação dos sinais tem precisão numérica finita.

e diminuiria a faixa dinâmica do espectro. Do ponto de vista computacional, entretanto, é interessante adotar um procedimento equivalente, denominado na G.728 como "correção de ruído branco".

Seja um sinal de voz $s(k)$ ao qual se associa uma função de autocorrelação $R_s(i), i = 1..M$, onde M é a ordem do filtro de predição. Ou seja, de maneira sintética:

$$s(k) \longleftrightarrow R_s(i)$$

Analogamente, a um ruído branco $n(k)$ associa-se uma função de autocorrelação $R_n(i)$:

$$n(k) \longleftrightarrow R_n(i)$$

No caso de haver adição desse ruído ao sinal, teremos a função de autocorrelação $R(i)$ resultante:

$$s(k) + G.n(k) \longleftrightarrow R(i) = R_s(i) + G.R_n(i)$$

onde G é uma constante de ganho denominada *"fator de correção de ruído branco"*. A relação sinal-ruído é dada em função de G por:

$$SNR_{dB} = 20 \log_{10} \left(\frac{1}{G} \right)$$

Como $n(k)$ é um ruído branco, a sua função de autocorrelação será dada por

$$R_n(i) = \begin{cases} 1, & \text{se i=0} \\ 0, & \text{c.c.} \end{cases}$$

Consequentemente:

$$R(i) = \begin{cases} R_s(0) + G, & \text{se i=0} \\ R_s(i), & \text{c.c.} \end{cases}$$

Deste modo, vê-se que basta acrescentar um certo valor G ao coeficiente $R_s(0)$ para diminuir o mal-condicionamento da matriz de autocorrelação. O valor de G é definido pelo nível de ruído que se deseja "adicionar" ao sinal.

No LD-CELP, esta técnica é utilizada para "adicionar" ruído aproximadamente 24dB abaixo do nível do sinal (G=1/256) e diminuir o mal-condicionamento da matriz de autocorrelação do filtro de síntese LPC e do filtro de ponderação perceptual, sem aumento da carga computacional do algoritmo.

Alargamento espectral: Uma técnica muito comum nos codificadores CELP consiste em atenuar o pico das formantes através da expansão da largura de faixa *(bandwidth expansion)* do espectro LPC[13]:

$$a_i = \lambda^i \hat{a}_i$$

onde $\hat{a}_i$ é o i-ésimo coeficiente calculado pelo preditor e λ é uma constante que satisfaz $\lambda < 1$ e $\lambda \approx 1$.

O efeito disso é fazer com que os pólos do preditor se afastem do círculo de raio unitário. Adicionalmente, a resposta impulsiva do modelo fica mais curta, diminuindo a propagação de erros de transmissão dentro do mecanismo de adaptação.

[13] Formantes de faixa muito estreita causam auditivamente um efeito de assobio (shirp), diminuindo portanto a qualidade subjetiva do sinal. Por outro lado, o alargamento da largura de faixa do filtro de síntese LPC pode diminuir o desempenho para sinais de dados, exigindo portanto um compromisso para sinais de voz e não-voz.

Formatos de Entrada e Saída: Como o LD-CELP é um algoritmo especificado para ser usado também na rede pública comutada, a representação do sinal de voz na entrada do codificador e na saída do decodificador deve seguir a Recomendação CCITT G.711 (em lei A ou μ)[14].

No entanto, as operações internas do algoritmo são realizadas no domínio linear (i.e., formato PCM linear). Portanto, o primeiro bloco do codificador na figura 2.2(a) consiste na expansão das amostras log-PCM, i.e., a conversão do formato log-PCM para o formato linear, para se iniciar o processamento das amostras. Já o último bloco do decodificador na figura 2.2(b), após o término dos processamentos, realiza a compressão das amostras lineares, i.e., a conversão do formato linear para o log-PCM.

2.4.2 Estruturas do Codificador

A seguir são descritos os blocos que compõem a estrutura do codificador do LD-CELP, ilustrado na figura 2.2(a).

Filtro de Síntese LPC: O filtro de síntese tem como função gerar o sinal quantizado a partir do vetor de excitação desnormalizado. É baseado em coeficientes LPC e possui ordem 50. A seguir, explica-se a razão desta alta ordem.

Uma técnica muito comum em codificadores de voz, em especial nos codificadores CELP, é o uso de predição de longo prazo *(long-term prediction)*, ou preditor de pitch adaptativo, com transmissão do parâmetro *(forward-adaptive pitch prediction)*. A predição LPC convencional é normalmente baseada num pequeno número de coeficientes LPC (de 10 a 12 [18, pp.419–420]), o que não permite efetuar uma predição de longo prazo que leve em conta o período de pitch. O uso de preditores de pitch é necessário para tornar mais branca a excitação obtida após a análise LPC convencional, explorando melhor a predizibilidade do sinal de voz. No entanto, como o LD-CELP transmite apenas a informação da excitação, a predição de pitch precisaria também ser retroativa, como acontece em [29, 34]. Esta técnica, porém, é muito sensível a erros de transmissão devido à alta ordem da filtragem, fazendo com que erros persistam (ou se propaguem) por um grande número de amostras. O uso de reinicializações artificiais poderia resolver o problema, mas como o algoritmo deve funcionar em altas taxas de erro (e.g. 10^{-2}, o que equivale a 160 erros por segundo), isso não funcionaria para o LD-CELP [42].

Aliado ao problema de propagação de erros de transmissão, percebeu-se que a melhoria introduzida pelo preditor de pitch para voz feminina era bem maior que para voz masculina[15] [42].

Assim, com os problemas do preditor de pitch retroativo com erros de transmissão e a maior importância da predição de pitch para vozes femininas, resolveu-se explorar a predizibilidade para vozes femininas usando-se um filtro de síntese com um preditor de alta ordem de modo que a maioria dos valores de período de pitch para vozes femininas fossem contemplados.

[14] Para as rotinas de verificação de implementação e de interoperabilidade [37, Apêndice 1], entretanto, o formato do sinal de entrada deverá ser linear. Portanto, deve-se evitar neste caso a passagem pelos blocos de expansão e de compressão do codec.

[15] Isso se explica por dois fatores. Primeiro, porque, sendo os períodos de pitch para vozes masculinas *bem maiores* que os para vozes femininas, o ganho de predição resulta maior para estas últimas. Além disso, a adaptação retroativa do preditor de pitch utiliza o sinal de erro *quantizado* como entrada, ao invés do sinal de erro original. O ruído de quantização resultante faz com que a correlação dentro de um período de pitch, que já é fraca para vozes masculinas, diminua ainda mais. Consequentemente, o ganho de predição fica ainda menor para as vozes masculinas [45].

Encontrou-se então experimentalmente que uma ordem 50 para o preditor LPC do filtro de síntese gerava bons resultados[16], substituindo com sucesso o conjunto "preditor LPC de baixa ordem e preditor de pitch" tradicionalmente empregado nos codificadores CELP.

Filtro de ponderação perceptual: A forma geral do filtro de ponderação perceptual é [46]:

$$W(z) = \frac{1 - Q(z/\gamma_1)}{1 - Q(z/\gamma_2)}, 0 < \gamma_2 < \gamma_1 \leq 1.$$

onde

$$Q(z/\gamma_j) = \sum_{i=1}^{M} \gamma_j^i q_i z^{-i}, j = 1, 2.$$

e q_i são os coeficientes LPC quantizados e M é a ordem do preditor LPC.

A sua função é moldar a envoltória espectral do sinal de erro, de modo que ela fique similar ao espectro do sinal de voz de entrada, daí mascarando a distorção que, sem essa ponderação, poderia ser percebida pelo usuário. Ao utilizar o erro ponderado perceptualmente na escolha do melhor vetor-código, este vetor-código será aquele que fornecerá, na recepção, o menor ruído de quantização *percebido pelo usuário*, aumentando (no decodificador) a qualidade subjetiva [47].

Para codificadores CELP em geral usa-se $(\gamma_1, \gamma_2) = (1.0, 0.8)$. No LD-CELP da Fase 1 e da Fase 2 usou-se $(\gamma_1, \gamma_2) = (0.9, 0.4)$ e $(\gamma_1, \gamma_2) = (0.9, 0.6)$, respectivamente, o que resultou num menor nível de ruído percebido. A mudança do coeficiente γ_2 foi devida à necessidade de melhoria do desempenho do codec da Fase 1 para 3 transcodificações em cascata, em que o codec apresentou uma qualidade muito abaixo da necessária [48].

A ordem M do preditor foi escolhida como 10 para evitar *artefatos*[17] que surgiram com o uso de ordens maiores (e.g. 50, como a do filtro de síntese). Para compensar a ordem baixa, ao invés de se utilizar um sub-conjunto dos coeficientes do preditor principal (que a princípio poderiam ser os mesmos), utilizou-se um preditor em separado operando sobre o sinal de entrada (lembre-se de que o preditor principal usa as amostras *quantizadas* em sua adaptação). Isto justifica-se por dois aspectos: como a ponderação perceptual não é necessária no decodificador, nada obriga o uso do sinal quantizado para o filtro perceptual; como justificativa principal, o uso do sinal sem quantização permite a obtenção de uma envoltória espectral mais precisa.

Busca do vetor-código de excitação ótimo: Para cada vetor de amostras s(n), é pesquisado qual dos 1024 vetores-código gera menor erro quadrático médio (ponderado perceptualmente). O vetor-código selecionado tem o seu índice transmitido para o decodificador e é realimentado para a parte adaptativa do codificador (réplica do decodificador no codificador), para ser utilizado como excitação do filtro de síntese. Na implementação do LD-CELP, foi utilizado um algoritmo computacionalmente eficiente para realizar a busca acima [49, 50], descrito em detalhes no corpo da Recomendação G.728 [37, pp.10–13].

[16] Adicionalmente, o preditor LPC retroativo de ordem 50 mostrou-se bem mais robusto a erros de transmissão que o preditor de pitch retroativo.

[17] *Artefatos* se referem aqui a instabilidades com sinais quase-periódicos de longa duração (maiores que 2 a 3 segundos), como acontece com voz artificial (Rec.CCITT P.50), certos tipos de passagens musicais (e.g. som sustentado de um violino) ou sons vocálicos sustentados (e.g. /a/). Essas instabilidades se reproduzem como distorção no timbre do som ouvido, o que dimunui a qualidade subjetiva do algoritmo [45].

Desnormalização da Excitação Quantizada: Para aumentar a eficiência da codificação, os 1024 vetores-código do *codebook* representam excitações típicas (obtidas pelo treinamento desse *codebook* a partir de vasto material de voz), cuja amplitude foi normalizada. Portanto, o vetor de excitação a ser utilizado pelo filtro de síntese precisa ser desnormalizado, o que é realizado pela unidade de desnormalização (ganho) da excitação. O mecanismo de adaptação deste ganho é descrito mais à frente.

Adaptação do Ganho da Excitação: O índice-ótimo transmitido para o decodificador representa o valor normalizado do vetor de excitação encontrado para o sinal de entrada. Portanto, para a reconstrução do sinal (tanto no decodificador como no codificador), é necessário se calcular o valor do ganho *(ganho da excitação)* a ser usado para a desnormalização da excitação. Isso pode ser feito em se utilizando um valor fixo, pré-determinado através da análise estatística de longo prazo dos valores possíveis para esse ganho (à semelhança de preditores ou quantizadores com coeficientes ou limiares fixos), utilizando-se vasto material de voz. Uma outra abordagem, que permite se encontrar fatores de normalização mais otimizados, é a de se utilizar um sistema adaptativo de cálculo do ganho. No LD-CELP, o cálculo do ganho é feito através de uma técnica mista, que utiliza um nível fixo *(offset)* de 32 dB para o ganho (obtido empiricamente), associado a um preditor LPC de ordem 10 que utiliza método da autocorrelação e janelamento híbrido. Este preditor adapta o ganho em torno desse offset, de modo a aumentar a sua precisão.

Seja o ganho da excitação $\sigma(n)$ que relaciona o erro normalizado $y(n)$ ao erro desnormalizado $\epsilon(n)$:

$$\epsilon(n) = \sigma(n)\ y(n)$$

Sejam ainda $\sigma_y^2(n)$ e $\sigma_\epsilon^2(n)$ respectivamente os valores quadrático médios de $y(n)$ e de $\epsilon(n)$. Então:

$$\log \sigma_\epsilon(n) = \log \sigma(n) + \log \sigma_y(n)$$

Pode-se predizer o valor presente do ganho da excitação $\sigma(n)$ a partir dos valores passados do ganho desnormalizado σ_ϵ através do preditor:

$$\log \sigma(n) = \sum_i^P p_i \log \sigma_\epsilon(n-i)$$

Neste caso, a ordem do preditor é P=10.

Note que esse preditor, ao se combinar as duas últimas equações, pode ser encarado como um preditor com P pólos e P zeros que usa $\log \sigma_y(n-1)$ como entrada:

$$\log \sigma(n) = \sum_i^P p_i \log \sigma(n-i) + \sum_i^P p_i \log \sigma_y(n-i)$$

A adaptação do ganho, como nas demais partes do LD-CELP, é feita retroativamente. O uso do método da autocorrelação para o cálculo dos coeficientes p_i garante a estabilidade do filtro acima, o que implica numa resposta impulsiva que cai assimptoticamente a zero. Isto faz com que erros introduzidos na transmissão tenham propagação limitada, indicando uma certa robustez deste bloco do algoritmo a erros. Porém, para aumentar essa robustez, aplica-se um expansão de faixa *(bandwidth expansion)*de 0.9 nos coeficientes (aproximando

os pólos e zeros da origem), encurtando a duração da resposta impulsiva do filtro do preditor e consequentemente diminuindo o tempo de propagação de erros dentro do algoritmo.

Adaptação do Filtro de Ponderação Perceptual: Para se adaptar os coeficientes do filtro de ponderação perceptual, aplica-se o método da autocorrelação sobre o sinal de entrada armazenado (não quantizado) com o janelamento híbrido, correção de ruído branco e o algoritmo de Levinson-Durbin. Os coeficientes LPC são então multiplicados pelos fatores γ_1 e γ_2, que definem o grau de ponderação perceptual.

Adaptação do Filtro de Síntese LPC: Os coeficientes do preditor LPC do filtro de síntese são adaptados usando-se como entrada o sinal quantizado. Do mesmo modo que para o filtro de ponderação perceptual, usa-se o janelamento híbrido, correção de ruído branco e o algoritmo de Durbin, porém com uma ordem maior (50) e diferentes parâmetros para a janela híbrida. Após o cálculo dos coeficientes LPC, estes passam por um alargamento espectral, sendo então fornecidos ao filtro de síntese.

2.4.3 Estruturas do Decodificador

A estrutura do decodificador do LD-CELP encontra-se na figura 2.2(b). Nela pode-se perceber que vários blocos do decodificador são idênticos aos do codificador. Entretanto, há alguns blocos originais, que são descritos a seguir.

Pós-filtro e sua adaptação: O pós-filtro *(post-filter)* [51] consiste de um filtro variante com o tempo colocado à saída de um decodificador com a finalidade de melhorar a qualidade percebida da voz decodificada.

A pós-filtragem, muito comum em algoritmos CELP, não foi utilizada no codificador da Fase 1 de testes devido a duas razões. A distorção do sinal introduzida pela pós-filtragem se acumula com múltiplas transcodificações (cascatas), o que pode comprometer seriamente a qualidade[18]. Além disso, a pós-filtragem introduz distorções de fase que podem trazer problemas na decodificação de alguns tipos de sinais de dados que contenham informação em sua fase (e.g. DPSK) [52, 44].

Como o algoritmo da Fase 1 não passou nos testes para 3 transcodificações em cascata, ele teve que ser modificado com a introdução da pós-filtragem no decodificador da Fase 2. Entretanto, o uso da pós-filtragem nos moldes convencionais implicaria em distorções inaceitáveis, pelas razões já citadas. Então, ao invés de otimizar a pós-filtragem para 1 transcodificação, otimizou-se para 3 transcodificações, o que implica numa menor quantidade de pós-filtragem para cada passo de transcodificação [42]. Com isso, a distorção introduzida pela pós-filtragem ficou dentro de níveis aceitáveis, enquanto que a qualidade subjetiva melhorou muito para as 3 transcodificações em cascata (testes subjetivos na época mostraram um aumento de 26% no MOS). Mesmo para somente 1 transcodificação houve uma boa melhora na qualidade subjetiva. Durante os testes do codec da Fase 2, a mudança mostrou-se eficaz na melhoria da qualidade subjetiva, mantendo o bom desempenho para sinais de dados.

[18] A principal distorção gerada pela pós-filtragem é a redução da largura de faixa dos picos espectrais. Seu acúmulo (por múltiplas transcodificações) introduz severas distorções na voz decodificada.

O pós-filtro usado no LD-CELP consiste de 3 blocos: um pós-filtro de longo prazo, um pós-filtro de curto prazo e em controle automático de ganho.

O *pós-filtro de longo prazo*, ou *pós-filtro de pitch*, é um filtro cuja resposta em freqüência é um pente com raias localizadas em freqüências múltiplas da freqüência fundamental (esta sendo recíproca do período de pitch) e é descrito pela equação:

$$H_l(z) = g_l(1 + bz^{-p})$$

onde p é o período de pitch obtido por um detector de pitch (para detalhes, ver [37, pp.18–20]) e g_l e b sãocoeficientes adaptados a partir de 240 amostras armazenadas da excitação. No LD-CELP, p varia de 20 a 140 (57 a 400 Hz).

Assim, ao se aplicar esta pós-filtragem ao sinal sintetizado, enfatiza-se o espectro na região em torno das freqüências múltiplas da freqüência fundamental. Perceptualmente, o efeito é de enfatizar a percepção do pitch e melhorar a qualidade subjetiva.

Já o *pós-filtro de curto prazo* tem a função de reduzir o ruído de codificação audível, através da ênfase dos picos e atenuação dos vales do espectro do sinal de entrada. Isto se baseia no fato de que os picos do espectro LPC (i.e., a região das formantes) é perceptualmente muito mais importante que a região dos vales.

O pós-filtro de curto prazo se compõe de um filtro de igual número de pólos e zeros com característica em geral passa-baixas, em cascata com um filtro passa-altas de primeira ordem. Este filtro de primeira ordem serve para compensar o efeito de "abafamento" causado pela atenuação das componentes de freqüência mais altas. A forma geral do pós-filtro de curto prazo, nos termos descritos, é dada por:

$$H_s(z) = \frac{1 - A(z/\gamma_1)}{1 - A(z/\gamma_2)}(1 + \mu z^{-1}), \text{ com } 0 < \gamma_1 < \gamma_2 \leq 1.$$

onde

$$A(z/\gamma_j) = \sum_{i=1}^{M} \gamma_j^i a_i z^{-i}, \text{ para } j = 1, 2.$$

$$\mu = \gamma_3 k_1$$

e γ_1,γ_2 e γ_3 são constantes, a_i são os coeficientes LPC, M é a ordem do preditor LPC e k_1 é o primeiro coeficiente de reflexão.

Note-se que o filtro passa-altas é implementado por um coeficiente adaptado, o coeficiente de reflexão k_1. Estatísticas para grande quantidade de amostras de voz (aproximandamente 10 minutos) mostram que na maior parte (92%) do tempo, $k_1 < 0$ (ver figura 2.4). Isto garante que o filtro assim implementado seja de fato um passa-altas para a maior parte dos sinais de voz processados pelo algoritmo.

No LD-CELP, a ordem do filtro $A(z)$ é 10 e os coeficientes a_i, bem como o coeficiente de reflexão k_1, são obtidos durante o processo de adaptação do filtro de síntese do decodificador. As constantes $(\gamma_1, \gamma_2, \gamma_3)$ foram empiricamente escolhidas como (0.65,0.75,0.15).

Finalmente, o *controle automático de ganho* tem a função de fazer com que a potência do sinal decodificado à saída do bloco de pós-filtragem seja aproximadamente a mesma do sinal antes da pós-filtragem. O AGC é calculado pela razão das magnitudes médias do sinal antes

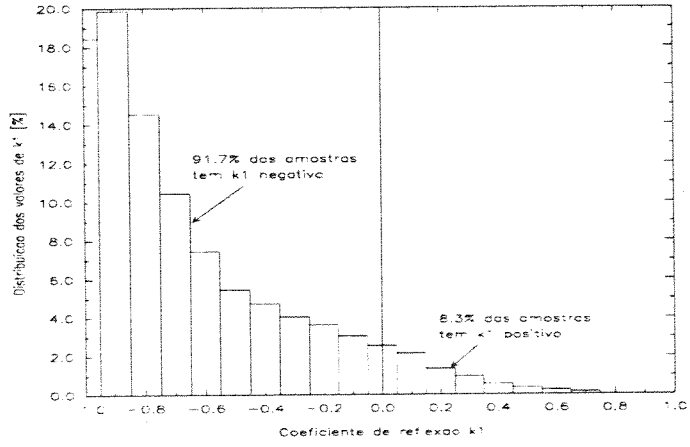


Figura 2.4: Histograma de valores de k_1 para 10 minutos de voz.

e após a pós-filtragem. O AGC evita a ocorrência de sinais muito atenuados, bem como a ocorrência de saturações. Fazendo o ganho variável, evitam-se as saturações que poderiam ocorrer para sinais de entrada de maior potência caso o ganho fosse fixo.

2.5 Sumário

Neste capítulo apresentamos os algoritmos de codificação de voz padronizados pelo CCITT e que são utilizados como pontos de referência quando da realização de testes subjetivos.

Dedicamos maior espaço à G.728, esmiuçando alguns de seus detalhes ainda obscuros na literatura. Esse detalhamento se justifica porque a G.728 foi recentemente aprovada pelo CCITT e porque apresenta uma grande complexidade, sendo o primeiro padrão internacional a empregar intensivamente técnicas de processamento digital de voz.

Capítulo 3

Infra-estrutura Laboratorial

3.1 Introdução

A realização de testes subjetivos dentro de padrões internacionais para a avaliação de equipamentos com uso pretendido para telefonia normalmente requer a disponibilidade de um conjunto de equipamentos e de instalações da ordem de milhares de dólares. Cada vez mais, implementações em software de algoritmos e de ferramentas vêm sendo utilizadas na realização de testes subjetivos.

As diversas fases de realização de um teste requerem uma parte comum e uma parte específica de equipamentos e instalações, que são descritas a seguir em detalhes. De um modo geral, uma sala controlada acusticamente é imprescindível, bem como voltímetros de voz, amplificadores de áudio de alta qualidade (baixa distorção), cabos e conexões de alta qualidade, conversores A/D e D/A com 16 bits de resolução, aparelho gerador de ruído modulado de referência *(Modulated Noise Reference Unit, MNRU)*, filtragem padrão (IRS, "anti-aliasing" e de reconstrução), telefones padronizados, entre outros.

Há ainda a consideração do uso de métodos de armazenamento e de processamento (condicionamento) analógicos ou digitais do material de voz. Isso implica em diversas ramificações que podem ser possíveis durante as fases de gravação, processamento e audição, isto é, o uso de diferentes filtros, filtragens analógicas ou digitais, por hardware ou software, de normalização de níveis de potência via hardware ou via software, adição de ruído analogicamente ou digitalmente (software), entre outras.

Independentemente do método, toda a parte do circuito por onde passam sinais analógicos deve ter um nível de ruído muito baixo (e.g. SNR maior que 60 dB), de modo a não comprometer a qualidade global dos sinais processados.

3.2 Sala Acústica

O primeiro requisito de infra-estrutura laboratorial é uma *sala isolada acusticamente ("sala acústica")*. Obviamente, esses termos são muito vagos, pois em algumas aplicações, 40 dBA de ruído ambiente é plenamente satisfatório, em outras 20 dBA pode ser insuficiente [15, p.313]. O nível de ruído ambiente normalmente encontrado em residências, escritórios e

outros ambientes onde o sistema telefônico é mais comumente usado, está na faixa dos 35 a 80 dBA. Então, para que os sinais gravados possuam ruído de fundo abaixo dessa faixa , geralmente a especificação de salas isoladas acusticamente diz que o ruído ambiente deve ser menor que 30 dBA (sem picos dominantes no espectro) e que o tempo de reverberação deve ser menor que 500 ms [91, p.12]. Esta instalação, no tocante à sua especificação, *independe* da abordagem digital ou analógica citada acima.

Uma sala acústica é uma versão simplificada de uma câmara anecóica, sendo mais econômica para se implementar. O projeto de uma sala acústica em geral envolve o uso de uma sala interna e de uma externa, formando um sistema. A sala externa deve ser feita com paredes duplas de tijolos maciços. Já a sala interna deve ser suspensa e se distanciar das paredes da sala externa em aproximadamente 10 a 20 cm, revestida com um material isolante acústico (e.g. lã de rocha revestida de um tecido grosso). O piso flutuante pode ser de concreto. Um cuidado importante se refere ao ar condicionado: sua vazão deve ser baixa, para não gerar turbulências e ruído na sala, além de prover controle de temperatura e de humidade. Em geral, um fluxo de 80m/min é adequado. O duto do ar condicionado deve apresentar placas metálicas alternativas revestidas de feltro (para absorver o ruído do ar em movimento). As portas da sala devem ser duplas, densas e bem vedadas com juntas de borracha.

Para garantir um baixo nível de ruído dentro da sala acústica, deve-se minimizar ao máximo a quantidade de equipamentos dentro dela. Isto torna necessária a criação de uma *sala de controle*, onde são alocados os equipamentos necessários. Deve-se também implementar uma forma de comunicação entre ambas as salas, de modo que o operador na sala de controle possa acompanhar a aplicação dos testes ou o processo de gravação de materiais de voz.

3.3 Infra-estrutura para gravação

Existem dois modos básicos para a gravação de sinais para serem utilizados em testes subjetivos: o analógico e o digital. O estado da arte hoje está numa fase de transição entre essas duas modalidades. Os laboratórios com tradição em testes subjetivos ainda persistem com metodologias mais analógicas que digitais, ao passo que os mais novos entram direto nas metodologias digitais. Mas a grande maioria dos laboratórios têm adotado uma estratégia híbrida de trabalho, ora usando a abordagem analógica, ora usando a abordagem digital nas diversas etapas de geração do sinal-fonte. A tendência, entretanto, é a de partir para abordagens puramente digitais.

3.3.1 Gravação Analógica

Consideremos inicialmente uma abordagem "*puramente analógica*" para a gravação de um material de voz (ver figura 3.1).

Para testes subjetivos de equipamentos, o material a ser processado precisa ter as características de freqüência típicas da rede onde esses equipamentos serão utilizados. No caso de codecs para aplicação em telefonia, o CCITT especifica na Rec.P.48 [113] o IRS (*Intermediate Reference System*), que representa a resposta em freqüência típica dos aparelhos telefônicos utilizados na rede. Por isso, ao se gravar materiais de voz para essa finalidade, deve-se de algum modo ponderar o sinal de voz pela máscara do IRS.

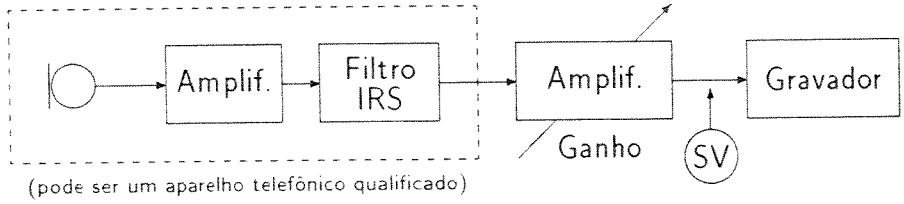


Figura 3.1: Exemplo de gravação puramente analógica.

Para a gravação do material dentro desse parâmetro, pode-se proceder de dois modos. Pode-se utilizar o monofone de um aparelho telefônico qualificado (atendendo à máscara da parte emissora do IRS), mas essa abordagem não é recomendada devido à sua baixa reprodutibilidade. Um outro modo seria utilizar um transdutor acústico-elétrico com resposta plana em freqüência para a faixa de áudio, ou pelo menos para a parte da faixa do áudio que se pretende utilizar[1]. Geralmente esse transdutor é um microfone dinâmico de boa qualidade e seu sinal deve ser filtrado com um filtro referente à parte emissora do IRS[2]. Na saída do filtro e entrada do gravador, é necessária a monitoração da potência (ativa) do sinal de voz e a modificação, se necessário, dos valores de ganho no amplificador ou filtro. Uma vez gravado o material, pode-se fazer uma edição desse material visando a uniformização das pausas entre as sentenças gravadas e uma equalização (em nível) mais precisa dos sinais gravados (uma vez que a potência do sinal de entrada nos codificadores é um dos parâmetros importantes a serem controlados na fase de processamento dos sinais).

Uma alternativa para fonte dos sinais seria a de se utilizar materiais previamente gravados em um sistema de alta qualidade, e.g. música em um toca-discos laser ou voz em um *DAT (Digital Audio Tape)*.

O gravador pode ser de vários tipos. Na década de 70, o mais usado era o gravador de rolo, que fornecia uma boa qualidade de reprodução, mas não impedia que corrupções dos sinais gravados ocorressem, especialmente com o passar do tempo ou diversas edições. Na década de 80, com a popularização dos equipamentos de vídeo-cassete, surgiram produtos para digitalizar o sinal de voz e modulá-lo na faixa de vídeo, de modo a permitir um armazenamento do sinal de uma forma mais imune a distorções. O mais famoso deles, fabricado pela Sony do Japão, conhecido genericamente como "Sony Betamax" (um abuso de notação, pois Betamax, ou β-Max, é o nome do sistema de vídeo-cassete desenvolvido pela Sony e que perdeu a batalha mercadológica para o sistema VHS, da JVC), foi adotado como um padrão de fato pelo CCITT e outros organismos internacionais para o intercâmbio de materiais de voz para testes subjetivos (por exemplo, para o ADPCM a 32 kbit/s, para o RPE-LTP e LD-CELP a 16 kbit/s).

A entrada do gravador possui um digitalizador com duas opções de resolução (16 bits sem proteção ou 14 bits com 2 bits de proteção), sendo utilizado geralmente o modo a 14 bits. O sinal de voz, após digitalizado, modula uma portadora de vídeo e o sinal resultante é gravado (geralmente com o sistema PAL-Europeu) numa fita de vídeo-cassete Betamax. Seu uso tende a ser descontinuado por três razões:

- Não ser controlável via computador: um dos aspectos importantes quando do embara-

[1] Por exemplo, se a taxa de amostragem para a digitalização de um material de voz for de 10 kHz, a linearidade do microfone é importante somente na faixa 0–5 kHz.

[2] Caso o procedimento de gravação seja para armazenamento digital, pode-se dispensar a ponderação pelo filtro IRS via hardware, caso haja disponível uma implementação em software do mesmo, como no Capítulo 3 de [11].

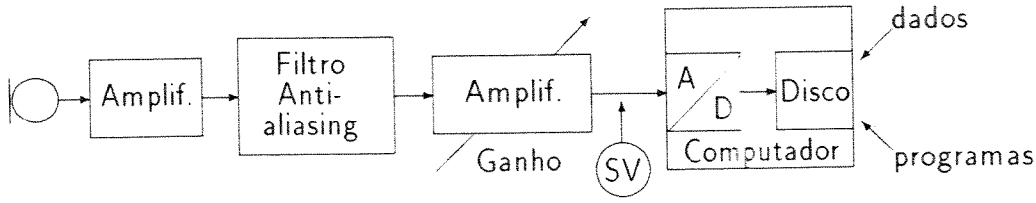


Figura 3.2: Exemplo de gravação digita

lhamento do material para audição, após o processamento é a edição do material. Como o Betamax não é controlável por computador, mas somente pelo painel frontal, toda a operação tem que ser manual. Isto significa longos tempos de preparo do material, sem contar os erros de edição (seleção e montagem de material errado) e as degradações que surgem (estalidos, entre outras).

- Susceptibilidade a erros: em geral as fitas de vídeo são altamente susceptíveis a terem seus registros alterados por campos magéticos e defeitos no material. Para um sinal de vídeo de um filme, essas degradações em geral passam desapercebidas pela alta redundância do sinal. No caso do sinal digitalizado isso não é verdade, pois toda informação é essencial, e perdas introduzem distorções que degradam o material gravado, muitas vezes o deixando inaproveitável para testes subjetivos. O uso de custosas fitas de altíssima qualidade e da opção de 14 bits com 2 bits de proteção contorna (mas não resolve) esse problema.

- Descontinuidade da produção do equipamento pela Sony: este é o mais forte dos argumentos de que mudar de padrão de gravação se faz necessário.

3.3.2 Gravação Digital

Por outro lado, partindo por uma abordagem *"puramente digital"* (veja figura 3.2) para as gravações, saindo do microfone com resposta linear, devemos ter um amplificador (para melhor explorar a faixa dinâmica dos circuitos utilizados, e.g. filtros ativos), um filtro "anti-aliasing" (para limitar a faixa do sinal), eventualmente um outro amplificador, e finalmente um conversor A/D de alta qualidade (geralmente de 16 bits) operando a uma taxa de amostragem adequada ao codec que se deseja avaliar. Esse conversor deve estar de alguma forma conectado a um dispositivo de armazenamento de massa de alta capacidade de dados, preferivelmente através de um computador (PC ou Sun, por exemplo).

A partir da digitalização, o ajuste do nível de gravação e a filtragem pelo IRS são feitos "off-line" via software, bem como a adição de ruídos e outras manipulações do sinal.

3.3.3 Filtragens

No caso de gravações analógicas que passarão para o domínio digital em algum momento ou para gravações digitais, um aspecto importante é o da filtragem do sinal. Além da filtragem *anti-aliasing*, podem surgir diversos contextos onde a filtragem adicional do sinal de voz pode ser necessária.

Um exemplo, já citado, é o da ponderação IRS. A curva do Sistema de Referência Intermediário da Recomendação CCITT P.48, parte de transmissão, é mostrada na figura 3.3.

Uma aplicação especial de filtragem ocorre no caso da simulação digital (software ou hardware) de transcodificações assíncronas. *Transcodificações assíncronas* são processamentos que envolvem a passagem de um sinal codificado no formato da G.711 para o domínio analógico e posterior re-codificação no formato anterior. Um exemplo disso seria um assinante servido por uma central CPA conectando-se a um outro assinante também servido por uma CPA, mas que em algum ponto da conexão entre eles a transmissão é feita por um canal analógico. A característica principal da transcodificação assíncrona é que, além de haver um potencial aumento do ruído de quantização, não há um sincronismo entre o D/A e o A/D das centrais. Essa falta de sincronismo entre os relógios gera no sinal uma distorção adicional, que pode ser simulada por uma filtragem com distorção de fase [114].

Outra aplicação é no caso de uso de sobre-amostragem para a geração de um material de voz de maior qualidade (ver discussão no Capítulo sobre Testes Subjetivos e [11, Cap.3]). Neste contexto, ao se digitalizar sinais de voz a 16 kHz para aplicações em telefonia, é empregada uma sobreamostragem de 2:1 para facilitar a filtragem anti-aliasing e garantir uma maior qualidade do sinal. Para ser utilizado em codificadores como o G.711, G.721 (G.726) e G.728, é necessário baixar-se a taxa para 8 kHz, i.e., utilizar-se um fator de dizimação de 2:1. Após os processamentos, para a audição, precisa-se voltar à taxa original, interpolando-se com um fator 1:2 para garantir uma boa qualidade de reconstrução do sinal. Um outro exemplo nesta mesma linha é o mecanismo de submissão de materias fontes de voz para as sessões laboratoriais para os testes subjetivos do codificador de segunda geração para o sistema rádio-móvel europeu. Nele, os laboratórios participantes dos testes enviaram fitas digitais DAT e, no laboratório central, os sinais digitalizados armazenados nessas fitas com uma taxa de amostragem de 48kHz (com 16 bits/amostra) foram dizimados por um fator de 6:1 e então processados pelos codificadores.

IRS da CCITT P.48

A máscara do IRS da Recomendação P.48 do CCITT tenta reproduzir a característica de amplitude de um telefone típico que pode ser encontrado na rede. A resposta média da parte de transmissão do IRS está mostrada na figura 3.3. Nela, pode-se ver que há uma ênfase na região em torno de 2kHz e uma atenuação forte abaixo dos 200Hz e acima dos 4kHz. A P.48, entretanto, não especifica a resposta de fase para um aparelho telefônico tradicional.

Um ponto interessante de se ressaltar é que, em testes subjetivos, normalmente não se especifica a filtragem pelo filtro de recepção da P.48 dos sinais (processados) a serem utilizados. Isto se deve ao fato de a máscara da recepção para a P.48 apresentar uma característica bem plana.

Na implementação da Biblioteca de Ferramentas de Software do CCITT, STL92 [11, pp.18–19,20–29], a ponderação IRS é feita através de uma filtragem do tipo FIR *(Finite Impulse Response)* com característica de fase linear, disponível para sinais digitalizados a 8 e a 16 kHz, com 151 e 209 "taps", respectivamente.

Mudança de taxa

Antes de se realizar a redução de taxa de N:1, é necessário limitar a faixa do sinal, para só então se descartar (N–1) a cada N amostras. Alternativamente, quando aumentando a taxa

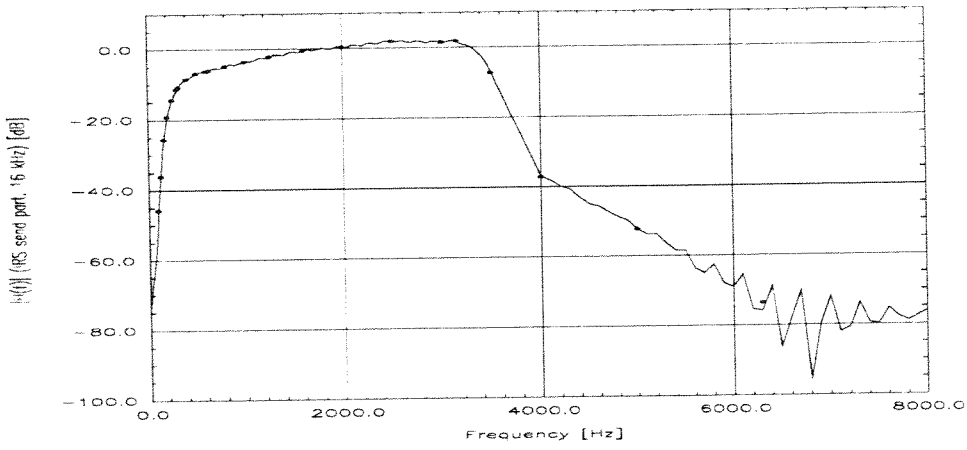


Figura 3.3: Resposta média do IRS, parte de transmissão.

de 1:N, após a inserção de (N–1) amostras nulas entre as amostras originais, deve-se efetuar uma filtragem, em geral do tipo passa-baixas. Além disso, deseja-se que a mudança de taxa seja feita introduzindo-se o mínimo de distorção possível. Para isso, deve-se utilizar filtragens passa-baixa extremamente planas na faixa de passagem e com fase linear. Na STL92 [11, pp.18,20–29], estão implementadas mudanças de taxa com alta qualidade para fatores 1:2, 1:3, 2:1 e 3:1 através de filtros FIR de fase linear. A ordem dos filtros é 118 para os fatores 1:2 e 2:1 e 168 para os fatores 1:3 e 3:1. Note-se que estes fatores não estão amarrados a uma taxa de amostragem específica, mas somente à relação entre a taxa de entrada e a de saída, ao contrário dos filtros para ponderação do IRS, que são específicos para cada taxa de amostragem. Na Figura 3.4 está a resposta em freqüência para os filtros com um fator de mudança de taxa de 2 (2:1 ou 1:2).

Para o caso especial de simulação de conversão para o domínio analógico, entretanto, o importante não é tanto a resposta em freqüência, mas a introdução de distorção de fase. Nesse caso, pode-se utilizar filtragem do tipo IIR *(Infinite Impulse Response)*. Na STL92 [11, pp.19–20,30–34], isso é feito pela filtragem que segue a máscara da Rec.CCITT G.712 [115], utilizada para sistemas log-PCM. Esta filtragem é implementada por um filtro passa-faixa IIR [66, Cap.4,p.153] com 3 células de segunda ordem em paralelo e está disponível para conversão entre 16kHz e 8kHz. À semelhança do filtro IRS, eles são amarrados à taxa do sinal. Na figura 3.5 está ilustrada a resposta em freqüência para o filtro PCM padrão da G.712 para sinais amostrados a 16 kHz.

3.4 Equipamentos

Os equipamentos necessários podem ser descritos de modo mais ou menos genérico para as duas classes de processamento descritas no item anterior e envolvem ou o interfaceamento

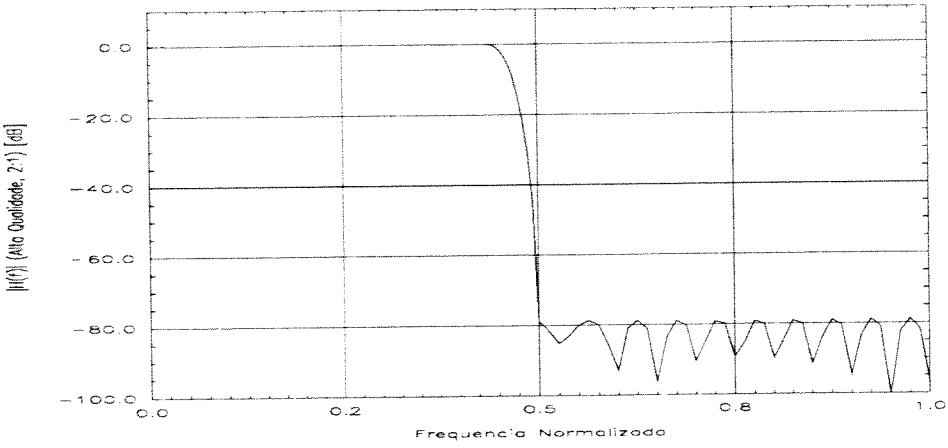


Figura 3.4: Resposta em freqüência do filtro de alta qualidade da STL92.

entre representações do sinal de voz processado e a ser processado, ou a geração de condições de referência necessárias para o teste subjetivo.

Para a determinação da qualidade de dispositivos e de circuitos, normalmente se utiliza oum *Analisador de Áudio*. Este dispositivo, em especial o model HP8903B [116], permite a medida de um modo simples e preciso de parâmetros como relação sinal-ruído, distorção e SINAD para sinais na faixa de áudio. Uma metodologia de qualificação de sistemas A/D e D/A do CCITT baseia-se nesse equipamento [117].

Para manter a qualidade dos sinais processados no circuito, deve-se utilizar amplificadores com baixo nível de ruído e distorção. Isto pode ser avaliado pela metodologia descrita em [117].

Quando da utilização de uma configuração de processamento analógico (ver Capítulo 4), é conveniente o uso de uma *Interface Analógica Comum*. Ela faria a conversão A/D e D/A para e de um padrão digital pré-estabelecido (como o formato log-PCM da CCITT G.711). Um exemplo de tal interface é a que foi utilizada nos testes para um padrão CCITT de codificação de voz a 16 kbit/s [91, 118], cujo nível máximo de entrada era +3dBm, com impedância de entrada e saída diferencial de 600Ω no lado analógico, convertendo o sinal analógico para um sinal digital serial codificado segundo a lei μ da G.711 (8 bits/amostra, 8 kHz de taxa de amostragem, perfazendo uma taxa de 64 kbit/s). Adicionalmente, a interface gerava um sinal de 8 kHz sincronizado com o início de cada palavra. Para a conversão de lei μ para analógico, a interface necessitava de um sinal externo a 8 kHz, sincronizado com o feixe a 64 kbit/s das palavras a serem convertidas.

Para o ajuste em um nível pré-definido dos sinais de voz a serem processados, o CCITT especifica o uso de um algoritmo de medição da potência ativa do sinal de voz descrito na Recomendação P.56 (Blue Book) ([119],[11, Cap.8]), comumente referido como *Voltímetro de Voz (Speech Voltmeter)*, descrito mais abaixo. Em termos de implementações do algoritmo, é clássico o equipamento analógico desenvolvido pela British Telecom e industrializado pela

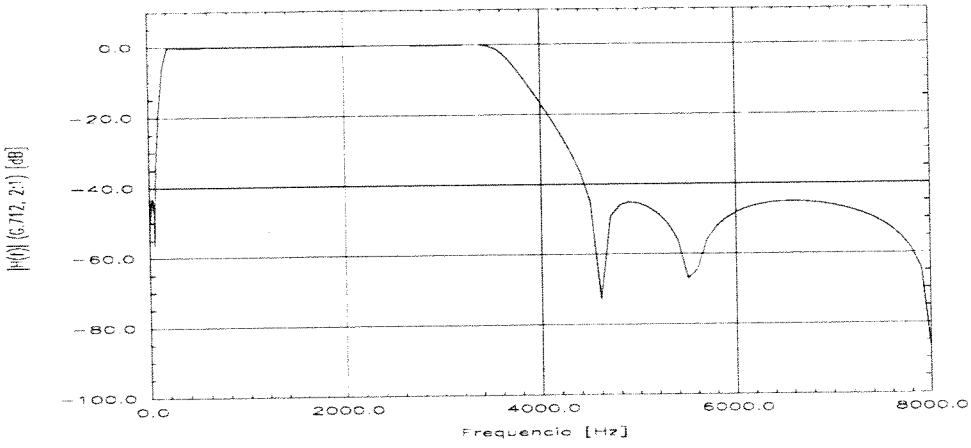


Figura 3.5: Resposta em freqüência do filtro G.712 da STL92

Malden Electronics, chamado SV6, que tem sido utilizado como um padrão em todos os testes subjetivos oficiais pelo CCITT e ETSI. Para a implementação digital, em especial em software, uma implementação de referência foi desenvolvida pelo Grupo de Usuários de Ferramentas de Software (UGST) do CCITT e publicada em maio de 1992 como parte da Biblioteca de Ferramentas de Software do CCITT [11, Cap.8].

Uma das mais importantes condições de referência são as geradas pelo *MNRU (Modulated Noise Reference Unit)* [120],[11, Cap.7]. O MNRU foi idealizado para ser um equivalente de referência (paramétrico) para ser utilizado na comparação de sistemas digitais.

Uma aplicação desta unidade é tornar possível a comparação de diferentes realizações de um mesmo teste subjetivo (por exemplo, um teste realizado em diferentes países com diferentes línguas). Não é possível a comparação direta dos resultados MOS, pois o valor médio em cada realização em geral apresenta um *"offset"* em relação a outras realizações. Assim, todos os resultados de cada um dos testes são convertidos para a escala Q do MNRU (ver mais à frente) e então são comparados entre si.

Adicionalmente, a conversão para Q do MOS encontrado para o codec sob teste é útil para a determinação de vários fatores de qualidade como o ruído de quantização e o ruído de circuito gerados pelo codec, ajudando portanto no estabelecimento de regras de planejamento para a introdução do codec na rede pública telefônica [121, 122, 123].

É importante também o estudo do comportamento dos codificadores com a presença de erros. Para isso, é necessário um mecanismo de inserção de erros. A nível de CCITT, não existe nenhuma recomendação que especifique um modelo, tampouco um equipamento de referência amplamante usado, como o caso do SV6 e do MNRU da BT. Entretanto, o UGST produziu uma implementação em software que foi publicada recentemente dentro da STL92 [11, Cap.4].

Outras condições de referência são outros padrões já criados e aos quais o desempenho do sistema em questão deve ser comparado. Exemplos necessários são, além da condição direta

(nenhum processamento), o padrão log-PCM utilizado em sistemas de comunicação (CCITT G.711) e o padrão ADPCM a 32 kbit/s (CCITT G.726). Com a padronização do LD-CELP pelo CCITT como o padrão de codificação de voz a 16 kbit/s (Rec.G.728) [37], este também deverá se tornar uma referência. Uma descrição destes codificadores é dada no Capítulo 2 deste trabalho.

3.4.1 O Algoritmo do Voltímetro de Voz

A especificação para a medida do nível ativo de um sinal de voz é dada na recomendação CITT P.56 [119]. Esse algoritmo é normalmente chamado de *Voltímetro de Voz (Speech Voltmeter)*, em função de sua implementação em hardware feita pela British Telecom e Malden Ltd., batizada de "SV6 Speech Voltmeter". Informações adicionais podem ser encontradas no *Handbook on Telephony* do CCITT [124] e em [11].

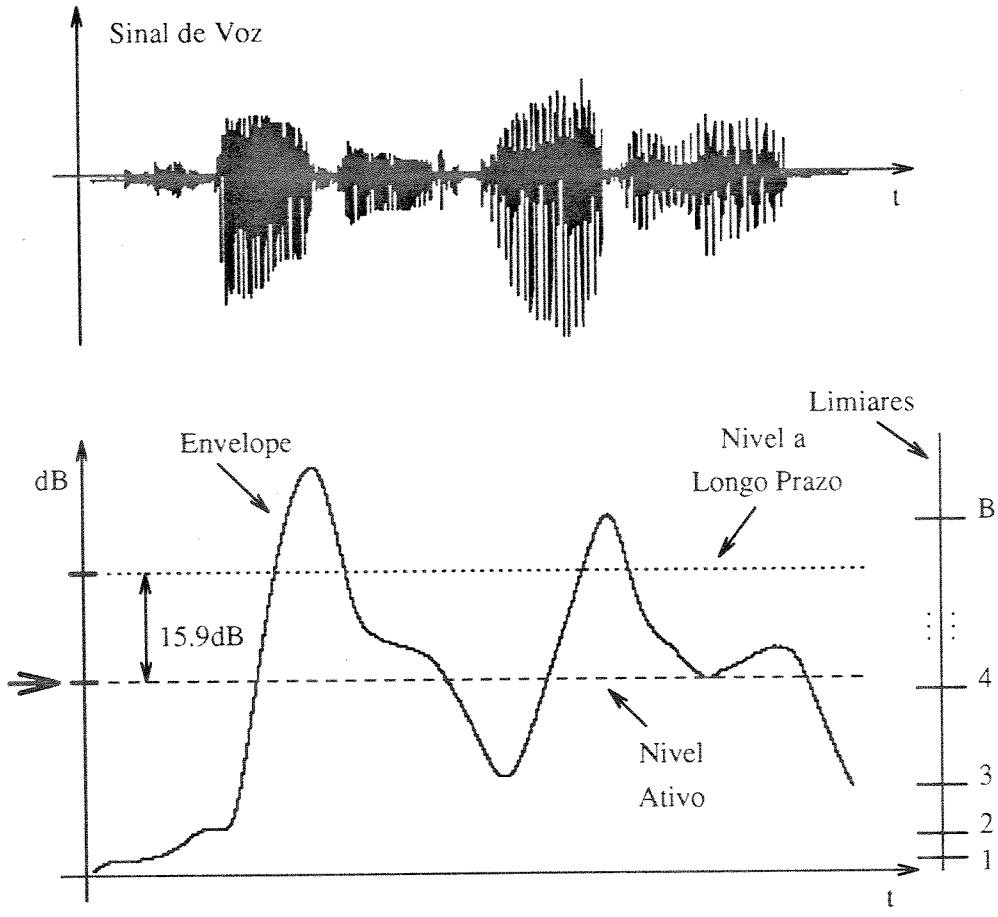


Figura 3.6: Um sinal de voz e sua envoltória, calculada pelo algoritmo da P.56. Estão indicados na figura o nível a longo prazo, o nível ativo e uma indicação dos B limiares.

Sumarizando, o algoritmo da P.56 toma amostras de um sinal na faixa de voz, calcula algumas estatísticas e, ao final do segmento de fala de interesse, calcula o seu nível ativo. Isto significa que momentos de silêncio e ruído de canal desocupado são desconsiderados no cômputo do nível médio do sinal. As pausas do tipo estrutural são consideradas no cálculo do nível ativo,

enquanto que as gramaticais são desconsideradas. *Pausas estruturais* são as pausas com duração média não superior a 250ms que são parte inerente ao processo de produção da voz. *Pausas gramaticais* são aquelas com duração superior a 300ms e que em geral são usadas para enfatizar palavras no processo conversacional. As pausas estruturais são um elemento imprescindível do processo de fala e contribuem de um modo importante para a sonoridade *(loudness)* da fala, ao contrário das pausas gramaticais, que não contribuem para a sonoridade. Este fato justifica a inclusão de uma e exclusão da outra.

Há na P.56 um vetor pré-definido de B limiares (ver a figura 3.6, existindo um contador de atividade associado a cada um deles. Para decidir entre a atividade ou não de um segmento de fala, o algoritmo calcula a envoltória (ou amplitude média de curto-prazo) do sinal, excluindo pausas superiores a 350ms. Para cada amostra de voz, analisa-se se a envoltória excede os diversos limiares do algoritmo ou se o tempo de *"hangover"* (200ms±5%) ainda não expirou, incrementando ou não os contadores de atividade correspondentes.

Após o término do segmento de interesse, o algoritmo passa à avaliação do nível ativo através da determinação do tempo em que o sinal foi considerado ativo (ou ainda, do número de amostras consideradas ativas). O número de amostras ativas será, por definição, o número de amostras em que a amplitude média de curto prazo ultrapassou um limiar que está 15.9dB abaixo da potência média de longo prazo. Em geral, esse número fica entre dois possíveis limiares, obrigando a interpolação entre eles para encontrar o valor aproximado do nível. A P.56 recomenda que o uso do *método da bissecção* [125], aceitando como tolerância uma margem de 0.5dB. Portanto, todas as medidas da P.56 têm uma margem de erro de ±0.5dB. Por fim, a potência a longo prazo do segmento medido é dividida pelo tempo de atividade (ou número de amostras ativas), definindo-se o seu nível ativo[3]. O fator de atividade do segmento de fala é dado pela razão entre o tempo de atividade (ou o número de amostras efetivamente ativas) e o tempo total de medida (ou o número total de amostras).

Um cuidado importante deve ser tomado quando normalizando sinais pelo algoritmo da P.56, quer pelo SV6, quer pelo algoritmo em software. A margem de 15.9dB acima foi otimizada para sinais de voz limpos, com baixo nível de ruído de fundo. Isto significa que no caso de geração de materiais de voz para testes subjetivos, em especial para o caso de materiais processados que possuam um nível significativo de ruído (e.g. MNRU para baixos valores de Q), o algoritmo da P.56 *não* deve ser utilizado para re-equalização desse material. Isto ocorre por que o nível de ruído introduzido está bem além do limiar de 15.9dB e levará o algoritmo da P.56 a gerar medidas incorretas do nível ativo da voz. Um modo prático de se observar se o algoritmo da P.56 pode ser utilizado em materias processados de voz é observar o fator de atividade: se o fator de atividade aumentou significantemente em relação ao do sinal original, então a P.56 não pode ser utilizada.

O algoritmo da P.56 é um processo de medida digital. A P.56 recomenda uma taxa de amostragem superior a 600Hz e o SV6 implementa uma taxa de amostragem em torno de 700Hz. Note que isto é feito sem filtragem *anti-aliasing*, provocando a sobreposição do espectro na faixa de voz. Entretanto, como as medidas se baseiam no cálculo da envoltória do sinal de voz, essa distorção não afeta significantemente a medida, dado que a tolerância para os resultados é de 0.5dB. Entretanto, para implementações em software, pode-se obter medidas mais precisas e o uso da sub-amostragem do sinal, como feita no SV6 e deixada em aberto pela P.56, deve ser evitado [126].

[3] Isto é feito sem se descontar as amostras abaixo do limiar.

3.4.2 O Algoritmo do MNRU

Historicamente, o MNRU foi idealizado para gerar uma distorção não-linear que simulasse a distorção produzida por sistemas de codificação PCM logarítmicos [127, 83]. Como se sabe, estes últimos produzem uma relação sinal-ruído de quantização independente do nível do sinal de entrada[4]. O resultado corresponde subjetivamente à adição ao sinal original de um ruído gaussiano (branco na faixa desse sinal) cuja amplitude seja proporcional à amplitude desse sinal (isto é, maior ou menor amplitude, mais ou menos ruído é injetado). Isto equivale à modulação do ruído pelo sinal de voz, advindo daí o nome "ruído modulado". A relação entre o sinal e tal ruído é referida como Q[5]. Com tal sistema, poder-se-ia comparar novos sistemas com a distorção que seria causada por um sistema de quantização log-PCM equivalente apenas variando-se a relação sinal-ruído deste último, dispensando o uso de um sistema log-PCM real (o que implicaria na necessidade do ajuste de diversos parâmetros).

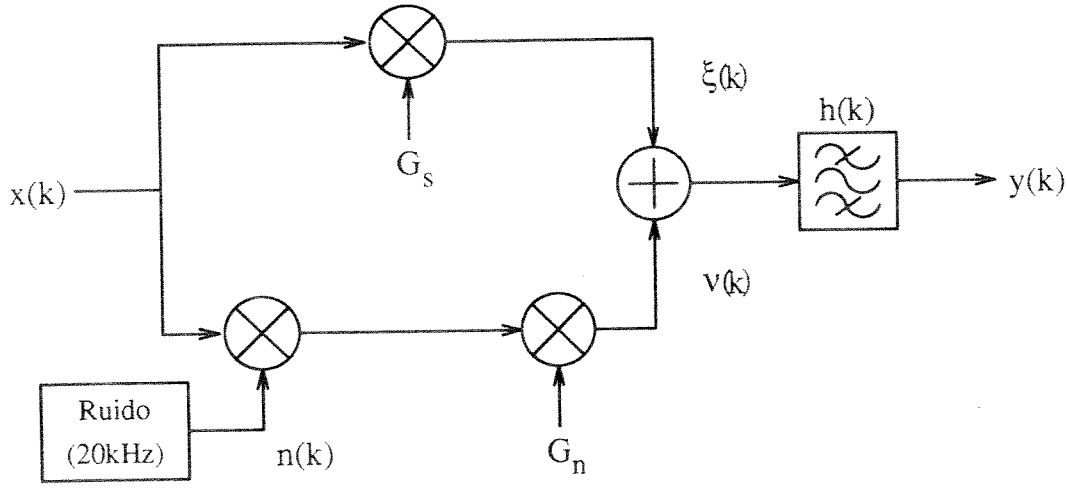


Figura 3.7: Diagrama do MNRU de faixa estreita da P.81

As primeiras idéias de um MNRU surgiram na referência [127]. Nela, a geração de ruído correlato ao sinal de voz era gerado baseado num modulador em anel controlado pelo sinal de voz de entrada, que modula uma portadora de ruído. Essa portadora possuía uma distribuição de potência relativamente uniforme na faixa de 0 a 20kHz. O ruído modulado era então adicionado ao sinal de entrada com um certo ganho aplicado, de modo a obter uma relação sinal-ruído controlada na saída, após uma filtragem passa-faixa (300–3400Hz). O esquema apresentado nesse artigo é basicamente o mesmo esquema implementado na P.81 hoje, bem como no MNRU da British Telecom. Esses algoritmo é chamado de MNRU de faixa estreita, pois se aplica à faixa de telefonia, em contraste com o MNRU de faixa larga (50–7000Hz), destinado a sistemas de faixa mais larga, como o da G.722 [128].

Devido a essa origem, a P.81 especifica alguns dos aspectos mais gerais do MNRU, deixando vagos alguns aspectos que, à época, viabilizaram uma implementação analógica. Entretanto, é possível que duas implementações diferentes não gerarão resultados compatíveis. Por essa

[4] Os sistemas PCM comerciais são uma aproximação da compressão logarítmica, mas a SNR é praticamente constante sobre uma vasta faixa de níveis de entrada, facultando esta comparação.

[5] Q é definida como sendo a relação sinal-ruído (em dB) do sinal de referência quando o MOS para a voz codificada é equivalente ao MOS da referência [121].

razão, a implementação da British Telecom foi tomada como padrão hardware de-facto.

O diagrama em blocos básico do MNRU da P.81 está na figura 3.7. Há nele dois caminhos: o caminho do sinal e o caminho do ruído. No caminho do ruído, o ruído gaussiano $n(k)$ é modulado pelo sinal de entrada $x(k)$ e tem um ganho G_n. O sinal resultante $\nu(k)$ é então adicionado à saída do caminho do sinal, $\xi(k)$, e filtrado, resultando no sinal corrompido por ruído modulado $y(k)$.

Em termos analíticos, o sinal corrompido por ruído modulado $y(k)$ é dado por:

$$y(k) = (G_s x(k) + G_n x(k) n(k)) * h(k)$$

Supondo que o filtro passa-faixa tem $|H(f)| = 1$ na região de passagem e chamando Q a relação sinal-ruído em dB à sua saída, teremos:

$$10^{Q/10} = \frac{\sigma_\xi^2}{\sigma_\nu^2} = \frac{E[\xi^2(k)]}{E[\nu^2(k)]} = \frac{G_s^2 E[x^2(k)]}{G_n^2 E[x^2(k) n^2(k)]}$$

Porém, como x e n são incorrelatos e o ruído é gaussiano com média 0 e variância 1:

$$10^{Q/10} = \left(\frac{G_s}{G_n}\right)^2 \frac{\sigma_x^2}{\sigma_x^2 \sigma_n^2} = \left(\frac{G_s}{G_n}\right)^2$$

ou

$$\begin{aligned} Q &= \Gamma_s + \Gamma_n \\ \Gamma_s &= 20 \log_{10}(G_s) \\ \Gamma_n &= -20 \log_{10}(G_n) \end{aligned}$$

Fazendo $\Gamma_s = 0$ $(G_s = 1)$, Q é exatamete Γ_n, i.e., a SNR é o ganho (em dB) do caminho do ruído.

Quando G_s e G_n são não-nulos, o MNRU estará no modo operacional normalmente chamado de *Modo de Ruído Modulado (Modulated-Noise Mode)*. Este é o modo de operação mais comum.

Alternativamente, se $G_s = 0$, a saída do algoritmo será somente o ruído modulado a um nível Γ_n dB abaixo do nível do sinal de entrada. Este é o *Modo Ruído (Noise-Only Mode)*.

Por outro lado, se $G_n = 0$, a saída do algoritmo será o sinal filtrado por $h(k)$, com um ganho G_s; este é o *Modo Sinal (Signal-Only Mode)*.

Atualmente, em testes subjetivos baseados em processamento analógico, usa-se o MNRU da British Telecom para a geração das condições de ruído modulado. Para procedimentos digitais, ainda não há uma implementação de referência como para o caso analógico. Entretanto, uma implementação de referência em software recentemente produzida é a da STL92 [11, Cap.7]. O Grupo de Especialistas em Qualidade de Voz (SQEG/XII) vem testando diversas implementações digitais do MNRU e dessa atividade pode surgir uma revisão da P.81, especificando-a melhor.

3.4.3 O Algoritmo de Inserção de Erros da STL92

O conceito de inserção de erros aparece quando se necessita estudar o comportamento de equipamentos e de sistemas de transmissão digital sob condições de erro.

Isso implica na necessidade de haver um modelo para o canal de transmissão e de um algoritmo de geração de erros. Em geral, precisa-se de geradores de erros aleatórios *(random error)* e em surto *(burst error)*. Em outras situações, pode se tornar importante o estudo do comportamento desses sistemas frente a perdas de quadro de transmissão *(frame erasure)*, especialmente para a avaliação de sistemas de comunicação móveis.

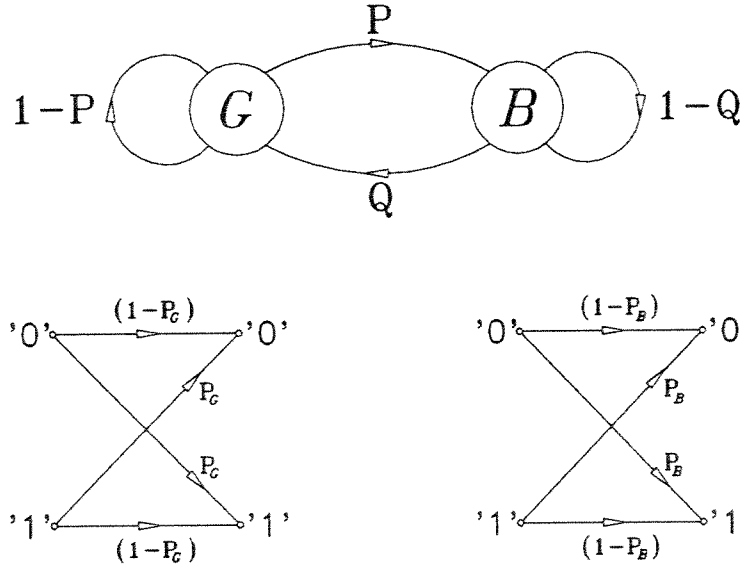


Figura 3.8: Modelo do Canal Discreto de Gilbert Elliott

O algoritmo de inserção de erros *(Error Insertion Device, EID)* baseia-se no modelo do Canal Discreto de Gilbert Elliott (GEC) e é mostrado na figura 3.8. Esse modelo tem dois estados: "Good" (G) e "Bad", ou "Burst", (B). Associado a esses dois estados, há quatro parâmetros (probabilidades): duas se relacionam à probabilidade do canal continuar no estado atual e duas relativas à probabilidade de haver a transição de estados.

As probabilidades associadas aos estados do canal são P e Q, sendo P a probabilidade de transição do estado G para o B e Q, a de transição de B para G. Consequantemente, a probabilidade de continuar no mesmo estado é $(1-P)$ e $(1-Q)$ para os estados G e B, respectivamente. Em um dado estado, há a probabilidade de ocorrência de uma mudança em um bit, que é P_G para o estado G e P_B para o estado B.

Portanto, o canal pode estar no estado G, onde a probabilidade de erro de bit média é bem baixa $(P_G \approx 0)$, ou no estado B, onde a probabilidade de erro de bit média P_B é relativamente alta $(P_B \approx 0.5)$.

A probabilidade de erro de bit BER gerada por esse modelo é:

$$BER = \frac{P}{1-\gamma} \cdot P_B + \frac{Q}{1-\gamma} \cdot P_G$$

onde

$$\gamma = 1-(P+Q)$$

Note-se que γ é uma medida da *correlação* entre os bits com erro. Portanto, γ indica se o canal apresenta erros aleatórios ou correlatos *(burst)*. Assim, se $\gamma \approx 0$, o canal apresentará

erros aproximadamente aleatórios, enquanto que $\gamma \approx 1$ implica num canal em que somente ocorrem erros em surtos *(burst)*.

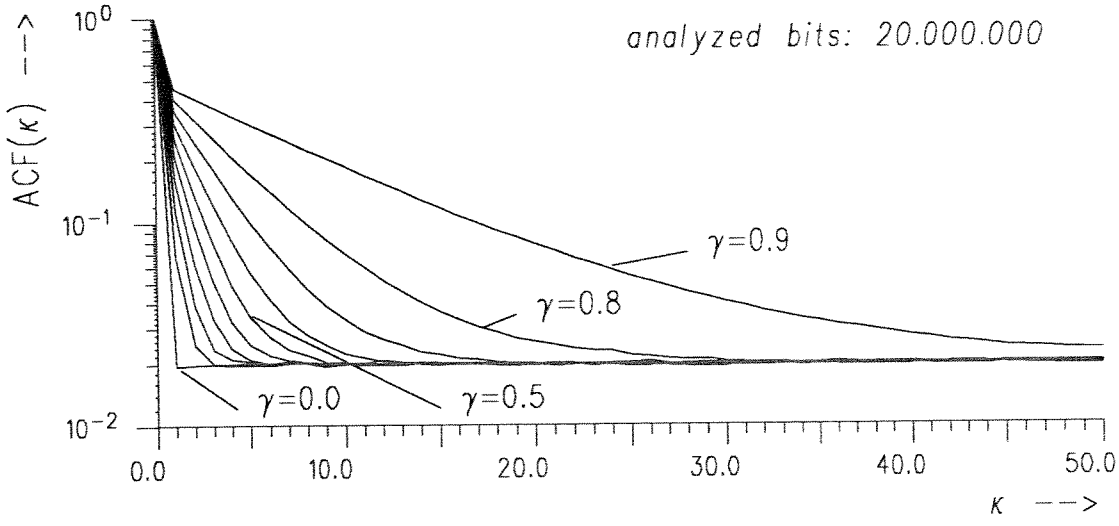


Figura 3.9: Autocorrelação ACF(κ) entre os bits com erro para diversos valores de γ com a BER fixa em 2%.

Há várias situações em que são interessantes seqüências com distintas probabilidades de erro BER e correlação γ. Das duas equações acima e escolhendo-se $0 \leq P_G < P_B \leq 0.5$, obtém-se:

$$P = (1-\gamma) \cdot \left(1 - \frac{P_B - BER}{P_B - P_G}\right)$$

$$Q = (1-\gamma) \cdot \frac{P_B - BER}{P_B - P_G}$$

No EID escolheram-se dois valores especiais: $P_G = 0$ e $P_B = 0.5$. Isto se relaciona ao fato de que no estado G, as mudanças de bit não são esperadas, daí $P_G = 0$. Para o estado B, espera-se que o canal esteja num estado totalmente incerto; por isso, $P_B = 0.5$. Com isso, reduzem-se as duas últimas equações a:

$$P = 2 \cdot (1-\gamma) \cdot BER$$

$$Q = (1-\gamma) \cdot (1 - 2 \cdot BER)$$

Na figura 3.9 pode-se ver que há um pico para a autocorrelação ACF(κ) para $\kappa = 0$ e que ela vai praticamente para zero quando $\kappa \neq 0$. Com γ crescendo, aumenta-se a correlação entre bits adjacentes. Para $\gamma \approx 1$, temos um aumento significativo na correlação entre os bits, levando a erros totalmente em surto, no limite.

3.5 Infra-estrutura para audição

Além da necessidade de uma sala isolada acusticamente (como descrito), é necessário haver equipamentos de audição adequados ao teste subjetivo em questão. Duas situações são de especial interesse: audição com equipamentos de alta-fidelidade e testes via aparelhos telefônicos.

3.5.1 Equipamentos de áudio

Em termos de equipamentos de alta-fidelidade, deve-se dispor de um amplificador de áudio de boa qualidade (baixo ruído e baixa distorção harmônica, etc) equipado com caixas acústicas com resposta relativamente plana.

É também conveniente que haja um gravador de alta qualidade, preferencialmente digital (e.g. Sony Beta-Max ou DAT – Ver Capítulo 4), de modo que materiais de demonstração possam ser gerados ou reproduzidos.

Pode ser importante o uso de equalizadores gráficos (de freqüência) para gerar sinais com uma ponderação de freqüência arbitrária (por exemplo, para gerar ruído colorido com espectro de Hoth [95, pp.266-267] a partir de um gerador de ruído branco), bem como o de misturadores *(mixers)*, para a combinação de sinais oriundos de diversas fontes, e.g. misturar um sinal de referência com um ruído de fundo.

3.5.2 Testes com Telefones

Os testes realizados com telefones em geral devem utilizar aparelhos que seguem as especificações mecânicas e elétricas do CCITT. Em relação à parte elétrica, devem seguir a Rec.CCITT P.48.

Deve-se providenciar também esquemas de conversão entre 2 e 4 fios (híbrida), bem como alimentação do telefone. Isto pode ser feito com um circuito equivalente ao da figura 3.10, normalmente chamado de *ponte de alimentação*.

Para permitir a audição em diferentes níveis, é conveniente que haja um amplificador com ganho variável entre a fonte de sinal e a ponte de alimentação. É desejável que esse amplificador possa ser controlado via computador, e.g. via interface GPIB, o que diminui os riscos de erros quando da aplicação de testes subjetivos.

Adicionalmente, geralmente é necessária a adição de ruído ambiente com espectro de Hoth na sala acústica. Em geral esse ruído é gerado a partir de um ruído branco ponderado em freqüência por um equalizador gráfico.

3.6 Software

A geração de sinais processados por algoritmos de referência e o processamento por algoritmos em teste podem também ser realizados através de processamento em software. Em especial

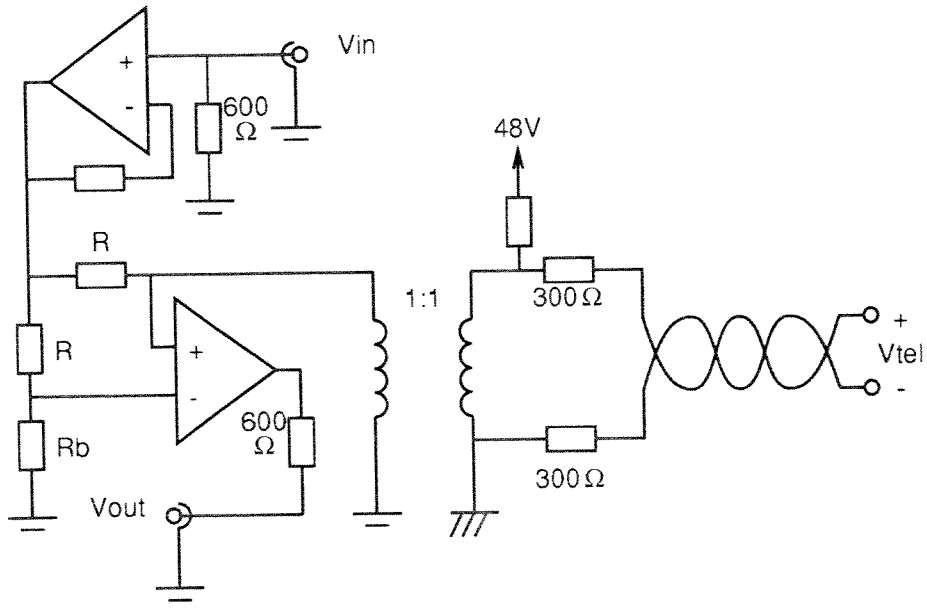


Figura 3.10: Ponte de alimentação para aparelhos telefônicos.

para testes de menor rigor (e.g. para testes informais ou em fases preliminares de desenvolvimento dos algoritmos, em que não estejam dosponíveis versões em hardware dos codecs) ou dentro de um ambiente de processamento puramente digital, o uso de implementações em software dos diversos algoritmos, referências e ferramentas de pré e pós-processamento é de extrema utilidade e conveniência.

Entre 1990 e 1992, o Grupo de Usuários em Ferramentas de Software do CCITT (UGST/XV) desenvolveu uma biblioteca de software para realizar essas funções. Dentro dela se encontram algoritmos de codificação de voz (G.711 e G.721), condições de referência (P.81, P.56) e ferramentas de pré e pós-processamento (mudança de taxa, filtragens, inserção de erros e conversão de formatos).

3.7 Sumário

Entendendo a infra-estrutura laboratorial como as ferramentas necessárias para a realização de avaliações de qualidade de algoritmos de codificação de voz, descrevemos aspectos de hardware e de software. Apresentamos a infra-estrutura de gravação, de processamento e de audição, bem como ferramentas de software necessárias tanto aos testes subjetivos como para as avaliações subjetivas.

Capítulo 4

Testes Subjetivos

4.1 Introdução

Discutimos a seguir diversos aspectos relevantes à metodologia de realização de testes subjetivos, desde sua motivação básica até sua análise, passando por sua história, seu projeto e aplicação.

4.1.1 Por que subjetivos?

Qualquer sistema de comunicação envolve três partes imprescindíveis: quem fala, quem ouve e o sistema de comunicação, por onde a informação flui.

Para um sistema de comunicação de dados, o objetivo a perseguir é diminuir o mais que se possa o número de bits com erro entre o transmissor e o receptor. Assim, uma medida conveniente para se avaliar o desempenho desse sistema é medir a taxa de erros: tanto menor, melhor. Esta é uma medida de desempenho objetiva, que plenamente avalia o sistema.

Porém , para sistemas onde o interlocutor é humano, há fatores envolvidos com o processo de análise da mensagem recebida que dificilmente poderão ser ponderados por medidas objetivas, como o incômodo que uma determinada distorção ou ruído causam, a inteligibilidade de um sinal ou mesmo o seu conteúdo cultural. Para esses sistemas, a análise fundamental de desempenho deve ser feita por meio de testes de opinião, onde pessoas membros de uma população-alvo (para a qual tal sistema de comunicação está sendo projetado) são convidadas a avaliar sinais por ele transportados, de acordo com uma certa metodologia; a análise estatística dessas opiniões será então utilizada para se estabelecer se o desempenho desse sistema está satisfatório ou não.

Obviamente, medidas objetivas são ferramentas importantes no processo de desenvolvimento de um sistema com interlocutores humanos, mas não são a avaliação definitiva: esta é e continuará sendo a avaliação subjetiva, feita pelo "elemento humano".

Neste trabalho, o nosso enfoque é para este tipo de sistema, mais especificamente sobre uma de suas partes: algoritmos de codificação digital de voz e de sinais na faixa de voz (especialmente voz), aos quais nos referimos intercaladamente como "algoritmos", "codificadores" ou "codecs".

4.1.2 O que avaliar?

Há diversos fatores importantes que devem ser analisados num teste subjetivo. A escolha dos mais prioritários pode variar de um algoritmo para outro e da aplicação pretendida para estes, mas a lista de fatores é mais ou menos a mesma:

- Variação do nível de entrada do sinal a ser codificado: por exemplo, se o codec estiver fisicamente numa central telefônica, dependendo da distância entre o usuário e ela, devido às perdas na linha, o sinal pode chegar mais ou menos atenuado na entrada do codificador;
- Ocorrência de erros no canal de operação do codec, como surtos de erro num canal de rádio;
- Possível ocorrência de múltiplas transcodificações ("tandem"): existência de codecs intra-rede, com conversão para outros padrões e de volta para o padrão original, ou a simples conversão para o mesmo padrão repetidas vezes;
- Diversidade de locutores: o conteúdo sonoro varia fortemente com fatores genéticos (e.g. sexo), étnicos e culturais;
- Presença de ruído ambiente em níveis variáveis: num escritório o nível de ruído é significantemente diferente daquele em frente ao telefone público situado numa avenida congestionada;
- Sinais dos quais o usuário requer boa qualidade, como por exemplo música instrumental;
- Atraso introduzido pelo esquema de codificação, que pode levar à percepção do eco em conexões de longa distância;

Desse modo, entre os fatores a serem avaliados, os mais importantes deverão ser investigados mais intensamente, bem como ter um peso maior na avaliação global do codec.

4.1.3 O conceito de qualidade

Qualidade é um conceito inerentemente relativo. Os fatores de que depende são inúmeros, segundo a abordagem que se dê. Do ponto de vista de nossa análise, devemos considerar os seguintes fatores:

- Aplicação pretendida: as exigências de "qualidade" para áudio profissional são bem diferentes das exigências para comunicações militares. Assim, um algoritmo cujo desempenho seja avaliado "bom" para comunicações militares (onde a maior preocupação é com a inteligibilidade do sinal), terá qualidade "inaceitável" para áudio profissional (quando a fidelidade e a faixa do sinal são mais importantes). Neste trabalho, o enfoque é para a qualidade de comunicações telefônicas ("Toll Quality").
- Público-alvo: quem utiliza o sistema de comunicação pode estar tão acostumado a sistemas com alto índice de degradação que, na avaliação de um codec, sejam extremamente "generosos" em suas notas. O outro extremo também existe: pessoas acostumadas com conexões de alta qualidade, quando chamadas a avaliar algoritmos para qualidade

telefônica, podem ser "rígidos" demais em suas avaliações. É curioso observar esta característica quando se consideram testes subjetivos realizados no passado e se constatou que as avaliações (utilizando-se obviamente os mesmos planos de teste subjetivo) feitas no Japão têm médias menores que para os países europeus, que por suas vez são mais rigorosos que os ouvintes norte-americanos [81].

Assim, de acordo com a abordagem, podemos estabelecer uma âncora, ou referência, para os níveis de qualidade, de modo a nos situarmos adequadamente frente a uma avaliação de qualidade.

Quanto à terminologia normalmente utilizada na literatura para classificar a qualidade de sistemas para sinais na faixa de voz, temos:

Broadcast ou Audio	Esta é a classe de sistemas que mantém a faixa plena dos sinais de áudio (sem distinção entre voz e música). Relações sinal-ruído típicas estão bem acima dos 30 dB.
Commentary	Como definida em [82], também denominada intra-CCITT como "wide-band", é a qualidade de sistemas com faixa até 7 kHz (CCITT: 50 a 7000 Hz), com mesma SNR e distorção harmônica que a qualidade de telefonia.
Toll	A mais conhecida, a qualidade de telefonia [82] é aquela que se obtém para sistemas cuja qualidade se compara à de um sinal analógico com faixa de 200 a 3400 Hz, relação sinal/ruído maior que 30 dB e distorção harmônica menor que 2 a 3 %. Para esta faixa de qualidade, o fator mais importante é a aceitação (ou grau de satisfação) do usuário [81].
Communication	Esta é uma qualidade intermediária, onde a faixa é a mesma da de comunicações telefônicas (200–3400 Hz), mas a qualidade é inferior (SNR<30 dB), mantendo a inteligibilidade [82].
Synthetic	Aqui, a qualidade é bastante reduzida, perdendo a naturalidade da fala ao ponto de ser difícil reconhecer o locutor. Nestes sistemas, o fator mais importante a analisar é a inteligibilidade [81].

4.1.4 Testes informais

A maioria dos testes realizados nas diversas fases de desenvolvimento de algoritmos de codificação de voz envolve testes de audição realizados por pessoas especializadas em codificação, geralmente o próprio pesquisador e membros da equipe em que ele trabalha. Esses testes não podem, devido à sua extensão limitada, ser utilizados para estabelecer de forma segura qual o desempenho (absoluto ou relativo) do esquema de codificação em questão.

Em geral, testes subjetivos informais carecem de várias características essenciais do ponto de vista estatístico:

1. **Tamanho do material fonte pequeno:** o material-fonte (sinal de voz a ser processado) utilizado para as avaliações de desempenho de algoritmos em fase de desen-

volvimento em geral não é estatisticamente significativo, tanto em termos de fonemas, como do conteúdo semântico e do nível social do locutor. Além disso, é mais comum que somente seja utilizada uma língua (a língua nativa do pesquisador que desenvolve o algoritmo).

2. **Tamanho da amostra limitado:** o número de pessoas utilizadas para as avaliações é reduzido, o que faz com que o nível de significância dos resultados seja muito pequeno.

3. **Condições de audição inadequadas:** nem sempre salas com as características acústicas necessárias são utilizadas para a audição de sinais processados. Além disso, muitas vezes se utilizam equipamentos de áudio de alta-fidelidade, não adequados à avaliação de algoritmos para sinais com "qualidade telefônica".

4. **Treinamento do ouvido:** o uso continuado dos mesmos avaliadores pode levar a um efeito de treinamento, quando sinais de boa qualidade podem vir a ser considerados insatisfatórios ou vice-versa.

No entanto, apesar dessas limitações, os testes informais são extremamente adequados na fase inicial de desenvolvimento, onde pequenas mudanças no esquema de codificação produzem mudanças significativas na qualidade do sinal processado. Isso porque com um material-fonte menor, o tempo de processamento é menor, gerando uma realimentação mais rápida para o projetista. Além disso, procedimentos informais de seleção dos ouvintes e de audição do material processado agilizam a realização dos testes, facilitando a realimentação ao projetista, o que é especialmente importante nessa fase do desenvolvimento.

Existe porém um ponto além do qual mudanças no algoritmo produzem alterações cada vez menos sensíveis, tornando necessária uma avaliação mais criteriosa e detalhada: é a hora de se partir para testes subjetivos formais.

4.1.5 Testes formais

Testes formais são testes subjetivos realizados de modo a manter o maior rigor estatístico possível para se atingir uma margem de segurança desejada para a análise de desempenho de um codec. Assim, o teste deve ser projetado com cuidado, de modo que as deficiências citadas acima não ocorram acima de um dado limite de segurança, que varia de teste para teste. Desse modo, o material-fonte deve conter uma amostragem significativa dos fonemas existentes, de pronúncias e vocabulários correntes da língua e de tipos de voz (masculina, feminina e infantil) e ser gravado em condições controladas de ruído ambiente e de circuito elétrico. O ambiente para audição também deve ser controlado (por exemplo, ruído ambiente abaixo de 30 dBA), bem como devem ser bem especificados os equipamentos de interface e audição (por exemplo, monofones com resposta em freqüência dentro da máscara IRS do CCITT).

A escolha do número mínimo de avaliadores é função do número de condições a serem testadas, do tamanho do material de voz processado (número de locutores e número de sentenças por locutor), do tipo de teste e do nível de significância estatística desejada para o teste (em geral não menos que 95%).

4.2 Histórico

A evolução dos métodos de avaliação está profundamente ligada à evolução dos próprios sistemas de comunicação, desde os mais simples do final do século XIX e princípio do século XX, até os mais complexos deste final de século.

No princípio, os métodos eram somente conversacionais, basicamente devido à maior praticidade deste tipo de método. Com o surgimento de gravadores de áudio e, mais recentemente, de sistemas de armazenamento de voz digitalizada, testes de opinião vêm substituindo os testes conversacionais.

Nesta seção, primeiro descreveremos os testes conversacionais, que não são o objeto de estudo nesse trabalho, mas que por razões históricas são incluídos [83]. Depois, um breve histórico do uso de testes de opinião será dado, até os dias atuais.

4.2.1 Testes Conversacionais

Para os testes conversacionais, sempre duas características do sistema de comunicação visavam ser analisadas: ou a sonoridade ou a articulação.

A medida de índices de sonoridade (*Loudness Ratings*) visava dar uma medida padrão da perda de transmissão de um dado caminho de voz (e.g, telefone–linha–telefone) da boca do locutor até o ouvido do ouvinte, expresso inicialmente em "Milha-Padrão" e depois em "Equivalentes de Referência". Assim, dado um locutor falando a um nível de voz constante, um número é associado à sonoridade que o ouvinte sente. O material fonético utilizado é sempre padronizado e inclui um pequeno número de palavras, que geralmente se escolhem por serem foneticamente balanceadas e supostamente representativas da língua como um todo.

A medida de índices de articulação (*Affaiblissement Équivalent pour la Netteté*, AEN) indica a capacidade de informação de um canal quando a informação que ele está transmitindo é na forma de voz, baseando-se no uso de logátomos[1]. Os logátomos são lidos sempre num mesmo nível sonoro e escolhidos de forma a gerar uma seqüência balanceada e aleatória foneticamente; o índice de articulação é obtido verificando-se a percentagem de logátomos interpretados corretamente. Esperava-se que o AEN pudesse substituir os sistemas baseados em índice de sonoridade, mas na década de 50 constatou-se sua "não-linearidade": circuitos com diferentes degradações, apresentando um desempenho que o usuário não considerasse como equivalentes, resultavam valores AEN muito próximos.

Ambas as medidas faziam uso de equipes especialmente treinadas para a realização dos testes, com um cuidado quanto ao nível de leitura do material de voz, a posição do monofone (para leitura e para audição), etc.

A seguir são descritos alguns sistemas utilizados historicamente para a avaliação conversacional subjetiva da qualidade telefônica. Note-se que na evolução dos sistemas está implícita uma mudança do parâmetro relevante: inicialmente, com linhas passivas, o aspecto mais importante era o volume que chegava ao usuário; com a evolução dos sistemas (sistemas ativos e cápsulas de maior eficiência), a inteligibilidade passou a ter maior importância. Mais tarde, com o

[1] Logátomo é uma combinação som vocálico–som consonantal–som vocálico, de modo que essa combinação não tenha sentido na língua considerada, como `ati`, `umi`, etc.

NOSFER, retornou-se aos testes de sonoridade.

Milha Padrão

No princípio do século XX, diversos países começaram a implantar suas redes telefônicas em âmbito nacional. Tais redes eram as mais simples possíveis: apenas um telefone na transmissão, outro na recepção, e uma linha passiva (sem amplificadores ou outros equipamentos). Assim, um modo de avaliação da qualidade da comunicação telefônica era simplesmente ver se o volume da voz no lado da recepção era satisfatório. Essa medida de **sonoridade** visava basicamente medir quanto uma evolução no terminal telefônico permitia aumentar a distância entre os terminais. Como uma conseqüência disso, a medida da qualidade de um terminal telefônico era expressa em "milhas-padrão" (*Mile Standard Cable*, msc), indicando que o sistema sob teste equivalia a um sistema de referência com uma linha com esse número de milhas.

SFERT

Em 1928, a AT&T Co. americana construiu um novo sistema de avaliação de qualidade de terminais telefônicos baseada na qualidade da conversação (aérea) entre duas pessoas distantes de 1 metro uma da outra (medida de índice de sonoridade). O sistema era composto de um sistema transmissor e receptor e de uma linha artificial[2], montada com componentes estáveis de alta qualidade, mas a característica em freqüência do sistema era altamente arbitrária devido às respostas do microfone capacitivo da parte transmissora e da bobina móvel da recepção, que variavam enormemente então. Esse sistema, obviamente melhor que o anterior, foi adotado como Sistema de Referência pelo organismo precursor do CCITT para telefonia, o CCIF, e foi denominado *SFERT (Système Fondamental Européen de Référence pour la transmission Telephonique)*.

Com o SFERT era calculado um número em dB, denominado Equivalente de Referência (ER), para um sistema sob testes, tal que, com esse número adicionado ou subtraído nos atenuadores do Sistema de Referência, a sonoridade avaliada na Recepção para ambos os sistemas (de Referência e em teste) fosse a mesma.

ARAEN

Em 1949 o CCITT padronizou um outro Sistema de Referência, chamado *ARAEN (Appareil de Référence pour la détermination de l' Affaiblissement Équivalent pour la Netteté)*. O objetivo era calcular o índice de articulação baseado na qualidade conversacional de uma boca distante 1 metro de um ouvido, sendo considerado apenas um caminho monoaural (audição por um ouvido, ao invés dos dois, que melhor corresponde a uma conversação utilizando-se o aparelho telefônico). O sistema era composto de uma parte transmissora, de uma linha artificial e de uma parte receptora, sendo os níveis monitorados no início da linha artificial por um "medidor de volume" (*volume meter*). Uma variante desse sistema, surgido com a inclusão pelo CCITT na rede telefônica da filtragem passa faixa (300–3400 Hz) na rede telefônica, foi

[2] Linha artificial é um circuito, normalmente passivo com resistores e capacitores, que visa simular uma linha de transmissão.

chamado de SRAEN (*Système de Référence pour la détermination de l' Affaiblissement Équivalent pour la Netteté*).

Assim, a principal diferença entre o ARAEN e o SFERT é que o SFERT mede a sonoridade e o ARAEN/SRAEN, a articulação.

NOSFER

Com a conclusão de que o ARAEN não se adequava ao uso como ferramenta de planejamento da rede telefônica, devido à sua não-linearidade, decidiu-se voltar a utilizar o índice de sonoridade. Porém, o SFERT (que era o sistema então existente para medidas de sonoridade) foi considerado obsoleto na época (1960), devido principalmente à descontinuação da fabricação de partes e componentes essenciais, além de reposições não serem mais disponíveis no mercado.

Estudou-se então que mudanças deveriam ser feitas no ARAEN de modo a permitir que ele servisse de substituto para o SFERT, inclusive gerando medidas consistentes com este. Identificadas as mudanças, o novo sistema foi chamado de NOSFER (*Nouveau Système Fondamental pour la détermination de l' Equivalent de Référence*). Esse sistema incorporou partes do ARAEN, mas mudou a posição de referência do locutor, introduziu amplificadores e equalizadores para mudar a resposta em freqüência do ARAEN para a do SFERT. A sistemática de realização dos testes conversacionais também foi mudada (nível de leitura, etc), de modo a adaptar os resultados aos anteriormente obtidos pelo SFERT, e se encontra descrita em [83].

4.2.2 Testes de Opinião

Com o início da introdução de sistemas mais complexos, os testes conversacionais deixaram de ser interessantes no tocante ao quão completos e controláveis eles poderiam ser. Assim, passou-se a utilizar cada vez mais testes utilizando-se um grande número de pessoas ouvindo um determinado material de voz.

Como o desejável para esse extenso material de voz é que ele sempre seja processado pelo mesmo equipamento sob teste e que os materiais de voz sejam sempre os mesmos para todos os ouvintes, é interessante que este material de voz esteja gravado de algum modo.

Com as tecnologias existentes antes do advento do gravador de áudio de rolo (década de 50–60), tais testes eram impossíveis. Com a popularização dos gravadores de alta qualidade, testes subjetivos de opinião começam a ser feitos, tornando-se mais populares no final dos anos 60 e início da década de 70. Mas ainda havia o problema de manejo do material processado gravado: cada sessão de audição, com um novo ouvinte, deve ser apresentada numa seqüência diferente, aleatorizada ("randomized"), o que é extremamente trabalhoso de ser feito sem introduzir degradações adicionais indesejadas no material a ser ouvido e avaliado. O advento nesta década de sistemas comerciais baratos para o armazenamento de sinais digitalizados de voz fez com que o uso de testes de opinião se difundisse ainda mais.

Também coincidente com o advento dos gravadores de alta qualidade é a própria evolução da eletrônica, em especial da digital: o advento das comunicações digitais e a introdução pela Bell Labs dos primeiros sistemas de transmissão digital PCM. No início da década de 70, novas

técnicas para redução de taxa para transmissão começam a ser estudadas, fazendo um uso cada vez mais intensivo de técnicas de processamento digital de sinais.

Nos anos 80, o CCITT promove testes internacionais para a qualificação de um padrão para codificação a 32 kbit/s, e é qualificado o esquema ADPCM proposto pela AT&T. Para isto, são realizados extensivos testes subjetivos formais em 7 países [84, 81].

Em 1987 e 1988, busca-se padronizar na Europa um codec para uso em rádio móvel digital celular num contexto pan-europeu (visando a integração econômica em 1992). Promovem-se então diversas fases de testes subjetivos, uma em âmbito nacional ("National Pre-selection") para a pré-seleção de 1 candidato por país e três de âmbito multinacional: a primeira ("Selection"), para classificar os candidatos; a segunda ("Optimization"), para escolher um candidato de compromisso (os dois melhores classificados na primeira dessas duas fases foram "misturados", gerando um único codificador); e a terceira ("Final Verification and Characterization"), para verificação final do codec escolhido na fase anterior e caracterização de seus diversos aspectos (mesmo aqueles considerados desejáveis mas não imprescindíveis) [85].

Em 1990 e 1991, o CCITT promoveu amplos testes subjetivos formais (em duas fases) para seleção de um codec a 16 kbit/s, processo que envolveu diversos países, incluindo pela primeira vez o Brasil. Foram recentemente realizados testes subjetivos para o codec para a segunda geração de rádio móvel digital celular pan-europeu (taxa líquida de 6.5 kbit/s) e em breve também serão realizados testes subjetivos para a identificação (padronização) de um codificador a 8 kbit/s pelo CCITT. Estes são somente alguns exemplos da aceleração do uso de metodologias de teste de opinião para a avaliação de sistemas de codificação de voz no cenário mundial.

4.3 Projeto de um teste

O projeto de um teste subjetivo bem feito é o passo mais importante para uma avaliação com sucesso do desempenho de um codec.

Uma análise interessante das características desejáveis para um plano de testes que deva ser realizado por diversos laboratórios independentes é esboçada em [81]. Um teste pode ser ponderado por três qualidades ou características:

- *conveniência*, isto é, a facilidade com que o teste pode ser implementado em um determinado local em função da infra-estrutura e metodologias existentes.
- *validade*, que é a capacidade do teste em fornecer respostas coerentes com a qualidade realmente encontrada (bons circuitos apresentarem boas notas, etc).
- *repetibilidade*, isto é, que diferentes realizações desses mesmos testes produzam os mesmos resultados.

A busca de padronização de metodologias de testes subjetivos vem de longa data [86] e encontra sua principal razão neste último item citado. A busca de procedimentos que apresentem uma boa repetibilidade mas mantendo um grau de conveniência à sua implementação é um compromisso entre procedimentos rígidos (que aumentam a repetibilidade, pois restringem o número de fatores que contribuem para a variação das respostas) e flexíveis (permitindo

que a infra-estrutura existente seja usada: material de voz, ambiente de gravação e audição, escolha dos ouvintes e outros detalhes), o que permite uma maior aceitação pela comunidade internacional dos procedimentos escolhidos.

Por outro lado, há muitos aspectos em jogo e que são determinados pelas aplicações pretendidas para o codec. Assim, para cada aplicação, um conjunto distinto de considerações devem ser feitas, considerações essas que em muitos casos são interdependentes.

De um modo geral, um plano de teste subjetivo deve especificar todas as etapas necessárias para se chegar à conclusão sobre o desempenho do codec, que são, de uma maneira geral: as condições que serão testadas, qual o material fonte a ser utilizado, qual o tipo de teste, como deverá ser feito o processamento (bem como as interfaces necessárias), como devem ser ministradas as audições do material processado, como devem ser coletadas as notas, como devem ser selecionados os ouvintes e, finalmente, qual deve ser a análise e a interpretação das estatísticas obtidas.

Nas seções a seguir, falaremos em linhas gerais sobre cada um desses tópicos. Muitas opções são mostradas e todas elas devem ser explicitadas o mais detalhadamente possível no Plano de Testes a ser elaborado, de modo a haver a menor margem possível para discrepâncias em diferentes realizações do(s) experimento(s) especificado(s) no Plano.

4.4 Tipos de teste

Os testes subjetivos podem ser divididos basicamente em três grupos: os conversacionais, os de locução e os de audição [83].

Os *testes conversacionais*, como citados anteriormente, envolvem duas pessoas que tenham sido especificamente treinadas, uma falando e a outra ouvindo (teste bidirecional "half-duplex"), dentro de uma certa metodologia. Como exemplo (utilizado na *British Telecom Research Laboratories, BTRL*), pode ser definido um conjunto de figuras numeradas comum a ambos os participantes. A tarefa consiste então em descrever uma das figuras, escolhida ao acaso, ao outro avaliador. Este, por sua vez, deve identificar qual a figura em questão, computando-se o tempo para tal. Após várias realizações do teste, calcula-se a o tempo médio necessário para a transmissão da informação, do qual se infere a qualidade do sistema.

Os *testes de locução* servem para a avaliação da perda de eficiência da comunicação telefônica devido à dificuldade do locutor em falar e.g. em decorrência do efeito local ou do eco. São testes unidirecionais em que o material de voz é gerado pelo próprio locutor, utilizando-se um texto padrão. Então, variam-se parâmetros de transmissão, como efeito local ou eco, e avalia-se, de acordo com uma escala, a dificuldade do locutor em falar.

Já os mais importantes testes subjetivos são os *testes de audição*, sendo sobre eles que esta seção discorrerá. Os testes de audição são testes unidirecionais que visam medir a capacidade de informação de um sistema. Baseiam-se na avaliação da sua qualidade (de acordo com uma escala apropriadamente escolhida), a partir da audição de sentenças simples, sem relação semântica entre elas, processadas pelo sistema sob avaliação e por sistemas (condições) de referência.

Há nestes testes dois objetivos. O primeiro, mais geral, é de obter avaliações que permi-

Nota	Opinião de Qualidade
5	A qualidade é excelente (muito boa)
4	A qualidade é boa
3	A qualidade é razoável
2	A qualidade é pobre
1	A qualidade é ruim (muito pobre)

Tabela 4.1: Escala de Qualidade Absoluta.

Nota	Opinião de Qualidade
A	A qualidade do primeiro é maior que a do segundo
B	A qualidade do segundo é maior que a do primeiro
C	A qualidade de ambos não se distingue

Tabela 4.2: Escala para Comparação de Pares.

tam predizer o grau de satisfação de usuários de um sistema cujos canais de comunicação sejam simétricos (isto é, tanto o caminho de ida e o de volta possuam características semelhantes, como ganho, resposta em freqüência, entre outras). O outro, mais específico, é de que os efeitos relativos de diferentes variedades de um tipo de degradação sejam graduados corretamente.

Há uma grande diversidade de testes de audição conhecidos. Atualmente os mais utilizados são três: comparação de pares, de categorias e de degradação.

Testes de comparação de pares (*Pair Comparison*) [87, pp.555-560] são realizados pela comparação sistemática de todo o material processado para um teste subjetivo (incluindo âncoras como a comparação contra referências). Nestes testes, os avaliadores usam uma escala ternária (ver tabela 4.2) para externar sua avaliação. Em um teste de comparação de pares com N condições, são necessárias N(N-1) comparações; se a comparação de A com B puder ser considerada *a priori* idêntica à de B com A, então esse número pode ser reduzido à metade. A deficiência desse tipo de teste é que somente desempenhos relativos podem ser obtidos. Porém, em certos casos, é o método mais adequado, quando o desejado é o desempenho relativo de um sistema em relação a outro. Como um exemplo, temos os testes subjetivos realizados para a recomendação CCITT para codificação de voz em faixa larga (7kHz), G.722, quando a comparação de desempenho foi feita em relação ao padrão CCITT para 32 kbit/s (G.721) [88].

Testes de categorias, que no jargão CCITT são chamados de ACR (*Assessment Category Rating*) [89], se baseiam na avaliação "absoluta" da qualidade do material processado usando uma escala como a da tabela 4.1. Um exemplo desse tipo de teste é o da Fase II da Padronização de um codec a 16 kbit/s pelo CCITT [90].

Já os *testes de degradação*, chamados no jargão CCITT de DCR (*Degradation Category Ra-*

Nota	Degradação
5	A degradação é inaudível
4	A degradação é audível mas não incomoda
3	A degradação incomoda um pouco
2	A degradação é incomodante
1	A degradação incomoda muito

Tabela 4.3: Escala de Degradação.

ting) [89], baseiam-se na avaliação da degradação relativa do material processado em relação ao material original (e.g., um arquivo processado pelo MNRU com Q=20 dB é apreciado pelo avaliador em conjunto ao mesmo arquivo sem qualquer processamento; em relação a esse par de arquivos, ele expressa o grau de degradação percebido); é, portanto, uma classe especial de comparação de pares. Utilizam uma escala como a da tabela 4.3, como aconteceu nos testes sobre a "Dependência com o Locutor" (Experimento 4) da Fase I da Padronização de um codec a 16 kbit/s pelo CCITT [91].

4.5 Tamanho do teste

O tamanho de um teste está relacionado a dois aspectos: o número de condições que se deseja testar e a confiabilidade que se deseja para o teste. Isto definido, pode-se determinar o número de ouvintes necessário para o teste.

4.5.1 Número de Condições

Obviamente o número de condições a serem testadas não é uma condição de contorno muito flexível, pois o que necessita ser caracterizado tem que ser caracterizado. Entretanto, o modo como essas condições são arranjadas no processo de avaliação de um codec pode permitir uma substancial redução no tamanho dos testes. Isto é feito alocando-se um conjunto de condições afins dentro de um teste específico, outro conjunto em outro teste, de modo que o conjunto de todos esses testes abranja todas as condições que se deseja testar.

Como um exemplo, considere-se a Fase I da Padronização do LD-CELP, o codec a 16 kbit/s padronizado pelo CCITT [91]: ela consistiu de um conjunto de 5 testes subjetivos completos (estanques em si mesmos), cujas informações eram complementares: o Experimento 1 media o "Efeito de Erros de Transmissão e Níveis de Audição"; o Experimento 2, o "Efeito de Múltiplas Transcodificações, Níveis de Entrada no codec e Níveis de Audição"; o Experimento 3, "Desempenho com Música e Ruído Ambiental"; o Experimento 4, "Dependência com o Locutor"; e o Experimento 5, "Desempenho com Sinalização".

Esse tipo de abordagem deve representar um compromisso de dois parâmetros: a comodidade do avaliador e a abrangência do teste.

Por *comodidade do avaliador*, queremos dizer que quanto menor o teste, melhor será a

qualidade da avaliação, pois o teste será menos extenso e menos cansativo, desde que uma quantidade suficiente de material de voz seja apresentada. Deve ainda conter um "continuum" de qualidade (isto é, deve haver condições de referência presentes representando de forma balanceada todos os níveis aos quais devem ser atribuídas notas de 5 a 1, no caso das tabelas 4.1 e 4.3), para que o ouvinte não seja tendenciado a opinar numa região fora da média da escala usada (nos casos citados acima, a nota 3).

Já a *abrangência do teste* se refere às condições que, por uma questão de segurança, devem estar incluídas dentro de um mesmo teste, de modo que possam ser comparados com uma baixa margem de erro. A origem deste problema se remete à dificuldade em se comparar diretamente resultados sobre um mesmo codec obtidos em testes diferentes, mesmo se realizados com um mesmo material de voz ou com os mesmos ouvintes. Caso o material de voz seja o mesmo, diferentes avaliadores deverão ser escolhidos; isto deixa margem a se utilizar amostras diferentes e não necessariamente representativas da população. Por outro lado, mantendo-se os mesmos avaliadores, deve-se mudar o material de voz, o que potencialmente altera o conteúdo fonético processado. Assim, testes estanques possuem, do ponto de vista teórico, potenciais problemas para a sua utilização como um conjunto. À parte destes aspectos teóricos, na prática é muito difícil reproduzir as mesmas condições de teste em esquemas de avaliação diferentes, pois podem ocorrer erros experimentais diversos [3]. Frente a esses impedimentos, sempre que determinadas condições terão que ser comparadas entre si, elas deverão constar de um mesmo teste. A quantidade de diferentes condições num mesmo teste determina, então, a abrangência do teste.

4.5.2 Confiabilidade do Teste

A confiabilidade de um teste é a margem de segurança com que um teste subjetivo pode predizer a qualidade de um codec. É função basicamente do número de graus de liberdade do teste e do tipo de distribuição das notas do teste.

A referência [92, pp.293-294] define "... os *Graus de Liberdade* de um modelo para os valores esperados de variáveis aleatórias ... [como] ... o excesso do número de variáveis sobre o número de parâmetros no modelo." Como um exemplo, se N votos foram colhidos, o número de graus de liberdade (geralmente referido como ν) é N–1 para a média (pois é um modelo de primeira ordem, onde o único parâmetro é a média) e N-2 para a variância (é um modelo de segunda ordem, i.e., com dois parâmetros, pois compreende a média e a variância).

A *distribuição* depende do tipo de análise que se deseja efetuar. Entretanto, de modo a se garantir uma margem de segurança alta, na maioria dos casos utiliza-se um grande número de amostragens (votos). Em conseqüência do grande número de amostras utilizado, a maioria das distribuições de interesse pode ser considerada uma boa aproximação da curva normal. Por isso, as expressões e tabelas relativas à distribuição normal são em geral aplicadas.

O número de votos necessários para se obter uma confiabilidade α para a média (MOS) de um teste cuja distribuição possa ser considerada normal é [93, p.229]:

$$N = (\frac{\tilde{z}\frac{\alpha}{2}}{\epsilon})^2.\sigma^2$$

[3] Note-se que se um erro sistemático ocorrer em um teste, a análise das estatísticas nesse teste pode ser feita de modo que os efeitos desse erro sistemático sejam isolados. Já o mesmo não pode ser garantido entre testes diferentes.

onde:
$z_{\frac{\alpha}{2}}$ é o valor encontrado nas tabelas de distribuição Normal para um nível $\frac{\alpha}{2}$;
σ^2 é o valor estimado da variância do teste *como um todo*;
ϵ é o desvio padrão desejado para cada uma das amostras *individuais* durante o teste.

Outra abordagem é a de tentativa-e-erro: estima-se um tamanho N para o teste e verifica-se o grau de confiabilidade obtido utilizando-se tabelas e a expressão acima.

De qualquer modo, as estimativas de σ^2 e ϵ se baseiam em haver um prévio conhecimento do comportamento do teste e de suas condições, de modo que previsões realistas possam ser feitas. De uma maneira concreta, σ^2 e ϵ somente serão conhecidos após realizado o teste.

4.5.3 Número de ouvintes

Uma vez definido o número de condições e um número mínimo de votos necessário, encontra-se o número de ouvintes necessário para o teste.

Esse número pode ser encontrado pela divisão do número de votos pelo número de condições do teste, porém isso não é adequado. Seria mais interessante que cada ouvinte (avaliador) pudesse ouvir todo o material de voz (isto é, ser exposto a todo o material fonético gravado para o teste), porém por distintas condições de processamento em relação aos outros ouvintes. Como um exemplo, suponhamos um teste com 4 condições de teste, numeradas de C1 a C4, e 4 ouvintes, numerados L1 a L4. Com 4 elementos (e.g. arquivos de voz, cada um com duas sentenças curtas, todas distintas entre si), numerados F1 a F4, processados pelos 4 fatores, geraremos 16 materiais processados: F1.C1, F1.C2, ..., F4.C4. Utilizando-se algum esquema de aleatorização *(randomization)*, atribui-se por exemplo a seguinte seqüência de audição:

Ouvinte	Seqüência de audição			
L1	F1.C1	F2.C3	F3.C4	F4.C2
L2	F2.C2	F1.C4	F4.C3	F3.C1
L3	F3.C3	F4.C1	F1.C2	F2.C4
L4	F4.C4	F3.C2	F2.C1	F1.C3

Pode-se verificar que L1 ouvirá todas as condições e todos os materiais de voz, bem como o farão L2, L3 e L4. No entanto, em nenhuma situação o mesmo par "*material de voz* × *condição de processamento*" será ouvido mais que uma vez.

De uma maneira mais geral, o ouvinte $\mathcal{X}$ ouviria uma sentença $\mathcal{L}$ processada pela condição C1; o ouvinte $\mathcal{Y}$ ouviria a mesma sentença $\mathcal{L}$, porém processada por uma outra condição, de modo que as matrizes "*material por ouvintes*" e "*processamento por ouvintes*" sejam ambas linearmente independentes. Desse modo, quadrados greco-latinos poderão ser utilizados nas sessões de testes subjetivos (ver secção 4.7.2). De fato, a tabela acima é um dos possíveis quadrados greco-latinos de ordem 4.

De uma maneira simples, essa alocação de audição dos materias pode ser feita em se encontrando um número de ouvintes que seja um *múltiplo* $\mathcal{R}$ do quadrado do número de condições de teste, de modo que o número de votos resultantes seja maior que o número mínimo necessário de votos[4]:

[4] Isso é válido para testes sem intercalamento ("interleaving"); se o intercalamento for utilizado, o número de ouvintes deverá ser proporcional ao quadrado da razão entre o número total de condições e do nível de intercalamentos.

$$\text{No.de Ouvintes} = \mathcal{R} \times (\text{No. de Condições})^2$$

onde

$$\mathcal{R}(\text{inteiro}) \times \text{No.de Ouvintes} \geq \text{No. Mínimo de Votos}$$

4.6 Geração do Material para Avaliação Subjetiva

4.6.1 Organização do Material de Voz

Uma vez definido o número mínimo de votos necessário e o número de condições a serem testadas, fica simples definir o material de voz a ser utilizado, baseado na descrição a seguir.

Estrutura dos Elementos: A estrutura do material gravado deve consistir de sentenças curtas (em torno de 3 a 4 segundos) e de sentido (significado) simples, de modo que o ouvinte não tenha que se esforçar para compreender as sentenças, minimizando assim um possível fator adicional que estaria envolvido no processo de avaliação. A escolha de tais sentenças deve ser tal que nenhuma delas seja repetida para um mesmo ouvinte, de modo que todo o material de voz a ser analisado nas sessões de testes subjetivos seja inédito, quando de sua audição, para todos os avaliadores.

Essas sentenças devem estar agrupadas num conjunto que garanta ao avaliador a audição de uma quantidade de voz mínima porém suficiente para que ele possa formar um conceito sobre a qualidade do processamento que será efetuado sobre esse material. Em geral, tem-se mostrado adequado agrupar duas sentenças (como as descritas no parágrafo anterior), gerando materiais de duração em torno de 7 segundos (incluindo uma pausa de aproximadamente 1 segundo entre sentenças) [91, 90]. Esse conjunto de 2 sentenças é chamado de *elemento*, pois se constitui na unidade básica de processamento (o mais curto material a ser processado). Os elementos são geralmente armazenados na forma de um arquivo de voz digitalizada. Um exemplo de sentenças agrupadas em elementos encontra-se no Anexo A.

Listas: Esses elementos em geral são agrupados em *listas*, sendo o número desses elementos igual ao número de variantes de um dos parâmetros do teste. Em geral, esse parâmetro é o número de níveis de audição a ser utilizado no teste. Como um exemplo, um teste com três níveis de audição apresentaria três elementos por lista.

Replicações: Se o material fonte alocado a um teste não tiver um número de listas igual ao número de ouvintes, a razão entre o número de ouvintes necessários e o número de listas deverá ser um número inteiro. Esse número inteiro é usualmente chamado de *número de replicações* $\mathcal{R}$ do teste e é o mesmo $\mathcal{R}$ usado na seção 4.5.3. Replicar significa que o teste será repetido um certo número de vezes para conjuntos diferentes de ouvintes, até que o número mínimo de votos (para um dado α) seja atingido. Isso implica que o material processado será ouvido de maneira única somente dentro de um conjunto de ouvintes de número igual ao de listas e que o experimento será repetido para um conjunto diferente de ouvintes tantas vezes quantas

forem as necessárias para se atingir o número de votos necessário para a potência escolhida para o teste. Assim, a técnica de replicação é um compromisso entre manter a potência de um teste, enquanto limitando a demanda sobre os ouvintes (número de votos e tempo dedicado à avaliação) e o tamanho do material de voz necessário.

Em resumo, dado um número de ouvintes e de condições[5], pode-se definir o número de listas como sendo o número de ouvintes ou de votos (no caso de replicações, de um sub-múltiplo destes) multiplicado pelo número de condições:

$$\text{No.de Listas} = \text{No.de Ouvintes} \times \text{No.de Condições} \div \mathcal{R}$$

4.6.2 Gravação do Material Fonte

Quando da geração do material fonte para ser processado e avaliado em testes subjetivos, dois são os tipos básicos de armazenamento: a gravação analógica e a gravação digital. Estas são descritas em mais detalhes no Capítulo 3.

Independentemente do modo de armazenamento dos sinais, deve-se garantir que um sinal limpo seja fornecido aos dispositivos de armazenamento. Como parte disso, o sinal fonte deve ser gravado num ambiente controlado acusticamente, como também descrito no Capítulo 3.

Cada dia mais e mais são utilizadas as técnicas de armazenamento digital. Por isso, detalharemos a seguir dois tópicos que permitem digitalizações de maior qualidade.

Digitalização de sinais

O nível ativo do sinal de entrada em conversores A/D deve ser cuidadosamente escolhido de forma a garantir um bom uso da faixa dinâmica do conversor. Sinais digitalizados com níveis muito baixos tendem a apresentar pouco uso dos bits mais significativos, reduzindo na prática a resolução do sinal de 16 para e.g. 11 bits.

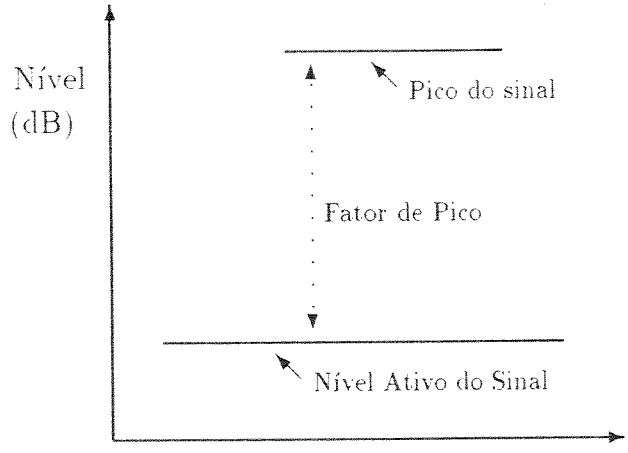


Figura 4.1: Conceito de fator de pico.

[5] E possivelmente de intercalamentos.

Por outro lado, se um nível ativo demasiado alto é escolhido, começará a ocorrer um número excessivo de 'clips' no sinal digitalizado, que é uma distorção não-linear de grande significância.

Estudos realizados pelo CCITT por longos anos com as principais línguas indo-européias indicaram que um nível (ativo) de gravação entre 22 e 23 dB abaixo do ponto de saturação do sistema de digitalização é um bom compromisso entre um bom uso da faixa dinâmica do conversor e no número de amostras saturadas[6]. Como o português é uma língua indo-européia, recomenda-se utilizar esse nível para a digitalização de sinais de voz.

Esse nível é resultado do fator de pico (ver figura 4.1), tipicamente em torno de 19 a 20 dB, mais uma margem de segurança de 3 dB.

Outra medida necessária é o ajuste do nível DC presente no conversor (quer interno, i.e., devido ao circuito de entrada do conversor, ou externo, i.e., associado aos equipamentos utilizados externamente ao sistema de digitalização). As calibrações necessárias para diminuir o nível DC devem ser feitas sempre que possível antes da digitalização, pois níveis DC diminuem a faixa dinâmica vista pelo conversor, além de aproximar os sinais do ponto de saturação do sistema[7].

O uso da sobre-amostragem

Quando se escolhe uma largura de faixa para o sinal de voz digitalizado, o mais natural é pensar em se utilizar uma taxa de amostragem igual à freqüência de Nyquist, isto é, o dobro da faixa do sinal.

Porém, diversos sistemas profissionais de alta qualidade utilizam técnicas de sobre-amostragem *(oversampling)*. Isto pode ser recomendado por diversas razões:

- Taxas mais altas reduzem a necessidade da correção da distorção $\mathtt{sinc(x)}$[8];
- Maior relação sinal/ruído de quantização (e.g, sobre-amostragem de 2 vezes implica numa S/R_q 3dB maior);
- Possibilidade de uma filtragem analógica menos rigorosa na faixa desejada (o que propicia um projeto de filtro que introduzirá menor distorção de freqüência no sinal amostrado), conquanto que a redução de faixa e de taxa seja feita a posteriori via software (com filtros FIR ou IIR [11, Cap.3]).

Um exemplo de sobre-amostragem é o material digitalizado para os testes subjetivos do LD-CELP, onde um fator de 2 vezes foi utilizado (i.e., taxa de amostragem em 16 kHz). Um exemplo comercial é o de toca-discos laser e DATs com sobre-amostragem de 4 ou 16 vezes.

Por outro lado, a desvantagem da sobre-amostragem é de se gastar mais espaço em disco, pois o aumenta-se o tamanho dos arquivos pelo fator de sobre-amostragem.

---

[6] Os recentes testes subjetivos do LD-CELP conduzidos pelo CCITT indicaram que diferente fatores de pico são encontrados para o japonês, onde 3 dB adicionais são necessários para se evitar um número excessivo de 'clips'.

[7] Adicionalmente, processamentos por diversos algoritmos (e.g., MNRU [94]) podem apresentar resultados errôneos se um nível DC muito grande estiver presente.).

[8] Esta correção é necessária para compensar o fato da reconstrução analógica não ser feita com impulsos, mas com degraus [17, pp.92–94].

Um bom compromisso para as aplicações em telefonia é de se utilizar um fator de 2 vezes, i.e., uma taxa de amostragem de 16 kHz, para posterior filtragem e sub-amostragem *(down-sampling)* via software. Uma outra opção seria a de se implementar a filtragem em linguagem assembly de um DSP e embutir no sistema de aquisição; neste caso, o "excesso de dados" existiria somente até que a filtragem fosse feita, sendo salvos para disco somente os dados pós-filtrados e sub-amostrados.

4.6.3 Processamento do Material Fonte

Os processamentos a serem realizados sobre o material fonte para gerar as condições de teste e de referência a ser utilizado no teste subjetivo devem estar bem definidos no Plano de Testes. Alguns pontos fundamentais são delineados a seguir.

Os sinais de entrada podem estar armazenados no formato analógico ou digital. Para o processamento analógico, os níveis de reprodução ("saída"), bem como as impedâncias de circuito, devem estar propriamente ajustados. Os sinais após os processamentos devem igualmente ter os seus níveis re-equalizados para o armazenamento do sinal com a mínima distorção possível. Um exemplo de ajuste de níveis é se utilizar -20 dBm como nível nominal de processamento e ter todas as impedâncias em 600Ω, balanceadas [91].

Outros pontos importantes ainda a se considerar são a necessidade de um circuito com baixo ruído e baixa distorção (SINAD da ordem de 80 dB) e alta estabilidade com o tempo.

No caso da realização de um experimento com diversos laboratórios envolvidos na aplicação de sessões de testes com diferentes materiais de voz, é muito importante que haja somente um laboratório responsável pelo processamento de todo o material (Laboratório Central, ou "Host Laboratory"), de modo a garantir igualdade de condições de processamento e minimizar o risco de erros não-sistemáticos, pois os erros sistemáticos de processamento podem ter seu efeito isolado quando da análise estatística dos dados e, ainda assim, manter a confiabilidade nos resultados.

Um diagrama esquemático de realização de algumas operações deve estar especificado no teste, como por exemplo, a realização de transcodificações síncronas e assíncronas e o uso e especificação de interfaces comuns.

Além disso, é conveniente que todo o material processado esteja, quando pronto para as sessões de avaliação, todos normalizados num mesmo nível e estejam numa mesma taxa de digitalização (e.g. 8 ou 16 kHz). Isto simplificará os procedimentos de aplicação do teste subjetivo. No caso de de testes com audição em diferentes níveis, a mudança de nível deverá ser feita preferivelmente por um sistema externo, já no domínio analógico, e.g. por um amplificador de ganho ajustável. Com relação à normalização do material processado, deve-se tomar cuidado para normalizar corretamente sinais com alto nível de ruído (e.g. MNRU com Q=5 ou 15 dB) através da P.56, como explicado no Capítulo 3.

Classes de Processamento

O processamento do material-fonte através dos codecs cuja qualidade se deseja avaliar e das condições de referência pode ser classificado em analógico, digital e misto, independentemente da forma de armazenamento do material ser analógica ou digital.

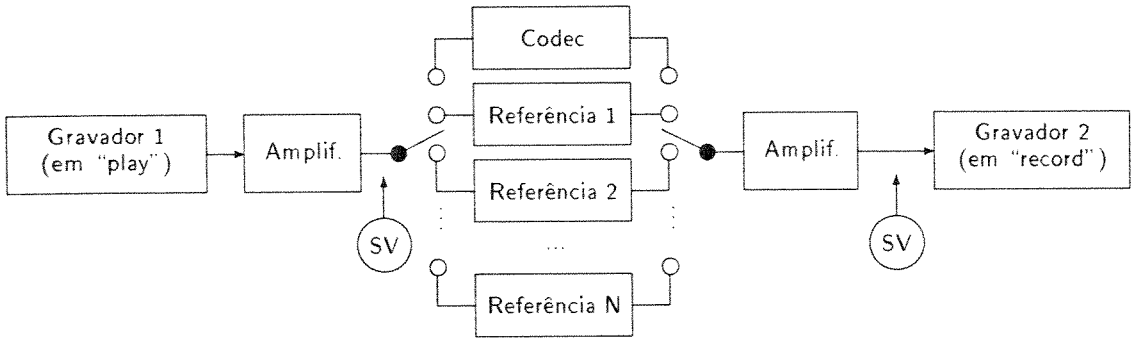


Figura 4.2: Exemplo de processamento puramente analógico.

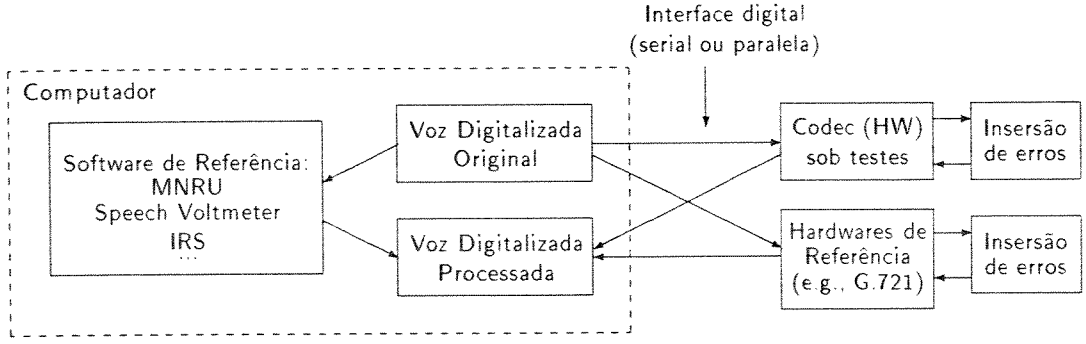


Figura 4.3: Exemplo de processamento puramente digital.

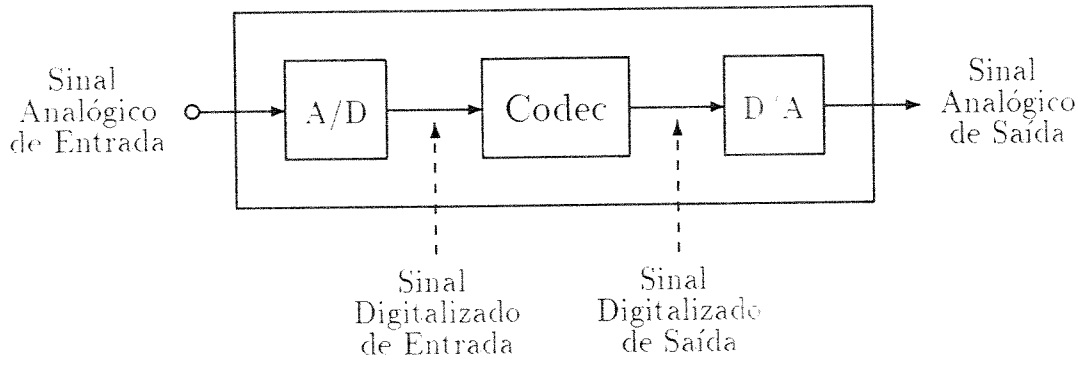
No *processamento analógico* nos referimos a sinais trafegando por um circuito de ajuste de níveis (insersão de perdas ou ganhos) de entrada e saída, no qual se coloca a *condição de processamento*[9] desejada.

Um exemplo de tal configuração está na figura 4.2. Nela pode-se ver que o sinal-fonte está armazenado num gravador (o Sony Betamax, por exemplo), é ajustado em nível (usando o "Speech Voltmeter"), é alimentado em uma das N+1 condições de processamento da figura, o nível de saída é novamente ajustado e o sinal condicionado é armazenado num outro gravador (que poderia ser um computador equipado com um conversor A/D de alta qualidade).

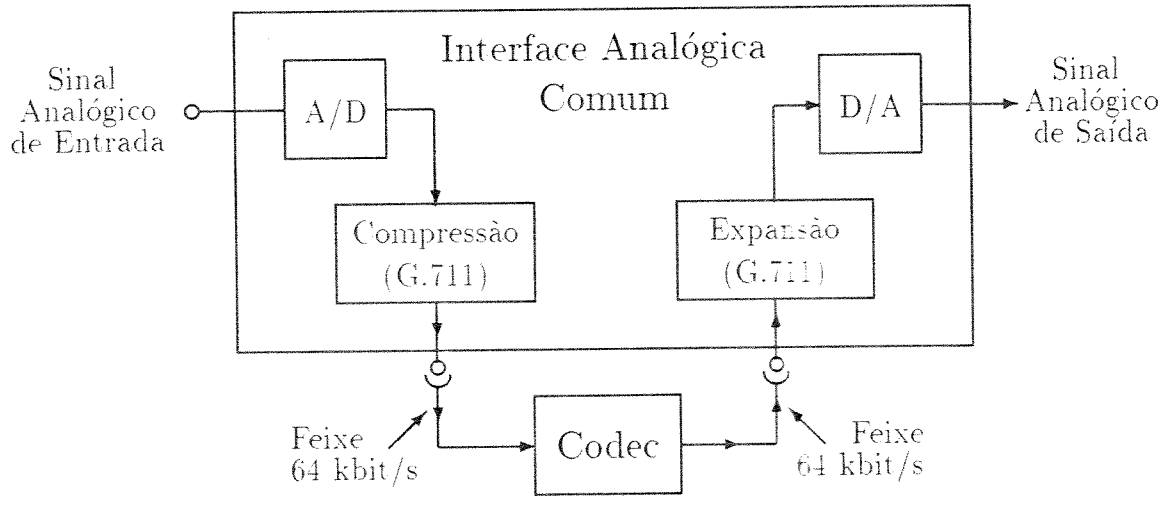
Numa outra abordagem, denominada *processamento digital*, o codec e as referências devem ter uma interface digital diretamente conectada a um computador (ver figura 4.3). Deste modo, não ocorrem conversões A/D e D/A adicionais e toda degradação introduzida nos sinais, a menos daquela própria do processamento, pode ser controlada de um modo muito preciso. Adicionalmente, o controle de ganho, a adição de ruído, etc., são feitos na forma digital.

No processamento analógico da figura 4.2, como os codecs sob teste são processos digitais,

[9] Por "condição" entendemos ou o codec com uma condição de contorno (por exemplo, taxa de erros de 10^{-3} e 1 transcodificação em cascata) ou uma configuração de referência (por exemplo, MNRU com Q=5dB). De um modo geral, deve representar a escolha, dentre todas as as situações de circuito possíveis para um dado sistema, daquelas que apresentem um papel-chave para a descrição (e definição) desse sistema. Em outras palavras, as condições escolhidas devem representar um subconjunto dos fatores que plenamente descreva a situação de circuito real.



(a) uso de interface analógica dedicada (integrada), possivelmente "proprietária".



(b) uso de interface analógica comum (especificação aberta).

Figura 4.4: Interfaceamento de um codec digital em uma abordagem de processamento puramente analógica (entrada e saída da "caixa-preta" são analógicas apesar do codec ser digital). Note que o codec compreende um par codificador-decodificador.

duas são as possibilidades: cada codec ter embutido o seu sistema de conversão A/D e D/A, de modo que a entrada e saída da "caixa-preta" seja analógica; ou então, suas entrada e saída serem digitais (por exemplo, um feixe serial a 64 kbit/s em lei A, segundo a CCITT G.711 [16]), como ilustrado na figura 4.4.

No primeiro caso, o codec sob teste deverá escolher o tipo de digitalização e ter ajustado o nível de entrada do sinal de acordo com regras estabelecidas a priori para o teste, em especial se os diversos protótipos forem oriundos de proponentes diversos.

No segundo caso, os proponentes não precisariam se preocupar com detalhes do interfaceamento analógico do protótipo a ser testado, mas somente com os da parte digital especificada para interfaceamento dos harwares propostos com uma "interface analógica comum". Tal interface com lado digital codificando em lei A ou lei μ de acordo com a CCITT G.711 é especialmente interessante no caso de codecs com aplicação em telefonia convencional, pois este é o padrão de codificação atualmente utilizado pelas centrais CPA, além de ser uma das condições de referência utilizadas em testes subjetivos[10].

Assim, no segundo caso, problemas decorrentes da diferença na qualidade dos conversores de um e outro hardware sob teste não ocorreriam e as comparações de desempenho seriam mais próximas do pretendido, isto é, a qualidade do algoritmo em si. Além disso, faz-se necessário somente calibrar a montagem de processamento para a interface analógica comum, e não para os diversos codecs sob teste. Isto aumenta a confiabilidade e reduz o tempo total de processamento dos materiais de voz. Além disso, a infra-estrutura laboratorial fica mais versátil, pois basta apenas trocar alguns cabos para processar sinais por novos codecs, sem ser necessário se preocupar com atenuadores, amplificadores, se a interface é diferencial ou não, se é balanceada ou não, etc.

Pelo acima exposto, a tendência hoje é de se ter os codecs sob teste com uma interface digital (dentro de um padrão pré-estabelecido para um determinado teste) e o laboratório (central) responsável pelos processamentos ter uma interface analógica comum, como aconteceu para os testes do codec a 16 kbit/s do CCITT[11].

Adicionalmente, ter-se um codec com interface digital torna mais fácil a evolução da metodologia atual (analógica) para metodologias futuras de processamento (puramente digital). Neste caso, não haveria uma interface analógica comum, mas uma interface digital para o computador (como na figura 4.3), de modo aos sinais digitalizados serem alimentados no codec em teste sem a necessidade de degradações adicionais introduzidas por conversões D/A e A/D, deixando a introdução de degradações como mais uma das condições do teste.

Esse tipo de abordagem foi utilizado pela primeira vez nos testes para a seleção de um codificador de segunda geração de telefonia móvel celular digital pan-européia, mas várias questões ainda estão abertas e deverão ser definidas, como: qual a implementação de MNRU que deverá ser escolhida (pois a recomendação CCITT que o especifica é muito vaga e foi moldada em termos de equipamento analógicos)? Como simular o retorno ao domínio analógico que ocorre no caso de múltiplas transcodificações assíncronas (por exemplo, 10 codificadores G.711 em cascata)?. Para esse teste em especial, soluções foram adotadas, mas sua adequação poderá ser verificada somente a posteriori. Apesar disso, espera-se atingir progressos significativos

[10] A G.711 é também utilizada para entrada e saída para o codec CCITT a 40, 32, 24 e 16 kbit/s, G.726 [24], e a 16 kbit/s, G.728 [37].

[11] Haverá uma recomendação CCITT da série P, no futuro, especificando tal interface, nos moldes aqui descritos.

nos próximos anos[12], indicando que essa técnica passará a ser a mais utilizada, ainda mais se considerarmos o barateamento e o aumento de velocidade e da confiabilidade dos recursos computacionais (principalmente memória de massa, como discos ópticos).

Obviamente, pode-se imaginar configurações mistas, quando partes do processamento (por exemplo, o codec) seriam feitas sem conversão para o domínio analógico, interfaceando-se diretamente a um computador, e parte das condições de referência seriam processadas analogicamente (por exemplo, o MNRU). Este tipo de configuração não nos parece ser muito recomendável para fins de testes subjetivos, pois as condições de processamento não serão as mesmas.

4.7 Audição do Material Processado

A audição do material processado envolve quatro pontos: qual o meio de audição a ser utilizado, qual a seqüência de audição que deve ser apresentada para cada ouvinte, como escolher esses ouvintes e quais os procedimentos durante as sessões de audição.

4.7.1 Meio de audição

Para a audição de sinais processados é sempre recomendável o uso de salas isoladas acusticamente, para se evitar a interferência de ruídos externos no processo de avaliação. Adicionalmente, dois aspectos básicos (que não se excluem mutuamente) devem ser levados em conta. O primeiro dos aspectos básicos é a aplicação para a qual se pretende avaliar os algoritmos, o que implica na definição de que faixa de voz será utilizada e de qual o mecanismo básico de interface com o usuário. O segundo aspecto se refere à fase de desenvolvimento do algoritmo em teste.

Em função da definição desses aspectos básicos, escolhe-se um dispositivo de audição adequado. Dois dos mais usuais são descritos a seguir.

Sistemas de alta-fidelidade

Muitas vezes é importante a avaliação do sinal processado por um algoritmo reproduzindo-o através de sistemas com resposta plana (como amplificadores de áudio e caixas acústicas de boa qualidade). Isto é especialmente verdade quando os sinais originais processados pelo algoritmo não hajam sofrido ponderações em freqüência (em geral a filtragem IRS).

Basicamente, esta avaliação é utilizada para verificar o comportamento do algoritmo em situações extremas durante a sua fase de desenvolvimento (nível de audição alto, para ressaltar distorções, e baixo, para medir inteligibilidade), quando se realizam testes informais.

Há outro caso, porém, em que sistemas de alta fidelidade são empregados: é quando a faixa de voz que o algoritmo pretende abranger é larga. Um exemplo dessa aplicação é a vídeo-conferência, em que se utilizam codecs operando em faixa-larga (50-7000 Hz). Neste caso,

[12] Um passo importante para essa definição foi a criação de uma nova questão no CCITT para o estudo de ferramentas de hardware e software dentro da Comissão de Estudos XV. Como passo anterior, foi desenvolvida uma biblioteca de ferramentas de software [11].

os usuários numa vídeo-conferência não utilizarão o monofone padrão da telefonia convencional, pois o sistema opera com mãos-livres ("hands-free"), com microfones e alto-falantes convencionais (de áudio) e monitores de vídeo. Assim, o uso de sistemas de alta fidelidade é necessário.

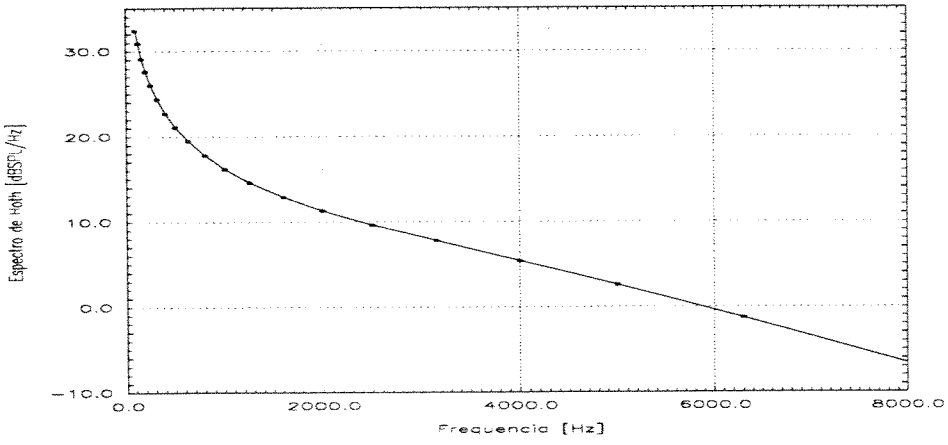


Figura 4.5: Densidade Espectral de Hoth.

Telefones

Quando da realização de testes subjetivos formais, todas as características do sistema devem estar bem definidas.

Do ponto de vista do terminal a ser utilizado para as avaliações, ele deve se adequar tanto às características de faixa de freqüências desejada, bem como às expectativas e preferências auditivas do avaliador.

Como um exemplo, para a qualidade telefônica, especifica-se um telefone com mecânica convencional e que tenha características elétricas seguindo a Recomendação CCITT P.48 (que define o Sistema Intermediário de Referência, normalmente referido como IRS). Além disso, o nível de audição deve ser ajustado de modo a estar no "nível preferido" *(Preferred listening level, PLL)*[13].

Um ponto a ressaltar, especialmente no caso de testes formais, é que como a sala acústica possui um ruído de fundo menor que o do sinal gravado (especialmente se este tiver sido processado), ocorrerá um contraste entre o nível de ruído percebido pelo ouvinte para cada um de seus ouvidos, causando-lhe um incômodo. Para reduzir esses efeitos de contraste de ruído de fundo, recomenda-se adicionar ao ambiente da sala acústica um ruído colorido

[13] Cabe ressaltar que o PLL não é o nível de audição preferido pelos avaliadores individualmente, mas sim um valor único para todos os avaliadores, obtido experimentalmente através de testes subjetivos, sendo especificado para cada teste subjetivo. Como um exemplo, nos testes para um codificador a 16 kbit/s do CCITT, o nível escolhido foi a pressão acústica de 15 dBPa no ponto de referência do ouvido (ERP) [90].

com espectro de Hoth [95, pp.266-267] (ver figura 4.5) a um nível de -68 dBmp através de alto-falantes uniformemente distribuídos na sala.

Adicionalmente, se o nível de ruído de fundo do circuito (analógico) entre o dispositivo de armazenamento e o aparelho telefônico for muito menor que o típico para os sinais processados a serem ouvidos, pode-se acrescentar um pouco de ruído (elétrico) de fundo, de modo a evitar contrastes muito grandes entre trechos de silêncio que ocorrem entre a reprodução de dois *elementos* e dos trechos de silêncio presentes no início e fim do material gravado. O nível desse ruído, entretanto, deve ser cuidadosamente escolhido de modo que, ao ser adicionado ao ruído do próprio material, não mascare aspectos que se deseja analisar no material de audição.

4.7.2 Seqüência de audição

A seqüência de audição deve ser bem escolhida de modo a garantir que, ao final do processo, as avaliações colhidas possam ser consideradas independentes umas das outras (minimizar a correlação entre notas adjacentes). Em outras palavras, deve-se minimizar as interações estatísticas entre fatores independentes (ver análise de variância).

Um modo simples de garantir essa independência é forçar que cada material de voz, processado por uma determinada condição, seja ouvido por um e tão somente um avaliador[14].

Uma estrutura que fornece esse tipo de arranjo é a dos quadrados greco-latinos *(Graeco-latin Squares, GLS)* [96, pp.469–474]. Porém, antes de sua definição, temos que conhecer o que é um quadrado latino.

Por definição, um quadrado latino ([97, pp.507–510], [98]) é uma matriz em que um determinado elemento aparece exatamente uma vez em cada linha e exatamente uma vez em cada coluna de uma matriz[15]. Abaixo são dados como exemplo dois quadrados latinos de ordem 4.

F1	F2	F3	F4
F2	F1	F4	F3
F3	F4	F1	F2
F4	F3	F2	F1

C1	C3	C4	C2
C2	C4	C3	C1
C3	C1	C2	C4
C4	C2	C1	C3

Porém, estes dois quadrados latinos têm uma propriedade especial: se combinados, formarão pares de elementos que aparecem uma e somente uma vez no quadrado combinado. Isto é, combinando cada elemento $a_{i,j}$ do primeiro quadrado latino com o correspondente elemento $\alpha_{i,j}$ do segundo, obtemos um quadrado de elementos $(a_{i,j}, \alpha_{i,j})$, cuja propriedade é:

$$(a, \alpha)_{i,j} = (a, \alpha)_{i',j'}$$

se e somente se $i = i'$ e $j = j'$. Esta é, exatamente, a definição de quadrado greco-latino: um conjunto de dois quadrados latinos em que os n^2 pares ordenados são todos distintos (de modo que todos os possíveis pares ordenados de símbolos envolvidos ocorram exatamente uma vez). Adicionalmente, o número de elementos distintos do quadrado deve ser igual ao número de linhas (ou colunas).

[14] Caso haja *replicações*, generaliza-se para "$\mathcal{R}$ *e tão somente* $\mathcal{R}$ avaliadores".

[15] De uma maneira genérica, um quadrado latino de ordem m é uma matriz $m \times m$ com tal propriedade.

Hoje, o arranjo da seqüência de audição do material processado na forma de um quadrado greco-latino se constitui numa solução clássica adequada a experimentos baseados em linhas e colunas.

Nessa abordagem, as colunas dos quadrados seriam a seqüência de apresentação dos materiais processados e as linhas seriam as sessões de audição (ou, equivalentemente, os avaliadores). Um dos quadrados latinos representaria as listas a serem apresentadas para cada combinação de linha e coluna. O outro quadrado, por sua vez, representaria as condições de processamento (ver um exemplo em [90] ou no Anexo B). Portanto, há quatro *fatores* presentes (avaliadores, seqüência de audição, listas e condições). Num outro tipo de experimento com quatro fatores (e.g. *experimento fatorial* [96]), haveria m^4 elementos a avaliar. Entretanto, quando se usa experimentos baseados em GLS, somente há m^2 avaliações, pois há m fatores de um dos quadrados combinados com os m fatores do segundo quadrado (e.g., m arquivos de voz original processadas por m condições resultam em m^2 arquivos processados). Isto implica numa grande economia quando da realização de testes subjetivos, daí a sua popularidade. Porém há um risco associado a essa redução no tamanho do teste: não pode haver *interação*[16] entre os fatores, senão a qualidade do teste ficará comprometida [99, pg.534].

Uma dificuldade adicional é que a geração de tais quadrados envolve o uso de algoritmos em geral complexos [100, 101, 102, 103]. Uma abordagem comum é o uso de quadrados publicados na literatura [104, 100], bem como sua combinação para obter quadrados de ordem maior [105].

4.7.3 Escolha dos avaliadores

Um aspecto importante a ser considerado quando da realização de testes subjetivos é a representatividade dos avaliadores escolhidos em relação à população a que se destina o algoritmo em teste. Em outras palavras, é importante garantir a representatividade da amostra em relação à população-alvo.

Por isso, cuidados simples porém essenciais devem ser tomados:

- representação equitativa entre homens e mulheres (também crianças, se for o caso), preferivelmente com distribuição homogênea de idade (por exemplo, entre 18 e 65 anos);
- as classes sociais a que se destina o serviço em que o codec será utilizado devem estar representadas homogeneamente;
- todos os ouvintes devem ter um mesmo nível de experiência em termos de avaliação, preferivelmente sendo leigos no aspecto em avaliação. Em nosso caso específico, leigos em processamento digital de voz;
- recomenda-se que os avaliadores não tenham participado de outros testes subjetivos pelo menos nos últimos 6 meses. Em especial para testes de audição, pelo menos um ano de espaçamento é desejável;
- todos devem ter audição normal, conforme testes audiométricos;

[16] Quando há vários fatores combinados num mesmo experimento, pode ser que a combinação de alguns deles resulte em alterações dos resultados globais, em relação à situação em que houvesse somente cada um dos fatores isoladamente. Essa alteração de resultado devido à combinação de fatores é chamada de *interação* entre os fatores [99, pg.473].

- a língua em que se conduzirá o teste determina a língua-mãe de todos os avaliadores;

4.7.4 Instruções

Para que todos os avaliadores partam de um mesmo ponto inicial em termos da expectativa de como será o teste (visto que deve se tratar de pessoas em sua grande maioria leigas), é importante a distribuição a priori de um texto com instruções e alguns detalhes operacionais do teste. Exemplos podem ser encontrados nos Anexos C e D para testes ACR e DCR respectivamente (ver seção 4.4).

4.7.5 Aplicação

Para a aplicação das sessões de audição, são necessários diversos cuidados.

O primeiro deles é que os avaliadores venham para a sessão de audição com algum conhecimento prévio da seqüência de operações que eles virão a fazer. Para garantir um acesso homogêneo às informações operacionais do teste, deve-se distribuir a priori as *Instruções* do item anterior aos avaliadores.

Adicionalmente, uma descrição verbal deve ser feita ao(s) avaliador(es) imediatamente antes de se começarem os testes, ressaltando os pontos importantes do que está contido nas Instruções.

Ainda, todo teste deve conter uma sessão de prática, que têm basicamente dois objetivos:

1. permitir ao avaliador um contacto com um gradiente de qualidade que ele encontrará nas audições subsequentes;

2. permitir também ao ouvinte que ele experencie a "seqüência de operações" para ouvir o material e marcar a nota.

Como essas avaliações servem como referência, os votos referentes serão descartados para efeito de cálculo dos valores MOS do teste.

Tendo os avaliadores passado pela sessão de prática, faz-se uma breve pausa para esclarecimento de eventuais dúvidas. Sendo estas esclarecidas, inicia-se o teste propriamente dito, quando todas as notas coletadas serão de fato utilizadas no cômputo dos valores MOS.

Deve-se lembrar aqui que sessões de audição muito longas causam o cansaço dos avaliadores, levando-os a darem notas de maneira irregular (sessões contínuas de mais de 20 minutos podem ser consideradas *longas*). Para que esse fator seja minimizado, a sessão de audição pode ser quebrada em sub-sessões, de modo que seu número seja um sub-múltiplo inteiro do número de listas a serem ouvidas pelo avaliador e que durem no máximo 20 minutos. Adicionalmente, as sub-sessões devem ser separadas por pausas de no mínimo 5 minutos. Apesar desse procedimento aumentar a duração do teste, ele é essencial para aumentar a sua eficácia.

4.8 Análise dos resultados

Uma vez coletados os dados, procede-se à sua análise. Aqui começa outra tarefa muitas vezes complexa, que se combina com a outra igualmente complexa tarefa de se projetar um teste subjetivo formal: para cada teste, deve ser definida uma análise estatística adequada, de modo a permitir que os aspectos relevantes sejam avaliados e isolados os efeitos que perturbem a análise.

A análise dos resultados deve incluir o cálculo das médias (MOS) e das variâncias, junto com seus intervalos de confiança. Pode ainda incluir uma análise das variâncias (ANOVA) e a conversão dos MOS para Q. Adicionalmente, outras análises específicas podem ser incluídas.

4.8.1 Médias, variâncias e intervalos de confiança

Nessa análise genérica, incluem-se basicamente o cálculo da média das avaliações, sua variância e desvio padrão e o intervalo de confiança dessas estatísticas. Além disso, é sempre seguro calcular a significância dos resultados, aspecto muito importante mas na maioria das vezes não considerado, especialmente em testes informais.

O cálculo do MOS de uma amostra [106, p.6] é feito por[17]:

$$\bar{x} = \frac{1}{N}\sum_{N} x_i, \tag{4.1}$$

onde x_i são as notas coletadas e N é o número total de notas para a condição que se deseja avaliar.

Igualmente, a estimativa da variância da amostra [106, p.10–11] é obtida da relação[18]:

$$\begin{aligned} s^2(x) &= \frac{1}{N-1}\sum_{N}(x_i - \overline{x})^2 \\ &= \frac{1}{N-1}(\sum_{N} x_i^2 - \overline{x}.\sum_{N} x_i) \end{aligned} \tag{4.2}$$

e a estimativa do desvio padrão[19] é dada por:

$$s(x) = \sqrt{\frac{1}{N-1}\sum_{N}(x_i - \overline{x})^2} \tag{4.3}$$

O intervalo de confiança do valor MOS encontrado é, num caso geral, função do nível de confiança $(1-\alpha)$ desejado para as medidas, da variância e do número de amostras utilizadas no cômputo das estatísticas. Se a distribuição das notas puder ser considerada normal (o que normalmente é verdade), então o intervalo de confiança $CI(x)$ [93, p.228] e a mínima diferença significativa [91], $MSD(x)$, são dados por:

$$CI(x) = \frac{z_{\frac{\alpha}{2}}.s(x)}{\sqrt{N}} \tag{4.4}$$

[17] Note que $\bar{x}$ é a média da amostra da população, não da população em si, μ; à medida em que N cresce, $\bar{x} \rightarrow \mu$ [107, pp.52,76].

[18] Note que $s^2(x)$ é uma estimativa da variância, enquanto que a variância de fato da população é σ^2 [107, pp.76–78].

[19] O desvio padrão de fato da população é σ.

$$MSD(x) = \sqrt{2}.CI(x) \tag{4.5}$$

onde $z_{\frac{\alpha}{2}}$ é o valor de distribuição normal para um nível $(1-\alpha)$ de confiança. Por exemplo, para $1-\alpha$=95%, $z_{\frac{\alpha}{2}}$=1.96 (ver exemplo no Anexo B e no Capítulo 6).

A análise de significância[20] da média e variância acima calculadas já é um pouco mais complicada. Este tipo de análise serve para se testar se uma média (ou variância) pode ser considerada não significativamente diferente de uma média (ou variância) esperada ou de referência, ou se aquela pode ser considerada maior ou menor que esta. Isto é feito através do estabelecimento de hipóteses e do cálculo da probabilidade associada a essas hipóteses; se a probabilidade associada à hipótese está acima de um certo limiar, aceita-se, senão rejeita-se (o que implica em aceitar o complemento da hipótese rejeitada). Esse limiar $alpha$ normalmente é escolhido como sendo de 5% (i.e., (1-α)=95%); em casos que se deseja maior ou menor precisão, pode-se utilizar α igual a 1% ou 10%, respectivamente. A significância de médias é feita através do teste t de Student e o das variâncias, pelo teste F de Fisher.

A distribuição de Student é definida pela função de distribuição de probabilidade [108]:

$$f_\nu(t) = \frac{1}{C_\nu}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \tag{4.6}$$

onde ν é o número de graus de liberdade, t é um dos parâmetros do modelo e C_ν é:

$$C_\nu = \sqrt{\nu}\int_0^1 t^{-\frac{1}{2}}(1-t)^{\frac{\nu}{2}-1}$$

Num caso genérico, t tem média 0 (distribuição simétrica) e variância $\sigma^2 = \nu/(\nu-2)$. Para valores grandes de ν (>30), t pode ser aproximada por uma rdistribuição Normal de média 0 e variância 1, $N(0,1)$. Para o caso mais comum, ν se iguala ao número de votos colhidos decrescido de 1.

A significância α das estatísticas é a integral $A(t|\nu)$ da eq. 4.6, que pode ser calculada numericamente ([106, p.61–64] ou [108, p.482–485]) ou retirada de tabelas, dado um valor de t e de ν.

O valor de t varia de acordo com o teste que se deseja realizar, mas para o nível de significância das médias obtidas, é calculado por

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\bar{s}(x_1,x_2)}\left(\sqrt{\frac{1}{N_{x_1}}+\frac{1}{N_{x_2}}}\right)^{-1} \tag{4.7}$$

onde

$$\bar{s}(x_1,x_2) = \sqrt{\frac{(N_{x_1}-1)s^2(x_1)+(N_{x_2}-1)s^2(x_2)}{N_{x_1}+N_{x_2}-2}}$$

e x_1, x_2 são dois conjuntos de amostras que se deseja comparar (e.g. MOS de duas condições de processamento) com N_{x_1} e N_{x_2} amostras (e.g. votos), respectivamente.

De posse dos valores de t e ν, e da probabilidade a eles associada, pode-se testar hipóteses sobre as médias em comparação.

[20] O conceito de significância relaciona-se à margem de segurança para que diferenças entre dois parâmetros (médias ou variâncias) possam ser considerados como comprovadores ou não da hipótese associada.

$A(t|\nu)$ é a probabilidade do t calculado ser menor que um nível $\tau_{\alpha,\nu}$ pré-escolhido (α sendo o nível de confiança) [108]:

$$Pr(|t| < \tau_{\alpha,\nu}) = A(t|\nu) \triangleq 1 - \alpha = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \int_0^{\frac{t^2}{\nu+t^2}} \frac{(1-t)^{\nu/2-1}}{\sqrt{t}} dt$$

onde $\Gamma(z)$ é a função gama definida por [108]:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

$A(t|\nu)$ é a área debaixo da função de distribuição de probabilidade no intervalo $(-t, t)$. Se $A(t|\nu)$ é menor que $A(\tau_{\alpha,\nu}|\nu)$, esta não contém a última, e isto reflete o fato de ser improvável (a um nível α) que a diferença das médias caia na faixa escolhida. Em outras palavras, pequenos valores de $A(t|\nu)$ implicam que há uma alta probabilidade de que a diferença encontrada entre $\bar{x}$ e $\bar{y}$ seja muito significativa.

Se rejeitarmos a hipótese de que a diferença entre $\bar{x}$ e $\bar{y}$ seja *não-significativa* (isto é, considerar que a diferença seja *significativa* ou *altamente significativa*), nós podemos verificar se $\bar{x} > \bar{y}$ ou $\bar{x} < \bar{y}$.

Primeiro, consideremos $Pr(t < \tau_{\alpha,\nu})$, ie, $\bar{x} \leq \bar{y}$. Se $(1-\alpha)$ é a área entre $(-t, t)$, então

$$Pr(t < \tau_{\alpha,\nu}) = (1-\alpha) + \frac{\alpha}{2} = 1 - \frac{\alpha}{2} = \frac{1 + A(t|\nu)}{2}$$

Para $\bar{x} \geq \bar{y}$, nós consideramos $Pr(t > \tau_{\alpha,\nu})$, que leva a:

$$Pr(t > -\tau_{\alpha,\nu}) = \frac{1 + A(t|\nu)}{2}$$

Alternativamente, para $\bar{x} < \bar{y}$,

$$Pr(t < -\tau_{\alpha,\nu}) = \frac{1-\alpha}{2} = \frac{1 - A(t|\nu)}{2}$$

e

$$Pr(t > \tau_{\alpha,\nu}) = \frac{1-\alpha}{2} = \frac{1 - A(t|\nu)}{2}$$

Dessas relações, podemos testar hipóteses usando o procedimento mostrado em [106]:

Se $A(t\|\nu) > \alpha$	.....	então $\bar{x}$ e $\bar{y}$ podem ser considerados iguais com uma confiabilidade α.
Se $\begin{cases} t > 0 \text{ e } A(t\|\nu) < 2\alpha \text{ ou} \\ t < 0 \text{ e } A(t\|\nu) > 2\alpha \end{cases}$	.....	então $\bar{x}$ pode ser considerada maior que $\bar{y}$ com uma confiabilidade α.
Se $\begin{cases} t < 0 \text{ e } A(t\|\nu) < 2\alpha \text{ ou} \\ t > 0 \text{ e } A(t\|\nu) > 2\alpha \end{cases}$	.....	então $\bar{x}$ pode ser considerada menor que $\bar{y}$ com uma confiabilidade α.

Fator	ν	Soma de Quadrados dos Fatores	Variancia de cada Fator	Teste-F(1)
Média	1	$SS_M = (\frac{1}{m}\sum_i \sum_j Y_{ij(kl)})^2$	-	-
Total	m^2-1	$SS_T = \sum_i \sum_j Y^2_{ij(kl)} - SS_M$	-	-
Linhas	$m-1$	$SS_R = \frac{1}{m}\sum_i R_i^2 - SS_M$	$MSQ_R = \frac{SS_R}{m-1}$	$\frac{MSQ_R}{MSQ_E}$
Colunas	$m-1$	$SS_C = \frac{1}{m}\sum_j C_j^2 - SS_M$	$MSQ_C = \frac{SS_C}{m-1}$	$\frac{MSQ_C}{MSQ_E}$
Listas	$m-1$	$SS_L = \frac{1}{m}\sum_k L_k^2 - SS_M$	$MSQ_L = \frac{SS_L}{m-1}$	$\frac{MSQ_L}{MSQ_E}$
Condições	$m-1$	$SS_{Tr} = \frac{1}{m}\sum_l T_l^2 - SS_M$	$MSQ_{Tr} = \frac{SS_{Tr}}{m-1}$	$\frac{MSQ_{Tr}}{MSQ_E}$
Resíduo	m^2-4m+3	$SS_E = SS_T - SS_R - SS_C - SS_L - SS_{Tr}$	$MSQ_E = \frac{SS_E}{(m-1)(m-3)}$	-
(1):	*Todos os testes* F_{ν_1,ν_2} *são* $F_{(m-1),(m-1)(m-3)}$ *e são definidos pela equação (4.8).*			

Onde:

R_i	$= \sum_j Y_{ij(kl)}$	L_k	$= \sum_l Y_{ij(kl)}$
C_j	$= \sum_i Y_{ij(kl)}$	T_l	$= \sum_k Y_{ij(kl)}$

i é o índice da linha do quadrado (sessão)

j é o índice da coluna do quadrado (ordem de apresentação)

k representa um dos fatores independentes (e.g. listas)

l representa o outro fator independente (e.g. condições)

$Y_{ij(kl)}$ é a avaliação recebida pela na i-ésima sessão na j-ésima ordem de apresentação; equivale à k-ésima lista processada pela t-ésima condição.

Tabela 4.4: Análise de Variança para experimentos baseados em Quadrados Latinos Ortogonais de ordem m.

4.8.2 Análise de variâncias (Anova)

A análise de variâncias aplicada a um teste subjetivo serve basicamente para se verificar se houve erros grosseiros na realização do teste. Baseia-se no cálculo da variância dos diversos fatores envolvidos num teste (como ouvintes, locutores, sexo dos locutores, listas, condições de processamento, ordem de apresentação), da variância associada à possível interação entre alguns dos fatores e da variância total [109, pp.69–77]. Uma vez calculadas, o teste F de R.A.Fisher [96] é aplicado para as variâncias de interesse, então se concluindo sobre a significância ou não da variância de alguns fatores e de suas interações.

No caso geral de um teste baseado em quadrados greco-latinos, a ANOVA pode ser feita através do conjunto de equações da tabela 4.4 [110, p.123].

O modelo no qual a Tabela 4.4 se baseia [99] supõe que haja quatro fatores sendo analisados num experimento de ordem m (m ouvintes, m condições, etc) que soma m^2 votos. Tenta-se então avaliar a significância das variâncias em relação ao resíduo (que é a diferença entre a variância total e as variâncias de cada um dos fatores). Para isso é necessário computar a soma dos quadrados para cada um dos fatores, como indicado na coluna 3 da Tabela 4.4. A partir destas, calculam-se as variâncias associadas a cada um dos fatores (coluna 4). Essas variâncias são usadas para se computar as razões F de Fisher (eq. (4.8)), com as variâncias da coluna 5 da Tabela 4.4. Com os valores de $F_{(m-1),(m-1)(m-3)}$ computados, pode-se calcular a probabilidade associada a cada um desses F, quer usando-se tabelas ou equações. Em [108, Cap.6, sec.1 e 3], é feita a formulação a seguir.

Sejam duas variâncias s_1^2 e s_2^2, com graus de liberdade associados ν_1 e ν_2, as quais se deseja

comparar. A razão F de Fisher é dada por:

$$F = \begin{cases} \frac{s_1^2}{s_2^2} = F_{\nu_1,\nu_2}, & s_1^2 \geq s_2^2 \\ \frac{s_2^2}{s_1^2}, = F_{\nu_2,\nu_1} & s_1^2 < s_2^2 \end{cases} \tag{4.8}$$

pode ser usada para se avaliar o grau de significância da diferença entre as variançcas, pela integral[21]:

$$\begin{aligned} Q(F|\nu_1,\nu_2) &= I_{\frac{\nu_2}{\nu_2+\nu_1 F}}\left(\frac{\nu_2}{2},\frac{\nu_1}{2}\right) \\ &= \frac{\int_0^{\frac{\nu_2}{\nu_2+\nu_1 F}} t^{\frac{\nu_2}{2}-1}.(1-t)^{\frac{\nu_1}{2}-1}dt}{\int_0^1 t^{\frac{\nu_2}{2}-1}.(1-t)^{\frac{\nu_1}{2}-1}dt} \end{aligned} \tag{4.9}$$

$$\tag{4.10}$$

e

$$P(F|\nu_1,\nu_2) = \begin{cases} 2[1-Q(F|\nu_1,\nu_2)]. & \text{se } Q(F|\nu_1,\nu_2) > \frac{1}{2} \\ 2.Q(F|\nu_1,\nu_2). & \text{caso contrário} \end{cases} \tag{4.11}$$

$P(F|\nu_1,\nu_2)$ acima indica a probabilidade de as variâncias serem "iguais". Se P for pequeno, a probabilidade de que as variâncias comparadas sejam diferentes é grande. Por outro lado, valores grandes de P implicam numa probabilidade pequena de que as variâncias sejam diferentes[22]. Limiares normalmente usados variam de 1% a 5%, como os aplicados para os testes de Student para a média.

4.8.3 Conversão de MOS para Q

A conversão de MOS para Q envolve basicamente duas etapas: encontrar uma função que mapeie MOS em Q e a aplicação desta função aos valores MOS encontrados.

Para relacionar MOS a Q, os valores MOS obtidos para as condições que envolvem o MNRU são utilizados para se encontrar uma função que relacione MOS a Q ($Q = \mathcal{F}[\text{MOS}]$)[23]. Correntemente, dois métodos têm sido utilizados para o ajuste de curva.

Um deles se baseia na estimação linear ponderada de dois parâmetros (L e M) e na estimação heurística de um terceiro, m. O método se utiliza da equação [90]:

$$Q_{equ} = \log(\frac{Y_q - 1}{5 - Y_q}) = a_0 + a_1.L + a_2.L^2 + a_3.L.M + a_4.M + a_5.M^2$$

[21] Se $F = 0$, $Q = 1$; se $F \to \infty$, $Q \to 0$. Portanto, Q está no intervalo [0,1].

[22] Alternativamente, $F \approx 1$ implica em $P \to 1$; já $F \gg 1$ implica em $P \to 0$.

[23] Para que a identificação dessa função seja possível, é importante haver um número razoável de condições do MNRU no teste, estabelecendo um "continuum" de qualidade (isto é, cobrindo os extremos de qualidade de uma maneira uniforme). Em geral, colocam-se de 6 a 8 valores de MNRU num teste, igualmente espaçados entre Q igual a 0 ou 5 dB até Q igual a 30 ou 35 dB. No teste subjetivo do LD-CELP da Fase II (ver Capítulo 6) havia 6 condições de MNRU com Q variando de 5 a 30 dB em passos de 5dB (5, 10, 15, ..., 30dB).

onde Y_q é o MOS obtido para cada nível de audição L
$M = 10^{-m \cdot Q}$, $m = 0.01..0.05$, para o m que gerar menor erro;
a referência [91] sugere $m = 0.03$.

A essa interpolação, associa-se a ponderação $\frac{(Y_q-1)(5-Y_q)}{16}$ em cada ponto e considera-se a condição direta (nenhum processamento) como sendo $Q \rightarrow \infty$, ou, equivalentemente, $M = 0$ [111].

O outro deles se utiliza da curva logística, que é uma técnica de estimação de três parâmetros (A, B e M) baseada em regressão (mínimos quadrados) não-linear ponderada. A curva logística é definida por [112]:

$$Q_{eqv2} = B \ln(\frac{Y_q - 1}{M - Y_q}) + A$$

onde Y_q é o MOS obtido para cada nível de audição L
A é o valor médio da curva logística;
M é o valor da assíntota superior da curva ($x \rightarrow \infty$);
$\frac{1}{4B}$ é a inclinação da curva no ponto $x = A$.

O uso de um ou outro método tem recomendações. O primeiro deve ser utilizado quando o valor M não precisar ser estimado, mas sim for um valor obtido diretamente do teste (por exemplo, um alto valor de Q – tipicamente 50 dB ou mais – ou a condição direta [111]); quando o valor de M não for conhecido a priori, o segundo método é mais adequado. Além disso, o segundo método também fornece uma idéia da média e da inclinação da curva MOS $\times$ Q (refletidos por A e B).

4.9 Sumário

Neste Capítulo apresentamos um histórico da avaliação subjetiva de qualidade envolvendo a transmissão telefônica. Descrevemos aspectos relacionados à gravação, processamento e avaliação desses sinais de voz processados, com ênfase especial nos chamados *testes de audição*. Apresentamos também métodos estatísticos normalmente utilizados na análise de testes subjetivos. No Apêndice a seguir, apresentamos dois exemplos bastante simples de planos de testes subjetivos.

A Exemplo de Sentenças para Testes Subjetivos

1. O abajur quebrou no meio da sala de visitas
   Abandonou os planos que lhe desagradaram
2. Acredito na honestidade do instituto
   O acusado se declarou inocente ao juiz
3. Eis aí uma escola que deu certo
   Ah, sim! só agora eu a reconheço
4. Airton Sena venceu mais uma corrida
   Alegria é uma dádiva que brota do coração
5. Alfredo entregou-se às ocupações diárias
   Alguém aqui me chamou de idiota?
6. Alguém aí quer que eu pegue um cafezinho?
   Quem viu a minha bota de caminhadas?
7. Dar à luz é renascer e fazer nascer
   O DDT está proibido nos Estados Unidos
8. De grão em grão a galinha enche o seu papo
   De quem são esses sapatos molhados?
9. Deputado deporá sobre o caso de Búzios
   Derramaram a refeição do príncipe
10. As férias coletivas sempre são em janeiro
    Fábio e Cabral são fanáticos por churrasco
11. Fiquei gripado na cachoeira de Souzas
    A Flávia faz aniversário hoje
12. O plano verao fracassou copiosamente
    As plantas secaram por falta d'água
13. Por favor, você poderia atender o telefone?
    Por que não vamos esperar o ônibus?
14. Sempre deveria ser fim de semana
    A todo momento vemos muita corrupção
15. Beleza é, hoje, sinônimo de saúde
    Bem-vindo ao nosso centro de estudos
16. O bigode dele é de pelo social
    Biscoito com chocolate engorda

B Exemplo de plano de teste subjetivo

A seguir são dados dois exemplos de plano de testes subjetivos utilizando quadrados greco-latinos. Seus tamanhos foram mantidos pequenos para facilitar sua compreensão.

O primeiro deles é o mais simples, em que todo o material fonte é processado por todas as condições (Experimento Fatorial, [96, pp.13–17]). O segundo utiliza a técnica de intercalamento, em que metade do material fonte é processado por metade das condições (e.g., condições de número par) e a outra metade pelas outras condições (e.g., condições ímpares).

Como pressuposto para o exemplo, o material de voz encontra-se digitalizado e armazenado em um computador equipado com um sistema de conversão A/D e D/A de alta qualidade; todas as condições de processamento encontram-se bem ajustadas e conectadas ao hardware de processamento. O material após processado será armazenado no mesmo computador e avaliadores serão chamados para apreciar os arquivos utilizando aparelhos telefônicos padrão em uma sala com ruído ambiente controlado. Suas notas, de 1 a 5. serão coletadas e analisadas posteriormente, descartando-se os votos obtidos durante a fase de treinamento. Antes do início das sessões subjetivas, um texto padrão será distribuído aos avaliadores com informações preliminares sobre o procedimento básico do teste.

Os arquivos estão batizados com a seguinte convenção: a raiz do nome contém informação de quem é o locutor, qual o seu sexo e a que lista o arquivo pertence; a extensão *ext* contém a informação sobre o tipo de processamento efetuado:

$$TxxLyyyz.ext$$

onde

T	identifica o sexo do locutor: M para voz masculina; F para voz feminina; C para crianças.
xx	numero do locutor dentro de um mesmo sexo (01..N);
yyy	número da lista em que o material foi alocado (001..L);
z	índice de um elemento dentro de uma lista (A,B,...);
ext	extensão do arquivo: .src, para arquivos originais (não processados); .ckk, para arquivos processados pela condição kk (kk=01..C).

Nos exemplos a seguir, o número de locutores de um mesmo sexo é N=2, o número de listas é L=4 e o número de condições é C=04. Com somente um elemento por lista, z=A. Escolhendo locutores somente do sexo masculino e femenino, T será M ou F, teremos a seguinte distribuição: locutor M01, lista 001; locutor M02, lista 002; locutor F01, lista 003; locutor F02, lista 004.

B.1 Experimento sem intercalamento

Condições do teste: As condições do teste são mostradas abaixo:

Condição	Descrição
1	codec sob teste
2	direto
3	G.721
4	MNRU com Q=20dB
Nota:	* *1 nível de audição*
	* *1 replicação*

Para quatro condições, é necessário um quadrado greco-latino de ordem 4. Neste experimento, onde todo o material será processado por todas as condições e somente haverá uma replicação, o número de listas será igual ao de condições.

Com somente 1 nível de audição, o número de elementos por lista é 1, isto é, cada lista compreenderá 1 arquivo de voz digitalizada (com duas frases).

Para 4 listas e 4 condições, serão gerados $4 \times 4 \times 1 = 16$ elementos (arquivos) processados. O seu arranjo para as sessões de audição utiliza um dos possíveis quadrados greco-latinos de ordem 4, como definido abaixo:

Sessão	Seqüência de apresentação			
	1	2	3	4
1	L001	L002	L003	L004
2	L002	L001	L004	L003
3	L003	L004	L001	L002
4	L004	L003	L002	L001

Sessão	Seqüência de apresentação			
	1	2	3	4
1	C01	C03	C04	C02
2	C02	C04	C03	C01
3	C03	C01	C02	C04
4	C04	C02	C01	C03

Suponha-se então que os arquivos de voz estejam batizados de acordo com a convenção estabelecida no preâmbulo. Então, o ouvinte 1 (ou equivalentemente a sessão 1) ouviria a seguinte seqüência de arquivos:

Arquivo	*Descrição*
M01L001A.C01	Lista 1 processada pela condição 1 (codec)
M02L002A.C03	Lista 2 processada pela condição 3 (G.721)
F01L003A.C04	Lista 3 processada pela condição 4 (MNRU)
F02L004A.C02	Lista 4 processada pela condição 2 (direto)

O próximo ouvinte (sessão 2):

Arquivo	*Descrição*
M02L002A.C02	Lista 1 processada pela condição 1 (codec)
M01L001A.C04	Lista 2 processada pela condição 3 (G.721)
F02L004A.C03	Lista 3 processada pela condição 4 (MNRU)
F01L003A.C01	Lista 4 processada pela condição 2 (direto)

E assim por diante. Pode-se notar que cada ouvinte vai avaliar uma combinação única de lista e condição, i.e., uma mesma combinação de lista e condição não será apresentada a dois ouvintes diferentes. Isto é decorrência direta da ortogonalidade dos quadrados greco-latinos, explicada anteriormente.

Ao final, seriam coletadas 4 notas por ouvinte, num total de 16 notas. Se se repetisse o experimento para outros conjuntos de 4 ouvintes, teríamos replicações, aumentando correspondentemente o número de notas.

Mais de um nível de audição: Uma extensão possível deste teste seria utilizar 3 níveis de audição. Neste caso, as listas seriam compostas não de 1, mas de 3 elementos (arquivos). Assim, o número total de arquivos originais *(.src)* a serem processados não seria 4, mas 12. Com 4 condições, teríamos $12 \times 3 = 48$ arquivos processados.

Neste ponto, o procedimento de audição pode variar, mas uma solução usual (que permite avaliar o efeito do nível de audição como uma variável independente do teste) é o de manter o esquema de audição lista/condição anterior, porém aleatorizando os níveis de audição durante a apresentação de cada lista. Para ilustrar isso, suponha-se a sequinte seqüência "aleatória" de níveis de audição:

N1	N2	N3	N2	N1	N3	... etc
N3	N2	N1	N1	N2	N3	...
N1	N2	N2	N3	N1	N2	
N3	N1	N2	N2	N3	N1	

Onde N1 representa e.g. o nível preferido de audição (79 dBSPL), N2 um nível 10 dB acima do nominal e N3 um nível 10 dB abaixo do nominal.

A seqüência de audição seria então:

Ouvinte	*Arquivo*	*Descrição*
1	M01L001A.C01	Lista 1/condição 1 (codec), nível N1
	M01L001B.C01	Lista 1/condição 1 (codec), nível N2
	M01L001C.C01	Lista 1/condição 1 (codec), nível N3
	M02L002A.C03	Lista 2/condição 3 (G.721), nível N3
	M02L002B.C03	Lista 2/condição 3 (G.721), nível N2
	M02L002C.C03	Lista 2/condição 3 (G.721), nível N1
	F01L003A.C04	Lista 3/condição 4 (MNRU), nível N1
	F01L003B.C04	Lista 3/condição 4 (MNRU), nível N3
	F01L003C.C04	Lista 3/condição 4 (MNRU), nível N2
	F02L004A.C02	Lista 4/condição 2 (direto), nível N3
	F02L004B.C02	Lista 4/condição 2 (direto), nível N1
	F02L004C.C02	Lista 4/condição 2 (direto), nível N2
2	M02L002A.C02	Lista 1/condição 1 (codec), nível N2
	M02L002B.C02	Lista 1/condição 1 (codec), nível N1
	M02L002C.C02	Lista 1/condição 1 (codec), nível N3
	M01L001A.C04	Lista 2/condição 3 (G.721), nível N1
	M01L001B.C04	Lista 2/condição 3 (G.721), nível N2
	M01L001C.C04	Lista 2/condição 3 (G.721), nível N3
	F02L004A.C03	Lista 3/condição 4 (MNRU), nível N3
	F02L004B.C03	Lista 3/condição 4 (MNRU), nível N1
	F02L004C.C03	Lista 3/condição 4 (MNRU), nível N2
	F01L003A.C01	Lista 4/condição 2 (direto), nível N2
	F01L003B.C01	Lista 4/condição 2 (direto), nível N3
	F01L003C.C01	Lista 4/condição 2 (direto), nível N1
..... etc		

Uma possível variante seria quebrar o teste em 3 sub-sessões, sendo que o nível de audição seria um parâmetro fixo dentro de cada sub-sessão. Assim, manter-se-ia o primeiro arranjo

descrito, que seria repetido três vezes, alterando-se da primeira para a segunda vez a letra A para B, e na terceira, para C (i.e., troca-se o elemento da lista).

Estatísticas: Voltando ao exemplo básico, com somente 1 nível de audição, teríamos para a média 3 graus de liberdade e, para a variância, 2 graus.

Denotando os votos colhidos por Y_{lc}, onde l é a lista e c é a condição de processamento, teríamos a média para cada condição c ($c = 1..4$):

$$MOS_c = \frac{1}{4}\sum_{l=1}^{4} Y_{lc}$$

e variância:

$$s_c^2 = \frac{1}{3}\sum_{l=1}^{4}(Y_{lc}^2 - \frac{MOS_c^2}{4})$$

Além disso, uma análise de variância pode ser implementada:

Fatores	Graus de liberdade
Ouvintes	3 (No. de ouvintes – 1)
Seqüência de Audição	3 (No. de seqüências – 1)
Condição	3 (No. de condições – 1)
Listas	3 (No. de listas – 1)
Resíduo (Erro)	$(4-1)\times(4-3) = 3$

Com o teste F pode-se então verificar se as variâncias individuais de cada fator do teste são significativas ou não em relação ao resíduo (erro experimental) do teste. Caso as variâncias sejam significativamente diferentes, então ou houve erro experimental, ou então existem outros fatores que foram deixados implícitos, ou *confundidos*, com o resíduo. Nestes casos, estudos mais aprofundados do teste são necessários para se descobrir que fatores são esses. Como um exemplo, poderia haver uma componente de variância entre locutores (como no caso de um locutor com voz irritante) ou uma componente de variância entre o resultado do processamento da voz de diversos locutores.

O intervalo de confiança para uma margem de 95% é dado por

$$CI_c = \pm 1.96\sqrt{s_c/4}$$

e a Mínima Diferença Significativa (MSD) é:

$$MSD_c = \sqrt{2} \times CI_c$$

Poderíamos ainda comparar o codec em teste para ver se o seu MOS (neste caso, MOS_1) poderia ser considerado igual ao de alguma âncora, e.g. G.721 (condição 3). Para isso usa-se o teste de Student para a média, dado um nível de significância α (e.g., 1% para diferenças não-significativas, até 5% para diferenças significativas e acima de 5% para diferenças muito

significativas). Calcula-se o valor de t para as duas médias em questão e aí a probabilidade $Pr(|t| < \tau_{\alpha,\nu})$ associada a esse t, que é por definição $1 - \alpha$. Com esse resultado classificam-se o codec e a referência como *subjetivamente* equivalentes, diferentes ou extremamente diferentes.

B.2 Experimento com intercalamento

Condições do teste: Como num teste com intercalamento metade das listas é processada pelas condições e.g. pares e a outra metade pelas e.g. ímpares, pode-se dobrar o número de condições do teste, para um mesmo tamanho de quadrado greco-latino[24]. As condições para este teste hipotético seriam então:

Condição	Descrição
1	Primeiro codec sob teste
2	Segundo codec sob teste
3	G.721
4	G.711
5	MNRU com Q=10dB
6	MNRU com Q=20dB
7	MNRU com Q=30dB
8	Direto
Nota:	* *1 nível de audição*
	* *1 replicação*

Para oito condições com intercalamento como descrito, são necessários dois quadrados greco-latinos de ordem 4, denotados como GL1 e GL2. Neste experimento, onde o material será processado por metade das condições e somente haverá uma replicação, o número de listas também será igual ao de condições.

Para n níveis de audição, o número de elementos por lista é também n; isto é, cada lista compreenderá n arquivos de voz digitalizada. No exemplo anterior descrevemos testes para n igual a 1 e 3; aqui, para simplificar, descreveremos apenas para $n = 1$.

Pode-se então relacionar listas, locutores e condições:

Locutor	Condições	
	Pares	Ímpares
M01	L001	L005
M02	L002	L006
F01	L003	L007
F02	L004	L008

Para 8 listas e 8 condições com intercalamento, serão gerados $8 \times (8/2) \times 1 = 32$ elementos (arquivos) processados[25]. O seu arranjo para as sessões de audição utiliza dois dos possíveis

[24] Numa outra abordagem, para um certo número de condições que se deseja testar, pode-se reduzir à metade o material de voz processado, o que implica em sessões de avaliação mais curtas. Em casos reais, isto pode significar uma redução do tempo de sessão de 1 hora para meia hora, bem como do espaço em disco para armazenamento do material.

[25] Note que se não houvesse intercalamento, o material processado compreenderia $8 \times 8 \times 1 = 64$ elementos.

quadrados greco-latinos de ordem 4 (GL1 e GL2, respectivamente), como definido abaixo, de forma intercalada. Entretanto, para não se diminuir a potência do teste, os avaliadores serão expostos às mesmas listas duas vezes, porém processadas por outras condições (ainda seguindo a mesma restrições anteriores).

GL1	Seqüência de apresentação				Sessão	Seqüência de apresentação			
Sessão	1	2	3	4		1	2	3	4
1	C01	C03	C05	C07	1	L001	L002	L003	L004
2	C03	C01	C07	C05	2	L004	L003	L002	L001
3	C05	C07	C01	C03	3	L002	L001	L004	L003
4	C07	C05	C03	C01	4	L003	L004	L001	L002

GL2	Seqüência de apresentação				Sessão	Seqüência de apresentação			
Sessão	1	2	3	4		1	2	3	4
1	C08	C06	C04	C02	1	L008	L007	L006	L005
2	C06	C08	C02	C04	2	L006	L005	L008	L007
3	C04	C02	C08	C06	3	L005	L006	L007	L008
4	C02	C04	C06	C08	4	L007	L008	L005	L006

O procedimento de audição consiste em apresentar o material processado de acordo com o intercalamento dos dois quadrados greco-latinos acima, isto é, primeiro uma lista/condição do GL1, então do GL2, volta-se ao GL1 e assim por diante. No nosso exemplo, teríamos a seguinte seqüência de audição:

Sessão 1	Sessão2	Sessão 3
M01L001.C01	F02L004.C03	M02L002.C05
F02L008.C08	M02L006.C06	...etc...
M02L002.C03	F01L003.C01	.
F01L007.C06	M01L005.C08	.
F01L003.C05	M02L002.C07	.
M02L006.C04	F02L008.C02	
F02L004.C07	M01L001.C05	
M01L005.C02	F01L007.C04	

Ao final, serão coletadas 8 notas por ouvinte, num total de 32 notas.

Variantes deste esquema são possíveis, para se aumentar a potência do teste. Pode-se, por exemplo, duplicar o material ouvido por um mesmo avaliador, fazendo com que o material que seria apreciado numa outra sessão por um outro ouvinte, também seja apreciado por este. Este aumento na potência do teste aumenta em conseqüência o tempo gasto pelo avaliador, mas não o número de arquivos de voz processados (isto pode ser significativo quando se têm, por exemplo, 24 ou 48 condições num experimento, como em [90] ou [91, pp.19–32]).

Estatísticas: As estatísticas básicas são computadas de mesmo modo:

$$\begin{aligned} MOS_c &= \frac{1}{8}\sum_{l=1}^{8} Y_{lc} \\ s_c^2 &= \frac{1}{7}\sum_{l=1}^{8}(Y_{lc}^2 - \frac{MOS_c^2}{8}) \\ CI_c = &= \pm 1.96\sqrt{s_c/8} \end{aligned}$$

Já a análise de variância é mais complicada e dependente das variantes implementadas. Bons exemplos são encontrados em [90] e [91, pp.66–76].

C Exemplo de instruções para um teste ACR

TESTES SUBJETIVOS - EXPERIMENTO 2 - FASE I
para a padronização de um codec a 16 kbps pelo CCITT

CPqD/Telebrás

INTRUÇÕES AOS AVALIADORES

Neste teste estarão sendo avaliados sistemas que poderão ser utilizados para serviços de telecomunicações entre lugares distantes.
Você logo estará ouvindo pelo aparelho telefônico um certo número de amostras de voz. Cada amostra consistirá de 2 sentenças. Por favor, ouça o par de sentenças e então anote a sua opinião sobre a qualidade geral do que você acabou de ouvir usando a seguinte escala de notas:

Nota	Opinião de Qualidade
5	Excelente (Muito Boa)
4	Boa
3	Razoável
2	Pobre
1	Ruim (Muito Pobre)

Após ouvir uma seqüência de sentenças, por favor tome nota de sua resposta na folha apropriada, especialmente distribuída para esse fim; essa nota deve representar a sua opinião sobre a qualidade das amostras que você acabou de ouvir.
Após cada par de sentenças haverá uma pausa para que você possa tomar nota de seu voto, da ordem de 5 segundos.
Para que você possa treinar, você vai ouvir preliminarmente duas amostras, uma representando o que se considera "Qualidade Excelente", e outra, "Qualidade Ruim". Após essa sessão preliminar haverá uma breve pausa para esclarecer eventuais dúvidas; isto feito, começará a sessão efetiva de coleta das suas avaliações.
O teste terá duração de aproximadamente uma hora e meia, e 3 pausas de aproximadamente 5 minutos deverão ser feitas para descanso.
Por favor, não discuta suas opiniões com outros avaliadores que estão ou estarão participando dos testes, para evitar a formação de pré-disposições para a avaliação. Durante a condução do teste, por favor não interaja com o avaliador ao lado, para evitar interferência mútua nas avaliações.

D Exemplo de instruções para um teste DCR

TESTES SUBJETIVOS - EXPERIMENTO 4 - FASE I
para a padronização de um codec a 16 kbps pelo CCITT

CPqD/Telebrás

INTRUÇÕES AOS AVALIADORES

Neste experimento você estará ouvindo pares de amostras de voz com e sem degradação, como aconteceria ao ouvi-los pela linha telefônica normal. O primeiro par de sentenças deve ser considerado como referência para o segundo par de sentenças aos quais você está sendo solicitado a avaliar a degradação de um para outro (isto é, quanto o segundo par piorou em relação ao primeiro), dando a cada um desses conjuntos uma nota de acordo com o seguinte critério:

Nota	Degradação
5	A degradação é inaudível
4	A degradação é audível mas não incomoda
3	A degradacão incomoda um pouco
2	A degradação é incomodante
1	A degradação incomoda muito

Após ouvir o par de sentenças de referência e o processado, queira por favor anotar sua avaliação no campo apropriado da folha que lhe foi distribuída para este fim; esta avaliação deverá refletir sua opinião quanto à degradação presente no segundo par de sentenças em relação ao primeiro par.
Entre cada par de sentenças haverá uma pausa de alguns segundos devido à necessidade do computador de carregar os sinais de voz em sua memória. Este tempo de espera deverá ser usado para você anotar a sua avaliação a cada dois pares de sentenças.
Para que você se acostume à prática, haverá inicialmente uma sessão com 12 avaliações, numeradas de 1 a 12 na folha de respostas. Após ela, uma pausa será feita para o esclarecimento de eventuais dúvidas; então começarão a primeira rodada e a segunda rodada de avaliações, entre as quais faremos uma pausa de 5 a 10 minutos, para sua maior comodidade.

E Uso de Unidades Relativas

Quando se avalia a qualidade de sistemas ou equipamentos, é importante expressar essa qualidade de modo que possam ser feitas comparações com outros sistemas ou equipamentos já conhecidos.

Um meio bastante utilizado é o uso de unidades relativas: a qualidade é expressa por meio de um valor único dentro de uma escala unidimensional, graduada numa unidade claramente definida e que possa ser considerada universal.

É importante notar que a unidimensionalidade não é suficiente para se definir uma boa unidade relativa. É necessário ainda que a escala seja inequívoca e com significado 'universal'. Com esse requisito, escalas como a ACR (é unidimensional de 5 pontos, Excelente, Bom, Razoável, Ruim e Péssimo) são inadequadas, pois apesar de apresentarem um *continuum* decrescente de qualidade (5, 4, 3, 2 e 1), o significado dos adjetivos (feitas as devidas traduções) não é universal. Como um exemplo, "excelente" para um ouvinte japonês tem um sentido muito mais restrito (mais próximo de "perfeito") que para um ouvinte norte-americano (mais próximo de "muito bom").

Com uma maior objetividade da medida, fica muito mais fácil o intercâmbio de informações sobre a avaliação de qualidade desses sistemas e equipamentos.

Quando essas unidades relativas visam relacionar o desempenho de um sistema em testes com o desempenho de um sistema de referência, normalmente elas são chamadas de "equivalentes de referência".

Há três métodos pelos quais essa comparação pode ser feita: da referência variável, da margem e o indireto.

No da referência variável, varia-se um parâmetro do sistema de referência, até que os sistemas de referência e de teste sejam considerados "balanceados" (i.e., possam ser considerados equivalentes).

No método da margem, deixa-se a referência fixa (por exemplo, simulando uma condição de circuito especialmente importante) e coloca-se um ganho ou perda no sistema sob teste, até que referência e teste estejam balanceados.

No método paramétrico (indireto), ambos os sistemas são descritos em termos de uma outra unidade (por exemplo, AEN ou Q), para várias condições de circuito; então, a relação entre eles é derivada indiretamente (e não por comparação direta, como nos casos anteriores).

Capítulo 5

Medidas Objetivas

5.1 Introdução

No capítulo sobre Metodologias de Testes Subjetivos serão mostrados diversos aspectos envolvidos com a avaliação da qualidade de algoritmos de codificação de voz. Da sua leitura pode-se concluir que os testes subjetivos, embora sejam a avaliação definitiva da qualidade de um sistema, são caros e complicados de serem realizados, além de demandar tempo para sua condução e para a alocação de avaliadores. Por isso, é de grande utilidade a identificação de métodos objetivos para a avaliação de algoritmos.

Esses métodos objetivos deveriam apresentar baixa complexidade computacional, bem como correlacionar-se bem com as avaliações humanas sobre a qualidade dos sinais codificados.

Verificou-se que medidas objetivas que não se baseiem em informações sintáticas e semânticas da linguagem nunca poderão correlacionar-se corretamente com a qualidade percebida pelo usuário para uma grande quantidade de sinais processados por algoritmos de codificação [53]. O estágio atual do compreensão do complexo processo de percepção da fala, entretanto, não permite a elaboração de medidas que plenamente modelem esse processo.

Apesar das diversas medidas objetivas estudadas até hoje falharem em maior ou menor grau nesse modelamento, muitas delas foram e têm sido utilizadas no processo de aprimoramento de diversos algoritmos de codificação.

Assim, é importante estar ciente das limitações dos métodos objetivos disponíveis hoje em dia, sem porém descartar o seu uso. Reforçando este aspecto, o CCITT vem estudando diversas técnicas objetivas para a avaliação algoritmos de codificação de voz dentro da Comissão de Estudos XII [13, 14].

5.2 Classes de Medidas

A grosso modo, as medidas objetivas podem ser classificadas em três classes: medidas temporais, medidas espectrais e medidas psicoacústicas.

As *medidas temporais* baseiam-se na análise temporal da forma de onda *(waveform)* do sinal. Um exemplo clássico é a razão ou relação sinal-ruído [54].

As *medidas espectrais* utilizam a estimação de parâmetros da representação espectral do sinal para avaliar sua qualidade. Um exemplo bastante comum na literatura é a distância cepstral [55].

Além dessas, há *medidas psicoacústicas*, que se baseiam em modelos de percepção da fala e levam em conta o mecanismo de audição do ouvido humano [56].

Pode-se ainda utilizar técnicas mistas, que utilizam aspectos de uma e de outra das classes acima. Um exemplo é a relação sinal-ruído ponderada em freqüência [57].

A seguir, serão descritas as medidas mais importantes.

5.3 Medidas Temporais

5.3.1 Relação Sinal-Ruído

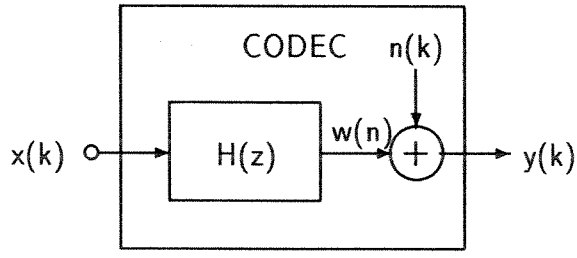


Figura 5.1: Modelo geral de um codificador de forma de onda. As distorções introduzidas pelo processo representado por H(z) são as distorções (lineares) de atenuação e atraso. Já n(k) modela as distorções não-lineares introduzidas pelo codificador.

A primeira das medidas objetivas de distorção é a relação sinal-ruído *(Signal-to-Noise Ratio, SNR)*[1]. Ela é uma herança dos sistemas analógicos, onde o tipo predominante de distorção é o ruído aditivo. Generalizando para sistemas digitais, a relação sinal-ruído passa a ser a relação entre a potência de um sinal de referência e a potência de um sinal de erro.

Em [58], a relação sinal-ruído *(SNR)* para um processo digital como o da figura 5.1 é dada por:

$$SNR_t = 10 \log_{10} \left(\frac{\sum_k x^2(k)}{\sum_k e^2(k)} \right), \text{ em [dB]}.$$

onde

$$e(k) = y(k) - x(k)$$

No caso de um arquivo de voz, o índice k acima cobre todas as amostras do sinal digitalizado. No caso mais geral, k abrangeria toda a duração do sinal. Por isso, esta SNR é chamada de *SNR total* ou *SNR de longo prazo*.

Apesar de adequada para avaliar a qualidade de outros sinais que não voz (como sinais de dados), há longa data conhece-se a baixa correlação da SNR total com a qualidade subjetiva

[1] Apesar da tradução mais exata do termo ingês *ratio* ser "razão", o termo "relação" é o mais utilizado em português, por razões históricas. Por isso, manteremos aqui o seu uso.

percebida para algoritmos de codificação[2], chegando mesmo a ser uma das piores medidas objetivas de distorção, segundo Quackenbush *et al.* [59, p.9]. Em particular, codificadores com SNR semelhantes podem apresentar um desempenho subjetivo significantemente diferente [58]. Como um exemplo, um sistema de codificação ADPCM apresentará em geral uma qualidade subjetiva superior à de um sistema de codificação PCM logarítmico com mesma SNR. Neste caso também se incluem equalizadores de fase e processos que implementam transformadas de Hilbert [60], bem como amplificadores com inversão de fase. Como o ouvido humano é insensível à resposta de fase do sistema, a qualidade subjetiva não se altera; entretanto, a forma de onda se altera e a SNR alterar-se-á significativamente.

5.3.2 Relação Sinal-Ruído Segmentada

Implementação clássica: As estatísticas de sinais de voz não podem ser conhecidas a priori, além dessas estatísticas serem variantes com o tempo. Por isso, não se conhece a priori a função de densidade de probabilidade *(pdf)* do sinal, tampouco a sua variância. Assim, é importante se calcular uma boa estimativa da variância do sinal. No caso de sinais variantes com o tempo, a estimativa tem que ser local e não global. Isso explica a discrepância básica da SNR total com a qualidade subjetiva. Adicionalmente, um fator que aumenta essa discrepância é o ruído de canal desocupado *(idle channel noise, ICN)*, que é bastante incômodo na avaliação subjetiva, mas é mascarado no cálculo da SNR total quando da elevação da potência do sinal.

Para tentar contornar os problemas de predição da qualidade subjetiva para codificadores digitais de forma de onda, Noll propôs a SNR segmentada *(SNR_s)* [54]. O cálculo da SNR seria feito localmente, isto é, para segmentos (ou blocos) consecutivos do sinal. A SNR_s para o sinal seria então o valor médio da SNR de todos os segmentos do sinal em análise. Em termos de equações,

$$SNR_s = \frac{1}{M}\sum_{m=0}^{M-1} SNR_s(m)$$

$$SNR_s(m) = 10\log_{10}\left(\frac{\sum_{k=0}^{N-1} x^2(k+m.N)}{\sum_{k=0}^{N-1} \epsilon^2(k+m.N)}\right), \text{ em [dB]}.$$

onde N é o número de amostras em cada segmento (bloco) e M é o número de segmentos em que o sinal analizado foi dividido. Note que x e ϵ ainda se referem à figura 5.1.

Com essa medida, segmentos de alta e baixa potência (nível) terão um peso semelhante no cálculo da SNR_s. Deste modo, o efeito do ICN estará refletido nesta medida. Paradoxalmente, esta é uma das limitações da SNR_s proposta por Noll: trechos de pequena amplitude, em especial de silêncio, tenderão produzir sinais de erro ϵ pequenos. Como neste caso a relação entre o sinal e o ruído será bem menor que 1, a SNR do segmento será grande e negativa e tenderá a dominar (diminuindo) o valor médio da SNR_s numa proporção maior que seu impacto em termos subjetivos.

[2] Isto pode ser explicado pelo fato de a SNR tratar o sinal de voz como um único vetor, como se o ouvinte fizesse uma comparação simples após "armazenar" uma fala completa [56]. Claramente, esta não é uma proposta razoável, pois as opiniões são formadas ao longo do processo de audição.

SNR segmentada com limiar: Para contornar esse efeito, Crochière [61] sugere o uso de um limiar. Quando a SNR de um segmento for inferior a um limiar T, ela não deveria ser incluída no cômputo da SNR_s. Ainda, para se evitar a dominância de outros segmentos com SNR muito alta, pode-se limitar a excursão da SNR entre um valor mínimo $SEG_{1(min)}$ e um valor máximo $SEG_{1(max)}$. Com essa modificação, a relação sinal-ruído segmentada com limiar passa a ser:

$$
\begin{aligned}
SEG_1 &= \frac{1}{M}\sum_{m=0}^{M-1} SEG_1(m) \\
SEG_1(m) &= \begin{cases} SEG_{1(min)}, & SEG_1'(m) \leq SEG_{1(min)}; \\ SEG_1'(m), & SEG_{1(min)} < SEG_1'(m) < SEG_{1(max)}; \\ SEG_{1(max)}, & SEG_1'(m) \geq SEG_{1(max)}; \end{cases} \\
SEG_1'(m) &= \begin{cases} 0, & \sum_k x^2(k+m.N) \leq T.N; \\ SNR_s(m), & \sum_k x^2(k+m.N) > T.N. \end{cases}
\end{aligned}
$$

Em [61], utiliza-se $SEG_{1(min)}$=*-10dB*, $SEG_{1(max)}$=*80dB* e T=*900* (para amostras excursionando entre -32768..32767). Este valor de T foi determinado empiricamente e corresponde a um segmento cuja potência está 61dB abaixo do ponto de saturação do sistema *(-61dBov)*. Nessa referência, utilizam-se segmentos de 20 ms de duração, o que corresponde a N=160 amostras para uma taxa de amostragem de 8kHz.

SNR segmentada sem limiar: Outra solução alternativa à SNR_s clássica de Noll é a descrita em [62] e em [63], dispensando o uso de um limiar pré-escolhido (como no caso da SEG_1)[3]:

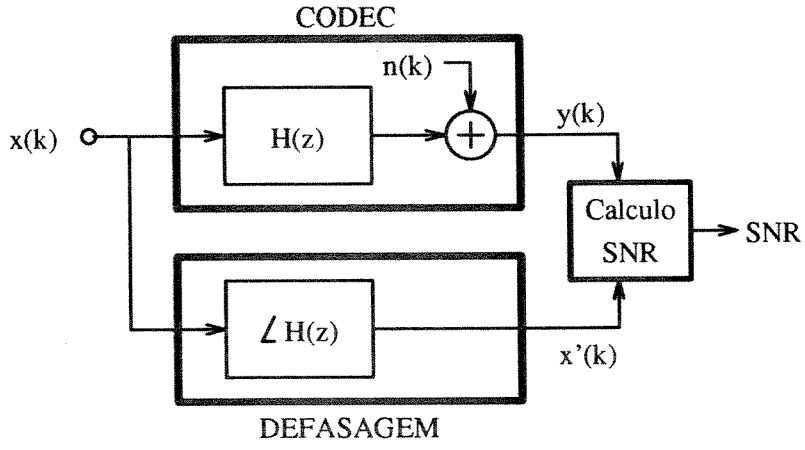
$$
\begin{aligned}
SEG_2 &= \frac{1}{M}\sum_{m=0}^{M-1} SEG_2(m) \\
SEG_2(m) &= 10\log_{10}\left(1 + \frac{\sum_{k=0}^{N-1} x^2(k+m.N)}{\sum_{k=0}^{N-1} \epsilon^2(k+m.N)}\right), \text{ em [dB].}
\end{aligned}
$$

Deste modo, um trecho de silêncio com ruído terá relação sinal-ruído segmentada $SEG_2(m) = 10\log_{10}(1) = 0$dB. Portanto, não contribuirá para o cálculo da SNR.

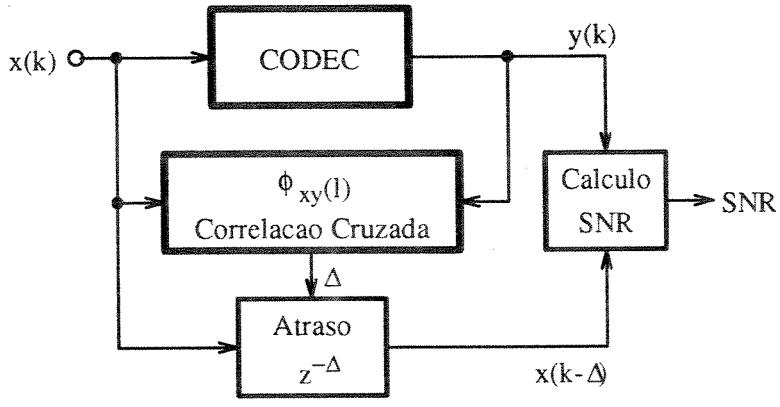
Esta abordagem aplica um critério suave para a inclusão ou não de segmentos, enquanto que a anterior, com limiar, aplica um critério brusco (inclui ou não inclui).

Tamanho do segmento: Mermelstein [62] reporta que o tamanho do segmento não altera significantemente o valor da SNR_s. Ele verificou que segmentos com duração entre 64 e 256 amostras geraram diferenças abaixo de 1dB no valor da SNR_s. Na referência [62] são empregados segmentos de 64 amostras (8ms), enquanto que em [63], assim como no caso da SEG_1, usam-se 160 amostras (20ms).

[3] A formulação em [62] é diferente daquela em [63]; porém, quando o sinal for *ativo* na maior parte do tempo, ambas serão equivalentes. Este é o caso de interesse para sinais de voz.



(a) Uso de filtro com mesma resposta de fase.



(b) Uso da correlação cruzada.

Figura 5.2: Métodos de cálculo da SNR compensando o atraso de processamento.

5.3.3 Compensação de Atraso

Todo processo digital introduz algum tipo de atraso no sinal codificado. Esse atraso pode ser originado pelos filtros embutidos no codificador, pelo armazenamento de amostras para a composição de quadros de processamento, ou pelo tempo necessário para o processamento do sinal por um DSP que implementa um codificador.

No caso de medidas objetivas de distorção temporal, os atrasos têm efeitos catastróficos, pois haverá intrinsecamente um descasamento entre o sinal processado $y(k)$ e o de referência, $x(k)$. Isto produzirá sinais de erro $e(k)$ artificialmente grandes. É, portanto, necessária a compensação do atraso entre $y(k)$ e $x(k)$ antes do cálculo da SNR.

Isto em geral é feito de dois modos. O mais simples é ilustrado na figura 5.2(a) e se baseia no modelo de codec da figura 5.1. Consiste no cálculo do sinal de erro utilizando-se uma versão atrasada do sinal de referência $x(k)$ pelo seu processamento por um filtro com resposta de fase igual à do filtro $H(z)$ do modelo do codificador. Uma desvantagem deste método consiste na dependência com o codificador, pois para cada codec em análise dever-se-á utilizar uma implementação (software) específica. Outra desvantagem é que este processo não modela atrasos provenientes do armazenamento de amostras e do tempo de processamento por DSPs.

Outro método que pode ser utilizado é o da correlação, que é ilustrado na figura 5.2(b). Calcula-se a correlação (cruzada) entre o sinal processado e o sinal de referência e encontra-se o pico (absoluto) na função de correlação. O atraso será o índice *(lag)* referente a esse pico. Como esse método não pressupõe qualquer modelo de processamento, ele é independente do algoritmo de codificação em questão, além de todos os atrasos descritos serem levados em conta. Além disso, é considerado como um dos métodos mais eficientes para a determinação do atraso [64].

Em termos gerais, a correlação cruzada entre o sinal processado e o de referência pode ser escrita como:

$$\phi_{xy}(l) \triangleq \sum_{j=-\infty}^{+\infty} x(j)y(j+l)$$

onde l é o índice da função de correlação cruzada. Ele se refere ao deslocamento entre as seqüências para as quais um determinado valor de correlação é calculado.

Para seqüências x e y de comprimento finito L_x e L_y, podemos expressar a correlação para valores positivos de l como sendo (ver figura 5.3):

$$\phi_{xy}(l) = \sum_{j=0}^{L_y-l-1} x(j)y(j+l)$$

Já para $l < 0$,

$$\phi_{xy}(l) = \sum_{j=-l}^{L_x-1} x(j)y(j+l)$$

Entretanto, a formulação acima é inconveniente para implementações em software, pois envolveria valores negativos de j. Efetuando-se substituição de variáveis na equação acima,

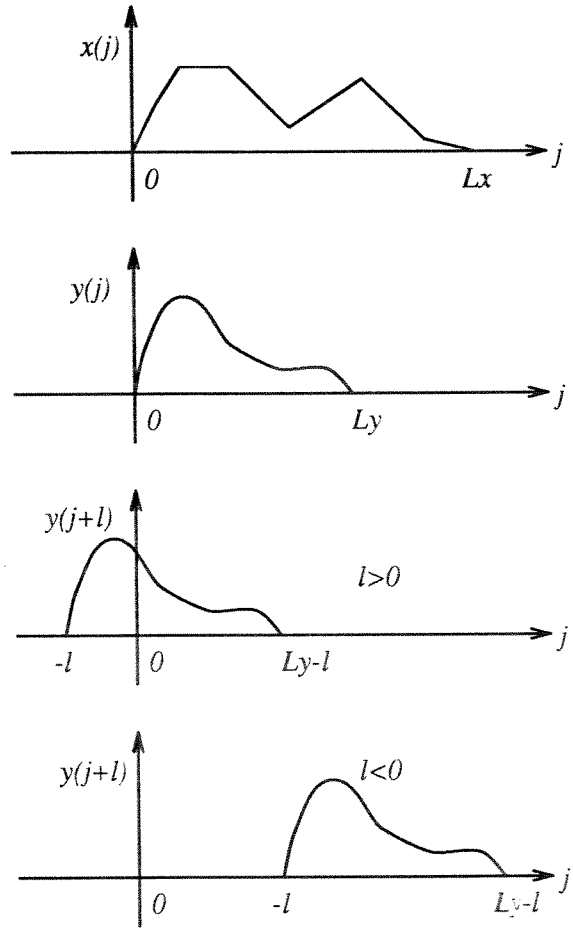


Figura 5.3: Cálculo da seqüência de correlação entre os sinais x e y de duração finita L_x e L_y.

teremos:

$$\phi_{xy}(l) = \sum_{m=0}^{L_x+l-1} x(m-l)y(m) = \sum_{m=0}^{L_x-\lambda-1} y(m)x(m+\lambda) \triangleq \phi_{yx}(\lambda)$$

onde $\lambda = -l$.

Como $l < 0$, $\lambda > 0$. Portanto, o cálculo de $\phi_{yx}(\lambda)$ utiliza somente índices positivos, facilitando a implementação em software. Assim, a correlação $\phi_{xy}(l)$ entre x e y para valores negativos de l pode ser encontrada pelo cálculo da função de correlação complementar, $\phi_{yx}(\lambda)$.

O valor do atraso Δ será o índice do pico absoluto da função de correlação cruzada:

$$|\phi_{xy}(\Delta)| \geq |\phi_{xy}(l)|, \text{para todo } l.$$

Se Δ for negativo, então a seqüência y estará atrasada em relação a x. Caso contrário, se Δ for positivo, então x é que estará atrasada em relação a y[4]. Complementarmente, se $\phi_{xy}(\Delta)$ for negativo, então além de atraso houve inversão de fase.

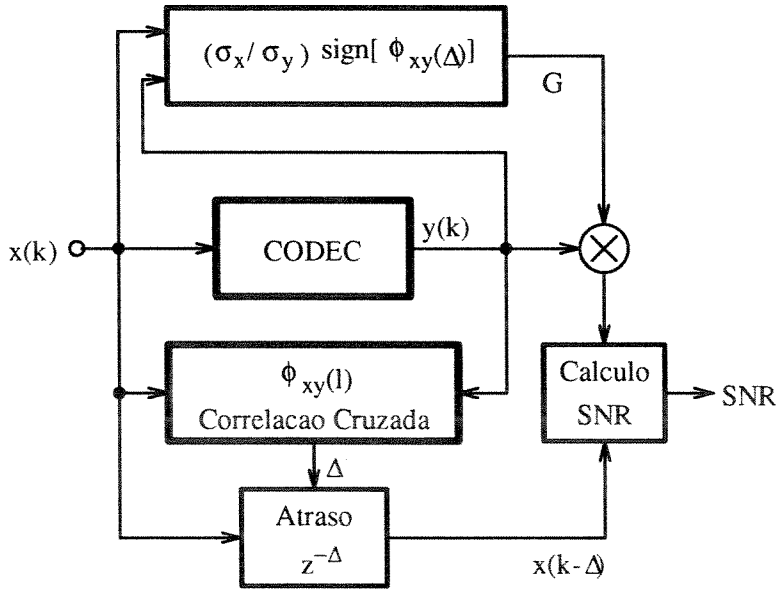


Figura 5.4: Compensação de atraso e equalização de nível. Método independente do codificador.

5.3.4 Equalização de Amplitude

É muito comum a alteração do nível do sinal à saída de codecs. Do mesmo modo que o atraso, a SNR será afetada por essa mudança de nível. Como um exemplo, se dois sinais x e y forem idênticos, porém y tendo o dobro da potência de x, a SNR será de 0dB, apesar da qualidade de ambos ser indistinguível.

[4] No caso de sinais processados por simulação de algoritmos, sempre ocorrerá o segundo caso. No entanto, para sinais processados em tempo real, poderá ocorrer o corte de trechos iniciais dos sinais, caindo no primeiro caso. Como um exemplo, se um sinal possuir um trecho inicial de 1s de silêncio e, após ser processado por um algoritmo que introduz 100ms de atraso, for armazenado com um corte de seus 0.5s iniciais, o sinal estará integro, porém o sinal original estará atrasado 400ms em relação ao processado armazenado. Por isso se justifica aqui a análise para atrasos positivos e negativos.

Como no caso anterior, poder-se-ia utilizar uma filtragem cuja resposta em amplitude reproduzisse aquela do filtro interno do codec. Outra vez, é um método dependente do codificador.

Alternativamente, poder-se-ia utilizar a razão entre a potência dos sinais processado e de referência para a equalização do sinal processado:

$$G = \frac{\sigma_x}{\sigma_y} sign(\phi_{xy}(\Delta))$$

onde G é o fator de equalização a ser aplicado no sinal y, $sign$ é a função-sinal definida por

$$sign(x) = \begin{cases} +1, & \text{se } x \geq 0 \\ -1, & \text{se } x < 0 \end{cases}$$

e σ_x e σ_y são dados por:

$$\sigma_x^2 = \sum_{k=0}^{L_x-1} x_k^2$$

$$\sigma_y^2 = \sum_{k=0}^{L_y-1} y_k^2$$

Como não foram pressupostos modelos para o codificador, este segundo método também é independente do codificador. Seu esquema básico encontra-se na figura 5.4, que inclui também a determinação do atraso entre as seqüências pelo método da correlação, explicado anteriormente.

5.4 Medidas Espectrais

As medidas espectrais de distorção são medidas de distância entre as magnitudes dos espectros de um sinal processado e de um sinal de referência. Elas têm sido usadas amplamente em reconhecimento e verificação de voz, bem como para a avaliação objetiva da qualidade de sinais processados.

Como essas medidas se baseiam na magnitude do espectro do sinal de voz, diferenças em fase e, consequentemente, o atraso entre os sinais comparados, não afetariam as medidas para cálculo de distorção espectral de longo prazo, caso as medidas fossem calculadas globalmente para o sinal. Entretanto, para explorar a propriedade de estacionaridade do sinal de voz para curtos segmentos de fala, faz-se necessário o cálculo das medidas espectrais utilizando-se medidas baseadas em segmentos do sinal. Em função do tamanho do bloco para medidas segmentais e do valor do atraso, pode haver um comprometimento da precisão da medida para sinais não-estacionários, como o sinal de voz. Isto se deve ao fato da segmentação poder ocorrer em trechos de transição da fala. Se o atraso for significativo, haverá inerentemente uma diferença na conformação espectral para aqueles segmentos do sinal. Portanto, apesar destas medidas serem *a princípio* insensíveis ao atraso de fase, a associação entre o caráter não estacionário (de longo prazo) e o uso de medidas *locais*, torna necessário o uso de técnicas de compensação de atraso, como as descritas anteriormente na Seção 5.3.3.

As medidas de distorção são inerentemente positivas, pois são distâncias euclidianas. Assim, valores próximos a zero refletem sinais semelhantes, enquanto que valores distantes de zero indicam uma distorção crescente, não necessariamente linear com a escala objetiva.

Além disso, verificou-se que as medidas de distorção baseadas diretamente na distância euclideana de parâmetros do modelo espectral de produção da fala, e.g. coeficientes LPC e de reflexão, apresentam baixa correlação com resultados subjetivos [55]. Entretanto, medidas baseadas no valor quadrático médio *(RMS)* da representação logarítmica do espectro apresentaram melhor correlação.

5.4.1 Medidas de Distância

Em geral, há uma série de propriedades desejáveis para uma medida de distância D entre dois sinais x e y [55, 59]:

1. Simetria: ..... $d(x,y) = d(y,x)$
2. Ser uma forma *definida positiva*: ..... $d(x,y) > 0; d(x,y) = 0 \Leftrightarrow x \equiv y$
3. Satisfazer a desigualdade triangular: ..... $d(x,y) \leq d(x,z) + d(z,y)$
4. Ter uma interpretação física no domínio da freqüência.
5. Poder ser implementada de forma eficiente.

As duas últimas se referem mais a aspectos práticos de utilização da medida de distância: ser interpretável fisicamente facilita o estabelecimento de uma correlação com as avaliações subjetivas, enquanto que uma implementação eficiente permite menores tempos de processamento.

Do ponto de vista matemático, uma medida de distância que satisfaça as três primeiras propriedades é chamada de *métrica* [65, pp.163-165]. Entretanto, isso raramente é necessário [59, p.37]. Em geral, não se exige que a propriedade 3 seja cumprida. Quackenbush *et al.* [59] questionam a obrigatoriedade da propriedade 1, mas sua conveniência é inequívoca: além de dispensar a preocupação de qual sinal é de teste e qual é o de referência, caso se conclua que x é N dB melhor que y, somente com a simetria pode-se inferir que y é N dB pior que x.

Assim, em geral exige-se de uma medida objetiva que ela atenda às propriedades 1 e 2 e deseja-se que satisfaça as propriedades 4 e 5.

Gray *et al.* [55] sugerem que seja utilizado um conjunto de normas ou distâncias definidas por:

$$(D_p)^p = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\mathcal{V}_{xy}(\omega)|^p d\omega$$

onde $\mathcal{V}_{xy}(\omega)$ é um sinal de erro no domínio da freqüência entre os sinais de referência x e de teste y. Aqui p é chamado de ordem da norma.

Quackenbush *et al.* sugerem uma versão discreta para esse conjunto de normas [59, p.210], que chamam de distância espectral *não ponderada*:

$$(\mathcal{D}_p)^p = \frac{1}{L} \sum_{l=0}^{L-1} |\mathcal{V}_{xy}(\omega_l)|^p$$

onde

$$\omega_l = \frac{2\pi l}{L}, \quad l = 0..L-1.$$

e L é o número de pontos em freqüência da DFT *(Discrete Fourier Transform, [66])* empregada, em geral utilizando-se L=128.

Dependendo da definição de $\mathcal{V}(\omega)$ ou $\mathcal{V}(\omega_l)$, surgem diversos tipos de distância. Quackenbush *et al.* [59, pp.210–211] ainda ampliam a medida com a inclusão de uma ponderação em freqüência. Numa forma generalizada, teríamos:

$$(D_p)^p = \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi} W(\omega)|\mathcal{V}_{xy}(\omega)|^p d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi} W(\omega) d\omega} \tag{5.1}$$

ou

$$(\mathcal{D}_p)^p = \frac{\frac{1}{L}\sum_{l=0}^{L-1} W(l)|\mathcal{V}_{xy}(\omega_l)|^p}{\frac{1}{L}\sum_{l=0}^{L-1} W(l)} \tag{5.1.a}$$

onde $W(\cdot)$ é uma função positiva de ponderação em freqüência.

Como esta medida é local (isto é, a análise é realizada para um bloco, ou segmento, do sinal), para o sinal como um todo calcula-se a média sobre todos os segmentos. Para manter a generalidade, Quackenbush *et al.* sugerem que se associe ao segmento m uma ponderação $W(m)$[5]. Então, a medida de distância média é:

$$\bar{D}_p = \frac{\sum_{m=1}^{M} W(m) D_p(m)}{\sum_{m=1}^{M} W(m)} \tag{5.2}$$

ou

$$\bar{\mathcal{D}}_p = \frac{\sum_{m=1}^{M} W(m) \mathcal{D}_p(m)}{\sum_{m=1}^{M} W(m)} \tag{5.2.a}$$

onde $D_p(m)$ e $\mathcal{D}_p(m)$ são as medidas de distância da equação 5.1 feitas para o m-ésimo segmento de análise.

Escolha de p: Em termos físicos, D_1 resulta no valor médio da magnitude do sinal de erro $|\mathcal{V}_{xy}(\omega)|$. Já D_2 é o valor rms do erro e representa uma distância quadrática média entre os sinais aos quais o erro se refere. Em D_∞ domina o valor de pico de $\mathcal{V}_{xy}(\omega)$, sendo portanto o valor da maior diferença (erro) entre x e y. Para as outras ordens das normas, não se associam interpretações físicas.

Gray *et al.* [55] reportam que existe uma alta correlação entre D_2 e D_∞. Isto indica que D_∞, bem como todas as normas com ordem acima de 2, podem ser satisfatoriamente aproximadas

[5] Isto serviria, por exemplo, ao propósito de eliminar trechos de silêncio da medida global, de ponderar mais trechos no meio do sinal e atenuar nas extremidades ou ainda outras finalidades desta natureza. No caso mais comum, onde todos os segmentos são de igual importância, $W(m) = 1, m = 1..M$.

por D_2. Com esse resultado, sugere-se que a escolha da ordem da norma a ser utilizada deva se pautar mais pela existência de um significado físico para a distância, bem como pela sua facilidade de implementação (computacional), do que pela busca de um valor de p que seja ótimo em algum sentido dentre todos os possíveis valores de p. Baseado nesse resultado, normalmente se utiliza $p = 2$ para as medidas objetivas de distorção.

Determinação de $\mathcal{V}_{xy}(\omega)$: A escolha do sinal de erro $\mathcal{V}_{xy}$ determinará o tipo de medida de distorção. O erro pode ser entre os coeficientes *LAR (Log Area Ratios)* [18, 67] ou ainda entre os coeficientes LPC ou de reflexão (ou entre sua representação logarítmica). Pode ser ainda definido em termos da representação espectral dos sinais (obtido via FFT), ou do logaritmo desta. As três primeiras são baseadas na *envoltória espectral* (às vezes chamado de *espectro suavizado*) dos sinais, enquanto que as últimas são baseadas na representação integral do espectro dos sinais. A seguir são apresentadas algumas das mais importantes definições de $\mathcal{V}_{xy}$, representando estes dois tipos de definição do erro.

5.4.2 Distância Cepstral

O Cepstro: O conceito de *cepstro*[6] vem das técnicas de processamento homomórfico [66, 18]. Sinais misturados por convolução podem ser decompostos em fonte e filtro através do processamento indicado na figura 5.5. Nela, $\mathcal{Z}$ indica a transformada Z, $\mathcal{Z}^{-1}$ indica a transformada Z inversa e $ln[\,]$ representa o logaritmo (neperiano) complexo de $X(z)$. O sinal processado $\hat{x}(k)$ é chamado de cepstro complexo do sinal $x(k)$ [66]. Para uma seqüência $x(k)$ de fase mínima[7], $\hat{x}(k)$ pode ser obtido a partir do logaritmo real do módulo de $X(z)$ [66].

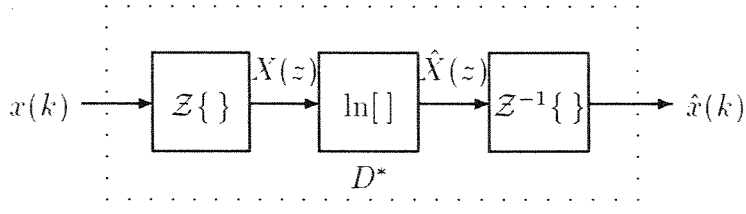


Figura 5.5: Processamento homomórfico para sinais misturados por convolução.

Matematicamente, modelando $x(k)$ como a convolução de uma excitação $v(k)$ por um filtro cuja resposta impulsiva é $h(k)$, poderemos escrever:

$$x(k) = v(k) * h(k)$$

[6] A palavra "cepstro" foi utilizada como a versão do termo *cepstrum*, utilizado em inglês. *Cepstrum* vem de um anagrama da palavra latina *spectrum* ao inverter as 4 primeiras letras, o que leva a *cepstro* em português.

[7] Uma seqüência $x(k)$ é dita de *fase mínima* quando todos os pólos e zeros de sua representação no plano Z, $X(z)$, caem dentro da circunferência de raio unitário.

e

$$\begin{aligned}\hat{x}(k) &= \mathcal{Z}^{-1}\{\ln[\mathcal{Z}\{x(k)\}]\} = \mathcal{Z}^{-1}\{\ln[X(z)]\} = \mathcal{Z}^{-1}\{\ln[V(z).H(z)]\} \\ &= \mathcal{Z}^{-1}\{\ln[V(z)] + \ln[H(z)]\} = \mathcal{Z}^{-1}\{\ln[V(z)]\} - \mathcal{Z}^{-1}\{\ln[H(z)]\} \\ &\triangleq \hat{v}(k) + \hat{h}(k)\end{aligned}$$

onde

$$\hat{v}(k) = \mathcal{Z}^{-1}\{\ln[\mathcal{Z}\{v(k)\}]\}$$

e

$$\hat{h}(k) = \mathcal{Z}^{-1}\{\ln[\mathcal{Z}\{h(k)\}]\}$$

Por analogia, $\hat{v}(k)$ é o cepstro complexo da excitação $v(k)$ e $\hat{h}(k)$ é o cepstro complexo da resposta impulsiva $h(k)$ do modelo de produção do sinal $x(k)$.

No caso de $x(k)$ ser um sinal de voz, $h(k)$ pode se referir à resposta impulsiva do filtro LPC que modela o sistema de produção da fala. Consequentemente, $\hat{h}(k)$ se refere ao cepstro complexo do modelo de produção da fala.

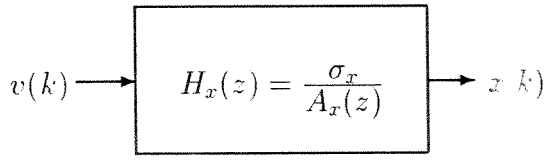


Figura 5.6: Modelo LPC de produção da fala

Modelo LPC: Se considerarmos o sinal de voz segundo o modelo LPC da figura 5.6, o filtro de síntese pode ser equacionado como:

$$H(z) = \frac{\sigma}{A_x(z)}$$

com

$$A_x(z) = 1 + \sum_{i=1}^{P} a_i z^{-i}$$

onde P é a ordem do preditor, $a_i (i = 1..P)$ são os coeficientes LPC e σ^2 é o ganho do modelo.

Podemos então expressar o logaritmo do modelo $H(z)$ por sua expansão em série de potências, utilizando o fato de que $\hat{h}(k) = \mathcal{Z}^{-1}\{\ln[H(z)]\}$:

$$\ln\left[\frac{\sigma}{A_x(z)}\right] = \sum_{k=0}^{\infty} \hat{h}(k) z^{-k} = \ln[\sigma] + \sum_{k=1}^{\infty} \hat{h}(k) z^{-k}$$

com $\hat{h}(0) = \ln[\sigma]$.

Por outro lado, o logaritmo de sua densidade espectral de potência $|H(z)|^2$ também pode ser expresso pela expansão em série de potências (lembrando que $\hat{h}(k)$ é causal [66]):

$$\begin{aligned}\ln\left[\left|\frac{\sigma}{A_x(z)}\right|^2\right] &= \ln\left[\left|\frac{\sigma}{A_x(z)}\right|\right] + \ln\left[\left|\frac{\sigma}{A_x^*(z)}\right|\right] \\ &= \sum_{k=1}^{\infty}\hat{h}(k)z^{-k} + \sum_{k=-\infty}^{-1}\hat{h}(-k)z^{-k} + \ln[\sigma^2] \\ &\triangleq \sum_{k=-\infty}^{\infty} c(k)z^{-k}\end{aligned}$$

onde $A^*(z)$ é o complexo conjugado de $A(z)$ e $c(0) = \ln[\sigma^2]$. Nesse contexto, $c(k)$ equivale aos coeficientes cepstrais da função de densidade espectral de potência. Podemos estabelecer as seguintes relações entre ambos os coeficientes desses dois cepstros:

$$\begin{aligned}c(k) &= \hat{h}(k), \quad k > 0 \\ c(k) &= c(-k) \\ c(0) &= 2\hat{h}(0)\end{aligned} \tag{5.3}$$

Recursão entre $\mathbf{a_i}$ e $\hat{\mathbf{h}}(\mathbf{k})$: Os coeficientes LPC definem unicamente $H(z)$ e, consequentemente, $h(k)$. Como $\hat{h}(k)$ representa uma transformação dos parâmetros do trato vocal, pode-se esperar uma relação entre a_i e $\hat{h}(k)$, ou, utilizando-se as relações acima, entre a_i e $c(k)$. De fato, Atal prova [68] que existe uma relação recursiva entre o cepstro complexo do modelo e os coeficientes LPC que o definem. Escrevendo essa relação em termos da definição de $A_x(z)$ acima e das relações dadas por (5.3), teremos:

$$c(k) = \begin{cases} -a_1, & \text{para } k = 1 \\ -\sum_{i=1}^{k-1}(1 - i/k)c(k-i)a_i - a_k, & \text{para } 1 < k \leq P \\ -\sum_{i=1}^{k-1}(1 - i/k)c(k-i)a_i, & \text{para } k > P \end{cases} \tag{5.4}$$

Distância Cepstral com infinitos termos Se $c(k)$ representa uma transformação logarítmica espectral do modelo do trato vocal, pode-se pensar num sinal de erro $\mathcal{V}_{xy}$ baseado em $\hat{H}(e^{j\omega})$. De fato, Gray *et al.* [55] definem o sinal de erro como a diferença entre as componentes de amplitude do espectro LPC, no domínio logarítmico, para o sinal original x e uma versão sua distorcida y como:

$$\mathcal{V}_{xy}^{(CD)}(\omega) = \ln\left[\frac{|H_x(e^{j\omega})|^2}{|H_y(e^{j\omega})|^2}\right] = \ln\left[\left|\frac{\sigma_x}{A_x(e^{j\omega})}\right|^2\right] - \ln\left[\left|\frac{\sigma_y}{A_y(e^{j\omega})}\right|^2\right] \tag{5.5}$$

Com essa definição, D_2 será uma distância espectral logarítmica entre os espectros de potência dos sinais de teste e de referência.

Fazendo $C(e^{j\omega}) \triangleq \ln|H(e^{j\omega})|^2$, poderemos escrever $\mathcal{V}_{xy}^{(CD)}$ como:

$$\mathcal{V}_{xy}^{(CD)}(\omega) = C_x(e^{j\omega}) - C_y(e^{j\omega})$$

Utilizando-se a definição da equação (5.1), gera-se uma família de distâncias

$$(D_p^{(CD)})^p = \frac{1}{2\pi}\int_{-\pi}^{\pi} |C_x(e^{j\omega}) - C_y(e^{j\omega})|^p d\omega$$

Em especial para $p = 2$,

$$(D_2^{(CD)})^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |C_x(e^{j\omega}) - C_y(e^{j\omega})|^2 d\omega$$

Aplicando-se o teorema de Parseval, podemos escrever D_2^2 em termos da transformada inversa de $C(e^{j\omega})$, $c(k)$, lembrando que $c(k)$ é uma seqüência real:

$$(D_2^{(CD)})^2 = \sum_{k=-\infty}^{+\infty} [c_x(k) - c_y(k)]^2$$

Como $c(k)$ é simétrica [55], podemos escrever:

$$(D_2^{(CD)})^2 = [c_x(0) - c_y(0)]^2 + 2\sum_{k=1}^{\infty}[c_x(k) - c_y(k)]^2, \tag{5.6}$$

Essa medida é uma métrica que representa a distorção espectral logarítmica do trato vocal entre um sinal y processado e um sinal de referência x e é chamada de *distância cepstral*. Além disso, ela satisfaz todas as 5 propriedades desejadas para uma medida de distorção e diversos estudos mostram a sua alta correlação com as medidas de avaliação subjetiva de qualidade [57].

Distância Cepstral Truncada: Desse equacionamento de $D_2^{(CD)}$ surge uma questão importante: como o cepstro de $h(k)$ tem comprimento infinito, o cálculo de $(D_2^{(CD)})^2$ envolverá o uso de um número infinito de coeficientes. Entretanto, para qualquer implementação prática, $c(k)$ terá que ser truncada. Sabendo que a amplitude de $c(k)$ decai hiperbolicamente com k [66, pp.502–503], Gray *et al.* verificaram que a seqüência cepstral truncada $[u(L)]$, definida por:

$$[u(L)]^2 = [c_x(0) - c_y(0)]^2 + 2\sum_{k=1}^{L}[c_x(k) - c_y(k)]^2$$

apresentou uma alta correlação ρ com $(D_2^{(CD)})^2$ para valores relativamente pequenos de L. Em especial, $u[P]$, $u[2P]$ e $u[3P]$ apresentaram ρ iguais a 0.98, 0.997 e 0.999, respectivamente. Adicionalmente, como $c(k)$ é obtida recursivamente a partir dos P coeficientes LPC, para que $c(k)$ possa ser univocamente definida por a_i, faz-se necessário o uso de todos os os coeficientes LPC[8]. Isso implica que o cálculo (recursivo) de $c(k)$ deve ser realizado para L pelo menos igual a P. Por isso, o limitante inferior para o comprimento da seqüência cepstral truncada é $L = P$.

[8] A definição unívoca de $c(k)$ por a_i é necessária para se manter a propriedade de univocidade para as medidas de distância (Propriedade 2), usando-se a definição de $\mathcal{V}_{xy}^{(CD)}$.

Compensação de Níveis Diferentes. Para evitar medidas errôneas de distorção, pode-se normalizar a energia do sinal de teste, em relação à do sinal de referência, como mostrado no esquema da figura 5.4. Isso implica em multiplicar todas as amostras do sinal de teste por um fator σ_x/σ_y, cujo cômputo envolve também uma multiplicação e uma soma para cada amostra de ambos os sinais. Entretanto, no caso da distância cepstral, isso não é necessário, bastando eliminar da somatória o termo $c_x(0) - c_y(0)$, que corresponde à razão da energia dos sinais (ver equação (5.3)). Isso resulta numa distância cepstral com normalização implícita do nível dos sinais, a partir da definição de $u(L)$ acima:

$$CD = \frac{10}{\ln(10)}\sqrt{2\sum_{k=1}^{L}[c_x(k) - c_y(k)]^2} \tag{5.7}$$

onde o termo $10/\ln(10)$ corresponde ao fator para exprimir a distância cepstral em dB. Essa distância é então calculada para cada segmento dos sinais e o seu valor médio é estabelecido pela equação (5.2.a). A definição de distância cepstral dada pela relação (5.7) tem sido estudada pelo CCITT [64, 57].

5.4.3 Distância Espectral

Seja um sinal x de referência e uma versão sua y distorcida, ambos de faixa limitada. A transformada de Fourier deles será:

$$T(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} t(n)e^{-j\omega n},\ T = \{X, Y\}, t = \{x, y\}$$

A sua versão discreta para L pontos em freqüência é:

$$\begin{aligned} T(l) &= \sum_{n=0}^{L-1} t(n)e^{-jn\omega_l} \\ \omega_l &= \frac{2\pi l}{N} \end{aligned}$$

O sinal de erro $\mathcal{V}_{xy}$ pode ser então definido em termos do logaritmo da magnitude espectral de x e y:

$$\mathcal{V}_{xy}^{(SD)}(\omega) = \log_{10}|X(e^{j\omega})| - \log_{10}|Y(e^{j\omega})|$$

O que gera outra distância D_2 (ainda a partir da equação 5.1)):

$$(D_2^{(SD)})^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\{\log_{10}|X(e^{j\omega})| - \log_{10}|Y(e^{j\omega})|\}^2 d\omega \tag{5.8}$$

ou, na versão discreta (pela equação 5.1.a)):

$$(\mathcal{D}_2^{(SD)})^2 = \frac{1}{L}\sum_{l=0}^{L-1}\{\log_{10}|X(l)| - \log_{10}|Y(l)|\}^2 \tag{5.8.a}$$

Na versão desta medida sugerida em [57], utiliza-se $L = 256$ e os espectros $X(k)$ e $Y(k)$ são obtidos via FFT.

Quackenbush *et al.* se referem a esta medida também como distância espectral logarítmica sem ponderação. Ela se distingue da distância cepstral em pelo menos dois aspectos fundamentais. Enquanto a distância espectral pressupõe apenas uma limitação de faixa dos sinais, a distância cepstral pressupõe um sinal que possa ser decomposto em excitação e filtro através de uma análise LPC. Além disso, a distância espectral utiliza o espectro *total* do sinal, enquanto que a distância cepstral apenas se baseia na envoltória do espectro de potência. Isto porque a parte referente ao espectro da excitação é removida implicitamente no processo de geração do cepstro a partir dos coeficientes LPC (que, afinal, modelam apenas o trato vocal e não o sinal de voz completo).

Compensação de Amplitude: O método de compensação de amplitude da figura 5.4 envolve o cômputo da energia local dos sinais de referência e de teste e a multiplicação de todas as amostras deste pela razão σ_x/σ_y. Como a distância espectral envolve a diferença dos logaritmos, é mais otimizado, após calcular-se o fator de normalização, subtraí-lo da distância espectral dos sinais sem normalização. Isto substitui multiplicações por somas. Assim, expressamos a distância espectral com compensação de amplitude, em dB, como sendo:

$$SD = 10\sqrt{\frac{1}{L}\sum_{l=0}^{L-1}\left\{\log_{10}|X(l)| - \log_{10}|Y(l)| - \log_{10}\left[\frac{\sigma_x}{\sigma_y}\right]\right\}^2} \qquad (5.9)$$

Como no caso da distância cepstral, a SD deve ser calculada para cada segmento dos sinais e o seu valor médio é encontrado pela equação (5.2.a).

Distância Espectral Ponderada: Itoh *et al.* [57] sugerem uma modificação da distância espectral através da divisão do espectro em bandas definidas pelo *Índice de Articulação* [69], conforme apresentado na tabela 5.1. No índice de articulação, cada uma das bandas foi determinada de modo a contribuir em igual monta à qualidade percebida do sinal, em especial quanto à sua inteligibilidade para sistemas de analógicos lineares corrompidos por ruído [59, pp.37–41]. Como são 20 bandas, diz-se que cada uma das bandas do espectro contribui *idealmente* com 5% da inteligibilidade do sinal [17].

Tabela 5.1: Bandas de Freqüência de Igual Contribuição ao Índice de Articulação.

Banda	Faixa [Hz]	Banda	Faixa [Hz]
1.	200 a 330	11.	1660 a 1830
2.	330 a 430	12.	1830 a 2020
3.	430 a 560	13.	2020 a 2240
4.	560 a 700	14.	2240 a 2500
5.	700 a 840	15.	2500 a 2820
6.	840 a 1000	16.	2820 a 3200
7.	1000 a 1150	17.	3200 a 3650
8.	1150 a 1310	18.	3650 a 4250
9.	1310 a 1480	19.	4250 a 5050
10.	1480 a 1660	20.	5050 a 6100

A medida proposta é:

$$D_2' = \frac{1}{B}\sum_{l=1}^{B} W_b D_2^{(SD)}(b) \qquad \text{ou} \qquad WSD = \mathcal{D}_2' = \frac{1}{B}\sum_{l=1}^{B} W_b \mathcal{D}_2^{(SD)}(b)$$

onde B é o número de bandas, W_b é a ponderação da banda b e $\{D_2(b),\mathcal{D}_2(b)\}$ são as distâncias $\{D_2,\mathcal{D}_2\}$ da equação 5.8 calculada *somente* para as componentes de freqüência da banda b. Note que D_2' é a versão contínua e $\mathcal{D}_2'$ é a versão discreta. Como cada uma das bandas do Índice de Articulação contribuem em igual monta, cada uma delas tem igual importância no cômputo da medida. Portanto, as ponderações W_b devem ser todas unitárias, i.e., $W_b = 1, b = 1..B$.

Itoh *et al.* [57] se referem ao uso de somente 16 das 20 bandas definidas para o Índice de Articulação. Isto pode ser explicado pelo fato dos sinais de voz para aplicações em telefonia terem seu espectro limitado na faixa dos 300 Hz aos 3400 Hz. Assim, somente se utilizariam as bandas de número 2 a 17, cobrindo o espectro de 330 Hz a 3650 Hz.

5.4.4 Razão de verossimilhança

A *Razão de Verossimilhança* é uma medida de distorção, inicialmente proposta por Itakura [63] para uso em reconhecimento de voz. Vários autores a estudaram para a predição da qualidade subjetiva de sinais de voz processados [59, 57, 70, 55].

Ela avalia a diferença dos modelos LPC de um sinal de teste e outro de referência através da comparação do erro de predição para esses sinais. Assim, ela presupõe que o sinal de voz possa ser representado pelo modelo LPC de ordem P da figura 5.6 e, à semelhança da distância cepstral, também reflete distorções não sobre o espectro total, mas sim sobre sua envoltória. A razão de verossimilhança é definida então como:

$$\delta/\alpha = \frac{\vec{a}_x R_y \vec{a}_x^t}{\vec{a}_y R_y \vec{a}_y^t}$$

onde t indica a operação de transposição, $\vec{a}_x$ e $\vec{a}_y$ são respectivamente os vetores de coeficientes LPC que modelam o sinal de referência x e de teste y,

$$\begin{aligned} \vec{a}_x &= [1 a_1 a_2 \ldots a_P] \\ \vec{a}_y &= [1 \hat{a}_1 \hat{a}_2 \ldots \hat{a}_P] \end{aligned}$$

e R_y é a matriz de autocorrelação para o sinal y, cujos elementos r_{ij} são calculados para um segmento com N amostras pela relação:

$$r_{ij} = r_{ji} = \sum_{n=1}^{N-|i-j|} y(n)y(n+|i-j|), \text{ para } |i-j| = 0..P-1$$

Como $\vec{a}_y$ é calculado a partir de y, $\alpha = \vec{a}_y R_y \vec{a}_y^t$ é o erro de predição mínimo. Portanto, $\delta = \vec{a}_x R_y \vec{a}_x^t > \alpha$.

Complementarmente, a razão de verossimilhança logarítmica, como definida por Itakura, é

$$l_{x,y} = \log(\delta/\alpha)$$

Gray *et al.* [55] sugerem um método eficiente para o cálculo da razão de verossimilhança. Podemos escrever α em termos do erro de predição $\epsilon(n)$:

$$\epsilon(n) = \sum_{i=0}^{P} \tilde{a}_i y(n-i)$$

Assim,

$$\alpha = \sum_{n=-\infty}^{+\infty} e(n)^2 = \sum_{n=-\infty}^{+\infty} \left[\sum_{i=0}^{P} \tilde{a}_i y(n-i) \right]^2$$

Separando os quadrados e trocando uma das variáveis, podemos escrever:

$$\begin{aligned} \alpha &= \sum_{n=-\infty}^{+\infty} \left[\sum_{i=0}^{P} \tilde{a}_i y(n-i) \right] \left[\sum_{j=0}^{P} \tilde{a}_j y(n-j) \right] = \sum_{i=0}^{P} \sum_{j=0}^{P} \tilde{a}_i \tilde{a}_j \left[\sum_{n=-\infty}^{+\infty} y(n-i) y(n-j) \right] \\ &\triangleq \sum_{i=0}^{P} \sum_{j=0}^{P} \tilde{a}_i \tilde{a}_j r_y(i-j) \end{aligned}$$

Expandindo a somatória em seus termos e identificando os termos da função de autocorrelação dos coeficientes $\tilde{a}_i$, obtemos:

$$\begin{aligned} \alpha &= r_y(-M) \sum_{i=0}^{P} \tilde{a}_i \tilde{a}_{i-M} + \cdots + r_y(-1) \sum_{i=0}^{P} \tilde{a}_i \tilde{a}_{i-1} + \\ &\quad + r_y(0) \sum_{i=0}^{P} \tilde{a}_i \tilde{a}_i + \\ &\quad + r_y(1) \sum_{i=0}^{P} \tilde{a}_i \tilde{a}_{i+1} + \cdots + r_y(M) \sum_{i=0}^{P} \tilde{a}_i \tilde{a}_{i+M} \\ &= r_{\tilde{a}}(-M) r_y(-M) + \cdots + r_{\tilde{a}}(0) r_y(0) + \cdots + r_{\tilde{a}}(M) r_y(M) = \sum_{k=-P}^{P} r_{\tilde{a}}(k) r_y(k) \end{aligned}$$

Lembrando $r(k) = r(-k)$, poderemos escrever:

$$\alpha = r_{\tilde{a}}(0) r_y(0) + 2 \sum_{k=1}^{P} r_{\tilde{a}}(k) r_y(k)$$

Analogamente,

$$\delta = r_a(0) r_y(0) + 2 \sum_{k=1}^{P} r_a(k) r_y(k)$$

Portanto, verificamos que α e δ podem ser obtidos facilmente a partir das seqüências de autocorrelação.

Em [55], Gray estabelece uma relação aproximada entre a métrica $D_2^{(CD)}$ e a razão de verossimilhança para pequenas distorções ($|\mathcal{V}_{xy}^{(CD)}(\omega)| \ll 1$):

$$(D_2^{(CD)})^2 \approx 2(\delta/\alpha - 1)$$

Gray *et al.* verificaram que para distorções na faixa de 0 a 2 dB, ambas eram praticamente idênticas. Porém, para distorções de até 6dB, eles identificaram um índice de correlação de 0.95. Isso implica que a relação entre $D_2^{(CD)}$ e δ/α acima vale, no caso de sinais de voz, para distorções maiores que as previstas teoricamente.

Dissimetria: Um problema da razão de verossimilhança é que ela não satisfaz o critério de simetria [55, 59]. Quackenbush *et al.* sugerem o uso da média entre as razões de verossimilhança, definindo uma razão de verossimilhança combinada:

$$l'(x,y) = \frac{l_{xy} + l_{yx}}{2}$$

onde

$$l_{yx} = \log(\delta'/\alpha') = \frac{\vec{a}_y R_x \vec{a}_y^t}{\vec{a}_x R_x \vec{a}_x^t}$$

A razão de verossimilhança é medida localmente. Para um sinal completo, pode-se calcular a média dos segmentos e associar a cada segmento m um peso $W(m)$ [59]:

$$LR = \frac{\sum_{m=1}^{M} W(m) l_{xy}}{\sum_{m=1}^{M} W(m)} \quad \text{ou} \quad LR' = \frac{\sum_{m=1}^{M} W(m) l'_{xy}}{\sum_{m=1}^{M} W(m)}$$

5.4.5 Medida cosseno hiperbólico

Itakura *et al.* [63] sugerem ainda uma outra definição da razão de verossimilhança, dada por:

$$\Xi = \frac{1}{2\pi} \int_{-\pi}^{+\pi} [e^{\mathcal{V}_{xy}(\omega)} - \mathcal{V}_{xy}(\omega) - 1] d\omega$$

onde $\mathcal{V}_{xy}$ é definida de acordo com o modelo LPC da figura 5.6 e com a equação 5.5.

Normalizando-se a energia de ambos os sinais de modo a fazer $\sigma_x = \sigma_y$, então Gray [55] demonstra que:

$$\delta/\alpha = 1 + \Xi$$

e

$$\delta'/\alpha' = 1 + \Xi'$$

onde:

$$\Xi' = \frac{1}{2\pi} \int_{-\pi}^{+\pi} [e^{-\mathcal{V}_{xy}(\omega)} + \mathcal{V}_{xy}(\omega) - 1] d\omega$$

Note-se que Ξ e Ξ' também são assimétricas. A sua média, entretanto, será simétrica e é conhecida como *medida cosseno hiperbólico*, ou simplesmente *cosh*:

$$COSH = \frac{1}{2}[\Xi + \Xi'] = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \{\cosh[\mathcal{V}_{xy}(\omega)] - 1\} d\omega = \frac{1}{2}[\frac{\delta}{\alpha} + \frac{\delta'}{\alpha'} - 2]$$

5.4.6 Outras medidas espectrais

Várias outras medidas espectrais podem ser elaboradas a partir dos coeficientes LPC ou de outros coeficientes transformados. Elas são de pequena importância, pois têm sido demonstrada a sua baixa correlação com as avaliações subjetivas [55, 63].

5.5 Medidas Psicoacústicas

Inicialmente descreveremos o modelo de percepção do som e depois descreveremos algumas medidas baseadas no espectro perceptual[9].

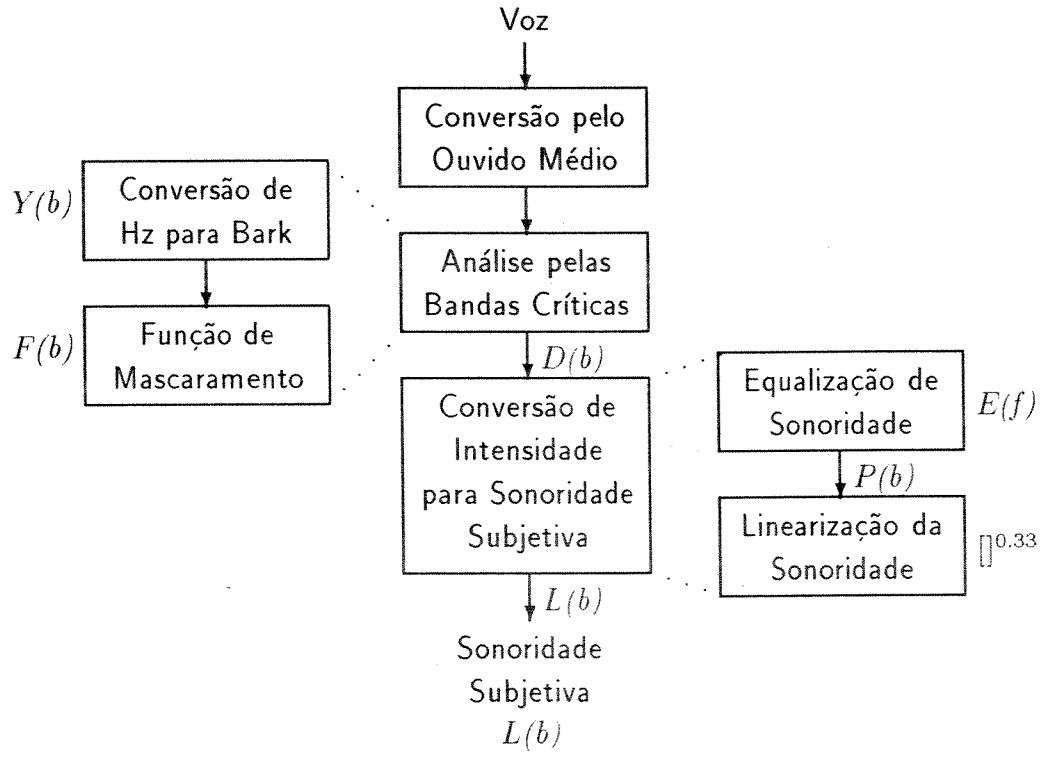


Figura 5.7: Diagrama em blocos do modelo psicoacústico de percepção da fala.

5.5.1 Modelos Psicoacústicos

Os modelos psicoacústicos tentam modelar o processo de percepção da fala através da emulação das funções biológicas associadas ao pré-processamento do sinal acústico pelo ouvido. Esse pré-processamento é uma atividade "objetiva", pois envolve a transformação do sinal acústico que chega ao ouvido externo em impulsos elétricos a nível dos feixes de neurônios distribuídos ao longo da cóclea. O processamento "subjetivo" será realizado pelas funções superiores do córtex cerebral, baseado neste sinal condensado gerado pelo ouvido [56]. Esses processos ocorrem de uma maneira bastante homogênea de uma pessoa para outra. Por isso,

[9] Ver nota à página 17

um bom modelo dos processos envolvidos na percepção da fala pode ser aplicado para um vasto número de pessoas [56]. Um diagrama em blocos do modelo discutido nessa seção encontra-se na figura 5.7.

Fisiologia. Na figura 5.8 temos um corte longitudinal do ouvido. Em termos fisiológicos, a onda acústica que chega ao pavilhão auditivo é transformada em movimento das estruturas ósseas que compõe o ouvido médio (martelo, bigorna e estribo). Os ossos do ouvido médio estimulam a cóclea através da *janela oval*, fazendo com que seu líquido interno se movimente. A cóclea pode ser modelada como um tubo com duas câmaras separadas por uma membrana chamada *membrana basilar*, conforme ilustrado na figura 5.9. Na extreminade oposta à janela oval, existe um orifício sobre a membrana basilar que comunica essas duas câmaras, chamado de *helicotrema* (ver Fig.5.9). A membrana basilar apresenta uma resistência que varia ao longo de sua extensão: próximo à janela oval ela é mais fina e tensa, ressoando em freqüências mais altas, enquanto que ao seu final (ápice), ela é espessa e flácida, ressoando então para freqüências mais baixas. Sobre a membrana basilar existem ainda duas estruturas: as fibras basilares e o órgão de Corti. As *fibras basilares* são cerca de 20000 espinhas delgadas com comprimentos que variam ao longo da membrana, sendo mais curtas junto à janela oval e mais longas no ápice da cóclea [71]. As ondas geradas pelo estribo viajam ao longo da cóclea, fazendo vibrar a membrana basilar na mesma freqüência do sinal de entrada [72]. Com isso, as fibras basilares vibram e estimulam as *células ciliadas* que compõem o *órgão de Corti*, que por sua vez transformam o movimento das fibras basilares em impulsos nervosos. Estes são então transmitidos pelo nervo coclear para a região específica do córtex cerebral [71].

Modelo do ouvido médio. O ouvido médio pode ser visto como um transformador que converte um sinal proveniente de um meio de baixa impedância acústica (ar) para um meio de alta impedância acústica (líquido coclear) [72]. Além disso, sua estrutura mecânica provoca a supressão de ondas acústicas de nível muito elevado, que danificariam as estruturas do ouvido interno [73].

O processo de transformação do sinal acústico nas ondas do líquido coclear é chamado de *função de transferência do ouvido médio*. Ele é equivalente a uma filtragem passa-baixas com corte em 5kHz, com um *"overshoot"* de 2000 a 5000 Hz e um pico em torno dos 3500Hz [74, 75]. Como essa filtragem não altera muito o espectro, ela é em geral desconsiderada para sinais com faixa até 5000Hz [74] e níveis acústicos não muito elevados.

A cóclea e as bandas críticas [73]. Cada ponto da membrana basilar é mais sensível a uma determinada freqüência, chamada de *freqüência característica*. Para um ponto específico da membrana basilar, a curva de resposta à freqüência de vibração do sinal presente na janela oval é equivalente à de um filtro passa-faixa com fator de qualidade[10] aproximadamente constante, resultando numa melhor resolução nas baixas freqüências. As fibras basilares localizadas na região de alta freqüência característica respondem a uma maior faixa de freqüências do que as fibras na região de baixa freqüência característica.

Um comportamento similar é obtido ao se traçar a curva de resposta ao longo da membrana basilar para um tom numa freqüência específica, o que é ilustrado na Figura 5.9. Para cada

[10] *Fator de Qualidade* é, por definição, a razão entre a freqüência central e a largura de faixa de um filtro passa-faixa.

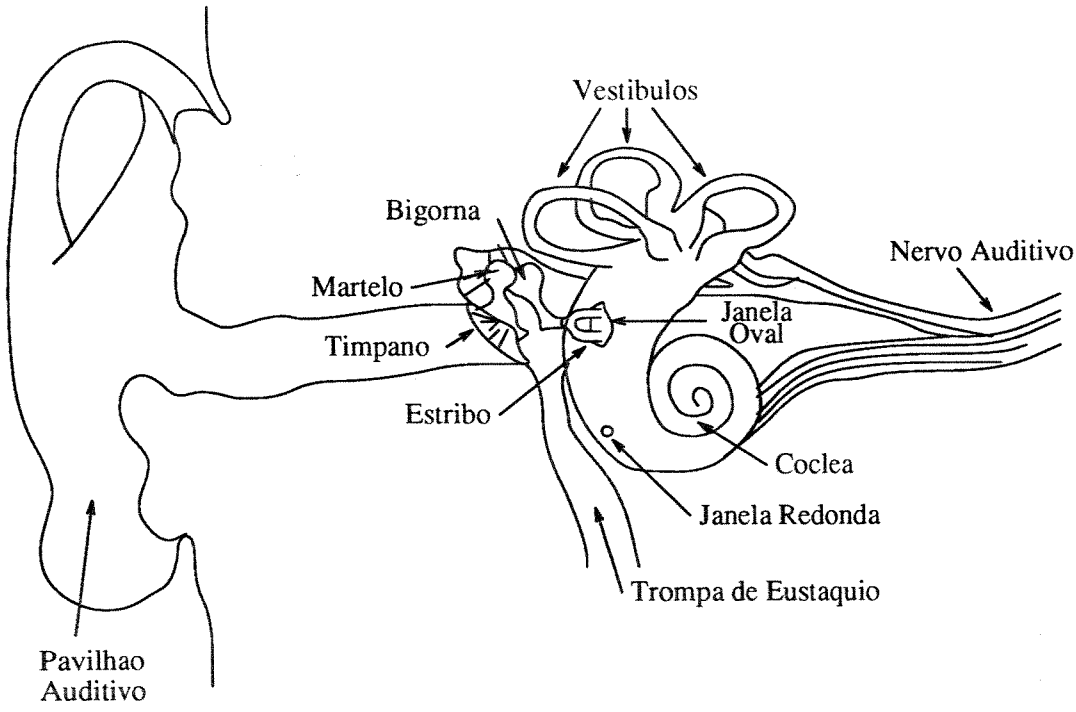


Figura 5.8: Diagrama de um corte longitudinal do ouvido, mostrando as principais estruturas: o ouvido externo, os ossículos do ouvido médio (martelo, bigorna e estribo), a cóclea, com sua janela oval e os feixes neuronais que levam o sinal acústico ao córtex.

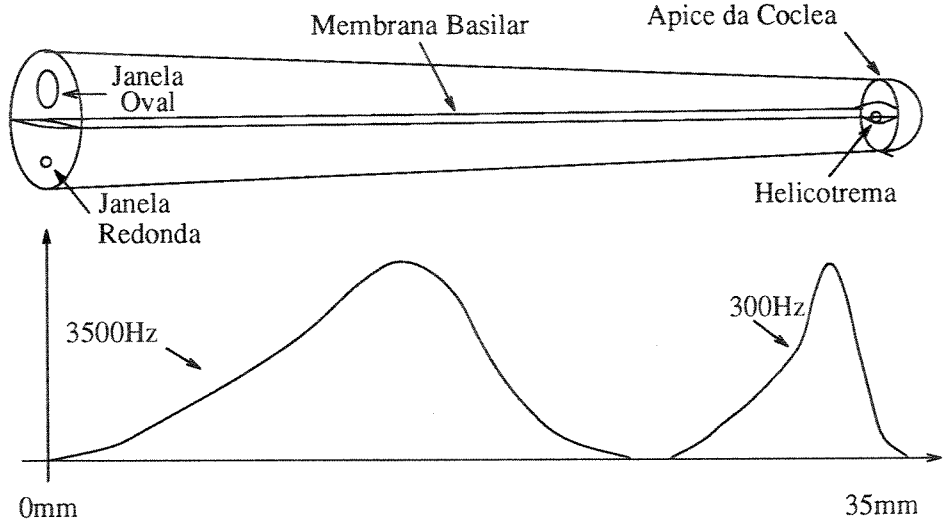


Figura 5.9: Diagrama mostrando a cóclea como um tubo mais largo em sua base (junto à janela oval) e mais estreito em seu ápice. Está ilustrada também a membrana basilar, com seus 35mm de comprimento. No gráfico está ilustrado a função característica de vibração da membrana basilar para duas freqüências diferentes. Note-se que freqüências mais baixas ressoam mais ao final da membrana, enquanto que freqüências mais altas ressoam mais ao seu início.

freqüência, há um ponto da membrana basilar em que a vibração é máxima. A posição desse ponto, medida a partir do helicotrema, é aproximadamente proporcional ao logaritmo da freqüência do som.

Um aspecto importante da audição é o *mascaramento*, em que um som fica inaudível na presença de outro som [76]. Muitos fenômenos de mascaramento podem ser explicados em termos de faixas de freqüências conhecidas como *bandas críticas* e que foram determinadas através de experimentos psicoacústicos. Uma banda crítica define uma faixa de freqüências em que a percepção de um estímulo muda abruptamente na medida em se modifica um estímulo sonoro de faixa estreita, de modo que este tenha componentes fora dessa banda. Quando dois sinais estimulam essa banda crítica, o de maior energia dominará a percepção e mascarará o outro estímulo sonoro[11].

Os filtros que definem as bandas críticas possuem cortes bem acentuados: 65 dB/oitava para as Bandas Críticas em torno de 500Hz e 100 dB/oitava em torno de 8kHz. A largura de faixa das bandas críticas é de aproximadamente 100Hz para freqüências abaixo de 500 Hz e de aproximadamente 1/6 da freqüência central da banda para freqüências acima de 1000Hz (i.e., em direção à janela oval) [75].

A largura de faixa das bandas críticas corresponde a um espaçamento uniforme de 1,5mm ao longo da membrana basilar, sugerindo que $35 \div 1,5 \approx 24$ filtros passa-faixa com largura de faixa crescente com a freqüência modelariam bem a membrana basilar. Zwicker [56] definiu uma escala perceptual, denominada *escala Bark*, que relaciona as freqüências acústicas à resolução perceptual de freqüência, de modo que 1 Bark cobre uma Banda Crítica. Schroeder definiu a relação entre as freqüências f em Hz e valores b na escala Bark através da equação [78, 75]:

$$Y(b) = f = 600 \sinh(b/6) \tag{5.10}$$

ou, equivalentemente:

$$\mathcal{Y}(f) = b = 6 \ln(f/600 + \sqrt{f^2/600^2 + 1}) \tag{5.11}$$

Com essa transformação, o formato de cada banda crítica é independente da freqüência [56] e dois picos adjacentes ficam, por definição, espaçados de 1 Bark. Adicionalmente, duas bandas críticas adjacentes devem se encontrar num ponto 3dB abaixo do valor de pico [56]. Esse formato "normalizado" é denominado *função de espalhamento da membrana basilar* [76] e é denotado por $F(b)$. Diversos autores apresentam diferentes formulações para $F(b)$, baseados nos dados experimentais de Zwicker. Essas formulações divergem consideravelmente entre si. Alguns aproximam a curva por segmentos de reta, outros usam funções analíticas. Como um exemplo, teríamos para cada uma das B bandas críticas [78]:

$$F_l(b) = \begin{cases} 0 & b - b_l < -1.3 \\ 10^{2.5(b-b_l+0.5)} & -1.3 \leq b - b_l \leq -0.5 \\ 1 & -0.5 < b - b_l < 0.5 \\ 10^{1.0(b-b_l-0.5)} & 0.5 \leq b - b_l \leq 2.5 \\ 0 & b - b_l > 0.5 \end{cases} \tag{5.12}$$

onde l se refere à l-ésima Banda Crítica e b_l é a freqüência central, em Bark, dessa Banda. O número total B de bandas críticas depende da faixa dos sinais considerados no modelo. Se

[11] Este efeito é empregado em vários codificadores de voz [37, 77, 35] através do uso de técnicas de pós-filtragem [51].

um sinal tiver largura de faixa de $F_N/2$ Hz, onde F_N é a sua freqüência de Nyquist, então o número B de bandas críticas será:

$$B = \mathcal{Y}(F_N/2) \tag{5.13}$$

Para sinais entre 0 e 5 kHz, a escala Bark varia de 0 a 16.9 Barks [78]. Para a faixa audível como um todo, normalmente considera-se a escala Bark indo de 0.5 a 24.5 Barks [74].

O padrão gerado ao longo da cóclea, $D(b)$ por um sinal de faixa estreita (e.g. um tom de freqüência f) pode ser modelado como a convolução desse sinal em Bark, $Y(b)$, com a função de espalhamento da membrana basilar, $F(b)$ [74]:

$$D(b) = F(b) * Y(b)$$

O sinal $D(b)$ é chamado de *padrão de excitação* [56, 74] e pode ser entendido como a distribuição de energia ao longo da membrana basilar [74]

A geração de padrões de excitação para sinais de faixa larga (e.g. voz), apesar de ser um fenômeno complexo e não linear, é geralmente aproximado pelo mesmo modelo linear da equação acima [76].

Escalas de Sonoridade. Com esse sinal em Bark, ainda é necessário considerar que nem todas as bandas críticas têm mesmo ganho, isto é, a sensibilidade não é uniforme ao longo da membrana basilar. De fato, conhece-se que a resposta auditiva humana começa a cair sensivelmente acima dos 5000Hz (–18dB/oitava) [78]. A ponderação do padrão de excitação pela sensibilidade da membrana basilar equivale à conversão de um sinal que está numa escala de *intensidade* (e.g. dBSPL) para uma escala de excitação, ou de *sonoridade*, expressa em *fonons*. Por definição, a sonoridade em fonons de um tom com certa freqüência e nível é a intensidade em dB de um tom de 1 kHz que *soe* igualmente a esse tom [56].

Essa última conversão se baseia em dados experimentais obtidos para tons [56], que mapeiam a potência do sinal para a sua sonoridade. Essa conversão tem sido modelada como um processo de filtragem do padrão de excitação pela *curva de correção de sonoridade*, $E(b)$ [78, 56]:

$$P(b) = E(b) \cdot D(b)$$

onde $P(b)$ é o sinal em fonons. Um exemplo de formulação da curva, expresso em Hz[12], é dada por [78]:

$$E(f) = (2\pi)^2 \frac{(f^2 + 1200^2) f^4}{(f^2 + 400^2)(f^2 - 3100^2)} \tag{5.14}$$

Verificou-se ainda que a escala de sonoridade não é linear em relação à sonoridade percebida pelo ouvido humano. Como um exemplo, se um sinal tem sonoridade em torno de 40 fonons, ao se adicionar 10 fonons ao sinal, a sonoridade percebida dobrará; entretanto, se o sinal estiver próximo ao limiar de audição, a sonoridade percebida decuplicará [56]! É útil então converter-se esta escala não linear de sonoridade em fonons, para uma outra, linear, em sonons. Por definição, 1 *sônon* é o aumento de potência que faz dobrar a sonoridade percebida [56]. A função de conversão de fonons para sonons é uma função de compressão não linear que em [78] é aproximada por:

$$L(b) = [P(b)]^{0.33} \tag{5.15}$$

[12] O modelo acústico é mais facilmente implementado se E estiver no domínio da freqüência, ao invés de na escala Bark [56].

onde $L(b)$ é a *sonoridade subjetiva* (percebida pelo ouvinte), em contraste com a sonoridade "objetiva" $P(b)$. Note-se que ambas são função da região b da membrana basilar em que estão mapeadas.

5.5.2 Medidas de Distorção baseadas no modelo Modelo Psicoacústico

A sonoridade subjetiva $L(b)$ representa a densidade espectral de potência perceptual, como é enviada para os processos de percepção do córtex cerebral. Portanto, a partir de $L(b)$ podem ser empregadas várias das propriedades aplicáveis a espectro de potência de sinais, como predição linear [78], distância espectral perceptual [56], distância cepstral, etc.

Distância Espectral Perceptual. Sejam os sinais de referência x e de teste y, cuja faixa é de $\mathcal{F}$ Hz, dos quais calculou-se a densidade espectral de potência perceptual, ou sonoridade subjetiva, L_x e L_y. Pode-se então definir a Distância Espectral Perceptual como para um sinal com B Bandas Críticas:

$$PSD^2 = \sum_{b=1}^{B} \mathcal{V}_{xy}^2(b) \triangleq \sum_{b=1}^{B} [L_x(b) - L_y(b)]^2$$

ou

$$PSD = \sqrt{\sum_{b=1}^{B} [L_x(b) - L_y(b)]^2} \tag{5.16}$$

Esta distância espelha a diferença entre a densidade espectral percebida para os sinais de teste e de referência.

Distância Cepstral Perceptual. Uma variante de medida perceptual pode ser obtida ao se definir $\mathcal{V}_{xy}$ como:

$$\mathcal{V}_{xy}(b) \triangleq \log_{10}[L_x(b)] - \log_{10}[L_y(b)]$$

Como $L(b)$ equivale à densidade espectral (perceptual) de potência, a $\mathcal{V}_{xy}$ acima é análoga à definição de $\mathcal{V}_{xy}^{(CD)}$ na página 100, podendo ser vista como sua versão perceptual com a ressalva de que aquela distância cepstral se referia somente ao modelo do trato vocal e aqui ela engloba tanto o modelo como a excitação. Portanto, podemos definir a distância cepstral perceptual em dB como sendo:

$$PCD = 10\sqrt{\sum_{b=1}^{B} \{\log_{10}[L_x(b)] - \log_{10}[L_y(b)]\}^2} \tag{5.17}$$

5.6 Outras medidas

Há ainda várias medidas objetivas de distorção que estão sendo estudadas pelo CCITT, além da distância cepstral. O PTT Holandês propôs a *Medida Perceptual de Qualidade de Voz (PSQM)* [79], que se baseia em critérios perceptuais e lembra a SNR_S da seção 5.3.2. Uma

outra, chamada de PRBA *(Pattern Recognition Based Assessment)*, utiliza técnicas de reconhecimento de padrões para, utilizando um "banco de distorções" *(distortion database)*, realizar a montagem de medidas compostas, sempre baseado em outras medidas já existentes.

Há ainda duas outras em estudo, que podem ser consideradas mistas por incorporarem estimações de freqüência, aspectos perceptuais e medidas temporais: o Índice de Informação *(II)*, proposto pelo CNET (França), e a Função de Coerência *(CF)*, proposto pela BNR (Canadá). Estas duas serão sucintamente descritas a seguir.

5.6.1 Índice de Informação

O Índice de Informação [80, 64] tenta considerar em sua formulação as perdas de transmissão, o ruído ambiente, atenuações, distorção de freqüência e o efeito local. Ela se baseia na teoria da informação de Shannon. O sistema auditivo aqui é modelado pela divisão do espectro em 16 bandas críticas, às quais se aplicam ponderações e limiares de audição obtidos empiricamente.

A relação sinal-distorção *(SDR)*, denotada $QS(i)$, é computada inicialmente para cada uma das 16 bandas do modelo:

$$QS_k(i) = 10\log_{10}\frac{\sum_{j\in b_i}|X(f_j)|^2}{|\sum_{j\in b_i}|X(f_j)|^2 - \sum_{j\in b_i}|Y(f_j)|^2|}$$

onde j cobre todas as freqüências especificadas para a i-ésima banda. Aqui, $X(f)$ e $Y(f)$ são a DFT de um dado segmento k dos sinais de referência e de teste. As 16 bandas de freqüência b_i são tabuladas. Tratando as bandas como canais independentes e separados, calcula-se a informação mútua (isto é, a capacidade máxima do canal) de cada banda, sendo então ponderada e somada para compor o Índice de Informação total, RII:

$$RII = \sum_{i=1}^{16} W_2(i)\frac{3}{0.1 + 10^{-[(\overline{QS}(i)+W_1(i))/10]}}$$

onde $\overline{QS}(i)$ é o valor médio de $QS(i)$ para todos os segmentos do sinal e $W_1(i)$ e $W_2(i)$ são funções de ponderação tabuladas que levam em conta respectivamente a largura de banda crítica e a importância perceptual da banda i.

A partir disso, o RII é mapeado em estimativas de qualidade através de uma série de equações não lineares oriundas da teoria da informação, que podem ser encontradas em [64].

5.6.2 Função de Coerência

A função de coerência [64] é uma medida da relação entre sinal e distorção *(SDR)* levando em conta a sensibilidade de audição, efeitos de limiar de ruído (mascaramento) e sensibilidade do terminal receptor. Neste método, segmentos de voz são classificados em quartis[13]. Em seguida, computa-se os espectros $S_{xx}(f)$ e $S_{yy}(f)$ dos sinais de referência e de teste, respectivamente,

[13] Os segmentos dos sinais são classificados de acordo com a sua potência em 4 classes *ou quartis*. Cada quartil corresponde a uma divisão uniforme da potência do segmento: o quartil 1 se refere aos segmentos com potência entre 0% a 25% do valor máximo possível, o quartil 2 para segmentos com potências entre 25% e 50%, etc.

bem como o espectro cruzado $S_{xy}(f)$, para todos os segmentos de cada cada quartil, após o que calcula-se a média para cada quartil. Os espectros médios de cada quartil q são usados para se computar a função de coerência $\gamma_q^2(f)$ do quartil q:

$$\gamma_q^2(f) = \frac{|S_{xy}^{(q)}(f)|^2}{S_{xx}^{(q)}(f)S_{yy}^{(q)}(f)}$$

A função de coerência pode ser interpretada como a correlação entre os sinais para uma dada freqüência f.

A seguir, a potência do sinal de referência:

$$GX_q(f) = \gamma_q^2(f)|S_{yy}^{(q)}(f)|^2$$

e de teste

$$GY_q(f) = [1 - \gamma_q^2(f)]|S_{yy}^{(q)}(f)|^2$$

é estimada a partir de $\gamma(f)$ e usada para se calcular a SDR modificada, $\zeta_q(f)$:

$$\zeta_q(f) = \frac{GX_q(f)W_{F2}(f)}{GY_q(f)W_{F2}(f) + W_{F1}(f)}$$

Nessas equações, $W_{F1}(f)$ e $W_{F2}(f)$ são funções de ponderação para o limiar de audição e para a sensibilidade do terminal receptor. A estimativa da qualidade subjetiva em termos de MOS é então dada por:

$$MOS_{est} = W_L \left(\sum_{q=1}^{4} W_{Qq}(f) \sum_{f=94Hz}^{4000Hz} W_{Fq}(f) W_{Gq}(\zeta(f)) \right)$$

onde W_L é uma função não linear para mapear a medida objetiva em um valor MOS estimado (MOS_{est}), $W_{Qq}(f)$ pondera a importância de cada quartil de amplitude, $W_{Fq}(f)$ é uma função de ponderação da importância perceptual de cada freqüência e $W_{Gq}(\zeta(f))$ é uma outra função não linear.

5.7 Sumário

Neste Capítulo descrevemos em detalhes as medidas de distorção mais significativas encontradas na literatura. Apresentamos desde a SNR clássica até as mais recentes medidas envolvendo modelos psicoacústicos de percepção da fala. Apresentamos também, mas de modo resumido, algumas das medidas em estudo atualmente pelo CCITT.

Capítulo 6

Análise de Alguns Algoritmos

6.1 Introdução

Neste Capítulo descreveremos inicialmente alguns dos testes subjetivos realizados para um vasto material de voz durante o processo de seleção e avaliação do codificador de voz a 16 kbit/s do CCITT, o LD-CELP da Recomendação G.728.

Após isso, descreveremos alguns detalhes de implementação e os resultados de algumas das medidas de distorção objetivas que descrevemos no Capítulo 5.

Será então apresentada uma análise comparativa entre as medidas objetivas e os resultados das avaliações subjetivas.

Já ao final, diversas conclusões sobre a aplicabilidade das medidas objetivas serão apresentadas.

Antes da leitura deste capítulo deve ser feita uma advertência: muitos dos termos e conceitos a serem utilizados nesta seção foram definidos e explicados nos Capítulos 4, sobre testes subjetivos, e 3, sobre aspectos de infra-estrutura. Convém que o leitor esteja familiarizado com esses capítulos, para melhor acompanhamento.

6.2 Testes Subjetivos

Como já descrevemos no Capítulo 2, o processo de definição do algoritmo de codificação de voz a 16 kbit/s do CCITT teve duas fases, chamadas de Fase I e Fase II. A segunda fase foi necessária para o aperfeiçoamento do algoritmo testado na Fase I.

O CPqD participou em 2 dos experimentos da Fase I (Experimento 2, "*Efeito de múltiplas transcodificações e do nível de entrada sobre a qualidade*", e Experimento 4, "*Avaliação da dependência com o locutor*") e no experimento único da Fase II. O material fonte de voz fornecido ao Laboratório Central de Processamento[1] encontra-se descrito a seguir, bem como um diagrama de seu processamento. Para fins da análise objetiva da qualidade e de sua correlação com a qualidade subjetiva, descreveremos somente um dos testes subjetivos, o referente à Fase II, por incluir uma maior variedade de tipos de distorção.

[1] Comsat Laboratories, Clarksburg, EUA.

6.2.1 Material de Voz

Os testes para o LD-CELP foram realizados a partir de sinais armazenados tanto no formato analógico (usando Sony-Betamax com padrão PAL) como no formato digital (voz digitalizada a 16 kHz com 16 bits de resolução segundo a especificação dada em[117]). O processamento foi do tipo analógico e seu diagrama básico está na Figura 6.1 [129]. O material gravado no formato analógico era processado pelo LD-CELP ou por condições de referência e então gravado em outro dispositivo analógico (neste caso, um DAT). Por outro lado, o material digitalizado armazenado em um PC (dotado de uma placa conversora A/D e D/A) precisava de duas etapas adicionais, como ilustrado na Figura 6.2: era convertido do formato digital para o analógico antes de ser processado e do formato analógico para o digital após o processamento[2]. Note-se que todo o material, antes de passar pelo *"circuito básico"*, tinha seu nível equalizado por um circuito passivo para ajustar o nível de saída de cada um dos dispositivos de armazenamento com o nível "esperado" pelo circuito básico. Além disso, efetuava-se uma filtragem passa-baixas em todos os materiais antes e depois de serem processados, independentemente da forma de armazenamento. Isto era feito para se garantir um processamento uniforme para todos os sinais-fonte submetidos.

No caso especial da Fase II de testes (ver descrição na seção 6.2.2), todos os materiais a serem processados estavam no formato digital [129]. No Plano de Teste [90], por sua vez, havia uma atribuição entre listas e condições pelas quais os processamentos seriam realizados e em [117] está definida a regra de batismo para o material de voz.

Cada *elemento* do material de voz era composto de pares de sentenças extraídas de noticiários de TV, de jornais e de outras fontes "populares"[3]. O conjunto total de sentenças foi embaralhado para tornar desconexas as sentenças de cada par e o conjunto de pares de sentenças foi dividido entre quatro locutores, escolhidos dentro do grupo de Processamento Digital de Voz do CPqD. O conjunto de sentenças lido por eles foi distinto. Com esse material em mãos, digitalizou-se a leitura num ambiente de baixo ruído. Esse material digitalizado foi filtrado em software pelo filtro IRS e armazenado em fitas magnéticas. A partir desta, no Laboratório Central de processamento para os Testes, foram processados de acordo com os Planos de Testes Subjetivos [91, 90]. Um diagrama em blocos dos processamentos pode ser visto na figura 6.1; aqueles referentes à Fase II encontram-se sumarizados na Tabela 6.2.

Todos os nossos arquivos gerados tinham 7 segundos de duração, porém em geral apresentando um trecho de silêncio em seu início e fim. Em termos de tamanho global, o material fonte para a Fase II compreendeu 144 arquivos que, a uma taxa de amostragem de 16 kHz, somavam 31MBytes (16m48s). Processado, esse volume de material cresceu para 378MBytes (3h22m).

6.2.2 Descrição da Fase II de Testes Subjetivos

A Fase II de Testes Subjetivos do LD-CELP visava verificar se a versão do LD-CELP que fora melhorada em relação àquela da Fase I passara a atender a todos os requisitos de qualidade especificados (ver Tabela 6.1).

O Plano de Testes foi mantido o menor possível, pois as modificações introduzidas no algoritmo

[2] O material digitalizado processado era temporariamente armazenado nos discos rígidos do PC e posteriormente transferido para discos ópticos do tipo WORM (*Write Once, Read Many*).

[3] Um exemplo de tais sentenças encontra-se no Anexo A do Capítulo 4.

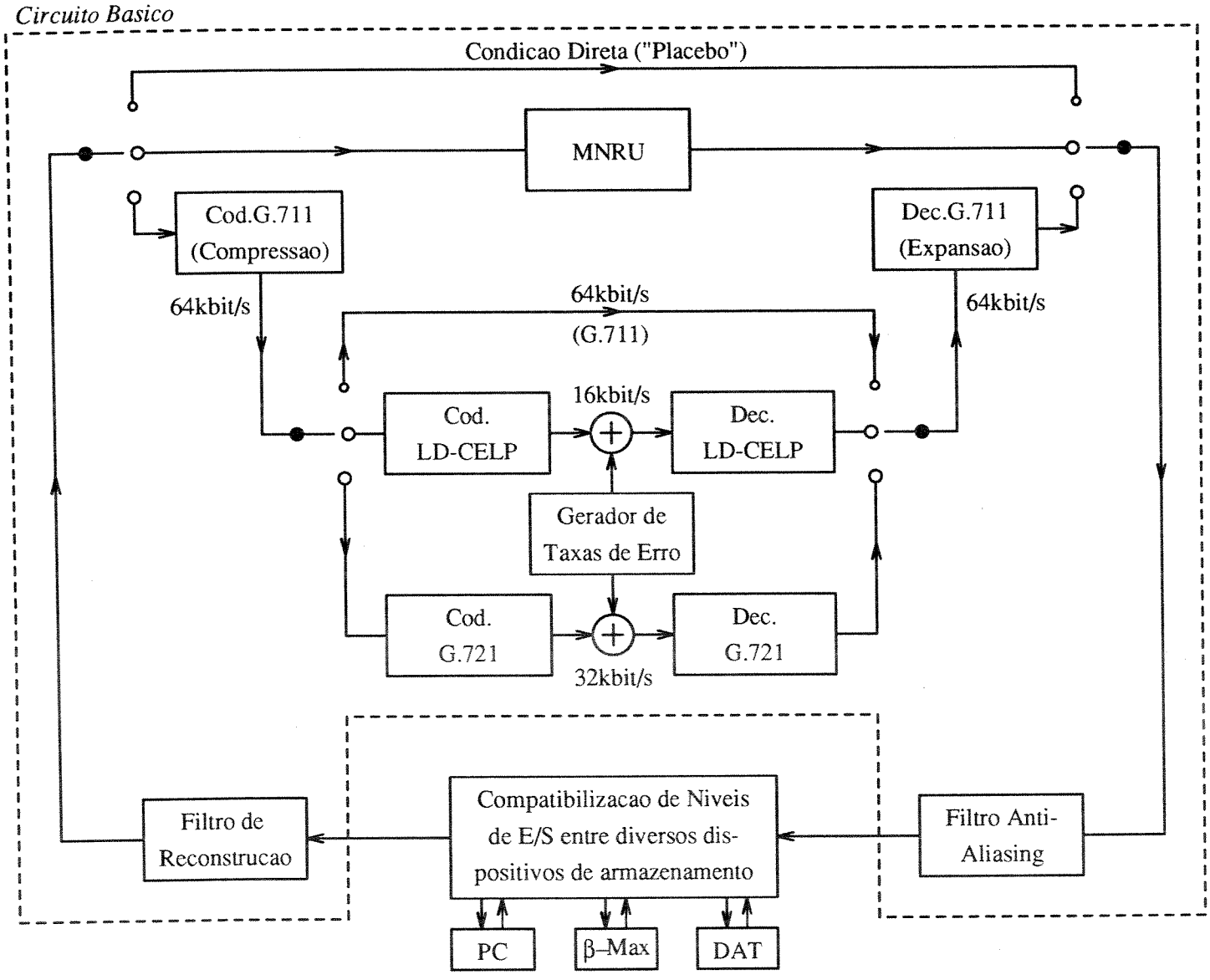


Figura 6.1: Esquema de processamento e armazenamento utilizado no Laboratório Central para a padronização do LD-CELP.

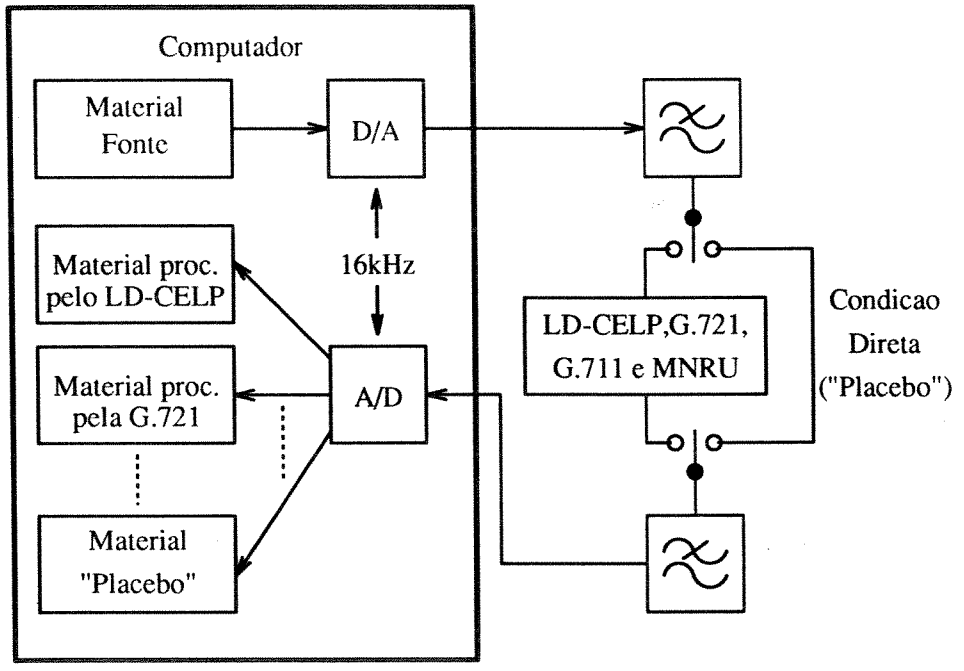


Figura 6.2: Procedimento simplificado de processamento do material fonte pelo Laboratório Central: condições de teste (LD-CELP, G.711, G.721 e MNRU) e de referência (direta ou *"placebo"*).

foram mínimas e havia severas limitações materiais e de prazos no cronograma do processo de padronização. Com isso, houve apenas 1 experimento (contra os 5 experimentos da Fase I) chamado "Efeito de Erros em Bits, de Transcodificações e do Nível de Entrada", em que se utilizou a técnica ACR.

A Fase II avaliou o desempenho do LD-CELP para 24 condições distintas em que se variavam os seguintes parâmetros:

- nível de entrada dos sinais originais no LD-CELP e nas condições de referência: nível nominal (–20dBm0), 10 dB acima e abaixo do nominal (–10dBm0 e –30dBm0);
- número de transcodificações síncronas e assíncronas (dependendo do algoritmo);
- desempenho em presença de erros de transmissão aleatórios.

Além disso, testou-se o LD-CELP contra 4 condições de referência:

- G.711 (para lei μ);
- G.721 (versão do Livro Azul, 1988);
- MNRU (Hardware da British Telecom/Malden Electronics segundo a P.81);
- condição direta (i.e., processamento somente pelo circuito básico, sem passar por qualquer um dos outros algoritmos acima ou pelo LD-CELP – ver Figura 6.2).

Tabela 6.1: Requisitos e Objetivos de Desempenho para o LD-CELP. *BER* é a taxa de erros de transmissão, *qdu* é a unidade de distorção de quantização da Rec.CCITT G.113 e *DTMF (Dual Tone Multi Frequency)* é a sinalização de chamada alternativa à decádica.

No.	Parâmetro	Requisito	Objetivo
1	Taxa de Bits	16kbit/s	
2	Atraso	$\leq$5ms	$\leq$2ms
3	Qualidade de voz para níveis de entrada e de audição nominais sem erros de transmissão	$\leq$4qdu	
4	Qualidade de voz com BER=1% e 0.1%	Não pior que a G.721	
5	Dependência da qualidade de voz com o nível de entrada	Não pior que a G.721	
6	Desempenho com múltiplas transcodificações	3 assíncronas $\leq$14qdu	Propriedade de tandem síncrono
7	Transmissão de Sinalização	Sinalização CCITT no.5, 6 e 7 e DTMF	A menor distorção possível
8	Transmissão de Sinais de Modem		A maior taxa possível com BER satisfatória
9	Transmissão de Música		Não introduzir efeitos desagradáveis

Uma lista sumarizando todas as condições e combinação de fatores encontra-se na Tabela 6.2. Nela podemos ver os 4 fatores definidos no Capítulo 4 para testes baseados em quadrados greco-latinos: Condições de Processamento (Algoritmos), Listas de Material (Locutores), Ouvintes e Sequencia de apresentação. Como o LD-CELP poderia apresentar dependência com o sexo do locutor, este também foi incluído como um fator para a Anova. Adicionalmente, por ser o teste baseado em *intercalamento* e ser dividido em duas Partes, acresceram-se outros dois fatores: Partes e Ouvintes dentro das Partes. Finalmente, como os materiais foram avaliados em três diferentes níveis de audição, este é um outro fator a ser considerado na Anova. Assim, neste teste houve oito fatores a avaliar na análise de variâncias, além das interações de 5 deles (Sexo, Ouvintes, Sequencia, Condições e Listas) com o Nível de Audição. No Capítulo 4 fizemos a definição para um teste mais geral, porém mais simples, em que somente 4 fatores estão presentes. Por isso, a Anova para a Fase II de Testes do LD-CELP teve que ser mais elaborada e encontra-se descrita em [90]. Os resultados de sua análise de variâncias está apresentada mais à frente.

O projeto deste teste subjetivo foi feito usando quadrados greco-latinos. Se se utilizasse um quadrado convencional, as 24 condições requeririam um quadrado de ordem 24. Duas seriam as conseqüências disso: o teste ficaria muito longo para ser aplicado e o volume de material de voz processado seria monstruoso. Para aliviar estes dois aspectos, usou-se um projeto de teste com intercalamento. Neste caso, utilizou-se dois quadrados latinos: um para as condições cujos números seqüenciais eram ímpares *(condições ímpares)* e outro para as *condições pares*. Isto permitiu reduzir à metade o volume de material de voz gerado. Do ponto de vista das medidas objetivas, isto é importante pois, com o biparticionamento do material processado em duas tabelas, surgem dois blocos de materiais processados: um com

Tabela 6.2: Lista de Fatores e Condições de Referência da Fase II de Testes do LD-CELP.

Algoritmos	Número	Comentário
LD-CELP:		
Taxas de Erro (BER)	3	BER=0%, 0.1% e 1%, aleatórios
Níveis de Voz de Entrada	3	−10, −20 e −30 dBm0 (Rec.CCITT G.711)
Transcodificações	3	1, 2 e 4 transcodificações
Modos de conexão	1	Somente assíncrono
MNRU:	6	Q de 5, 10, 15, 20, 25 e 30 dB, todas a -20dBm0
G.711:		
Níveis de Voz de Entrada	1	−20 dBm0 (Rec.CCITT G.711)
Transcodificações	4	2, 4, 8 e 16 transcodificações
G.721:		
Níveis de Voz de Entrada	1	−20 dBm0 (Rec.CCITT G.711)
Transcodificações	3	1, 2 e 4 transcodificações
Modos de conexão	1	Somente assíncrona
Condição Direta:	1	Somente para -20dBm0
Variáveis Comuns:		
Número de Locutores	4	2 vozes masculinas e 2 femininas
Níveis de Audição	3	Preferido *(PLL=79dBSPL)* e 10 dB acima e abaixo do Nível de Audição Preferido.
Ouvintes	24	12 homens e 12 mulheres (leigos).
Partes	2	12 ouvintes na Parte A e 12 na Parte B, balanceados em relação ao sexo.

as condições pares e outro com as condições ímpares. Como um exemplo, a figura de mérito objetiva para o LD-CELP com 1 transcodificação, ao ser comparado com a da G.711 para 4 transcodificações (respectivamente as condições 1 e 12 da Fase II), serão baseadas em materiais de voz diferentes, enquanto que a mesma condição do LD-CELP, ao ser comparada com a da G.711 para 2 transcodificações (condição 11 da Fase II), basear-se-á no mesmo material fonte. Apesar de todos os codificadores em questão serem independentes do locutor, pode surgir aí uma fonte de erro quando das comparações. Um exemplo desse fenômeno é ilustrado por Sekey *et al.* [56], onde um algoritmo é considerado (subjetivamente) indistinto quanto ao sexo do locutor, mas sua medida objetiva apresenta diferenças. Esses dois conjuntos de material original de voz serão denominados **GL1** para as condições ímpares e **GL2** para as condições pares.

Outro aspecto a considerar é que somente as condições de operação *nominais* do algoritmo são de interesse para esta pesquisa inicial de medidas objetivas. Isto sugere que sejam consideradas nos estudos deste Capítulo somente as condições cujo processamento se deu no nível nominal (−20dBm0). Isto reduz as condições de interesse de 24 para 22.

Na Tabela 6.3 estão as condições de interesse da Fase II para nossos estudos, mostrando as condições processadas com nível de entrada nominal, o algoritmo, o número de transcodificações, o modo de operação, a taxa de erro *(BER)* e o grupo do material de voz original (**GL1** ou **GL2**), como definidas no Plano de Testes original do CCITT [90].

Tabela 6.3: Lista de Condições para o Nível Nominal de Processamento para a Fase II de Testes do LD-CELP. O termo *"tandem"* abaixo se refere ao número de transcodificações em cascata referente a cada condição.

Condição	Algoritmo	Tandem	Modo	BER	Material
1	LD-CELP	1	-	0	GL1
2	LD-CELP	1	-	1%	GL2
3	LD-CELP	1	-	0.1%	GL1
6	LD-CELP	2	assíncrono	0	GL2
7	LD-CELP	3	assíncrono	0	GL1
8	LD-CELP	4	assíncrono	0	GL2
9	LD-CELP+G.721+LD-CELP	-	assíncrono	0	GL1
10	G.721+LD-CELP+G.721	-	assíncrono	0	GL2
11	G.711	2	assíncrono	0	GL1
12	G.711	4	assíncrono	0	GL2
13	G.711	8	assíncrono	0	GL1
14	G.711	16	assíncrono	0	GL2
15	G.721	1	assíncrono	0	GL1
16	G.721	2	assíncrono	0	GL2
17	G.721	4	assíncrono	0	GL1
18	MNRU	1	Q=30dB	0	GL2
19	MNRU	1	Q=25dB	0	GL1
20	MNRU	1	Q=20dB	0	GL2
21	MNRU	1	Q=15dB	0	GL1
22	MNRU	1	Q=10dB	0	GL2
23	MNRU	1	Q=5dB	0	GL1
24	Direta	1	-	0	GL2

Pode-se perceber na Tabela 6.3 que o número de condições entre um grupo de listas e outro é balanceado, havendo 11 das condições sendo processado pelo bloco **GL1** e 11 pelo Bloco **GL2**.

6.2.3 Resultados dos Testes Subjetivos

O Experimento da Fase II foi realizado com 24 ouvintes utilizando-se a sala acústica disponível no CPqD da Telebrás. Os ouvintes foram selecionados entre empregados do CPqD/Telebrás de vários níveis de formação (técnicos, engenheiros e analistas de sistemas), todos leigos em relação a codificação de voz. De acordo com o Plano de Testes [90], o teste foi dividido em duas Partes, cada uma envolvendo 12 ouvintes.

Cada condição foi avaliada com 48 votos para cada nível de audição. No total foram coletados 3456 votos, dos quais nossa análise utilizará 3168 votos, relativos às 22 condições de interesse das tabelas 6.3 e 6.5 avaliadas nos 3 níveis de audição.

Na Tabela 6.4 está o resultado da análise de variâncias (Anova) [130, 99] para a Fase II de testes subjetivos do LD-CELP.

Na Tabela 6.5 encontram-se respectivamente o valor MOS e o intervalo de confiança *(CI)* associado, relativos às condições de interesse para a Fase II de Testes Subjetivos do LD-CELP. Essas estatísticas foram calculadas pelas relações descritas no Capítulo 4. Desses resultados

Tabela 6.4: Análise de variâncias para a Fase II de Testes do LD-CELP para a língua portuguesa. *HS* é Altamente Significativo, *S* é Significativo and *NS* é Não-Significativo; *N/A* é Não Aplicável. *Q* é a probabilidade das variâncias não serem significativamente diferentes (cf.Capítulo 4).

Fator do Teste	*Q*	Efeito	
		Observado	Esperado
Níveis de Audição:	0.00%	HS	HS
Sexo dos Locutores:	17.97%	NS	NS
Interação (Sexo × Nível):	23.44%	NS	NS
Ouvintes:	0.00%	HS	S/HS
Interação (Ouvintes × Níveis):	0.00%	HS	S/HS
Seqüência de Apresentação:	0.00%	HS	NS
Interação (Seqüência × Níveis):	0.31%	HS	NS
Condições:	0.00%	HS	S/HS
Interação (Condições × Níveis):	1.16%	S	S/HS
Listas:	0.00%	HS	NS
Interação (Listas × Níveis):	33.27%	NS	NS
Partes:	0.90%	HS	NS
Ouvintes dentro das Partes:	0.00%	HS	NS

ressaltaremos os aspectos mais significativos, lembrando que estes resultados são somente os obtidos para a língua portuguesa e portanto não são os resultados globais compilados pelo CCITT.

Análise de variâncias para a Fase II

Da Tabela 6.4 temos a análise de variâncias feita para a Fase II de testes, de acordo com o especificado no Plano de Testes Subjetivos [90]. A probabilidade mostrada na Tabela é a probabilidade de que as variâncias observadas para cada um dos fatores experimento não sejam significativas. Variâncias serão "Significativas" quando sua probabilidade for menor que 5%. "Altamente Significativo" é usado para probabilidades menores que 1%. "Não Significativo" é usado para probabilidades acima de 5%.

Através da tabela, podemos identificar os efeitos relativos aos Níveis de Audição, ao Sexo dos Locutores, aos Ouvintes, à Seqüência de Apresentação, às Condições, às Listas, às Partes e aos Ouvintes dentro das Partes, bem como aos efeitos das *interações* entre todos esses fatores e o Nível de Audição.

Observa-se que variâncias *Altamente Significativas* foram encontradas para Níveis de Audição, Ouvintes, Interação entre Ouvintes e Níveis, Seqüência de Apresentação, Condições, Listas, Partes e Ouvintes dentro de partes. Variâncias Significativas ocorreram para Interação entre Condições e Níveis. Mostraram-se *Não Significativas* para Interações entre Sexo e Nível, Interação entre Seqüência de Apresentação e Níveis e Interação entre Listas e Níveis.

Destes resultados, contrariaram o esperado a variância altamente significativa para a Seqüência de Apresentação e para sua Interação com o Nível de Audição, para as Partes e para os Ouvintes dentro das Partes. Isto ocorreu provavelmente devido a diferenças sistemáticas entre a primeira e a segunda parte do experimento. Esse fato ocorreu também para os dados globais do CCITT

Tabela 6.5: Valores de MOS, Intervalo de Confiança (CI) e Mínima Diferença Significativa (MSD) da Fase II de Testes Subjetivos do LD-CELP, para todos os locutores. Resultados somente para as condições de interesse avaliadas no PLL. Estatísticas calculadas de acordo com o Capítulo 4.

Condição	Descrição	Estatísticas	
		MOS ± CI	MSD
1	LD-CELP	3.96 ± 0.22	0.31
2	BER=1%	2.85 ± 0.22	0.31
3	BER=0.1%	3.94 ± 0.24	0.33
6	2 × LD-CELP	3.94 ± 0.22	0.31
7	3 × LD-CELP	3.77 ± 0.27	0.38
8	4 × LD-CELP	3.60 ± 0.25	0.35
9	LD-CELP+G.721+LD-CELP	3.94 ± 0.22	0.31
10	G.721+LD-CELP+G.721	3.71 ± 0.23	0.32
11	2 × G.711	4.21 ± 0.23	0.33
12	4 × G.711	4.02 ± 0.22	0.31
13	8 × G.711	3.88 ± 0.17	0.24
14	16 × G.711	3.42 ± 0.22	0.31
15	1 × G.721	4.08 ± 0.24	0.34
16	2 × G.721	3.96 ± 0.23	0.33
17	4 × G.721	3.63 ± 0.22	0.32
18	MNRU(Q=30dB)	4.13 ± 0.22	0.32
19	MNRU(Q=25dB)	4.13 ± 0.22	0.32
20	MNRU(Q=20dB)	3.71 ± 0.21	0.30
21	MNRU(Q=15dB)	3.23 ± 0.22	0.31
22	MNRU(Q=10dB)	2.35 ± 0.21	0.29
23	MNRU(Q=5dB)	1.69 ± 0.21	0.30
24	Direta	4.17 ± 0.19	0.27

[131]. Por outro lado, foi discrepante também a ocorrência de variâncias significativas para as Listas, apontando para a ocorrência de diferenças sistemáticas no material de voz utilizado, não relacionadas com o sexo dos locutores, pois este não foi um fator significativo.

Resultados da Fase II

A seguir é apresentada uma análise das avaliações dadas para as condições do teste em grupos de interesse. A análise completa para os desempenho com a língua portuguesa está nos documentos CCITT [132, 133]. A análise global dos dados da Fase II está em documentos CCITT [131, 134, 135] e no artigo por Usai e South [36].

Desempenho para múltiplas transcodificações. Como na Fase I de testes detectaram-se problemas para 3 transcodificações do LD-CELP, neste teste foi incluída a condição com três transcodificações (condição 7). Comparando-se os MOS das condições 1, 6, 7 e 8 da Tabela 6.5, vê-se que a qualidade cai monotonicamente. Um aspecto interessante é que a qualidade para 1 e 2 LD-CELP é praticamente indistinta.

Para a G.721, a qualidade decai monotonicamente para 1, 2 e 4 cascatas (condições 15, 16 e 17 da tabela 6.5). Os valores MOS deste teste e os do Experimento 2 da Fase I estão na mesma faixa e com intervalos de confiança semelhantes, como seria de se esperar.

Testou-se ainda a qualidade da cascata envolvendo a G.721 e o LD-CELP nas condições 9 e 10 da Tabela 6.5. Pode-se verificar que o desempenho da condição 9 (LD-CELP, G.721 e LD-CELP) está muito próximo ao da condição para 2 LD-CELP, enquanto que a condição 10 (G.721, LD-CELP e G.721) apresenta um desempenho entre 2 e 4 G.721 e é muito próximo ao de 3 LD-CELP. Observa-se então que enquanto há 2 LD-CELP na cascata, a G.721 praticamente não introduz distorção, enquanto que quando há 2 G.721 e 1 LD-CELP, a distorção é equivalente à de 3 LD-CELP. Isto indica uma não linearidade no acúmulo de distorção em qdu[4] para o LD-CELP.

No Experimento 2 da Fase I [48] verificaram-se algumas inconsistências nas cascatas da G.711, mais provavelmente se relacionando a problema ocorrido durante os processamentos do Laboratório Central. Neste experimento não se incluiu a condição para 1 G.711, para poder incluir todas as condições de teste necessárias, enquanto mantendo pequeno o teste. Mantiveram-se entretanto as outras 4 condições para a G.711 (condições 11, 12, 13 e 14 da Tabela 6.5). Em função da ausência da condição para 1 G.711, não é possível verificar a hipótese de que os problemas tenham se manifestado novamente neste teste, mas como os resultados aqui foram praticamente idênticos aos do Experimento 2 da Fase I, é de se esperar que haja problemas nas análise envolvendo cascatas da G.711. Por isso, elas devem ser consideradas com ressalvas e outras análise comparativas devem ser buscadas.

Desempenho com taxas de erro. As condições 1, 3 e 2 da Tabela 6.5 mostram o desempenho do LD-CELP com 0%, 0.1% e 1% de erros (aleatórios). O desempenho com 0.1% de erros é praticamente indistinto da condição livre de erros, enquanto que com 1% de erros a qualidade cai bastante.

Comparação entre LD-CELP e G.721. Neste experimento ficou difícil distinguir entre o LD-CELP e a G.721, pois os MOS e os CI se sobrepõem para 1, 2 e 4 transcodificações em cascata. Em associação à análise global do CCITT, concluiu-se que a qualidade do LD-CELP é indistinta da qualidade da G.721. Isso permite induzir que a qualidade relativa às 3 cascatas do LD-CELP da condição 7 seja equivalente a 3 cascatas da G.721, infelizmente não presente no teste. Porém, a G.113 [122] estabelece 3.5 qdu para 1 G.721, ou 13.5 qdu para 3 G.721. Assim, o LD-CELP também teria 13.5 qdu para 3 transcodificações em cascata e satisfaria o requisito 6 da Tabela 6.1. Adicionalmente, 1 LD-CELP também apresentaria 3.5 qdu e satisfaria o requisito 3 da mesma tabela.

Comparação entre G.721 e G.711. Comparando-se a condição 15 às condições 11 e 12, podemos inferir que a qualidade da G.721 situa-se entre 2 e 4 qdu. Testes anteriores situam

[4] A Recomendação CCITT G.113 [122] define 1 *qdu* como sendo a distorção equivalente à de 1 conversão A/D–D/A segundo a lei de codificação na Recomendação G.711 [16]. Esta unidade é utilizada para o planejamento de sistemas de telecomunicações: por isso é desejável que a distorção total de uma conexão possa ser medida como a soma dos qdu introduzidos pelos dispositivos intermediários. Neste caso específico, vê-se que uma conexão com G.721 e LD-CELP não pode ter sua distorção avaliada com segurança pela simples adição de seus qdu individuais.

o valor em qdu para a G.721 em 3.5 qdu, o que coincide em termos qualitativos com os resultados obtidos

Comparação entre LD-CELP e G.711. Comparando a condição 1 com a condição 12, observamos que a qualidade do LD-CELP está em torno dos 4 qdu, reforçando a conclusão acima de que o LD-CELP satisfaria o requisito 3 da Tabela 6.1. Já para 3 transcodificações do LD-CELP (condição 7 da Tabela 6.5) a qualidade ficou entre 8 e 16 G.711 (condições 13 e 14). Isso também reforça a conclusão de que o LD-CELP testado satisfaz também o requisito 6.

Resultados para o MNRU. Neste experimento, a condição direta apresentou um MOS superior ao do maior Q avaliado (Q=30dB), como esperado[5].

A G.721 fica com uma relação sinal-ruído modulado equivalente $(Qeqv)$ entre 20 e 25 dB, ao compararmos as condições 15, 19 e 20. O LD-CELP para 1 passo de codificação também ficou com Q_{eqv} entre 20 e 25 dB, bem como 3 LD-CELP, LD-CELP+G.721+LD-CELP e G.721+LD-CELP+G.721.

Conclusões sobre a qualidade subjetiva

Dessas duas Fases de testes, as principais conclusões para a língua portuguesa, que coincidem com as observadas pelo CCITT, são:

- A qualidade do LD-CELP da Fase II satisfez os requisitos para 1 transcodificação em cascata (requisito 3 da Tabela 6.1);
- A qualidade do LD-CELP para 3 transcodificações atendeu ao requisito 6 da Tabela 6.1 em sua versão da Fase II;
- As cascatas mistas do LD-CELP com a G.721 estão abaixo dos 14 qdu e satisfazem o requisito 6. Elas também indicaram uma não linearidade no acúmulo de distorção em qdu para o LD-CELP;

Assim, o CCITT estabeleceu que o desempenho do LD-CELP e da G.721 são indistingüíveis em condições livres de erros de transmissão. Em função destes resultados, as regras de planejamento da G.721 constantes da Recomendação CCITT G.113 [122] devem provisoriamente ser aplicadas para o LD-CELP quando se estudar a introdução do LD-CELP na rede telefônica.

6.3 Medidas Objetivas

Nesta seção abordaremos a análise de qualidade de algoritmos a partir de medidas objetivas. Das medidas objetivas do Capítulo 5, selecionamos uma medida temporal, duas medidas espectrais e duas medidas com motivação psicoacústica.

[5] Amiúde faz-se a suposição de que a condição direta é a assíntota para $Q \to \infty$.

Inicialmente descreveremos o material de voz utilizado e o pré-processamento realizado. A seguir apresentaremos os resultados para as medidas selecionadas, realizando uma análise comparativa.

6.3.1 Material de voz

O material de voz utilizado pelas medidas objetivas deve ser o mesmo utilizado pelos testes subjetivos. De fato, utiliza-se um sub-conjunto deste, pois o material processado para os testes subjetivos foi avaliado em três diferentes níveis de audição. Como as medidas objetivas devem preferencialmente utilizar somente sinais avaliados utilizando-se condições nominais [90], parte do material processado não foi utilizado aqui. O volume de material usado está mostrado na Tabela 6.6.

Tabela 6.6: Material de voz utilizado para as medidas objetivas. A coluna "src" abaixo se refere ao material fonte e as condições são as definidas para a Fase II de Testes do LD-CELP. O tempo total de material processado é de 1h34m, equivalendo a 86 MBytes (MB).

Condição	01	02	03	06	07	08	09	10
Número de arquivos	40	38	40	44	38	38	42	40
Tamanho (MB)	4.3	4.1	4.3	4.8	4.1	4.1	4.5	4.3
Duração	4m36s	4m23s	4m36s	5m04s	4m23s	4m23s	4m50s	4m36s
Condição	11	12	13	14	15	16	17	18
Número de arquivos	42	44	36	42	36	40	38	46
Tamanho (MB)	4.5	4.8	3.9	4.5	3.9	4.3	4.1	4.9
Duração	4m50s	5m04s	4m09s	4m50s	4m09s	4m36s	4m23s	5m18s
Condição	19	20	21	22	23	24	src	
Número de arquivos	44	38	38	44	42	40	144	
Tamanho (MB)	4.8	4.1	4.1	4.8	4.5	4.3	15.5	
Duração	5m04s	4m23s	4m23s	5m04s	4m50s	4m36s	16m35s	

Como o volume de material processado era muito grande e como o material processado teve a sua faixa limitada em 4kHz pelo filtro IRS (pré-processamento) e pelos filtros passa-baixa do circuito de processamento da Figura 6.2, o material de voz foi dizimado por um fator de 2:1 antes de ser submetido às medidas objetivas utilizando-se as ferramentas de redução de taxa da STL92 [11], descritas no Capítulo 3. A dizimação permitiu um cômputo mais rápido das medidas sem que houvesse perda de informação, pois o conteúdo espectral acima dos 4kHz fora eliminado. Para garantir compatibilidade, os sinais-fonte foram também dizimados pelo mesmo processo.

Sinais de Referência e de Teste. Todas as medidas objetivas se baseiam na comparação de dois sinais, exprimindo um valor que quantifique o grau de diferença entre eles. Um deles, distorcido, chamamos de *sinal de teste*; o outro, o *sinal de referência*, é a referência de qualidade para a avaliação da quantidade de distorção presente no sinal de teste.

É desejável que os sinais de referência e de teste passem sempre pelo mesmo processamento básico, de modo que a única diferença de manipulação entre eles é que o de testes tenha passado por uma condição de processamento (e.g. um codec) e o de referência tenha "curto-circuitado" esse processamento. Em referência à figura 6.2, o sinal de referência seria a

condição direta ("placebo") e o de teste, as outras condições (MNRU, G.721, etc.).

Como explicado na seção 6.2.2, devido ao projeto do teste empregar dois conjuntos *distintos* de materiais de voz, o **GL1** foi processado pelas condições ímpares e o **GL2** pelas pares. Como a condição direta é a de número 24, somente as condições pares possuem arquivos que podem ser comparados aos da condição direta, pois os sinais originais para as condições pares são sempre da **GL2**. Assim, sinais processados pelas condições ímpares não poderiam ser comparados aos da condição direta porque os seus sinais-fonte são distintos. Para permitir a avaliação dos sinais processados por ambos os grupos de condições, não foi possível utilizar os sinais da condição direta (que seriam os mais indicados como sinais de referência). Ao invés deles, os sinais originais foram utilizados, pois eles estão disponíveis tanto para o **GL1** como para o **GL2**. Existe aqui a desvantagem deles não haverem sido processados pelo circuito básico do Laboratório Central. Por isso, eles apresentam um nível de distorção menor que o material processado pela condição direta (1 conversão A/D, 1 D/A e 2 filtragens passa-baixa – ver Figura 6.2).

6.3.2 Pré-processamento do material a ser avaliado

Compensação de Atraso e Amplitude. Antes de se avaliar a distorção de um material de voz em relação a uma referência, deve-se:

- compensar o atraso de fase entre os sinais;
- equalizar o nível dos sinais, por exemplo fazendo a potência do sinal de teste se igualar à do sinal de referência.

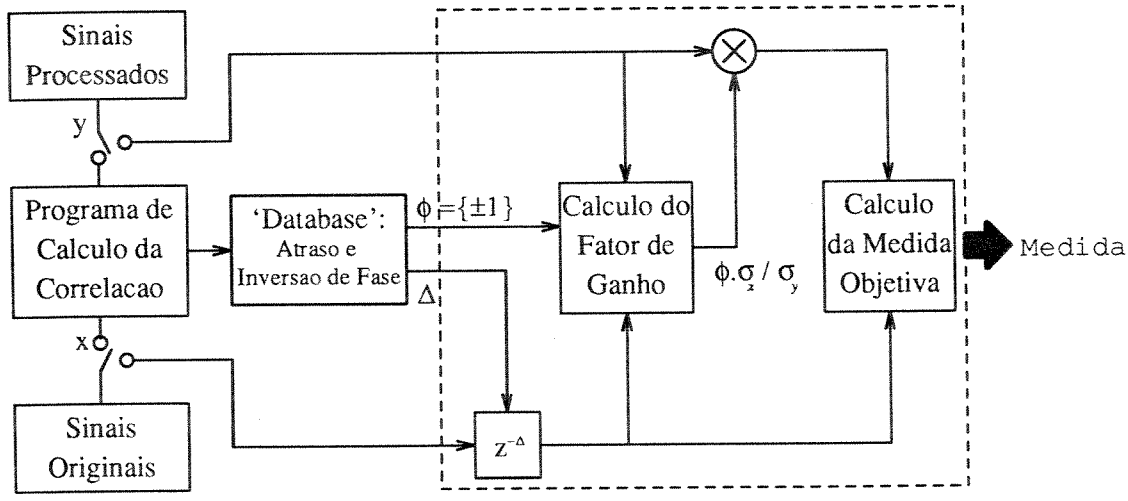


Figura 6.3: Esquema alternativo de compensação de atraso, inversão de fase e ganho. Aqui os atrasos e a informação de inversão fase são calculados "off-line" quando as chaves conectam os dados ao programa de cálculo de correlação, criando o banco de dados. Com a chave na outra posição, as medidas de distorção são calculadas, sendo o sinal atrasado e o ganho calculado localmente.

Uma indicação de como este processo pode ser implementado foi feita no Capítulo 5. Embora esse esquema seja válido do ponto de vista formal, ele foi implementado de uma forma

alternativa, esquematizada na Figura 6.3.

Como um primeiro passo calculou-se o atraso existente entre todos os sinais processados e os sinais originais, através da determinação da localização do pico na correlação cruzada entre cada um deles. Além do atraso, verificou-se se a houve inversão de fase entre eles (sinal do valor de pico da correlação cruzada). Esses resultados foram armazenados num banco de dados e eram fornecidos aos programas que implementavam cada uma das medidas de distorção. Como as medidas de distorção seriam calculadas repetidas vezes para os mesmos arquivos e o cálculo da correlação cruzada é um processo extremamente demorado, essa abordagem permitiu uma economia substancial de tempo no processamento das medidas[6].

Já a equalização de amplitude era realizada "on-line" pelos programas de cálculo de distorção através da estimativa da razão da energia local de cada segmento dos sinais de referência e teste. Essa razão definia o fator a ser aplicado ao sinal de teste. Mecanismos específicos de equalização de amplitude encontram-se explicados no Capítulo 5.

Janelamento do Sinal. Uma outra consideração a ser feita relaciona-se às técnicas espectrais, onde o janelamento dos sinais de referência e de teste é necessário. Seguindo a indicação em [64], utilizamos o janelamento de Hamming para segmentos de 32ms de duração (N=256 amostras a uma taxa de amostragem de 8 kHz ou 512 amostras a 16 kHz), sem sobreposição *(overlap)* entre segmentos, segundo a equação:

$$w(k) = 0.54 - 0.46\cos\left(\frac{2\pi k}{N-1}\right), \qquad \text{para } 0 \leq k \leq N-1 \tag{6.1}$$

6.3.3 Medidas Objetivas selecionadas

Neste trabalho não consideramos qualquer das medidas temporais descritas no Capítulo 4 porque o material de voz processado, que aqui é o sinal de teste para as medidas objetivas, passou por várias filtragens adicionais em relação ao sinal de referência, que, como explicado, aqui teve que ser o sinal original (ao invés daquele processado pela condição direta). Essas filtragens provocaram distorções de fase que não afetam medidas espectrais e perceptuais, mas invalidam o uso de medidas temporais. Por isso, selecionamos apenas essas medidas imunes.

Uma das medidas espectrais é a distância cepstral em sua versão truncada com compensação de ganho implícito, para $L = 2P$:

$$CD = \frac{10}{\ln(10)}\sqrt{2\sum_{k=1}^{L}[c_x(k) - c_y(k)]^2} \qquad \text{[dB]} \tag{6.2}$$

A outra é a distância espectral em sua versão discreta:

$$SD = 10\sqrt{\frac{1}{L}\sum_{l=0}^{L-1}\left\{\log_{10}|X(l)| - \log_{10}|Y(l)| - \log_{10}\left[\frac{\sigma_x}{\sigma_y}\right]\right\}^2} \qquad \text{[dB]} \tag{6.3}$$

[6] A mais demorada das medidas de distorção tomou cerca de 5 horas numa estação de trabalho Sun Sparc 4/330 para processar todos os sinais (ver a Tabela 6.7). Já o cálculo dos atrasos pela função de correlação tomou, na mesma máquina, 2 dias e 6 horas, correspondendo a um fator de 11 vezes.

Por fim, duas medidas com motivação perceptual foram testadas. Definidas no Capítulo 5, a distância espectral perceptual de um segmento é dada por:

$$PSD = \sqrt{\sum_{b=1}^{B}[L_x(b) - L_y(b)]^2} \tag{6.4}$$

e a distância cepstral perceptual de um segmento:

$$PCD = 10\sqrt{\sum_{b=1}^{B}\{\log_{10}[L_x(b)] - \log_{10}[L_y(b)]\}^2} \quad \text{[dB]} \tag{6.5}$$

A figura de mérito de um sinal será então o valor médio de uma dada distorção para todos os seus segmentos.

Tabela 6.7: Exemplo de tempos de processamento para cada uma das medidas objetivas para processar todo o material de interesse selecionado da Fase II de testes do LD-CELP. Os valores apresentados são o tempo de cpu e o tempo total gasto por uma Sparc Sun 4/330, de 16 MIPS. Os valores entre parênteses são o tempo gasto, normalizado em relação à distância cepstral.

Parâmetro	SD	CD	PSD e PCD
cpu	4h15m12s (4.15)	1h01m30s (1.00)	2h52m38s (2.81)
total	4h56m10s (4.23)	1h09m58s (1.00)	3h02m30s (2.61)

6.3.4 Resultados das Medidas Objetivas

As medidas objetivas e as análises estatísticas descritas anteriormente foram implementadas em linguagem C [136] e desenvolvidas em ambiente Unix de estações de trabalho Sun. A manipulação de arquivos de dados foi feita usando ferramentas nativas do Unix, como o `awk` [137, 138] e o sed [139]. A complexidade computacional de cada uma das implementações das medidas pode ser estimada pelo tempo gasto para o processamento de todo o material de voz, mostrado na Tabela 6.7. Pode-se observar que a mais demorada delas é a SD (por envolver FFTs) e a mais rápida é a CD. Os tempos relativos são mais relevantes que os valores absolutos.

Os valores obtidos para todas as medidas encontram-se na Tabela 6.8. A seguir, são descritos os resultados para cada uma das medidas de distorção, agrupadas em classes de processamento.

LD-CELP em presença de erros

As condições que se referem à análise do LD-CELP em presença de 0%, 1% e 0.1% de erros de transmissão são respectivamente as condições 1, 2 e 3 da Tabela 6.8.

Tabela 6.8: MOS e medidas objetivas para as condições selecionadas do material processado para a Fase 2 de testes do LD-CELP. Os valores apresentados são o valor médio seguido de seu desvio padrão. SD, CD e PCD estão em dB. PSD é adimensional. 1 *qdu* corresponde a 1 × G.711 (Rec.CCITT G.113)

Cond	Descrição	MOS	SD	CD	PSD	PCD
1	LD-CELP	3.94±0.23	0.28±0.01	1.59±0.07	0.07±0.01	2.96±0.17
2	BER=1%	2.91±0.24	0.35±0.01	3.36±0.07	0.20±0.01	5.62±0.11
3	BER=0.1%	3.95±0.25	0.29±0.01	1.87±0.07	0.09±0.01	3.34±0.17
6	2 × LD-CELP	3.91±0.22	0.33±0.01	2.42±0.11	0.11±0.01	4.16±0.20
7	3 × LD-CELP	3.76±0.29	0.36±0.01	2.91±0.11	0.13±0.01	5.00±0.17
8	4 × LD-CELP	3.59±0.27	0.39±0.01	3.50±0.12	0.17±0.01	6.14±0.21
9	LD-CELP+G.721+LD-CELP	3.94±0.21	0.34±0.01	2.44±0.10	0.11±0.01	4.29±0.23
10	G.721+LD-CELP+G.721	3.70±0.24	0.32±0.01	2.05±0.10	0.09±0.01	3.59±0.25
11	2 × G.711 (≡ *2 qdu*)	4.14±0.26	0.23±0.01	1.26±0.12	0.05±0.01	3.04±0.26
12	4 × G.711 (≡ *4 qdu*)	4.01±0.23	0.25±0.01	1.69±0.13	0.07±0.01	3.93±0.27
13	8 × G.711 (≡ *8 qdu*)	3.88±0.18	0.45±0.01	4.22±0.26	0.24±0.01	7.37±0.43
14	16 × G.711 (≡ *16 qdu*)	3.48±0.22	0.53±0.01	6.19±0.15	0.36±0.01	10.44±0.24
15	1 × G.721	4.10±0.25	0.25±0.01	1.31±0.10	0.04±0.01	2.61±0.26
16	2 × G.721	3.95±0.24	0.29±0.01	1.93±0.12	0.07±0.01	3.59±0.32
17	4 × G.721	3.63±0.25	0.34±0.01	2.59±0.19	0.10±0.01	4.47±0.38
18	MNRU(Q=30dB)	4.14±0.21	0.23±0.01	0.80±0.07	0.03±0.01	1.39±0.15
19	MNRU(Q=25dB)	4.09±0.23	0.24±0.01	0.91±0.05	0.03±0.01	1.57±0.12
20	MNRU(Q=20dB)	3.70±0.20	0.26±0.01	1.12±0.06	0.04±0.01	1.95±0.16
21	MNRU(Q=15dB)	3.18±0.23	0.30±0.01	1.64±0.07	0.06±0.01	2.86±0.13
22	MNRU(Q=10dB)	2.31±0.21	0.36±0.01	2.41±0.10	0.13±0.01	4.67±0.26
23	MNRU(Q=5dB)	1.71±0.23	0.46±0.01	3.54±0.11	0.30±0.01	7.94±0.15
24	Direta	4.12±0.20	0.14±0.01	0.28±0.03	0.01±0.01	0.70±0.10

SD. Consistentemente com os resultados subjetivos, a SD para 0.1% de erros é um pouco maior que aquela para 0% e a SD para 1% é bem maior que para essas duas outras condições. Entretanto, o aumento na SD entre 0% e 0.1% pode ser condiderado significativo, enquanto que a queda no MOS não é significativa.

CD. Também consistente com os testes subjetivos, a CD para 0% de erros é um pouco menor que aquela para 0.1%, enquanto que há um aumento pronunciado de distorção para 1% de erros. Esse aumento de distorção foi relativamente maior para a CD que para a SD.

PSD. Do mesmo modo que as duas anteriores, a PSD também foi consistente com com as medidas subjetivas, sendo porém a queda de qualidade entre as condições 2 e 3 ainda mais pronunciada que no caso da SD e da CD.

PCD. Outra vez em acordo com os resultados subjetivos, a PCD apresentou uma queda intermediária entre a PSD e as SD e CD.

LD-CELP em tandem

As condições 1, 6, 7 e 8 da Tabela 6.8 apresentam as avaliações para o LD-CELP para 1, 2, 3 e 4 transcodificações assíncronas.

SD. A SD comportou-se coerentemente em relação à qualidade subjetiva, sendo monotonicamente crescente em incrementos de aproximadamente 0.05dB entre 1 e 2 LD-CELP e entre 2 e 4 LD-CELP[7]. A SD para 1 LD-CELP situou-se entre 4 e 8 qdu e a SD para 3 LD-CELP em cascata aponta para uma distorção entre 4 e 8 qdu. Comparando com o MNRU, 1 LD-CELP situou-se entre a SD para Q de 15dB e 20dB e a SD para 3 LD-CELP caiu em torno daquela para Q=15dB.

CD. A CD também foi monotonicamente crescente, com uma inclinação diferente. O aumento em distorção de 1 para 2 LD-CELP foi de 0.8dB, enquanto que entre 2 e 4 LD-CELP foi de 1.1 dB, o que é aproximadamente constante se considerarmos os intervalos de confiança associados. O valor apontado para 1 LD-CELP cai entre aquele para 2 e 4 qdu, enquanto que a CD para 3 LD-CELP situa-se entre as condições equivalentes a 4 e 8 qdu. Em relação ao MNRU, 1 LD-CELP estaria com um Q entre 20dB e 15dB e 3 LD-CELP entre 10dB e 5dB.

PSD. A PSD mostrou um aumento monotônico e significativo na distorção para as condições 1, 6, 7 e 8, sendo que, à semelhança do ocorrido com a SD, houve um incremento de 0.05 dB entre 1 e 2 LD-CELP e entre 2 e 4 LD-CELP. Segundo a PSD, a distorção equivalente a 1 LD-CELP em cascata situa-se em torno de 4 qdu; a PSD para 3 LD-CELP em cascata, a exemplo da SD e da CD, também está entre 4 e 8 qdu. A PSD associada a 1 LD-CELP situou-se em torno da PSD para o MNRU com Q=15dB, enquanto que a PSD para 3 LD-CELP caiu em torno da condição com Q=10dB, semelhantemente ao ocorrido com a SD.

PCD. Houve um aumento monotônico da distorção segundo a PCD, não se mantendo o incremento constante entre condições representando potências de 2 de número de transcodificações em cascata. A PCD encontrada para 1 LD-CELP situou-se em torno de 2 qdu. Como nos casos anteriores, a distorção para 3 LD-CELP cai entre 4 e 8 qdu. Já comparando com o MNRU, a PCD para 1 LD-CELP situou-se em torno daquela para Q=15dB e a PCD para 3 LD-CELP está entre a PCD para Q entre 5dB e 10dB.

Comparação entre LD-CELP e G.721

O desempenho relativo entre o LD-CELP e a G.721 pode ser encontrado comparando-se as condições 1 e 15 ou 7 e 17.

SD e PSD. Tanto a SD como a PSD para 1 LD-CELP confundem-se com aquela para 2 G.721. A SD e a PSD para 3 LD-CELP situam-se abaixo da SD para 4 G.721.

[7] É desejável que o aumento em dB seja constante quando se compara condições com o dobro de transcodificações, o que permitiria estabelecer um lei log-linear.

CD e PCD. A CD e a PCD comportam-se de maneira idêntica nesta comparação. Ambas apresentam uma distorção para o LD-CELP entre 1 e 2 G.721, enquanto que para 3 LD-CELP a distorção está acima daquela para 4 G.721.

G.711 em tandem

As condições 11 a 14 da Tabela 6.8 apresentam os valores MOS e de distorção para 2, 4, 8 e 16 transcodificações em cascata da G.711.

SD. A SD foi monotonicamente crescente para as condições envolvendo cascatas da G.711, mas apesar de envolver cascatas em potências de 2, o aumento equivalente em dB mostrou-se bastante não-linear, notadamente entre 4 e 8 cascatas. A distorção referente a 16 transcodificações em cascata da G.711 situou-se entre os Q de 5 e 10 dB do MNRU (condições 22 e 23), sendo maior que a verificada nos testes subjetivos.

CD. A CD também apresentou um incremento monotônico com o número de cascatas, sendo o incremento não linear, especialmente entre 4 e 8 transcodificações. Também aqui a distorção atribuída para 16 G.711 foi maior que a verificada subjetivamente, apresentando aqui uma distorção maior que a para Q=5dB.

PSD. Acontece também aqui um aumento monotônico da PSD, sendo porém em incrementos não constantes, sendo também maior entre 4 e 8 transcodificações em cascata. O valor associado a 16 G.711 apresentou como nos casos anteriores uma distorção maior que à associada à condição 23 (MNRU com Q=5dB), o que não casa com os resultados subjetivos.

PCD. A PCD cresceu monotônicamente com o número de transcodificações em cascata. O incremento foi aproximadamente constante entre 4 e 8 cascatas, da ordem de 3.4 dB, e entre 8 e 16 cascatas, com 3.1 dB, mas significativamente inferior entre 2 e 4 cascatas (0.9dB). Comparando com valores MNRU, outra vez a distorção para 16 G.711 foi bem maior que a apontada para Q=5dB.

G.721 em tandem

O desempenho da G.721 com 1, 2 e 4 transcodificações assíncronas está apresentado nas condições 15, 16 e 17 da Tabela 6.8.

SD. A SD foi monotônica com o número de cascatas da G.721, porém o incremento não foi muito constante, sendo um pouco maior entre 2 e 4 transcodificações. A SD para 1 G.721 situou-se entre a SD para Q de 20dB e 25dB. Para 3 G.721, a qualidade situou-se entre os Q de 10dB e 15dB.

CD. A CD também foi monotônica, mas o incremento foi constante de 0.6dB entre 1 e 2 e entre 2 e 4 transcodificações da G.721. A distorção cepstral associada a 1 G.721 caiu entre as associadas às condições para Q de 15dB e 20dB. Já a CD para 3 G.721 iguala-se à equivalente a um Q de 10dB.

PSD. A PSD cresceu monotonicamente em incrementos constantes de 0.03. A PSD para 1 G.721 situou-se em torno da PSD para MNRU com Q=20dB, enquanto que a PSD para 3 G.721 caiu entre a PSD para Q de 10dB e 15 dB.

PCD. A PCD apresentou um aumento constante de 0.93dB da distorção para as condições envolvendo G.721 em cascata. A PCD associada a 1 G.721 está em torno daquela para Q=15dB, enquanto que para 3 G.721 a PCD situou-se em torno da PCD para Q=10dB.

Tandem misto com G.721 e LD-CELP

A condição 9 da Tabela 6.8 mostra o desempenho do tandem misto entre 2 LD-CELP e 1 G.721. Já a condição 10 apresenta a combinação complementar, 2 G.721 e 1 LD-CELP.

SD. Coerentemente com os resultados subjetivos, a cascata da condição 10 apresentou uma distorção menor que a da condição 9. Em especial, a distorção atribuída a 2 LD-CELP foi equivalente àquela para a condição 9 (LD-CELP, G.721 e LD-CELP).

CD. A CD também foi coerente com os resultados subjetivos. À semelhança do ocorrido para a SD, a distorção encontrada para a condição 9 também foi equivalente à da condição para 2 LD-CELP. Além disso, a condição 10 (G.721, LD-CELP e G.721) apresentou distorção equivalente à de 2 G.721 em cascata.

PSD. A PSD para a condição 10 foi menor que a para a condição 9, como para a SD e a CD. Como para esta última, a PSD para a cascata entre LD-CELP, G.721 e LD-CELP situou-se em torno daquela para 2 LD-CELP em cascata. Já a PSD para G.721 em cascata com LD-CELP e G.721 situou-se em torno da PSD para 3 G.721.

PCD. Como no caso da CD, a distorção para o LD-CELP em cascata como G.721 e LD-CELP situou-se em torno daquela para 2 LD-CELP e a CD para a cascata entre G.721, LD-CELP e G.721 também situou-se em torno da CD para 2 G.721. Outra vez, a condição 9 apresentou uma distorção maior que a condição 10.

MNRU e Condição Direta

As condições 18 a 23 da Tabela 6.8 dão as avaliações para sinais processados pelo MNRU respectivamente para Q de 30, 25, 20, 15, 10 e 5dB. A condição 24, por sua vez, é a condição direta.

SD. A SD, coerentemente, comportou-se monotonicamente crescente com o aumento de ruído modulado do MNRU (Q decrescente) e a condição de Q=30dB apresentou-se com mais distorção que a condição direta. Entretanto, o aumento na distorção não foi constante, sendo o incremento crescente em direção a Q menores. Também o incremento entre a distorção da condição direta e para Q=30dB não foi proporcional à queda equivalente em qualidade subjetiva.

CD. A CD também comportou-se monotonicamente crescente com a queda em Q e, à semelhança da SD, os incrementos não foram constantes. Outra vez, o aumento de distorção entre a condição direta e a condição com Q=30dB foi mais abrupta do que parece indicar a equivalente queda em MOS.

PSD. A PSD cresce monotonicamente com a queda em Q, sendo menor para a condição direta do que para Q=30dB. Outra vez, o incremento da PSD entre condições não foi constante.

PCD. A PCD cresceu monotonicamente em incrementos não constantes. Os incrementos foram monotonicamente crescentes para as condições 18 a 23, mas o aumento da PCD entre a condição direta e aquela para Q=30dB foi significativamente maior que o aumento de distorção entre Q de 30dB e 25dB e de 25dB e 20dB. Isso não é coerente com a qualidade subjetiva.

6.4 Transformação de medidas objetivas em valores MOS

Quando se estabelece uma medida objetiva, para que ela possa predizer com algum grau de precisão a qualidade subjetiva, é necessário que se identifique uma função aproximadora que mapeie essa medida objetiva em valores MOS a partir de um sub-conjunto das avaliações subjetivas. Esses valores MOS serão chamados de *MOS estimados*. A seguir, são descritos o método empregado neste trabalho para se encontrar esse mapeamento através de uma função mapeadora, a escolha dessa função, bem como o método para se avaliar a qualidade do ajuste entre valores reais e estimados.

6.4.1 Identificação da função aproximadora

A determinação da função de aproximação envolve três etapas [108, p.518]:

1. Determinação dos parâmetros da função de aproximação;
2. Cálculo dos parâmetros e estimativa do erro entre o modelo e os dados experimentais;
3. Uma avaliação estatística da qualidade do ajuste dos parâmetros;

Cálculo dos parâmetros do modelo.

O ajuste das medidas objetivas com as avaliações subjetivas pode ser feito através do método dos mínimos quadrados [125, 140, 108]. O método dos mínimos quadrados consiste na busca de um conjunto de coeficientes $a_k, k = 1..M$ que aproxima um conjunto de N dados $(x_i, y_i), i = 1..N$ a uma *função aproximadora* $y(x)$ que seja a combinação linear de uma família de M *funções de base* $X_k(x)$. Isto é, achar

$$y(x) = \sum_{k=1}^{M} a_k X_k(x) \tag{6.6}$$

que minimiza

$$\chi^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - \sum_{k=1}^{M} a_k X_k(x_i) \right]^2 \tag{6.7}$$

onde σ_i é o desvio padrão relativo à i-ésima amostra. Em nosso caso, corresponderá ao intervalo de confiança dos valores MOS.

Essa formulação pode ser re-escrita na forma matricial:

$$\mathbf{A}.\mathbf{a} = \mathbf{b} \tag{6.8}$$

onde

$$\begin{aligned} A_{ij} &= \frac{X_j(x_i)}{\sigma_i} \\ \mathbf{a} &= [a_1\ a_2 \cdots a_M] \\ \mathbf{b} &= \frac{y_i}{\sigma_i} \end{aligned}$$

Aqui $\mathbf{A} = [A_{ij}]$ é uma matriz $M \times N$ cujos elementos são o valor da j-ésima função de base no ponto x_i normalizado pelo desvio padrão associado, a é um vetor-coluna de dimensão M com os coeficientes que ajustam os dados experimentais à função aproximadora e b é um vetor-coluna de dimensão N cujos elementos são os pontos y_i normalizados pelo desvio padrão associado.

O problema é então encontrar a que minimiza

$$\chi^2 = |\mathbf{A} \cdot \mathbf{a} - \mathbf{b}|^2$$

Há vários métodos de resolução para esse problema, como o método da resolução das equações normais [125, 140, 108], pela ortogonalização de Gram-Schmidt [140] ou pelo método da decomposição em valores singulares [108]. Destes, o menos indicado é o método das equações normais, devido à sua alta sensibilidade a erros de arredondamento e ao fato de em muitos casos $\mathbf{A}$ ser singular [140, 108]. Utilizaremos aqui o método indicado por Press *et al.* em [108, Cap.14], que é o da decomposição em valores singulares *(SVD, Singular Value Decomposition)*.

O SVD parte do princípio de que a matriz $\mathbf{A}$ pode ser decomposta em três outras matrizes:

$$\mathbf{A} = \mathbf{U}.\mathbf{W}.\mathbf{V}^t$$

onde $\mathbf{U}$ é uma matrix $M \times N$ cujas colunas são ortonormais[8], $\mathbf{W}$ é uma matriz diagonal de ordem $N \times N$ cujos elementos são não-negativos e $\mathbf{V}$ é uma matriz ortonormal de ordem $N \times N$. O índice t indica a operação de transposição. Se $\mathbf{A}$ fosse uma matriz quadrada, sua inversa poderia ser escrita como [108]:

$$\mathbf{A}^{-1} = \mathbf{V}.[\mathrm{diag}(1/w_j)].\mathbf{U}^t$$

de tal modo que poderíamos escrever:

$$\mathbf{a} = \mathbf{V}.[\mathrm{diag}(1/w_j)].(\mathbf{U}^t.\mathbf{b}) \tag{6.9}$$

Essa decomposição explicita na matriz $\mathbf{W}$ quais as linhas da matriz $\mathbf{A}$ que a tornam singular ou mal-condicionada através de seus elementos $w_{ii} = w_{jj} \triangleq w_j$ que sejam ou nulos ou muito próximos a zero, pois a inversa de $\mathbf{W}$ é uma matriz diagonal cujos elementos são $1/w_j$. Do ponto de vista da álgebra linear, $\mathbf{A}$ é uma transformação linear cujo espaço nulo é gerado pelas colunas j de $\mathbf{A}$ que correspondem a $w_j \approx 0$. O SVD explora esse fato a partir do raciocínio de que, já que como essas colunas de A correspondem ao espaço nulo, não importa o coeficiente $1/w_j$ associado, pois existe uma família de valores que solucionarão o sistema. Então, na inversa acima, troca o valor $1/w_j \rightarrow \infty$ por $1/w_j = 0$. Essa artimanha faz com que exista sempre uma solução para o sistema, independentemente de $\mathbf{A}$ ser inversível, singular, bem ou mal condicionada. Do ponto de vista do método dos mínimos quadrados, isso equivale a zerar a contribuição de uma dada função base, ao invés de se utilizar um valor aleatoriamente grande[9] $1/w_j$.

No caso dos mínimos quadrados, $\mathbf{A}$ não é quadrada, mas a solução da equação 6.9 ainda vale, bem como a substituição dos valores $(1/w_j \rightarrow \infty)$ por zero. Então, podemos re-escrever 6.9 como sendo:

$$\mathbf{a} = \sum_{i=1}^{M} \left[\frac{1}{w_j}.\mathbf{U}^t_{(\mathrm{i})}.\mathbf{b} \right] \mathbf{V}_{(i)} \tag{6.10}$$

onde $\mathbf{U}_{(i)}$ e $\mathbf{V}_{(i)}$ são vetores-linha que correspondem respectivamente à i-ésima linha de $\mathbf{U}$ e de $\mathbf{V}$.

Pode-se ainda estimar a variância associada a cada parâmetro a_j estimado para o modelo:

$$\sigma^2(a_j) = \sum_{i=1}^{M} \left(V_{ji} \frac{1}{w_j} \right)^2$$

Apesar deste método ser computacionalmente mais intensivo que o método das equações normais, ele é mais vantajoso para o método dos mínimos quadrados pelo menos por dois motivos:

- A solução da equação 6.8 sempre apresentará números "razoáveis" do ponto de vista prático;
- A inspeção da matriz $\mathbf{W}$, ou dos valores w_j, permite perceber se há funções-base inadequadas na função de aproximação.

[8] Uma matriz é dita ortonormal quando $\mathbf{U}.\mathbf{U}^t = \mathbf{1}$, sendo $\mathbf{1}$ uma matriz diagonal cujos elementos são 1.

[9] A princípio, w_j pode ser zero. Porém, no caso mais comum, $w_j \approx 0$, devido aos erros de arredondamento. Como os erros de arredondamento são aleatórios e pequenos, $1/w_j$ pode assumir valores grandes e aleatórios.

Determinação do Erro e Avaliação do Ajuste

A avaliação estatística da qualidade do ajuste *(Goodness of fit)* entre as medidas objetivas e as avaliações subjetivas pode ser feito pela avaliação da probabilidade de que o erro observado para o modelo tenha ocorrido "por acaso". Pequenos valores para essa probabilidade indicariam que, mais provavelmente, os erros seriam determinísticos e não aleatórios, isto é, o modelo tem necessariamente um erro grande.

Definimos na equação 6.7 o erro como sendo χ^2. Neste caso, χ^2 é conhecido como a distribuição qui-quadrada para $\nu = N - M$ graus de liberdade. A probabilidade associada dos erros serem aleatórios é calculada através da função gama incompleta [108, pp.171,177]:

$$Q(\chi^2|\nu) = \frac{\int_{\chi^2/2}^{\infty} e^{-t} t^{\nu/2-1} dt}{\int_0^{\infty} e^{-t} t^{\nu/2-1} dt} \tag{6.11}$$

Um valor de $Q(\chi^2|\nu) \approx 0$ indicaria que o modelo está mal-ajustado ou que os desvios-padrão σ_i reais são maiores que os utilizados. Um erro "tolerante" no modelo normalmente permite aceitar modelos com Q acima de 0.001. Modelos rigorosos podem exigir Q acima dos 10^{-18} [108, p.522]! Por outro lado, se a probabilidade estiver muito próxima de 1, provavelmente o σ_i utilizado é bem maior que o desvio padrão real. Valores "razoáveis" para Q para um ajuste razoável ocorrem para $\chi^2 \approx \nu$ [108, p.522].

Escolha dos parâmetros do modelo.

A função aproximadora deve ser montada a partir do comportamento do sinal. Componentes periódicas podem ser modeladas por senóides e sinais que saturam por exponenciais. No caso do mapeamento de medidas objetivas em valores MOS, é comum utilizar-se uma função polinomial do tipo:

$$y(x) = \sum_{k=1}^{M} a_k x^{k-1} \tag{6.12}$$

A determinação da ordem M do modelo pode ser feita a partir de predições teóricas ou a partir de tentativa-e-erro, isto é, buscando-se a ordem M que resulte no melhor ajuste. No caso das medidas objetivas aqui consideradas, seria complexo definir teoricamente a ordem necessária para cada uma delas, apesar de haver valores utilizados na literatura (e.g., Itoh *et al.* [57] utilizam $M = 3$ para a distância cepstral). Entretanto, frente à diversidade de medidas e um desconhecimento a priori de seu comportamento, é mais versátil utilizar-se a segunda abordagem. Os resultados para diversas medidas e diversas ordens encontram-se na seção 6.3.4.

6.4.2 Qualidade de predição da função aproximadora

Obtidos os coeficientes $a_k . k = 1..M$ da função aproximadora, pode-se calcular o valor estimado para cada avaliação subjetiva. A partir destes valores estimados, pode-se calcular seu

valor médio, isto é, um MOS estimado. Assim, para cada condição, pode-se avaliar a qualidade do ajuste através da correlação r de Pearson [108, pp.503-506]:

$$r = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}} \tag{6.13}$$

onde y_i representa o valor MOS da condição i, $\hat{y}_i$ o valor MOS estimado da condição i e N é o número de condições. As variáveis com barra são os valores médios das variáveis y_i e $\hat{y}_i$ para todos os i considerados.

A probabilidade de que a hipótese (H_0) de que y e $\hat{y}$ sejam incorrelatos é definido pela estatística

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

que é igual à distribuição t de Student para $\nu = N - 2$ graus de liberdade, cujo grau de confiança α (ou probabilidade de H_0 ser verdadeira) é dado por:

$$1 - \alpha = A(t|\nu) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \int_0^{\frac{t^2}{\nu+t^2}} \frac{(1-t)^{\nu/2-1}}{\sqrt{t}} dt \tag{6.14}$$

onde $\Gamma()$ é a função gama [108].

Assim, valores de $\alpha \to 1$, ou $(1 - \alpha) \to 0$, indicam um bom ajuste do modelo aos dados reais, isto é, que a função aproximadora identificada prediz adequadamente valores subjetivos a partir de medidas objetivas.

6.5 Correlação entre avaliações objetivas e subjetivas

Após a escolha do método pelo qual as medidas objetivas serão transformadas em valores MOS estimados e qual a família de funções que implementará esse mapeamento, resta determinar quais os coeficientes da função aproximadora, o conjunto de dados a ser utilizado para se gerar esses coeficientes (que chamaremos de *conjunto de treinamento*) e o conjunto de dados para se avaliar o ajuste do modelo aos valores subjetivos reais (que chamaremos de *conjunto de avaliação de ajuste*).

Na seção anterior descrevemos o Método dos Mínimos Quadrados através da Decomposição dos Valores Singulares (LS-SVD) como um método adequado ao objetivo deste trabalho. Escolhemos também um polinômio como função aproximadora. Resta definir, antes de passarmos aos resultados, os aspectos pendentes.

Após a realização das medidas objetivas, criou-se um banco de dados que relacionava a avaliação subjetiva (nota entre 1 e 5) e as medidas objetivas de distorção (SD, CD, PSD e PCD) para cada sinal processado. A partir dos valores individuais de cada sinal, para todas as condições ou para classes de distorção, pôde-se associar valores médios, que para as avaliações subjetivas chamaremos de MOS.

A seleção dos conjuntos de treinamento e de avaliação de ajuste a partir desse banco de dados poderia ser feito de várias formas:

1. Obtendo os coeficientes polinomiais a_k (treinamento) a partir do banco de dados como um todo e avaliando a qualidade de predição usando os valores médios da Tabela 6.8;
2. Treinando usando os valores médios (Tabela 6.8 e avaliando a partir dos valores individuais constantes do banco de dados;
3. Dividindo o banco de dados em dois subconjuntos, A e B. O conjunto A é usado para a identificação dos coeficientes a_k (treinamento) e o sub-conjunto B é usado para avaliar o modelo definido por esses a_k.

As duas primeiras abordagens acima possuem uma severa restrição: como os valores individuais (votos e distorção para cada arquivo) e os médios são interdependentes, a avaliação do ajuste a partir de um treinamento com dados dependentes provavelmente estará *viciada*. Isso comprometeria a possibilidade de se utilizar os mesmos coeficientes para um conjunto distintos de sinais. Em função disso, parece mais razoável do ponto de vista da identificação de métodos objetivos de estimação da qualidade que o treinamento e a avaliação do ajuste sejam feitos por conjuntos independentes de dados. Portanto, neste trabalho utilizamos a terceira opção, dividindo o banco de dados em dois subconjuntos. Por simplicidade, ele foi dividido em duas partes com *mesmo número* de dados. Em conseqüência, os dados apresentados nesta seção possuem igual significância por terem utilizados o mesmo número de amostras.

Função aproximadora universal ou especializada? Uma questão que surge neste ponto é se é possível identificar uma função aproximadora para cada medida que sirva para todas as condições ou então se é necessário identificar-se conjuntos distintos de funções aproximadoras para cada medida objetiva e para cada classes de distorção[10]. Do ponto de vista formal, seria altamente desejável a identificação de uma medida que pudesse ser aplicada através de uma única função aproximadora para a estimação da qualidade subjetiva sem distinção do tipo de distorção envolvida. Mais à frente será estudada a precisão de ajuste para ambas as abordagens.

6.5.1 Método I

Nesta primeira abordagem, o conjunto de treinamento (sub-conjunto A do banco de notas e distorções) foi utilizado sem qualquer tipo de distinção entre os tipos de distorção. Assim, gerou-se um conjunto de coeficientes para diversas ordens M do polinômio aproximador, que foram avaliados a partir dos dados do sub-conjunto B do banco de notas e distorções: uma vez calculados os coeficientes, a partir de cada medida objetiva, foi calculado o seu voto estimado. A partir dos votos estimados, para cada condição de processamento a que se referiam, foi calculado um valor médio, que equivale a um MOS estimado. Após isso, com esse conjunto de 22 MOS estimados, verificou-se a correlação entre esses MOS estimados e os MOS reais *para os votos reais do sub-conjunto B*. Isto avalia a capacidade da função aproximadora, treinada

[10] Chamaremos de *classes de distorção* ou *classes de processamento* o conjunto de condições de processamento que se assemelhem quanto ao *tipo* de distorção que introduzem (e.g., LD-CELP em cascata ou MNRU para vários Q).

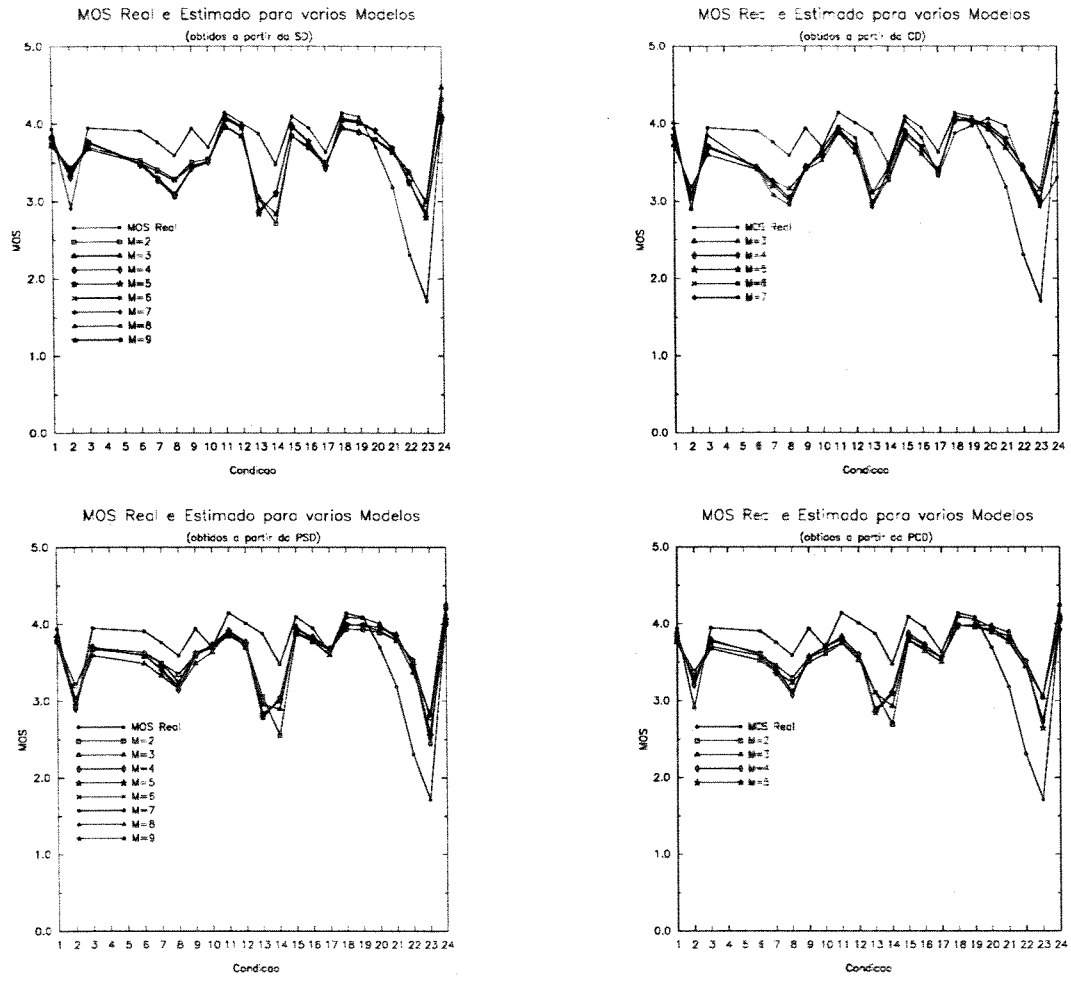


Figura 6.4: MOS real e estimado a partir das 4 medidas objetivas para o Método I considerando diversas ordens do modelo. Somente foram traçadas estimativas para ordens com Q não nulo e α superior a 95%.

a partir de um conjunto heterogêneo de dados, em estimar o MOS para uma condição (ou tipo de distorção) específica.

Os resultados desta abordagem estão na Tabela 6.9 e uma visão qualitativa dos ajustes pode ser vista na Figura 6.4. Na tabela estão apresentados a ordem M da equação (6.12), a qualidade $Q(\chi^2, \nu)$ do ajuste da equação (6.11) e o nível de confiança α da equação (6.14).

Na Figura 6.4 os valores do MOS real são apresentados com um ponto negro e os MOS estimados para diversas ordens estão marcados com outros ícones. Cada gráfico se refere a uma das medidas objetivas (SD, CD, PSD e PCD). As ordens M que possuem Q nulo ou $\alpha < 95\%$ na Tabela 6.9 não foram mostradas, pois seu ajuste mostrou-se muito ruim. Da figura pode-se observar que, para cada medida, existem várias ordens que produzem curvas praticamente idênticas. Outro aspecto, qualitativo, é que em geral as estimativas estão distantes dos valores MOS reais. Em geral, o valor MOS estimado esteve inferior ao valor real, exceto para as condições envolvendo o MNRU. Também pode-se notar que a tendência é de não haver boas estimativas para condições com baixos MOS reais. Assim, o Método I mostrou-se pouco eficiente para mapear as medidas objetivas em valores MOS estimados.

Tabela 6.9: Qualidade do ajuste $Q(\chi^2, \nu)$ e nível de confiança α, em percentagem, associado ao ajuste das medidas objetivas aos valores objetivos utilizando o Método I. Um ajuste ruim é indicado por $Q \rightarrow 0\%$. Um valor de $\alpha \rightarrow 100\%$ representa um bom ajuste e valores próximos a zero indicam um mau ajuste.

Ordem	SD		CD		PSD		PCD	
(M)	Q	α	Q	α	Q	α	Q	α
2	70.78	99.42	38.65	94.74	77.77	99.56	58.49	98.30
3	71.21	99.45	65.16	98.70	87.62	99.73	62.89	98.93
4	85.95	99.84	72.00	99.03	93.66	99.89	79.16	99.63
5	85.71	99.87	74.11	99.20	93.44	99.92	81.01	99.73
6	84.96	99.87	73.24	99.20	92.94	99.91	0.00	39.39
7	84.31	99.88	44.35	98.19	94.07	99.93	0.00	29.85
8	83.48	99.88	0.00	98.84	93.83	99.93	0.00	99.52
9	82.62	99.88	0.00	99.57	93.49	99.93	0.00	8.22

6.5.2 Método II

A outra abordagem que foi considerada é a de assumir que não se possa encontrar uma medida de uso amplo para as diversas classes de distorção. É necessário então gerar uma função aproximadora para cada classe $\mathcal{C}$ de distorção, assim definidas[11]:

- i Condições envolvendo o LD-CELP com erros de transmissão (condições 1, 2 e 3);
- ii Condições envolvendo o LD-CELP em tandem (condições 1, 6, 7, 8 e 9);
- iii Condições envolvendo conexões tandem da G.711 (condições 11 a 14);
- iv Condições envolvendo a G.721 em tandem (condições 10 e 15 a 17);
- v Condições envolvendo o MNRU e a condição direta (condições 18 a 24).

A identificação dos coeficientes foi realizada a partir dos dados constantes no sub-conjunto A do banco de dados. Diferentemente do Método I, somente os dados referentes às condições pertencentes a cada classe foram utilizados no treinamento. Após a determinação, para cada classe, da função aproximadora para diversas ordens, selecionava-se do sub-conjunto B, à semelhança do treinamento, as condições referentes somente à classe em questão. Então, para cada ordem do modelo, calculava-se o valor dos votos estimados para cada condição e então o seu valor MOS. O conjunto de valores MOS estimados eram então comparados com os MOS reais correspondentes, calculando-se a correlação e o α associado. Em termos qualitativos, avaliava-se, para diversas ordens, a capacidade da função aproximadora prever o MOS para distorções de uma dada classe.

A Tabela 6.11 mostra a qualidade do ajuste $Q(\chi^2, \nu)$ e o nível de confiança α para diversas ordens M do modelo para cada uma das cinco classes de distorção. A partir dessa tabela, pode-se excluir ordens de modelo inadequadas ($\alpha < 95\%$ e $Q(\chi^2, \nu) \approx 0$), que se encontra sumarizado na Tabela 6.10. Para a SD, a ordem M pode estar entre 3 e 9 para as Classes i e ii, entre 2 e 9 para as Classes iv e v e tem que ser 3 para a Classe iii. A CD tem bom ajuste para as ordens de 2 a 9 para as Classes ii e v, de 3 a 9 para a Classe i, de 2 ou 3 para a Classe iii e tem que ser 3 para a Classe iv. A PSD apresenta bom ajuste para M entre 3 e 9 para a

[11] Note-se que as condições com transcodificações mistas (9 e 10) foram colocadas respectivamente nas Classes ii e iv. Isto foi feito para se obter melhores resultados. Em estudos preliminares onde havia uma classe adicional, somente com as condições 9 e 10, praticamente todas as condições apresentaram ou $Q \approx 0$ ou $\alpha < 95\%$. A que melhor satisfez estes critérios (SD), mesmo assim apresentou um ajuste muito pobre. A divisão delas nos moldes apresentados foi a que melhor resultado apresentou.

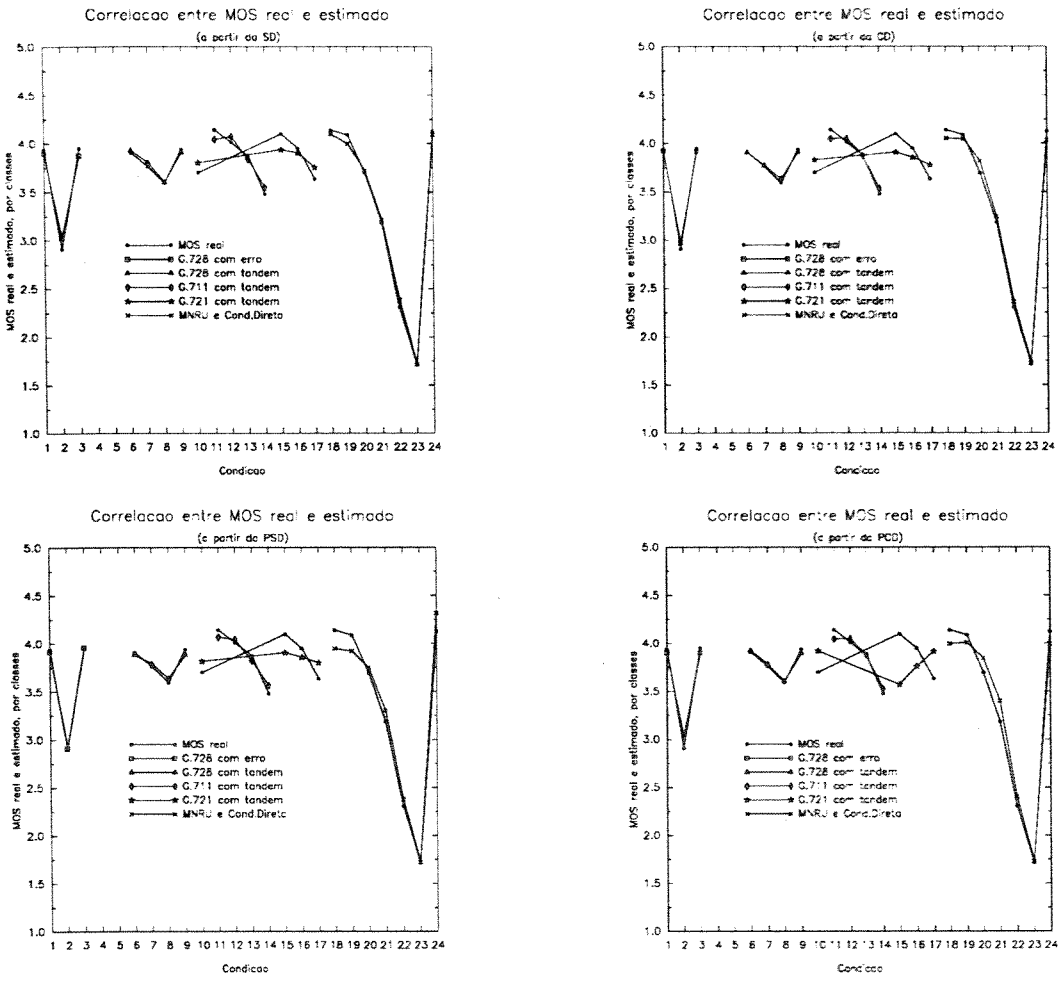


Figura 6.5: MOS real e estimado a partir das 4 medidas objetivas para o Método II para as diversas classes de distorção. As curvas se referem ao polinômio aproximador com a ordem que permitiu melhores resultados.

Classe i, entre 2 e 9 para as Classes ii e v e pode ser 2 ou 3 para as Classes iii e iv. Finalmente, a PCD aproxima bem os valores MOS para ordens de 3 a 7 para a Classe i, de 3 a 6 para a Classe 2, de 2 a 4 para a Classe iii, de 2 a 6 para a Classe v e pode ser 3 ou 7 para a Classe 4. Entre as várias ordens possíveis para a maioria das classes e medidas, pode-se observar que os valores de Q e de α muitas vezes são próximos. O resultado disso é, à semelhança do que ocorre na Figura 6.4, muitas curvas se sobrepõem. Isto nos permite selecionar para cada medida uma única ordem; um critério pode ser escolher os maiores valores de α e, quando muito próximos, optar pela menor ordem. Com esse critério, foram escolhidas para cada classe e medida as ordens indicadas entre parênteses na Tabela 6.10.

Na Figura 6.5 estão mostrados os valores MOS reais e estimados destacando-se cada uma das condições. Já a Figura 6.6 mostra um diagrama de dispersão dos valores previstos, destacando-se cada uma das classes. Quanto mais os pontos caem sobre a diagonal do gráfico, melhor é o ajuste entre os dados. O conjunto de coeficientes encontrados para as cinco funções aproximadoras está na Tabela 6.12. Estes são os valores efetivamente utilizados para a geração das figuras e tabelas descritas nesta seção.

Tabela 6.10: Ordens M que apresentaram $Q \neq 0$ e $\alpha > 95\%$ utilizando-se o Método II. Valores entre parênteses indicam as ordens com melhor ajuste.

Classe	SD	CD	PSD	PCD
i	3..9 (3)	3..9 (3)	3..9 (5)	3..7 (3)
ii	3..9 (4)	2..9 (7)	2..9 (3)	3..6 (5)
iii	3 (3)	2..3 (3)	2..3 (3)	2..4 (3)
iv	2..9 (3)	3 (3)	2..3 (3)	3,7 (7)
v	2..9 (6)	2..9 (5)	2..9 (3)	2..6 (5)

6.5.3 Análise dos resultados

Dois métodos de ajuste foram estudados a partir da bipartição do banco de notas e distorções: Método I e Método II. O Método I tentou identificar uma única função aproximadora que pudesse prever o MOS para qualquer das 22 condições da Tabela 6.3. O Método II identificou uma função aproximadora para cada uma das 5 classe de distorção em que as mesmas 22 condições foram agrupadas. O desempenho do Método I pode ser visto na Figura 6.4 e na Tabela 6.9. O desempenho do Método II encontra-se descrito nas figuras 6.5 e 6.6 e na Tabela 6.11.

Comparando figuras 6.4 e 6.5, pode-se concluir a baixa capacidade de predição dos valores MOS a partir de medidas objetivas do Método I e a alta correlação dos MOS estimados pelo Método II com os dados reais.

As outras medidas objetivas estudadas comportaram-se de modo muito semelhante para ambos os métodos de mapeamento, bem como na sua capacidade de predição de qualidade a partir de desempenhos relativos, como descrito na seção 6.3.4. Em especial para o Método II, destas 4 medidas de distorção espectral, a PCD foi a que apresentou pior desempenho devido ao comportamento inconsistente para as condições envolvendo a G.721 em cascata. Em geral, o pior ajuste verificado para todas as 4 medidas foi de fato com essas mesmas condições. Todas elas previram bem o comportamento do LD-CELP em presença de erros e em tandem e previram razoavelmente bem as condições com MNRU e a condição direta.

Observando-se a Figura 6.4, visualmente o melhor ajuste acontece para a SD, depois para a CD e então para a PSD. Entretanto, do ponto de vista de significância estatística, as 4 mostraram um desempenho semelhante.

6.6 Conclusão

Apresentamos inicialmente neste Capítulo um dos testes subjetivos realizados durante a avaliação para a língua portuguesa do algoritmo hoje padronizado pelo CCITT para codificação de voz a 16 kbit/s, o LD-CELP da Recomendação G.728. Junto à análise de sua qualidade apresentamos também o desempenho observado para outros algoritmos e condições de referência, descritos em mais detalhes nos Capítulos 2 e 3. Neste ponto concluímos pela adequação do LD-CELP a seus requisitos de qualidade e descrevemos o desempenho relativo de várias das condições de processamento.

Tabela 6.11: Qualidade do ajuste Q e nível de confiança α, em percentagem, associado ao ajuste pelo Método II das medidas objetivas aos valores objetivos. Um ajuste ruim é indicado por $Q \to 0\%$. Um valor de $\alpha \to 100\%$ representa um bom ajuste e valores próximos a zero indicam um mau ajuste.

Classe	Ordem	SD		CD		PSD		PCD	
($\mathcal{C}$)	(M)	Q	α	Q	α	Q	α	Q	α
i	2	98.82	92.02	99.62	89.51	99.82	88.16	99.00	89.57
	3	99.20	97.96	99.75	99.07	99.83	95.89	99.53	99.23
	4	99.94	96.38	99.87	95.77	99.83	98.24	99.75	97.55
	5	99.92	96.61	99.83	95.77	99.79	98.35	99.70	97.72
	6	99.90	96.73	99.79	95.77	99.73	98.35	99.63	97.70
	7	99.87	96.84	99.74	95.78	99.67	98.35	99.45	95.47
	8	99.83	96.84	99.67	95.88	99.59	98.35	95.12	88.11
	9	99.79	96.90	99.59	95.69	99.49	98.35	13.85	79.38
ii	2	99.99	92.77	99.99	95.23	99.99	96.55	99.99	95.93
	3	99.99	99.18	99.99	99.70	99.99	99.41	99.99	99.14
	4	99.99	99.37	99.99	99.97	99.99	97.66	99.99	99.22
	5	99.98	99.37	99.99	99.98	99.99	97.67	99.99	99.75
	6	99.98	99.34	99.98	99.98	99.99	97.67	99.98	97.63
	7	99.98	99.34	99.98	99.98	99.98	97.67	97.86	56.87
	8	99.97	99.34	99.98	99.98	99.98	97.67	5.29	4.22
	9	99.97	99.34	99.96	99.87	99.97	97.67	0.00	48.36
iii	2	100.00	91.73	100.00	96.10	100.00	96.05	100.00	96.91
	3	100.00	96.68	100.00	98.01	100.00	97.87	100.00	98.04
	4	100.00	88.12	100.00	93.12	100.00	93.83	100.00	95.73
	5	100.00	90.40	100.00	92.08	100.00	81.88	100.00	89.65
	6	100.00	90.40	100.00	89.23	100.00	81.25	96.02	48.47
	7	100.00	90.32	99.99	75.89	100.00	89.44	0.00	59.06
	8	100.00	90.34	0.00	47.82	100.00	89.55	0.00	15.48
	9	100.00	90.34	0.00	76.57	100.00	89.65	0.00	96.97
iv	2	99.89	99.55	99.87	94.88	99.88	98.60	99.85	92.66
	3	99.89	98.73	99.84	95.06	99.85	98.77	99.82	95.16
	4	99.87	97.29	99.85	76.30	99.87	83.61	99.88	72.19
	5	99.84	97.28	99.83	78.23	99.84	83.94	99.86	65.78
	6	99.80	97.26	99.85	64.12	99.80	83.97	99.83	71.59
	7	99.83	98.67	99.81	62.66	99.76	83.97	10.11	96.88
	8	99.80	98.61	99.75	62.23	99.71	83.97	0.00	96.43
	9	99.75	98.61	98.26	41.41	99.65	83.97	0.00	97.20
v	2	100.00	99.84	100.00	99.99	99.99	99.71	100.00	99.97
	3	100.00	99.93	100.00	99.99	100.00	100.00	100.00	99.99
	4	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.99
	5	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	6	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	7	100.00	100.00	100.00	100.00	100.00	100.00	0.00	20.01
	8	100.00	100.00	100.00	100.00	100.00	100.00	0.00	98.60
	9	100.00	100.00	100.00	100.00	100.00	100.00	0.00	45.77

Tabela 6.12: Conjunto de coeficentes calculados pelo Método II para as ordens de modelo escolhidas para as 4 medidas objetivas.

Medida	Classe	Ordem	Coeficientes
SD	i	3	(-10.3147; 100.8421; -178.6246)
	ii	4	(-12.2859; 133.6049; -353.2522; 295.0115)
	iii	3	(3.2594; 5.6603; -9.6108)
	iv	3	(2.2021; 13.7135; -27.0006)
	v	6	(-2.0226; 85.1125; -350.5664; 410.1680; 114.8819; -309.1200)
CD	i	3	(2.9594;1.1712;-0.3478)
	ii	7	(2.0303;1.4143;0.4180;-0.3675;-0.0500;0.0515;-0.0066)
	iii	3	(3.9782; 0.0943;-0.0270)
	iv	3	(4.0521;-0.1030;-0.0018)
	v	5	(3.6771; 1.9068;-2.1545; 0.5921;-0.0510)
PSD	i	5	(2.3908;42.3958;-330.8988;583.6711;370.4579)
	ii	3	(3.9173;2.1710;-23.0247)
	iii	3	(4.0983;-0.4241;-3.0280)
	iv	3	(3.9821;-1.4713;-4.7291)
	v	3	(4.6151;-23.5601;46.3281)
PCD	i	3	(2.5639;0.8977;-0.1453)
	ii	5	(-1.5222;4.7233;-1.4208;0.1780;-0.0081)
	iii	3	(3.8940;0.0908;-0.0121)
	iv	7	(0.1456;0.2833; 0.4561;0.4245;-0.2734;0.0480;-0.0027)
	v	5	(3.5818;1.0242;-0.6079;0.0916;-0.0044)

Após a descrição dos resultados subjetivos, descrevemos o material de voz utilizado nas análises objetivas e o pré-processamento nele realizado e apresentamos as medidas objetivas de interesse: duas baseadas no espectro real (SD e CD) e duas com motivação perceptual (PSD e PCD). Estas foram descritas em detalhes no Capítulo 5. Foi apresentada uma análise objetiva da qualidade através do estudo do desempenho relativo entre os diversos processamentos para cada um das medidas.

Descreveu-se um método estatístico para realizar a conversão das medidas objetivas em valores MOS estimados e estudamos duas abordagens para determinar as funções aproximadoras e avaliar a qualidade de seu ajuste. Concluímos que para os dados analisados não foi possível determinar uma única função aproximadora para qualquer das medidas objetivos que fosse capaz de predizer adequadamente a qualidade subjetiva associada a qualquer das classes de processamento. Por outro lado, verificamos a ocorrência de bons ajustes quando agrupando as condições de processamento em classes e determinando uma função aproximadora para cada uma dessas classes.

Na análise realizada, concluímos que as medidas baseadas na distorção do conteúdo espectral do sinal apresentaram bons resultados para indicar a qualidade objetiva (comparando o desempenho relativo entre as diversas condições de processamento) e a qualidade subjetiva (através da conversão dos valores objetivos em subjetivos estimados) quando as condições de processamento foram agrupadas em cinco classes. Entre as medidas espectrais de distorção, a que apresentou o pior ajuste aos valores MOS foi a PCD. As outras 3 não foram significativamente diferentes do ponto de vista estatístico.

Analisando-se o desempenho de todas as medidas e utilizando a complexidade computacional associada a cada medida como um fator de compromisso, poder-se-ia optar pela distância cepstral truncada com compensação de ganho implícito como um bom estimador de baixa

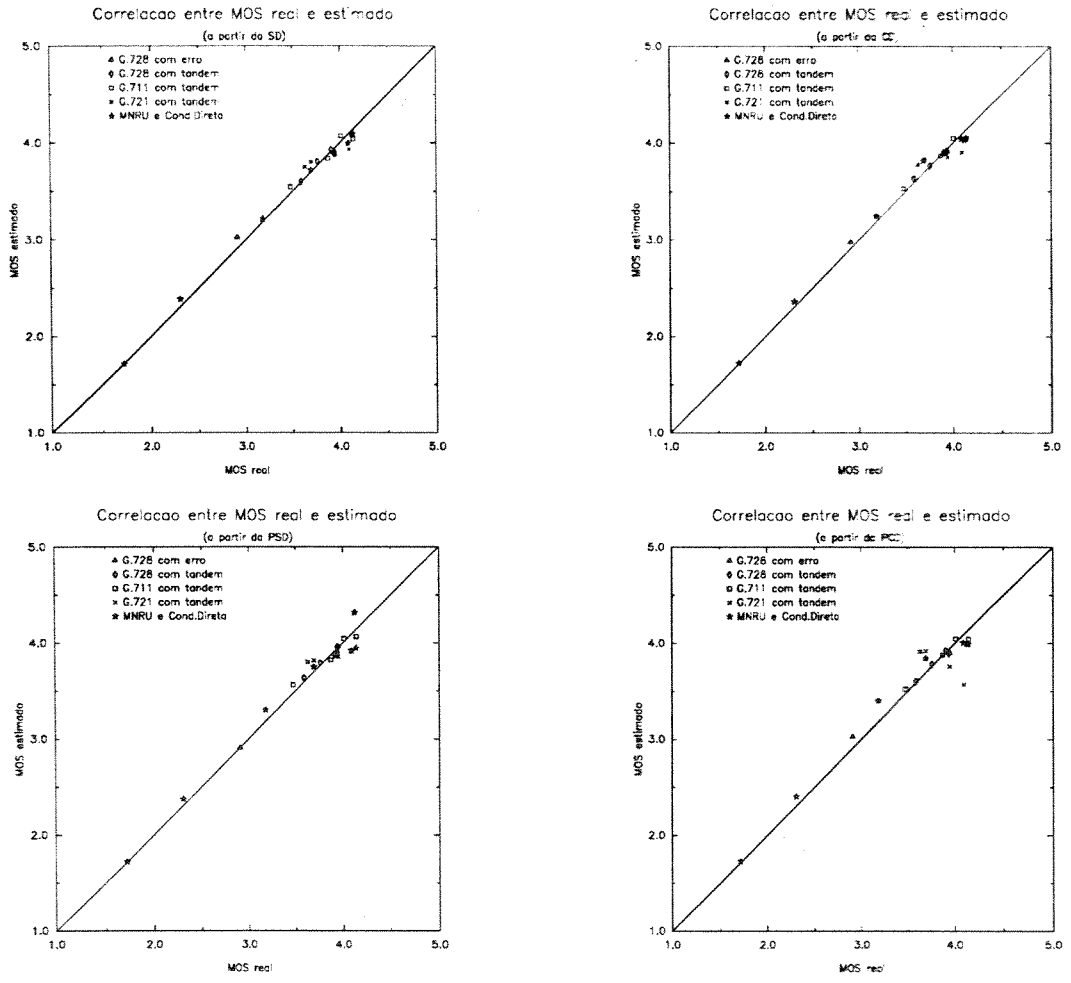


Figura 6.6: Diagrama de dispersão entre o MOS real e estimado a partir das 4 medidas objetivas para o Método II por classes de distorção. As curvas se referem à ordem que permitiu melhores resultados. O acúmulo dos pontos sobre a reta $x=1$ indica um bom ajuste entre dados estimados e reais.

complexidade para qualidade subjetiva quando se utilizando os coeficientes poliniomiais apresentados na Tabela 6.12.

Capítulo 7

Conclusão

Nos capítulos iniciais deste trabalho apresentamos uma visão geral dos diversos aspectos envolvidos na análise de qualidade de algoritmos de codificação de voz para aplicação em telefonia.

Inicialmente descrevemos três algoritmos importantes padronizados pelo CCITT. A G.711 especifica o formato digital mais comum até hoje em sistemas de transmissão digital em todo o mundo e se constitui na referência básica de qualidade, em termos da qual normalmente se expressa a distorção introduzida por processamento digitais na rede (qdu). A G.721, hoje substituída pela G.726, é o primeiro padrão mundial em que se utilizam técnicas de codificação digital de sinais e é uma importante referência de qualidade. A G.728 recebeu destaque no Capítulo 2; foi recentemente aprovada e estabelece o LD-CELP para a codificação a 16 kbit/s de sinais na faixa de voz para uso na rede telefônica. É um algoritmo complexo e o seu processo de padronização foi o fato originador deste trabalho. Apesar de codificar sinais em metade da taxa da G.721, este algoritmo apresenta uma qualidade equivalente porque utiliza eficiente e massivamente técnicas de PDS. Chega até a ser melhor que a G.721 em alguns aspectos, como em presença de erros de transmissão.

Porém descrevermos somente estes algoritmos de referência não era suficiente: foi necessário também descrever outros algoritmos normalmente utilizados em avaliações de qualidade, em especial o MNRU e o EID. Junto com esses algoritmos surge a necessidade de descrever uma infra-estrutura tanto de laboratório como de software. Isto foi feito no Capítulo 3.

Descritos os algoritmos e colocada uma infra-estrutura, introduzimos no Capítulo 4 a metodologia para a realização de testes subjetivos, de longe o mais denso dos capítulos deste trabalho. Começamos apresentando um histórico de avaliações subjetivas e passamos então por diversos dos aspectos envolvidos com o projeto de um tipo específico de teste subjetivo, os testes de audição. Detalhamos aspectos relacionados à definição de seu tipo e tamanho, à geração do material fonte, ao processamento deste e aos procedimentos para audição do material processado. Isto feito, apresentamos diversos métodos estatísticos normalmente empregados na análise dos resultados deste tipo de testes. Ainda nesse capítulo apresentamos exemplos de sentenças a serem utilizadas em testes subjetivos, de intruções aos ouvintes, bem como de dois planos bastante simples de testes subjetivos.

Para completar a parte tutorial, descrevemos no Capítulo 5 diversas medidas objetivas de avaliação de qualidade baseadas em técnicas temporais, espectrais e com motivação psicoacústica. Também descrevemos sumariamente algumas das medidas em estudo pelo CCITT.

Escolhemos quatro dessas medidas de distorção, de forma que representassem as diferentes abordagens mais frequentemente encontradas na literatura: a distância espectral (SD) e a distância cepstral (CD) como medidas de distorção a partir do conteúdo espectral "objetivo" dos sinais processados e finalmente duas medidas baseadas na distorção do espectro "subjetivo", a partir da aplicação de um modelo psicoacústico aos sinais e o cálculo da contrapartida perceptual das distâncias espectral e cepstral, chamadas respectivamente de distâncial espectral perceptual (PSD) e a distância cepstral perceptual (PCD). Neste trabalho não foram testadas medidas de distorção temporais devido às características dos sinais de teste e de referência utilizados, que invalidavam o uso desta classe de medidas, conforme descrevemos no Capítulo 6.

Para avaliar a aplicabilidade de cada uma dessas quatro medidas, apresentamos no Capítulo 6 o seu desempenho baseado no material de voz usado realmente na Fase II de testes subjetivos do LD-CELP (G.728) para a língua portuguesa. Antes disso, descrevemos os aspectos relevantes da Fase II de testes subjetivos como um embasamento para a análise das medidas objetivas e apresentamos o desempenho subjetivo do LD-CELP, bem como das condições de referência (G.711, G.721, MNRU e condição direta). Em seguida, analisamos cada uma das quatro medidas objetivas em termos do seu desempenho relativo para as diversas condições de processamento.

O próximo passo foi avaliar, ainda no Capítulo 6, a capacidade de cada uma das medidas para prever a qualidade subjetiva. Para tanto, o material processado avaliado na Fase II de testes do LD-CELP foi dividido em 2 sub-conjuntos de igual tamanho. Ao primeiro deles foi aplicado o método dos mínimos quadrados para se identificar polinômios que mapeassem as medidas objetivas em valores MOS estimados. A outra parte do material processado foi utilizada para se avaliar se o polinômio era de fato capaz de estimar a qualidade subjetiva. Exploramos duas abordagens básicas: identificação de um polinômio *único* para cada medida, que pudesse ser aplicado a qualquer uma das condições de processamento *(Método I)* e a identificação de vários polinômios para cada medida, cada qual correspondendo a uma Classe de Processamento *(Método II)*.

Como resultados importantes temos que não foi possível identificar boas funções aproximadoras pelo Método I. Já o Método II apresentou bons resultados para as medidas consideradas.

As quatro medidas espectrais comportaram-se de maneira muito semelhante tanto pelo Método I como pelo Método II. Em especial para este último, em que bons ajustes foram obtidos, a distância cepstral perceptual foi a que apresentou o pior comportamento, em especial quando prevendo a qualidade da G.721 em tandem. Considerando a complexidade computacional como um fator de escolha da medida, a *distância cepstral* parece ser um bom estimador da qualidade subjetiva quando se utiliza o Método II e os coeficientes da Tabela 6.12.

Um aspecto desapontador foi o desempenho das medidas baseadas no modelo psicoacústico do Capítulo 5. De fato, seu desempenho foi estatisticamente equivalente ao das medidas baseadas no espectro "objetivo". A comparação com os resultados apontados pela literatura, que são altamente promissores para a medidas com motivação psicoacústica, indica que talvez o modelo psicoacústico considerado seja inadequado e precise ser refinado.

Antes de finalizar, é importante ressaltar o caráter tutorial deste trabalho. Os testes de algoritmos são normalmente definidos sem o rigor necessário. Com esse trabalho esperamos poder fornecer fundamentos para a definição mais precisa de metodologias de avaliação de

algoritmos de codificação de voz em desenvolvimento por pesquisadores nos diversos grupos de pesquisas no Brasil.

Finalizando, neste trabalho sumarizamos um esforço que vem sendo desenvolvido há mais de quatro anos, desde que nos envolvemos com as atividades de padronização do codificador a 16 kbit/s pelo CCITT, apresentando os conceitos básicos que tivemos que elaborar em seu decurso e buscamos, em seu final, identificar métodos que possam facilitar o complexo trabalho que é, como o leitor deve agora concordar, avaliar a qualidade de algoritmos de codificação de voz.

Bibliografia

[1] D.A. Mindell. Dealing with a digital world. *Byte*, pages 246–256, August 1989.

[2] L. Gun. DSP uproots traditional analog jobs. *Electronic Design*, pages 49–58, September 28, 1989.

[3] Texas Instruments. *First Generation TMS320 User's Manual*, 1987. SPRU013.

[4] Texas Instruments. *Second Generation TMS320 User's Manual*, 1987. SPRU014.

[5] Texas Instruments. *TMS320C3x User's Manual*, 1990. SPRU031A 2558539-9702 Rev.A.

[6] Texas Instruments. *TMS320C4x User's Manual*, 1991. SPRU063 2564090-9761 Rev.A.

[7] Motorola. *DSP56000/DSP56001 Digital Signal Processor User's Manual*, 1990. DSP56000UM/AD Rev.2.

[8] AT&T. *DSP32C Digital Signal Processor*, January 1990. DS89-060DMOS Issue 3.0.

[9] M.L. Fuccio, R.N. Gadenz, C.J. Garen, J.M. Huser, B. Ng, S.P. Pekarich, and K.D. Ulery. The DSP32C: AT&T's second generation floating point DSP. *IEEE Micro*, pages 30–48, December 1988.

[10] R. Weiss. 32-bit floating point DSP processors. *EDN*, pages 127–146, November 7, 1991.

[11] User's Group on Software Tools. *CCITT Software Tool Library Manual*. CCITT SG XV, May 1992. COM XV-R 87-E.

[12] CCITT. *Recommendation G.191, Software Tools for Speech and Audio Coding Standards*. ITU, Geneva, 1993. To be approved by the Xth Plenary Assembly, March 1993. Current text found in COM XV-R 86-E.

[13] IXth Plenary Assembly. Question 13/XII, Methods for the evaluation of non-linear distortions. In *List of questions proposed for study during the 1989-1992 Study Period*, pages 20–22. CCITT, Melbourne, January 1989. Doc.No.COM XII-R 1-E.

[14] SG XII. Question X3/XII, Methods for the measuring and modelling the effects of non-linear processes on the speech quality of transmission systems. In *Report to the Xth CCITT Plenary Assembly - Wording of questions proposed for study during the 1993-1996 Study Period*, pages 15–16. CCITT, Helsinki, March 1992. Doc.No.AP X-8-E.

[15] J. Fennick. *Quality Measures and the design of telecommunications systems* . Artech House, 1988.

[16] CCITT. *Recommendation G.711. Pulse code molulation (PCM) of voice frequencies*, volume Fascicle III.4 of *Blue Book*, pages 175–184. ITU, Geneva, 1989.

[17] N.S. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice-Hall. 1984.

[18] Rabiner,L.R. and Schafer,R.W. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, 1978.

[19] CCITT. *Recommendation G.721, 32 kbit/s adaptive differential pulse code modulation (ADPCM)*, volume Fascicle III.3 of *Red Book*, pages 125–159. ITU, Geneva, 1985.

[20] W.R. Daumer, X. Maitre, M. Mermelstein, and I. Tokizawa. Overview of the ADPCM coding algorithm. *Proc.Globecom*, pages 774–777, 1984.

[21] CCITT. *Recommendation G.721, 32 kbit/s adaptive differential pulse code modulation (ADPCM)*, volume Fascicle III.4 of *Blue Book*, pages 231–268. ITU, Geneva, 1989.

[22] CCITT. *Recommendation G.763, Digital Circuit Multiplication Equipment using 32 kbit/s ADPCM and digital speech interpolation*. ITU, Geneva, Jun 1991.

[23] CCITT. *Recommendation G.723, Extensions of Recommendation G.721 Adaptive Differential Pulse Code Modulation (ADPCM) to 24 and 40 kbit/s for Digital Circuit Multiplication Equipment Application*, volume Fascicle III.4 of *Blue Book*, pages 341–358. ITU, Geneva, 1989.

[24] CCITT. *Recommendation G.726, 40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)*. ITU, Geneva, 1991.

[25] M. Bonnet, O. Macchi, and M. Jaidane-Saidane. Theoretical analysis of the ADPCM CCITT algorithm. *IEEE Trans.on Communications*, 38(6):847–858, June 1990.

[26] M. Bonnet, O. Macchi, and M. Jaidane-Saidane. Theoretical analysis of the ADPCM CCITT algorithm. *IEEE Trans.on Communications*, 38(6):847–858, June 1990.

[27] M.H.. Sherif. Série de Comunicações Privadas, 18–29/Set/1992.

[28] CCITT SG XV. Q.21/XV- Requirements and Objectives. Contribution COM XV-R -E, CCITT, July 1990.

[29] Iyengar,V. and Kabal,P. A Low Delay 16 kbit/sec Speech Coder. In *ICASSP*, pages 243–246, 1988.

[30] AT&T. Description of 16 kbit/s Low-Delay Code-Excited Linear Predictive Coding (LD-CELP) algorithm. Contribution AH.89-D02, CCITT, March 1989.

[31] BNR-Canada. Letter to the Chairman of CCITT Ad-hoc Group for Q.21/XV. Contribution AH.89-D13, CCITT, March 1989.

[32] Consortium for Speech Coding. Letter to the Chairman of CCITT Ad-hoc Group for Q.21/XV. Contribution AH.89-D13, CCITT, March 1989.

[33] Consortium for Speech Coding. A High Level Description of the Consortium's Low Delay Vector Excitation Coder (LD-VCX). Contribution AH.89-D08, CCITT, March 1989.

[34] Consortium for Speech Coding. Description of the Consortium's Low Delay Vector Excitation Coder (LD-VCX), Version 2. Contribution AH.89-D21, CCITT, July 1989.

[35] V. Cuperman, A. Gersho, R. Pettigrew, J.J. Shynk, and J.-H. Yao. *Backward Adaptive Configurations for Low-delay Vector Excitation Coding*, pages 13–23. Kluwer Academic Publishers, 1989.

[36] South,C.R. and Usai,P. Subjective Performance of CCITT's 16 kbit/s LD-CELP algorithm with voice signals. In *Globecom*. IEEE, 1992.

[37] CCITT. *Recommendation G.728, Coding of Speech at 16 kbit/s using Low Delay Code Excited Linear Prediction (LD-CELP)*. ITU, Geneva, 1992.

[38] Schroeder,M.R. and Atal,B.S. Code-Excited Linear Prediction (CELP): high speech quality at very low rates. In *ICASSP*, pages 937–940. IEEE, 1985.

[39] Cox,R.V. and Johansen,F.T. Test Verification of LD-CELP: an objective measurement approach to a non-bit exact standard. In *Globecom*. IEEE, 1992.

[40] Chen,J.-H. and Lin,Y-C. and Cox,R.V. A fixed-point 16 kb/s LD-CELP algorithm. In *ICASSP*, pages 21–24. IEEE, 1991.

[41] Barnwell III,T.P. Recursive windowing for generating autocorrelation coefficients for LPC analysis. *Transactions on Acoust., Speech, Signal Processing*, ASSP-29(5):1062–1066, October 1981.

[42] J.-H. Chen, R.V. Cox, Y.-C. Lin, N. Jayant, and M.J. Melchner. A Low-Delay CELP Coder for the CCITT 16 kb/s Speech Coding Standard. *IEEE Journal on Selected Areas in Communications*, 10(5):830–849, June 1992.

[43] AT&T. Improvements to AT&T's 16 kbit/s Low-Delay Code-Excited Linear Predictive Coding (LD-CELP) algorithm. Contribution AH.89-D26, CCITT, July 1989.

[44] J.-H. Chen. *A robust Low-Delay CELP Speech Coder at 16 kb/s*, pages 25–35. Kluwer Academic Publishers, 1989.

[45] J.-H. Chen and R.V. Cox. Série de Comunicações Privadas, 21–28/Set/1992.

[46] Yao,J.-H. and Gersho,A. Real-time vector APC speech coding at 4800 bps with adaptive post-filtering. In *ICASSP*, pages 2185–2188. IEEE, 1987.

[47] Grant,D. and Young,M. and Gersho,A. Real Time Vector Excitation Coding of Speech at 4800 bps. In *ICASSP*, pages 2189–2129. IEEE, 1987.

[48] Campos Neto, S.F. and Irii, H. and Rosenberger,J. and Sotcsheck,J. and Usai,P. Effect of Tandeming and Input Level. *CSELT Technical Reports*, XXI(2–2):7, 1993.

[49] J.-H. Chen and R.V. Cox. LD-CELP: A high quality 16 kb/s speech coder with low delay. In *Globecom*, pages 528–532. IEEE, 1990.

[50] J.-H. Chen. A robust Low-Delay CELP Speech Coder at 16 kb/s. In *Globecom*, pages 1237–1241. IEEE, 1989.

[51] Ramamoorthy,V. and Jayant,N.S. and Cox,R.V. and Sondhi, M.M. Enhancement of ADPCM Speech Coding using Backward-Adaptive Algorithms for Post-Filtering and Noise Feedback. *Journal on Selected Areas in Communications*, 6(2):364–382, February 1988.

[52] AT&T. Detailed description of AT&T's LD-CELP algorithm. Contribution AH.89-D02, CCITT, November 1989.

[53] III Barnwell, T.P. Frequency variant spectral distance measures for speech quality testing. In *National Electronics Conf.*, pages 246–250, Chicago, USA, Oct 1981.

[54] P. Noll. Adaptive quantizing in speech coding systems. In *Int. Zurich Seminar on Digital Comm.*, pages B3.1–B3.6. IEEE, 1974.

[55] A.H. Gray and J.D. Markel. Distance Measures for Speech Processing. *IEEE Trans.on Acous,Speech and Sig.Proc.*, ASSP-24(5):0, Oct 1976.

[56] S. Wang, A. Sekey, and A. Gersho. An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications*, 10(5):819–829, June 1992.

[57] K. Itoh, N. Kitawaki, and K. Kakehi. Objective quality measures for speech waveform coding systems. *Review of the Electrical Communication Laboratories*, 32(2):220–228, 1984.

[58] R.E. Crochière, L.R. Rabiner, N.S. Jayant, and J.M. Tribolet. A Study of Objective Measures for Speech Waveform Coders. In *Int.Zurich Seminar on Digital Comm.*, pages H.1.1–H1.7. IEEE, March 1978.

[59] S.R. Quackenbush, III Barnwell, T.P., and M.A. Clements. *Objective Measures of Speech Quality*. Signal Processing Series. Prentice-Hall, 1988.

[60] T. Moriya and M. Honda. Speech coder using phase equalization and vector quantization. In *ICASSP*, pages 1701–1704, 1986.

[61] R.E. Crochière. An analysis of 16 kb/s sub-band coder performance: dinamic range, tandem connections, and channel errors. *Bell Syst. Tech. J.*, 57(8):2927–2952, Oct 1978.

[62] P. Mermelstein. Evaluation of a segmental SNR measure as an indicator of the quality of ADPCM coded speech. *J.Acoust.Soc.Am.*, 66(6):1664–1667, 1979.

[63] D.J. Goodman, C. Scagliola, R.E. Crochière, L.R. Rabiner, and J. Goodman. Objective and subjective performance of tandem connections of waveform coders with an LPC vocoder. *Bell Syst.Tech.J.*, 58(3):601–629, Mar 1979.

[64] CCITT Study Group XII. Annex G to Supplement no.3, 'Objective Method of Estimating the Quality of Speech Degraded by Non-linear Distortion'. Technical Report COM XII-R 30-E, UIT, Geneva, April 1992.

[65] Hygino H. Domingues, Carlos A. Callioli, and Roberto C.F. Costa. *Álgebra Linear e Aplicações*. Editora Atual, São Paulo, 1982.

[66] Alan V. Oppenheim and Ronald W. Schafer. *Digital Signal Processing.* Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[67] N. Kitawaki, M. Honda, and K. Itoh. Speech Quality Assessment Methods for Speech Coding Systems. *IEEE Comm.Mag.*, 22(10):26–33, Oct 1984.

[68] B.S Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J.Acoust.Soc.Am.*, 55(6):1304–1312, Jun 1974.

[69] K.D. Kryter. *Methods for the calculation of articulation index.* ANSI, NY, 1969. Std.No. S3.5-1969.

[70] R.E. Crochière, J.M. Tribolet, and L.R. Rabiner. An interpretation of the log likehood ratio as a measure of waveform coder performance. *IEEE Trans. Acoust., Speech. Signal Processing*, ASSP-28(3):318–323, Jun 1980.

[71] Arthur C. Guyton. *Fisiologia Humana.* Ed.Guanabara, 6a. edition, 1985.

[72] Harvey Fletcher. *Speech and Hearing in Communication.* D.Van Nostrand Co., Toronto, 1953.

[73] Douglas O'Shaughnessy. *Speech Communication: Human and Machine.* Addison-Wesley, 1987.

[74] Schroeder, B.S Atal, and Hall. Optimizing Dig. Speech Coders by Exploting Masking Properties of the Human Ear. *J.Acoust.Soc.Am.*, 66(6):1647–1652, Dec 1979.

[75] A.J. Fourcin et al. *Speech Processing by Man and Machine: Group Report*, pages 307–351. Bullock,T., 1977.

[76] Y. Mahieux. High-quality audio transform coding at 64 kbit/s. *Ann.Télécommun.*, 47(3–4):95–106, 1992.

[77] Ramamoorthy,V. and Jayant,N.S. Adaptive Post-Filtering of 16 kb/s-ADPCM Speech. In *ICASSP*, pages 829–832. IEEE, 1986.

[78] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J.Acoust.Soc.Am.*, 87(4):88–94, April 1990.

[79] PTT The Netherlands. Measuring the quality of audio devices. Contribution SQ-27.91, CCITT Study Group XII – Speech Quality Experts Group, Florence, July 1991. Reprint of paper presented at the 90th AES Convention, Paris, Feb.1991.

[80] J. Lalou. The information index: an objective measure of speech transmission performance. *Annales des Telecommunications*, 45(1–2):47–65, 1990.

[81] D.J. Goodman and R.D. Nash. Subjective quality of the same speech transmission conditions in seven different countries. *IEEE Trans. on Communications*, COM-30(4):642–654, Apr 1982.

[82] J.L. Flanagan, M.R. Schroeder, B.S. Atal, R.E. Crochière, N.S. Jayant, and J.M. Tribolet. Speech Coding. *IEEE Transactions on Communications*, COM-27(4):710–736, Apr 1979.

[83] D.L. Richard. *Telecommunication by Speech.* John Wiley, NY and Butterworths, London, 1973.

[84] Proc.Globecom. *Subjective Performance Evaluation of the 32 kbit/s ADPCM algorithm*, 1984.

[85] A. Coleman, N. Gleiss, H. Scheuermann, J. Sotscheck, and P. Usai. Subjective performance evaluation of the RPE-LTP codec for the pan-european cellular digital mobile radio system. *CSELT Technical Reports*, XVIII(2):89–94, April 90.

[86] D.J. Goodman. An IEEE Meeting on Subjective Testing of voiceband codecs. *IEEE Communications Society Magazine*, 15:4–5, Nov 1977.

[87] S. Kotz, N.L. Johnson, and C.B. Read. *Encyclopedia of Statistical Sciences*, volume 6. John Wiley & Sons, New York, 1982.

[88] G. Modena, A. Coleman, P. Usai, and P. Coverdale. Subjective Performance Evaluation of the 7 KHz Audio Coder. In *GLOBECOM*, Houston, Texas, Dec.1–4 1986.

[89] CCITT. *Recommendation P.80, Methods for the subjective determination of transmission quality*, volume V of *Blue Book*, pages 197–198. ITU, Geneva, 1989.

[90] BTRL Colin R.South. Phase II Subjective Test methodology for a 16 kbit/s speech coder. Contribution SQ-12.91, CCITT Study Group XII, April 30 1991.

[91] C.R.South, G.J.Barnes, N.Gleiss, H.Irii, P.Combescure, D.Pascal, and P.Usai. Subjective Test methodology for a 16 kbit/s speech coder. Contribution SQ-10.89R, Study Group XII Speech Quality Experts Group, March 1 1990.

[92] S. Kotz, N.L. Johnson, and C.B. Read. *Encyclopedia of Statistical Sciences*, volume 2. John Wiley & Sons, New York, 1982.

[93] E.E. Bonini and S.E. Bonini. *Estatística. teoria e exercícios.* FMU, Sõ Paulo, 1972.

[94] Brasil. Comparison of UGST and CSELT MNRU software implementations. Contribution, SG XII/CCITT, 1992.

[95] S. Armbrust. *Telefonometria Básica.* Telebrás, Brasília, 2a. edition, 1992.

[96] S. Kotz, N.L. Johnson, and C.B. Read. *Encyclopedia of Statistical Sciences*, volume 3. John Wiley & Sons, New York, 1982.

[97] S. Kotz, N.L. Johnson, and C.B. Read. *Encyclopedia of Statistical Sciences*, volume 4. John Wiley & Sons, New York, 1982.

[98] J. Dénes and A.D. Keedwell. *Latin squares and their applications.* Akadémiai Kiadó, Budapest; English Universities Press, London; Academic Press, New York, 1974. O principal livro de referência sobre quadrados latinos.

[99] R.M. Bethea, B.S. Duran, and T.L. Boullion. *Statistical Methods for Engineers and Scientists*, volume 57 of *Statistics: textbooks and monographs.* Marcel Dekker, New York, 2 edition, 1985.

[100] G.H. Freeman. Complete Latin Squares and Related Experimental Designs. *Journal of the Royal Statistics Society. Series B*, 41(2):253–262, 1979.

[101] R.C. Bose. On the application of properties of Galois fields to the problem of construction of hypergraeco latin squares. *Sankhyā*, 3:323–338, 1938.

[102] R.C. Bose, S.S. Shrikhande, and E.T. Parker. Further results on the construction of mutually orthogonal latin squares and the falsity of Euler's conjecture. *Canadian Journal of Mathematics*, XII(2):189–203, 1960.

[103] H.B. Mann. The Construction of Orthogonal Latin Squares. *The Annals of Mathematical Statistics*, XIII(4):418–423, December 1942.

[104] R.A Fisher and F. Yates. *Tabelas Estatísticas para Biologia, Medicina e Agricultura.* Ed.Univ.São Paulo e Ed.Polígono, S.Paulo, 6th. edition, 1971.

[105] B.S. Boob and H.L. Agrawal. A note on the construction of mutually orthogonal Latin squares. *Biometrics*, 32(1):191–193, March 1976.

[106] W. Volk. *Engineering statistics with a programmable calculator.* McGraw-Hill, 1982.

[107] R.J. Barlow. *Statistics - A Guide to the use of statistical methods in the physical sciences.* The Manchester physics series. John Wiley & Sons, Chichester, UK, 1989.

[108] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes in C - The Art of Scientific Programming.* Cambridge University Press, Bilbao, 1988.

[109] S. Kotz, N.L. Johnson, and C.B. Read. *Encyclopedia of Statistical Sciences*, volume 1. John Wiley & Sons, New York, 1982.

[110] N.L. Johnson and F.C. Leone. *Statistics and Experimental Design*, volume 2. John Willey & Sons, New York, 1964.

[111] CCITT. *Recommendation P.83, Subjective performance assessment of Telephone Band and Wideband Digital Codecs.* ITU, Geneva, 1992.

[112] AT&T, BT, CNET, and NTA. Proposed Section on Curve fitting for P.84. Contribution SQ-57.91, CCITT Study Group XII – Speech Quality Experts Group, Arizona, Dec 3-6 1991.

[113] CCITT. *Recommendation P.48, Specification for an intermediate reference system* , volume V of *Blue Book*, pages 81–86. ITU, Geneva, 1989.

[114] User's Group on Software Tools. Software Tools for the Qualification Test of a Codec at 8 kbit/s. Contribution, CCITT SG XV, May 1992. COM XV-R 88-E.

[115] CCITT. *Recommendation G.712, Performance Characteristics of PCM channels.* ITU, Geneva, 1992. This is the new version to be approved by the Xth Plenary Assemby.

[116] Hewlett-Packard, USA. *HP8903B Audio Analyser: Operating and Calibration Manual*, August 1990.

[117] Study Group XII Speech Quality Experts Group. Digital Interchange Method, v1.0. Contribution SQ-02.90, CCITT, 1990.

[118] Study Group XII. An interface for measuring ISDN digital telephone set performance based on the concepts of an ideal codec and a direct approach. Contribution COM-XII-7-E, CCITT, 1989.

[119] CCITT. *Recommendation P.56. Objective measurement of active speech level*, volume V of *Blue Book*, pages 110–120. ITU, Geneva, 1989.

[120] CCITT. *Recommendation P.81. Modulated Noise Reference Unit (MNRU)* , volume V of *Blue Book*, pages 198–203. ITU, Geneva, 1989.

[121] T. Watanabe, H. Nagabushi, and N. Kitawaki. Law of addition for subjective Quality of Voiceband codecs. *Review of Electrical Comm.Labs*, 5(4):0, 1987.

[122] CCITT. *Recommendation G.113, Transmission Impairments*, volume Fascicle III.1 of *Blue Book*, pages 63–84. ITU, Geneva, 1989.

[123] Study Group XII Speech Quality Experts Group. Definition of QDU and use of MNRU. Contribution SQ-15.88, CCITT, 1988.

[124] CCITT. *Handbook on Telephony* . ITU, Geneva, 1987.

[125] Ana F.P.C. Humes, Wagner T. Martins, Inês S.H. Melo, and Luzia K. Yoshida. *Noções de Cálculo Numérico*. McGraw-Hill, São Paulo, 1984.

[126] CPqD/Telebrás-Brazil. Measure of active level with and without decimation. Contribution D.387(XV/2), CCITT, WP XV/2, Geneva, November 1991.

[127] H.B Law and R.A. Seymour. A reference distortion system using modulated noise. *Proc.Institution of Electrical Engineers (IEE)*, 109B:484–485, Nov 1962.

[128] CCITT. *Recommendation G.722, 7 kHz audio-coding withing 64 kbit/s*, volume Fascicle III.4 of *Blue Book*, pages 269–341. ITU, Geneva, 1989.

[129] Comsat. Report on the Activities of the 16kbit/s Subjective Evaluation Host Processing Laboratory. Contribution SQ-18.91, CCITT Speech Quality Experts Group. July 1991.

[130] Simão Ferraz de Campos Neto. Metodologias de avaliação de Algoritmos de Codificação de Voz. Tese de mestrado, FEE/ Unicamp, Abril 1992.

[131] CSELT. Subjective Assessment of the CCITT 16 kbit/s coding algorithm (Listening Tests in Portuguese, Italian, Japanese and English American) – Phase II. Contribution SQ-21.91R, CCITT, SQEG/XII, Florence, July 1991.

[132] CPqD/Telebrás. Subjective Assessment of the CCITT 16 kbit/s coding algorithm. Brazil: Listening Tests in Portuguese Language – Phase II. Contribution SQ-30.91, CCITT, SQEG/XII, Florence, July 1991.

[133] CSELT. Subjective Assessment of the CCITT 16 kbit/s coding algorithm. Brazil: Listening Tests in Portuguese Language – Phase II. Contribution SQ-22.91, CCITT, SQEG/XII, Florence, July 1991.

[134] SQEG/XII. Liaison Statement to the SG XV Ad-hoc Group on Phase II Testing of the 16 kbit/s LD-CELP codec. Contribution SQ-42.91, CCITT, Florence, July 1991.

[135] Chairman SQEG/XII. Report of the meeting. Contribution SQ-49.91, CCITT, Florence, July 1991.

[136] Brian W. Kernighan and Dennis M. Ritchie. *The C Programming Language*. Prentice Hall Software. Prentice Hall, Englewood Cliffs, 2 edition, 1988.

[137] Alfred V. Aho, Brian W. Kernighan, and Peter J. Weingerger. *The AWK Programming Language*. Computer Science. Addison-Wesley, October 1988.

[138] Diane Barlow Close et al. *The GAWK Manual*. Free Software Foundation, Cambridge, 0.1β edition, March 1989. GNU Project.

[139] Ramkrishna S. Tare. *UNIX utilities*. Prentice Hall Software. McGraw-Hill, Englewood Cliffs, 1987.

[140] John R. Rice. *Matrix Computation and Mathematical Software*. McGraw-Hill Computer Science Series. McGraw-Hill, Singapure, International Student edition, 1985.

Sobre o autor

Simão Ferraz de Campos Neto graduou-se em Engenharia Elétrica em 1986 pela Universidade Estadual de Campinas (Unicamp). Trabalha no CPqD/Telebrás desde 1986, começando com o Grupo de Pesquisa Aplicada sobre a Interface U da RDSI. Em 1987, juntou-se ao então recente Grupo para Processamento Digital de Voz, dentro do Programa de Pesquisa Aplicada em Transmissão Digital, trabalhando inicialmente com codificação de voz. Atualmente é responsável pelas atividades de avaliação subjetiva decodificadores de voz no CPqD, como a realização dos testes subjetivos para a língua portuguesa do padrão CCITT a 16 kbit/s, G.728. É o coordenador do "User's Group on Software Tools" da Comissão de Estudos 15 do CCITT (UGST/15) e coordenador do sub-grupo A dentro da CBTT (Com.Bras. de Telegrafia e Telefonia) XII.