Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Métricas de Qualidade para Sinais Áudio-Visuais

Helard Becerra Martinez

Documento apresentado como requisito parcial
para a conlusão do Mestrado em Informática

Orientadora
Prof.ª Dr.ª Mylène Christine Queiroz de Farias

Brasília
2013

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Mestrado em Informática

Coordenador: Prof. Dr. Ricardo Pezzuol Jacobi

Banca examinadora composta por:

Prof.ª Dr.ª Mylène Christine Queiroz de Farias (Orientadora) — UnB
Prof. Dr. Alexandre de Almeida Prado Pohl — UTFPR
Prof. Dr. Bruno Luigi Macchiavello Espinoza — UnB

CIP — Catalogação Internacional na Publicação

Martinez, Helard Becerra.

Métricas de Qualidade para Sinais Áudio-Visuais / Helard Becerra Martinez. Brasília : UnB, 2013.
82 p. : il. ; 29,5 cm.

Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2013.

1. qualidade áudio-visual, 2. multimídia, 3. métrica objetiva de qualidade.

CDU 004

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Métricas de Qualidade para Sinais Áudio-Visuais

Helard Becerra Martinez

Documento apresentado como requisito parcial
para conclusão do Mestrado em Informática

Prof.ª Dr.ª Mylène Christine Queiroz de Farias (Orientadora)
UnB

Prof. Dr. Alexandre de Almeida Prado Pohl
UTFPR

Prof. Dr. Bruno Luigi Macchiavello Espinoza
UnB

Prof. Dr. Ricardo Pezzuol Jacobi
Coordenador do Mestrado em Informática

Brasília, 13 de Dezembro de 2013

Dedicatória

Dedico este trabalho aos meus pais, por serem eles a base fundamental de tudo o que eu sou hoje, por serem eles o meu melhor exemplo de como se tem que amar e apoiar um filho. Aos meus irmãos e família, que com muito carinho e apoio, não mediram esforços e sacrifícios para que eu chegasse até esta etapa da minha vida. A minha orientadora Prof.ª Dr.ª Mylène C. Q. Farias exemplo de disciplina e dedicação, pela enorme paciência que teve comigo, pelas muitas vezes que deixou de ser a orientadora para virar a mãe que cuida do filho, pelo seu apoio e motivação para a culminação do mestrado.

Agradecimentos

Quero Agradecer a todas as pessoas que me apoiaram desde o primeiro momento que tomei a decisão de começar o mestrado, quando cheguei nesse novo país, e no decorrer dos estudos até a culminação do mestrado. Sem o apoio deles nada do que consegui ate hoje houvesse sido possível.

A decisão de fazer estudos fora do país de origem não é simples, por isso quero agradecer a minha família pelo seu apoio incondicional, aos meus pais David e Yolanda e irmãos Danny e Renzo, que sacrificaram a presença do seu filho e irmão somente por ter a satisfação de ver-me superado numa nova etapa da minha vida. Aos meus amigos Lucero e Oscar, a sua amizade foi muito importante nesta etapa da minha vida.

A todos os professores do departamento de Computação, que foram tão importantes na minha vida acadêmica e no desenvolvimento deste trabalho, de maneira especial a minha orientadora Prof. Mylène C. Q. Farias, pelas aulas, pelas sugestões, pelos conselhos e dicas de pesquisa, pelo material emprestado, pela paciência que teve comigo, pela participação e pela ajuda incondicional, quem com seus conhecimentos e experiência soube me encaminhar no mestrado. Torço pra que esta parceria continue por muitos anos. Ao professor Alexandre de Almeida Prado Pohl e o professor Bruno Luigi Macchiavello Espinoza que são parte da minha banca, pela sua presença, suas sugestões e contribuições para com meu trabalho.

Agradeço também aos meus colegas do mestrado com quem compartilhei muito durante meus estudos, de maneira especial a todos que me ajudaram nos momentos de duvidas e dificuldades, muito obrigado Ruben Cruz, Harley Vera, Toni Serrano, Stephanie Alvarez, Henrique Freitas, Yang Liu, Ariane Alvez, Amanda Cristina, Lucas Araujo, Jonathan Alis, Paula Leticia e Gabriela Quirino. Agradeço também aos meus colegas e amigos do Grupo de Processamento Digital de Sinais (GPDS). Aos amigos e colegas voluntários que participaram dos experimentos, obrigado pelo apoio.

Aos meus amigos peruanos, com quem foi mais fácil matar as saudades de nossas famílias e nosso país, pela companhia nos bons e maus momentos, em especial aos meus amigos Harley Vera, Ruben Cruz, Maria Laura Chavez, Jose Luis Soncco e Juan Cruz.

Resumo

Nesta dissertação é estudada a avaliação da qualidade em sinais áudio-visuais, especificamente a avaliação subjetiva e objetiva de sinais em alta definição. O procedimento mais preciso para a aferição da qualidade é a avaliação utilizando experimentos psicofísicos (subjetivos) com observadores humanos. Esta metodologia demanda um longo período de tempo e um elevado custo operacional. Uma alternativa consiste em utilizar métricas objetivas para obter uma estimação da qualidade do sinal multimídia. As métricas objetivas podem ser classificadas segundo a quantidade de informação necessária para estimar a qualidade do sinal: (1) Referência Completa (*Full Reference (FR)*), (2) Referência Reduzida (*Reduce Reference (RR)*), e (3) Sem Referência (*No-Reference (NR)*). O objetivo desta dissertação é propor modelos de qualidade objetivos e subjetivos para sinais multimídia (áudio e vídeo), ou seja modelos áudio-visuais. Os modelos subjetivos são baseados nos dados experimentais, enquanto que os modelos objetivos (métricas FR e NR) são obtidos através de uma função de combinação de métricas de áudio e vídeo. São descritos os três experimentos psico-físicos realizados com o fim de estudar a relação entre as componentes de áudio e vídeo. Com o objetivo de obter um modelo áudio-visual objetivo, duas métricas sem-referência (NR) de áudio e vídeo foram propostas. Ao todo, são propostos neste trabalho três modelos subjetivos, três modelos objetivos com-referência (FR) e três modelos objetivos sem-referência (NR). Os resultados apresentados mostram que os modelos conseguem estimar os valores de qualidade áudio-visual de forma aceitável. O desempenho destes modelos foi comparado com o desempenho de propostas existentes na literatura.

Palavras-chave: qualidade áudio-visual, multimídia, métrica objetiva de qualidade.

Abstract

In this work, we studied audio-visual quality assessment models, focusing on the subjective and objective quality assessment of high definition signals. The most accurate method to determine the quality of a video is by using psychophysical experiments with human subjects (subjective metrics). However, these kinds of methods are expensive and time-consuming. Objective metrics represent a good alternative for measuring video quality. They can be classified according to the amount of reference (original) information used to stimate the signal quality: Full Reference (FR), Reduced Reference (RR), and No-Reference (NR) metrics. The main objective of this dissertation is to propose subjective and objective quality models to calculate the quality of multimedia (audio and video) signals, i.e. audio-visual signals. Subjective models are designed by collecting data from psychophysical experiments; meanwhile, objective models (FR and NR metrics) use quality measures of audio and video (from an audiovisual signal) and combine them into a single measure. Three psychophysical experiments were performed, with the goal of studing the relation between the audio and video components of an audio-visual signal. Moreover, in order to model the audio-visual quality metrics, two quality metrics (NR) for audio and video were proposed. In sumary, we introduce three subjective models, three FR objective models, and three NR objective models to estimate the audio-visual quality of a signal. The performance of these models was compared with the performance of other metrics available in the literature.

Keywords: quality assessment, audio-visual quality, qoe, compression.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

Capítulo 1

Introdução

1.1 Contextualização

O grande progresso alcançado pelas aplicações multimídia está evidenciado no aumento do número de serviços e produtos oferecidos atualmente. Uma aplicação multimídia é definida como uma aplicação que pode combinar várias componentes, tais como texto, gráficos, áudio, vídeo, etc. Posteriormente, este conteúdo poderá ser digitalmente transmitido através de um canal de comunicação e, mais frequentemente, pela Internet [15]. Entre os exemplos de aplicações multimídia que são transmitidas na Internet (IP-based), podemos citar o vídeo sob demanda (e.g., *Netflix*), a televisão IP (*Internet Protocol Television (IPTV)*), a televisão móvel (*Mobile Television (MoTV)*) e o vídeo web (e.g., *YouTube, Facebook, Google*, etc.). A qualidade destas aplicações não depende apenas de quais destas componentes (texto, gráficos, áudio, vídeo) estão incluídas no pacote, mas principalmente das características intrínsecas destas componentes e como elas contribuem para sua qualidade.

O grande aumento nos serviços multimídia oferecidos acarreta um maior interesse e demanda de ferramentas para avaliar, controlar e melhorar a qualidade do material multimídia entregue aos usuários. A necessidade de ferramentas e métodos de aferição é ainda maior, dado o fato que os consumidores atuais têm se tornado mais exigentes e conhecedores das capacidades dos serviços multimídia. Logo, é importante o desenvolvimento de métricas que ajudem a garantir que estas aplicações proporcionem um serviço de alta qualidade e que englobem noções da percepção humana (tanto visual como auditiva) para assim satisfazer as expectativas dos usuários. Da mesma forma, a implementação de métricas de qualidade em tempo real é de grande utilidade para os provedores do serviço. Estas métricas podem fornecer informação relacionada à qualidade do sinal transmitido, permitindo aos operadores da rede controlar seus recursos em tempo real, mantendo desta forma a satisfação do usuário. O desempenho destas aplicações é tradicionalmente medido em termos de qualidade de serviço (*Quality of Service (QoS)*), que levam em conta características técnicas da rede como por exemplo, taxas de perda de pacotes e transmissão de bits, *jitter* etc. Nos últimos, anos esta abordagem está sendo gradativamente substituída por medidas de qualidade de experiência (*Quality of Experience (QoE)*). As medidas de QoE tentam estimar a qualidade dos sinais de acordo com a qualidade percebida pelo usuário, ou seja levando em conta aspectos do Sistema Visual Humano (SVH) e do Sistema Auditivo Humano (SAH), assim como aspectos do comportamento humano [16].

O procedimento mais preciso para a aferição da qualidade multimídia é a avaliação utilizando experimentos psico-físicos (subjetivos) com observadores humanos. Infelizmente, estes experimentos são caros, exigindo tempo e recursos físicos. Além disso, estas técnicas não podem ser utilizadas para monitorar uma aplicação em tempo real. Uma alternativa consiste em utilizar métodos computacionais, ou seja métricas objetivas para obter uma estimação da qualidade do sinal multimídia. Nos últimos anos, várias métricas de qualidade de áudio e vídeo têm sido propostas [4] [17] [18]. Entre estas, as métricas objetivas de qualidade com melhores resultados [7] [19] são aquelas que utilizam um sinal de referência (sinal 'original' ou o sinal livre de degradações) para a aferição da qualidade. Essas métricas são chamadas de métricas com-referência ou referenciadas (*Full reference* – FR). Infelizmente, num ambiente típico de comunicações, tal referência não está disponível do lado do usuário ou do receptor, o que torna difícil a sua utilização para o monitoramento da qualidade em tempo real. Portanto, o desenvolvimento de métricas automáticas, que não utilizem o sinal de referência ou utilizem apenas alguns parâmetros sobre este sinal, continua sendo um problema de pesquisa em aberto [20]. Estas métricas são chamadas de métricas com referência reduzida (*Reduce Reference* – RR) e métricas sem-referência ou não-referenciadas (*No-Reference* – NR).

Entres as técnicas mais importantes da literatura para a aferição da qualidade de vídeo, podemos citar o modelo de Sarnoff [21], o algoritmo proposto por Wang [7] e a métrica de qualidade de vídeo da *National Telecommunications and Information Administration (NTIA)* [19]. Todas estas propostas seguem uma abordagem FR para o cálculo da qualidade do sinal de vídeo. Entre as métricas propostas que utilizam a abordagem RR, estão o modelo proposto pela Universidade Yonsei [8] e o algoritmo de Força Harmônica Local (FHL) [9]. Entre as métricas do tipo NR, podemos citar as métricas de Caviedes [22], a métrica de Farias [23] e o algoritmo BLIINDS-II proposto por Bovik-Saad [10]. Em suma, existem um bom número de métricas (FR, RR, e NR) que incluem diversas propiedades do sistema visual humano (SVH) nos seus algoritmos de aferição da qualidade. Não obstante, a maioria dos avanços foram feitos na área de FR e há muito por fazer na área de RR e NR [20]. Novas tendências na área incluem o desenvolvimento de métricas híbridas (métricas que incluem informação da rede e aspectos do SVH), métricas para sinais de vídeo 3D e métricas para sinais multimídia. Este último tópico é a ênfase desta dissertação, onde focamos a nossa atenção no desenvolvimento de uma métrica áudio-visual para sinais de vídeo.

A quantidade de modelos de qualidade de áudio propostos é menor. Entre os mais representativos temos: (1) Métrica FR para a aferição da qualidade perceptiva de áudio (PEAQ) [24] e (2) Métrica NR para a aferição da qualidade de voz (SESQA) [11], com foco na qualidade de voz em transmissões telefônicas. Ambas as propostas foram adotados como padrões da *International Telecommunications Union (ITU)*. Atualmente, não existe uma métrica de áudio NR genérica. Mas, dois projetos estão sendo desenvolvidos pela ITU: *Parametric non-intrusive assessment of audiovisual media streaming quality (P.NAMS)* e *Opinion Model for Video Streaming applications (G.OMVAS)*. Espera-se que, uma vez concluídos estes projetos, novas métricas para a aferição do áudio sejam propostas.

Como pode ser observado, um grande progresso foi feito no desenvolvimento de métricas de áudio e vídeo (isoladamente) [4] [17] [18] [25]. Isto poderia nos levar a pensar que para estimar a qualidade áudio-visual de um sinal de vídeo bastaria 'somar' as estimações

obtidas para as qualidades do áudio e do vídeo. Entretanto, estudos [26] mostram que há uma interação entre os valores das qualidades de áudio e vídeo. Ou seja, o valor da qualidade multimídia ou da qualidade geral do sinal não é simplesmente a 'soma' da qualidade das componentes de áudio e vídeo. Na literatura, existem alguns modelos (*métricas subjetivas*) de qualidade áudio-visual [26] [27]. Mais especificamente, para avaliação da qualidade áudio-visual destacam-se o modelo proposto por Hands [26], o modelo proposto por Winkler [28] e o modelo proposto por Garcia [27]. Estes três modelos, entretanto, são baseados no cálculo *subjetivo* da qualidade multimídia, isto é a utilização de medidas obtidas em experimentos psico-físicos. Um quarto modelo importante é a métrica objetiva adotada pela ITU para aplicações em vídeo-telefonia [29], que calcula a qualidade do sinal em termos de parâmetros de QoS, tais como perda de pacotes, bitrate e taxa de quadros.

Em suma, a grande maioria das métricas de qualidade objetivas propostas não consideram a componente de áudio, não podendo ser utilizadas para aplicações multimídia. Além disso, a maioria das métricas propostas é com referência (FR), não podendo ser utilizadas para aplicações reais. Sendo assim, até o nosso conhecimento não existe um modelo computacional para a aferição da qualidade áudio-visual de sinais multimídia ou até mesmo áudio-visual (áudio+vídeo). A maioria dos modelos na literatura utiliza dados subjetivos (valores coletados experimentalmente) como entrada para calcular os valores de qualidade do sinal. Acredita-se que o desenvolvimento de uma métrica objetiva para aferir a qualidade áudio-visual representará uma contribuição importante na área do processamento multimídia.

O objetivo principal desta dissertação do mestrado é propor uma métrica de qualidade (FR e NR) para aferir a qualidade áudio-visual em aplicações multimídia. Com este objetivo, busca-se integrar métricas de qualidade de áudio e vídeo já existentes para desenvolver uma métrica áudio-visual objetiva. É muito importante que a métrica utilize aspectos dos sistemas visual e auditivo humano para o cálculo da qualidade, garantindo resultados que possuam boa correlação com os valores obtidos em testes subjetivos.

1.2 Objetivos

Este trabalho tem como objetivo geral propor modelos de qualidade objetivos e subjetivos para sinais áudio-visuais. Os modelos subjetivos são baseados nos dados experimentais, enquanto que os modelos objetivos (métricas FR e NR) são obtidos através de uma função combinação de métricas de áudio e vídeo.

No intuito de atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

- Realizar experimentos subjetivos para estudar a relação entre as qualidades das componentes de vídeo e áudio das sequências e a relação destas qualidades com a qualidade áudio-visual;
- Selecionar métricas (FR e NR) de áudio e vídeo de bom desempenho para estimar a qualidade dos sinais de áudio e vídeo;
- Propor um modelo subjetivo para a qualidade áudio-visual;
- Propor um modelo objetivo FR para a qualidade áudio-visual;
- Propor um modelo objetivo NR para a qualidade áudio-visual;

- Testar os modelos propostos e compará-los com os resultados de modelos já existentes na literatura.

1.3 Metodologia Utilizada

A metodologia utilizada para a obtenção dos modelos de qualidade áudio-visuais inclui diferentes tarefas. As principais tarefas realizadas nesta dissertação foram:

- Escolha das sequências áudio-visuais: Um conjunto de sequências é escolhido para as avaliações subjetivas e objetivas. A escolha foi feita de acordo com as recomendações dos relatórios do *Video Quality Experts Group (VQEG)*.
- Processamento das sequências: As sequências são processadas e degradadas para os experimentos e testes. São utilizados os padrões H.264 para compressão de vídeo e MPEG-1 para compressão de áudio.
- Experimentos Subjetivos: São realizados 3 experimentos psico-físicos (subjetivos) diferentes, nos quais são coletadas as notas sobre a qualidade dos sinais. O primeiro experimento contém apenas sinais de vídeo, o segundo apenas sinais de áudio e o terceiro sinais com componentes de áudio e vídeo.
- Avaliação Estatística e Subjetiva: Nesta fase são avaliados estatisticamente os dados dos experimentos e propostos modelos subjetivos para a qualidade dos sinais áudio-visuais.
- Avaliação Objetiva: São desenvolvidos modelos objetivos de qualidade (FR e NR) para os sinais áudio-visuais, utilizando métricas de áudio e vídeo.
- Validação das métricas: Para os testes dos modelos, são utilizadas as sequências de teste e dados subjetivos dos três experimentos. O modelos são comparados com modelos já propostos na literatura.

Na Figura 1.1 é apresentado o diagrama simplificado deste processo e as tarefas feitas nesta dissertação.

As principais contribuições desta dissertação são:

1. Proposta de modelo subjetivo para qualidade áudio-visual;
2. Proposta de métrica de qualidade áudio-visual com-referência;
3. Proposta de métrica de qualidade áudio-visual sem-referência;
4. Disponibilização de um banco inédito de sequências multimídia (áudio, vídeo, áudio+vídeo), degradadas por algoritmos de compresssão, com suas correspondentes medidas subjetivas de qualidade.

1.4 Estrutura do Trabalho

A estrutura desta dissertação está dividida da seguinte forma. No Capítulo 2 são descritos conceitos introdutórios sobre o sistema visual, o sistema auditivo humano, os

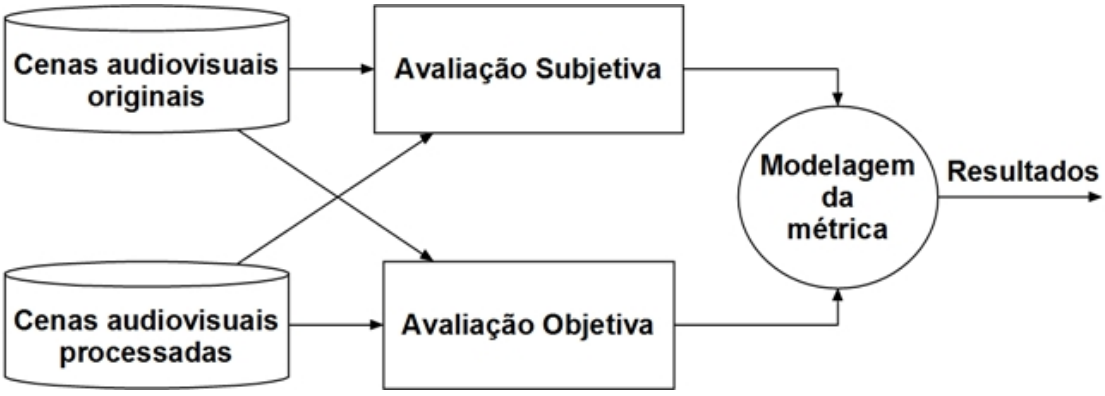


Figura 1.1: Fases da metodologia utilizada

sistemas digitais de áudio/vídeo e a sua relação com a percepção da qualidade áudio-visual. No Capítulo 3, são descritas as formas de avaliação da qualidade dos sinais de vídeo, áudio e áudio-visual, os métodos subjetivos e as diversas abordagens objetivas existentes na literatura.

O Capítulo 4 descreve os experimentos subjetivos realizados neste trabalho, incluindo a geração das sequências de teste, a metodologia experimental, a análise estatística e os resultados obtidos. O Capítulo 5 detalha os modelos objetivos de qualidade áudio-visual propostos neste trabalho. Finalmente, no Capítulo 6 são apresentadas as conclusões e propostas de trabalhos futuros.

Capítulo 2

Conceitos Básicos

Neste capitulo são apresentados ao leitor alguns dos conceitos básicos utilizados neste trabalho, com o fim de familiarizá-lo com o tópico de pesquisa. São abordados conceitos relacionados ao funcionamento dos sistemas visual e auditivo humanos. Descrevemos também características importantes dos sistemas digitais de áudio e vídeo.

2.1 Sistema Visual Humano (SVH)

A percepção visual é muito importante para o ser humano, que recebe informação de maneira constante e deve processá-la para poder interagir com o ambiente que o rodeia. Logo, muitas das métricas de qualidade de vídeo levam em consideração aspectos do sistema de visão humano e noções psico-físicas, com o objetivo de melhorar a sua correlação com os testes subjetivos [19]. Nesta seção, apresentamos alguns dos conceitos básicos do funcionamento do olho humano e as características psico-físicas de percepção do sistema visual humano (SVH).

2.1.1 Mecanismo da Visão

O mecanismo da visão pode ser mais bem entendido se compararmos o globo ocular a uma câmara fotográfica: o cristalino seria a lente; a íris, a abertura da lente; e a retina seria a placa ou película. Desta maneira, os raios luminosos chegam ao cristalino (lente) após penetrarem na córnea e na câmera anterior (humor aquoso) e passarem pela pupila.O cristalino leva a imagem mais para trás ou para frente, permitindo que ela se projete sobre a retina.

O olho humano é aproximadamente esférico, com diâmetro em torno de 2,5 cm e peso de 7 gramas. O olho pode ser considerado um dispositivo que captura a luz e a focaliza em uma superfície de fundo (retina). A Figura 2.1 representa um corte transversal do olho humano destacando seus componentes principais. O que o olho percebe em uma cena é determinado pelos raios de luz emitidos ou refletidos a partir da cena. Quando estes raios são suficientemente fortes e estão no intervalo de luz visível, ou seja, dentro do espectro eletromagnético entre 380 – 780nm, o olho reagirá enviando um sinal elétrico ao cérebro através do nervo ótico [30].

O raio de luz atravessa primeiro a córnea, que é uma capa transparente protetora que atua como uma lente que refrata a luz. Logo, a luz passa através do humor aquoso

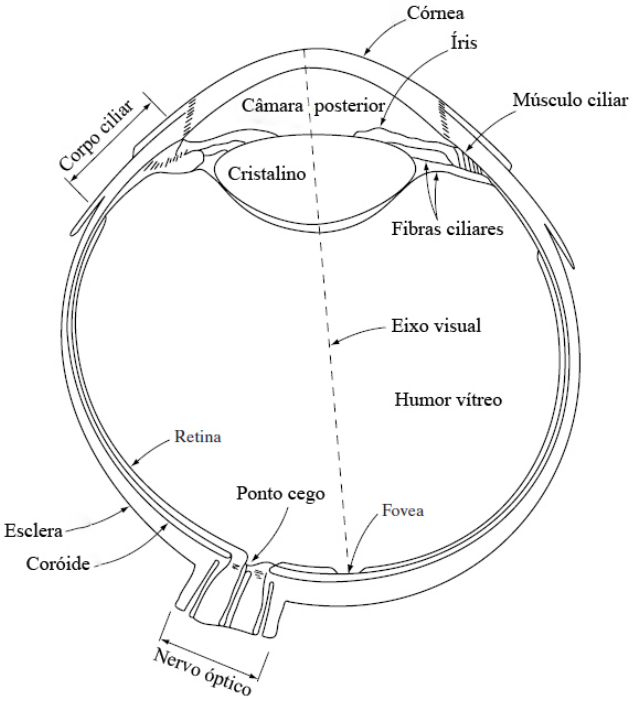


Figura 2.1: Diagrama simplificado de um corte transversal do olho humano [1].

(lente) e da íris. Esta última forma uma abertura redonda que pode variar de tamanho para determinar a quantidade de luz que pode atravéssá-la. Em seguida, o raio de luz atravéssa a lente e o humor vítreo para chegar finalmente à retina. O cristalino varia sua forma para ajudar a focar a imagem na retina.

Na retina, os raios de luz são detectados e convertidos em sinais elétricos por fotorreceptores: os bastonetes e os cones. Existem aproximadamente cem milhões de bastonetes no olho humano. Com exceção da região da fóvea (área central da retina), os bastonetes estão espalhados uniformemente na retina. A fóvea é a parte da retina onde a visão humana atinge sua mais alta resolução espacial. Esta área possui uma maior densidade de cones, com aproximadamente seis ou sete milhões localizados em torno da fóvea. Existe um ponto na retina onde não existem fotorreceptores, que é onde o nervo ótico está conectado ao olho. Esta região é denominada ponto cego, uma vez que não adquirimos nenhuma informação dessa região. Na Figura 2.2 é apresentada a distribuição espacial dos cones e bastonetes na retina, destacando-se a quantidade de fotorreceptores em cada posição (grau do eixo visual).

Finalmente, estes sinais elétricos são conduzidos pelo nervo ótico para o centro da visão, no cérebro, onde são decodificados e interpretados.

2.1.2 Percepção de Cores

A percepção de cores está relacionada com a sensibilidade dos fotorreceptores (cones e bastonetes) ao nível de luminosidade. Os bastonetes são mais sensíveis à luz que os cones e, numa situação de penumbra, só os bastonetes estão ativos [31]. Além disso, os bastonetes não conseguem distinguir cores. Logo, em uma situação de baixa luminosidade, só é possível perceber tons de cinza. Este tipo de visão é chamada de escotópica ou visão

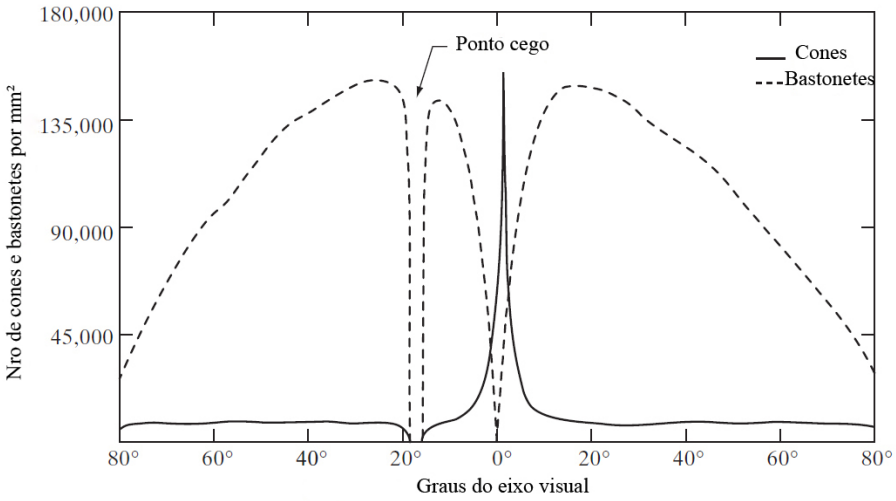


Figura 2.2: Distribuição dos bastonetes e cones na retina [1].

noturna. Por outra lado, em circunstâncias de grande luminosidade, os cones estão mais ativos e experimentamos uma visão fotótica [31]. Em circunstâncias de luminosidade intermédia, os cones e bastonetes estão ambos ativos por igual. Esta visão é denominada visão mesópica.

A existência de três tipos de cones, cada um deles sensível a uma diferente banda no espectro eletromagnético, permite ao ser humano a percepção de cores. Este modelo é denominado Teoria Tri-Cromática. Estes três tipos de cones são denominados: (1) Curtos (*Short* – S), comprimento de onda de 440 - 485 nm, (2) Médios (*Medium* – M, comprimento de onda de 500 - 565 nm e (3) Longos (*Long* – L), comprimento de onda de 625 - 740 nm. A sensibilidade espectral relativa dos cones (representada como uma função do comprimento de onda) é apresentada na Figura 2.3. Pode-se observar que um tipo de cone está ativo quase exclusivamente dentro do espectro eletromagnético entre 400 - 500 nm (cores violeta e azul) e os outros dois tipos de cones permanecem ativos na região entre 500 - 600 (ciano - vermelho) [1].

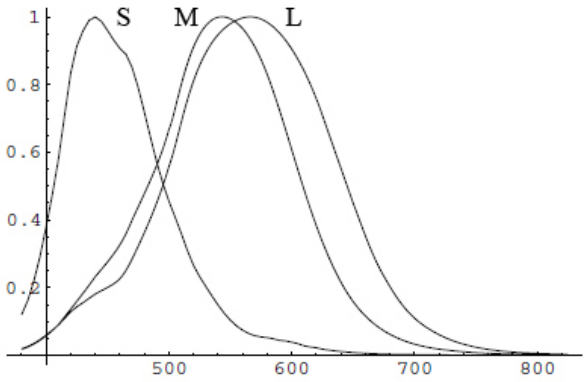


Figura 2.3: Sensibilidade espectral dos cones: S (*short*), M (*medium*) e L (*long*) [2].

2.1.3 Sensibilidade ao Contraste

A habilidade do ser humano perceber detalhes em objetos ou em uma cena visual é determinada pela capacidade do seu sistema visual em distinguir contraste, isto é, a diferença de brilhos de áreas adjacentes. Por outro lado, tem-se demostrado que a sensibilidade ao contraste está relacionada com atividades que implicam em uma discriminação visual, tais como o reconhecimento da faces.

A sensibilidade ao contraste é representada pela Função de Sensibilidade ao Contraste (FSC) [3]. Esta função é obtida através de experimentos subjetivos com indivíduos. O valor do contraste é definido em função da luminosidade máxima e mínima, conforme a equação:

$$\mathrm{C} = \frac{L_{max} - L_{min}}{L_{max} + L_{min}}. \tag{2.1}$$

A curva da sensibilidade ao contraste versus a frequência espacial (obtida experimentalmente) é apresentada na Figura 2.4 [3].

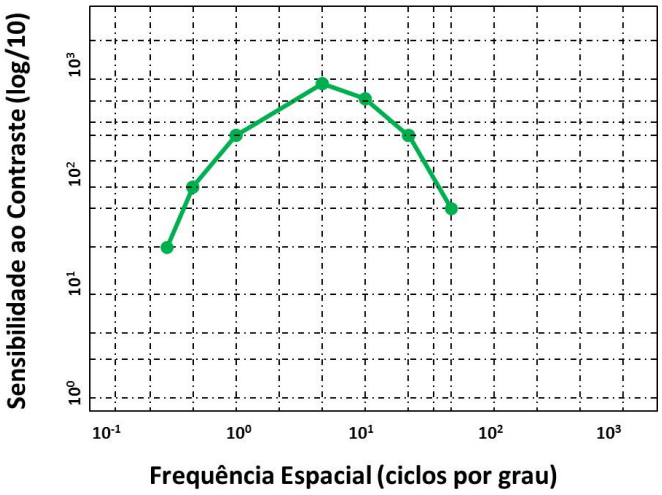


Figura 2.4: Curva da sensibilidade ao contraste FSC (Adultos) [3].

2.1.4 Outras Características Perceptuais

Algumas outras características do funcionamento do olho humano, que são levadas em conta no desenvolvimento de métricas de qualidade [32] [33], são descritas a seguir.

- Visão foveal e periférica – A densidade nos fotorreceptores não é uniforme (ver Figura 2.2), temos valores altos na parte central da retina (fóvea) e valores baixos no sentido contrario. Como consequência, a resolução dos objetos focados na parte central da retina é maior que dos objetos na área periférica.
- Adaptação à intensidade de luz – A intensidade de luz varia muito e o SVH se adapta a esta variação ajustando a quantidade de luz que entra ao olho. Isto é feito mediante adaptações no diâmetro das pupilas e variações do ganho dos fotorreceptores na retina. Como resultado, em lugar de codificar intensidades absolutas de luz, a retina

codifica o contraste do estímulo visual. O fenômeno que mantém a sensibilidade do contraste em um intervalo amplo de intensidade de luz é conhecido como a regra de Weber:

$$\Delta I/I = K, \tag{2.2}$$

no qual I é a luminosidade do fundo, ΔI é a variação da luminosidade no fundo e K é uma constante chamada de fração de Weber.

- Mascaramento e facilitação – São aspectos importantes no SVH para a modelagem das interações entre componentes diferentes de uma imagem presentes em um mesma posição espacial. Mais especificamente, refere-se ao fato de que a presença de um componente da imagem (máscara) reduz a capacidade de visão de outro componente (sinal de teste). Para alguns casos, a máscara pode facilitar a capacidade de percepção do sinal de teste. Muitas das métricas de qualidade incorporam modelos para mascaramento e/ou facilitação.

2.2 Sistema Auditivo Humano (SAH)

2.2.1 Mecanismo da Audição

A audição é um fenômeno tanto mecânico (de propagação de onda) como sensorial e perceptivo. Em outras palavras, quando uma pessoa percebe um som, este sinal chega ao seu ouvido através do ar como uma onda mecânica. No interior do ouvido, esta onda se transforma em impulsos nervosos e viaja para o cérebro. Por isso, em muitos dos problemas de acústica, como o processamento de áudio, é vantajoso levar em conta, não apenas a mecânica do meio ambiente, mas também o fato de que tanto o ouvido como o cérebro estão envolvidos na experiência de audição de uma pessoa [34].

Na Figura 2.5 é apresentada uma ilustração da anatomia do ouvido humano. O ouvido humano pode ser separado em três grandes segmentos, de acordo com a função desempenhada e a localização: o ouvido externo, o ouvido médio e o ouvido interno. A seguir descrevemos cada um desses segmentos.

- Ouvido externo – Fazem parte do ouvido externo o pavilhão auricular e o canal auditivo, cujas funções são recolher e encaminhar as ondas sonoras até o tímpano.
- Ouvido médio – Também denominado caixa timpânica, é uma cavidade com ar localizada por detrás da membrana do tímpano. É através do ouvido médio que a energia das ondas sonoras é transmitida, do ouvido externo à janela oval na cóclea (localizada no ouvido interno). Essa transmissão de energia é efetuada através de três ossos minúsculos: o martelo, a bigorna e o estribo, que vibram com o tímpano.
- Ouvido interno – No ouvido interno, a cóclea é responsável pela nossa capacidade de diferenciar e interpretar sons. De fato, desenrola-se na cóclea uma função complexa de conversão de sinais. Como resultado, os sons mecânicos recebidos na cóclea são transformados em impulsos elétricos que chegam até o cérebro pelo nervo auditivo, onde são decodificados e interpretados.

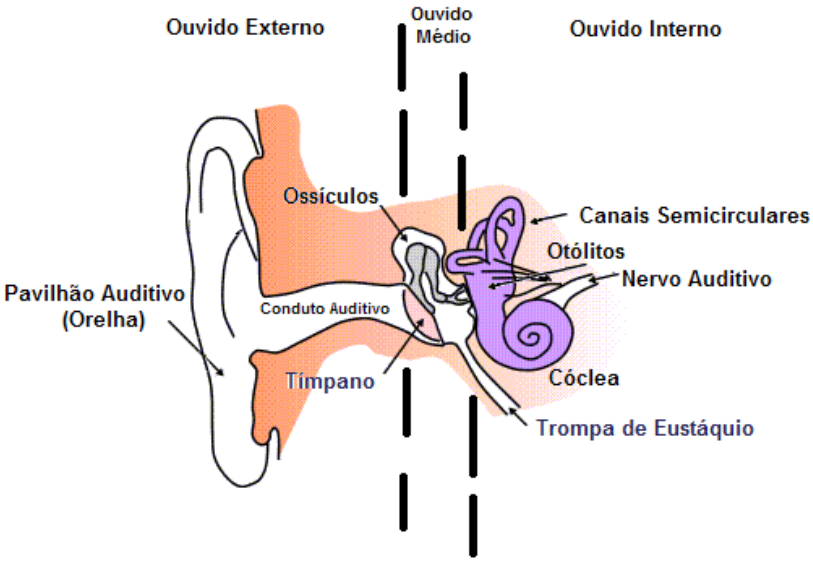


Figura 2.5: Anatomia do ouvido humano (*Wikimedia Commons, 2005*).

2.2.2 Características Psico-Acústicas

A relação entre sensações auditivas e as características físicas do som é estudada pela psico-acústica. Estas sensações são determinadas principalmente por duas características: a frequência e amplitude. Na qualidade de áudio, o modelamento de características psico-acústicas do sistema auditivo humano é muito importante. Muitas destas características podem ser utilizadas como bases para calcular a qualidade de áudio. Consideram-se as seguintes sensações entre as mais importantes:

- Sonoridade – Sensação subjetiva de volume sonoro, determinada pela pressão sonora e frequência do sinal.
- Altura – Refere-se à forma como o ouvido humano percebe a frequência fundamental dos sons, isto quer dizer, baixas frequências são percebidas como sons graves e as mais altas como sons agudos.
- Timbre – É a característica sonora que permite distinguir se sons da mesma frequência foram produzidos por fontes sonoras diferentes. Em outras palavras, é a capacidade de diferenciar uma mesma nota musical produzida por diferentes instrumentos.
- Mascaramento – O efeito de mascaramento descreve o comportamento do ouvido quando dois ou mais sons diferentes o estimulam simultaneamente ou num curto intervalo de tempo. Ele consiste no "apagamento" parcial ou total de algumas componentes do sinal de áudio devido à proeminência de outras componentes.

Há divergências sobre a independência destas sensações. Acredita-se que estas sensações guardam uma forte inter-dependência .

2.3 Sistemas Digitais de Vídeo

A seguir, são apresentados alguns conceitos sobre compressão de vídeo e os principais métodos de compressão utilizados. Estudam-se também alguns dos artefatos mais comumente encontrados.

2.3.1 Compressão do Vídeo Digital

A compressão de dados é uma importante tarefa no processamento de sinais digitais de vídeo. Os objetivos principais são reduzir o espaço de armazenamento necessário dos dados e facilitar a transmissão deles em um canal de comunicação estabelecido. Para atingir estes objetivos, muitas estratégias são utilizadas para comprimir a informação sem afetar negativamente a qualidade dos dados processados. Existem dois tipos de compressão: com perda (*lossy*) e sem-perda (*lossless*)[35]. Os algoritmos de compressão sem perda oferecem uma reconstrução perfeita do sinal original. Infelizmente, a taxa de compressão neste tipo de algoritmo é muito baixa, principalmente para sinais de vídeo. Uma compressão com perda oferece uma taxa muito mais alta e mais útil para sinais de vídeo, mas existe uma perda na qualidade do sinal comprimido.

Os algoritmos de compressão de vídeo procuram eliminar a redundância na informação do sinal para atingir uma taxa de compressão alta. O objetivo de um codificador de vídeo é diminuir a informação em um sinal de vídeo, removendo a informação redundante. São levados em conta, principalmente, quatro tipos de redundância:

- Redundância perceptiva – É definida como o tipo de informação que pode ser eliminada sem que a perda seja visivelmente perceptível. Por exemplo, considerando que existe um maior número de bastonetes do que de cones, o SVH é mais sensível às informações de brilho do que às informações de cor. Desta forma, não há necessidade de utilizar a mesma quantidade de bits para representar as componentes de luminância e de cor. Pode-se modificar a relação dos pixels das componentes de cores (do formato de cores YCbCr, por exemplo) de 4:4:4 para 4:2:2 ou 4:2:0. Esta alteração é chamada de sub-amostragem (*downsampling*). Utilizando uma relação de 4:2:0, para cada 4 pixels de luminância (Y) temos 2 pixels de cor (CbCr), o que implica em uma economia de 50% dos pixels. Neste caso, os dados de cor que foram removidos são considerados irrelevantes.

- Redundância espacial – Este tipo de redundância, conhecida também como redundância intra-quadro ou interpixel [36], resulta da correlação entre pixels espacialmente distribuídos em um quadro. É possível observar esta correlação tanto no domínio espacial, quanto no domínio de frequências. No domínio espacial, esta correlação é percebida observando pixels vizinhos em um quadro. A tendência é achar valores semelhantes em superfícies grandes do quadro. Esta redundância pode ser reduzida utilizando a codificação intra-quadro, utilizada em alguns dos padrões de codificação na atualidade. A imagem deve ser transformada do domínio espacial para o domínio de frequências. A transformação para o domínio de frequências tem como objetivos a remoção entre as amostras e a agrupação da maior quantidade de energia em poucas amostras. A quantização consiste em uma divisão inteira dos coeficientes gerados pela transformação de domínio, ela reduz um grande número

de coeficientes à zero. Dado que os resíduos das divisões não são armazenados, esta operação produz perdas na informação, o que a torna irreversível.

- Redundância temporal – É conhecida também como redundância inter-quadros [36], e é causada pela correlação existente entre quadros sucessivos em uma sequência de vídeo. Muitos dos quadros consecutivos em uma sequência de vídeo tendem a ser muito similares (fundo da imagem, objetos estáticos, etc.), mantendo pixels com valores iguais ou próximos. Outros pixels apresentam uma variação leve de valores causada por uma deslocação de um quadro para outro (movimento de um objeto em uma cena). A redução da redundância temporal está presente na grande maioria dos padrões de compressão atuais. Uma adequada redução deste tipo de redundância resulta em altas taxas de compressão.

- Redundância Estatística – Também chamada de redundância entrópica, está relacionada com as probabilidades de ocorrência dos símbolos codificados. A entropia é uma medida da quantidade média de informação transmitida por um símbolo de vídeo [37]. A quantidade de informação nova transmitida por um símbolo diminui à medida em que a probabilidade de ocorrência deste símbolo aumenta. As abordagens que exploram este tipo de redundância visam transmitir a maior quantidade de informação possível por símbolo codificado e, deste modo, representar mais informação com menos bits. Para reduzir este tipo de redundância podem ser utilizadas algumas abordagens de codificação sem-perda, como a codificação de Huffman ou a codificação aritmética [1].

Em cada estágio de um algoritmo de compressão, procura-se reduzir um tipo de redundância específica. A Figura 2.6 mostra os componentes funcionais de um algoritmo de compressão genérico.

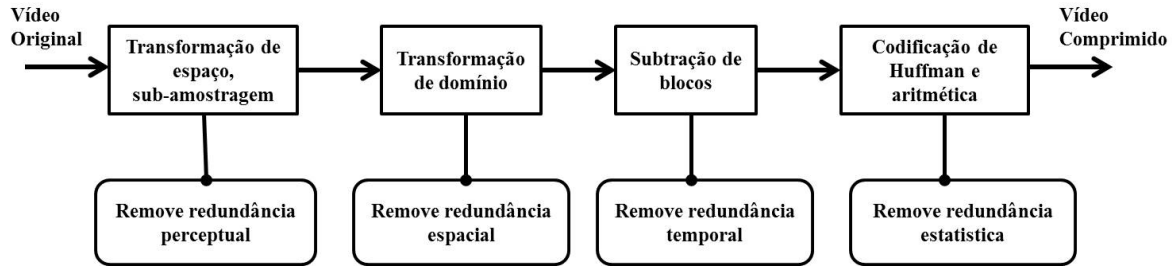


Figura 2.6: Componentes funcionais de um algoritmo de compressão de vídeo (Adaptação [4]).

Os padrões de compressão mais populares são desenvolvidos principalmente por dois organismos internacionais: *International Telecommunications Union* (ITU) e *International Standards Organization (ISO)*. Esta última através dos grupos: *Joint Photographic Experts Group (JPEG)* e *Moving Pictures Experts Group (MPEG)*.

Entre os padrões produzidos pelo MPEG está o MPEG-1, o primeiro método de compressão com perda proposto pelo MPEG. O MPEG-1 ainda é utilizado para compressão de vídeos em CD-ROM. O padrão MPEG-2 é um esquema muito popular utilizado para

Tabela 2.1: Padrões de compressão de vídeo [13].

Padrão	Aplicação	Taxa de bits
H.261	Vídeo-telefonia e teleconferência	64 Kb/s
MPEG-1	Vídeo em CD-ROM	1.5 Mb/s
MPEG-2	Televisão digital, SD-DVD	$>$ 2 Mb/s
H.263	Vídeo-telefonia	$<$ 33.6 Kb/s
MPEG-4	Codificação baseada em objetos	Variável
H.264	HD-DVD, vídeo-segurança, vídeo-conferência	Variável
H.265	Vídeo-conferência, armazenamento digital e *Internet Streaming*	Variável

transmissão e para compressão em DVDs. A sua principal vantagem é o baixo custo. Finalmente, um outro padrão popular é o MPEG-4, que é utilizado em aplicações de vídeo segurança dada sua facilidade para trabalhar com resoluções e taxas de bits diferentes [13].

Entre os padrões desenvolvidos pela ITU tem-se o H.261 e o H.263. Um trabalho em conjunto das duas organizações, ITU e ISO, teve como resultado o padrão H.264, conhecido também como MPEG-4 *Advanced Video Coding (AVC)*. Este padrão representa o maior avanço, até agora, na tecnologia de compressão para vídeos. Em comparação ao MPEG-2, o H.264 consegue comprimir com aproximadamente metade da taxa de bits a um custo de qualidade idêntico [38]. A Tabela 2.1 mostra uma comparação entre os diferentes padrões de compressão de vídeo disponíveis, destacando as aplicações e as taxas de bits para cada um deles.

2.3.2 Artefatos Comuns

Os artefatos são definidos como características indesejáveis em uma sequência de vídeo, resultantes de erros visíveis aos observadores/usuários. Eles podem ser introduzidos no processo de produção, nas fases de captura, compressão, transmissão, recepção e entrega para o usuário final (Figura 2.7). A presença de artefatos é indesejável, uma vez que pode afetar o valor da qualidade percebida pelo usuário [22].

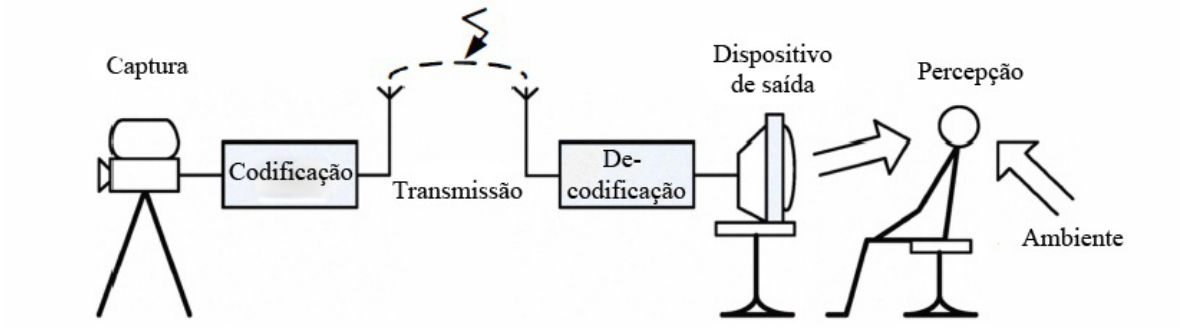


Figura 2.7: Diagrama ilustrando os diversos estágios onde artefatos podem ser adicionados ao sinal [5].

Alguns dos artefatos visuais que encontramos nos sinais de vídeos digitais são:

- Borrado -– O artefato borrado é definido como a perda de energia e detalhe espacial, como consequência da redução de nitidez nas bordas. Ele é ocasionado quando a taxa de bits (*bitrate*) é reduzida pela quantização [1] (Figura 2.8).

Figura 2.8: Exemplo de borrado causado pela redução da taxa de bits [2].

- Serrilhado ou Anelamento – O serrilhado é um artefato introduzido no processo de quantização dos coeficientes transformados, sendo relacionado com o fenômeno de Gibbs. Ele é mais perceptível nas bordas com alto contraste e tem a aparência de duplicações dos contornos da imagem. Na Figura 2.9 apresenta-se uma imagem com efeito de serrilhado reduzido (direita) e a mesma imagem com o efeito serrilhado mais forte(esquerda).

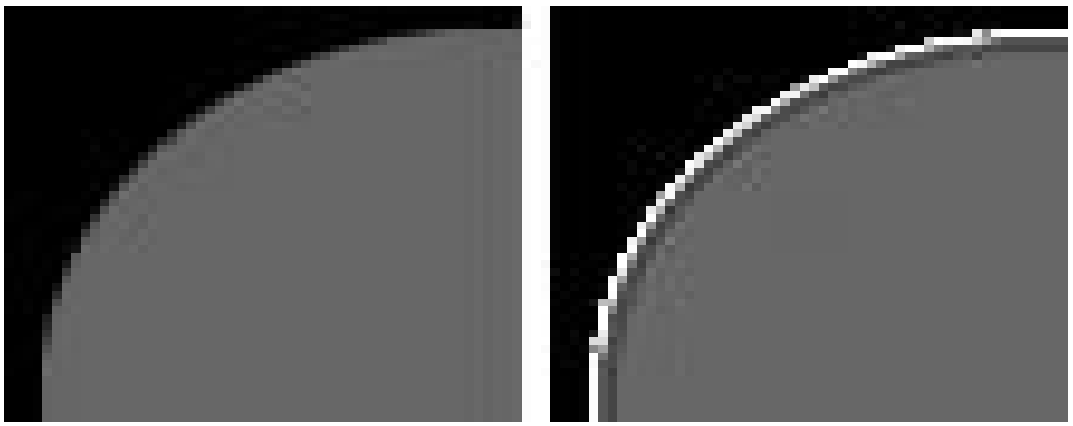


Figura 2.9: Exemplo de imagens sem (esquerda) e com (direita) serrilhado.

- Ruído – O ruído é definido como um padrão de flutuações de intensidade, que pode afetar a qualidade da imagem ou vídeo. Existem diversos tipos de ruído (*noise*) produzidos comumente pelos algoritmos de compressão, entre eles pode-se citar os ruídos do tipo mosquito e quantização (ver Figuras 2.10 e 2.11 respetivamente).
- Recorte (*Clipping*) – É o resultado dos cortes abruptos nos valores dos picos superior e inferior do intervalo dinâmico dos pixels de uma imagem/vídeo, produzindo efei-

(a) Original

(b) Com ruído

Figura 2.10: Exemplo do ruído mosquito.

tos de serrilhado (*aliasing*) que são causados pelas altas frequências criadas nestas descontinuidades, como ilustrado na Figura 2.12. Técnicas de aumento na nitidez (*sharpness*) da imagem podem produzir *clipping* [22].

- Blocagem – Os efeitos de blocagem são o resultado de uma quantização grosseira em transformadas feitas bloco a bloco (com a DCT). A quantidade de defeitos visíveis aumenta com a taxa de compressão. A Figura 2.12 (esquerda) apresenta a imagem com uma taxa de bits alta (taxa de compressão baixa), enquanto que a Figura 2.13 (direita) apresenta a mesma imagem com o efeito de blocagem evidente, causado pela redução da taxa de bits (aumento da taxa de compressão).

2.4 Sistemas Digitais de Áudio

Nesta seção, são apresentados alguns conceitos sobre a compressão de sinais de áudio e os padrões de compressão de áudio mais utilizados na atualidade. No final da seção, discutimos brevemente os artefatos que influenciam a qualidade do áudio.

2.4.1 Compressão do Áudio Digital

A grande maioria dos algoritmos de compressão de áudio tem como objetivo comprimir o sinal de áudio. Isto é feito utilizando aspectos de percepção de som (psico-acústica) para eliminar sons relevantes ou pouco perceptíveis e, assim, reduzir o tamanho do sinal [14].

(a) Original

(b) Com ruído

Figura 2.11: Ruído de Quantização.

Entre os formatos de som comprimido mais importantes podem ser citados: o *Dolby Digital*, o DTS (*Digital Theater System*) e o MPEG (*Moving Pictures Experts Group*). *Dolby Digital* e DTS são formatos muito populares em sistemas de entretenimento caseiros e cinemas. Os padrões propostos pelo MPEG são amplamente utilizados para representar áudio e vídeo com taxas baixas de bits. O primeiro padrão desenvolvido pelo MPEG foi o MPEG-1 (definido em ISO/IEC 11172-3), que é utilizado em modelos psico-acústicos para a redução de dados. O MPEG Layer-3 (mp3) baseado no padrão anterior, utiliza um tipo de compressão com perda que reduz bastante a quantidade de dados necessários para representar o sinal, ainda assim mantendo uma qualidade aceitável. Dois modelos populares de áudio *surround* baseados em MPEG são o MPEG-2 BC (*Backward Compatible*, ISO/IEC 13818-3) e o MPEG-2 AAC (ISO/IEC 13818-7). Alguns dos padrões de áudio são listados na Tabela 2.2.

Figura 2.12: Imagem Lena: (a) original e (b) com artefato recorte.

Figura 2.13: Efeito de blocagem produzido pela redução do *bitrate*.

2.4.2 Artefatos Comuns

Os artefatos mais comuns nos sinais de áudio digital são:

- Pre-eco – É um artefato produzido pelo processo de compressão de áudio. Ocorre quando o som é ouvido antes que este ocorra. Este artefato é mais perceptível em sons impulsivos de instrumentos de percussão, como castanholas ou címbalos.
- *Birdie Effect* – Este artefato aparece no processo de compressão devido a uma política de alocação de bits inadequada. Gera a sensação de ouvir um som falso.
- Recorte – De forma similar ao caso do vídeo digital, os recortes nos picos dos sinais de áudio no nível máximo produzem distorções no sinal de áudio. Uma ilustração deste artefato é apresentada na Figura 2.14, onde a amplitude do sinal é incrementada na

Tabela 2.2: Padrões de compressão de áudio [14].

Padrão	**Aplicação**	**Taxa de bits**
Dolby Digital	Cinemas, HDTV, DVDs	448 Kb/s
DTS	Espetáculos musicais, sistemas de cinema em casa, DVDs.	512 kb/s
MPEG-1 Layer 3 (mp3)	formato de áudio para transferências e armazenamento.	320 kb/s

metade da linha do tempo. A informação perdida por causa deste tipo de artefato é difícil de recuperar.

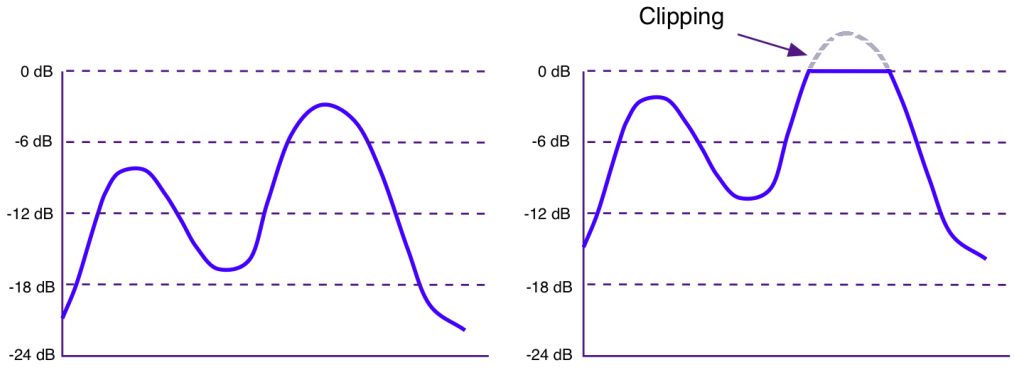


Figura 2.14: Efeito de Recorte num sinal de áudio (*clipping*).

Capítulo 3

Qualidade de Sinais

Neste capítulo são descritas as diferentes abordagens de aferição de qualidade de áudio e vídeo. Os métodos subjetivos mais representativos são descritos em detalhe, assim como algumas das métricas objetivas mais importantes. Finalmente, as atuais propostas para a aferição da qualidade áudio-visual são descritas brevemente. Também está incluída uma descrição do projeto *Audiovisual HD Quality (AVHD)*, realizado pelo *Video Quality Experts Group* (VQEG).

3.1 Avaliação Subjetiva de Vídeo e Áudio

Nesta seção, apresentamos uma breve introdução das principais metodologias de avaliação subjetiva da qualidade de áudio e vídeo. Os experimentos subjetivos ou psico-físicos representam as técnicas mais precisas para a avaliação da qualidade de áudio e vídeo. Em um teste subjetivo, um grupo de observadores (participantes ou avaliadores) assiste e/ou ouve a uma série de sequências de vídeo e/ou áudio e atribui um valor à qualidade do áudio e/ou vídeo. Tirando-se a média dos valores atribuídos pelos participantes, obtém-se o *Mean Opinion Score (MOS)* de cada uma das sequências exibidas.

Embora esta estratégia seja a mais precisa, ela exige muito tempo e recursos para seu planejamento, execução e análise de resultados. Além disso, é necessária a disponibilidade de participantes, equipamento físico sofisticado, *software* específico, ambiente adequado para o teste. Além disso, para garantir a qualidade e precisão do experimento, deve-se escolher uma metodologia experimental adequada. Os testes têm que ser conduzidos seguindo uma das metodologias recomendadas pela *International Telecommunication Union* (ITU) ou a *European Broadcasting Union (EBU)*. A utilização de metodologias padronizadas nas diferentes recomendações publicadas por estas organizações garantem que os seus resultados possam ser reproduzíveis e comparados entre os diferentes laboratórios.

Os métodos já estabelecidos como padrões para os experimentos que envolvem áudio, vídeo e audio-visuais são:

- Áudio: P.800 [39], BS.1116 [40] e BS.1534 [41].
- Vídeo: BT.500 [42], BT.710 [43] e P.910 [6].
- Áudio-visual: P.911 [44] e P.920 [45].

Estas recomendações descrevem as metodologias utilizadas, especificando aspectos tais como as escalas de avaliação, conforme ilustrado na Figura 3.1. Outros parâmetros especificados nas recomendações incluem a escolha do material de teste (formato, duração, conteúdo), o ambiente físico onde o teste será realizado, o equipamento, o número mínimo de participantes, etc.

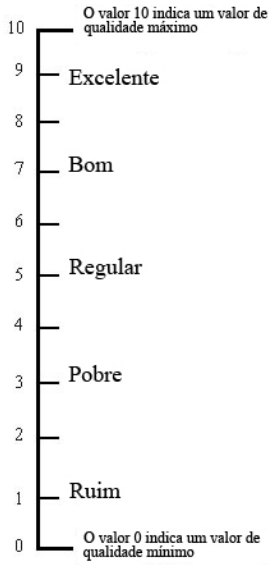


Figura 3.1: Escala numérica de qualidade com 11 degraus [6].

Os procedimentos de avaliação podem ser classificados segundo a forma de apresentação das sequências de teste (estímulos), podendo ser de estímulo simples ou duplo. Na abordagem simples, a sequência é apresentada unicamente, enquanto que no caso do duplo ela é apresentada junto com uma sequência de referência (original). Alguns dos procedimentos mais importantes encontrados nas recomendações [44] utilizados em experimentos de áudio e vídeo são:

- Avaliação de Categoria Absoluta (*Absolute Category Rating (ACR)*) – Neste método, as sequências são apresentadas uma depois de outra, sem utilizar sequências de referência. Este método é conhecido também como o método de estímulo simples (Single Stimulus Method – SSM). O método especifica que, depois de apresentada a sequência, o participante tem que pontuá-la imediatamente. Uma escala de cinco categorias é utilizada para a avaliação da qualidade percebida. Elas são: 'ruim', 'pobre', 'regular', 'bom', e 'excelente'. Também é possível utilizar uma escala com 11 categorias quando é necessária uma maior precisão na pontuação (ver Figura 3.1).
- ACR com Referência Oculta (*ACR with Hidden Reference (ACR-HR)*) – Neste método, as sequências são apresentadas de forma similar ao método ACR, sendo utilizada a mesma escala. A diferença é a inclusão de uma sequência de referência (livre de erros) para cada uma das sequências de teste apresentadas. Esta sequência de referência deve ser apresentada como qualquer outra sequência (sem especificar para o avaliador que ela é a referência). Isto é denominado "condição de referência oculta".

Pontuações diferenciais (*Differential* MOS) de qualidade são calculadas utilizando a pontuação da sequência de teste e a pontuação da sua referência (oculta) correspondente. Ou seja, subtrai-se a nota atribuída à sequência de referência da nota da sequência de teste e adiciona-se o valor '5' ao resultado (no caso de ser utilizada uma escala de '11' pontos é adicionado o valor '10'). Se o resultado final for igual a '5' isto significa que a sequência tem uma 'excelente' qualidade. Entretanto, para aquelas sequências com resultado igual a '1' teremos uma qualidade 'ruim'. Para garantir a precisão dos resultados é necessário que um especialista escolha uma ótima sequência de referência.

- Avaliação com Categoria de Degradação (*Degradation Category Rating (DCR)*) – É também denominada *Double Stimulus Impairment Scale* (DSIS). Neste método, a sequência de referência é sempre apresentada antes da sequência de teste (a presença da referência não está oculta para o avaliador). Os participantes pontuam os defeitos na sequência de teste utilizando uma escala de cinco categorias: 'imperceptível', 'perceptível, mas não irritante', 'ligeiramente irritante', 'irritante' e 'muito irritante'. Este método é útil para avaliar os defeitos claramente perceptíveis nas sequências. Na Figura 3.2 é apresentado um diagrama do metodo de avaliação DCR. Observe que as sequências, de aproximadamente 10 segundos, são apresentadas em intervalos de 2 segundos e são pontuadas em 10 segundos.

- Comparação de Pares (*Pair Comparison (PC)*) – Para este método, todos os possíveis pares de combinações das sequências de teste são apresentados (referências e videos degradados). Os observadores têm que indicar qual das duas sequências eles acham que têm a melhor qualidade. Com este método pode-se obter uma maior distinção entre as condições das sequências, mas é necessário mais tempo para o experimento em comparação aos outros métodos.

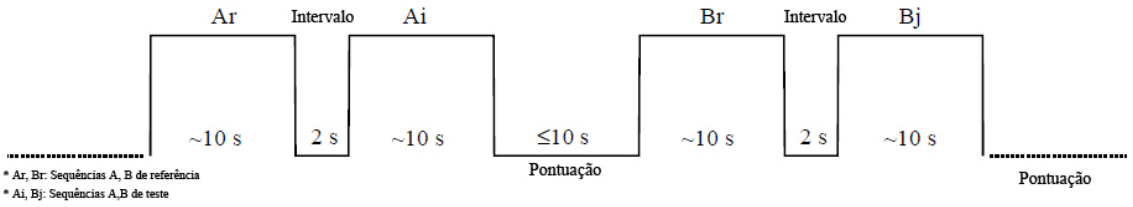


Figura 3.2: Apresentação do estímulo no método DCR (r: referência, i:teste) [6].

3.2 Avaliação Objetiva de Áudio e Vídeo

A avaliação objetiva refere-se aos métodos computacionais utilizados para aferir a qualidade dos sinais de vídeo e áudio. Estes métodos representam uma boa alternativa aos experimentos subjetivos estudados na seção anterior, uma vez que os métodos objetivos podem ser utilizados para monitorar a qualidade dos sinais trasmitidos em aplicações reais. Além disso, utilizando a informação fornecida pelas métricas objetivas de qualidade, podem ser feitas comparações entre sistemas e algoritmos de processamento de vídeo, permitindo aperfeiçoá-los.

As métricas objetivas de qualidade podem ser classificadas em três tipos, segundo a quantidade de informação necessária para estimar a qualidade do sinal. São estes:

- Referência Completa (*FR*) – O primeiro tipo inclui as métricas de "Referência Completa" ou Referenciadas que utilizam todo o sinal original para obter uma estimação da qualidade. Este é o tipo de métrica com melhor desempenho, o que se deve, em grande parte, à disponibilidade da sequência original. As métricas FR são desenvolvidas para aplicações *off-line*, incorporando vários aspectos do SVH e do SAH considerados relevantes para estimação de qualidade. A utilização destas métricas para aplicações em tempo real (tais como vídeo-conferências, transmissões de Internet, etc.) não é possível devido à exigência do sinal de referência e a sua alta complexidade computacional. Outros fatores limitantes são a grande quantidade de informação e a dificultade de sincronização entre as sequências de referência e teste. Na Figura 3.3 é apresentado o diagrama de blocos correspondente a um sistema genérico de aferição de qualidade com referência completa. É possível observar que a referência inteira está disponível no ponto de aferição.

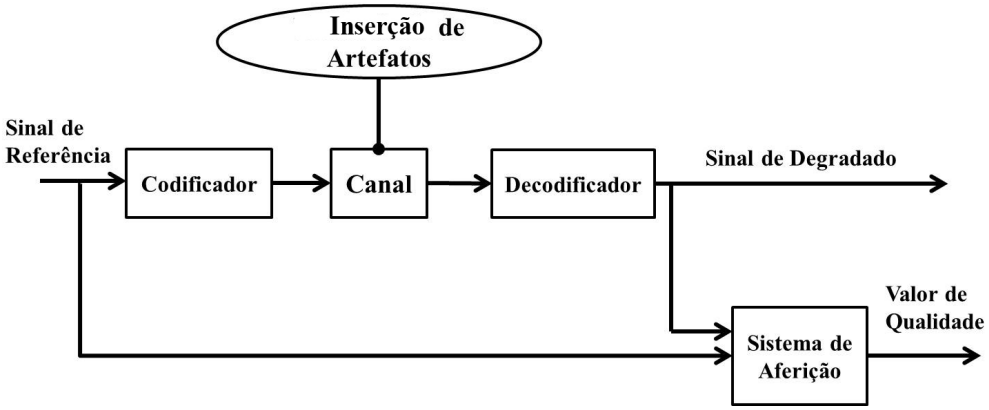


Figura 3.3: Sistema de aferição de qualidade com referência completa (Adaptação [4]).

- Referência Reduzida (*RR*) – O segundo tipo inclui as métricas de "Referência Reduzida" que utilizam uma pequena parte da informação do sinal original (em alguns casos só alguns parâmetros) [8]. Uma caraterística importante das métricas RR é a forma de transmissão do sinal de referência. É assumido que existe um canal lateral, livre de erros de transmissão, por onde o sinal de referência é transmitido, enquanto que a sequência de teste é transmitida por um canal com maior largura. A suposição de ter um canal lateral livre de erros é realista no sentido que o canal seja o suficientemente pequeno (tamanho da informação minimamente necessária para referência) para garantir a proteção dele a erros de transmissão. Como a quantidade de informação utilizada pelas métricas RR é bem menor, elas são menos precisas que as métricas FR. Por outro lado, elas são computacionalmente menos complexas, o que torna a sua implementação mais fácil. Na Figura 3.4 é apresentado o diagrama de blocos de uma métrica RR. Neste caso, a referência está representada por algumas características extraídas do sinal original, que são transmitidas através de um canal auxiliar.

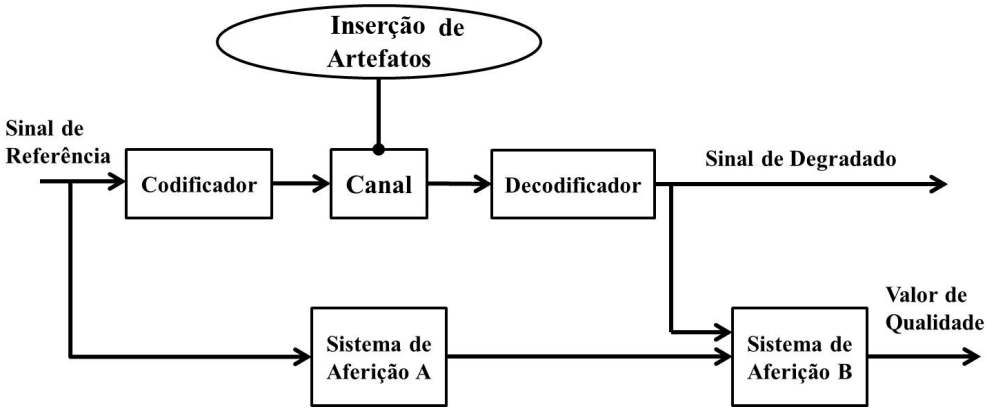


Figura 3.4: Sistema de aferição de qualidade com referência reduzida (Adaptação [4]).

- Sem-Referência (*NR*) – O terceiro tipo inclui as métricas "Sem-Referência" ou Não-Referenciadas que não utilizam nenhuma informação do sinal original. Neste contexto, muitas das métricas estudam várias das características do sinal que são relevantes para a aferição da qualidade. Estas características são, na maioria das métricas, artefatos como o borrado, a blocagem e o serrilhado. Estas características são identificadas e aferidas para que, posteriormente, o algoritmo calcule a intensidade de cada distorção. Ao final, estes valores são combinados de forma a se obter um valor final para a qualidade. A Figura 3.5 apresenta o diagrama de blocos de uma métrica objetiva sem-referência.

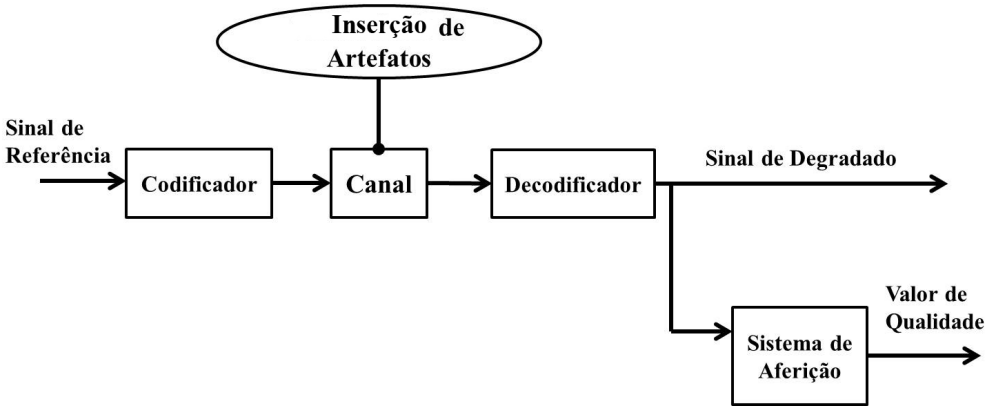


Figura 3.5: Sistema de aferição de qualidade sem-referência (Adaptação [4]).

As métricas objetivas de qualidade também podem ser classificadas segundo a abordagem utilizada para estimar a qualidade do sinal. Existem dois tipos principais de abordagens: sensíveis a erro ou extração de características. A primeira abordagem analisa as diferenças visíveis entre as sequências originais e de teste, obtendo as estimações de qualidade com base nos graus de similaridade entre as duas sequências. Esta abordagem é utilizada geralmente pelas métricas do tipo FR, já que a sequência de referência original para este tipo de métrica está disponível. Neste caso, assume-se que o original está livre de defeitos ou artefatos. A abordagem de extração de características procura extrair do

vídeo características que são consideradas relevantes para estimação da qualidade. Algumas das características consideradas são o contraste, o borrado, etc. Esta abordagem é particularmente utilizada pelas métricas do tipo RR e NR, uma vez que estas características podem ser aferidas utilizando alguns parâmetros do sinal original ou apenas o sinal de teste.

3.2.1 Métricas Objetivas de Vídeo

Métricas de Vídeo Com-Referência Completa (FR)

- *Just Noticeable Differences JND* – O modelo de Sarnoff, também conhecido como modelo de discriminação visual (VDM) [21], foi patenteado pela *Tektronix Company* e comercializado mediante o analisador de imagens *PQA600*. O modelo leva em consideração aspectos do SVH, como cor e variação temporal. Ele foi projetado para estimar a probabilidade de detecção de artefatos num quadro do vídeo. É utilizado o conceito de diferenças minimimanente perceptíveis (JNDs), que são limiares para percepção de mudanças de níveis de luminosidade nas imagens. O JND é definido como a menor diferença detectável entre os níveis inicial e secundário de um estímulo. É possível interpretar as JNDs do seguinte modo:

  - Uma unidade de JND representa diferenças pouco visíveis.
  - Três unidades JND representam diferenças visíveis mediante observação detalhada.
  - Cinco unidades JND representam diferenças claramente visíveis.

- Métrica de qualidade de Vídeo (*Video Quality Metric* VQM) – Esta métrica foi recentemente adotada como padrão para a qualidade de vídeo objetiva pelo ANSI (*American National Standards Institute*). No relatório do VQEG (*Video Quality Experts Group*) [19], o VQM obteve um dos melhores resultados entre as métricas participantes. O algoritmo utilizado pelo VQM inclui medições para os diferentes artefatos de vídeo, como o borrado (*blurring*), ruído, blocagem (*blockiness*) e distorção de cor. Os resultados para cada artefato são combinados para gerar um valor para a qualidade.

  A métrica VQM possui quatro estágios principais. São estes:

  - Calibração - No estágio de calibração são corrigidos os deslocamentos de espaço e tempo e as variações de contraste e brilho entre as sequências comparadas.
  - Extração de características - Na fase de extração, são aplicados filtros específicos para reconhecer as características de qualidade (que representam mudanças na percepção entre os quadros comparados). Um limiar de visibilidade é aplicado na fase final.
  - Estimação de parâmetros - No estágio de estimação de parâmetros são obtidos os parâmetros de qualidade mediante uma comparação entre as características obtidas para a sequência de teste e a sequência original.
  - Combinação - No estágio de combinação, a qualidade global é calculada utilizando uma combinação linear da estimação das intensidades dos artefatos obtidas nos estágios anteriores.

- Similaridade Estrutural e qualidade de Imagem (*The Structural SIMilarity (SSIM) index*) – Esta métrica [7] se baseia na ideia que as imagens naturais são altamente estruturadas, isto é, os sinais das imagens mantêm uma forte relação entre si, carregando informação sobre a estrutura dos objetos na imagem. O algoritmo do SSIM calcula três características das imagens original x e de teste y: a luminância $l(x, y)$, o contraste $c(x, y)$ e a estrutura $s(x, y)$. Para calcular estas características, são utilizadas as seguintes expressões:

$$l(x, y) = \frac{2 \cdot \mu_x \cdot \mu_y + C_1}{{\mu_x}^2 + {\mu_y}^2 + C_1}, \tag{3.1}$$

e

$$\mathrm{c(x, y)} = \frac{2 \cdot \sigma_x \cdot \sigma_y + C_2}{{\sigma_x}^2 + {\sigma_y}^2 + C_2}, \tag{3.2}$$

e

$$\mathrm{s(x, y)} = \frac{\sigma_{xy} + C_3}{\sigma_x \cdot \sigma_y + C_3'}. \tag{3.3}$$

nas quais C_1, C_2 e C_3 são constantes pequenas. Os valores de μ e σ estão denotados pelas expressões:

$$\mu_\mathrm{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{3.4}$$

e

$$\sigma_\mathrm{x} = \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{\frac{1}{2}}. \tag{3.5}$$

Em seguida, o valor da qualidade global é calculado com base nesses resultados. O resultado da qualidade varia entre '0' e '1', sendo '1' a melhor qualidade possível. A fórmula geral da métrica SSIM é dada pela seguinte expressão:

$$\mathrm{SSIM(x, y)} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{({\mu_x}^2 + {\mu_y}^2 + C_1)({\sigma_x}^2 + {\sigma_y}^2 + C_2)}. \tag{3.6}$$

Na Figura 3.6 é apresentado o diagrama de blocos da métrica SSIM. O processo para atingir um valor de similaridade começa com a comparação da luminância. Subsequentemente, uma função de comparação para o contraste é obtida. O sinal é normalizado, obtendo-se a função para a comparação estrutural. Finalmente, as três componentes são combinadas e o valor da similaridade é obtido.

Métricas de Vídeo Com Referência Reduzida (RR)

- Modelo com Referência Reduzida da Universidade Yonsei – Este modelo foi apresentado pela Universidade de Yonsei para a avaliação no ITU-R [8]. O modelo está baseado no fato de que o SVH é mais sensível às degradações em torno das bordas de um objeto. Assim, o modelo computa valores dos pixels da borda para a sequência de referência e de teste e compara as diferenças entre os dois conjuntos de pixels para obter um valor da qualidade do vídeo. No primeiro estágio, um algoritmo de reconhecimento de bordas é aplicado ao vídeo de referência e, depois, um limiar

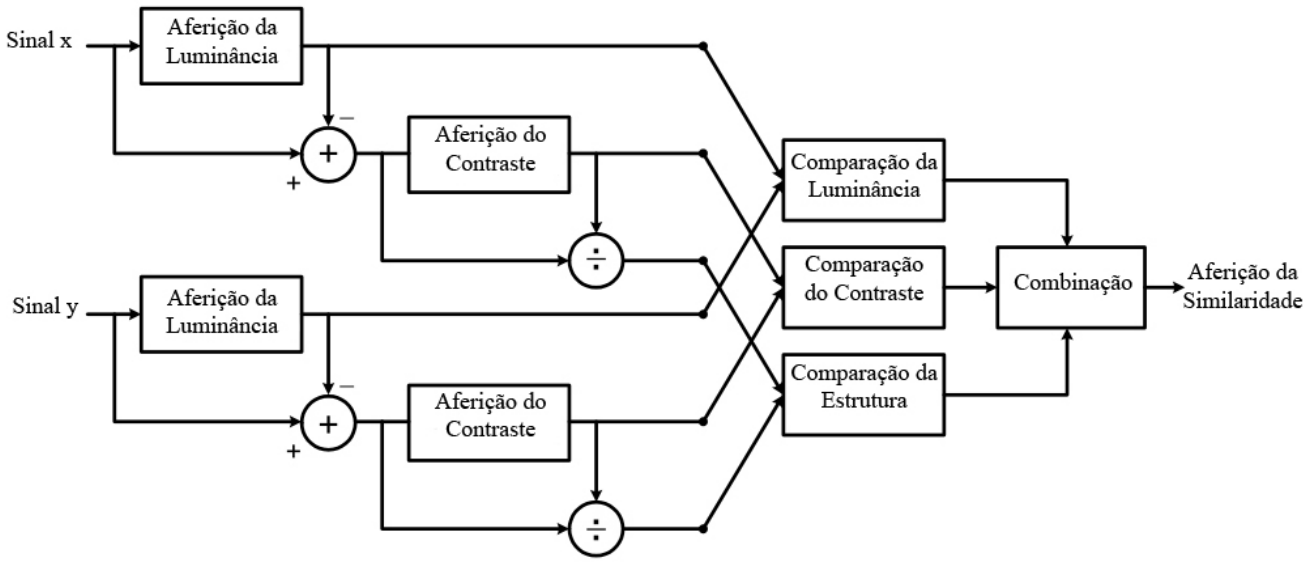


Figura 3.6: Diagrama em blocos da métrica SSIM [7].

para obter as bordas finais. A Figura 3.7 apresenta o processamento da imagem utilizado para calcular os pixels das bordas.

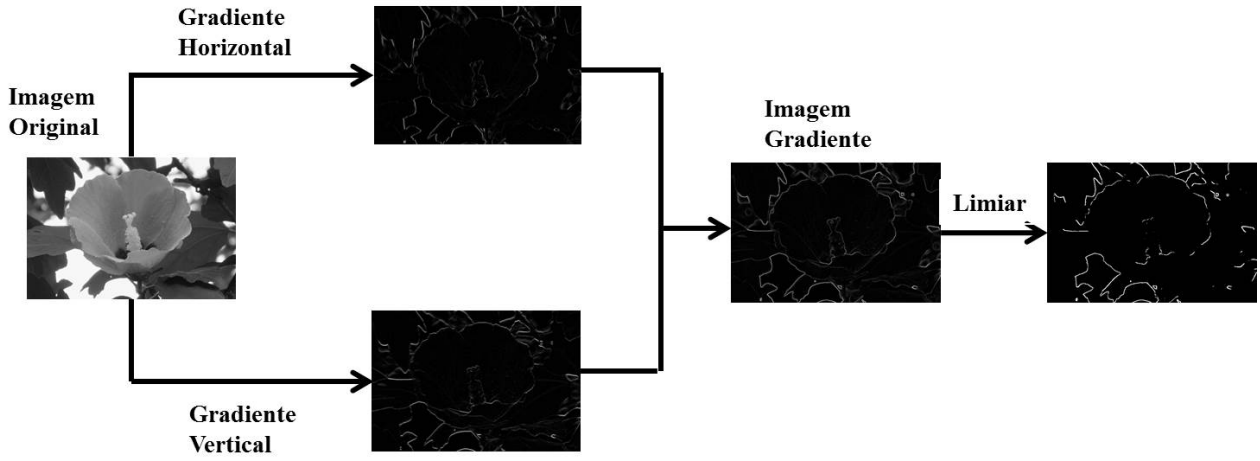


Figura 3.7: Processamento da imagem para o cáculo das bordas [8].

Certo número de pixels das bordas é selecionado do frame obtido ao final do processamento da imagem original. Em seguida, são computados os valores e posições desses pixels e, posteriormente, codificados e transmitidos por um canal (livre de erros). No receptor, o sinal de teste (transmitido por um canal convencional) passa pelo mesmo processo que o sinal original passou para o cálculo das bordas. Finalmente, são comparadas as posições e valores dos pixels do sinal original e de teste. Com base nas diferenças entre estes valores, obtém-se o valor da qualidade do sinal de vídeo.

- Força Harmônica Local (*Local Harmonic Strength* LHS) – Esta métrica [9] está baseada na força harmônica local (FHL), uma característica que pode ser interpretada como a aferição da atividade espacial, estimada em termos das bordas horizontais ou verticais de uma imagem. Em outras palavras, o cálculo da qualidade da imagem está baseada nas estimações de ganho ou perda harmônica, obtidas após uma análise da força harmônica local das bordas da imagem (imagem gradiente). A Figura 3.8 apresenta um diagrama com os estágios da métrica.

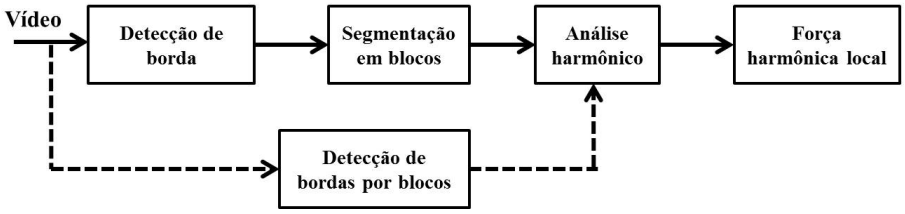


Figura 3.8: Diagrama de blocos da métrica de qualidade FHL [9].

O primeiro passo do algoritmo é a detecção das bordas mediante um operador *Sobel* de 3×3 pixels de tamanho, que gera uma imagem gradiente. Em seguida, a imagem é processada por um algoritmo de segmentação e dividido em blocos de 32×32 pixels de tamanho. A força harmônica local é calculada para cada bloco e os valores obtidos são organizados numa matriz. Um procedimento similar é aplicado à imagem de teste, gerando outra matriz. É realizada uma análise para diferenciar o incremento (ganho) ou decremento (perda) da força harmônica de ambas as matrizes. O ganho corresponde ao valor da blocagem e a perda ao valor do borrado. Para obter o valor da qualidade global uma média aritmética pode ser utilizada.

Métricas de Vídeo Sem-Referência (NR)

- Métrica de Qualidade Sem-Referência de Caviedes [22] -– A métrica de Caviedes estima características consideradas "desejáveis" e "não desejáveis" para uma imagem [22]. Entre essas características consideradas pelo algoritmo de Caviedes estão os artefatos de blocagem, serrilhado, recorte (*clipping*), ruído, contraste e nitidez. As métricas utilizadas para computar a intensidade dos artefatos de ruído, recorte e contraste estão baseadas na métrica de qualidade utilizada em um sistema de otimização para vídeo [46]. As métricas para estimar a intensidade dos artefatos de blocagem e *ringing* foram desenvolvidas pelo autor. Uma vez calculadas as intensidades, os valores são refinados e ponderados. Em seguida, é feita uma combinação de todos estes resultados para se obter um valor da qualidade global do quadro. Os resultados obtidos mostraram uma alta correlação com os resultados obtidos em experimentos subjetivos.

- Métrica de Qualidade Sem-Referência Baseada na Aferição de Artefatos [23]– Esta métrica é baseada na suposição que a qualidade percebida pelo usuário é afetada pelo tipo, quantidade e intensidade dos artefatos presentes no quadro [23]. A métrica compreende três sub-métricas de artefatos:

- Blocagem - É uma modificação da métrica de Vlachos [47] que estima a intensidade da blocagem com base na razão entre a correlação entre os pixels em regiões inter e intra blocos.
- Borrado – Baseada no algoritmo de Marziliano [48]. Esta métrica identifica as bordas fortes de um quadro e detecta a largura destas bordas. Quanto mais larga a borda, mais borrado o quadro.
- Ruído/Ringing – baseado no trabalho de Lee [49]. A métrica identifica as bordas fortes da imagem e mede a atividade espacial em torno destas bordas.

A simplicidade destas sub-métricas permite executá-las em tempo real. Um modelo da distorção final é obtido utilizando a métrica Minkowski [4]. A Figura 3.9 apresenta um diagrama de blocos simplificado do modelo da métrica.

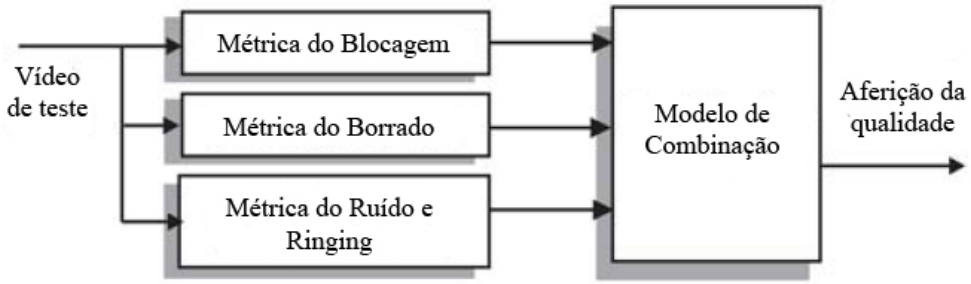


Figura 3.9: Diagrama de blocos da métrica de vídeo sem-referência [4].

- BLIINDS-II – Este algoritmo está baseado em um modelo de estatísticas de coeficientes DCT (*discrete cousine transform*) para cenas naturais. A Figura 3.10 mostra a estrutura e os estágios do modelo BLINDS-II. No primeiro estágio, a imagem é particionada em blocos mediante um algoritmo de segmentação, produzindo blocos de tamanho $n \times n$. A ideia da segmentação em blocos é que o SVH processa o espaço visual de maneira local [1]. Em seguida, é aplicada uma transformada DCT a cada um dos blocos. Para poder obter a informação direcional, estes blocos são particionados em três sub-regiões orientadas. Isto se deve ao fato que algumas distorções modificam a orientação local de energia de maneira irregular.

  No segundo estágio é aplicado uma função generalizada de densidade gaussiana a cada um dos blocos de coeficientes DCT, bem como às partições específicas em cada bloco. No terceiro estágio, são computadas as funções dos parâmetros do modelo. Este estágio corresponde a extração das características para aferição da qualidade. No último estágio, é utilizado um modelo bayesiano para estimar a qualidade da imagem.

- Métrica Combinatória Baseada em Borrado e Blocagem (Proposta neste trabalho) – Esta proposta combina duas métricas que calculam o nível de Borrado e Blocagem em uma imagem, descritas a seguir.

  - Métrica de Borrado (Narvekar e Karam [50]) – Esta métrica calcula a probabilidade de detecção do borrado nas bordas da imagem. O processo começa com a detecção das bordas horizontais da imagem. Em seguida, é calculada

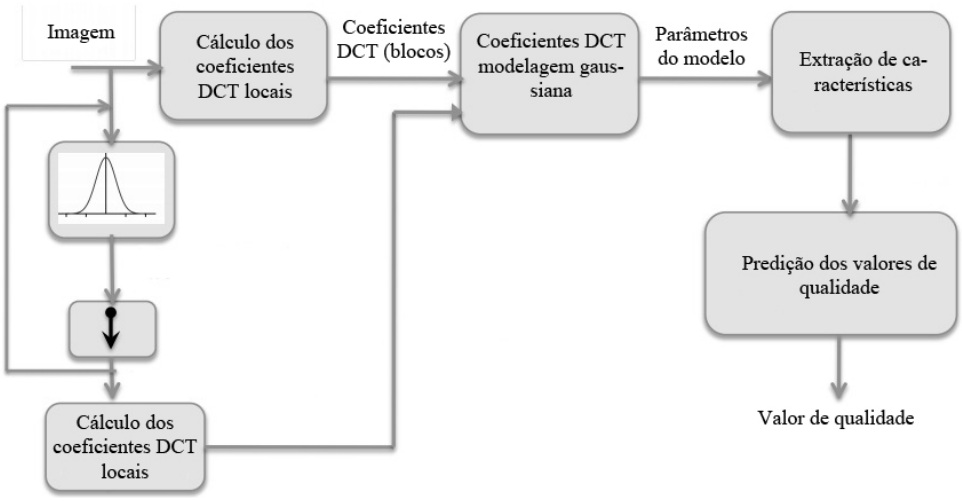


Figura 3.10: Visão de alto nível da estrutura do algoritmo BLINDS-II [10].

a probabilidade de detecção do borrado para cada pixel que pertence a uma das bordas. Com esta informação, um histograma normalizado da detecção do borrado é obtido. Utilizando este histograma é possível calcular o nível de borrado da imagem.

– Métrica de Blocagem (Wang e Bovik [51]) – O efeito de blocagem é calculado em direções vertical e horizontal separadamente. Para calcular o nível de blocagem vertical, a imagem é tratada como uma matriz. Para calcular a blocagem horizontal, a imagem é reorganizada para formar um sinal 1-D. É possível identificar o efeito de blocagem nos picos das frequências características DCT 1/8, 2/8, 3/8, e 4/8. A média dos níveis de blocagem horizontal e vertical determina o nível global da blocagem.

Os valores de borrado e blocagem são combinados utilizando uma regressão linear simples, de forma a se obter uma estimação para a qualidade da imagem. O valor da qualidade de vídeo está dado pela seguinte equação:

$$\mathrm{Qv} = -195.08 \cdot \mathrm{Borrado} + -55.23 \cdot \mathrm{Blocagem} + 320.94, \tag{3.7}$$

3.2.2 Métricas Objetivas de Áudio

Em seguida, são apresentadas algumas métricas objetivas representativas para a aferição da qualidade de áudio. As métricas apresentadas nesta seção são do tipo FR e NR. Não foram encontradas na literatura métricas de áudio RR.

Métricas de Áudio Com-Referência (FR)

- PEAQ [24] — PEAQ é uma métrica de áudio FR recomendada pela ITU-R para a avaliação da qualidade [24]. Existem duas versões do algoritmo PEAQ: a versão

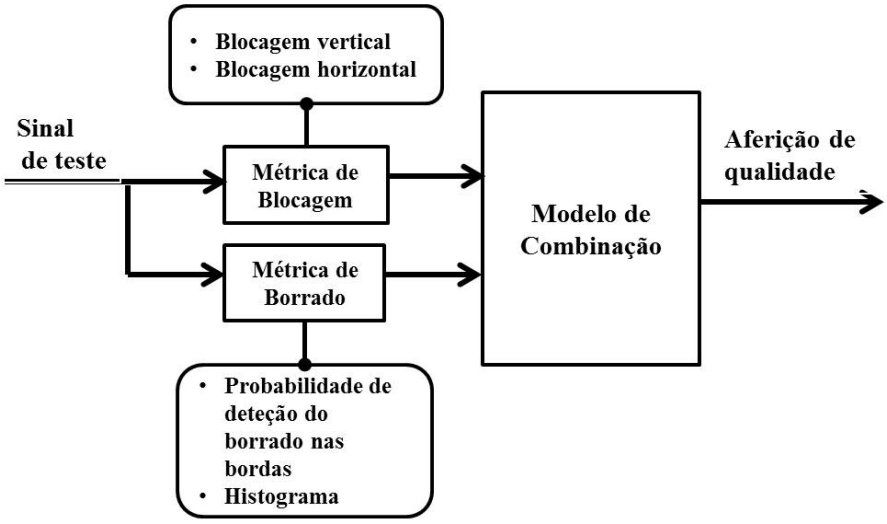


Figura 3.11: Métrica Combinatória Baseada em Borrado e Blocagem.

básica e a avançada. A versão básica é utilizada em aplicações onde a eficiência computacional é um requisito importante. Na versão avançada, o cálculo da qualidade é mais preciso, mas o algoritmo requer quatro vezes mais recursos computacionais que na versão básica [24].

A maior diferença estrutural entre as duas versões é que a versão básica possui apenas um modelo acústico periférico baseado na *Fast Fourier Transform* (FFT), enquanto que a versão avançada possui dois modelos acústicos baseados na FFT e em um banco de filtros. As variáveis de saída do modelo, chamadas de MOVs, têm características baseadas na sonoridade, modulação, mascaramento e adaptação. A versão avançada possui 11 MOVs e a versão básica possui apenas 5. As MOVs servem de entrada para uma rede neural treinada para mapear as MOVs em uma escala de diferença global (EDG). Esta escala varia de '0' a '4', onde '0' representa um sinal com distorções imperceptíveis e '4' um sinal com distorções muito irritantes.

Métricas de Áudio Sem-Referência (NR)

- SESQA (*Single Ended Speech Quality Metric*) [11] - A métrica SESQA foi proposta originalmente para a avaliação da qualidade de sinais em aplicações telefônicas. A primeira fase do algoritmo consiste de um pré-processamento do sinal. Um Detector de Atividade Vocal (DAV) é utilizado para estimar o nível de dialogo no sinal. Logo, o sinal é analisado e um conjunto de 51 características paramétricas do sinal é obtido (ver Anexo I). Depois, baseados num conjunto de 8 parâmetros chaves, é feita uma distribuição dos parâmetros em 4 classes de distorção. Estas classes de distorção incluem: (1) 'Descritores básicos de voz', (2) 'Voz não natural', (3) 'Análise do ruído', e (4) 'Interrupções - Silêncios'. Estes 8 parâmetros são listados a seguir:
  - Altura media (*PitchAverage*)
  - Robotização (*Robotization*)
  - Curtose do coeficiente de predição linear (*LPCcurt*)

– Relação Sinal-Ruído (*Signal to noise ratio, SNR*)
– SNR estática e segmentada (*EstSegSNR*)
– Interrupções de dialogo (*SpeechInterruptions*)
– Reduções abruptas (*SharpDeclines*)
– Duração de silêncios (*MuteLength*)

Os parâmetros chaves e estas classes de distorção são utilizados pelo modelo para calcular a qualidade do diálogo nos sinais. A Figura 3.12 mostra o diagrama de blocos da métrica.

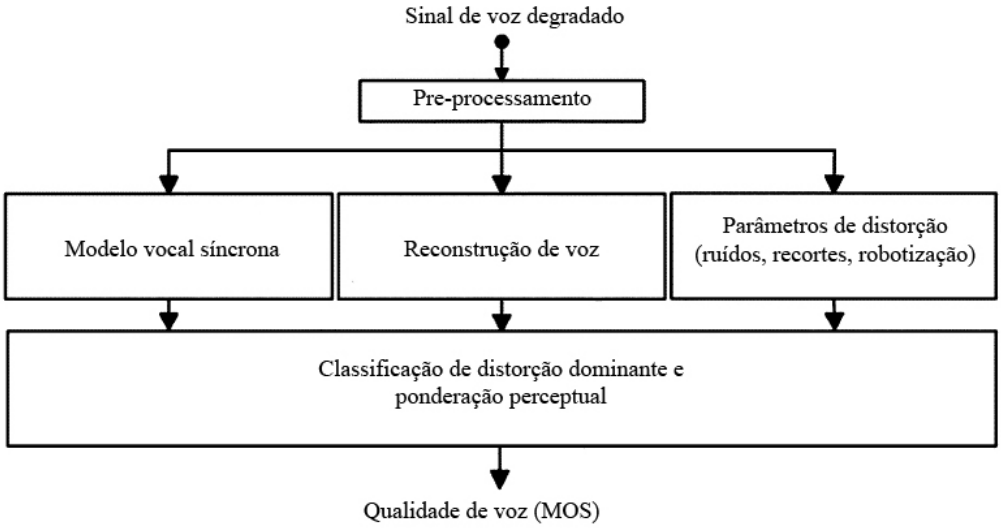


Figura 3.12: Diagrama de blocos do algoritmo SESQA [11].

3.3 Qualidade Áudio-Visual

A qualidade de uma sequência áudio-visual é influenciada por vários fatores, entre estes as qualidades individuais das componentes de áudio e vídeo. Experimentos subjetivos são os métodos mais utilizados para estudar a interação entre as qualidades das componentes de vídeo e áudio, avaliadas individualmente. O estudo desta interação é fator importante na estimação da qualidade global de sinais áudio-visuais e a grande maioria dos estudos nestas área tenta explicar qual é o grau de influência de cada uma destas componentes na qualidade global.

Há uma lista extensa de trabalhos focados em explorar este tópico [26, 27, 28]. Cada um destes experimentos produz um modelo que mapeia a qualidade de áudio e a qualidade do vídeo para a qualidade áudio-visual. Entre os modelos diponíveis na literatura, destacamos os três estudos descritos a seguir:

- Um estudo bastante completo foi realizado por Hands [26]. Hands realizou dois experimentos subjetivos com o objetivo de encontrar um mo-delo (subjetivo) da

qualidade áudio-visual [26]. Ele utiliza dois tipos de cenas nos experimentos: (1) sequências de diálogos simples e (2) cenas com muito movimento. Os experimentos sugerem bons resultados quando se utiliza um elemento multiplicativo na combinação dos valores médios de qualidade (MOS) das componentes de áudio e vídeo. Um analise de regressão múltipla foi realizada nos dados do experimento 2, este analise deu como resultado dois modelos, que são apresentados a seguir:

$$\mathrm{QSav}_{\mathrm{H1}} = 0.25 \cdot \mathrm{MOS}_v + 0.15 \cdot (\mathrm{MOS}_a \times \mathrm{MOS}_v) + 0.95, \tag{3.8}$$

e

$$\mathrm{QSav}_{\mathrm{H2}} = 0.17 \cdot (\mathrm{MOS}_a \times \mathrm{MOS}_v) + 1.15, \tag{3.9}$$

no qual $\mathrm{QSav}_{\mathrm{H1}}$ e $\mathrm{QSav}_{\mathrm{H2}}$ representam os valores estimados para a qualidade subjetiva áudio-visual e MOS_a e MOS_v são as qualidades subjetivas (experimentais) das componentes de áudio e vídeo, respectivamente. Os valores de correlação obtidos foram bons, apesar da pouca diversidade de conteúdo. Os coeficientes de correlação para ambos modelos foi de 0.72.

- Na proposta de Garcia [27], são estudadas duas abordagens distintas para a aferição da qualidade de sinais áudio-visuais. O primeiro consiste de um modelo objetivo que integra a qualidade de vídeo e áudio utilizando um modelo linear simples, baseado em fatores de degradação da qualidade. Estes fatores de degradação são calculados utilizando parâmetros de QoS (*Quality of Service*) disponibilizados na rede. Os resultados sugerem que existe uma forte contribuição de ambas as componentes (áudio e vídeo) na qualidade global, dependendo do conteúdo, formato e o tipo de degradação do material utilizado. O segundo modelo consiste de um modelo subjetivo multiplicativo. Mediante um analise de regressão múltipla, realizada nos dados experimentais, um grupo de modelos são propostos por Garcia, deste grupo de modelos pode-se destacar o denotado pela seguinte equação:

$$\mathrm{QSav}_{\mathrm{G}} = 0.13 \cdot \mathrm{MOS}_v + 0.0006 \cdot (\mathrm{MOS}_a \times \mathrm{MOS}_v) + 28.49, \tag{3.10}$$

onde $\mathrm{QSav}_{\mathrm{G}}$ é a estimação para a qualidade subjetiva áudio-visual e MOS_a e MOS_v são as qualidades subjetivas (experimentais) das componentes de áudio e vídeo, respectivamente. O coeficiente de correlação para este modelo foi de 0.94.

- Winkler [28] propôs um modelo subjetivo de qualidade áudio-visual destinado a sinais com baixas taxas de bits. Este modelo é destinado a aplicativos móveis. Foram realizados experimentos subjetivos para estudar a contribuição das componentes de áudio e vídeo na qualidade áudio-visual. Os modelos resultantes apresentaram um bom desempenho e mostraram uma significativa contribuição de cada um das componentes na qualidade global. Os dois modelos propostos por Winkler, resultado de um analise de regressão múltipla, são apresentados a seguir:

$$\mathrm{QSav}_{\mathrm{W1}} = 0.103 \cdot (\mathrm{MOS}_a \times \mathrm{MOS}_v) + 1.98, \tag{3.11}$$

e

$$\mathrm{QSav}_{\mathrm{W2}} = 0.77 \cdot \mathrm{MOS}_v + 0.456 \cdot \mathrm{MOS}_a - 1.51, \tag{3.12}$$

onde $\mathrm{QSav}_{\mathrm{W1}}$ e $\mathrm{Qav}_{\mathrm{W2}}$ representam o valor predito para a qualidade subjetiva áudio-visual e MOS_a e MOS_v são as qualidades subjetivas (experimentais) das componentes de áudio e vídeo, respectivamente. O coeficiente de correlação para este modelo foi de 0.94.

Existem muitas iniciativas que visam estudar a qualidade de sinais áudio-visuais e desenvolver métodos confiáveis para estimá-la. Em 2012, os projetos do *Video Quality Experts Group* (VQEG) denominados *High Definition Television* (HDTV) *Phase II* e *Multimedia Phase II* foram unidos para criar o projeto *Audio-Visual HD Quality* (AVHD). Este projeto tem como finalidade a validação de métodos e modelos objetivos para o cálculo da qualidade em sequências áudio-visuais. Atualmente, o projeto encontra-se numa fase de planejamento. Com exceção das componentes de áudio, muito do conteúdo disponível coincide com o material utilizado nos projetos antigos de qualidade de vídeo (sem áudio) do VQEG e, portanto, precisa ser reavaliado.

Capítulo 4

Experimentos Subjetivos

Neste capitulo, são descritos os três experimentos subjetivos realizados neste trabalho. São detalhados a metodologia experimental, as condições físicas e os resultados experimentais. A Tabela 4.1 contém um resumo dos experimentos realizados. No Experimento I, avaliadores humanos pontuaram a qualidade de sequências de vídeo (sem a componente de áudio) comprimidas a diferentes taxas de bit. No Experimento II, a qualidade de sequências de áudio (sem a componente de vídeo) comprimidas a diferentes taxas de bit foi avaliada. Finalmente, no Experimento III, os avaliadores pontuaram a qualidade de sequências áudio-visuais. As componentes de áudio e vídeo destas sequências foram comprimidas de forma independente a diferentes taxas de bit.

4.1 Procedimentos Experimentais

Nesta seção, são descritos os procedimentos utilizados para a execução dos experimentos. São detalhadas as condições físicas, a seleção de conteúdo, o processo de geração das sequências de teste, a metodologia experimental e os métodos estatísticos utilizados.

4.1.1 Condições Físicas

Os experimentos foram projetados de acordo com as recomendações do ITU publicadas nos documentos ITU-R BT-500-8 e ITU-R P.911 [42, 44]. Estes documentos incluem recomendações para o ambiente onde serão realizados os experimentos e a instalação do equipamento. Neste trabalho, os experimentos foram realizados com dois avaliadores de cada vez. Foram utilizados dois computadores, dois monitores LCD (com características

Tabela 4.1: Especificações dos Experimentos Subjetivos I, II e III.

Experimento	Componente	Bitrate	Codec	Número Sequências de Teste	Número Avaliadores	Duração (Semanas)
Experimento I	Vídeo	30, 2, 1, 0.8 MB/s	H.264	30	16	2
Experimento II	Áudio	128, 96, 48 KB/s	MPEG-1 Layer-3	24	16	2
Experimento III	Áudio+Vídeo	Áudio: 128, 96, 48 KB/s Vídeo: 30, 2, 1, 0.8 MB/s	H.264 MPEG-1 Layer-3	78	17	3

Tabela 4.2: Especificações técnicas dos monitores e fones de ouvido utilizados nos experimentos.

Equipamento	Especificações
Monitor 1	Samsung SyncMaster P2370 Resolution: 1,920x1,080; Pixel-response rate: 2ms; Contrast ratio: 1,000:1; Brightness: 250cd/m2
Monitor 2	Samsung SyncMaster P2270 Resolution: 1,920x1,080; Pixel-response rate: 2ms; Contrast ratio: 1,000:1; Brightness: 250cd/m2
Fones de ouvido	Philips SHL580028 Headband Headphones Sensitivity: 106dB; Maximum power input: 50mW; Frequency response: 10–28,000Hz; Speaker diameter: 40mm.

similares) e dois pares de fones de ouvido. O contraste dinâmico dos monitores foi desligado e os valores do contraste e do brilho foram fixados em 100 e 50 respetivamente. As especificações dos monitores e dos fones de ouvido são apresentadas na Tabela 4.2.

Seguindo com as recomendações da ITU-R para o ambiente experimental, foi escolhido o estúdio de gravação do Núcleo de Multimídia e Internet (NMI). O NMI está vinculado ao Departamento de Engenharia (ENE) da Faculdade de Tecnologia da Universidade de Brasília (UnB). Esta sala (à prova de som) cumpre maioritariamente com os requisitos de ambiente listados na recomendação da ITU-R [42]. As luzes da sala foram desligadas para evitar que qualquer luz fosse refletida nos monitores e luzes laterais.

Os dois computadores e os dois monitores foram colocados numa mesa na sala. Duas cadeiras foram colocadas em frente dos monitores e duas lâmpadas foram instaladas perto dos teclados para ajudar aos participantes inserir suas respostas. A intensidade e direção das lâmpadas foram fixadas de modo a não afetar as condições visuais dos participantes. Cada participante estava sentado em frente do monitor, centrado ou ligeiramente abaixo da altura dos olhos para a maioria dos indivíduos. A distância entre os olhos do participante e o monitor foi definida em três vezes a altura da tela do monitor, conforme ilustrado na Figura 4.1. A escolha desta medida é conservadora. Segundo as diretrizes do ITU-R [42], esta distância deve satisfazer as regras para o tamanho da tela do monitor e a Distância de Visualização Preferida (*Prefered Viewing Distance* – PVD).

O *software* "Presentation" distribuído pelo *Neurobehavioral Systems Inc.* foi utilizado como plataforma para executar os experimentos e coletar os dados subjetivos. Nossos avaliadores foram estudantes de graduação e pós-graduação dos Departamentos de Ciências da Computação e Engenharia Elétrica da Universidade de Brasília. Os participantes são considerados não-especialistas em tarefas de qualidade de vídeo digital. Não foram feitos testes de acuidade de visão nos participantes, mas foi requisitado que eles usassem óculos ou lentes de contato no caso de ser necessária correção visual. É recomendada a participação de um mínimo de 15 pessoas para cada um dos experimentos subjetivos. Com estas considerações, foram recrutadas 16 pessoas para participar do Experimento I, outras 16 para o Experimento II e, finalmente, um total de 17 pessoas para o Experimento III. É bom destacar que nenhum dos participantes foi recrutado para mais de um dos experimentos.

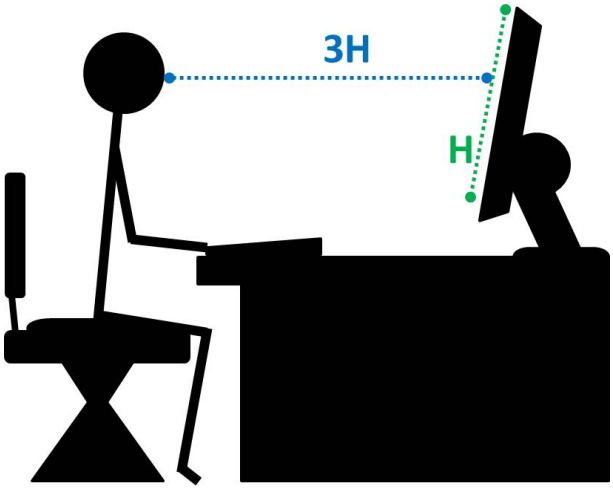


Figura 4.1: Distância entre os olhos do participante e o monitor.

4.1.2 Seleção do Conteúdo

Para seleção do conteúdo das sequências áudio-visuais utilizadas nos experimentos, seguiu-se as recomendações do VQEG (Video Quality Experts Group) [52]. Estas recomendações incluem, por exemplo, a escolha de um conteúdo de vídeo diversificado (filmes, esportes, música, publicidade, animação) e especificações técnicas desejáveis para as sequências de vídeo (resolução espacial, resolução temporal, amostragem da cor, duração). Os mesmos princípios são considerados para escolher o conteúdo de áudio, onde visamos escolher um conteúdo de áudio variado (diálogo, música e som ambiental).

As sequências utilizadas neste trabalho foram obtidas do site The Consumer Digital Video Library (CDVL, http://www.cdvl.org/). As características técnicas selecionadas, com base nestas recomendações são: (1) Resolução Espacial de 1280×720, (2) Resolução Temporal de 30 quadros por segundo (*frames per second*, fps), (3) Amostragem de Cor 4:2:0, e (4) Tempo de duração de 8 segundos.

No total, 6 sequências foram utilizadas na sessão principal do experimento. Quadros representativos das 6 sequências utilizadas na sessão principal do experimento são apresentadas na Figura 4.2.

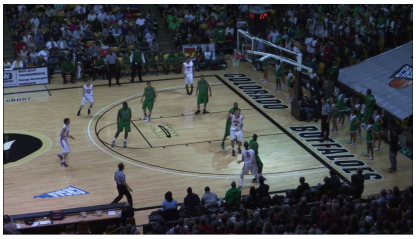
O relatório do VQEG [52] recomenda ainda que o conjunto de sequências utilizadas nos experimentos deve ter uma boa distribuição de atividade espacial (*Spatial Information (SI)*) e temporal (*Temporal Information (TI)*) [52]. Sequências com valores altos deste tipo de características representam cenas mais complexas (alta atividade espacial) ou cenas com um grau de movimentação alto (alta atividade temporal). Na Figura 4.3 são apresentados os valores de atividade espacial e temporal (computados conforme definido por Ostaszewska em [12]) para as seis sequências do experimento. Pode-se observar que a sequência 'Reporter' apresenta o maior valor de atividade temporal. No entanto, o valor de atividade espacial para esta mesma sequência foi o menor valor registrado. A sequência 'Music' registrou valores altos na sua atividade temporal e espacial, enquanto que a sequência 'Park Run' obteve valores baixos para a sua atividade temporal e espacial.

Do mesmo modo, foi utilizado o algoritmo proposto por Giannakopoulos [53] para descrever o conteúdo do *áudio* das seis sequências áudio-visuais. Este algoritmo divide o

(a)‘Boxer’

(b) ‘Park Run’

(c) ‘Crowd Run’



(d) ‘Basketball’

(e) ‘Music’

(f) ‘Reporter’

Figura 4.2: Quadros representativos das seis sequências utilizadas nos experimentos.

fluxo de áudio em diversos segmentos não sobre-postos. Em seguida, o algoritmo classifica cada um dos segmentos em uma das seguintes classes:

- Diálogo (*Speech*)
- Musica (*Music*)
- Sons ambientais baixos: vento, chuva (*Others1*)
- Sons com mudanças abruptas (*Others2*)
- Sons fortes: maquinas e carros (*Others3*)
- Tiros (*Gunshots*)
- Gritos (*Screams*)

Na Figura 4.4 é apresentada a classificação da componente de áudio das seis sequências utilizadas no experimento.

Essa classificação do áudio permite compreender melhor a componente de áudio das sequências utilizadas. O vídeo ‘Reporter’, por exemplo, foi classificado como Speech e Others1. Já o vídeo ‘Park Run’ foi classificado unicamente como música. Por outro lado, a sequência ‘Music’ foi classificada como música, Others2 e Screams. Finalmente, os vídeos ‘Basketball’ e ‘Crowd Run’ foram classificados como Others1.

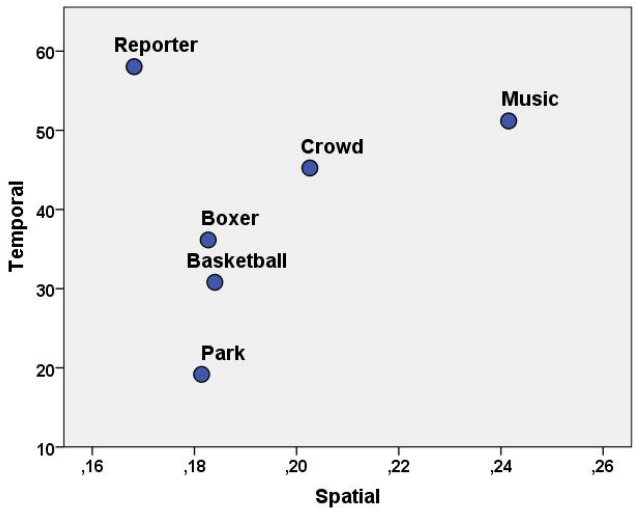


Figura 4.3: Atividade Espacial (SI) e Atividade Temporal (TI) (computados conforme definido por Ostaszewska em [12]) para as sequências do experimento.

4.1.3 Geração das Sequências

A plataforma FFMPEG foi utilizada para processar as seis sequências utilizadas no experimento. Foram escolhidas sequências de 8 segundos cada uma, já que 8 segundos pode ser considerado um intervalo de tempo aceitável para uma pessoa fazer um julgamento da qualidade do sinal. Experimentos semelhantes foram feitos utilizando sequências de 6 segundos e apresentaram bons resultados [26].

No Experimento I, cada uma das sequências originais de vídeo (sem áudio) foi comprimida utilizando o codec H.264. Quatro valores diferentes de taxa de bits de vídeo **(vb)** foram utilizados: 30, 2, 1, 0.8 Mbps. Este esquema de processamento gerou 6 (sequências originais) × 4 (valores de taxa de bits) = 24 sequências de teste. Estas 24 sequências, juntamente com as 6 sequências originais, resultam nas 30 sequências de teste do Experimento I. Estas 30 sequências foram apresentadas durante a sessão principal do Experimento I.

Para o Experimento II, só a componente de áudio foi utilizada. Os sinais foram comprimidos utilizando o codec MPEG-1 *layer* 3. Três valores de taxa de bits de áudio **(va)** foram utilizados: 128, 48, e 32 Kbps. Este esquema de processamento gerou 6 (sequências originais) × 3 (valores de taxa de bits) = 18 sequências de teste. Estas 18 sequências, juntamente com as 6 sequências originais, resultaram nas 24 sequências de teste do Experimento II. Estas 24 sequências foram exibidas na sessão principal do Experimento II.

Finalmente, no Experimento III, ambas componentes (áudio e vídeo) foram comprimidos. As componentes de vídeo foram processadas utilizando o mesmo codec H.264 e as mesmas taxas de bits do Experimento I. As componentes de áudio foram comprimidas com o codec MPEG-1 *layer* 3. Foram utilizadas as mesmas taxas de bits do Experimento II (**va** e **vb**). Levando em consideração os 3 valores de taxa de bits de áudio e os 4 valores de taxas de bits de vídeo (3 taxa de bits de áudio × 4 taxa de bits de vídeo) para as 6 sequências, são geradas um total de 3×4×6 = 72 sequências de teste. Estas 72 sequên-

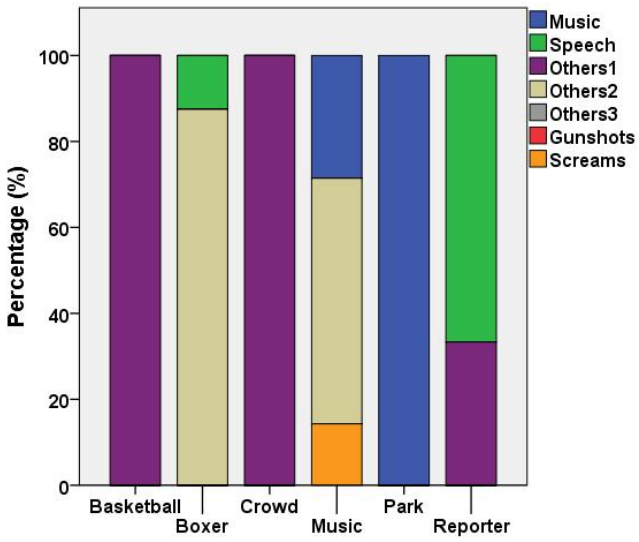


Figura 4.4: Classificação da componente de áudio das seis sequências utilizadas nos experimentos.

cias, juntamente com as 6 sequências originais, completam as 78 sequências experimentais apresentadas durante a sessão principal do Experimento III (Ver Tabela 4.1).

4.1.4 Metodologia Experimental

A metodologia *Absolute Category Rating with Hidden Reference* (ACR-HR) foi utilizada em todos os três experimentos [54, 44]. Nesta metodologia, duas sequências são apresentadas em cada ensaio, onde cada uma das duas sequências possui o mesmo conteúdo. Destas duas sequências, uma delas é a referência (sequência original) e a outra é a sequência de teste (sequência processada). Os avaliadores, nesta metodologia, não sabem qual das sequências é a referência e qual é a processada, isto se deve a apresentação aleatória das sequências. Os avaliadores são instruídos a pontuar cada uma das duas sequências apresentadas em cada ensaio.

Se bem a metodologia ACR com referencia oculta (ver secção 3.1) indica que as duas pontuações para ambas as sequências (teste e a referência oculta) devem de ser processadas para calcular o DMOS, neste trabalho foram levadas em conta apenas as pontuações das sequências de teste, é dizer, foi calculado o MOS utilizando as pontuações das sequencias de teste somente. A decisão de descartar as pontuações das sequências de referência oculta foi tomada devido ao próximo que ficaram estas pontuações dos valores de qualidade máxima do experimento. Acredita-se que a não inclusão destas pontuações não representa maiores mudanças nos resultados finais.

Antes de começar o experimento, pediu-se ao avaliador para ficar na posição correta e na distância definida, segundo as recomendações descritas previamente. A lâmpada é ligada e o software para a apresentação é executado. O *software* "Presentation" distribuído pelo *Neurobehavioral Systems Inc.* foi utilizado como plataforma para executar os experimentos e coletar os dados subjetivos. Foi necessário elaborar três scripts para controlar cada um dos três experimentos.

Para os experimentos que incluíam a componente de áudio (Experimentos II e III) os participantes utilizaram os fones de ouvido. Uma breve explicação verbal foi dada pelo experimentador aos participantes, durante o experimento. As mesmas instruções são apresentadas na tela do monitor, garantindo o total entendimento dos participantes das tarefas experimentais.

Os experimentos foram divididos em três sessões: (1) Treinamento, (2) Ensaio e (3) Sessão Principal. Na sessão de Treinamento, são apresentados um conjunto de sequências originais e as suas correspondentes sequências processadas aos avaliadores. O objetivo desta sessão é familiarizar os participantes com o intervalo de qualidade utilizado no experimento. Em outras palavras, após esta sessão, os participantes devem conseguir distinguir melhor os níveis de qualidade de cada sequência.

Na sessão de Ensaio, os participantes pontuam um conjunto de sequências de teste. O objetivo desta sessão é apresentar aos usuários as sequências, da mesma maneira como elas serão apresentadas na sessão principal. O objetivo é treinar e familiarizar o participante na inserção de respostas. A ITU-R recomenda descartar os primeiros cinco resultados obtidos num experimento subjetivo, já que é assumido que um participante não consegue pontuar corretamente as primeiras sequências [54]. Mediante a inclusão desta sessão de ensaio, evita-se o descarte de pontuações na sessão principal. Foram incluídos 5 ensaios. Não foram computadas as pontuações obtidas durante esta sessão para a análise estatística.

Na sessão Principal, são apresentados aos participantes conjuntos de duplas sequências. Estas sequências são apresentadas de forma aleatória. Após a apresentação de duas sequências, os participantes pontuam cada uma delas utilizando a escala de pontuação apresentada na tela. A escala de pontuação vai de '0' a '10', como ilustrado na Figura 4.5. Para evitar a fadiga visual dos participantes, esta sessão foi dividida em duas partes.

É importante enfatizar que, apesar da escala utilizada nos experimentos ter um intervalo entre 0 e 10, na fase de análise os valores coletados foram convertidos para o intervalo entre 0 e 100. Logo, apesar de continuarmos utilizando a nomenclatura MOS, os valores estão convertidos em relação aos dados coletados.

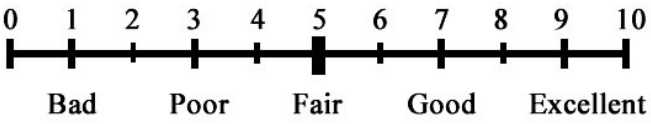


Figura 4.5: Escala de pontuação contínua utilizada nos experimentos.

4.1.5 Métodos Estatísticos de Análise

A pontuação dada pelos participantes para qualquer uma das sequências do experimento são denominadas de pontuações subjetivas. Estes dados subjetivos são processados através do cálculo do valor de opinião média (*Mean Opinion Score*, MOS). Para calcular o MOS, é calculada a média dos valores das pontuações de todos os participantes para

cada uma das sequências de teste:

$$\text{MOS} = \overline{S} = \frac{1}{L} \cdot \sum_{i=0}^{L} S(i), \tag{4.1}$$

no qual o valor $S(i)$ é a pontuação reportada pelo i-ésimo participante e L é o número total de participantes que pontuaram uma sequência. Foram calculados também o desvio padrão das pontuações:

$$STD = \left(\frac{1}{L} \cdot \sum_{i=0}^{L} \left(S(i) - \overline{S} \right)^2 \right)^{1/2}, \tag{4.2}$$

e o erro padrão interno do $\overline{S}$:

$$\overline{STD} = \frac{STD}{\sqrt{L}}. \tag{4.3}$$

O intervalo de confiança para o MOS de uma sequência de teste é dado por:

$$\overline{S} \pm \ t_{L,\alpha/2} \ \overline{STD} \tag{4.4}$$

no qual $t_{L,\alpha/2}$ corresponde ao coeficiente t, da distribuição t de Student.

4.2 Resultados Experimentais

Nesta seção são apresentados os resultados obtidos nos três experimentos subjetivos realizados neste trabalho.

4.2.1 Experimento I

Neste experimento foram apresentadas sequências de vídeo sem a componente de áudio. Estas sequências foram comprimidas em diferentes taxas de bit (vb) utilizando o codec H.264. Participaram deste experimento um total de 16 pessoas que avaliaram as sequências gerando um único valor de MOS_v. A Figura 4.6 apresenta os valores de MOS_v obtidos para o Experimento I versus os quatro valores de taxa de bit (vb1 = 800 Kbps, vb2=1 Mbps, vb3=2 Mbps, vb4=30 Mbps).

Observa-se na Figura 4.6 que o MOS_v aumenta à medida que aumenta a taxa de bits. Isto mostra que os participantes do Experimento I foram capazes de perceber as variações na taxa de bits de vídeo (vb), as quais resultam em variações na qualidade de vídeo percebida (MOS_v).

Os vídeos 'Basketball' e 'Park Run', que têm pouca atividade temporal e espacial, apresentaram, em média, os mais baixos valores para MOS_v. No entanto, os vídeos 'Music' e 'Crowd Run', com uma alta atividade temporal e espacial, apresentaram, em média, os mais altos valores para MOS_v.

Estes resultados estão em concordância com os resultados disponíveis na literatura [52] que relatam que, em cenas complexas (alta atividade temporal e espacial) os artefatos são

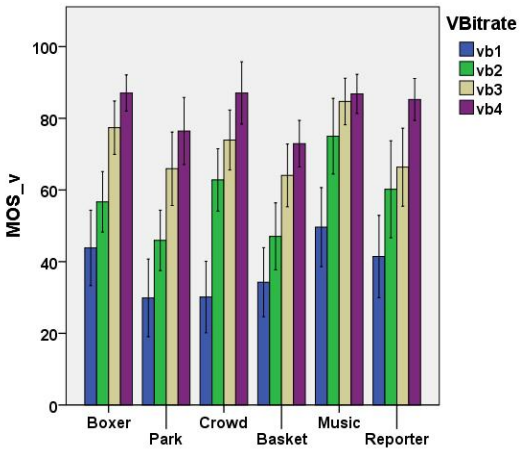


Figura 4.6: Experimento I: (MOS_v) versus taxa de bits de vídeo (vb1 = 800 Kbps, vb2=1 Mbps, vb3=2 Mbps, vb4=30 Mbps).

mais difíceis de perceber. É por isto que este tipo de cena é percebida como tendo uma qualidade mais alta.

4.2.2 Experimento II

No segundo experimento, foram apresentadas sequências de áudio sem a componente de vídeo. Estas sequências foram comprimidas utilizando três diferentes taxas de bits (ab), utilizando o codec MPEG-1 *layer* 3. No total, 16 participantes avaliaram a qualidade de áudio, gerando um valor de MOS_a para cada uma das sequências deste experimento.

A Figura 4.7 apresenta os valores de MOS_a obtidos no Experimento II versus os três valores de taxa de bit utilizados (ab1=48 Kbps, ab2=96 Kbps, ab3=128 Kbps). É possível observar na Figura 4.7, como aconteceu com o Experimento I, que o valor do MOS_a aumenta à medida que aumenta a taxa de bits.

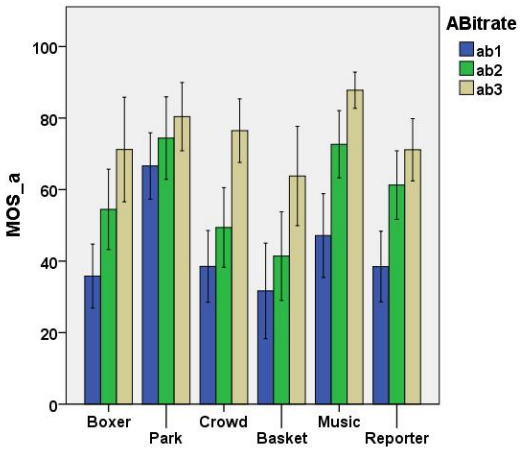


Figura 4.7: Experimento II: (MOS_a) versus taxa de bits de áudio (ab1=48 Kbps, ab2=96 Kbps, ab3=128 Kbps).

Segundo os resultados apresentados na Figura 4.7, a sequência de áudio 'Basketball', que foi classificada como Others1 (sons ambientais, ver secção 4.1.2), apresentou o MOS_a mais baixo. Por outra lado, sequências como 'Music' e 'Park Run' (classificadas como música, gritos e Others2, ver secção 4.1.2) apresentaram os valores mais altos de MOS_a. Estes resultados pareceam indicar que a qualidade de áudio em sequências que contêm tipos de sons mais variados foi menos afetada pela compressão de áudio.

4.2.3 Experimento III

No Experimento III, ambas as componentes de áudio e vídeo foram incluídas. Três taxas de bits de áudio e quatro de vídeo foram utilizadas. No total, 17 pessoas participaram deste experimento, gerando um único valor MOSav para cada uma das sequências áudio-visuais apresentadas.

A Figura 4.8 apresenta como os valores MOS_{av} mudam entre os quatro valores de taxas de bits do vídeo (vb). Esta mudança é observada para cada uma das seis sequências originais e os valores de taxa de bits de áudio (ab). Pode ser observado que os valores de MOS_{av} aumentam à medida que aumentam os valores das taxas de bits de vídeo para as sequências 'Park', 'Crowd' e 'Reporter' (como foi observado nos experimentos I e II). No entanto que sequências como 'Boxer', 'Basketball' e 'Music' não apresentam o mesmo comportamento. Isto pode ser observado na Figura 4.8.

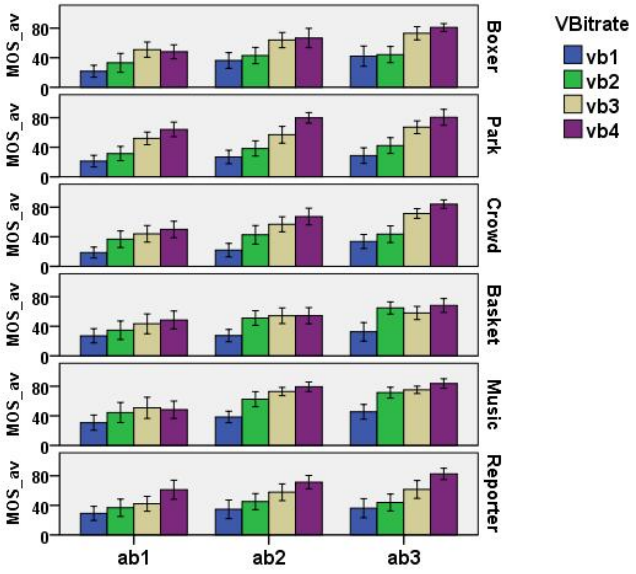


Figura 4.8: Experimento III: (MOS_{av}) versus taxa de bits de áudio (vb1 = 800 Kbps, vb2=1 Mbps, vb3=2 Mbps, vb4=30 Mbps, ab1=48 Kbps, ab2=96 Kbps, ab3=128 Kbps).

A Figura 4.9 apresenta como os valores MOS_{av} mudam entre os três valores de taxas de bits do áudio (ab). Novamente, pode ser observado o mesmo comportamento dos valores MOS_{av} em relação aos valores de taxas de bits, neste caso, de áudio. Observa-se novamente diferenças no comportamento das inclinações causadas pelo incremento das

taxas de bits de áudio. Porém, em geral, estas diferenças são menores quando comparadas com as inclinações observadas na Figura 4.8. Em outras palavras, a compressão do vídeo (vb) tem um maior impacto na qualidade global do que a do áudio (ab).

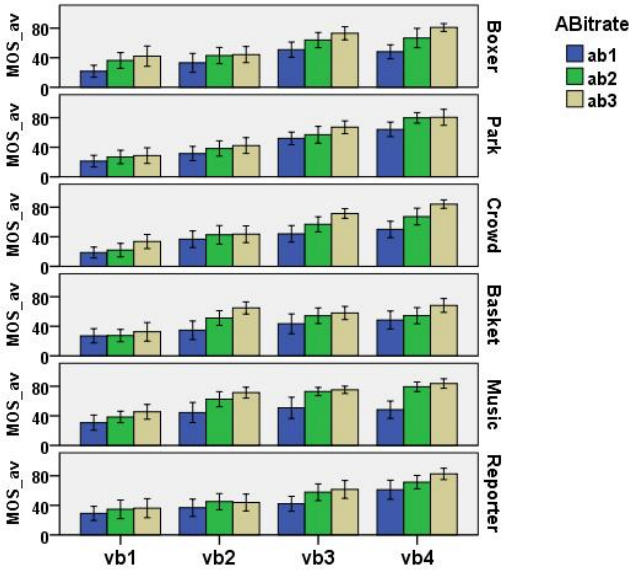
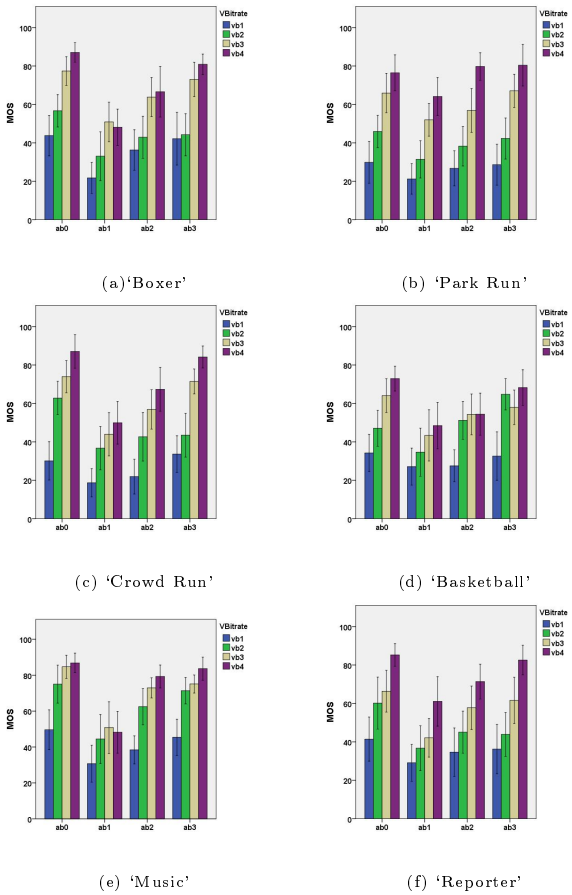


Figura 4.9: Experimento III: (MOS_{av}) versus taxa de bits de vídeo (vb1 = 800 Kbps, vb2=1 Mbps, vb3=2 Mbps, vb4=30 Mbps, ab1=48 Kbps, ab2=96 Kbps, ab3=128 Kbps).

4.2.4 Comparações e Discussão

A última análise consiste em entender melhor a contribuição da componente de áudio para a qualidade global das sequências áudio-visuais. Com este objetivo, na Figura 4.10 são apresentados os dados dos experimentos I e III. Neste gráficos, os dados do experimento I (experimento sem a componente de áudio) são apresentados como 'ab0' (primeiras quatro colunas na esquerda de cada gráfico). Pode-se observar que em comparação às sequências com áudio os avaliadores pontuaram as sequências de vídeo sem áudio (ab0) com maiores valores de MOS na maioria dos casos (exceto na sequência 'Park Run'). A Figura 4.10 mostra que a ausência da componente de áudio influencia positivamente a qualidade áudio-visual. Este tipo de resultado sugere que a componente de áudio pode ser um fator distrator para os avaliadores durante o teste subjetivo, já que neste caso os avaliadores precisam prestar atenção em outras características.

Pode-se concluir que as características do conteúdo das cenas, no caso da componente de vídeo, são importantes ao determinar o MOS. Isto comprova que existe uma correlação entre a atividade temporal e espacial e os valores MOS obtidos nos experimentos. Mediante a análise do conteúdo de áudio nas sequências utilizadas nos experimentos, pode se pensar que as sequências com conteúdo do tipo Others1 (sons ambientais) foram mais sensíveis à compressão do que outras sequências com outros tipos de conteúdo. Esta afirmação ainda pode ser avaliada mediante a realização de novos experimentos utilizando



(a)'Boxer' (b) 'Park Run'

(c) 'Crowd Run' (d) 'Basketball'

(e) 'Music' (f) 'Reporter'

Figura 4.10: Experimento I e III: MOS_v e MOS_{av} versus taxas de bits de áudio (e vídeo) (vb1 = 800 Kbps, vb2=1 Mbps, vb3=2 Mbps, vb4=30 Mbps, ab1=48 Kbps, ab2=96 Kbps, ab3=128 Kbps).

novos conteúdos de áudio. Foram analisados os resultados de áudio e vídeo separadamente e foi possível concluir que a compressão da componente de vídeo tem um impacto maior na qualidade áudio-visual. Finalmente, comparando os resultados dos Experimentos I e III, foi possível perceber que a componente de áudio pode atuar como um fator de distrator durante a avaliação subjetiva, diminuindo o MOS.

Capítulo 5

Modelos de Qualidade áudio-visual

Neste capítulo, são descritos os modelos áudio-visuais propostos neste trabalho. São apresentados quatro modelos subjetivos, três modelos objetivos com-referência (FR) e três modelos objetivos sem-referência (NR). Além de uma descrição de cada modelo, o capítulo contém uma analise dos resultados obtidos para os dados obtidos através dos experimentos.

5.1 Modelo Subjetivo

Para obter o modelo de qualidade áudio-visual, a qualidade objetiva de vídeo e a qualidade objetiva de áudio, respectivamente, das sequências dos Experimentos I e II foram obtidas utilizando as métricas VQM e SESQA: Qv e Qa. Quando comparadas com os resultados subjetivos dos Experimentos I e II, estas medidas objetivas da qualidade do vídeo e do áudio apresentaram coeficientes de correlação iguais a 0.82 e 0.94, respectivamente. Estes resultados serviram como base para construir os modelos áudio-visuais.

Com o objetivo de se obter um modelo de qualidade áudio-visual, analisaram-se os dados obtidos nos três experimentos descritos no Capitulo 3. Mais especificamente, se utilizou uma análise de regressão para analisar se é possível estimar a qualidade subjetiva áudio-visual (Experimento III, MOS_{av}) utilizando a qualidade subjetiva de áudio (Experimento II, MOS_{a}) e a qualidade subjetiva de vídeo (Experimento I, MOS_{v}). Para isto, testaram-se vários modelos de combinação para obter MOS_{av} a partir de MOS_{v} e MOS_{a}.

O primeiro modelo testado foi uma função linear descrita pela seguinte equação:

$$\mathrm{PrMOS}_1 = \alpha_1 \cdot \mathrm{MOS}_v + \beta_1 \cdot \mathrm{MOS}_a + \gamma_1. \tag{5.1}$$

no qual PrMOS_1 é o valor de MOS estimado para a sequência áudio-visual utilizando o modelo linear. O ajuste retornou os coeficientes escalares $\alpha_1 = 0.76$, $\beta_1 = 0.41$, e $\gamma_1 = -21.92$. Os coeficientes de correlação de Pearson e Spearman [55] da estimação obtidos com este modelo são iguais a 0.9110 e 0.9173, respectivamente. Na Figura 5.1, é apresentado o gráfico do valor estimado de MOS para as sequências áudio-visuais do Experimento III, PrMOS_1, versus o MOS_{av} obtido neste experimento.

Como a componente de áudio teve uma importância muito inferior à componente de vídeo no modelo linear ($\alpha_1 > \beta_1$), testaram-se um segundo modelo no qual incluiu-se apenas a componente de vídeo. O objetivo foi estimar a influência da componente de

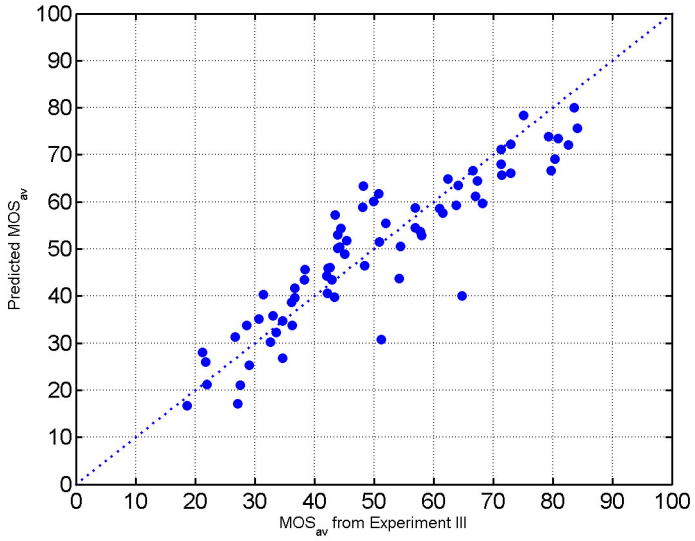


Figura 5.1: Valor estimado do MOS_{av} utilizando o modelo linear, versus MOS_{av} obtido no Experimento III.

vídeo na qualidade áudio-visual global e, portanto, fazer uma análise da eficiência do uso de métricas de vídeo para aferição da qualidade áudio-visual. O segundo modelo testado tem a seguinte equação:

$$\mathrm{PrMOS}_2 = \alpha_2 \cdot \mathrm{MOS}_v + \gamma_2. \tag{5.2}$$

no qual PrMOS_2 é o valor de MOS estimado para a sequência áudio-visual utilizando o modelo linear simplificado. O ajuste para este modelo retornou um coeficiente α_2 igual a 0.78 e uma constante aditiva γ_2 igual a 1.52. Os valores dos coeficientes de correlação Pearson e Spearman obtidos foram 0.8191 e 0.8378, respectivamente. Na Figura 5.2, é apresentado o gráfico do valor estimado de MOS. Embora este modelo não tenha um desempenho tão bom quanto o anterior, a sua boa correlação evidencia a importância da componente de vídeo na predição da qualidade áudio-visual.

O terceiro modelo testado foi a função ponderada de Minkowski [56], descrita pela equação:

$$\mathrm{PrMOS}_3 = (\alpha_3 \cdot \mathrm{MOS}_v^{p_1} + \beta_2 \cdot \mathrm{MOS}_a^{p_1})^{\frac{1}{p_1}}. \tag{5.3}$$

no qual PrMOS_3 é o valor de MOS estimado para a sequência áudio-visual utilizando o modelo ponderado de Minkowski. Observe que, se o valor de p_1 fosse igual a 1, o modelo se comportaria como uma função linear simples. O ajuste para este modelo retornou um valor igual a 0.0001 para o exponente p_1. Os valores dos coeficientes escalares α_3 (vídeo) e β_2 (áudio) foram iguais a 0.7024 e 0.2976, respectivamente. O coeficiente de correlação Pearson obtido foi 0.9197 e o coeficiente de correlação Spearman foi 0.9267. É possível perceber que o modelo Minkowski ponderado apresenta um maior poder preditivo, em comparação com os outros dois modelos testados. Na Figura 5.3, é apresentado o gráfico do valor estimado de MOS para as sequências áudio-visuais do Experimento III, PrMOS_3, versus o MOS_{av} obtido neste experimento.

Finalmente, o quarto modelo testado foi uma função Produto de Potências, descrita pela equação:

$$\mathrm{PrMOS}_4 = (\gamma_3 + \alpha_4 \cdot \mathrm{Qv}^{p_2} \cdot \mathrm{Qa}^{p_3}), \tag{5.4}$$

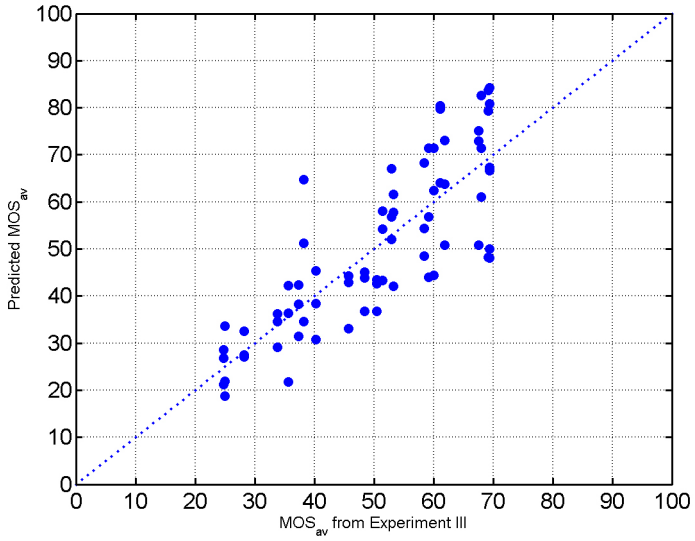


Figura 5.2: Valor estimado do MOS_{av} utilizando o modelo linear simplificado, versus MOS_{av} obtido no Experimento III.

no qual PrMOS_4 corresponde à qualidade áudio-visual estimada pelo modelo. O ajuste para este modelo retornou exponentes $p_2 = 1.3213$ e $p_3 = 0.6533$, correspondendo às componentes de vídeo e áudio, respectivamente. Também são obtidos os coeficientes escalares $\alpha_4 = -0.0109$ e $\gamma_3 = -12.9734$. Para este ajuste, os valores dos coeficientes de correlação Pearson e Spearman obtidos foram 0.9285 e 0.9270, respectivamente. Na Figura 5.4, é apresentado o gráfico dos valores de PrMOS_4 estimados para as sequências áudio-visuais do Experimento III versus o MOS_{av} obtidos neste experimento.

Com base nestes resultados, é possível observar a importante contribuição da componente de vídeo na predição da qualidade áudio-visual dos três modelos descritos. Observe que, tanto para o modelo linear como para o modelo ponderado de Minkowski, os coeficientes correspondentes à componente de vídeo (α_1 e α_2) são maiores do que os coeficientes correspondentes à componente de áudio (β_1 e β_2). Mas, embora menos importante, a componente de áudio não pode ser ignorada. Isto é evidenciado pela queda no desempenho do modelo linear simplificado.

5.2 Modelo Objetivo Com-Referência

A fim de ser projetados modelos objetivos com referência, foram escolhidas duas métricas objetivas de qualidade: (1) uma métrica de qualidade de áudio e (2) uma métrica de qualidade de vídeo. A métrica de áudio escolhida foi a SESQA (*Single Ended Speech Quality Metric*, ver Seção 3.2). A métrica qualidade de vídeo (com-referência) escolhida foi a VQM (*Video Quality Metric*, ver Seção 3.2). Vale a pena mencionar que, a métrica SESQA foi proposta originalmente para a avaliação da qualidade de sinais em aplicações telefônicas. No intuito de adaptar esta métrica para o cálculo da qualidade de sinais de áudio (diálogo, música sons genéricos, etc.), a métrica foi alterada. Ao invés de utilizar todos os 51 parâmetros considerados no algoritmo original (ver Anexo 1), foram selecionados apenas 17 destes parâmetros. A escolha destes 17 parâmetros foi feita mediante

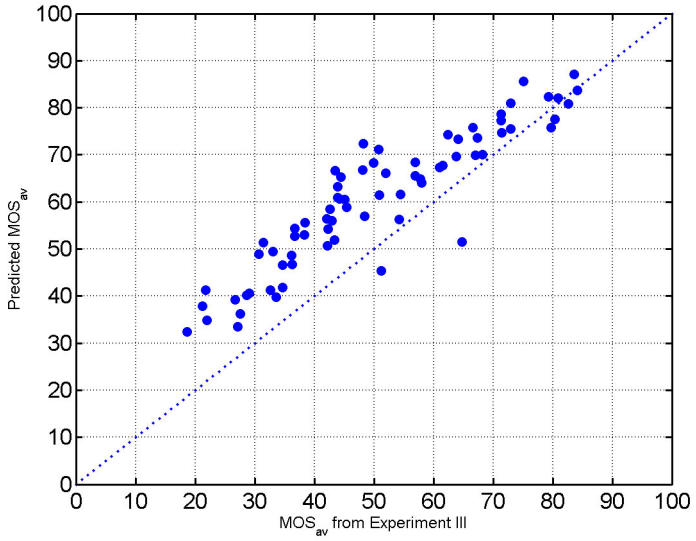


Figura 5.3: MOS_{av} estimado com o modelo Minkowski versus MOS_{av} obtido no Experimento III.

uma regressão linear simples, utilizou-se a ferramenta WEKA 3.8 para obter o modelo de qualidade de áudio. Foram introduzidos primeiro os 51 parâmetros do algoritmo original, mediante a ferramenta de classificação de regressão linear do WEKA foi gerado um modelo para o cálculo da qualidade de áudio que utiliza apenas 17 dos 51 parâmetros. Estes 17 parâmetros mostraram resultados coerentes na estimação da qualidade de sinais de áudio degradados. O modelo resultante e os 17 parâmetros são apresentados no Anexo II. O conteúdo das sequências testadas incluía musica, diálogos, explosões, sons ambientais, etc. Nenhuma outra alteração foi feita no algoritmo.

Nesta seção, são apresentados três novos modelos objetivos de qualidade áudio-visual baseados na combinação destas duas métricas.

Realizam-se uma análise de regressão utilizando os dados de qualidade objetiva de vídeo e áudio, (Qv e Qa), e os resultados subjetivos do Experimento III ($\mathrm{MOS}_{\mathrm{av}}$). De forma semelhante ao que foi realizado na seção anterior, testaram-se três modelos de combinação para obter $\mathrm{MOS}_{\mathrm{av}}$ a partir de Qv e Qa.

O primeiro modelo testado foi uma função linear simples, descrita pela seguinte equação:

$$\mathrm{Qav}_1 = \alpha_5 \cdot \mathrm{Qv} + \beta_3 \cdot \mathrm{Qa} + \gamma_4, \tag{5.5}$$

na qual Qav_1 corresponde à qualidade áudio-visual estimada, Qv representa o valor de qualidade de vídeo obtido com a métrica VQM e Qa o valor de qualidade de áudio obtido com a métrica SESQA. O ajuste retornou os coeficientes escalares α_5=0.45, β_3=0.48, para vídeo e áudio, respectivamente, e o coeficiente independente γ_4=-8.9275. Para este ajuste, os coeficientes de correlação Pearson e Spearman foram de 0.8472 e 0.8337, respectivamente. Na Figura 5.5, é apresentado o gráfico dos valores de Qav_1 estimados para as sequências áudio-visuais do Experimento III versus o MOS_{av} obtidos neste experimento.

O segundo modelo testado foi o modelo ponderado Minkowski dado pela seguinte equação:

$$\mathrm{Qav}_2 = (\alpha_6 \cdot \mathrm{Qv}^{p_4} + \beta_4 \cdot \mathrm{Qa}^{p_4})^{\frac{1}{p_4}}. \tag{5.6}$$

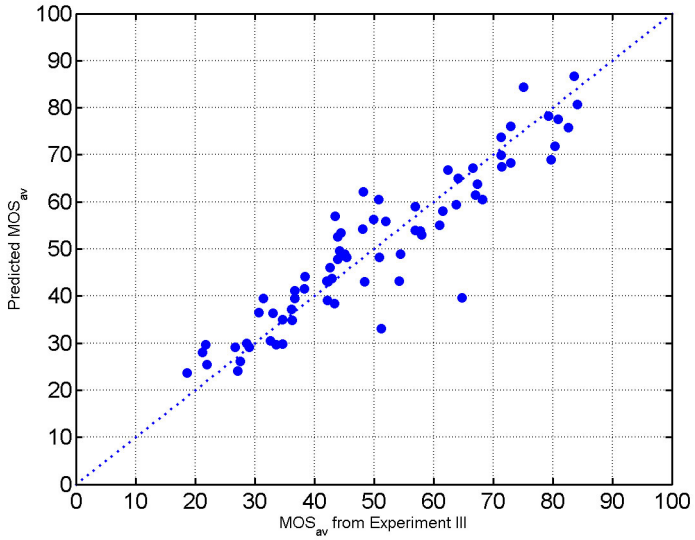


Figura 5.4: MOS$_{av}$ estimado com o modelo de Produto de Potências versus MOS$_{av}$ obtido no Experimento III.

no qual Qav$_2$ corresponde à qualidade áudio-visual estimada pelo modelo Minkowski ponderado. O ajuste para o modelo Minkowski retornou um exponente $p_4 = 0.9165$ e coeficientes escalares $\alpha_6 = 0.4148$, e $\beta_4 = 0.3999$, correspondentes às componentes de vídeo e áudio, respectivamente. Para este ajuste, o valor do coeficiente de correlação Pearson obtido foi 0.8448, enquanto que o coeficiente de correlação de Spearman foi igual a 0.8392. Na Figura 5.6, é apresentado o gráfico dos valores de Qav$_2$ estimados para as sequências áudio-visuais do Experimento III versus o MOS$_{av}$ obtidos neste experimento.

Finalmente, o terceiro modelo ajustado foi o modelo do produto de potências utilizado por Wang e Bovik em [51], descrito pela seguinte equação:

$$\text{Qav}_3 = (\gamma_5 + \alpha_7 \cdot \text{Qv}^{p_5} \cdot \text{Qa}^{p_6}), \tag{5.7}$$

no qual Qav$_3$ corresponde à qualidade áudio-visual estimada pelo modelo. O ajuste para este modelo retornou exponentes $p_5 = 1.5837$ e $p_6 = 0.9524$, correspondendo às componentes de vídeo e áudio, respectivamente. Também são obtidos os coeficientes escalares $\alpha_7 = 0.0006$ e $\gamma_5 = 26.9240$. Para este ajuste, os valores dos coeficientes de correlação Pearson e Spearman obtidos foram 0.8545 e 0.8384, respectivamente. Na Figura 5.7, é apresentado o gráfico dos valores de Qav$_3$ estimados para as sequências áudio-visuais do Experimento III versus o MOS$_{av}$ obtidos neste experimento.

Os resultados apresentados mostram que os modelos conseguem estimar os valores de qualidade áudio-visual de forma aceitável. Além disso, é interessante observar que, ao contrário do que acontece no caso dos modelos subjetivos, para modelos objetivos há um equilíbrio na contribuição das componentes de áudio e vídeo no que diz respeito à qualidade áudio-visual estimada. Embora seja uma diferença leve, o modelo produto de Potências (Qav$_3$) apresenta a melhor correlação dentre os três modelos.

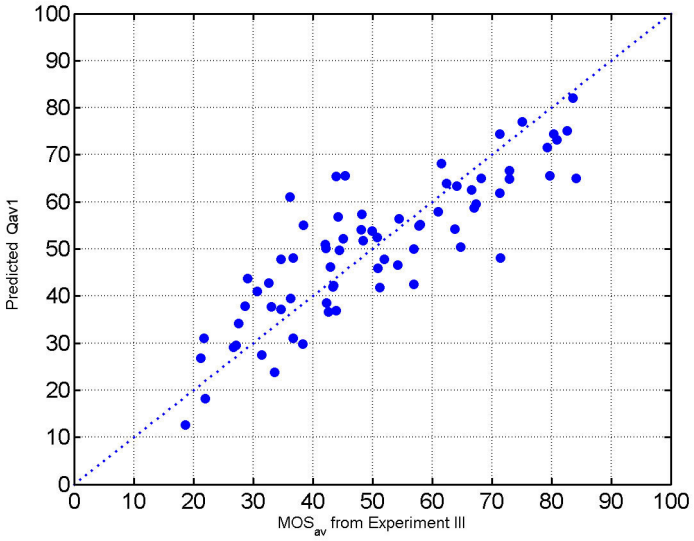


Figura 5.5: Qav_1 estimado com o modelo Linear versus MOS_{av} obtido no Experimento III.

5.3 Modelo Objetivo Sem-Referência

O objetivo desta seção é obter um modelo objetivo sem-referência para qualidade áudio-visual. Com este fim, a métrica de vídeo utilizada na seção anterior foi substituída, utilizando duas métricas de aferição de artefatos: Blocagem e Borrado (descritas na seção 3.2.1). Com o intuito de saber quais dos artefatos são mais fácies de identificar foram realizados experimentos piloto. Logo, foram escolhidos apenas os artefatos de Blocagem e Borrado para o cálculo da qualidade de vídeo. É bom mencionar que, para obter os valores de qualidade de vídeo foi construída uma métrica sem referência denominada: Métrica Combinatória baseada em Borrado e Blocagem. Duas métricas para o cálculo de artefatos de imagens foram combinadas. A métrica que calcula o nível de borrado foi a métrica proposta por Narvekar e Karam[50]. No enquanto, para calcular o nível de blocagem foi utilizada a métrica proposta por Wang e Bovik [51]. Estas métricas foram escolhidas com base aos experimentos realizados utilizando um grupo de métricas que calculam os níveis de blocagem e borrado [57] [58] [59] [60]. Depois de combinar vários dos resultados obtidos, as métricas propostas por Narvekar, Karam e Wang e Bovik foram escolhidas por apresentar a melhor correlação. Estas métricas são combinadas mediante uma função linear para obter um valor da qualidade do vídeo. Como a métrica de áudio SESQA é uma métrica sem-referência, usamos esta métrica para calcular a qualidade objetiva de áudio.

Da mesma forma que na seção anterior, para obter o modelo de qualidade áudio-visual, a qualidade objetiva de vídeo e a qualidade objetiva de áudio, respectivamente, das sequências dos Experimentos I e II foram obtidas utilizando as métricas de Borrado e Blocagem e SESQA: Qv e Qa. Quando comparadas com os resultados subjetivos dos Experimentos I, as medidas objetivas da qualidade do vídeo apresentaram coeficientes de correlação iguais a 0.61. Estes resultados serviram como base para construir os modelos áudio-visuais.

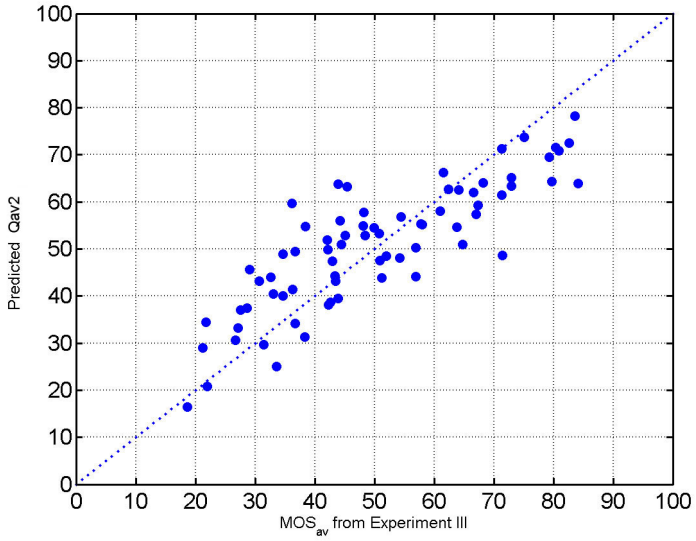


Figura 5.6: Qav_2 estimado com o modelo Minkowski versus MOS_{av} obtido no Experimento III.

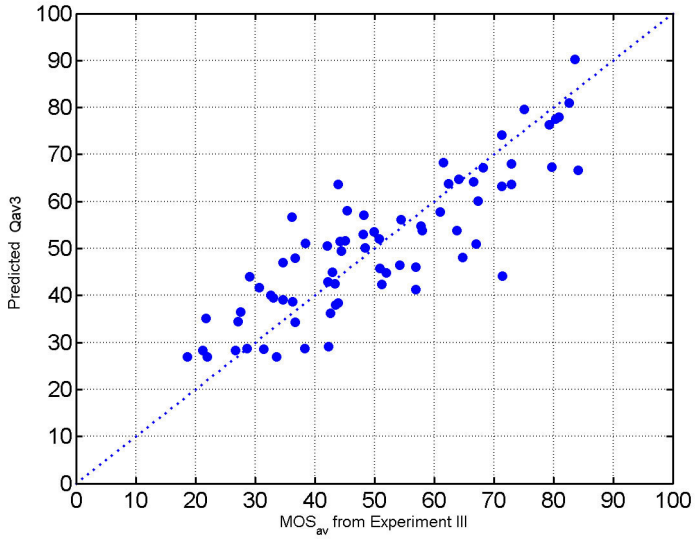


Figura 5.7: Qav_3 estimado com o modelo Minkowski versus MOS_{av} obtido no Experimento III.

Realizamos uma análise de regressão utilizando os dados de qualidade objetiva (sem-referência) de vídeo e áudio, (Qv e Qa), e os resultados subjetivos do Experimento III (MOS_{av}). De forma semelhante ao que foi realizado na seção anterior, testaram-se três modelos de combinação para obter MOS_{av} a partir de Qv e Qa.

O primeiro modelo foi um modelo linear simples, descrito pela seguinte equação:

$$\text{Qav}_4 = \alpha_8 \cdot \text{Qv} + \beta_5 \cdot \text{Qa} + \gamma_6, \tag{5.8}$$

no qual Qav_1 corresponde à qualidade áudio-visual estimada, Qv corresponde ao valor

de qualidade obtida com a métrica de vídeo, e Qa representa o valor da qualidade obtida utilizando a métrica de áudio. O ajuste retornou coeficientes de escala α_8 =0.866, β_5 =0.5242, e $\gamma_6 = -35.6387$. Para este ajuste, os coeficientes de correlação Pearson e Spearman foram iguais a 0.7929 e 0.7972, respectivamente. Na Figura 5.8, é apresentado o gráfico dos valores de Qav_4 estimados para as sequências áudio-visuais do Experimento III versus o MOS_{av} obtidos neste experimento.

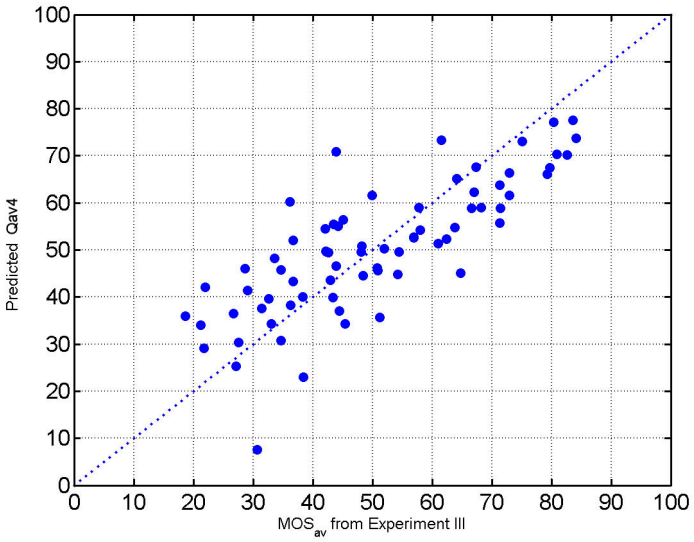


Figura 5.8: Qav_4 estimado com o modelo Linear versus MOS_{av} obtido no Experimento III.

O segundo modelo testado foi o modelo Minkowski ponderado, dado pela seguinte equação:

$$\text{Qav}_5 = (\alpha_9 \cdot \text{Qv}^{p_7} + \beta_6 \cdot \text{Qa}^{p_7})^{\frac{1}{p_7}}. \tag{5.9}$$

no qual Qav_5 corresponde à qualidade áudio-visual estimada pelo modelo. O ajuste para o modelo minkowski retornou um exponente $p_7 = 0.0003$ e coeficientes escalares $\alpha_9 = 0.6160$ e $\beta_6 = 0.3840$, correspondentes aos valores de vídeo e áudio, respectivamente. Para este ajuste, os coeficientes de correlação Pearson e Spearman obtidos foram de 0.7779 e 0.7920, respectivamente. Na Figura 5.9, é apresentado o gráfico dos valores de Qav_5 estimados para as sequências áudio-visuais do Experimento III versus o MOS_{av} obtidos neste experimento.

Finalmente, o terceiro modelo ajustado foi o modelo do produto de potências utilizado por Wang e Bovik [51], descrito pela seguinte equação:

$$\text{Qav}_6 = (\gamma_7 + \alpha_{10} \cdot \text{Qv}^{p_8} \cdot \text{Qa}^{p_9}), \tag{5.10}$$

no qual Qav_6 corresponde à qualidade áudio-visual estimada pelo modelo de produto de Potência. O ajuste para este modelo combinatório retornou exponentes $p_8 = 1.9904$ e $p_9 = 0.9762$, correspondendo às componentes de vídeo e áudio, respectivamente. São retornados também os coeficientes escalares $\alpha_{10} = 0.0001$ e $\gamma_7 = 20.1468$. Para este ajuste, os coeficientes de correlação Pearson e Spearman foram iguais a 0.8100 e 0.8068, respecti-

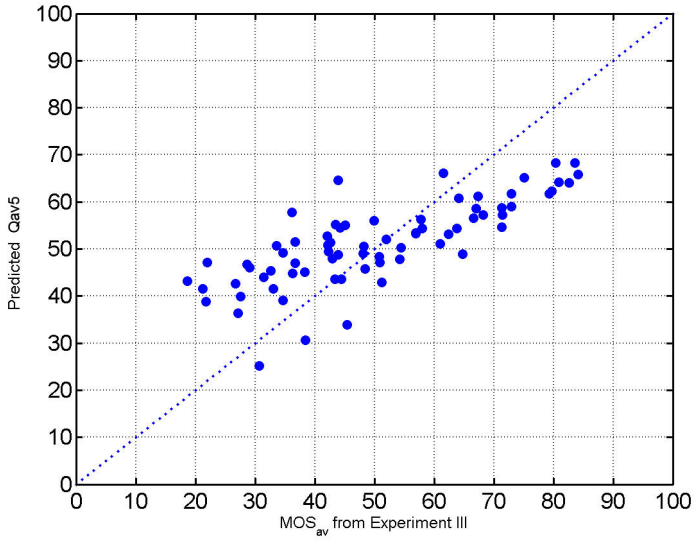


Figura 5.9: Qav_5 estimado com o modelo Minkowski versus MOS_{av} obtido no Experimento III.

vamente. Na Figura 5.10, é apresentado o gráfico dos valores de Qav_6 estimados para as sequências áudio-visuais do Experimento III versus o MOS_{av} obtidos neste experimento.

Os resultados apresentados mostram que modelos conseguem estimar os valores de qualidade áudio-visual de forma aceitável. Assim como os modelos objetivos com-referência, há um equilíbrio na contribuição das componentes de áudio e vídeo no que diz respeito à qualidade áudio-visual estimada. Pode-se observar que a contribuição das componentes de áudio e vídeo nos modelos objetivos com referência é relativamente similar. Por outra lado, nos modelos objetivos sem referência a componente de vídeo apresenta uma contribuição muito mais forte. Assim como no caso dos modelos objetivos com referência completa, o modelo de combinação que apresentou o melhor desempenho foi o produto de potências (Qav_6).

5.4 Comparação dos Resultados

Na Tabela 5.1 é apresentado um resumo dos resultados dos modelos propostos nesta dissertação. Também são apresentados os resultados de 5 modelos da literatura testados com as sequências do Experimento III. Os modelos considerados foram (ver Seção 3.3):

- 2 modelos propostos por Hands [26];
- 1 modelo proposto por Garcia [27]; e
- 2 modelos propostos por Winkler [28].

Todos estes modelos são modelos subjetivos. É possível observar da Tabela 5.1 que os modelos subjetivos propostos neste trabalho apresentam melhor desempenho que estes modelos. Em particular, os modelos PrMOS_4 e PrMOS_3 foram os dois modelos com melhor desempenho entre todos os modelos testados.

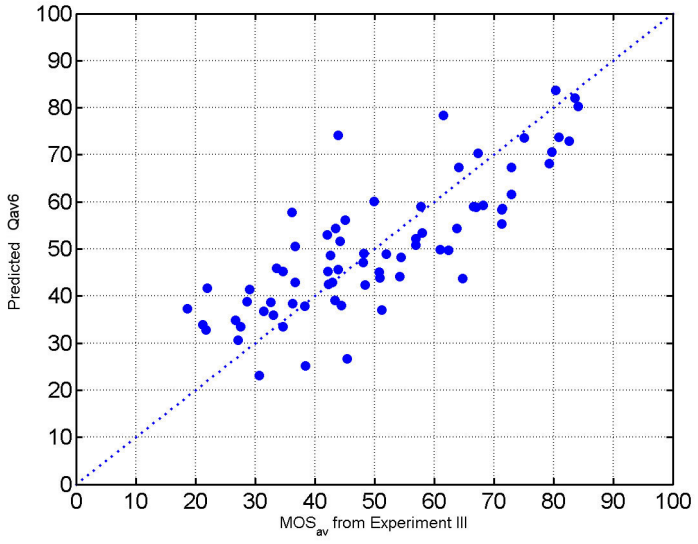


Figura 5.10: Qav_6 estimado com o modelo de Potencia versus MOS_{av} obtido no Experimento III.

Entre os modelos objetivos (propostos), as métricas FR apresentaram melhor desempenho, quando comparadas às métricas NR. Entretanto a diferença de desempenho entre as métricas FR e NR não é muito grande. Experimentos futuros utilizando métricas com referência reduzida (RR) podem ser realizados.

Mas, como esperado, os modelos objetivos apresentaram um desempenho um pouco pior que os modelos subjetivos. Observe que, em geral, o modelo de produto de potências apresentou o melhor desempenho entre os modelos propostos (objetivos e subjetivos). Cabe mencionar que, não foram citados nesta dissertação os testes que incluíam apenas o fator multiplicativo de áudio e vídeo (Qv $\times$ Qa). Estes modelos apresentaram uma correlação muito baixa (da ordem de 0.60). Por esta razão estes modelos não foram incluídos. Isto mostra que, as componentes de áudio e vídeo possuem definitivamente um termo de interação exponencial. As Figuras 5.11 e 5.12 apresentam comparações entre os valores de correlação Pearson e Spearman obtidos pelos modelos, tanto propostos como da literatura.

O modelo Qav_3 apresentou o melhor desempenho entre todos os modelos objetivos, com bons valores de correlação. É importante lembrar que, dado que o número de propostas para a aferição da qualidade áudio-visual na literatura é baixo, estes modelos representam uma importante contribuição à pesquisa na área da qualidade áudio-visual.

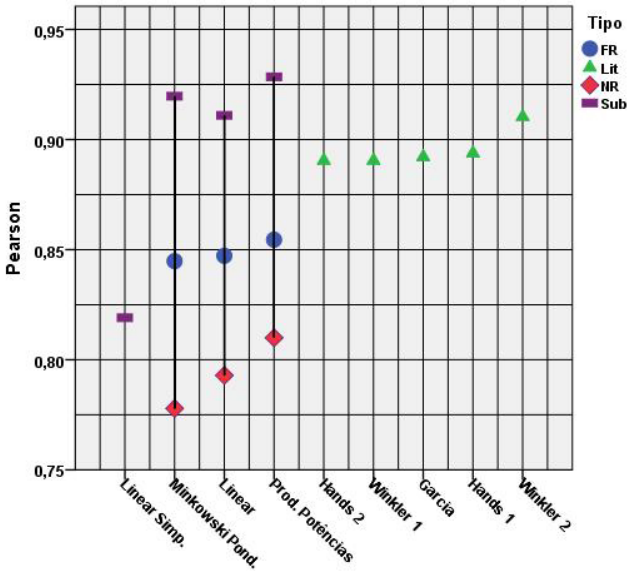


Figura 5.11: Coeficientes de correlação Pearson (FR, Lit=Literatura, NR, Sub=Subjetivo).

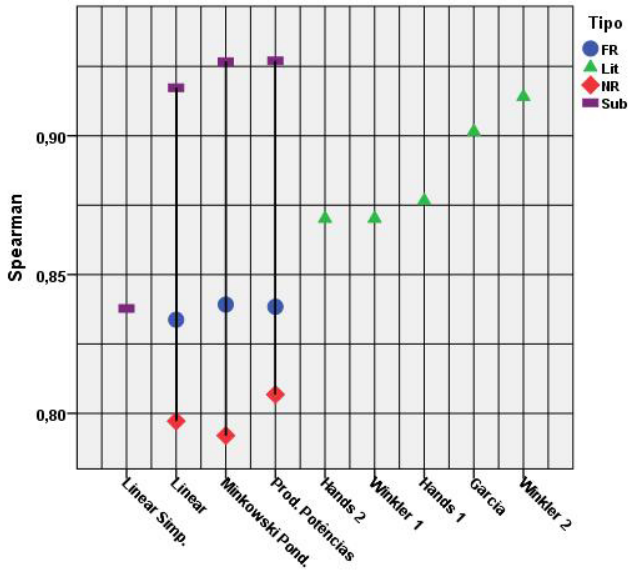


Figura 5.12: Coeficientes de correlação Spearman (FR, Lit=Literatura, NR, Sub=Subjetivo).

Tabela 5.1: Resumo dos resultados obtidos para os modelos de qualidade áudio-visual, testados nas sequências do Experimento III.

Modelo	**Tipo**	**Pearson**	**Spearman**
$PrMOS_1$	SUBJ.	0.9110	0.9173
$PrMOS_2$	SUBJ.	0.8191	0.8378
$PrMOS_3$	SUBJ.	0.9197	0.9267
$PrMOS_4$	SUBJ.	**0.9285**	**0.9270**
Qav_1	FR	0.8472	0.8337
Qav_2	FR	0.8448	**0.8392**
Qav_3	FR	**0.8545**	0.8384
Qav_4	NR	0.7929	0.7972
Qav_5	NR	0.7779	0.7920
Qav_6	NR	**0.8100**	**0.8068**
Hands-1 [26]	SUBJ.	0.8938	0.8766
Hands-2 [26]	SUBJ.	0.8905	0.8702
Garcia [27]	SUBJ.	0.8921	0.9015
Winkler-1 [28]	SUBJ.	0.8905	0.8702
Winkler-2 [28]	SUBJ.	**0.9104**	**0.9141**

Capítulo 6

Conclusões

Nesta dissertação, apresenta-se um conjunto de modelos para avaliar a qualidade subjetiva e objetiva de sequências áudio-visuais. O desempenho destes modelos foi comparado com o de várias das propostas existentes na literatura. Nestas comparações, foram utilizandos os dados subjetivos obtidos a partir da realização de três experimentos psicofísicos. Foi observado que o desempenho dos modelos é bom em comparação aos outros modelos. Ainda que os valores de correlação obtidos estejam abaixo de 0,9, os modelos apresentaram resultados coerentes e próximos aos modelos subjetivos. Este pode ser um resultado bom, considerando a alta complexidade das sequências testadas e o fato da proposta incluir métricas objetivas.

Como parte deste trabalho, desenvolveu-se uma métrica de qualidade de vídeo NR e uma métrica de qualidade de áudio NR:

- Foi apresentado um modelo sem-referência para calcular a qualidade de sinais de vídeo. O modelo combina os níveis de borrado e blocagem mediante um modelo linear simples. Estes valores são calculados previamente utilizando as métricas descritas em [50] e [51], respetivamente.
- Com o objetivo de encontrar uma métrica que possa calcular a qualidade de áudio sem utilizar sinais de referência, foi apresentada uma modificação do algoritmo de aferição da qualidade SESQA. Esta nova versão do algoritmo SESQA foi testada com sequências de áudio com conteúdo variado.

As métricas mencionadas acima foram utilizadas no modelo objetivo da qualidade áudio-visuais.

As principais contribuições deste trabalho na área de qualidade de áudio e vídeo foram:

- Um conjunto de modelos subjetivos para a predição da qualidade subjetiva em sinais áudio-visuais. Foram apresentados três modelos subjetivos: (1) um modelo linear, (2) um modelo linear simplificado, onde foi incluído apenas a componente de vídeo, (3) um modelo minkowski ponderado, e (4) um modelo de produto de potências. O modelo (4) apresentou a melhor correlação obtendo um valor igual a 0.92.
- Um conjunto de modelos objetivos com referência para estimar a qualidade de sinais áudio-visuais. Três modelos objetivos foram apresentados: (1) um modelo linear, (2) um modelo minkowski ponderado, e (3) um modelo de produto de potências. Todos

os três modelos apresentaram uma boa correlação com os resultados subjetivos. O modelo (3) obteve a melhor correlação, com um valor de 0.85;

- Um conjunto de modelos objetivos sem-referência para a aferição da qualidade de sinais áudio-visuais. Três modelos objetivos foram apresentados: (1) um modelo linear, (2) um modelo minkowski ponderado, e (3) um modelo de produto de potências. Este último modelo apresentou a melhor correlação obtendo uma correlação de 0.81;
- Banco de dados de sequências de teste com degradações de compressão e avaliações subjetivas de qualidade. As sequências de teste consistem de: (1) seis sequências originais, (2) 24 sequências do Experimento I, (3) 18 sequências do Experimento II, e (4) as 72 sequências do Experimento III. Todas as sequências possuem valor de qualidade subjetivo associado (mínimo de 15 observadores).

6.1 Trabalhos Futuros

Para dar continuidade a este trabalho, as seguintes abordagens são sugeridas:

- Testar os modelos subjetivos propostos neste trabalho utilizando a informação subjetiva disponibilizada na comunidade cientifica. O objetivo é utilizar sequências áudio-visuais como conteúdo novo, com o fim de avaliar e observar o comportamento dos modelos propostos;
- Estudar novas funções combinatórias que possam integrar melhor as qualidades de áudio e vídeo.
- Estudar modelos de qualidade que considere outros aspectos do sinal multimídia, além do áudio ou vídeo, como texto ou efeitos gráficos.
- Levando em consideração as métricas descritas no Capitulo 3, utilizar métricas com referência reduzida (RR) para gerar novos modelos de qualidade áudio-visual.
- Utilizar um abordagem híbrida para aferir a qualidade do sinal multimídia. Esta abordagem consiste em combinar os parâmetros comumente utilizados no cálculo da qualidade de vídeo e dados relacionados como o desempenho da rede (QoS).

6.2 Conclusões Finais

Algumas das conclusões importantes deste trabalho são:

- As características do conteúdo das componentes de áudio e vídeo são importantes para determinar a qualidade percebida;
- A compressão da componente de vídeo tem um impacto levemente maior na qualidade áudio-visual;
- A componente de áudio pode atuar como um fator de distração durante a avaliação subjetiva, influenciando negativamente a opinião dos avaliadores;

- A contribuição das componentes de áudio e vídeo nos modelos objetivos com referência é relativamente similar;
- A contribuição da componente de vídeo nos modelos objetivos sem-referência é maior do que a contribuição da componente de áudio;
- Na ausência de um sinal de referência, no caso dos modelos objetivos sem-referência, a qualidade de vídeo ganha uma maior importância no cálculo da qualidade global áudio-visual;
- Modelos objetivos e subjetivos apresentam melhores resultados na inclusão de um termo de interação exponencial.

Tendo como base os resultados da Tabela 5.1, os experimentos psico-físicos continuam sendo a melhor abordagem para a aferição da qualidade áudio-visual. Eles são de muita ajuda no estudo do relacionamento entre a qualidade de áudio e vídeo. Ainda existe um caminho longo para percorrer no modelamento de métricas áudio-visuais objetivas. As métricas apresentadas representam uma importante contribuição para futuros estudos na área de qualidade áudio-visual.

Referências

[1] R. C. Gonzalez, R. E. Woods, and P. Hall, *Digital Image Processing (2nd Edition).* Tom Robbins, 1987. x, 7, 8, 13, 15, 29

[2] J. P. L. Velasco, "Video quality assessment," in *Video Compression, Edited by Amal Punchihewa* (T. Smiljanic, ed.), pp. 129–154, InTech, amal punch ed., 2012. x, 8, 15

[3] F. O. D. L. F. R. C. d. L. S. N. A. d. Gadelha. Maria Jose Nunes, Andrade. Michael Jackson Oliveira, "Sensibilidade ao contraste acromatico para grades senoidais verticais em adolescentes e adultos," *Psicologia: teoria e pratica*, vol. 12, pp. 59 – 70, 00 2010. x, 9

[4] M. Farias, "Video quality metrics," in *Digital Video - Edited by Floriano De Rango* (F. D. Rango, ed.), no. February, ch. 16 Video Q, pp. 330–358, InTech, 2010. x, 2, 13, 23, 24, 29

[5] J. Korhonen, "Audiovisual quality assessment in communications applications: current status, trends and challenges," *Signal Processing*, pp. 6–9, 2010. x, 14

[6] ITU-R, "Recommendation p.910: Subjective video quality assessment methods for multimedia applications," tech. rep., 1999. x, 20, 21, 22

[7] Z. Wang, L. Lu, and A. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Comm.*, vol. vol19, pp. 121–132, 2004. x, 2, 26, 27

[8] ITU-R, "J.246 : Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference," tech. rep., 2008. x, 2, 23, 26, 27

[9] G. M. Gunawan I., "Image quality assessment based on harmonics gain/loss information," *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 1, pp. I – 429–32, 2005. x, 2, 28

[10] M. A. Saad, A. C. Bovik, and C. Charrier, "Dct statistics model-based blind image quality assessment," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 3093–3096, 2011. x, 2, 30

[11] L. Malfait, J. Berger, and M. Kastner, "The itu-t standard for single-ended speech quality assessment," Tech. Rep. 6, 2006. x, xii, 2, 31, 32, 67

[12] A. Ostaszewska and R. Kloda, "Quantifying the amount of spatial and temporal information in video test sequences," in *Recent Advances in Mechatronics, Springer*, pp. 11–15, 2007. x, 37, 39

[13] A. Aksay, *Motion Wavelet Video Compression.* PhD thesis, Middle East Technical University, 2001. xii, 14

[14] D. J. M. Robinson, *Perceptual model for assessment of coded audio.* PhD thesis, University of Essex, 2002. xii, 16, 18

[15] L. J. Najjar, "Multimedia information and learning," in *Journal of Educational Multimedia and Hypermedia*, pp. 129–150, 1996. 1

[16] K. Piamrat, C. Viho, J.-M. Bonnin, and A. Ksentini, "Quality of Experience Measurements for Video Streaming over Wireless Networks," *Information Technology: New Generations, 2009. ITNG '09. Sixth International Conference on*, pp. 1184–1189, 2009. 1

[17] D. Campbell, E. Jones, and M. Glavin, "Audio quality assessment techniques—a review, and recent developments," *Signal Processing*, vol. 89, pp. 1489–1500, aug 2009. 2

[18] U. Engelke and H.-J. Zepernick, "Perceptual-based quality metrics for image and video services: A survey," *Next Generation Internet Networks, 3rd EuroNGI Conference on*, pp. 190–197, 2007. 2

[19] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," tech. rep., VQEG, 2003. 2, 6, 25

[20] Z. Wang and A. Bovik, "Reduced- and No-Reference Image Quality Assessment," *Signal Processing Magazine, IEEE*, vol. 28, pp. 29–40, 2011. 2

[21] J. Lubin, "Sarnoff jnd vision model," 1997. 2, 25

[22] J. E. Caviedes, "No-reference quality metric for degraded and enhanced video," in *Proceedings of SPIE* (T. Ebrahimi and T. Sikora, eds.), vol. 5150, pp. 621–632, SPIE, 2003. 2, 14, 16, 28

[23] M. Farias and S. Mitra, "No-reference video quality metric based on artifact measurements," *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 3, no. 2, pp. III – 141–4, 2005. 2, 28

[24] ITU-R, "Recommendation bs.1387 : Method for objective measurements of perceived audio quality," tech. rep., 1998. 2, 30, 31

[25] J. Gustafsson, S. Argyropoulos, M.-n. Garcia, D. Lindegren, G. Heikkila, M. Pettersson, P. List, and B. Feiten, "Ip-based mobile and fixed network audiovisual media services," *Signal Processing Magazine, IEEE*, vol. 28, pp. 68–79, nov 2011. 2

[26] D. S. Hands, "A Basic Multimedia Quality Model," *Multimedia, IEEE Transactions on*, vol. 6, no. 6, pp. 806–816, 2004. 3, 32, 33, 39, 55, 58

[27] M. N. Garcia, R. Schleicher, and a. Raake, "Impairment-factor-based audiovisual quality model for iptv: Influence of video resolution, degradation type, and content type," *EURASIP Journal on Image and Video Processing*, pp. 1–14, 2011. 3, 32, 33, 55, 58

[28] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *Multimedia, IEEE Transactions on*, vol. 8, no. 5, pp. 973–980, 2006. 3, 32, 33, 55, 58

[29] ITU-T, "Recommendation itu-t g.1070: Opinion model for video-telephony applications," tech. rep., 2007. 3

[30] C. Starr, C. Evers, and L. Starr, *Biology: Concepts and Applications*. Brooks/Cole biology series, Thomson, Brooks/Cole, 2006. 6

[31] F. V. Rodrigues, "Fisiologia sensorial," vol. 5, pp. 24–32, 2010. 7, 8

[32] Y. Zhao, L. Yu, Z. Chen, and C. Zhu, "Video quality assessment based on measuring perceptual noise from spatial and temporal perspectives," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, pp. 1890–1902, dec 2011. 9

[33] A. Liu, M. Verma, and W. Lin, "Modeling the masking effect of the human visual system with visual attention model," *Information, Communications and Signal Processing, 2009. ICICS 2009. 7th International Conference on*, pp. 1–5, dec 2009. 9

[34] H. Fastl, "Psycho-acoustics and sound quality," in *Communication Acoustics* (J. Blauert, ed.), pp. 139–162, Springer Berlin Heidelberg, 2005. 10

[35] R. E. Bosi, M. Goldberg, *Introduction to Digital Audio Coding and Standards*. Springer International Series in Engineering and Computer Science, 2003. 12

[36] M. Ghanbari, *Standard Codecs: Image compression to advanced video coding*, vol. 49. Iet, 2003. 12, 13

[37] Y. Q. Shi and H. Sun, *Image and Video Compression for Multimedia Engineering*. Boca Raton, FL, USA: CRC Press, Inc., 1st ed., 1999. 13

[38] P. Lambert, W. de Neve, I. Moerman, P. Demeester, and R. V. de Walle, "Rate-distortion performance of h.264/avc compared to state-of-the-art video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 1, pp. 134–140, 2006. 14

[39] ITU-R, "Recommendation p.800 : Methods for subjective determination of transmission quality," tech. rep., 1996. 20

[40] ITU-R, "Recommendation bs.1116 : Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," tech. rep., 1997. 20

[41] ITU-R, "Recommendation bs.1534 : Method for the subjective assessment of intermediate quality levels of coding systems," tech. rep., 2003. 20

[42] ITU-R, "Recommendation bt.500-8: Methodology for subjective assessment of the quality of television pictures," tech. rep., 1998. 20, 35, 36

[43] ITU-R, "Recommendation bt.710 : Subjective assessment for image quality in high-definition television," tech. rep., 1997. 20

[44] ITU-R, "Recommendation P.911 : Subjective audiovisual quality assessment methods for multimedia applications," 1998. 20, 21, 35, 40

[45] ITU-R, "Recommendation P.920 : Interactive test methods for audiovisual communications," tech. rep., 2000. 20

[46] K. Van Zon, "Automated video chain optimization," in *Consumer Electronics, 2001. ICCE. International Conference on*, pp. 296–297, 2001. 28

[47] T. Vlachos, "Detection of blocking artifacts in compressed video," *Electronics Letters*, vol. 36, no. 13, pp. 1106–1108, 2000. 29

[48] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to jpeg2000," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163–172, 2004. 29

[49] J. Lee and K. Hoppel, "Noise modeling and estimation of remotely-sensed images," *Geoscience and Remote Sensing Symposium, 1989. IGARSS 89. 12th Canadian Symposium on Remote Sensing., 1989 International*, vol. 2, pp. 1005–1008, 1989. 29

[50] N. Narvekar and L. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (cpbd)," *Image Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2678–2683, 2011. 29, 52, 59

[51] Z. Wang, H. R. Sheikh, and A. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, pp. I–477–I–480 vol.1, 2002. 30, 51, 52, 54, 59

[52] VQEG, "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, Phase I," tech. rep., 2008. 37, 42

[53] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A multi-class audio classification method with respect to violent content in movies using bayesian networks," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pp. 90–93, 2007. 37

[54] ITU Recommendation BT.500-8, *Methodology for subjective assessment of the quality of television pictures.* 1998. 40, 41

[55] J. L. Rodgers and A. W. Nicewander, "Thirteen Ways to Look at the Correlation Coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988. 47

[56] S. G. Revesz, "The Generalized Minkowski Functional with Applications in Approximation Theory," *J. Convex Anal.*, vol. 11, pp. 303–334, Mar 2007. 48

[57] J. Chen, Y. Zhang, L. Liang, S. Ma, R. Wang, and W. Gao, "A no-reference blocking artifacts metric using selective gradient and plainness measures," in *Advances in Multimedia Information Processing - PCM 2008*, vol. 5353 of *Lecture Notes in Computer Science*, pp. 894–897, 2008. 52

[58] H. Liu and I. Heynderickx, "A no-reference perceptual blockiness metric," in *in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, 2008*, pp. 865–868. 52

[59] G. Yammine, E. Wige, and A. Kaup, "A no-reference blocking artifacts visibility estimator in images," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 2497–2500, 2010. 52

[60] A. Leontaris and A. R. Reibman, "Comparison of blocking and blurring metrics for video compression," in *Journal of Semantics*, pp. 585–588, 2005. 52

Anexo A

Parâmetros utilizados no algoritmo SESQA

Tabela A.1: Visão geral de todos os parâmetros utilizados no algoritmo SESQA [11].

Descritores de voz	**Voz não natural**		**Análise do ruído**		**Interrupções - silêncios**
	Análise do trato vocal	**Estatísticas da voz**	**SNR estática**	**SNR por segmentos**	
PitchAverage	Robotization	LPCcurt	SNR	EstSegSNR	SpeechInterruptions
SpeechSection LevelVar	ConsistentArt Tracker	LPCskew	EstBGNoise	SpecLevel Dev	SharpDeclines
SpeechLevel	VTPMaxTube Section	LPCskewAbs	NoiseLevel	SpecLevel Range	MuteLength
LocalLevelVar	FinalVtpAverage	CepCurt	HiFreqVar	RelNoise Floor	Unnatural Silence
	VTPPeakTracker	CepSkew	SpectralClarity		Unnatural SilenceMean
	ArtAverage	CepADev	GlobalBGNoise		Unnatural SilenceTotEnergy
	VtpVadOverlap		GlobalBGNoiseTotEnergy		
	PitchCrossCorrlOffset		GlobalBGNoise RelEnergy		
	PitchCrossPower		GlobalBGNoise AffectedSamples		
	BasicVoiceQuality		LocalBGNoise Log		
	BasicVoiceQuality Asym		LocalBGNoise Mean		
	BasicVoiceQuality Sym		LocalBGNoise Stddev		
	FrameRepeats		LocalBGNoise		
	FrameRepeats TotEnergy		LocalBGNoise AffectedSamples		
	UnnaturalBeeps				
	UnnaturalBeeps Mean				
	UnnaturalBeeps				
	AffectedSamples				

Anexo B

Parâmetros utilizados no modelo de qualidade de áudio sem referência

Tabela B.1: Parâmetros utilizados no modelo de qualidade de áudio sem referência

Parâmetro	**Nome**	**Classificação**
1	PitchAverage	Descritores de voz
3	SpeechLevel	Descritores de voz
4	MuteLength	Interrupções silêncios
10	LocalBGNoiseLog	Análise do ruído
13	RelNoiseFloor	Análise do ruído
14	SNR	Análise do ruído
15	SpecLevelDev	Análise do ruído
16	SpecLevelRange	Análise do ruído
17	SpectralClarity	Análise do ruído
18	BasicVoiceQuality	Voz não natural
19	ArtAverage	Voz não natural
21	CepCurt	Voz não natural
22	FinalVtpAverage	Voz não natural
24	LPCCurt	Voz não natural
25	LPCSkew	Voz não natural
27	PitchCrossCorrelOffset	Voz não natural
28	PitchCrossPower	Voz não natural

Anexo C

Modelo de qualidade de áudio sem referência

$$\begin{aligned}\text{Qaudio} = Parametro_1 \cdot 6.03 + Parametro_3 \cdot 23.62 + Parametro_4 \cdot 1.43 - Parametro_{10} \cdot 0.03 \\ + Parametro_{13} \cdot 3.85 - Parametro_{14} \cdot 0.31 + Parametro_{15} \cdot 18.23 - Parametro_{16} \cdot 4.68 \\ - Parametro_{17} \cdot 8.00 + Parametro_{18} \cdot 2.10 + Parametro_{19} \cdot 12.39 - Parametro_{21} \cdot 3.66 \\ - Parametro_{22} \cdot 365.91 - Parametro_{24} \cdot 2.04 - Parametro_{25} \cdot 321.54 - Parametro_{27} \cdot 7.37 \\ - Parametro_{28} \cdot 3.43 - 315.94.\end{aligned} \tag{C.1}$$