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

# A Three Layer System for Audio-visual Quality Assessment

Helard Alberto Becerra Martinez

Dissertação apresentada como requisito parcial para
conclusão do Doutorado em Informática

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

Brasília
2019

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

# A Three Layer System for Audio-visual Quality Assessment

Helard Alberto Becerra Martinez

Dissertação apresentada como requisito parcial para
conclusão do Doutorado em Informática

Prof.ª Dr.ª Mylene Christine Queiroz de Farias (Orientadora)
CIC/UnB

Prof. Dr. Eduardo Peixoto Fernandes da Silva
University of Brasília

Prof. Dr. Teófilo Emidio de Campos
University of Brasília

Prof. Dr. Alexandre de Almeida Prado Pohl
Federal University of Technology - Paraná

Prof. Dr. Bruno L. M. Espinoza
Coordenador do Programa de Pós-graduação em Informática

Brasília, 11 de Janeiro de 2019

# Resumo

As métricas objetivas de avaliação de qualidade de sinais tem o objetivo de prever a qualidade dos sinais percebida pelo ser humano. Uma das áreas de qualidade de maior interesse nos últimos anos é o desenvolvimento de métricas de qualidade para sinais áudio-visuais. A maioria das propostas nesta área estão baseadas na aferição da qualidade individual das componentes de áudio e vídeo. Porém, o modelamento da complexa interação existente entre as componentes de áudio e vídeo ainda é um grande desafio. Nesta tese, o objetivo é desenvolver uma métrica, baseado em ferramentas de aprendizado de máquina (Machine Learning - ML), para a aferição da qualidade de sinais áudio-visuais. A proposta utiliza como entrada um conjunto de características descritivas das componentes de áudio e vídeo e aplica Deep Autoencoders para gerar um novo conjunto de características descritivas que representa a interação entre as componentes de áudio e vídeo. O modelo está composto por várias fases, que realizam diferentes tarefas. Primeiramente, são extraídos um conjunto de características descritivas que descrevem características das componentes de áudio e vídeo. Na próxima fase, um autoencoder de duas camadas produz um novo conjunto de características descritivas. Em seguida, uma função de classificação mapeia as características descritivas em escores de qualidade audiovisual. Para garantir a precisão nos resultados, o modelo é treinado utilizando um conjunto de dados que representa todos os artefatos considerados no modelo. O modelo foi testado tanto com no banco de dados gerado neste trabalho, como em uma base de dados extensa e pública. Os resultados mostraram que esta abordagem obtém predições de qualidade, cujos valores estão altamente correlacionadas com os escores de qualidade obtidos em experimentos subjetivos.

**Palavras-chave:** Qualidade Audiovisual, Metricas Objetivas de Qualidade, Multimidia

# Abstract

The development of models for quality prediction of both audio and video signals is a fairly mature field. But, although several multimodal models have been proposed, the area of audiovisual quality prediction is still an emerging area. In fact, despite the reasonable performance obtained by combination and parametric metrics, currently there is no reliable pixel-based audiovisual quality metric. The approach presented in this work is based on the assumption that autoencoders, fed with descriptive audio and video features, might produce a set of features that is able to describe the complex audio and video interactions. Based on this hypothesis, we propose a set of multimedia quality metrics: video, audio and audiovisual. The visual features are natural scene statistics (NSS) and spatial-temporal measures of the video component. Meanwhile, the audio features are obtained by computing the spectrogram representation of the audio component. The model is formed by a 2-layer framework that includes an autoencoder layer and a classification layer. These two layers are stacked and trained to build the autoencoder network model. The model is trained and tested using a large set of stimuli, containing representative audio and video artifacts. The model performed well when tested against the UnB-AV and the LiveNetflix-II databases.

**Keywords:** Audiovisual Quality, Objective Quality Metrics, Multimedia

# Contents

# List of Figures

# List of Tables

# Dedicatória

Para David, Yolanda, Danny, e Renzo.

# Agradecimentos

A todas as pessoas que encontrei neste longo caminho, todas elas contribuíram de alguma forma para eu estar neste momento da minha vida. Muito obrigado.

Aos meus pais David e Yolanda e meus irmãos Danny e Renzo. A distância que nos separou foi provavelmente o maior desafio que tive que superar nestes anos. Muito obrigado.

A Profa. Mylène C. Q. Farias, 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 doutorado. Muito obrigado.

Aos professores do departamento de Computação, que foram tão importantes na minha vida acadêmica e no desenvolvimento deste trabalho. Também aos professores Alexandre de Almeida Prado Pohl, Eduardo Peixoto, e Teófilo de Campos, que fazem parte da minha banca, pela sua presença, suas sugestões e contribuições para com meu trabalho.

Ao Prof. Andrew Hines, pelas valiosas contribuições e incentivo ao longo deste trabalho.

Aos amigos que encontrei nesta cidade, obrigado pela companhia e por compartilhar seu dia a dia comigo. Muito obrigado Maria Paz, Toni, Harley, Armando Vanessa e Dulce, Victor, Stephanie, Amanda, Gabriela, Nathaly, Carlos, Daniela, Julian, Juliana, Mariano, Ruben, Yang, Zheng, Iago, Danny, Esther, Sofia, Alessandro, Felipe, Marcelo, Maria Clara, Dina, Philipe.

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.

# Publications


1. Martinez, Helard B, and Mylène CQ Farias. "A no-reference audio-visual video quality metric." Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European. IEEE, 2014.

2. Martinez, Helard B, and Mylène CQ Farias. "Full-reference audio-visual video quality metric." Journal of Electronic Imaging 23.6 (2014).

3. Martinez, Helard B, and Mylène CQ Farias. "An objective model for audio-visual quality." IS& T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2014.

4. Martinez, Helard A. Becerra, and Mylène CQ Farias. "Using the immersive methodology to assess the quality of videos transmitted in UDP and TCP-based scenarios", Image Quality and System Performance (IQSP) XV (2018).

5. Martinez, Helard A. Becerra, and Mylène CQ Farias. "Combining audio and video metrics to assess audio-visual quality." Multimedia Tools and Applications (2018): 1-20.

6. Martinez, Helard, Mylène CQ Farias, and Andrew Hines. "Perceived quality of audio-visual stimuli containing streaming audio degradations." 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018.

7. Martinez, Helard, and Mylène CQ Farias. "Analyzing the influence of cross-modal IP-based degradations on the perceived audio-visual quality." Electronic Imaging 2019.01 (2019).

# Chapter 1

# Introduction

The great progress achieved by communication technology in the last twenty years is reflected by the amount of multimedia services available nowadays, such as digital television, IP-based video transmission, mobile services, etc. One of the most popular services is Internet-based transmission, which has recently gained a huge popularity among consumers of entertainment services. Recent advances on smarthphones technology have transformed services like video conference (Skype, Google Hangout, Facebook Video, FaceTime) and on-demand streaming media (Netflix, iTunes, Amazon) into an essential tool for the common user. Yet, it is understood that the success of these kind of services relies on its trustworthiness and the quality of experience of the provided service [1]. Under these circumstances, the development of efficient real-time monitoring quality tools, which can quantify the audio-visual experience (as perceived by the end user), can bring real benefits to Internet Service Providers (ISP) and broadcast companies.

The most accurate way to determine the quality of an audio-visual content is by measuring it using psychophysical experiments with human subjects (sometimes referred as subjective experiments)[2]. These experiments are usually conducted in a controlled environment (e.g., soundproof laboratories), where a set of test stimuli (e.g., audio-visual sequences) are presented to a group of non experts human subjects. In order to reproduce these experiments among different labs, researchers design the experiments following a set of recommendations that vary according to the type of experiment and test stimuli under study. These recommendations include several instructions on topics like experimental methodology, viewing conditions, test material and grading scale. This type of recommendations are compiled in several documents by different communication agencies such as the International Telecommunications Union (ITU) and the European Broadcasting Union (EBU) [3]. By following these recommendations experiments can be reproduced in different laboratories with a guaranteeing a certain level of reliance in the results.

Although subjective experiments represent the most accurate way of measuring the

signal quality, they are expensive, time-consuming, and hard to incorporate into a design process or an automatic control of quality. Therefore, the ability to measure audio and video quality accurately and efficiently, without using human observers, is highly desirable in practical applications. With this in mind, the development of fast algorithms (objective metrics) that give an accurate prediction of the subjective quality of the media is an area that has much to be explored.

Objective metrics use computational methods to analyse the characteristics of video and audio signals and obtain an estimate for the perceived quality. Unfortunately, within the signal processing community, quality measurements have been largely limited to a few objective measures, such as peak signal-to-noise ratio (PSNR) and total squared error (TSE). Although these metrics have a reasonable performance for signals in which every bit is equally important, they do not provide good estimates for audio and video signals, i.e., their estimates do not always have a good correlation with the human judgement of quality [4, 5].

Depending on the amount of reference information used by the algorithm, objective metrics can be classified as Full Reference (FR), Reduced Reference (RR), or No-Reference (NR) metrics. In the FR approach, the entire reference is used to obtain an estimate of the quality. In the RR approach, the algorithm uses only part of the reference, which generally consists of a set of features extracted from it. In this case, the information available at the measurement point is transmitted through an auxiliary channel. Finally, in the NR approach, the quality estimation is obtained blindly using only the test video.

There is an ongoing effort to develop video quality metrics that are able to estimate quality as perceived by human viewers [6, 4, 7]. Unfortunately, metrics with better results are often FR metrics. Usually, the best performing quality metrics incorporate models of the human visual system (HVS), such as contrast sensitivity functions, motion models, pooling strategies, and visual attention models. To date, most of the achievements have been in the development of complex FR video quality metrics [8, 9, 10] and much remains to be done in the development of real-time metrics that do not require the reference signal (NR or RR). A new trend in video quality is the development of hybrid and parametric metrics, which are metrics that use a combination of packet information, bitstream headers, and decoded video to estimate the quality [11]. Parametric metrics estimate quality using only the information available at the receiver, like for example bitrate, frame rate, QP, motion vectors, and network information. These metrics are generally faster than pixel-based video quality metrics and, depending on the level of access to the bitstream, can produce reliable results [12]. It is worth pointing out that parametric metrics are coding and transmission dependent, reducing their applications. For example, parametric

metrics cannot be used to predict quality for content transcoded among different compression standards or bitrates.

There is also a great need for metrics that estimate quality of experience in multimedia applications. So far, very few metrics have addressed the issue of simultaneously measuring the quality of multimedia content (e.g., video, audio, and text). In fact, only a small number tackle the simpler problem of developing audio-visual objective metrics [13, 14, 15]. Among the most relevant works, we can cite the parametric NR objective quality metrics proposed by Garcia et al. [15] and Yamagishi and Gao [16].

In this work, our goal is to develop an accurate model to assess the audiovisual quality of a video sequence. The proposed model is based in an Autoencoder Network approach which is composed of two main stages. The first stage consists of an audio and video feature extraction phase, where measures that describe the audio and video signal characteristics are computed. The second stage consists of a training phase formed by an autoencoder and a classification layer trained with the sets of features extracted before. During the training phase of the model, the autoencoder is trained using the audio and video features as input, resulting in a low-dimensional representation of the features. Then, using the classification function, a mapping between this new set of features and the subjective scores associated with the audiovisual sequence is obtained. The assumption is that by training the autoencoders using these audio and video features a stronger representation of the signals can be obtained and, consequently, a more faithful description of the distortions affecting the signal. This might lead to a more precise prediction of the perceived signal quality.

## 1.1 Problem Statement

The area of multimedia quality assessment is a multi-disciplinary area, which combines knowledge from several domains, such as psychology, physiology, image and audio signal processing. Although the specific area of Visual Quality is fairly mature [4, 7, 17], there are still several challenges to be solved in the broader area of multimedia quality. In particular, as pointed out by Pinson et al. [18], the issue of simultaneously measuring the quality of multimedia contents (e.g. video, audio, and text) is still an open problem. In the simpler case of audio-visual content, some work has been done on trying to understand audio-visual quality, what resulted in a couple of subjective models [13, 14] and a few audio-visual objective quality metrics [16, 19, 20, 21]. But, so far, few works have studied the interaction between different audio and video components [22, 23, 24], a research topic that has become very relevant given the popularity of audio-visual content.

Modelling how humans perceive audio and video signals is a challenging task. This gets even more complex when the interaction between audio and video requires a mathematical model to represent it. The difficulty lies in the little knowledge about the cognitive processing that humans use to interpret the interaction of this stimulus. This interpretation is key in order to develop an accurate audiovisual quality assessment model. Considering these issues, Machine Learning paradigms arise as an appealing option to tackle the audiovisual quality assessment problem from a different perspective. Quality assessment methods based on ML are capable of mimicking human reactions to media distortions, instead of explicitly modelling it. Traditionally, methods that are based on ML are composed of two basic stages: (1) the computation of features describing the media distortion and (2) a mapping of those features into quality scores. As a result, the model learns the complex non-linear function that maps features into quality scores. Some important aspects need to be covered in order to successfully model these complex mapping functions. These aspects are the definition of the feature set that describes the signal and the ML tool to implement the mapping function.

Audio and visual descriptive features have been studied for several years and they have been applied to different research fields, such as speech intelligibility and pattern recognition. Its effectiveness relies on how good they are able to describe the signal characteristics in terms of human perception. For the quality assessment field, several audio and video quality metrics have exploited these features in order to predict the perceived quality with very good results. As for the ML tool, it is key to select the technique that best suits the assigned task. For the particular task of finding a way to describe the audio and video stimulus interaction, Autoencoders can be used to find relationships between both audio and video sets of features. This type of strategy has been successfully used on studies to reduce and find stronger descriptive features.

It is assumed, based on the previous information, that a model composed of a set of audio and video features applied over an autoencoder technique might produce a way to describe the complex interaction between both audio and video stimulus. Given the nature of this approach, its application to an audiovisual quality assessment scenario would represent a valuable contribution.

## 1.2 Proposed Approach

A previous work by Soni et al. presented a deep autoencoder based method for non-intrusive speech quality assessment [25]. The metric adopts a two-layer approach to treat speech background noise distortions and uses audio information in the form of spectrograms. In the first layer, a speech spectrogram is passed on a two-layer autoencoder in

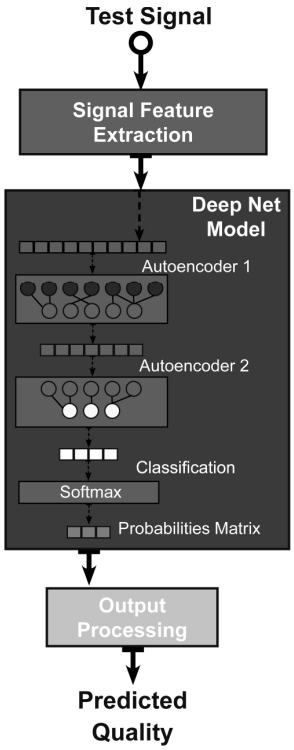



Figure 1.1: Diagram of the proposed Deep Autoencoder Network for audio-visual quality assessment.

order to produce a low-dimensional set of new features. A mapping function between the features and subjective scores is found using an artificial neural network (ANN). Results showed that an autoencoder approach produced better descriptive features than Filterbank Energies (FBEs) and more accurate speech quality predictions [25].

In this work, we aim to extend the idea proposed by Soni et al., adapting it to assess the quality of audiovisual signals. A diagram sketch of the proposed system is depicted in Figure 1.1. First, a set of features that describe the characteristics of the audio and video components are computed. In the next stage, a two-layer autoencoder produces a low-dimensional set of features. At this stage, it is expected that these low-dimensional set of features are able to describe the complex interaction between audio and video stimulus. Then, a classification function maps the features into audiovisual quality scores. Finally, the model output is processed and the overall audiovisual quality is computed.

## 1.3 Document Structure

The structure of this work is organized as follows. In Chapter 2, some basic concepts related to the development of this work are presented. In Chapter 3, a brief revision of the literature regarding the signal quality assessment is presented. In Chapter 4, a set of three subjective quality experiments for this work is described. In Chapter 5, the

proposed audiovisual quality assessment model is described. Finally, Chapter 6 presents some conclusions of this work and the main contributions.

# Chapter 2

# Basic Concepts

This chapter presents some general concepts that are employed in the development of this work. The main idea is to familiarize the reader with the research topic and introduce some relevant information employed on this research. Some basic concepts related to the performance of the human visual and auditory systems are presented, in addition, characteristics of the video and audio digital systems are described. Finally, a brief explanation of some of the machine learning techniques employed in this work is included.

## 2.1 Human Visual and Auditory System

While our sensory system is constantly collecting information from our surrounding environment, it is the way we interpret that information that influences our interaction with the world. Human perception is referred to the way we organize, identify, and interpret any sensory information captured from the surrounding environment [26]. Visual and auditory stimuli are very important for humans because they provide essential environment information and permit a proper interaction between humans and their surroundings. A very brief description of the human visual and auditory perception phenomena is presented in this section.

### 2.1.1 Visual Perception Phenomena

Visual perception is very important for humans, who constantly receive and process information to interact with the surrounding environment. With the objective of obtaining quality predictions that are highly correlated with the quality as perceived as human viewers [27], most video quality metrics take into consideration aspects of the human visual system, and some psychophysical concepts. Next, some of the basic concepts of the human visual system are presented, along with some of its psychophysical characteristics.

### Color Perception

The color perception in the human visual system is related to luminance sensitivity of the photoreceptors (cones and rods) located in the retina. As it was mentioned before, rods are more sensitive to light and are not able to distinguish between colors, i.e. they can only recognize greyscale tones and provide information related to the shape of objects. Given that rods are more sensitive to light, in a penumbra scenario (dim light), only rods are actives, this type of vision is called scotopic vision. On the other hand, on a scenario with great light exposure, cones are more active and this is known as a photopic vision. Finally, a mesopic vision corresponds to a midterm scenario (combination of photopic and scotopic vision) where light exposure is low but not quite dark (0.001 to 3 $cd/m^2$), both rods and cones are active [31].

As mentioned previously, cones are divided into three different classes according to their sensitivity to different bands of the electromagnetic spectrum. This particular organization, also known as the trichromatic theory, is what allows humans to perceive colors. The three types of cones are denoted as (1) Short –S, with a wavelength of 440 – 485 nm, (2) Medium – M, with a wavelength of 500 – 565 nm, and (3) Long – L, with a wavelength of 625 – 740 nm. The relative spectral sensitivity of the cones S, M, and L (presented as a wave-length function) is depicted in Figure 2.1. Depending on the range they occupy on the electromagnetic spectrum, a particular color can be attached to that particular cone: Short (blue), Medium (green), and Long (red).

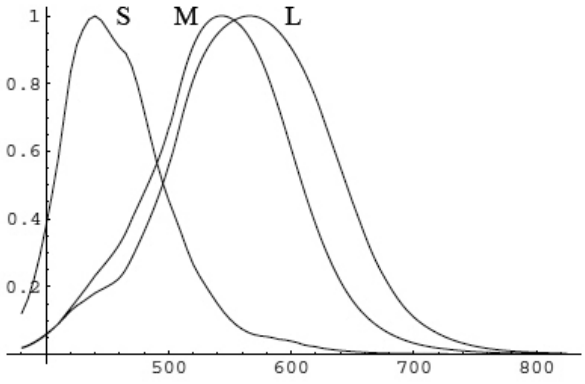



Figure 2.1: Cones relative spectral sensitivity: S (*short*), M (*medium*) e L (*long*). Original illustration from [32].

### Contrast Sensitivity

The ability of humans to perceive details in a particular scene is determined by the capacity of the visual system to detect contrast, i.e. the difference in brightness of contiguous

areas. Contrast sensitivity of humans is represented by the Contrast Sensitivity Function (CSF) [33], which is obtained through subjective experiments with human participants. The contrast value is defined as the radio between the highest and lowest luminance, as defined by the following equation:

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

The curve of the contrast sensitivity versus the spatial frequency (gathered from experiments) is presented in Figure 2.2. The graph measures the contrast threshold gathered from comparing two stimulus at different spatial frequencies.

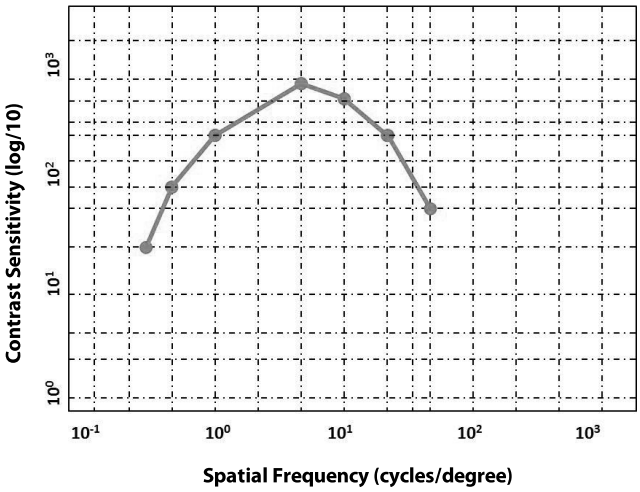



Figure 2.2: Curve of the Contrast Sensitivity versus the Spatial Frequency (Adults) [34].

### 2.1.2 Auditory Perception Phenomena (Psychoacoustics)

Human's sense of hearing is one of the most complex and important systems in the human body. It allows the interaction between humans and the surroundings, as well as with other humans through speech or any particular sound [35]. It depends basically on the processing of vibrations (which produces sound waves) and its later interpretation by the human brain. Despite its complexity, a tremendous research effort has made possible to understand its mechanism and, later on, modelling it for computational simulations [36]. As with the visual quality metrics, several audio quality metrics or methods employ aspects of the human auditory system with the objective of predicting the audio quality perceived by humans [22]. Following, we present some basic concepts regarding the human auditory system along with a number of psychoacoustic characteristics.

### Psychoacoustics

Psychoacoustics studies how humans perceive different types of sound. More particularly, it covers the human psychological and physiological responses to sound phenomena (speech, music, environmental sound, etc.) [37]. These responses are mostly determined by the sound's wave frequency and amplitude. The frequency of a sound is referred to the number of waves that pass a certain point in a given time. Meanwhile, the amplitude is defined as the difference between the high and low pressures created in the air by that sound wave. Sound itself can be analyzed by means of a number of characteristics related to its frequency and its amplitude. Next, some of these characteristics are briefly described. Figure 2.3 depicts some waveform representations of them.

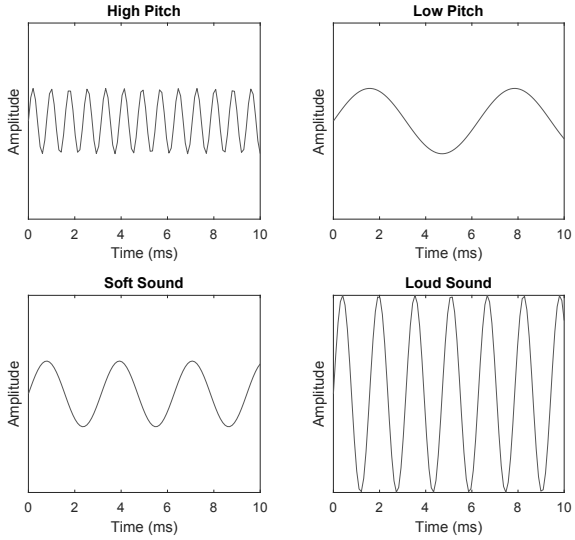



Figure 2.3: Waveform representation of some psychoacoustics.

- *Pitch:* Pitch is a term used to describe the perceived frequency of a sound, i.e., how high or low frequency present in the sound signal. For instance, a high pitch sound is the result of short waves passing very fast by a certain point, while fewer slower waves result in a lower pitch sound. That is, the pitch of a sound will be determined by its frequency. Sound waves occurring at fairly consistent frequencies will be perceived as having a definite pitch (musical tones), in contrast sound waves that present irregular frequencies will be perceived as having an indefinite pitch (noise).

- *Duration:* Duration is a term used to describe how long or short a sound is. It is related to the time length from the moment the sound has been perceived until the moment the sound has been identified as changed or ceased.

- *Loudness:* Loudness refers to how loud or soft listeners perceive a particular sound. How loud the sound is will depend (partially) on the sound wave's intensity. Intensity is a measure of the existing energy in the sound waves and it is directly proportional to the square of the sound wave's amplitude. The higher the amplitude the higher the volume of the perceived sound. On the other hand, a smaller amplitude will be associated with a softer sound.

- *Timbre:* Timbre can be interpreted as the quality of different sounds sources (e.g., people clapping, a train scraping on tracks, a musical instrument, human voice, etc.). The sound timbre describes characteristics that make possible to humans distinguish between different wave sounds with the same pitch and loudness. For example, the timbre will permit differentiating the sound of a flute and a clarinet playing the same note at the same volume.

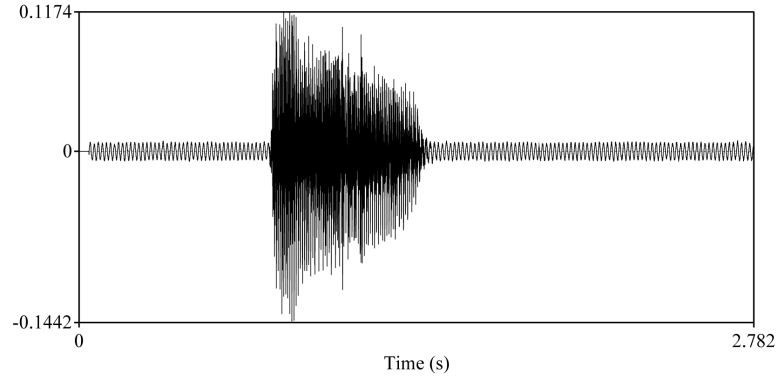
Different audio representations present audio properties in a different way. Spectrograms are capable of representing some audio characteristics that help the analysis of some type of distortions. This makes them particularly relevant to the audio quality assessment area.

### Spectrograms

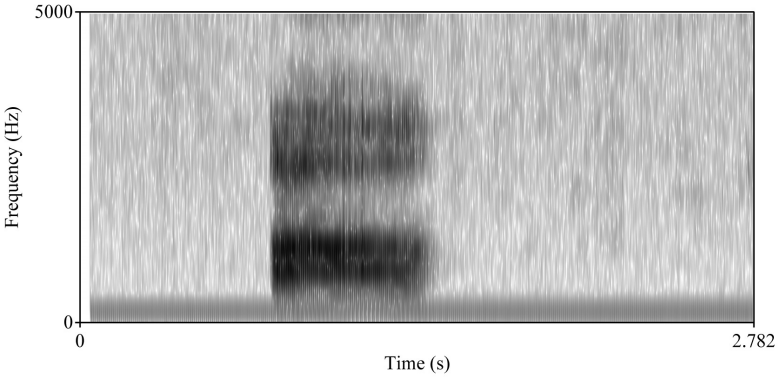
Spectrograms are the visual representation of sounds (or any signal) which displays the amplitude of the frequency components over time. This representation is obtained by using a Fast Fourier Transform (FFT), which basically decomposes the signal into their frequency components. Figure 2.4 presents a spectrogram of a vowel letter 'a' along with its waveform representation.

A spectrogram representation offers several advantages compared to other types of visual representations like waveforms [36]. For instance, complex signals that contain more than one frequency component are more easily analyzed using spectrograms. A Spectrogram displays time on the horizontal axis, and frequency on the vertical Y-axis (pitch). Additionally, the volume is represented by color depending on the color scheme used by the spectrogram.

Spectrograms are a useful tool to analyze the timbre of a sound due to its overtones way of representation. Due to the human speech mechanism that permits using the shape of the mouth to produce different overtones, spectrograms representations can be exploited in the analysis of human speech signals. Figure 2.5 presents a spectrogram comparing vowels sounds 'a', 'e', and 'o', where differences on the overtones are easily observed. Furthermore, spectrogram representations have been applied into research areas such as speech intelligibility, speech quality, and audio quality with very good results [38, 39, 40].




(a) 'Waveform'



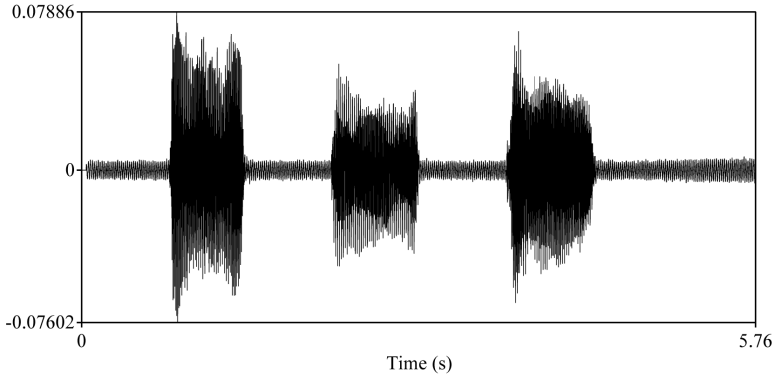

(b) 'Spectrogram'

Figure 2.4: Waveform and Spectrogram representation of vowel 'a'.

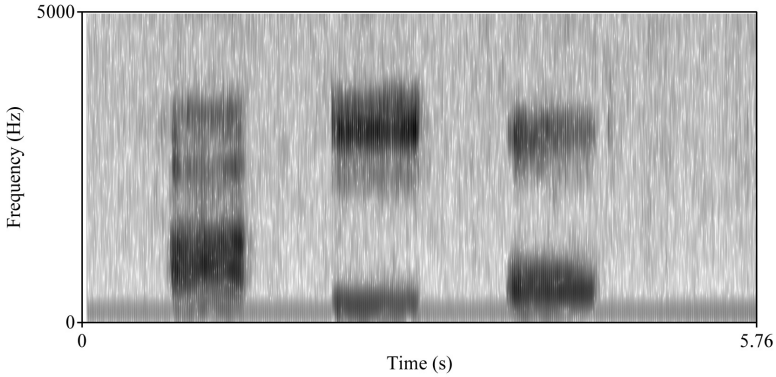
## 2.2 Digital Communication Systems

In its most basic form, a digital communication system is formed by three main entities: a transmission unit, a receiver unit, and, in between, a communication channel. Its main purpose is to transfer information from a source to a recipient through a channel or medium. This section will briefly describe the stages of a common digital communication system, along with some basic concepts related to the digital processing of signals. Furthermore, two types of information, which are the main focus of this work, are described in detail: the video and audio digital systems.

Figure 2.6 presents a basic diagram block of a digital communication system. As a starting point, the first block is called the digital signal (or message source), which in this context, is interpreted as a binary representation (0s and 1s) of the data to be transmitted (e.g., human speech in a digital form). Such digital signal is passed on as input to the source-coding block. At this stage, it is known that the signal samples are highly correlated, i.e., differences between nearby samples are very little. This property is exploited on this block in order to reduce the bitrate of the transmitted signal. In practice, there are two types of approaches to deal with this task: a lossy coding approach (some




(a) 'Waveform'




(b) 'Spectrogram'

Figure 2.5: Waveform and Spectrogram representation of vowels 'a', 'e', and 'o'.

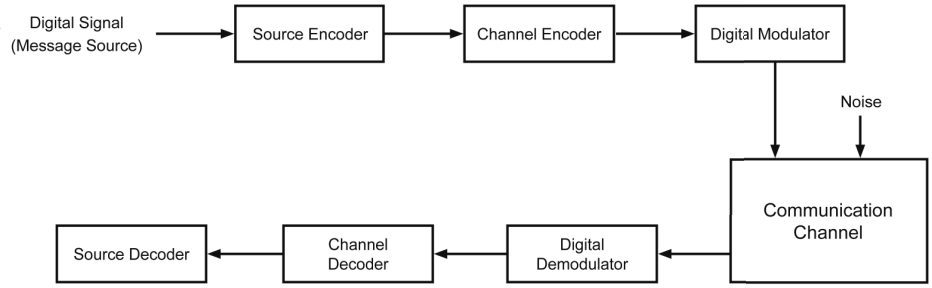



Figure 2.6: Basic diagram block of a digital communication system.

level of degradation is accepted), and the lossless coding [41]. Next, the output digital signal, with a reduced bitrate, is passed on to the channel-coding block.

The main purpose of the channel-coding block is to safeguard the information transmitted. Its task is to ensure, as much as possible, that there is no error in the transmitted information when the signal is recovered and delivered to the end user. In order to complete this task, some additional information is included in the signal; this is broadly known as error control coding. Next, the digital information sequence, along with some safeguard against possible errors, is passed on to the digital modulator block. In this stage the digital signal is transformed into analog continuous pulses that are transmitted

through the air, or the communication channel. The channel is either wired or wireless, i.e. a co-axial cable, an optical fiber, the air, or the space. At this stage, the digital signal is exposed to transmission errors and noise. From there, the transmitted signal, in form of analog pulses, passes through a similar circuit in order to go back to its original form: demodulator, channel decoder, and source decoder.

In the following lines, some important characteristics of the digital representation of video and audio are presented, along with some common artifacts that result from its digital processing and transmission.

### 2.2.1 Video Digital System

#### Video Coding

Data coding is an important task in the processing of digital video signals. The main objectives are to reduce the amount of required storage space and to facilitate its transmission through a stablished communication channel. In order to achieve these objectives, several strategies are employed to compress the digital information with a minimum effect in the quality of the processed data. There are two types of compression techniques used with digital video: lossles and lossy. A lossless compression technique offers a perfect reconstruction of the original signal. However, its compression rate is very low. A lossy compression technique, on the other hand, offers a higher compression rate, which is of great value for video digital signals. Yet, this type of strategy carries the loss of information from the transmitted signal, which might affect the quality of the compressed signal [42].

Most broadly used coding standards are mainly developed by two international agencies: International Telecommunications Union (ITU) and the International Standards Organization (ISO). This last one through two sub-groups, the Joint Photographic Experts Group (JPEG) and the Moving Pictures Experts Group (MPEG). Table 2.2 presents a summarized comparison of the above mentioned video coding standards.

MPEG-1 is known as the first, lossy compression, coding standard developed by the MPEG. This standard is considered as being highly compatible and it is still being used for compression using Compact Disks Read-only Memory (CD-ROM). The MPEG-2 was the second coding standard developed by the MPEG. This format is commonly used for transmission of digital television signals, as well as movies and software distributed in Digital Video Discs (DVD). Although newer standards are more efficient, MPEG-2 is still very much used due to its backwards compatibility with existing hardware and software [43].

Over the years the ITU presented several video coding standards grouped in the H.26x family. The first coding standard used in practical terms was the H.261. Design improvements in new coding standards led H.261 to be almost obsolete, however, it is still used as a backward compatibility feature in several video conferencing systems [44]. Another important coding standard is the H.263, also a member of the H.26x family. This standard was originally designed as a low-bitrate format for compressed signals in videoconferencing transmissions. It was also used in several internet applications in the form of Flash videos. Due to the advances in new standards, H.263 is now mainly used as a compatibility feature in the implementation of newer standards [44].

A collaboration effort between ISO and ITU agencies resulted in the H.264 video coding standard, also known as MPEG-4/AVC (Advanced Video Coding). This standard is, by far, the most commonly used video coding standard nowadays, as well as the most widely supported. The H.264 provides a significant better compressing rate compared to its predecessors (almost 50 percent at a similar quality cost) [45].

The H.265 standard, also known as High-Efficiency Video Coding (HEVC), was developed by the ITU as the successor of the H.264/AVC standard. H.265 reaches a compression rate that is almost double the value achieved by the H.264 coding standard at the same level of video quality [46]. This particular feature is very important for video resolutions above 2K, as well as for high-quality video streaming. However, its coding process is much more complex and it requires much more resources [47]. Although adoption of HEVC is growing, it is still far from being as popular as H.264.

Table 2.1: Summarized comparison of some video coding standards.

| Year | Standard | Agency | Implementations |
|---|---|---|---|
| 1988 | H.261 | ITU | Videoconferencing, videotelephony |
| 1993 | MPEG-1 | ISO-MPEG | Video CD-ROM |
| 1995 | MPEG-2 | ISO-MPEG | Digital Television Transmission, Video DVD |
| 1996 | H.263 | ITU | Videoconferencing, videotelephony, mobile-phone videos |
| 2003 | MPEG-4/AVC (H.264) | ITU, ISO-MPEG | High Definition DVD, Digital TV, videoconferencing, Blu-ray, iPod Video |
| 2013 | HEVC (H.265) | ITU | Ultra HD Blu-ray, UHD streaming |

### Common Artifacts

For the present context, a video artifact is defined as an unwanted characteristic present in the video signal that might affect the quality of the signal perceived by a particular user. Artifacts might be introduced to the video signal during capture, coding, transmission, reception, and delivery to the final user, as it is shown in Figure 2.7. Because of this, one important requirement for each of these stages is to keep the negative quality impact at a minimum in order to maintain a certain level of satisfaction. Next, some of the most common video artifacts are listed and briefly described.

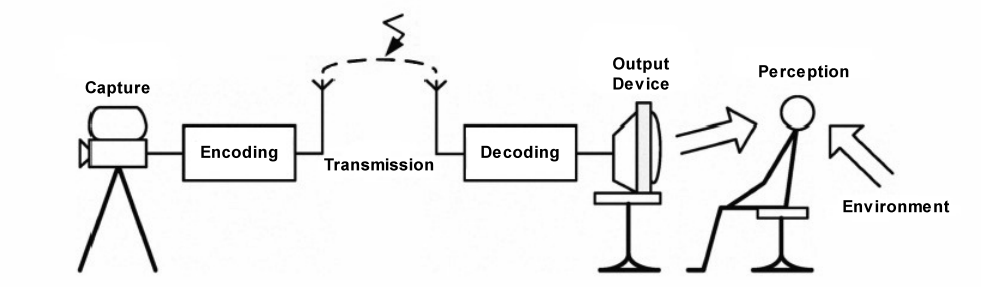



Figure 2.7: Video signal on the production processing phases.

**Blocking** This type of artifact is considered the most common visual degradation. Figure 2.8 presents a sample of a video frame containing a blocking artifact. It is produced due to the division of frames into macroblocks of rectangular shape. These macroblocks are coded separately from one another without considering the existence of spatial correlation between them, as a result, horizontal and vertical borders appear [48].

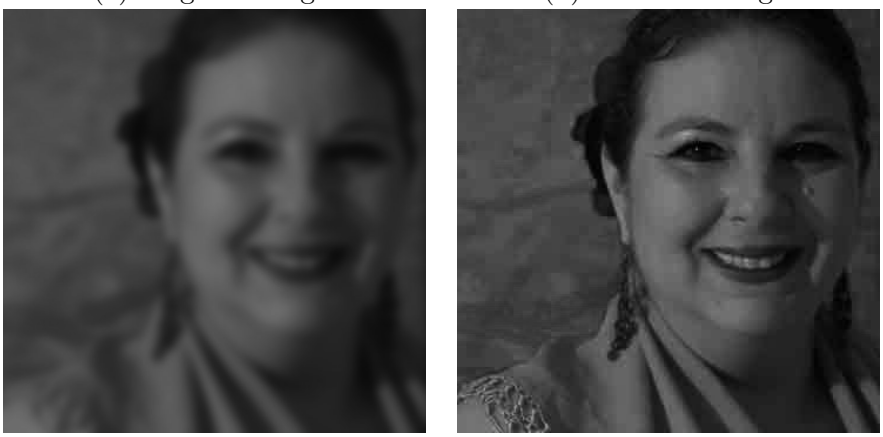
**Blur** A blur distortion is shown as a reduction of edge sharpness and spatial detail [49]. Figure 2.8 presents a sample of a video frame containing a blur artifact. It can be introduced during the processing phase of the video (coding) as a result of a loss of high frequency information. In addition, strong de-blocking can expose blurring artifacts during the attempt to flatten block edges [48].

**Ringing** Ringing is a common form of artifact which are perceived as "halos" around sharp edges [48]. Figure 2.8 presents a sample of a video frame containing a ringing artifact. They are visible for most compression techniques, especially when the signal is transformed into frequency domain. This distortion is also known as a Gibbs Phenomenon. Ringing results from a poor reconstruction of pixel values and is more noticeable along high contrast edges. This effect is stronger if the edges are located in areas with a generally smooth texture [50].

**Block Loss** Block loss artifacts are characterized by the presence of one or several flat color blocks in the video frame. The effect is caused by the loss of data packets during the transmission stage. These blocks might also be substituted by an approximation of the original blocks if an error concealment algorithm is used [21]. Figure 2.8 presents a sample of a video frame containing a Block loss artifact.

**Blackout** A blackout causes the whole frame to disappear. It is produced when all data packets of a frame are lost during the transmission. It can also be a consequence of incorrect video recording [21].

(a)‘Original Image’ (b) ‘Blocked Image’



(c)‘Blurred Image’ (d) ‘Ringing Effect’

(c)‘Blockloss Effect’ (d) ‘Slicing Effect’

Figure 2.8: **(a)** Original Image. **(b)** Image containing a Blocking artifact. **(c)** Image containing a Blur artifact. **(d)** Image contaning a Ringing artifact. **(e)** Image containing a Block loss artifact. **(f)** Image containing a Slicing artifact.

**Freezing** A frame freezing effect can be categorized as a basic frame freezing or a frame freezing effect skipping. The basic frame freezing effect is composed of time-discrete “snapshots” of the original continuous scene. This effect is also known as jerky motion effect and it is associated with an inadequate sampling or display rate which is commonly

used to accommodate a reduced temporal bandwidth [21]. Meanwhile, a frame freezing without skipping corresponds to a pause of the video that does not discard any of the following frames. This effect is produced when the available throughput is lower than the bitrate of the media and, as a result, the media stalls until enough data is downloaded. When the pause occurs before the media starts playing this freezing is known as "initial loading". But, when the pause occurs in the middle of the media reproduction it is known as 'stalling' [51].

**Slicing** This artifact appears when a limited number of video lines (stripes) is severely damaged. The artifact is caused by a loss of video data packets. The decoder replaces the lost slices by using previous slices [21]. Figure 2.8 presents a sample of a video frame containing a slicing artifact.

### 2.2.2 Audio Digital System

As in video digital coding, the main objective of audio digital coding is to reduce the bitrate of the audio signal in order to reduce storage space and, more importantly, to facilitate signal transmission. Lossy and lossless compression techniques, as it was explained before, are the approaches used to develop audio compression algorithms (audio codecs). The basic requirements for such techniques, besides a low bitrate, are robustness against random channel errors (packet loss) and low encoder/decoder delays, all of this at a minimum quality impact. As in video coding, a temporal redundancy is exploited to achieve lower bitrates. This type of redundancy is also called inter-sampling redundancy. In general, information redundancy is reduced employing different types of methods like coding, pattern recognition, and linear prediction [35].

In order to provide higher compression rates at a low fidelity cost, lossy compression algorithms take advantage of some psychoacoustics characteristics. Consequently, the fidelity of less audible sounds is sacrificed to reduce the size of the data for storage and transmission. On the other hand, lossless compression algorithms are capable of producing signal representations that can be decompressed to the exact digital copy of the original audio signals. However, they can only achieve limited compression rates (around 50 – 60 percent of the original) due to the complexity of the waveforms and its rapid variations in sound forms [52].

Over the years, the MPEG working group, mentioned previously, presented some important multimedia coding standards, most of them performing both audio and video coding. The MPEG-1 audio part, developed for the CD-ROM quality multimedia storage, is sub-divided into three layers. These three layers are increasingly complex and efficient. MPEG-Layer III, also known as MP3, is one of the most famous (and widely

supported) audio codecs in the market. This lossy audio codec exploits the limitations of human hearing in order to achieve very high compression ratios at a minimal quality impact [53]. Despite new audio coding improvements, MP3 continues to be a widely popular format for sharing and playing audio content. The MPEG-2 audio part, also known as Advanced Audio Coding (AAC), targets HDTV applications. Compared to MP3, some of the benefits of AAC are its widely support and better sound for the same bitrate. These features made AAC the most popular audio codec for videos. As for surround experiences, the Dolby laboratories presented the AC-3 audio standard, which fully preserves surround sound settings making it very popular for movie theaters and high fidelity musical equipment. Table 2.2 presents a summarized comparison of the above mentioned audio coding standards.

Table 2.2: Summarized comparison of some audio coding standards.

| **Year** | **Standard** | **Agency** | **Implementations** |
|---|---|---|---|
| 1993 | MPEG-1 Layer 3 (MP3) | ISO-MPEG | Audio CD-ROM |
| 1995 | MPEG-2 (AAC) | ISO-MPEG | Audio DVD, audio streaming |
| 1999 | AC-3 | Dolby Laboratories | Cinema, TV broadcast |

### Common Artifacts

Video and audio digital signals go through similar processing phases, as a result, they are affected by similar errors that might occur during such phases. However, these errors have a different impact on the actual data that is transmitted, resulting in different types of artifacts that affect the perceived quality of the transmitted signal. For this particular work, a few types of audio degradations have been considered common in a voice over IP transmission environment. These artifacts are platform independent, that is, they occur independently of the codec, hardware, or network [54]. Next, these artifacts are listed and briefly described.

### Background Noise

Background noise is described as any sound other than the sound being monitored. It can be characterized as stationary or non-stationary background noise. Non-stationary noises are commonly found in our sound environment, like traffic noise, alarms, and people talking. Audio signals can also be corrupted by static noise in the transmission channel. For example, additive white Gaussian noise can interfere with a signal by spectrally masking its features [55].

Figure 2.9 presents the spectrogram of an original audio file and its distorted version. The audio clip corresponds to two sentences from a male speaker, separated by a silence of

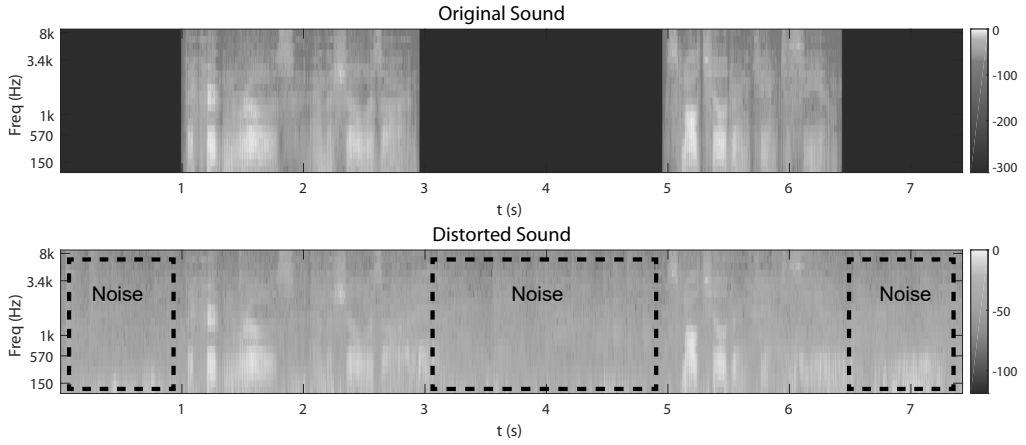



Figure 2.9: Spectrogram representation of Background-noise distortion (original versus distorted).

two seconds. Additionally, a one-second silence at the beginning and at the end of the clip is present in the audio signal. Figure 2.9 shows the spectrogram of a clean audio file and of the same file affected by background noise. Notice that the background noise occupies the silence gaps between sentences. Moreover, the actual signal suffers some variations due to the noise added to the signal.

### Chop Speech

This type of degradation consists of speech signals in which samples are missing. Regarding the VoIP scenario, choppy speech is referred to speech that is affected by missing samples. This is commonly caused by packet loss in the VoIP network. Packet Loss Concealment (PLC) can be used to smooth the effects of the missing samples. As a result, missing samples are replaced by either silence, previous samples repeated, or they are simply skipped [54].

Figure 2.10 compares the spectrogram of a clean audio file against the same file affected by chop speech. By observing both spectrograms, we can notice the missing samples in the distorted version of the sound. These samples are illustrated as vertical lines in the middle of the signal and they represent the chop in the audio.

### Mute

Mute might be the audio equivalent of the Block loss artifact. Interruptions such as mutes are among the most common distortions produced by packet loses. The detection of mute artifacts depends on two thresholds: (1) the minimum level of signal noticeable by the human ear and (2) the duration of the shortest silent interval perceptible as a

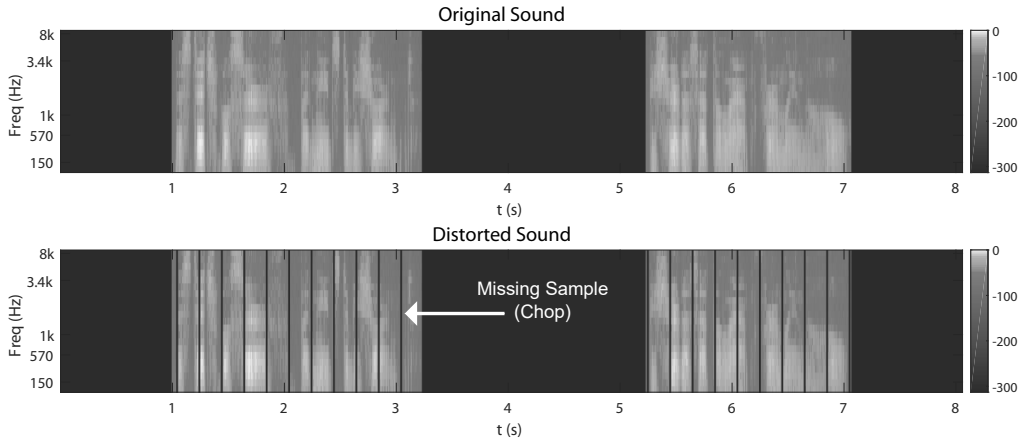



Figure 2.10: Spectrogram representation of Chop distortion (original versus distorted).

mute artifact. It is assumed that the spectrum of the audio signal is inside the hearing frequency range(20 Hz and 20 kHz) [56].

### Clipping

A digital audio signal can be subjected to a clipping process in situations in which the amplitude of the signal exceeds a maximum intensity level. Clipping consists of attenuating the incoming signal amplitudes to maintain them below the maximum allowed intensity level. As a result, unwanted effects such as intermodulation, aliasing, and harmonic distortions are inserted [57]. Also, the presence of additional frequency components might reduce the perceptual quality of the audio signal. During a VoIP call, amplitude changes can arise due to a person's high voice volume when speaking into the microphone.

Figure 2.11 presents the spectrogram of a clean audio file and of a file affected by a clipping distortion. It can be observed that the distorted sound presents higher intensity, which is the result of adding a frequency component to the original sound.

### Echo

An echo effect is a reflection of sound, arriving at the listener some time after the original sound. Echo effects in a voice call generally occur due to the transmitted speech being picked up in the receiving unit's microphone, creating a feedback loop. Strategies for echo cancellation [58] are not completely effective since they create their own problems in the audio signal. The ITU recommendation G.131 [59] offers guidance on how to mitigate talker echo in transmission systems.

Figure 2.12 depicts the spectrogram of a clean audio file and of a file affected by an echo type of distortion. By comparing both spectrograms it can be observed that the distorted

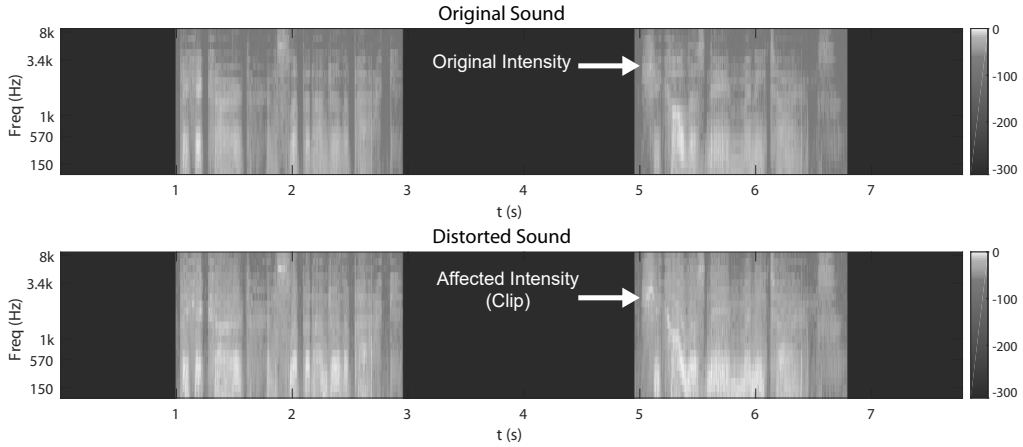



Figure 2.11: Spectrogram representation of Clipping distortion (original versus distorted).

signal presents a certain type of propagation (signal repetitions), which was produced by the echo. Additionally, these repeated signals occupy the gap silences between sentences and they change the duration of the original sound.

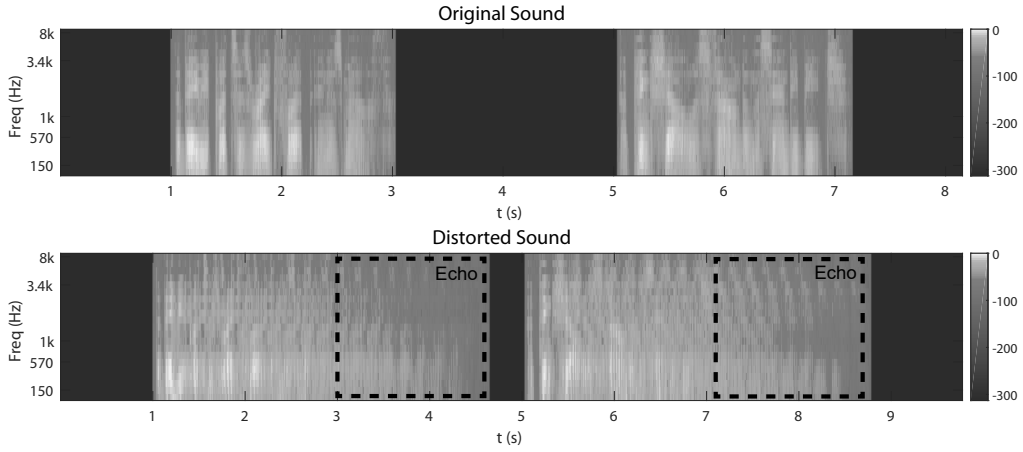



Figure 2.12: Spectrogram representation of Echo distortion (original versus distorted).

## 2.3 Machine Learning

Machine learning paradigms have gained a very important role in several research areas, including multimedia quality assessment [60]. A machine learning approach tackles the quality assessment problem by imitating different aspects of the HVS and HAS, rather than modelling very complex non-linear functions. Regarding its computational demands, most of the resources are only required during the training phase, producing light and fast models [61]. Next, some of the most basic concepts related to machine learning are

presented. Additionally, two techniques, which are used in the present work, are briefly described: AutoEncoders and SoftMax function.

### 2.3.1 Machine Learning Basics

Machine learning algorithms can be defined as the type of algorithms that are able to learn from certain data without being explicitly programmed. A common definition by Mitchell [62] states that: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". Machine learning is mainly focused on developing computer algorithms that are able to teach themselves to evolve and change whenever new data is presented. Considering the rate at which data is growing nowadays, machine learning tools that can help process this data in an efficient and elegant way are very much needed [63].

### 2.3.2 Types of Algorithm

Most machine learning algorithms can be classified into two categories: Supervised (task driven) and Unsupervised (data driven). Next, some characteristics of these categories are briefly described, additionally, Table 2.3 depicts a summarized list of some machine learning algorithms.

#### Supervised Learning

Supervised learning algorithms use labelled data for training. That is, both the input and output are known. The algorithm basically learns by comparing the input data with the correct outputs, minimizing the errors, then it modifies the model accordingly. Supervised learning exploits the data patterns in order to predict the outputs based on the labels used during the training. This type of approach can be used on applications where historical data is able to predict likely upcoming events. Regarding the task of the algorithm, supervised learning algorithms can be classified as Regression or Classification algorithms.

- *Regression:* This type of supervised learning uses the labelled data in order to make predictions in a continuous form. Regression is a form of predictive modelling which investigates the relationship between a dependent variable (output) and independent variables (input). Common applications of this technique are forecasting the weather, time series modelling, and process optimization [64].

- *Classification:* This type of supervised learning uses the labelled data to make predictions in a non-continuous form. The output of the model is not always continuous and the graph is non-linear. A classification technique learns from the input data and then specifies which of the classes a certain input data belongs. One common application is the task of object recognition, where the input is an image (described as pixel brightness values) and the output is a numeric value that represents a certain object [64].

**Unsupervised Learning**

Unsupervised learning uses data without any labels to train a model. The output must be figured out by the algorithm itself. To do so, the algorithm seeks a particular structure in the data. Once the algorithm recognizes the data structure, it makes clusters of data with different labels. This particular approach is commonly used to identify common attributes on a large set of items. These items are then grouped on different clusters that can be treated or classified using some particular criteria. Considering the task they perform, this type of algorithm can be sub-divided into two groups: Clustering and Dimensionality Reduction.

- *Clustering:* Clustering uses unlabeled data in order to group similar entities together by identifying common attributes within the data. Then, the data is organized in clusters depending on its similarity. Once the model is trained it is capable of identifying the cluster that any new data should belong to.

- *Dimensionality Reduction:* This type of algorithm aims to reduce the dimension of the input data by removing irrelevant information from the original structure. This technique identifies the most stronger features, in terms of information, and removes those that are considered to carry less relevant information. This type of technique is very much important for a pre-processing phase of the input data [64]. Its more appealing benefit is that a reduced version of the data (in terms of dimensionality) can be used with very little information loss.

Among the different machine learning tools available in the literature, autoencoders drive the attention due to its capability of finding relationships among a set of descriptive features. Next, some basic properties of this technique are presented.

### 2.3.3 Autoencoders

Data compression is an important topic that is used in computer vision, computer networking, and several other areas. As it was pointed out before, the main goal of compres-

Table 2.3: Summarized list of Machine Learning algorithms.

| Type | Task | Algorithm | Applications |
|---|---|---|---|
| **Supervised Learning** | **Regression** | Simple linear regression<br>Polynomial Regression<br>Support Vector Regression<br>Ridge Regression<br>Lasso Regression<br>ElasticNet Regression<br>Bayesian Regression<br>Decision Tree Regression<br>Random Forest Regression | Weather forecasting,<br>predict housing prices,<br>predicting sales of particular<br>product next month, etc. |
| | **Classification** | K-Nearest Neighbours<br>Support Vector Machines<br>Kernel Support Vector Machines<br>Naive Bayes<br>Decision Tree Classification<br>Random Forest Classification | Customer segmentation,<br>audio and image categorization,<br>text analysis, etc. |
| **Unsupervised Learning** | **Clustering** | K-Means Clustering<br>Hierarchical Clustering | Document classification,<br>customer segmentation,<br>insurance fraud detection, etc. |
| | **Dimensionality Reduction** | Principal Component Analysis<br>Linear Discriminant Analysis<br>Kernel Principal Component Analysis<br>AutoEncoders | Feature selection,<br>image denoissing,<br>audio denoissing, etc. |

sion is to convert input data into a smaller representation. The smaller representation of the data can be used later to reconstruct an approximation of the original version. Autoencoders are unsupervised neural-networks that use machine learning to do this compression [65]. In other words, they are trained with the goal of copying their input to their output. However, copying the input perfectly might not be especially useful, this is why autoencoders are designed so that the copies generated are not perfect copies. This particular design forces the model to prioritize aspects that should be copied, which often leads to learning important properties of the data [64].

Traditional applications for autoencoders include dimmentionality reduction and feature learning. Another very popular technique to deal with the dimensionality reduction is the Principal Component Analysis (PCA). There are several scenarios where using Autoencoders can be a better approach, for instance non-linear transformations like the one depicted in Figure 2.14. Unlike PCA, Autoencoders can learn non-linear transformations by using a non-linear activation function and multiple layers [66]. They are also more efficient in the sense that they can learn from several layers rather than deal with one huge transformation as in PCA [66]. These characteristics made Autoencoders gain attention in several research areas such as data denoising and dimensionality reduction.

Due to its ability to deal with image processing (image compression and denoising applications), autoencoders can also be used to solve audio processing tasks in which an image representation of the audio is used (spectrograms or neurograms). In a work presented by Soni [25], deep autoencoders were used to extract low-dimensional features from a speech spectrum. These features were later mapped to corresponding subjective scores using an artificial neural network (ANN). Results showed that autoencoders were

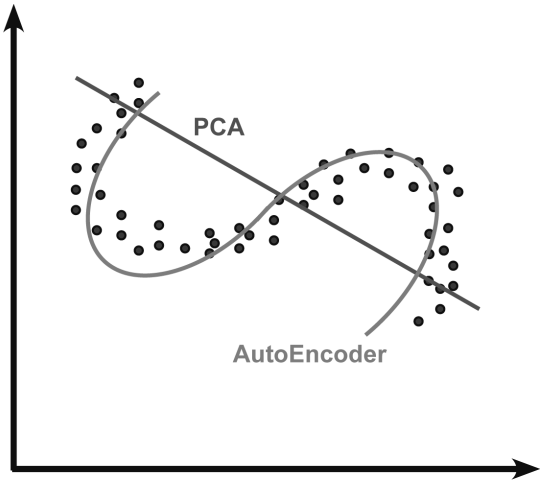



Figure 2.13: PCA versus Autoencoder (linear versus non-linear dimensionality reduction).

able to capture noise information better than Filterbank Energies (FBEs) [67].

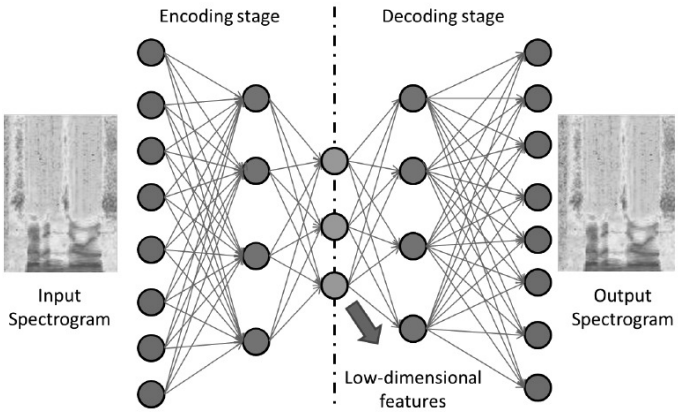



Figure 2.14: Speech spectrum through an autoencoder. Original image extracted from [25].

Most recently, theoretical similarities and connections shared by autoencoders and latent variable models have granted autoencoders an important role in deep generative modeling [64]. They have become a very powerful tool to build deep models and to solve complex tasks, such as audio and speech processing. It is also a very appealing approach to solve some other problems, such as the ones related to the video quality assessment.

Figure 2.15 presents a basic structure of an Autoencoder depicting its three components: Encoder, Code, and Decoder. In an Autoencoder, middle layers are inserted between the input and the output. These layers have a lower dimension compared to the input data. These three components of the autoencoder are described next.

- *Encoder:* The Encoder is the first component of the autoencoder, its task is to compress the input into a latent space representation. The encoder layer produces a compressed representation in a reduced dimension that is usually a distorted version

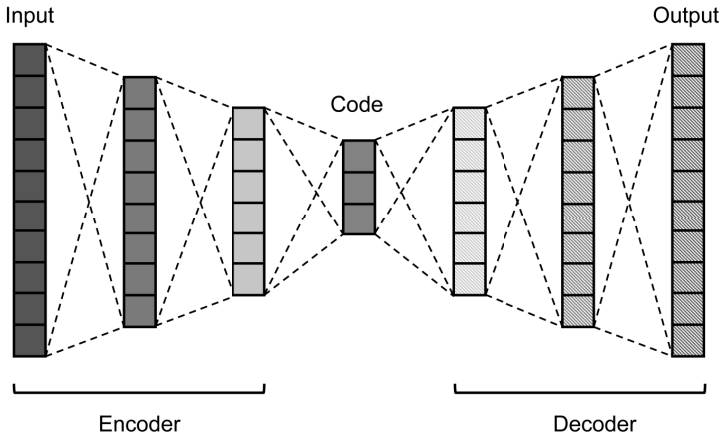



Figure 2.15: Basic structure of an Autoencoder.

of the original data. The encoder is basically a neural network that receives as input $x$, where $x$ is a vector of the form $x \in [0,1]^d$, with $d$ as the input's dimension. Mathematically, the encoding operation can be represented as follows:

$$\mathrm{z} = f_\theta = a(xW + b) \tag{2.2}$$

where the output $z$ is a hidden representation of $x$ with a dimension $d'$, such that, $d'$ is lower than $d$. This representation is associated with a variable $\theta = \{W, b\}$, where $W$ and $b$ represent the weights and the biases of the network, respectively. Finally, $a$ is the activation function applied to every neuron in the layer.

- *Code:* The next layer represents the latent space, which is a layer known as the Code. It represents the compressed input that is fed to the next layer.

- *Decoder:* The third layer is called the Decoder. Its main function is to decode the input back to its original dimension. The decoded data is a lossy reconstruction of the original input. The decoder is another neural network that receives as input the representation $z$ and produces an output function whose parameters are optimized to make the output as close as possible to the input $x$. Mathematically, the decoding procedure can be represented as follows:

$$\mathrm{x}' = g_\phi = a(zW' + b') \tag{2.3}$$

where $x'$ is associated with a variable $\phi$ that represents the weights ($W'$) and biases ($b'$) of the network.

An Autoencoder is a type of model that seeks to minimize the reconstruction error between the input value $x$ and the reconstructed value $x'$. Consequently, the training process will focus on the minimization of a loss function $\mathcal{L}(x, g(f(x))$. In this context, a loss function is a measure of how good a model predicts an expected outcome value. Regarding the objective task they perform, loss functions can be classified into Classification and Regression.

Additionally, sparsity can be encouraged to the autoencoder by adding a regulariser term to the loss function. Sparsity is important whenever the model is required to perform another task, for instance, classification. Then, the new cost function is formed by the loss function plus a sparsity term, which is given by the following equation:

$$\mathcal{L}' = \mathcal{L} + \beta \cdot \Omega_{sparsity}, \tag{2.4}$$

where $\beta$ is the coefficient for the sparsity regularization term. Moreover, an $L_2$ regularization can be added when training a sparse autoencoder. Adding an $L_2$ regularization term to the cost function prevents the sparsity regulariser get smaller whenever the associated weights increase and the $z$ values decrease [68]. The resulting cost function is given by the following equation:

$$\mathcal{L}'' = \mathcal{L} + \beta \cdot \Omega_{sparsity} + \lambda \cdot \Omega_{weights} = \mathcal{L}' + \beta \cdot \Omega_{sparsity}, \tag{2.5}$$

where $\lambda$ is the coefficient of the $L_2$ regularization term. Depending on the implementation characteristics, autoencoders can be organized in several classes. Next, some of these types of autoencoders are briefly described.

### Types of Autoencoders

Although the objective of an autoencoder is to approximate its output to its input, in practice, it is expected that, as a result of the training, the representation $z$ holds some useful properties. In order to achieve this, $z$ is forced to have a lower dimension compared to $x$. This type of representation constrains the model to capture the most salient features of the data.

For an *Undercomplete Autoencoder*, the learning process is restricted to minimizing the loss function $\mathcal{L}$. One particular problem with undercomplete autoencoders is that if the encoder and decoder are given too much power capacity, then the model can perform the copying without learning useful information about the data distribution. *Regularized autoencoders* employs the loss function in order to make the model find other additional properties instead of just copying its input.

A *Sparse Autoencoder* is basically a type of autoencoder that includes a sparsity penalty $\Omega(z)$ in its training criterion (See Equation 2.4). Sparse autoencoders are commonly used to learn features that are going to be used for another task, such as classification. Autoencoders that have been regularized using sparsity penalties are trained to respond to certain statistical features of the data that it is been trained on.

*Denoising autoencoders* modify the reconstruction error function, instead of just adding a sparsity penalty $\Omega$. It is understood that by changing the reconstruction error term the model might learn some useful information. Taking as basis the loss function $\mathcal{L}$, a denoising autoencoder uses the term $\mathcal{L}(x, g(f(\hat{x}))$ to force the denoising, where $\hat{x}$ is a corrupted copy of $x$. As a result, denoising autoencoders are trained to undo this corruption instead of just simply copying their input. Denoising autoencoders represent a good example of how different useful properties can arise from varying the loss function associated with the training model.

#### Training Autoencoders

There are four parameters that can be set before training an autoencoder: code size, number of layers, loss function, and number of nodes per layers.

1. *Code size:* The code size represents the number of nodes in the middle layer (also named Code). In other words, it is the target dimension of the input data, a smaller size will result in a higher compression rate.

2. *Number of layers:* This parameter defines how many times the input data will be encoded (and decoded). That is, it sets the number of encoding, as well as the decoding procedures. The number of layers sets how deep the autoencoder is going to be.

3. *Loss Function:* Among the different loss functions, two of the most important are: the Mean Square Error (MSE) and the Binary Cross Entropy. If input values are in the range of 0 and 1, it is common to use the binary cross entropy, otherwise, MSE is selected.

4. *Number of nodes per layer:* The number of nodes per layer decreases with each subsequent layer in the encoder and increases back in the decoder. The decoder and encoder are symmetric in terms of the layer structure.

### 2.3.4 Softmax Function

Commonly, machine learning classification task relies on functions that calculate probabilities to predict a target class. The softmax function is a very popular technique to deal

with the classification problem. Basically, a softmax function calculates the probability distribution of a particular event for $n$ different possible outcomes. This distribution helps to estimate the corresponding target class for a given input. A softmax function has a range of output probabilities in the interval $[0, 1]$, with the sum of all probabilities being equal to 1. Then, in a multiclass problem, the class with the highest probability will be the target class.

The mathematical representation of the softmax function is given by the following equation:

$$\text{softmax}(\text{x}_\text{i}) = \frac{exp(x_i)}{\sum_{j=1}^{n} exp(x_j)}, \tag{2.6}$$

where the numerator represents the exponential function of a given input value and the denominator is the sum of all exponential values of the inputs.

Softmax functions are commonly used in several multiclass classification models, like softmax regression, multiclass linear discriminant analysis, naive Bayes classifiers, and artificial neural networks (ANN).

# Chapter 3

# Multimedia Quality Assessment Methodologies

As new types of codecs, distribution schemes, and application scenarios evolve, the quality assessment of different types of media signals (audio and video) become an even more significant issue in consumer electronics. Since the emerge of a Quality of Experience (QoE) approach, the traditional Quality of Service (QoS) approach is no longer the only measurement technique for media signals. A QoE approach takes into account (in addition to QoS features) characteristics of the Human Visual System (HVS) and the Human Auditory System (HAS). The usage of this type of approach resulted in several objective quality metrics for digital TV [69], lower-resolution video [70], speech [59], or audio signals in general [71]. The performance of these quality metrics is gauged by measuring their correlation with human quality responses. Human perceived quality is assessed by carrying out subjective experiments, where a group of human participants is asked to rate a series of signal stimuli (audio, video, or audio-visual) using a particular scale. Recommendations for conducting these subjective experiments have been published by telecommunication agencies (ITU, EBU) and research organizations (Video Quality Experts Group - VQEG). Although these experimental recommendations are widely accepted and used, they have trouble representing an authentic user experience. This is why several researchers have modified or created unique methods to deal with these particularities. The immersive methodology proposed by Pinson [72] seeks to put the human participant into a more natural scenario and obtain results that are more realistic.

Since the main goal of objective quality metrics is to provide quality estimates that are highly correlated with subjective responses, it is expected that the usage of this new immersive approach in the development of objective quality metrics will result in more accurate quality predictions. Moreover, subjective experiments that apply this methodology and assess the overall multimedia perceived quality are key to develop quality metrics that

include all media components involved. However, by revising the current literature, it is possible to observe that most quality assessment research has been focused on individual components (audio and video separately) [24]. On the other hand, very few proposals deal with the audiovisual problem from an integrated perspective. Several authors tackled the audiovisual quality problem by combining individual audio and video objective responses [14, 15, 73, 74]. This type of approach serves as a starting point to understand how audio and video interact and how the overall audiovisual quality is perceived. However, given its low complexity level they are far from modeling the quality perception of a multimodal process that involves both visual and auditory human systems. As an alternative to these type of limitations, the quality assessment problem has been tackled from a different angle using machine learning algorithms. This new type of approach exploits the descriptive features used by several objective metrics to model the complex, non-linear mapping functions between signal features and their quality scores [61]. The development of new machine learning algorithms, as the rise of stronger descriptive audio and video features helps understand the complex interaction between both modalities and promotes the development of more accurate audiovisual quality metrics.

The remainder of this chapter is divided as follows. Section 3.1 presents a brief revision of some subjective quality assessment methodologies in the literature. The basic structure of these methodologies is discussed and a description of the Immersive methodology is presented. In Section 3.2, the state of the art of various signal quality metrics (video, audio, and audio-visual) is presented. Moreover, two important video and audio quality metrics, that are the base for this work, are described in detail.

## 3.1 Subjective Quality Assessment

Traditional subjective experiments usually consist of presenting a great number of test sequences to a set of observers. Often, a very narrow range of contents is used. These experimental methodologies generally cause fatigue and content memorization, which may generate less accurate rating results. Pinson et al. [72] proposed an immersive experimental methodology to tackle these problems. The proposed immersive methodology increases the content diversity (number of original sequences) and makes sure that each original content is viewed, or heard, only once by each participant.

### 3.1.1 Traditional Methods

Over the years, many different subjective test standards have been published by information and communication agencies. Two of the most important around the world are the International Telecommunications Union (ITU), and the European Broadcasting Union

(EBU). Depending on the target application, several recommendations have been published. For example, ITU-T Rec. 910 and ITU-R Rec. BT.500 establish subjective assessment methods for evaluating video quality [75]. Similarly, EBU developed SAMVIQ (Subjective Assessment Methodology for Video Quality) to assess the quality of video codecs in Internet applications [76]. For the goal of measuring speech and audio quality, ITU proposes Rec.P.1301 and Rec.P.800 for audio and speech, respectively [77, 78]. EBU Rec.274 presents a number of methods for the subjective quality assessment of audio signals [79]. Finally, for audio-visual signals quality, ITU-T Rec.P.911 and P.913 describe audio-visual quality assessment methods [80, 81]. Table 3.1 presents a list of the more relevant recommendations for subjective experiments.

Table 3.1: Recommendations for subjective experiments.

| Year | Agency | Code | Name | Signal Modality |
|---|---|---|---|---|
| 1994 | ITU | P.85 | A method for subjective performance assessment of the quality of speech voice output devices | Speech |
| 1996 | ITU | P.800 | Methods for subjective determination of transmission quality | Audio |
| 1998 | ITU | P.911 | Subjective audiovisual quality assessment methods for multimedia applications | Audio-visual |
| 1998 | EBU | R274 | Tech Review: Subjective assessment of audio quality | Audio |
| 1999 | EBU | R22 | Technical Recommendation: Listening conditions for the assessment of sound programme material | Audio |
| 2000 | VQEG | FRTV1 | Final Report: Full Reference Television (FRTV) Phase I | Video |
| 2003 | VQEG | FRTV2 | Final Report: Full Reference Television (FRTV) Phase II | Video |
| 2005 | EBU | SAMVIQ | Subjective Assessment Methodology for Video Quality | Video |
| 2007 | ITU | BT.1788 | Methodology for the subjective assessment of video quality in multimedia applications | Video |
| 2008 | ITU | P.910 | Subjective video quality assessment methods for multimedia applications | Video |
| 2008 | VQEG | MM1 | Final Report: Multimedia Phase I | Video |
| 2010 | VQEG | HDTV | Final Report: High Definition Television (HDTV) | Video |
| 2012 | ITU | BT.500 | Methodology for the subjective assessment of the quality of television pictures | Video |
| 2012 | ITU | P.1301 | Subjective quality evaluation of audio and audiovisual multiparty telemeetings | Audio |
| 2016 | ITU | P.913 | Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment | Audio, Video, Audio-visual |
| 2016 | ITU | P.807 | Subjective test methodology for assessing speech intelligibility | Speech |

Minor differences aside, most documents coincide on their basic structure, for instance, the source stimuli selection. Considering that, one of the main objectives of the subjective quality assessment experiments is the analysis of the quality of test videos (processed, compressed, transmitted, etc.), the source stimuli selected must possess specific characteristics that better represent the media capability (e.g., spatial-temporal characteristics for video and phonemes for audio and speech). This type of criteria compels the selection of stimuli content often considered "artificial", as exemplified by Pinson [72]. Additionally, content diversity is usually limited to a small set of sources and the length stimuli is quite short (from 6 to 10 seconds for each stimuli).

Usually, for this type of tests, subjects are asked to rate the source stimuli and the hypothetical reference circuit (HRC) combination. The HRC refers to a particular test condition of the source, e.g., a fixed combination of a video (audio) bitrate level, at a network condition, and a video (audio) encoding algorithm. This type of method maximizes measurement accuracy for each individual stimulus and allows a systematic comparison between all HRCs. Only one task is performed by the participants: to rate the perceived quality of the current stimulus. Most of the recommendations encourage participants to

ignore the stimulus content and focus only in the visual (or audio) quality. Given that it has been shown in several studies that the stimulus content strongly influences human perception of quality [82], new methodologies that take this effect into account need to be developed and used in multimedia quality assessment.

All these limitations induced researchers to propose new methods to assess the subjective quality of signals. One of the main aspects where these variations have to take into account is the duration of the test stimuli. Staelens based his subjective quality assessment methodology on full-length movies [83]. The main goal of the method proposed by Staelens is to present the test stimuli in the same environment and under the same conditions end-users watch it. A first group of participants were asked to watch a full-length DVD movie at home. Blocking effects, caused by packed loss and frame freezing effects were included in test sequences. No audio degradation was included. After watching the movie, subjects were asked to immediately fill up a questionnaire, reporting their opinion on the visual quality of the movie. Another subjective experiment was carried out using a different group of human observers, this time using the traditional single stimulus (SS) ACR method described in ITU-R Rec. BT.500 [84]. Shorter video sequences from the same movies were created with the same visual impairments. These sequences were presented to another group of subjects in a controlled laboratory environment. Results showed a significant difference related to the detection and annoyance of the visual impairments. Environment and experience conditions proved to be important factors in quality assessment.

Another interesting methodology was proposed by Borowiak et al. [85]. Borowiak's methodology takes into consideration requirements to assess the QoE of multi-modal systems: 1) use of continuous sessions of long duration material, 2) suppression of an explicit quality reference, 3) minimization of participant's fatigue, and 4) focus on stimuli content instead of the assessment task itself. This quality assessment methodology allows participants to calibrate the quality level of the stimuli during playback, while degradations are occurring. Participants were presented with long video sequences (30 minutes in average). During the reproduction of the video, automatic changes in the video quality were presented periodically. Once the quality drop was noticed by the participant, he/she was able to turn a knob to request a higher quality level. Rotating the knob too far made the quality drop again, this might be considered as a penalty mechanism. This method is based on a purely perceptual judgment.

Both Staelens and Borowiak experiments argue that traditional methodologies might not accurately represent the quality perceived by end-users. Long duration stimuli helped capture the attention of participants and encourage them to focus on the experience itself. Also, presenting audio-visual content to assess only video (or audio) degradations seemed

to be a more realistic way to capture the real experience.

### 3.1.2 Immersive Methodology

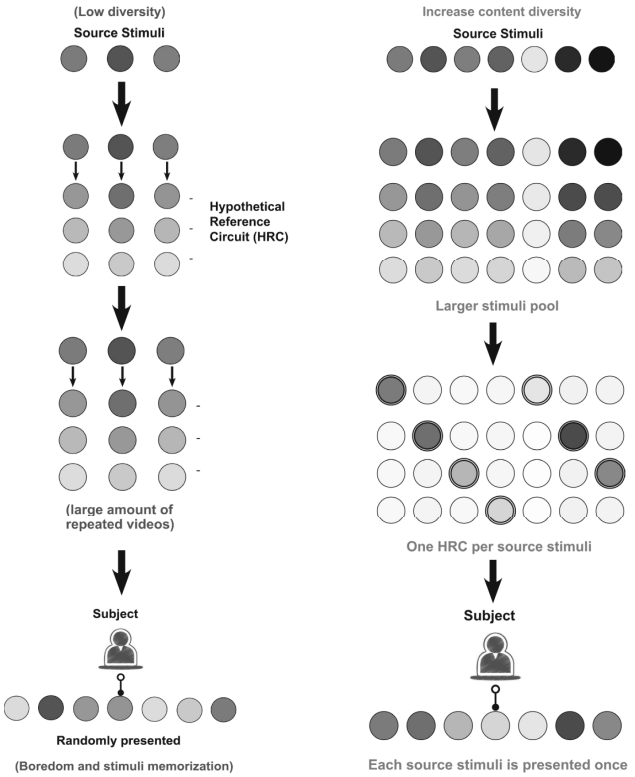
The immersive methodology, proposed by Pinson et al.[72], takes into consideration some of the previously mentioned aspects and proposes a new approach for subjective experimentation. This new methodology has the goal of capturing the perceived quality for different HRCs, putting the subject in a more natural scenario.

In order to reproduce a natural scenario, certain variations for the experiment setup were included, for instance, longer stimuli. Capturing the attention and engaging the participant in the content matter are the main goals of using longer stimuli. Neither full-length movies nor 30-minute clips are considered for the immersive methodology, given that their inclusion might result in extremely long tests sessions. Instead, sequences of 30 to 60 seconds length are considered sufficient to transmit an entire idea and capture the subjects attention, while maintaining an acceptable test session duration.

Another important consideration for the immersive methodology is the usage of audio-visual stimuli to evaluate video-only or audio-only impairments. A video-only stimuli provides a poor representation of the user experience for an audio-visual application (consumers rarely watch videos with no sound). Certain exceptions can be made depending of the objective of the immersive test, for example immersive tests for cell phones (audio-only) and immersive tests for surveillance videos (video-only).

Using audio-visual stimuli has certain consequences. For instance, in an immersive test, subjects must always be asked to rate the overall audio-visual quality. Beerends and Caluwe at [86] showed that participants had trouble separating the audio quality from the video quality when an audio-visual stimuli is presented. The impact of audio quality on video quality can be controlled by evaluating impairments for one component while keeping the quality of the other component constant. Other important consequence of using audio-visual stimuli is the variation in the range of the mean opinion score (MOS) values. Evaluating a component while keeping constant the other component decreases the quality range and could cause saturation of the rating scale.

Immersive methodology seeks to reduce participant's fatigue. As previously mentioned, on traditional subjective experiments a large set of stimuli processed at a number of HRCs is presented to the subjects. Subjects have to assess the quality of stimuli corresponding to the same content, which leads indefectibly to boredom and stimuli memorization. Figure 3.1 (a) depicts a illustration of what a traditional method would be. In the immersive methodology, each source stimulus is presented only once to each subject. This strategy prevents fatigue and assures that results are not influenced by stimulus memorization. As a recommendation, the number of sources used for the experiment




(a) 'Traditional Methods' (b) 'Immersive Method'

Figure 3.1: Traditional Methods vs Immersive Method.

should be an integer multiple of the number of HRCs under study. More preferably, for each HRC, each subject should see five to ten stimuli, which leads to a good estimate of subject's opinion about each HRC [82]. Figure 3.1 (b) presents an illustration of what an immersive method would be.

The basic setup in an immersive experiment is given by a number of source stimuli (w), a set of Hypothetical Reference Circuit - HRC (y), and a number of subjects (n). The combination of every source stimuli and HRC results in a total of $w \cdot y$ stimuli. Each subject rates $w/y$ of these stimuli for each HRC. When all subject scores are pooled, approximately $n/y$ subjects rate each individual stimuli.

An immersive experiment will produce the traditional MOS (per-stimulus measurement) and an $MOS_{HRC}$ (per-HRC measurement). For a per-stimulus measurement experiment, the accuracy of the MOS value will depend on the number of subjects (n) included in the experiment. Therefore, in a immersive experiment, its accuracy is reduced because only a group of participants will rate a particular stimulus. Meanwhile, for a per-HRC measurement experiment, the accuracy of the $MOS_{HRC}$ depends mainly on the number of sources (w) used for the experiment. The impact of increasing the number of subjects have a minor effect when compared to the increase of the number of source stimuli. To give an illustration of this effect, suppose that for a certain experiment a set of five videos

depicting sports are chosen. Increasing the number of participants will not improve the understanding of video generally, such as movies, news, cartoons, music videos, sports or home videos, it just increases the knowledge of those five sport videos.

One last consideration refers to the type of question made after the stimuli is presented. The immersive methodology formulates two target questions and three distractor questions. The first target question refers to the overall perceived quality of the visual and audio components. This question is used to calculate the MOS ($MOS_{HRC}$). The second target question refers to the stimuli content. The participant is asked to give its opinion on the stimuli content. This type of questions helps the researcher investigate the influence of the stimuli content on the perceived quality. Although they are not strictly required in an immersive test, distractor questions have the goal of determining whether or not the stimuli is acceptable for a particular application. Common distractor questions could be related to the topic presented in the sequence (e.g., *What topic was this person discussing?*) or a particular detail about the content matter (e.g., *What attracted your attention the most?*).

In order to use the immersive methodology, a researcher must pay special attention to the number of HRCs that must be included in the experiment. Stable results have been observed on experiments using 30 to 40 participants to rate four HRCs [82]. However, the inclusion of a high number of HRCs might result in long experimental sessions, one possible solution is to increase the number of subjects to obtain a balance.

An immersive methodology to assess speech quality was performed by Pinson in [82]. Twenty audio-visual sequences with different content were included for the test, plus the two sequences that were used in the training session. The test material consisted of audio-visual sequences with a variety of people discussing various topics in response to an interviewer. All sequences contained a dialogue containing a complete idea that could be understood without having prior information or context of the interview. The stimuli depicts a traditional head and shoulder format with a gray background. The selected stimuli had a duration of 34 to 50 seconds. Four HRCs (impairment levels) were selected for the test, which were a combination of narrow-wide band channels conditions and four bitrate compression levels (4.75, 8.85, 12.2, and 24.0 kb/s). A total of 16 subjects took part in the experiment. For this particular experiment, a total of 80 stimuli were produced. Each participant rated 5 stimuli for each HRC. After pooling all subject's scores, each particular stimuli was rated by 4 participants. Results show the capability of the immersive methodology to replicate results from quality experiments conducted with traditional methodologies. Immersive $MOS_{HRC}$ values differed by a gain and offset, which can be explained by the prescense of high quality videos [72].

The immersive methodology is specially tailored for multimedia applications that require longer sequences for a better analysis, a type of application in which traditional methodologies have limitations. For example, Garcia *et al.* showed the importance of using an immersive methodology to measure the quality of long videos in adaptive streaming applications [51]. Moreover, Robitza *et al.* used the immersive methodology to study the impact of quality variations and stalling events[87]. Although this experiment used 66 source sequences of 1-minute, leading to experimental sessions of over an hour, results showed that the participants's alertness was not affected. Finally, Staelens *et al.* obtained good results using the immersive methodology to perform an experiment that included camera angle changes [88].

Although the immersive methodology cannot replace traditional methods and recommendations, it provides a promising alternative for certain applications that are hard to analyze using traditional subjective testing methods. The usage of distractor questions can help infer the minimum level of quality that is acceptable for a particular application. Commercial decisions on video products and services, where the vendor needs to decide between perceived quality and cost, might benefit from using the immersive methodology. Another application for which the immersive methodology can be used are video systems for sign language, where the layered interaction between different linguistic elements makes it difficult to create artificial stimuli.

## 3.2 Objective Quality Assessment

Objective quality assessment are computational algorithms (objective metrics) that have the goal of predicting the perceived quality of a signal stimuli. As mentioned before, the performance of objective metrics is estimated by comparing their results with the results gathered from subjective experiments. At the present time, the vast majority of objective quality metrics estimate the perceived quality of the independent media components, i.e. audio quality and video quality are measured separately. Regarding audio-visual quality metrics, current proposals are limited to a combination of separate audio and video quality estimations. Due to the great influence of machine learning techniques, feature-based metrics have gained great importance in recent years. Several metrics are now being used to provide with strong descriptive features to predict the quality of the transmitted signal [89, 90].

This section presents a brief description of several video and audio quality objective metrics from the literature. Additionally, one objective metric for video and one for audio, which are the basis of this work, are described in detail. Finally, the progress attained on the development of audio-visual quality metrics is analyzed.

### 3.2.1 Video Quality Metrics

Regarding the amount of information required for quality assessment, video quality metrics can be organized into three categories: 1) Full-Reference (FR), 2) Reduced Reference (RR), and 3) No-Reference (NR).

FR metrics have access to both the original and test video signals. This type of metrics have been widely studied and they usually present good performance in predicting the perceived quality. However, they cannot be implemented for a monitoring type of service. Two of the most common FR metrics are the Mean Square Error (MSE) and the related Peak Signal to Noise (PSNR). These two metrics are commonly used because of their simplicity and straightforward mathematical definition, still, they have been criticized for not taking into consideration aspects of the HVS. More advanced versions of these two metrics, which include characteristics of the HVS, have been presented in the literature [91], alongside with more complex models. For example, the Structural Similarity Index (SSIM) predicts the perceived visual quality by comparing the luminance, contrast, and structure information of the original and distorted image [10]. Several variations and adaptations for video, based on SSIM, were later presented in the literature [92, 93, 94]. Another image metric is the Visual Information Fidelity (VIF) [95], the VIF uses three models to calculate the quality of distorted images, such models are the Natural Scenes Statistics (NSS), distortion and HVS models. An extended version that works on videos is denoted by V-VIF. FR metrics that were originally designed to work on video sequences are the standardized ITU-T J.144 [69] Video Quality Metric (VQM) and the Motion based Video Integrity Evaluation (MOVIE) [96]. Table 3.2 presents an extended list of several FR video quality metrics.

RR metrics calculate the video quality by extracting a limited amount of information from the original video. Commonly, some quality features are extracted from the original video and they are compared with the ones extracted from a distorted or modified version. The Reduced Reference Entropic Differencing (RRED) metric [101] measures the information changes between the original and distorted images by finding differences in the entropy of their wavelets coefficients. Extended versions that work with spatial and temporal entropic differences (SRRED and TRRED) were also proposed [102]. The algorithm proposed by Wang and Simoncelli at [97] predicts the quality score based on a natural image statistic model in the wavelet domain. The Kullback-Leibler distance between the marginal probability distributions of wavelet coefficients of both original and distorted images is used to measure the image distortion. Additionally, the RR image quality assessment system proposed by Redi on [98], exploits color information on second order histograms (color correlograms) to estimate the image quality. Table 3.2 presents an extended list of several RR video quality metrics.

Table 3.2: Overview objective quality metrics

| Year | Name | Reference Information | Information Extracted | Signal Modality |
|---|---|---|---|---|
| 2007 | SSIM | Full-Reference | Signal-Based | Video |
| 2003 | MS-SSIM | Full-Reference | Signal-Based | Video |
| 2011 | IW-SSIM | Full-Reference | Signal-Based | Video |
| 2011 | FSIM | Full-Reference | Signal-Based | Video |
| 2006 | VIF | Full-Reference | Signal-Based | Video |
| 2004 | VQM | Full-Reference | Signal-Based | Video |
| 2010 | MOVIE | Full-Reference | Signal-Based | Video |
| 2012 | RRED | Reduced-Reference | Signal-Based | Video |
| 2005 | RR IQA [97] | Reduced-Reference | Signal-Based | Video |
| 2010 | RR IQA [98] | Reduced-Reference | Signal-Based | Video |
| 2005 | NR VQM [99] | No-Reference | Signal-Based | Video |
| 2003 | NR VQM [100] | No-Reference | Signal-Based | Video |
| 2010 | BIQI | No-Reference | Signal-Based | Video |
| 2011 | BLIINDS-II | No-Reference | Signal-Based | Video |
| 2014 | DIVINE | No-Reference | Signal-Based | Video |
| 2012 | BRISQUE | No-Reference | Signal-Based | Video |
| 2016 | VIIDEO | No-Reference | Signal-Based | Video |
| 1998 | PEAQ | Full-Reference | Signal-Based | Audio |
| 2013 | POLQA | Full-Reference | Signal-Based | Speech |
| 2012 | VISQOL | Full-Reference | Signal-Based | Audio |
| 2006 | P.563 | No-Reference | Signal-Based | Speech |
| 2013 | P.1201 | No-Reference | Parametric | Audio-visual |
| 2011 | NR AVQM [15] | No-Reference | Parametric | Audio-visual |
| 2014 | NR AVQM [73, 14, 15, 74] | No-Reference | Audio-Video Combination | Audio-visual |

NR metrics, on the other hand, have a more difficult task since no information about the original signal is available. NR metrics in general consist of measures of the several features and characteristics that are common in distorted signals. Commonly, the features used to calculate visual quality are artifact signals, such as blockiness, blurriness, and ringing. For instance, the algorithms proposed by Wang [103] and Wu [104] estimate image quality using only a blockiness measurement. Some other authors included other feature measurements, such as noise and contrast, to calculate the overall visual quality [99] [100]. NR metrics that use distortion artifacts and coding parameters settings are considered hybrid metrics. Another group of NR metrics analyses the Natural Scene Statistics (NSS). For example, the Blind Image Quality Index (BIQI), is based on NSS and requires a training stage before it can be used. This means that no knowledge of the distortion is needed. Similarly, the BLind Image Integrity Notator using Discrete Cosine Transform (DCT) Statistics (BLIINDS) [105] uses NSS of DCT coefficients to predict visual quality. Likewise, the Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [106], uses NSS to assess image quality in the spatial domain. Finally, the Video Intrinsic Integrity and Distortion Evaluation Oracle (VIIDEO) was presented as a completely blind video quality method. The VIIDEO metric exploits the statistic naturalness of the video frames in order to detect some irregularities and hence predict the quality of the video

sequence [107].

Table 3.2 presents an extended list of several NR video quality metrics currently available in the literature. At present, a considerable amount of FR, RR, and NR video quality metrics was developed. However, most of the current video quality metrics are FR metrics.

Commonly, machine learning methods for video quality assessment rely on feature sets derived from several objective video quality metrics in the literature. For instance, the Distortion Identification-based Image Verity and Integrity Evaluation (DIIVINE) index [108] uses NSS to identify the distortion type and quantify its presence on the affected image. This particular metric was used as the base to the development of the current audiovisual quality metric. As some of the formerly mentioned metrics, it is based on the extraction of natural scene statistic features (NSS). Next, a detailed description of the metric is presented, putting special care on the feature extraction process.

The divine metric, originally developed as an image quality metric, bases its approach on the assumption that images possess certain statistical properties that are perturbed in the presence of certain distortions. Hence, this metric attempts to predict the quality of images by measuring the level of naturalness of these statistical properties. Given that this analysis requires only the distorted image properties (test phase), this type of approach represents an interesting No-Reference solution for image and video quality.

The divine metric is a 2-stage method that involves a feature extraction phase and a distortion-specific quality assessment. After the feature extraction phase, a vector that describes the image is passed on to perform two tasks. First, identify the probability that the image is affected by one of the five types of distortion that the metric considers. Second, map the feature vector into a quality score for each type of distortion, then use the probabilistic distortion estimate to build the final quality score of the image. Figure 3.2 presents a simplified diagram of the DIIVINE metric. Next, the main stages that compose the diivine metric are described in detail.

- *Feature Extraction*

  Before extracting the descriptive features, the image under observation is subject to a wavelet decomposition using the steerable pyramid method [109]. As a result, the image is decomposed into 12 sub-bands, denoted as $s_{\alpha}^{\theta}$, across two scale and six orientation values, where $\alpha \in \{1, 2\}$, and $\theta \in \{0°, 30°, 60°, 90°, 120°, 150°\}$. Next, a divisive normalization is applied to the set of sub-bands [110]. The main objective is to reduce statistical dependencies between neighboring sub-bands. After this pre-processing phase, marginal and joint statistics are computed across all sub-bands to extract the descriptive features of the observed image. The following five procedures are applied to extract the descriptive features:

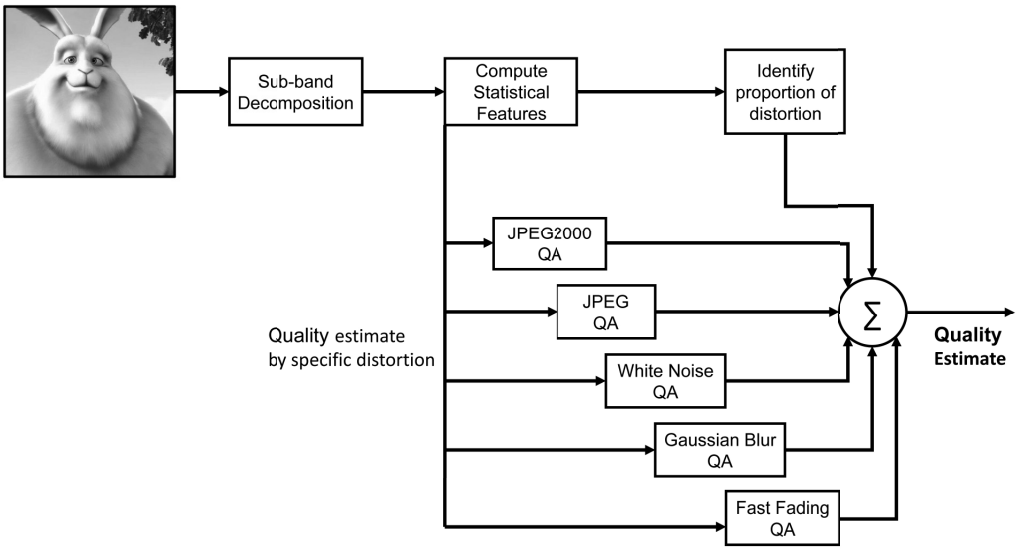



Figure 3.2: Diivine metric block diagram. Adapted from [108].

1. *Scale and orientation selective statistics (features 1 to 24):* Sub-bands coefficients are parametrized using a generalized Gaussian distribution (GGD) [111]. Coefficients for each of the 12 sub-bands are computed using the variance and the shape-parameter values. The variance and shape-parameter produce one feature for each of the sub-bands, leading to a set of 24 features.

2. *Orientation selective statistics (features 25 to 31):* Again, a GGD fitting is used to calculate the second set of features. The GGD is fitted to the coefficients obtained by stacking together coefficients from the sub-bands with the same orientation ($\theta$). Additionally, another feature is computed when all sub-bands are stacked together. Then, features from 25 to 30 correspond to features across scales over different orientations and feature 31 correspond to statistics across sub-bands.

3. *Correlation across scales (features 32 to 43):* Given that edges are of great importance in image quality assessment tasks, it is assumed that statistical properties between high-pass (HP) responses and their band-pass (BP) counterparts hold valuable information. The structural correlation between BP and HP is computed for each of the 12 sub-bands. As a result, 12 new features are computed.

4. *Spatial Correlation (features 44 to 73):* Based on the observation that natural images are highly structured and that the presence of distortions affects this structure, spatial correlation statistics of the sub-bands spatial structure are computed. This results in 30 new descriptive features that represent the spatial

correlation across sub-bands.

5. *Across orientation statistics (features 74 to 88):* This last procedure computes features based on the statistical correlations of images across orientations. Similar to the correlation across scales, structural correlations are calculated for all possible pairs of adjacent orientations at the same scale. This combination leads to a total of 15 features.

- *Distortion identification and distortion-specific quality assessment*

  At this stage, the DIIVINE metric is able to perform two tasks. First, use the descriptive features to estimate the probability that an observed image has one of the distortions considered by the metric. Second, for each type of distortion considered, a regression model maps the descriptive features of the image onto a quality score. Finally, each distortion-specific quality score is weighted by the probability that a specific distortion is present in the observed image. In the end, an overall quality score of the observed image is computed.

### 3.2.2 Audio Quality Metrics

Audio quality metrics can be separated into two categories, intrusive and non-intrusive; that is, Full-Reference and No-Reference respectively, if compared to video quality metrics. Intrusive audio quality metrics compare the original signal with a degraded version that has been processed. Early methods like the Signal to Noise Ratio (SNR) are unable to emulate human's judgment on the perceived audio quality. The Perceptual Evaluation of Audio Quality (PEAQ) [71], which was standardized as ITU-R BS.1387, uses a psychoacoustic model and a cognitive model to estimate the perceived quality. The psychoacoustic model provides the cognitive model with a number of model output variables that are mapped to an objective difference grade (ODG) quality score via a multi-layer neural network. Two versions of this technique were presented: a basic version, which is optimized for speed, and an advanced version, which has an improved accuracy. An advanced version of the Perceptual Evaluation of Speech Quality (PESQ) [112] resulted on the Perceptual Objective Listening Quality Assessment (POLQA) [113]. The same logic used in PESQ was used in POLQA, that is, an alignment of the original and the distorted signals is made, and then both metrics are compared using a perceptual model. POLQA was designed for speech quality assessment and it can be used on a narrowband mode (300 – 3400 Hz) or superwideband mode (50 – 14000 Hz). Currently, the authors are working to develop an adapted audio quality version of POLQA.

Though promising results have been presented, there is still a lot of work on the development of accurate non-intrusive audio quality metrics. Table 3.2 presents an extended

list of several intrusive and non-intrusive audio quality metrics currently available. To the present, none non-intrusive audio quality metrics has been standardized by the ITU. The development of such type of metrics is still an active area of research. The ITU-T standard P.563 for single-ended speech assessment [114] represents the most important achievement for this area. The first step of the P.563 algorithm consists on processing the test signal using a voice activity detector (VAD). This first step serves to identify speech signals and estimate their speech levels. Then, the signal is analyzed and a set of 51 characteristic signal parameters is obtained. Next, it classifies the signals using a set of distortion classes that are based on a restricted set of key parameters. The main distortion classes include 'unnatural speech', 'noise', and 'interruptions, mutes, clippings'. The key parameters and the assigned main distortion class are used to estimate the speech quality.

The capacity of spectrograms to represent important audio and speech characteristics makes them an important tool for audio and speech quality assessment. For instance, an intrusive approach named Virtual Speech Quality Objective Listener (VISQOL) [39] has been adapted for audio quality testing resulting in VISQOLAudio. Both VISQOL and its audio version measures the signal quality by comparing the similarity of the spectrograms obtained from the original and degraded signals. The algorithm uses the Neurogram Similarity Index Measure (NSIM), a metric inspired on the visual SSIM metric. This audio metric was used as the base for the development of the current audiovisual quality metric. Next, a detailed description of the VISQOL metric is presented.

The VISQOL metric, originally developed to assess speech quality, compares 2-D representations (spectrograms) of the speech signal in order to predict the speech quality. Overall, the metric compares the spectrograms of the distorted signal and a clean reference version of the same signal. The level of similarity of both spectrograms is measured by using an NSIM index, later on, such similarity response is mapped to an objective quality scale, refered as $Q_{MOS}$. Figure 3.3 depicts a block diagram presenting the five major stages of the VISQOL metric, which are described next.

- *Pre-processing*

  First, the degraded signal y(t) is scaled to match the power level of the reference signal x(t). Next, spectrogram representations of both degraded and reference signals are extracted by using a Short-term Fourier transform (STFT). The degraded and reference spectrograms, denoted as 'd' and 'r' respectively, are passed on as input to the second stage of the metric.

- *Patch alignment*

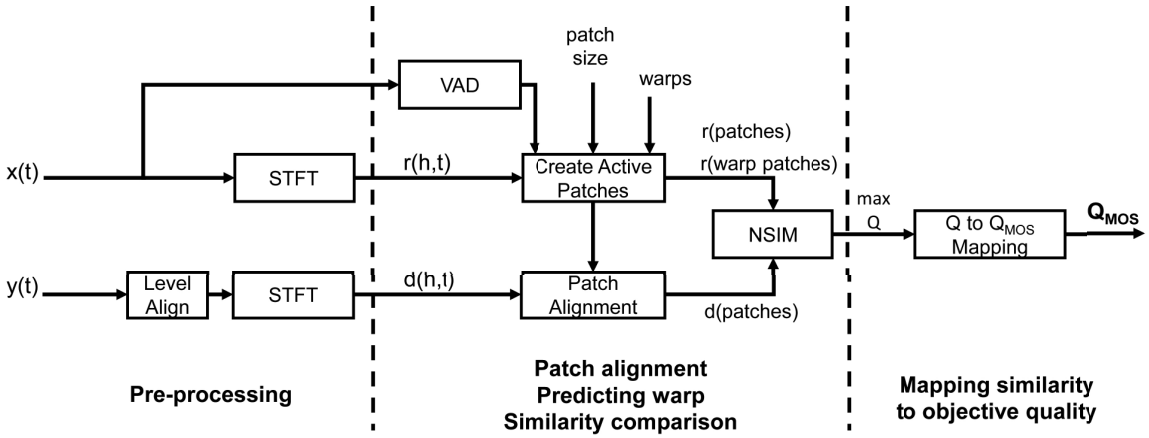



Figure 3.3: Visqol metric block diagram. Adapted from [115].

At this phase, the reference spectrogram is segmented into patches of 30 frames (480 ms), as illustrated in Figure 3.4. Additionally, a simple voice activity detector is used to identify active patches. Then, the NSIM is used to time align the patches from the reference signal with the corresponding areas of the degraded spectrogram. The NSIM is calculated for each reference patch and the test spectrogram patch (frame by frame), thus identifying the higher similarity value for NSIM. Figure 3.4 illustrates how a patch from the reference signal is tested along the distorted signal, additionally, NSIM values for each frame are plotted.

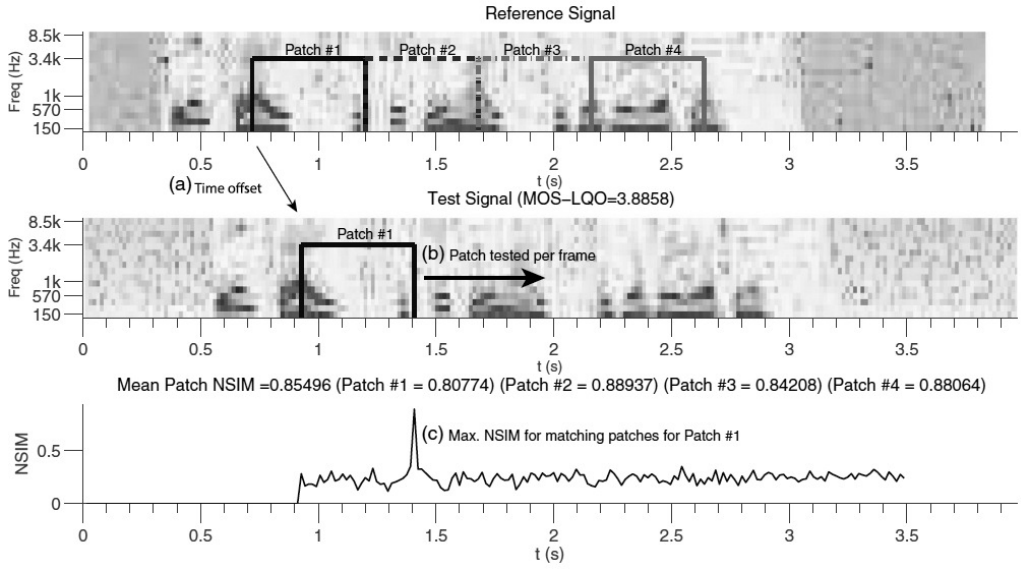



Figure 3.4: Visqol patch alingnment and NSIM similarity. Original illustration from [115].

- *Predicting warp*

As an additional process, alternative patches are created 1% and 5% longer and shorter than the reference patches. These patches are denoted as warped patches and they are included in the NSIM comparison. If the similarity score of the warped patch is higher than a regular patch, then the warped patch similarity is kept.

- *Similarity comparison*

  All higher similarity values corresponding to all patches are then averaged to form the signal similarity estimate. The similarity value between two spectrograms, $r$ and $d$, is defined with a weighted function of intensity ($l$), contrast ($c$), and structure ($s$), given by:

$$\mathrm{NSIM(r,d)} = l(r,d)^{\alpha} \cdot c(r,d)^{\beta} \cdot s(r,d)^{\gamma} \tag{3.1}$$

  where $\alpha, \beta$, and $\gamma$ are set to 1 for a basic version of the metric.

- *Mapping similarity to objective quality*

  A sigmoid mapping function is used to translate the similarity estimate of the signal into an objective quality score $Q_{MOS}$. As in traditional MOS scores, $Q_{MOS}$ ranges from 1 to 5.

### 3.2.3 Audio-visual Quality Metrics

At the present, there is no reliable metric available for measuring the overall audio-visual quality of signals. A parametric model that uses information extracted from packet headers and network information has been standardized as ITU-T Rec. P.1201 [116]. A similar approach [15], uses impairment factors, which quantify the quality-impact of different types of degradations. Such impairment factors are computed using information extracted from the bitstream or packet headers. Since parametric models are codec and transmission dependent, these methods are less generally applicable.

A large group of audio and video quality metrics has been revised in the preceding sections. All these metrics assume only a single modality, either audio or video. Given the progress on the assessment of audio and video quality separately, several studies have proposed models for audio-visual quality that consist of simple linear combinations of audio and video quality scores. Several studies rely on a commonly used combination model given by the following equation [14, 15, 73, 74]:

$$\mathrm{Quality}_{av} = \alpha_1 + \alpha_2 \cdot \mathrm{Quality}_{a} + \alpha_3 \cdot \mathrm{Quality}_{v} + \alpha_4 \cdot \mathrm{Quality}_{a} \cdot \mathrm{Quality}_{v}. \tag{3.2}$$

Moreover, different types of combination strategies have been proposed in order to obtain more accurate results. Becerra et al. [74] proposed two different strategies are used to

combine the audio and video quality values. The first strategy uses a weighted minkowski function given by the following equation:

$$\text{Quality}_{av} = (\alpha_1 \cdot \text{Quality}_a^p + \alpha_2 \cdot \text{Quality}_v^p)^{\frac{1}{p}}. \tag{3.3}$$

while the second strategy uses a power-based model given by the following equation:

$$\text{Quality}_{av} = (\alpha_1 + \alpha_2 \cdot \text{Quality}_a^{p_1} \cdot \text{Quality}_v^{p_2}). \tag{3.4}$$

Despite the strategies used for combination, it is important to understand how the auditory and visual stimuli are perceived and at what stage in the human perceptual process they are fused. Table 3.2 presents an extended list of the several combination-based audio-visual quality metrics.

More recently, a large group of artifact indicators was developed by the Monitoring Of Audio-Visual quality by key Indicators (MOAVI) [21]. MOAVI is a subgroup of the VQEG formed to develop NR models for monitoring audio-visual service quality. These artifact indicators are classified in four groups (based on their origin), such groups are capturing, processing, transmission, and displaying. They can be calculated by analyzing the media signal, or by using parametric (bit-stream) measurements. The list of considered artifacts includes blocking, blurring, ringing, freezing and block missing, for video signals; and clipping, noise and mute for audio signals.

## 3.3 Databases for Multimedia Quality Assessment

The availability of databases with diverse media content is a key factor in the media quality assessment field. These databases are fundamental to the development of computational quality assessment methods, more specifically on tasks like training, testing, and benchmarking. It is desired that media databases possess: 1) relevant types of degradations commonly found on a real transmission scenario, 2) signal characteristics from common multimedia applications, and 3) subjective quality ratings from human observers gathered in psychophysical experiments. However, most of these material remains private, and the few publicly available databases are not adequate to the research demands. More particularly, most databases are restricted to audio-only and video-only content, disregarding the need for audio-visual material [24]. Table 3.3 depicts a summarized list of some of the publicly available databases in the literature. Next, some of those databases are briefly described in the following lines.

- *Audio Databases*

Table 3.3: Summarized list of some of the publicly available databases in the literature.

| | | Subjective ratings | | | |
|---|---|---|---|---|---|
| **Component** | **Database** | **Audio** | **Video** | **Audiovisual** | **Year** |
| **Audio** | ITU93 [117] | Yes | No | No | 1993 |
| | MPEG95 [118] | Yes | No | No | 1995 |
| | Live Music [119] | Yes | No | No | 2013 |
| | Blizzard Challenge [120] | Yes | No | No | 2016 |
| | TCD-VoIP [54] | Yes | No | No | 2015 |
| **Video** | Live VQ [121] | No | Yes | No | 2010 |
| | VQEG HDTV [122] | No | Yes | No | 2010 |
| | CVD2014 [123] | No | Yes | No | 2014 |
| **Audiovisual** | VQEG-MM2 [124] | No | No | Yes | 2012 |
| | UnB-Audiovisual Database [125] | Yes | Yes | Yes | 2013 |
| | INRS [126] | No | No | Yes | 2016 |
| | Live-NFLX-II [127] | No | No | Yes | 2018 |

1. *ITU93 [117]:* This database is composed of seven audio sequences (Asa Jinder, bagpipe, bass clarinet, castanets, harpsichord, German male speech, and violin). The original audio sequences are processed at different coding algorithms and bitrate values to generate the test sequences of the database. A total of 42 sequences, which were rated by 33 human listeners, are available in the database.

2. *Live Music Dataset [119]:* This database is composed of two sets of live music recordings containing four types of music gender: rock, pop, electronic, and country. The first set corresponds to 500 original music recordings, while the second set corresponds to 2400 synthetically degraded music recordings. Sixty (60) subjects with normal hearing provided subjective responses using a web-based interface.

3. *TCD-VoIP [54]:* The TCD-VoIP dataset includes some common degradations encountered in a voice over IP transmission. Degradations are considered as "platform-independent" as they are not influenced by the codec, hardware, or network in use. The dataset contains five types of degradations: 1) background noise, 2) competing speakers, 3) echo effects, 4) amplitude clipping, and 5) choppy speech. For each type of degradation, a number of test conditions are set. These test conditions are applied to a set of speech samples, resulting in the TCD-VoIP dataset. A total of 384 audio sequences were rated by 24 human listeners.

- *Video Databases*

1. *Live VQ [121]:* The Live Video Quality database includes a set of 15 video-only source sequences. These sequences are then processed to a number of conditions, including different codes (MPEG-2 and H.264) and simulated trans-

mission over IP and wireless network conditions. Video sequences were rated by 38 human observers.

2. *VQEG HDTV [122]:* This database is composed of five publicly available subsets. Test conditions include several bitrate compression values, compressed using two codecs: MPEG-2 and H.264. Additionally, two network impairments are included: slicing error and freeze error.

- *Audiovisual Databases*

1. *UnB AudioVisual Database [125]:* Six source high definition videos, with accompanying audio, were used to build this database. The videos were 8 seconds long, had a resolution of 1280x720, a color space of 4:2:0, and a frame rate of 30 frames per second (fps). The database was sub-divided into three subsets. For the first subset, sequences had only the video component with no audio and they were compressed using an H.264 codec at different (video) bitrate values. For the second subset, sequences had only the audio component with no video and they were compressed using an MPEG-1 layer 3 codec, at different (audio) bitrate values. Finally, for the third subset, both audio and video components were compressed using the bitrate values from the previous setups, both components were processed individually. All three sub-sets were rated by a group of 45 human observers and their responses were collected.

2. *VQEG-MM2 [124]:* The database consists of data gathered from six different international laboratories associated with VQEG, resulting in ten sets of audiovisual subjective values. The database sequences contained audio and video components and they were degraded using different levels of audio and video rate compression. Audiovisual sequences were rated by almost 189 participants (from all six laboratories).

3. *Live-NFLX-II [127]:* This database was built using a set of 15 source high definition videos with accompanying audio. The selected audiovisual content covers a number of genres such as documentary, sports, music, and video games. Different network conditions were simulated in order to recreate common transmission errors. Additionally, client adaptation strategies were included such as bitrate adaptation, buffering adaptation, and quality adaptation. A total of 420 video sequences were rated by 65 human observers and their responses were gathered.

The development of more public available audio-visual databases is crucial for the area of video quality and quality of experience. This work tries to bridge this gap by

presenting three large new audio-visual databases containing several types of audio and video degradations. The three databases are considered as up-to-date material and it is expected to contribute to the development of new audio-visual quality methods. In the next chapter, we describe the experiments performed with the goal of creating this dataset.

# Chapter 4

# Immersive Audio-visual Quality Experiments

Subjective responses from human participants are key to the development of media quality metrics. The collected data is fundamental during the training and testing of the proposed method. Moreover, the test material associated with these responses must reflect the scope of the metric being developed. That is, the test material must cover some particular characteristics such as the type of component (e.g., audio, video, audiovisual, etc.), the context under test (types of degradation), the content under test (e.g., video conferencing, movies, sports transmissions, documentaries, etc.), etc. Given the limited number of databases and subjective responses available in the literature, and considering the need for a tailored test material for the development of an audiovisual quality metric, we conducted three subjective experiments in this work. It is expected that these experiments will contribute to the development of the audiovisual quality assessment field.

For all three experiments, groups of human observers rated the audio-visual quality of a set of video sequences. All three experiments applied the immersive method described before. For the first experiment, visual artifacts degraded the video component, meanwhile, the audio component didn't suffer any type of degradation. In the second experiment, the audio component was subject to signal artifacts while the video component remained untouched. Finally, in the last subjective experiment, both audio and video components were subject to the same types of degradation used for the previous two experiments. For all three experiments, subjects were asked to rate the overall audio-visual quality.

The remainder of this chapter is divided as follows. In Section 4.1 a brief summary of the related work is presented. In Section 4.2, the source material used in all three experiments is described. In Section 4.3, the visual and audio degradations considered for this study are presented. In Section 4.4, the experimental apparatus and the physical conditions are described in detail. In Section 4.5, the experimental methodology is pre-

sented. In Section 4.6, some statistical analysis methods are described. Sections 4.8, 4.9, and 4.8 present the experimental results for experiments 1, 2 and 3 respectively. Finally, Section 4.11 presents a general discussion on the results from all three experiments.

## 4.1 Related Work

Several subjective experiments have been conducted with the purpose of better comprehending the impact of different impairments on perceived quality of different media components such as video, audio, and audio-visual. Regarding the video component, a number of studies have explored the effect of packet-loss and frame-freezing errors on perceived quality. Staelens *et al.* [128] presented a methodology to evaluate the effects of frame-freezing and packet-loss errors using full-length movies. They performed a subjective experiment with 56 non-expert viewers, who rated a total of 80 DVDs on typical home viewing conditions. Results from the study showed that frame-freezing errors were less noticeable when compared to packet-loss errors, suggesting that participants were more tolerant towards visual impairments (packet-loss errors) when placed on a more natural viewing context. Moorthy *et al.* [129] conducted a broader study on different mobile platforms addressing several types of impairments including: video compression, wireless-channel packet-loss, frame-freezing, rate adaptation, and temporal dynamics. Responses from a group of 30 participants were gathered using a video-only dataset. The study concluded that participants preferred few longer stalling events than many shorter stalling events. Nevertheless, the results also suggest that the consumer preference depends on the type of content being displayed (e.g., sports and video conference).

As for the impact of audio distortions in the perceived quality, several studies have been conducted with the objective of comparing different noise scenarios and their corresponding impact [130, 131]. For instance, Wendt et al [132] explored the speech intelligibility comparing two different scenarios. The level of comprehension was measured using speech sentences syntactically complex under different levels of background noise. It was observed that participants were more affected by the level of noise than the complexity of the sentences. It is understood that a background noise type of distortion remains as a determinant factor in the perceived audio quality. Moreover, the TCD-VoIP database was used on a subjective experiment with the objective of studying several VoIP degradations [54]. Speech samples were subjected to five (5) types of distortion: background noise, clipping, competing speaker effects, echo effects, and chop speech. The study treated degradations in isolation. The main focus was on how the distortions impact varied at different levels. Results showed that echo and background noise distortions had a heavier

impact on the perceived quality. Meanwhile, clip, chop, and competing speaker effect had a midterm impact.

With regards to the audio-visual quality, several subjective experiments have been conducted in order to contribute to the theoretical and practical understanding of the perceived audio-visual quality [9, 14, 15]. Although early experiments have suggested a dominant influence of the video component in the overall audio-visual quality, it has been argued that this influence is not the same on all types of applications, e.g., video conference services [14]. What's more, additional studies have confirmed that the interaction between the audio and video components is heavily influenced by some other factors (human, technological, and contextual) that are detailed further in this paper. Researchers have tried to tackle these influential factors by proposing new methods to assess the subjective audio-visual quality. For example, Staelens and Borowiak have tried to capture the attention of participants and encourage them to focus on the experiment itself by using long duration audio-visual stimuli [85, 88]. There is a limited number of experiments, aiming to study the overall audiovisual quality, where both the audio and video components are processed and degraded. Usually, only the video component is subjected to degradations leaving the audio component unimpaired, then, audiovisual subjective scores are collected under these conditions [127]. Some few studies have explored the overall audiovisual quality in a context where both audio and video component suffered individual distortions. Pinson and Becerra [125, 133] conducted subjective experiments employing audiovisual sequences on which both audio and video components suffered distortions due to heavy audio and video compression. Results showed the dominance of the video component in the overall perceived quality, however, it was also observed the impact of the audio component on several types of media content.

Based on these previous results, it is possible to conclude that further studies are needed to analyze the relationship between different types of impairments and their effect on the perceived quality. There are several studies attempting to explore the audiovisual quality on common network scenarios. However, most of them ignore the audio component and the effect of the distortions on the overall quality. This work targets these particular issues and explores a number of distortions (audio and video) often neglected in the current literature. More specifically, we use the immersive methodology to analyze the quality of a set of audio-visual sequences, with a considerable variety of content, degraded by some of the aforementioned types of audio and video impairments. Next, details about the designing of three audio-visual quality experiments are presented.

## 4.2 Source Stimuli

One-hundred and forty (140) high-definition video sequences (with accompanying audio) were used as the source to build the three datasets of this study. They were distributed among all experiments in the following manner: sixty (60) video sequences for experiment 1, forty (40) for experiment 2, and forty (40) for experiment 3. Some of these 140 sequences were generated from parsing larger videos. These videos were gathered from four (4) different websites (listed on Table 4.1). Table 4.1 presents a list of all twenty seven (27) types of video content and the sequences produced using these type of sequences. A pre-processing phase was necessary to standardize some of the video characteristics, such as spatial and temporal resolution, and color space configuration. For this study, we considered a spatial resolution of 1280x720 (720p), a temporal resolution of 30 frames per second (fps), and a color space format of 4:2:0. As for the audio component, the bit-depth and sample frequency were set to 16 bits and 48 kHz, respectively. It is worth mentioning that none of the gathered videos had characteristics below the ones mentioned (Table 4.1 describes all original video characteristics). The stimuli were 19 to 68 seconds long, with an average length of 36 seconds. Representative frames of all 140 videos are depicted in Annex A.

The selection of the source stimuli was made following some of the recommendations found on the Final Report on the validation of objective models multimedia quality assessment (phase 1) of the Video Quality Experts Group (VQEG) [134]. The document highlights the importance of a good distribution of the spatial and temporal activity of the video stimuli. Figure 4.1 presented the spatial and temporal measures computed for all one-hundred videos from experiments 1, 2 and 3 respectively, as defined by Ostaszewska and Kloda [135].

As for the audio component, special attention was paid to the diversity of the content. Stimuli containing a variety of music, speech, smooth and rough sounds were considered during the selection stage. An audio classification was made using the algorithm proposed by Giannakopoulos [136]. The algorithm divides the audio streams into several non-overlapping segments and classifies each segment into one of the following classes: music, speech, others1 (low environmental sounds: wind, rain, etc.), others2 (sounds with abrupt changes, like a door closing), others3 (louder sounds, mainly machines, and cars), gunshots, fights, and screams. Figure 4.2 presents the audio classification for all three experiments to form a better idea of how the different types of audio are distributed. As it can be observed, there is a good distribution of audio content among all video sequences for all three experiments.

Annex C lists all source stimuli and a brief description of the audio and video content, along with the length of each sequence. The length difference observed among the

Table 4.1: Original source stimuli gathered from the 4 different websites. Some of these videos were parsed to produced the one-hundred and forty (140) high-definition sequences used on all three experiments.

| Content Id | Sequence | Experiment 1 | Experiment 2 | Experiment 3 | Spatial Resolution | Scanning | Temporal Resolution (fps) | Copyright | Available |
|---|---|---|---|---|---|---|---|---|---|
| c1 | Guy Sleeping | v01 | | | 1920x1080 | Progressive | 30 | Public | http://www.libde265.org/ |
| c2 | Flamenco | v02, v60 | v18, v19 | v24 | 1920x1080 | Interlaced | 59.94 | Public | http://www.cdvl.org/ |
| c3 | Big Buck Bunny | v03, v04, v57 | | | 1920x1080 | Progressive | 50 | Public | http://www.libde265.org/ |
| c4 | Elephant | v05, v06, v59 | | v01, v23 | 1920x1080 | Progressive | 50 | Public | http://www.libde265.org/ |
| c5 | France Tourism | v07, v19, v53 | | | 3840x2160 | Progressive | 60 | Public | http://www.4ever-project.com/ |
| c6 | WomanDay | v08, v18, v23, v45 | | v08 | 4K | Progressive | 50 | Public | http://4ksamples.com/ |
| c7 | Taiwan | v09, v13, v24, v32 | | | 3840x2160 | Progressive | 30 | Public | http://4ksamples.com/ |
| c8 | Barca vs Atletic | v10, v17, v27, v40, v50 | v01, v05, v07, v11 | v02, v07, v10, v18 | 3840x2160 | Progressive | 30 | Public | http://4ksamples.com/ |
| c9 | FootMusic | v11, v55 | v02, v16 | v03, v21 | 1920x1080 | Interlaced | 30 | Public | http://www.cdvl.org/ |
| c10 | Atlanta Betline | v12, v30, v36, v37 | v03, v10 | v04, v14 | 1920x1080 | Progressive | 59.94 | Public | http://www.cdvl.org/ |
| c11 | Netflix El Fuente | v14, v26, v29, v38, v46, v49, v51, v54 | v13, v15 | v05, v15, v17, v19 | 4K | Progressive | 59.94 | Public | http://www.cdvl.org/ |
| c12 | Box interview NTIA | v15, v21, v58 | v04, v06, v17 | v06, v22 | 1920x1080 | Interlaced | 59.94 | Public | http://www.cdvl.org/ |
| c13 | Honey Bees | v16, v35 | | | 4K | Progressive | 30 | Public | http://4ksamples.com/ |
| c14 | Kenpo Strikes NTIA | v20, v43 | | | 1920x1080 | Progressive | 30 | Public | http://www.cdvl.org/ |
| c15 | Taipei Fireworks | v22, v44 | | v09 | 3840x2160 | Progressive | 30 | Public | http://4ksamples.com/ |
| c16 | Old Town Car NTIA | v25 | | | 1920x1080 | Progressive | 30 | Public | http://www.cdvl.org/ |
| c17 | NTIA Violin | v28, v47 | v12 | v16 | 4K | Progressive | 59.94 | Public | http://www.cdvl.org/ |
| c18 | Puppies | v31, v48 | | v11 | 4K | Progressive | 30 | Public | http://4ksamples.com/ |
| c19 | Big Green Rabbit | v33 | v08 | v12 | 1920x1080 | Progressive | 30 | Public | http://www.cdvl.org/ |
| c20 | Movie Trailer Sintel | v34 | v09 | v13 | 4096x1720 | Progressive | 30 | Public | http://www.libde265.org/ |
| c21 | Landscape Fast | v39 | | | 3840x2160 | Progressive | 30 | Public | http://4ksamples.com/ |
| c22 | FoxBird | v41 | | | 1920x1080 | Progressive | 30 | Public | http://www.cdvl.org/ |
| c23 | Fishing Florida | v42, v56 | | | 1920x1080 | Progressive | 30 | Public | http://www.cdvl.org/ |
| c24 | Food | v52 | v14 | v20 | 3840x2160 | Progressive | 30 | Public | http://4ksamples.com/ |
| c25 | Interview | | v20, v21, v22, v23, v24, v25, v26, v27, v28,v29 | v25, v26, v27, v28, v29, v30, v31 | 1920x1080 | Progressive | 30 | Public | http://www.cdvl.org/ |
| c26 | Reporter | | v30, v31, v32, v33, v34, v35, v36, v37, v38 | v32, v33, v34, v35, v36, v37, v38 | 1920x1080 | Progressive | 30 | Public | http://www.cdvl.org/ |
| c27 | Travel Clip | | v39, v40 | v39, v40 | 1920x1080 | Progressive | 30 | Public | http://4ksamples.com/ |

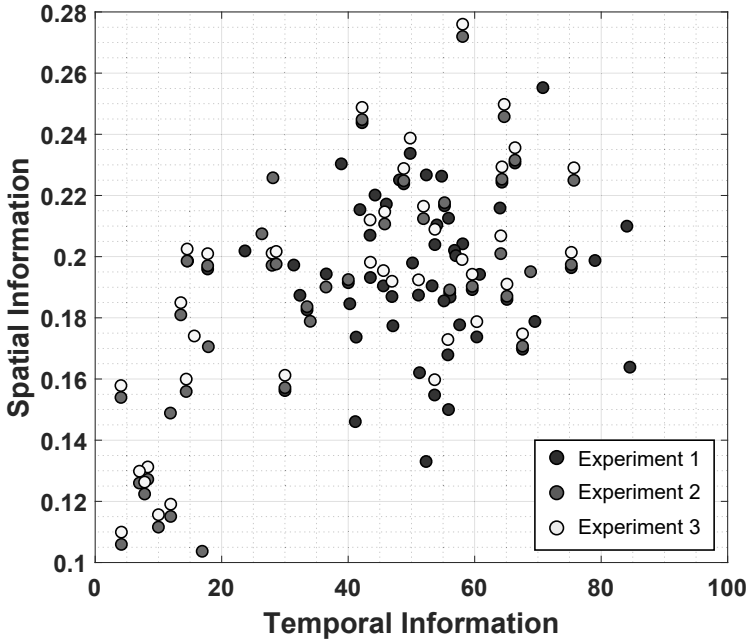



Figure 4.1: Source videos spatial and temporal information measures

sequences backs up the intention of presenting videos capable of transmitting an entire idea.

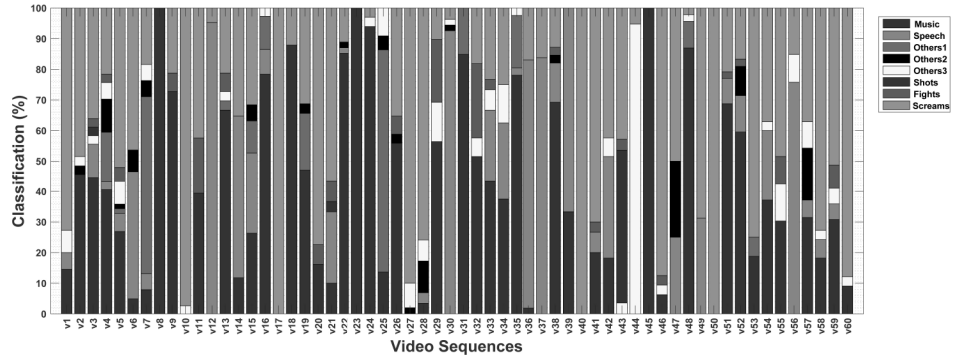
## 4.3 Media Degradations

The source sequences were subjected to some video and audio types of distortions. Such distortions were selected by the researchers based on previous studies from the literature and a particular interest in studying some specific types of distortions. The video component of the sequence was subjected to three types of distortions: video coding, packet loss, and frame freezing. As for the audio component, source sequences were subjected to four types of distortions: background noise, clipping, echo, and chop. This section describes all these types of degradations and reports the sequence processing used to generate the stimuli pools for all three databases.

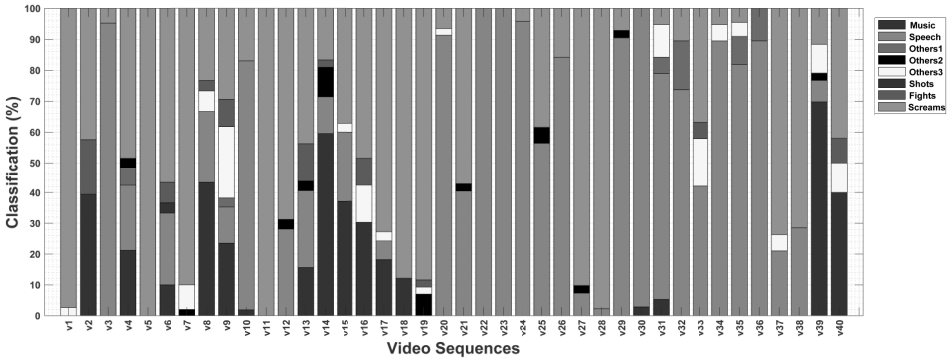
### 4.3.1 Video Degradations

#### Coding Artifacts (compression)

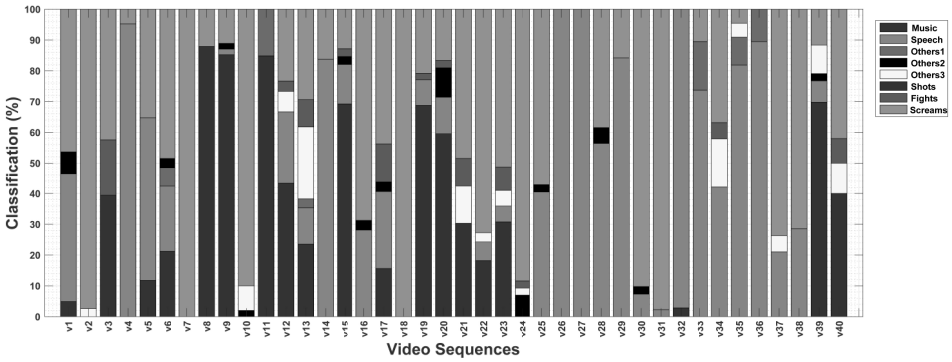
Coding artifacts are the result of the application of lossy data compression. Among the most common coding artifacts we can cite blocking, blurring and ringing artifacts [32]. In this work, we selected two coding standards to compress each of the source stimuli: the H.264/MPEG-4 Advance Video Coding (AVC) and the H.265 High Efficiency Video




(a) 'Experiment 1'




(b) 'Experiment 2'




(c) 'Experiment 3'

Figure 4.2: Audio classification of video sequences. Eight (8) audio classes are considered: music, speech, others1 (low enviromental sounds: wind, rain, etc.), others2 (sounds with abrupt changes, like a door closing), others3 (louder sounds, mainly machines, and cars), gunshots, fights, and screams. **(a)** Experiment 1 (60 sequences). **(b)** Experiment 2 (40 sequences). **(c)** Experiment 3 (40 sequences)

Coding (HEVC) [137, 138]. Four bitrate levels were chosen for each coding standard which were labeled as Low, Medium, High, and Very High. An empirical criteria was used to select these bitrate values, which consisted of visually examining video sequences compressed at a number of bitrate levels and chosing four very clear quality levels, taking into account previous works found in the literature [51, 139]. Table 4.2 presents all four bitrate values used for each codec. We used the reference implementations of AVC and

HEVC presented in [140, 141]. The encoder parameters used for both coding standards are listed in the Annex D.

Table 4.2: Bitrate values for each codec

| | **Low** | **Medium** | **High** | **Very High** |
|---|---|---|---|---|
| **H.264/AVC** | 500 Kb | 800 Kb | 2 Mb | 16 Mb |
| **H.265/HEVC** | 200 Kb | 400 Kb | 1 Mb | 8 Mb |

### Packet Loss

Packet loss occurs when one or more packets fail to reach their destination during transmission or storage. The impairment caused by a packet loss depends on the encoding parameters, how the decoder handles errors, the packetization strategy, and the video content. This might cause flickering and blocking artifacts, which typically last for a few seconds, depending on the number and type of lost packet [142]. For the present experiment, all videos were first encoded using AVC (H.264) and HEVC (H.265) codecs. Then, packet loss artifacts were generated by dropping Network Abstraction Layer (NAL) packets from the video bit-stream similarly to what was previously done in other works [143]. In this experiment, we used the software NALTools, which was developed to insert a packet loss distortion in a video bitstream and has been used in several packet loss related studies [143, 144]. In order to avoid the generation of unrealistic strong artifacts, the standard error concealment algorithm of the corresponding codec [143] was used, which basically replaces a lost packet by the co-located packet from the previous frame during decoding. Five packet loss ratios were considered for this experiment: 1%, 3%, 5%, 8%, and 10%. These values replicate a real transmission scenario found in video streaming applications [142, 145].

### Frame Freezing

A frame freezing effect can be experienced on a progressive download of a video service, such as Video on Demand (VoD) and Youtube. Such types of video services are based on reliable transport mechanisms like the Transmission Control Protocol (TCP). In TCP, any lost or delayed packets are detected and a resend request is sent by the client. As a consequence, any user of progressive download services does not experience packet loss distortions contrary to what happens in the case of non-reliable transport mechanisms, such as User Datagram Protocol (UDP). When the available throughput is lower than the bitrate of the media, the reproduction will stall until enough data has been downloaded. This effect is perceived by the end users as freezing without skipping, commonly known

as rebuffering or stalling. The freezing effect is also experienced before the media starts its reproduction, this is known as the 'initial loading'.

Table 4.3: Organization of all Frame Freezing parameters

| | Level | Events | Pos1 | Pos2 | Pos3 | Len1 | Len2 | Len3 |
|---|---|---|---|---|---|---|---|---|
| **Low** | S1 | 1 | 2 | | | 2 | | |
| | S2 | 2 | 1 | 3 | | 1 | 3 | |
| **Medium** | S3 | 2 | 2 | 3 | | 2 | 2 | |
| | S4 | 3 | 1 | 2 | 3 | 2 | 2 | 3 |
| **High** | S5 | 3 | 1 | 2 | 3 | 3 | 3 | 2 |

For the present experiment, three parameters were considered for creating a frame freezing effect: 1) number of freezing events, 2) position of the freezing events in the sequence, and 3) length of the freezing event. Each video sequence was likely to have one, two or three freezing events. As for the position of the events, three possible options were chosen: "1", "2" and "3". The positions resulted from dividing by three the total length of the video sequence and multiplying it by: zero, one and two. A freezing event located at position "1" represents the initial loading experienced before the video starts playing. Finally, the length for the freezing events were fixed at 1, 2 and 4 seconds. All three parameters (number, position and length) were then organized and combined in order to represent the level of discomfort perceived by the user. The levels were set as "S1", "S2", "S3", "S4" and "S5", going from the least annoying combination (S1) to the most annoying combination (S5). Table 4.3 presents the organization of all parameters and their representation on this scale.

The initial loading and the stalling were inserted in the 480 codified videos using Avisynth (http://avisynth.nl). Avisynth is a powerful tool for video post-production; it is based on a script system allowing advanced non-linear editing. Regarding the audio component, silence was inserted using a faded in and out effect to avoid artifacts at the silence boundaries. Figure 4.3 presents a graphic illustration sample of all five levels of freezing distortion.

### 4.3.2 Audio Degradations

Four types of audio degradations were selected for this study: background noise, clipping, echo, and chop. The TCD-VoIP dataset [54] served as a reference to produce this set of audio distortions. The study had the goal of recreating some of the streaming audio degradations from the TCD-VoIP dataset on an audio-visual scenario.

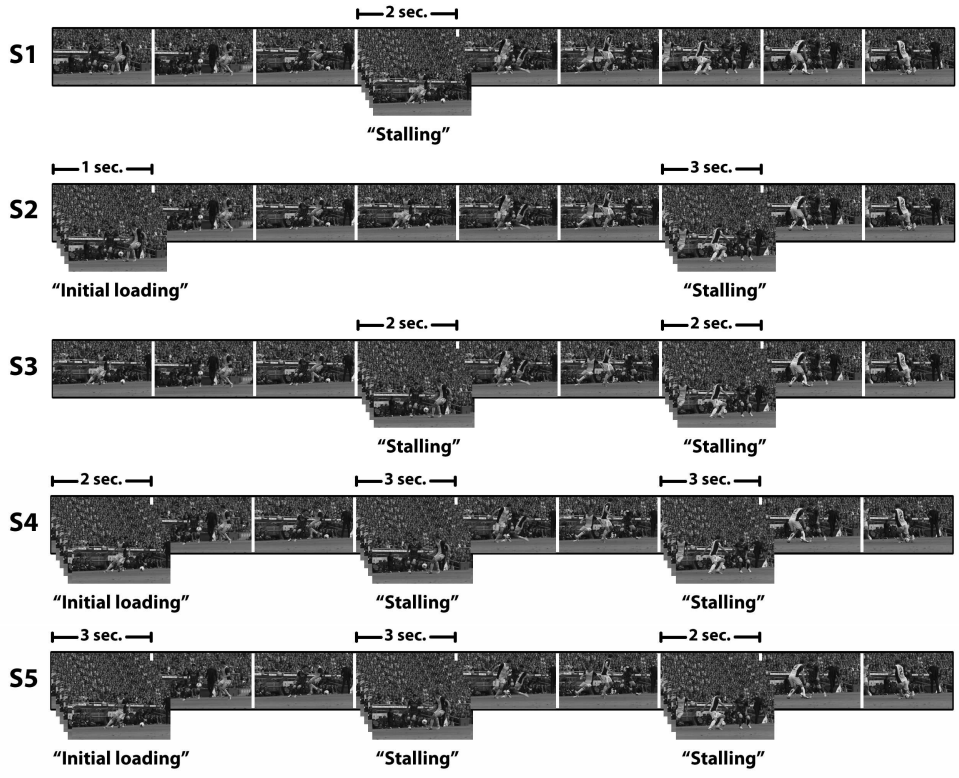



Figure 4.3: Freezing levels of distortion.

### Background Noise

As defined previously, background noise describes any sound that is not the sound under study. This study focuses on the called Non-stationary background noise, which is commonly found in our sound environment (e.g., traffic noise, alarms, and people talking). As mentioned before, this study takes the TCD-VoIP setup to recreate some of their test conditions by following the same processing method. The TCD-VoIP database uses four types of common noise: car noise, road noise, speech babble noise, and office noise. Samples for the car, road, and office noise were taken from a database of noise samples built by The European Telecommunications Standards Institute (ETSI). Meanwhile, the speech babble noise was created by combining random speech samples from the TSP speech database [146].

Four types of Background Noise (e.g. babble, car, road, and office) were added to the original signal at different SNR levels. Thus, two varying parameters were considered for this type of degradation: the type of background noise and the SNR level associated with the noise. Four combinations were selected, each one corresponding to a particular test condition. Table 4.4 details the four test conditions and their corresponding parameters.

Table 4.4: Audio Degradations and Parameters.

| Degradation | Conditions | Parameters | Range |
|---|---|---|---|
| Chop | 4 | Rate | 1, 2, 5 (chops/s) |
| | | Period | 0.02, 0.04 (s) |
| | | Mode | previous, zeros |
| Clip | 4 | Multiplier | 11, 15, 25, 55 |
| Echo | 4 | Alpha | 0.175, 0.3, 0.5 (%) |
| | | Delay | 25, 100, 140, 180 (ms) |
| | | Feedback | 0, 0.8 (%) |
| Noise | 4 | Noise type | car, babble, office, road |
| | | SNR | 15, 10, 5 (dB) |

### Clipping

As described before, a clipping type of distortion appears when a transmitted signal exceeds the maximum amplitude level permitted. This situation is handled by cutting the signal (clipping) to maintain a permitted level of amplitude. As a result, some samples become clipped and the signal quality gets compromised. On a VoIP call, the amplitude level might rise above the permitted limit due to a person's high voice volume when speaking into the microphone. The TCD-VoIP dataset creates this effect raising the amplitude level of the sequences by some constant, making that some sequence samples get clipped.

For this study, the clipping effect was generated by amplifying the signal using a multiplying factor. Four values of the amplitude multiplier were used to generate the test conditions. Table 4.4 details these four test conditions and their corresponding parameters.

### Echo

In a voice call, an echo effect normally occurs when a microphone picks up audio signals and send them back to its origin, thus creating a feedback loop. The TCD-VoIP database uses an echo scenario where copies of the signal being transmitted are picked up at the receiving microphone and then added to a returning signal. To simulate an echo effect, delayed versions of the signal at different SNR values were added to the original signal.

Following the TCD-VoIP dataset processing, the echo effect was produced by adding delayed versions of samples to the original signal. Three parameters were varied to generate different levels of distortion: 1) Alpha, the amplitude percentage of the first delayed version with respect to the original, 2) Delay, the time length between the first delayed version and the original, and 3) Feedback, the percentage reduction of the subsequent delayed versions. Four combinations were selected, each one corresponding to a partic-

ular test condition. Table 4.4 details the four test conditions and their corresponding parameters.

### Chop

A chop type of degradation is referred to transmitted signals with missing samples. The TCD-VoIP focus its study on the effect of missing samples due to hardware overload. One particular example of this scenario might be a CPU being overloaded during a voice call (e.g., video conferencing, smartphone call, etc.) causing the loss of some samples. For this setup, missing samples were handled with three types of approach: substitute the missing samples by silence, substitute the missing samples with previous repeated ones, or skyping the missing samples.

For this study, three parameters were varied to produce different levels and types of choppy speech: 1) Period, which sets the length of the discarded samples, 2) Rate, which indicates the frequency of the sample discard, and 3) Mode, which states how the discarded samples are handled. Four combinations were selected, each one corresponding to a particular test condition. Table 4.4 details the four test conditions and their corresponding parameters.

The above-described video and audio degradations, and the test conditions associated with each of them were used as the base to build all three audiovisual datasets. Source stimuli were processed according to different test conditions and they received a particular HRC number for each of the three experiments. These test conditions (HRCs) will be further presented on the subjective experiments description.

## 4.4 Apparatus and Physical Conditions

All three experiments were conducted at the University of Brasília (UnB), in a recording studio of the Núcleo Multimedia e Internet (NMI) of the Department of Engineering (ENE). Sound isolation was guaranteed during the experiment and only one participant was allowed during each experimental session. Hardware equipment consisted of a desktop computer, an LCD monitor, and a set of earphones. Additionally, a dedicated sound card (Asus Xonar DGX 5.1) was used to provide participants with an optimal sound experience (in terms of hardware). Detailed specifications of the equipment are presented in Table 4.5. The dynamic contrast of the monitor was turned off, the contrast was set at 100 and the brightness at 50. The room had the lights dimmed to avoid any light reflected on the monitor. The subjects were seated straight ahead of the monitor, centered at or slightly below eye height for most subjects. The distance between the subject's eyes and

Table 4.5: Equipment specifications

| Equipment | Technical Details |
|---|---|
| Monitor | Samsung SyncMaster P2370<br>Resolution: 1,920x1,080; Pixel-response rate: 2ms;<br>Contrast ratio: 1,000:1; Brightness: 250cd/m2 |
| Earphones | Sennheiser Hd 518 Headfone<br>Impedance: 50 Ohm; Sound Mode: Stereo;<br>Frequency response: 14–26,000Hz; |
| Sound Card | Asus Xonar DGX 5.1 |

the video monitor was set at three screen heights, which is a conservative estimate of the viewing distance according to the ITU-T Recommendation BT.500.1 [75].

The experiments were run using a quality assessment software developed by the Grupo de Processamento Digital de Sinais (GPDS), which was also used to record the subject's data (source code available at http://www.gpds.ene.unb.br/). The experimental interface was design using a client server model based on the HTML standard (version 5), using PHP, javascript, and a Postgresql database. The client-server model consists of a web server and a postgresql database running on the same station where the content is reproduced (HTML5 player). For all three experiments, the experimental session was controlled and started by the browser using a HTML5 interface to communicate with the server.

All three experiments were performed with volunteers from the University of Brasília, most of them were graduate students from the Computer Science and Electrical Engineering Departments. They were considered naïve of most kinds of digital video defects and the associated terminology. No vision or hearing tests were performed on the subjects, unimpaired hearing was a pre-requirement, moreover, participants were asked to wear glasses or contact lenses if they needed them to watch TV. Details about participants gender and age are presented in Table 4.6.

Table 4.6: Details about participants.

| Experiment | Participants | Female | Male | Age Range |
|---|---|---|---|---|
| Experiment 1 | 60 | 18 | 42 | 19-36 |
| Experiment 2 | 40 | 15 | 25 | 21-36 |
| Experiment 3 | 42 | 16 | 26 | 20-34 |

## 4.5 Experimental Methodology

As mentioned before, recommendations presented in the immersive method [72] were used for the set of experiments. Overall, the entire experimental session was divided into three sub-sessions: 1) Display Session, 2) Training Session, and 3) Main Session.


(a)'Quality scale' (b) 'Content scale'

Figure 4.4: ACR Quality and Content scales.

- *Display Session*

  For the display session, participants were presented with a set of original source video and their corresponding degraded versions (test conditions). The objective of this session was to familiarize the participant with the quality interval of the test sequences in the experiment. The display session considered an original source stimuli and the corresponding degraded versions of the sequence associated with a test condition (HRC). This procedure was repeated for each type of degradation considered in the experiment. As soon as the display session was over, a brief pause was made by the researcher to ask participants if they have perceived the difference between all test conditions and degradations, with the purpose of guaranteeing the consistency of the participants grading.

- *Training Session*

  In the training session, subjects performed the same tasks performed in the main session. The goal of the training session was to expose subjects to sequences with impairments and give them a chance to try out the data entry procedure. After observers were presented with the test stimuli, they were asked to answer two questions using two rating scales. The first question concerned the participant's perception of the overall audio-visual quality. To answer this question, participants were presented with a five-point Absolute Category Rating (ACR) scale ranging from 1 to 5. The five-point on this quality scale were labelled (in Portuguese) as "Excellent", "Good", "Fair", "Poor", and "Bad". Figure 4.4 depicts an image of the scale used for this experiment. The second question aims to gather information about the participant's personal opinion about the content. To answer this question, participants

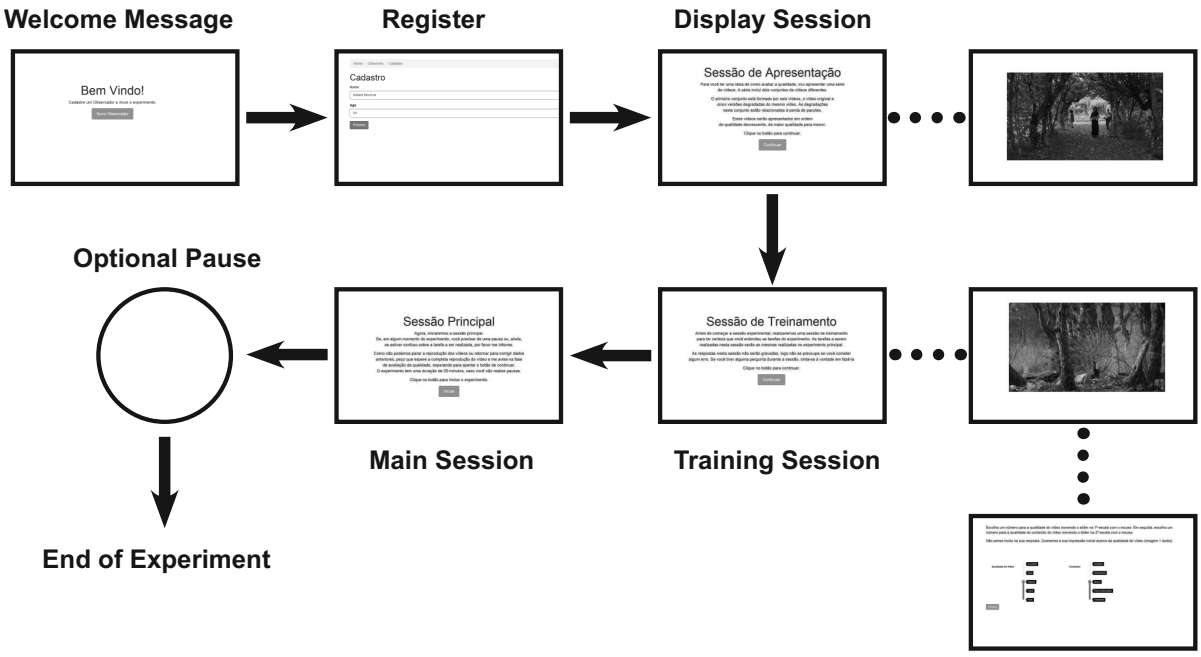



Figure 4.5: Steps of the video quality assessment experiment.

were presented with another five-point ACR scale, in which the five points of this content scale are labelled as “Intriguing”, “Interesting”, “Neutral”, “Uninteresting”, and “Boring”. These labels were inspired by the immersive speech experiment presented by Pinson et al. [72]. Figure 4.4 depicts an image of the two scales used in the experiments. Two training trials were included for this session. Once the training session was over, the participants were asked if they fully understood the functionality of the score entry interface.

- *Main Session*

  In the main session, the actual experimental task was performed. Figure 4.5 presents an illustration of the several steps of the experiments. Participants were presented with a number of sequences, out of the entire stimuli pool of the corresponding experiment. None of the presented videos had content equal to another video. For each session, participants were able to assess five stimuli for each HRC. Approximately five subjects rated each single stimuli, from the entire pool of test videos. The time of the experimental session was limited to 50 minutes. A break was introduced in the middle of the main session to allow the subjects to rest.

## 4.6 Statistical Analysis Methods

The judgments given by the subjects to any test sequence are called subjective scores. Traditionally, the data is first processed by calculating the mean opinion score (MOS). To obtain this value, subjective scores of all observers are averaged for each of the test stimuli. For this group of experiments, two different scores where gathered: the **quality** and **content** scores. These scores were averaged according to the type of HRC (ten HRCs and two anchors) and the original video sequences (sixty different sequences).

The **Mean Quality Score** (**MQS**) with respect to the set of HRCs is given by:

$$\mathrm{MQS}_{\mathrm{HRC(j)}} = \frac{1}{n} \cdot \sum_{i=0}^{n} S_j(i), \tag{4.1}$$

where $S_j(i)$ is the score reported by the $ith$ subject for the $jth$ element of the set $HRC = \{1, 2, \ldots, 12\}$ and $n$ is the total number of subjects. In other words, $\mathrm{MQS}_{\mathrm{HRC(j)}}$ gives the average average quality score for the $j-th$ HRC, measured over all subjects and contents originals. A similar notation is used for the **Mean Content Score** (**MCS**):

$$\mathrm{MCS}_{\mathrm{HRC(j)}} = \frac{1}{n} \cdot \sum_{i=0}^{n} S_j(i). \tag{4.2}$$

Therefore, $\mathrm{MCS}_{\mathrm{HRC(j)}}$ gives the average content scores for the $j-th$ HRC, measured over all subjects and contents originals.

## 4.7 Internal Consistency of the Results

The confidence levels are calculated, for each of the scores (MQS and MCS) of all three experiments. A high level of variability in the scores given by different subjects may indicate a low confidence level, thereby a low reliability of the results. Therefore, in order to evaluate the reliability of the results of the immersive methodology, we analyse the agreement among subjects on the questions. More specifically, we analyze the variation of: 1) the quality scores among all HRCs ($MQS_{HRC}$) and 2) the content score among all HRCs ($MCS_{HRC}$).

One of the most common measure techniques for internal consistency (reliability) is the Cronbach's alpha [147]. This coefficient is used as an estimate of the reliability of a psychometric test [148, 149]. The $\alpha$ coefficient ranges from 0 to 1, a greater value is interpreted as a greater internal consistency, i.e. more reliability. For coefficients in the range from 0.00 to 0.69 the internal consistency is considered poor, from 0.70 to 0.79 fair, from 0.80 to 0.89 good, and from 0.90 to 1 excellent. Table 4.7 presents the Cronbach's alpha coefficients for all $MQS_{HRC}$ and $MCS_{HRC}$ of all three experiments.

Table 4.7: Cronbach's $\alpha$ of both $MQS_{HRC}$ and $MCS_{HRC}$ questions for all three experiments.

| **Score** | **Analysis** | **Cronbach's $\alpha$** | **Experiment** |
|---|---|---|---|
| $MQS_{HRC}$ | per-HRC | 0.924 | Experiment 1 |
| $MCS_{HRC}$ | per-HRC | 0.858 | Experiment 1 |
| $MQS_{HRC}$ | per-HRC | 0.893 | Experiment 2 |
| $MCS_{HRC}$ | per-HRC | 0.841 | Experiment 2 |
| $MQS_{HRC}$ | per-HRC | 0.896 | Experiment 3 |
| $MCS_{HRC}$ | per-HRC | 0.864 | Experiment 3 |

For Experiment 1, the coefficient value for the $MQS_{HRC}$ on the per-HRC analysis was 0.924. The coefficient of $MCS_{HRC}$ for the same analysis (per-HRC) was 0.858. This suggests that subjects agreed more on the quality score than on the content score when the quality levels, represented by the HRCs, were shifted. As for Experiment 2, the coefficient for $MQS_{HRC}$ was 0.893, meanwhile the coefficient of $MCS_{HRC}$ was 0.841. Although the level of agreement is lower, this is still considered a good level of consistency. Finally, Experiment 3 coefficients of $MQS_{HRC}$ and $MCS_{HRC}$ were 0.896 and 0.864 respectively. As previous experiments, the level of consistency is good.

Given these results, it can be concluded that the MQS and MCS scores gathered during the group of experiments are highly reliable. This validates the use of the immersive method and encourages the execution of more experiments using this type of methodology.

## 4.8 Subjective Experiment 1 (video-only)

In this experiment, a group of volunteers was presented with a set of audio-visual sequences and were asked to rate the perceived quality of those sequences. The sequences were subjected to three types of distortions: video coding, packet loss, and frame freezing. The source pool used for the experiment consisted of a set of high definition audio-visual sequences. Impairments were only inserted into the video component, while the quality of the audio component remained constant. The objective of this particular experiment is to analyse different types of source degradations and compare the transmission scenarios where they occur. Given the nature of these degradations, the analysis is focused on the visual component of the sequence. The experiment was conducted using the basic directions of the immersive methodology described in the previous section. Although the experiment used the guidelines of the immersive methodology, some of the traditional recommendations were also considered for certain aspects of the experiment [84, 80]. This section describes the aim of the experiment, a list of the test conditions used to build the test pool stimuli, and an analysis of the gathered data from the experiment.

### 4.8.1 Test Conditions

For this experiment, video sequences were subject to impairments caused by compressing the original video at different bitrate levels (and codec algorithms), introducing packet losses simulating errors in the transmission, and frame freezing simulating degradations caused by delays in transmission. Users of progressive download services, in which any lost or delayed packets are detected and requested back, do not experience packet loss related video distortions. They do, however, experience playout pauses (frame freezing) when the available throughput is lower than the bitrate of the media. Since frame freezing and packet loss related video distortions do not occur simultaneously in a real transmission context [51], two groups of HRCs were considered. The first group combines artifacts produced by compression with packet loss video distortions (HRC1 to HRC5). The second group combines artifacts produced by compression with frame freezing effects (HRC6 to HRC10). Additionally, two video sequences compressed at extremely high bitrate levels, with no packet loss video distortions or frame freezing effects, worked as anchors to help participants recognize the entire range of quality used for the experiment. These anchors represented the equivalent of a no degraded sequences. The inclusion of these anchors might ease one of the drawbacks of the immersive method, which states that presenting audio-visual stimuli in an only-video study might cause saturation of the range scale.

Sixty (60) source stimuli were considered for the experiment, they all were subject to compression at four (4) different bitrate levels (low, medium, high, and very high) using two (2) coding algorithms (H.264 and H.265). This process resulted in four-hundred and eighty (480) video sequences (source stimuli x bitrate levels x codecs). Since the packet loss and the frame freezing cannot be present at the same transmission scenario (they both use different transmission mechanisms), two (2) groups of HRC combinations were formed. The first group considers the coding artifacts and the packet loss distortions, while the second group combines the coding artifacts and the frame freezing effects.

Regarding the first HRC group, five (5) combinations of bitrate levels and codecs were chosen, these five combinations represented five levels of quality. For each of these combinations a packet loss ratio was assigned (1%, 3%, 5%, 8%, and 10%). This resulted in five (5) HRCs which are presented in Table 4.8. These five HRCs are replicated for all sixty (60) source stimuli, resulting in three hundred (300) test stimuli.

As for the second HRC group, another five combinations of bitrate levels and codecs were used. It is worth mentioning that no combination used for the first group was used in the second group. Each of these five encoding combinations was paired with one of the five levels of the frame freezing discomfort scale (S1, S2, S3, S4, and S5). Five HRCs resulted from this combination. A more detailed display of these combinations is presented on Table 4.9. These five HRCs are replicated for all sixty source stimuli, resulting in three

Table 4.8: First group of HRCs.

| HRC | Codec | Bitrate (kb/s) | PLR |
|---|---|---|---|
| **HRC1** | H.264 | 500 | 10% |
| **HRC2** | H.265 | 400 | 8% |
| **HRC3** | H.264 | 2000 | 5% |
| **HRC4** | H.265 | 1000 | 3% |
| **HRC5** | H.265 | 8000 | 1% |

Table 4.9: Second group of HRCs.

| HRC | Codec | Bitrate (kb/s) | Freezing |
|---|---|---|---|
| **HRC6** | H.265 | 200 | S5 |
| **HRC7** | H.264 | 800 | S4 |
| **HRC8** | H.265 | 1000 | S3 |
| **HRC9** | H.264 | 2000 | S2 |
| **HRC10** | H.264 | 16000 | S1 |

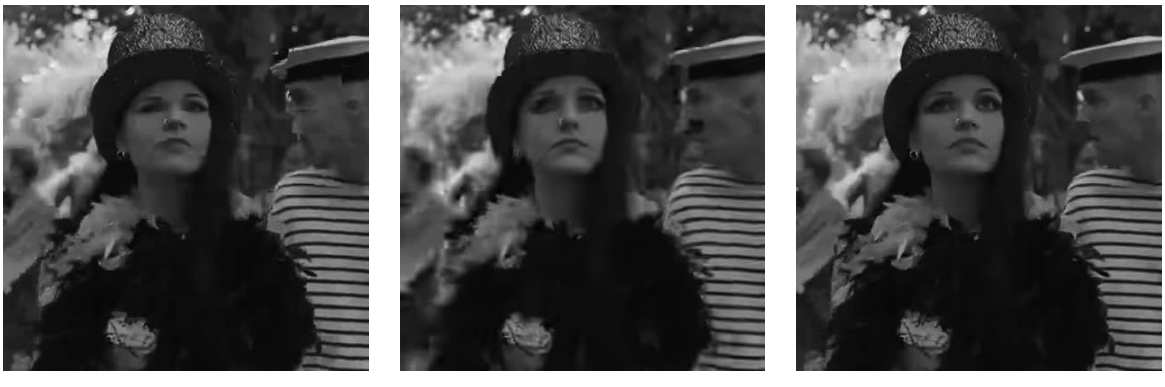
hundred (300) test stimuli.

As it was mentioned before, in order to ease a possible saturation on the range scale due to the usage of audio-visual stimuli, two anchors were considered. These two anchors were encoded using the H.264 and H.265 codecs at extremely high bitrate levels. These anchors are replicated for all sixty source stimuli, resulting in one hundred and twenty (120) test stimuli. Figures 4.7 and 4.7 presents sample frames of the two HRCs groups used for this experiment.

Pooling all test stimuli, seven hundred and twenty (720) test videos were generated for this experiment. It is important to mention that for each test session, the participant was presented with only 60 test stimuli of the 720 available. Each participant observed the content corresponding to an original sequence only once.

### 4.8.2 Experimental Results

Results of the experiment are presented and discussed in the following lines. As pointed out before, participants answered two questions about the video sequences they watched. The first question had the goal of collecting the opinion of the participant with respect to the audio-visual quality of the sequence (**quality score**). The second question gathered information about the participant's opinion about the content of the sequence (**content score**). Results were organized in terms of the HRCs considering two scenarios: Coding-PacketLoss scenario and Coding-Freezing scenario.



(a)'HRC 1' (b) 'HRC 2' (c) 'HRC 3'

(d) 'HRC 4' (e) 'HRC 5' (f) 'ANC 1'

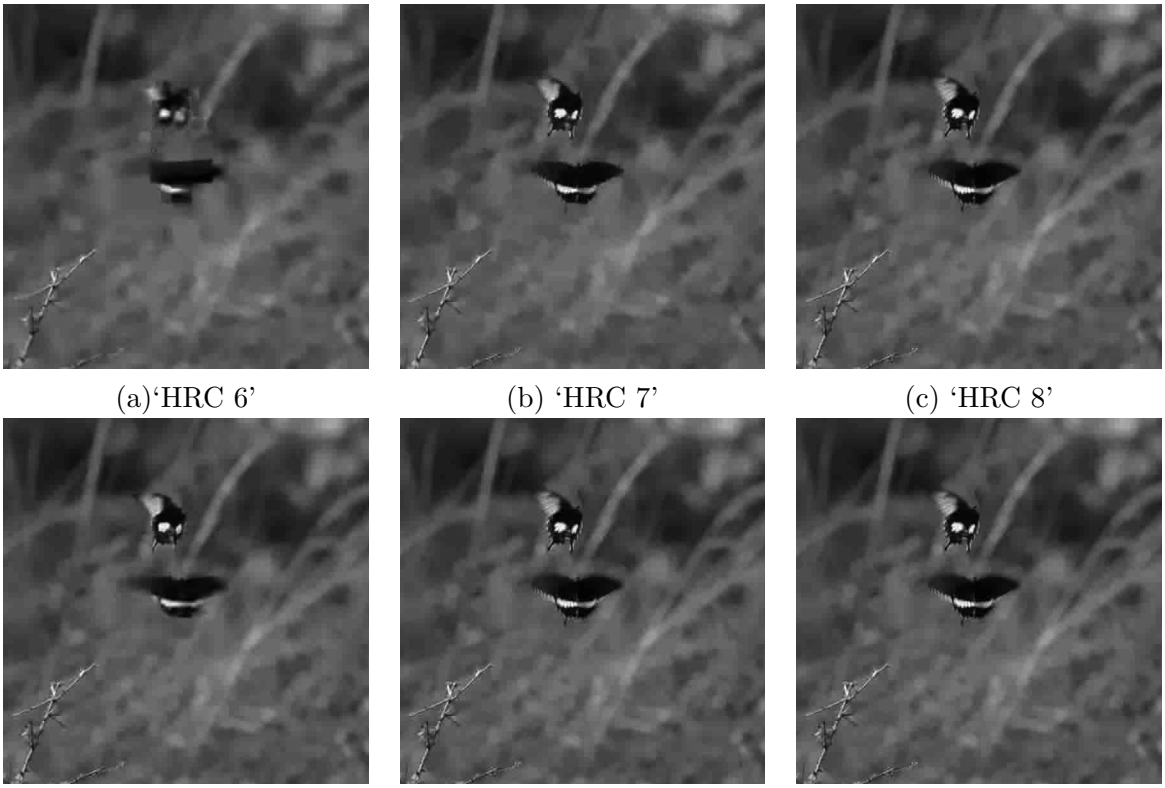
Figure 4.6: Sample frames of the firts group of HRCs.

### Results

Two main scenarios were considered for the organization of the HRCs used in this experiment. The first scenario, which corresponds to HRCs from 1 to 5 (including the anchor 1), presented coding impairments and distortions due to packet loss (see Table 4.8). The second scenario, which corresponded to HRCs from 6 to 10 (including the anchor 2), presented coding impairments and frame freezing distortions (see Table 4.9).

- *Coding-PacketLoss Scenario*

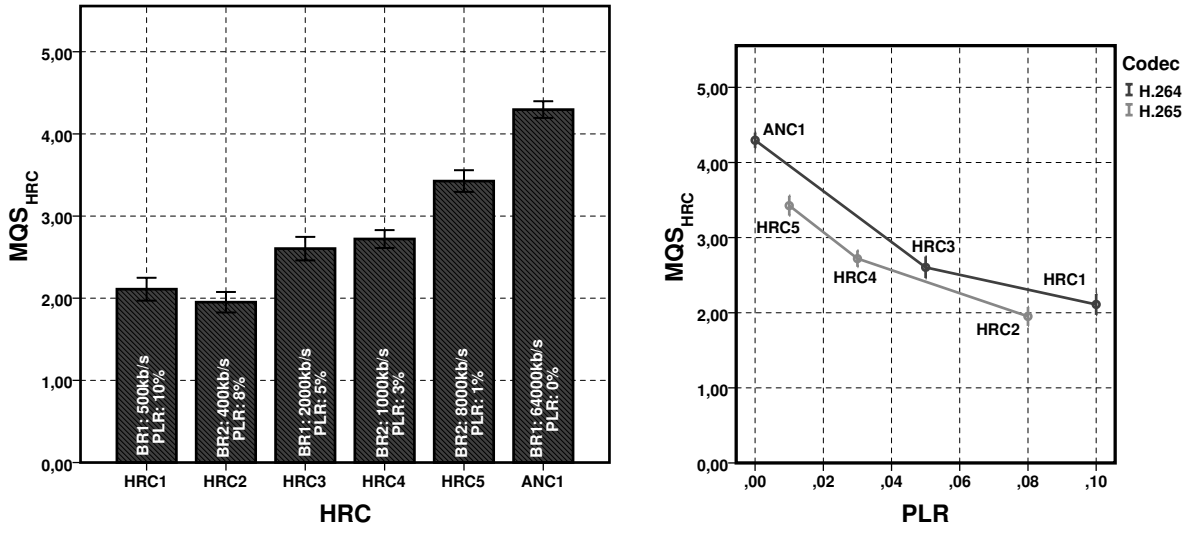
  Figure 4.8 (a) presents the $MQS_{HRC}$ values, including a 95% confidence interval, for the coding-packeloss scenario. Each HRC is paired with a bitrate level and a packet loss rate (HRC1 = 500kb/s - 10%, HRC2 = 400kb/s - 8%, HRC3 = 2000kb/s - 5%, HRC4 = 1000kb/s - 3%, HRC5 = 8000kb/s - 1%). As it can be observed, the $MQS_{HRC}$ increases along with most of the BR and PLR combinations. Such increase is not seen for HRCs 1 and 2, in fact, only a small difference (with no statistical significance) is observed between them. An early analysis might suggest that this difference is caused by the coding algorithm used on the HRCs and its response to the packet loss insertion algorithm. The $MQS_{HRC}$ values fall on the range of 1.95



(a)'HRC 6' (b) 'HRC 7' (c) 'HRC 8'

(d) 'HRC 9' (e) 'HRC 10' (f) 'ANC 2'

Figure 4.7: Sample frames of the second group of HRCs.




(a) (b)

Figure 4.8: **(a)** $MQS_{HRC}$ for the coding-packeloss scenario. **(b)** $MQS_{HRC}$ according the Packet loss rate. **HRC1**: BR = 500kb/s, PLR = 10%. **HRC2**: BR = 400kb/s, PLR = 8%. **HRC3**: BR = 2000kb/s, PLR = 5%. **HRC4**: BR = 1000kb/s, PLR = 3%. **HRC5**: BR = 8000kb/s, PLR = 1%. **ANC1**: BR = 64000kb/s, PLR = 0%. Legend: **BR1** = bitrate coded with H.264, **BR2** = bitrate coded with H.265, **PLR** = packet loss rate.

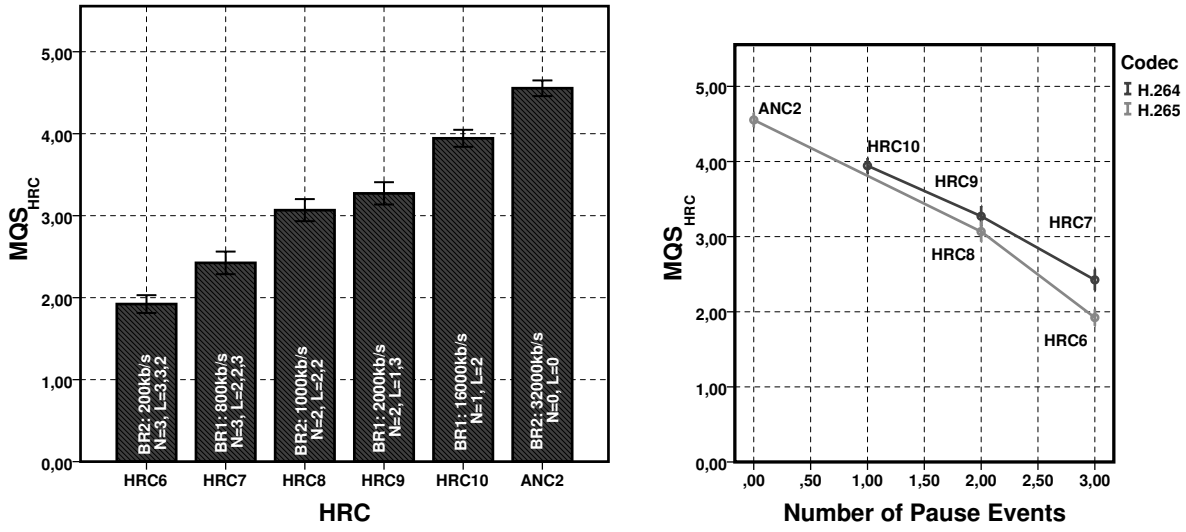
and 4.30 with no evidence of scale saturation. This suggests that participants were able to distinguish between the different levels of impairments used for this scenario.

Figure 4.8 (b) depicts the $MQS_{HRC}$ as a function of the packet loss rate values (PLR). The figure displays the different HRCs for both H.264 and H.265 codecs. It can be observed that the $MQS_{HRC}$ drops as the PLR is increased and the bitrate is decreased. However, a very similar $MQS_{HRC}$ value is observed for two different cases. The $MQS_{HRC}$ for HRC4 (PLR = 3% , BR = 1000kb/s, and Codec = H.265), does not exhibit a statistical difference when compared to the HRC3 (PLR = 5% , BR = 2000kb/s, and Codec = H.264). From previous studies [150, 139], it is known that a similar subjective quality is expected for a video encoded with H.265 with a 50% bitrate savings compared to a video encoded with H.264. Such behaviour is observed for HRC4 and HRC3 (1000kb/s, H.265 and 2000kb/s, H.264), although it is worth pointing out that, they both differ on their packet loss rate values (3% and 5%). This might indicate that the coding algorithms responded differently to packet loss impairments. From the literature [151, 152], it has been shown that H.265 is very sensitive to packet losses and less error resilient when compared to H.264. This might explain why a higher PLR (5% for HRC3) does not exhibit a significant difference when compared to a lower PLR (3% for HRC4). On these grounds, it is easy to explain the difference observed on the $MQS_{HRC}$ for HRC2 (PLR = 8% , BR = 400kb/s, and Codec = H.265) and HRC1 (PLR = 10% , BR = 500kb/s, and Codec = H.264). For this case, the sensitiveness of the H.265 to a packet loss is having a greater effect on $MQS_{HRC}$ than the bitrate.

For this first scenario, it has been observed that the video bitrate, the coding algorithm, and the PLR all have an important impact on the perceived audio-visual quality ($MQS_{HRC}$). However, for certain rates of packet looses, the coding algorithm is proven to be highly determinant.

- *Coding-Freezing Scenario*

  Figure 4.9 (a) presents the $MQS_{HRC}$ values, including a 95% confidence interval, for the coding-freezing scenario. Each HRC is paired with a frame freezing level of distortion, denoted by the number of pause events (N), the position of the pause event (P), and the length of the pause events (L). Detailed descriptions of the frame freezing levels of distortion are presented in Table 4.3 and Figure 4.3. It can be observed that the $MQS_{HRC}$ increases for a high bitrate (BR) level and a low pause frequency, i.e. number of pause events (N). This steep increasing pattern is observed along all HRCs. The $MQS_{HRC}$ values are on the range of 1.92 and 4.55




(a) (b)

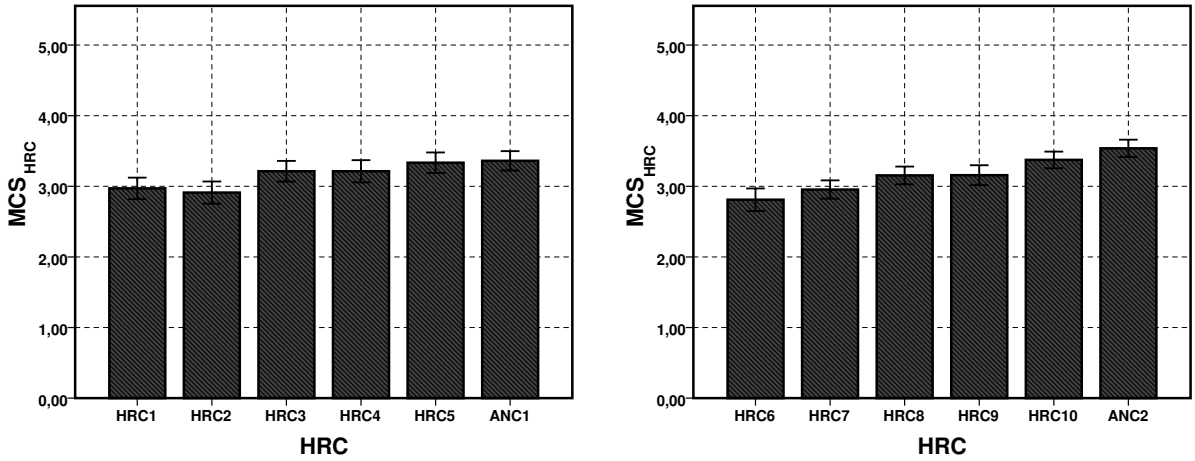
Figure 4.9: **(a)** $MQS_{HRC}$ for the coding-freezing scenario. **(b)** $MQS_{HRC}$ according the Number of pause events. **HRC6**: BR = 200kb/s, N = 3, P = 1-2-3, L = 3-3-2. **HRC7**: BR = 800kb/s, N = 3, P = 1-2-3, L = 2-2-3. **HRC8**: BR = 1000kb/s, N = 2, P = 2-3, L = 2-2. **HRC9**: BR = 2000kb/s, N = 2, P = 1-3, L = 1-3. **HRC10**: BR = 200kb/s, N = 1, P = 1, L = 2. **ANC2**: BR = 32000kb/s, N = 0, P = 0, L = 0. Legend: **BR1** = bitrate coded with H.264, **BR2** = bitrate coded with H.265, **N** = Number of pause events, **P** = Position of the pause events, **L** = Length of he pause events.

with no evidence of scale saturation. This suggests that participants were able to distinguish between the different levels of impairments used for this scenario.

Figure 4.9 (b) presents the $MQS_{HRC}$ as a function of the number of pause events (N). The figure presents the different HRCs for both H.264 and H.265 codecs. For the particular case of HRC8 and HRC9 (same number of pause events), it can be inferred that the $MQS_{HRC}$ difference was determined by the position (P) and length (L) of the pause events, since a certain equivalence is expected in terms of bitrate [150, 139]. For HRC9, the pause events were located at positions "1" and "3", and their durations were 1 and 3 seconds respectively. For HRC8, the pause events were located at positions "2" and "3", and their durations were 2 seconds for both pauses. By comparing these values, we can see that a short pause at the beginning of the playout (initial loading) is less annoying than a pause during the playout. Such affirmation is verified by several studies in the literature [153]. For the case of HRC6 and HRC7 (same number of pause events), the higher difference might be attributed to their bitrate levels (200kb/s and 800kb/s) and the positions and duration of both HRCs. For HRC6, the pause events were located at positions "1" , "2" and "3", and their durations were 3, 3, and 2 seconds respectively. For HRC7, the pause events were located at positions "1" , "2" and "3", and their durations were 2, 2, and 3 seconds respectively. Clearly, a higher initial loading affected the

perceived quality.

On the basis of these results, we conclude that there is an additive impact of pauses and video bitrate on the perceived audio-visual quality. Such impact can be determined by the number of pause events and their position and duration.




(a)'Coding-Packetloss Scenario' (b) 'Coding-Freezing Scenario'

Figure 4.10: $MCS_{HRC}$ for both scenarios. All MCS were averaged in terms of HRC.

- *Content Score Results*

  Regarding the MCS, Figures 4.10 (a) and 4.10 (b) present the $MCS_{HRC}$ for each HRC corresponding to both coding-packetloss and coding-freezing scenarios. As pointed out before, the five points of the content scale are labeled as "Intriguing" = 5, "Interesting" = 4, "Neutral" = 3, "Uninteresting" = 2, and "Boring" = 1. It is observed that the range of values for the $MCS_{HRC}$ values gets reduced, for both scenarios, and it fluctuates around a "Neutral" value. Such drop is caused by the averaging of all content responses over all of HRCs. This averaging helps to distinguish among different HRC levels in terms of $MQS_{HRC}$, but it does not provide a good representation of the $MCS_{HRC}$ when all video contents are "averaged".

  Although the $MCS_{HRC}$ range is smaller, it is possible to observe a pattern on both values of the figures. The $MCS_{HRC}$ varies as the level of impairment varies (HRC). This behaviour suggests that participant's opinion about the content is in accordance with its opinion about its quality. Such behaviour is better visualised in Figure 4.11 where the evolution of both $MQS_{HRC}$ and $MCS_{HRC}$ are plotted along all HRCs. These results reinforce the premise that participant's perception of quality are influenced, at a certain level, by the video content [82, 154].

  In spite of the existence of a mutual impact of the video content and quality level, it is not yet possible to uncover the mechanisms behind this impact. It is clear though,

that the usage of a content analysis methods, combined with quality assessment for audio and video, would certainly improve the performance of the audio-visual quality estimation.

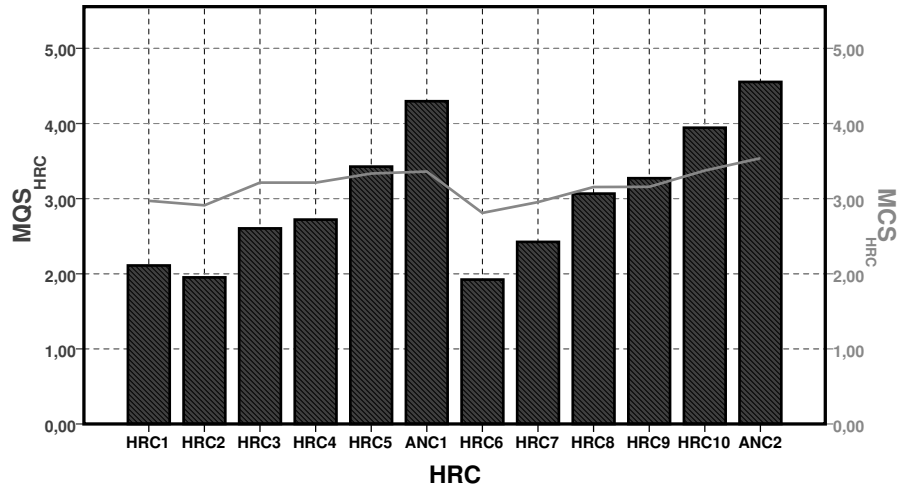



Figure 4.11: Evolution of both $MQS_{HRC}$ and $MCS_{HRC}$ scores along all HRCs.

Figure 4.12 presents the MQS values obtained for each of the HRCs, along with the single user scores for experiment 1. It can be observed that for most of the test conditions, the results gathered are more consistent, i.e., the spread of points is smaller. More particularly, test conditions where the perceived quality got higher responses presented more consistent results.

### Objective Comparison

Additionally, the NR video quality metric VIIDEO [107] was used to predict the MQS of the sequences. For presentation purposes, results from the VIIDEO metric were scaled in the interval (1,5). Figure 4.13 shows a scatter plot comparing the predicted quality using the VIIDEO metric and the $MQS_{HRC}$ results organized according to the packet loss and frame-freezing scenarios. The overall Pearson correlation coefficient achieved is $\rho = 0.87$. It is observed that the coding-freezing scenario presents a subtle advantage on the $MQS_{HRC}$.

Finally, Figure 4.14 (a) presents a scatter plot comparing subjective results of the $MQS_{HRC}$ and $MCS_{HRC}$ for the two scenarios (coding-packetloss and coding-freezing). It is observed that, participants gave higher $MQS_{HRC}$ and $MCS_{HRC}$ responses to the coding-freezing scenario compared to the coding-packetloss scenario. These results show that participants were more tolerant to pauses during the video playout than to severe visual distortions (blocking, slicing, blockloss) caused by packet loss. Such tolerance is reflected

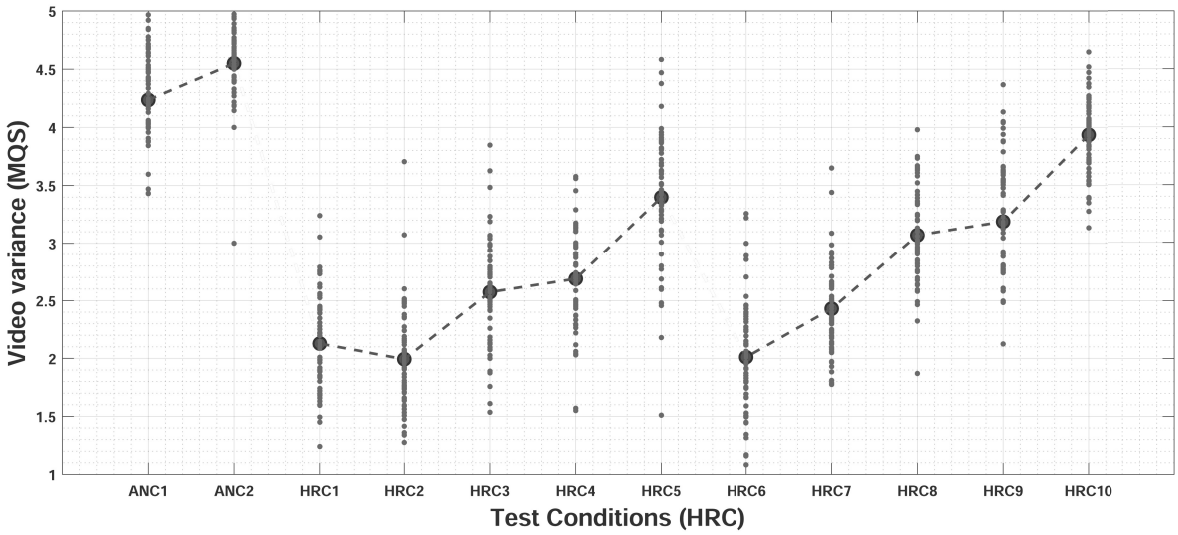



Figure 4.12: Mean Quality Score (MQS), and its respective spread of scores, for the different Hypothetical Reference Circuit (HRC) degradations.

also on the $\mathrm{MCS}_{\mathrm{HRC}}$, suggesting that the participant's opinion of the content is affected by visual distortions present in the videos.

## 4.9 Subjective Experiment 2 (audio-only)

In this experiment, we used the immersive methodology to perform a subjective experiment with the goal of estimating the quality of audio-visual sequences. Quality scores were gathered for a set of audio-visual sequences with distortions only in the audio component. The TCD-VoIP dataset [54] served as a reference to produce a new audio-visual dataset with only-audio distortions: the Im-AV-Exp2. The experiment had the goal of recreating some of the streaming audio degradations from the TCD-VoIP dataset on an audio-visual scenario and analyzing the effect of such degradations on the perceived audio-visual quality. More importantly, the experiments had the goal of testing the effect of the visual content on the overall quality. Findings from these experiments will be used to analyse the relationship between streaming and compression artifacts on audio-visual quality. This section describes the aim of the experiment, list the test conditions used to build the test pool stimuli, and performs an analysis of the gathered data from the experiment.

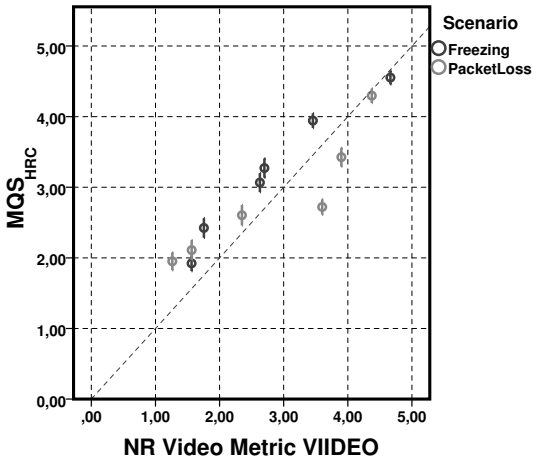



Figure 4.13: Scatter plot showing the objective estimates from the NR video metric VIIDEO versus the $\text{MQS}_{\text{HRC}}$ for both scenarios. Overall correlation $\rho = 0.87$.

### 4.9.1 Test Conditions

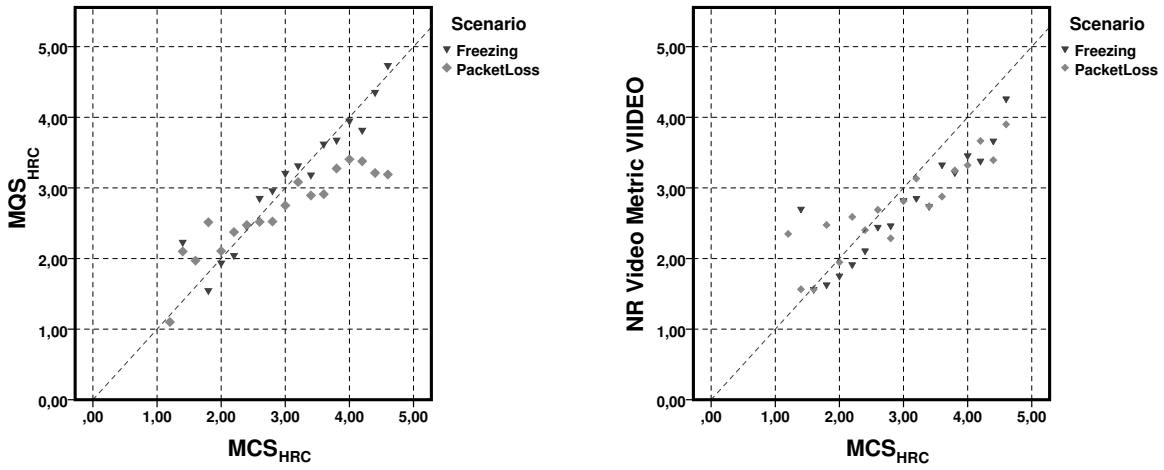
As mentioned before, four common streaming types of degradations were considered for this particular experiment. The Im-AV-Exp2 dataset was built following the same processing method used in the TCD-VoIP dataset. For each type of degradation, four test conditions were selected from the TCD-VoIP dataset and presented as a particular Hypothetical Reference Circuit (HRC). These test conditions were selected empirically with the goal of covering the entire range of quality observed in the TCD-VoIP dataset. As a result, sixteen (16) HRC arrangements were obtained. The HRCs were organized according to the type of degradation. Additionally, one test condition without degradations was used as an anchor (ANC) to help participants establish the range of quality used in the experiment. Next, a brief description of the degradations and the procedure used in the experiment is presented.

- *Background Noise*

  Four types of Background Noise (e.g. babble, car, road, and office) were added to the original signal at different SNR levels. Four combinations were selected, each one corresponding to a particular HRC (HRC1 to HRC4). Table 4.10 details the four HRCs, their corresponding parameters, and the anchor test condition (ANC1).

- *Chop*

  Three parameters were varied to produce different levels and types of choppy speech: 1) Period, which sets the length of the discarded samples, 2) Rate, which indicates the frequency of the sample discard, and 3) Mode, which states how the discarded samples are replaced. Four combinations were selected, each one corresponding




(a) (b)

Figure 4.14: **(a)** Scatter plot comparing subjective results of $MQS_{HRC}$ and $MCS_{HRC}$ for the two scenarios. **(b)** Scatter plot comparing the predicted $MQS_{HRC}$ (VIIDEO) versus the subjective $MCS_{HRC}$ for the two scenarios.

to a particular HRC (HRC5 to HRC8). Table 4.10 details the four HRCs, their corresponding parameters, and the anchor test condition (ANC2).

- *Clipping*

  A clipping effect was produced by amplifying the signal using a multiplying factor. Four values of the amplitude multiplier were used to generate the HRCs (HRC9 to HRC12). Table 4.10 details these four HRCs, their corresponding parameters, and the anchor test condition (ANC3).

- *Echo*

  An echo effect was produced by adding delayed versions of samples to the original signal. Three parameters were varied to generate different levels of distortion: 1) Alpha, the amplitude percentage of the first delayed version with respect to the original, 2) Delay, the time length between the first delayed version and the original, and 3) Feedback, the percentage reduction of the subsequent delayed versions. Four combinations were selected, each one corresponding to a particular HRC (HRC13 to HRC16). Table 4.10 details the four HRCs, their corresponding parameters, and the anchor test condition (ANC4).

### 4.9.2 Experimental Results

Results of the experiment are presented and discussed in the following lines. The participant's opinion with respect to the audio-visual quality of the sequences in Experiment 2 is organized in terms of the HRCs considering all four audio degradations: Background

Table 4.10: HRC corresponding parameters used in Im-AV-Exp2. Anchor test conditions (ANC).

| **BG Noise** | **Noise** | **SNR (dB)** | |
|---|---|---|---|
| HRC1 | car | 15 | |
| HRC2 | babble | 10 | |
| HRC3 | office | 10 | |
| HRC4 | road | 5 | |
| ANC1 | - | - | |
| **Chop** | **Period (s)** | **Rate (chops/s)** | **Mode** |
| HRC5 | 0.02 | 1 | previous |
| HRC6 | 0.02 | 2 | zeros |
| HRC7 | 0.04 | 2 | previous |
| HRC8 | 0.02 | 5 | zeros |
| ANC2 | - | - | - |
| **Clipping** | | **Multiplier** | |
| HRC9 | | 11 | |
| HRC10 | | 15 | |
| HRC11 | | 25 | |
| HRC12 | | 55 | |
| ANC3 | | - | |
| **Echo** | **Alpha (%)** | **Delay (ms)** | **Feedback (%)** |
| HRC13 | 0.5 | 25 | 0 |
| HRC14 | 0.3 | 100 | 0 |
| HRC15 | 0.175 | 140 | 0.8 |
| HRC16 | 0.3 | 180 | 0.8 |
| ANC4 | - | - | - |

noise, Clipping, Chop, and Echo. Additionally, results are compared against the subjective results of the TCD-VoIP database.

### Results

This section presents the analysis of the degradation conditions (i.e. Echo, Chop, Clip, Noise degradations), which are considered service aspects that may be affected during streaming. Figure 4.15 presents the *Mean Quality Score* (MQS), including a 95% confidence interval, for all HRCs corresponding to the four audio distortions. Results are grouped according to the corresponding audio distortion type. Anchor test conditions are highlighted in white.

For the Background Noise distortion type, each HRC corresponds to a combination of a noise type and an SNR value, as detailed in Table 4.10. It can observed that the quality scores rarely reach 3 points in the MQS scale. These results are in accordance with previous results that showed that, for noise SNR values bellow 20dB, the quality scores are around 3 points or less [54]. The MQS values vary from 2 to 3 points. Analyzing the parameters, it can be observed that sequences with an SNR value below 15dB obtained quality scores smaller than 3 points. For the particular case of HRC2 and HRC3, which both present the same SNR value (10dB), it can be observed that the babble noise was more annoying than the office noise. Such behavior is again in accordance with results from

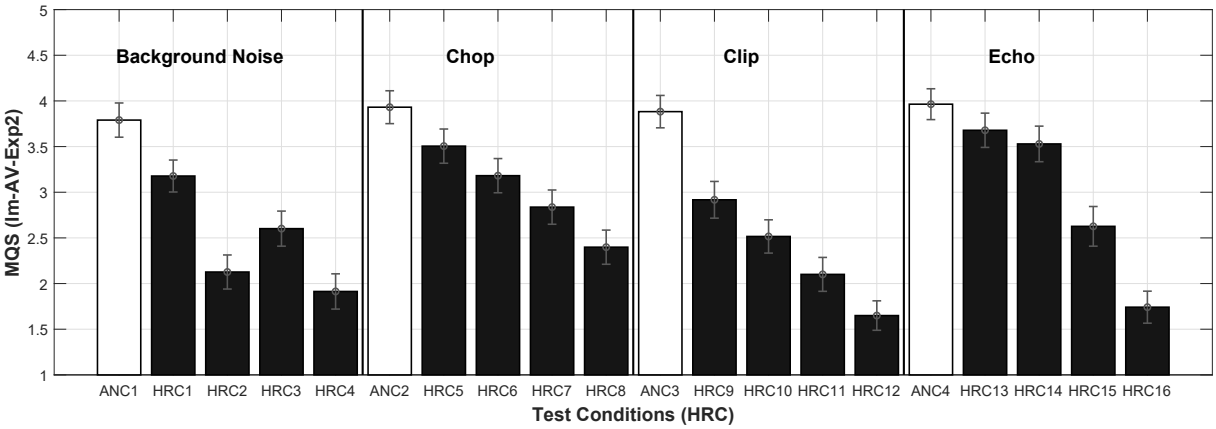



Figure 4.15: MQS for all four distortions. See HRC specifications in Table 4.10.

previous (audio-only) experiments [54]. Regarding the MQS corresponding to sequences affected by road noise, the poor quality scores might be attributed to the low SNR noise value (5dB).

For the Chop distortion, each HRC corresponds to a combination of three parameters (rate, period, and mode), as detailed in Table 4.10. It can be noticed that the MQS values vary from 2.5 to 3.5, with the MQS values decreasing from HRC5 to HRC8. This behavior seems to be closely related to the chop rate value. An analysis of the parameters suggests that the perceived quality decreases as the chop rate increases, independently of the chop mode. In particular, for a fixed rate of 2 chops/second, repeating previous portions of samples (*previous* mode) is slightly more annoying than inserting silence portions (*zeros* mode). For the particular case of HRC8, where the chop rate corresponds to 5 chops/seconds, MQS fluctuates around 2.5 points. This result is again in accordance with earlier (audio-only) experiments, which have shown that a chop rate of 3 chops/second leads to quality scores below 3 points [54]. Comparing the MQS in terms of the chop mode and of the chop period, it can be observed that inserting silence portions (*zero* mode) with a period of 0.02s is the equivalent of repeating portion samples (previous mode) with a period of 0.04s. Comparing both chop modes, at a fixed period of 0.02s, showed that using a *zero* mode produces lower quality scores than using a *previous* mode.

For the Clip distortion type, the MQS values vary between the 3 and 1.5 points in the MQS scale. All four HRCs values decrease from HRC9 to HRC12. As it can be observed, clipped distortions have quality scores below 3 for all four condition levels. Such results might suggest that clipped distortions are perceived as more severe. For the particular case of HRC9 and HRC10, where the multipliers values are 11 and 15, respectively, quality scores below the 3 points are observed. These results are particularly interesting since previous (audio-only) experiments found similar quality scores for multiplying factors above 18 [54].

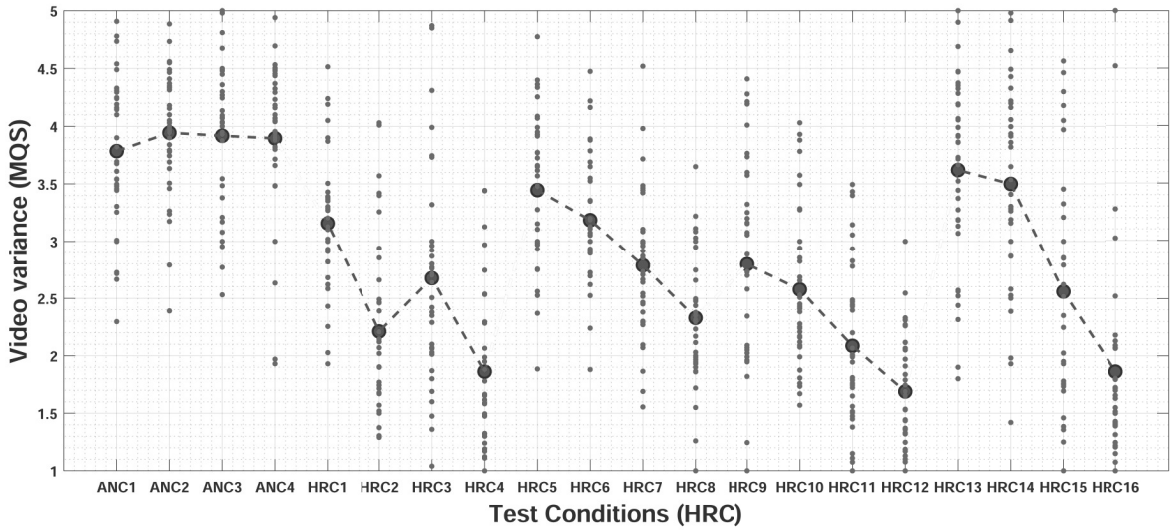



Figure 4.16: Mean Quality Score (MQS), and its respective spread of scores, for the different Hypothetical Reference Circuit (HRC) degradations.

For the Echo distortion type, each HRC corresponds to a combination of three parameters (alpha, delay, and feedback), as detailed in Table 4.10. The MQS values vary between 3.7 and 1.7 points in the MQS scale. Although the HRCs quality values decrease from HRC13 to HRC16, an abrupt drop in MQS is observed between HRC14 and HRC15. For this particular case, it can be observed that the presence of a feedback affects considerably the perceived quality. These results were also observed in previous audio-only studies, where the inclusion of a feedback produced the lowest quality scores [54].

An analysis of the parameters shows that the feedback factor has a strong influence on the perceived quality. Sequences with a feedback factor of 0.8% have quality scores below 3 points in the MQS scale. It can also be observed that the variation of the echo alpha value impacts only sequences with a feedback factor. Perceived quality values fluctuate between 1.5 and 4 in the MQS scale. By comparing the quality scores obtained for HRC13 and HRC14, we notice that a certain balance can be reached by using a large amplitude factor with a short delay (0.5% and 25 ms) or a lower amplitude with a larger delay (0.3% and 100 ms). Regarding the results corresponding to HRC15 and HRC16, a large amplitude factor combined with a large delay (0.3% and 180 ms) results in lower quality scores (HRC16).

Figure 4.16 presents the MQS values obtained for each of the HRCs, along with the single user scores for experiment 2. It can be observed that for most of the test conditions, the results gathered are not as consistent as the ones obtained in experiment 1. With the exception of the anchor test conditions (ANC1, ANC2, ANC3 and ANC4), the spread of the score points was high. This might suggest that there was less agreement regarding

audio types of distortion compared to the agreement observed for experiment 1 where distortions were inserted only in the video component.

### Comparison of Datasets

As mentioned earlier, the Im-AV-Exp2 dataset was built by recreating, in the audio component of the audio-visual stimuli, a number of test conditions of the TCD-VoIP dataset. In this section, we compare the objective and subjective quality responses for both datasets. It is worth pointing out that there are obvious differences between the two datasets. First, the TCD-VOIP dataset contains only speech audio sequences, while the Im-AV-Exp2 dataset contains speech, sport, movies, and music audios in audio-visual sequences. Second, the two datasets used different experimental methodologies to collect the subjective scores. Despite these differences, a comparison of theses two datasets can provide interesting insights regarding the impact of the visual component on the overall quality perception, when the stimuli contains streaming degradations (only) in the audio component.

To perform this comparison, we used two versions of an objective quality metric to establish a similar measure for both datasets. In TCD-VoIP, the VISQOL speech model [115] was used to estimate the speech quality of the stimuli. Meanwhile, in Im-AV-Exp2, the VISQOLAudio quality metric [40] was used to obtain the quality of the audio component of the stimuli. Then, we compared the subjective quality scores, **MQS (Im-AV-Exp2)** and **MOS (TCD-VoIP)**, of both datasets with the corresponding VISQOL objective scores, **VISQOL (Im-AV-Exp2)** and **VISQOL (TCD-VoIP)**. Figure 4.17 depicts scatter-plots showing comparisons of these objective and subjective scores. In the graphs, data from both datasets are plotted, with points corresponding to the different types of degradations being identified by different colors.

Figures 4.17 (a) and (b) show the subjective scores versus the VISQOL scores for the Im-AV-Exp2 and TCD-VoIP datasets, respectively. Notice that the VISQOL metric tends to over-estimate the quality for all degradations in both datasets. Interestingly, we observe that VISQOL ranked all degradations in the same order for both datasets, i.e. Chop degradations were rated as less annoying, followed by Clip, Echo, and Noise degradations. These results show that the characteristics of the audio degradations seem to be affecting the perceptual quality of the stimuli of both datasets in a similar way.

Figure 4.17 (c) depicts a scatter-plot of the VISQOL scores for TCD-VOIP versus the VISQOL scores for Im-AV-Exp2. From the plot in this figure, we can notice that, in both datasets, there is a consistency of the results corresponding to the Chop degradations (identical results would correspond to points in the diagonal traced line). It is worth pointing out that the VISQOL scores for the Chop degradations had high values (over

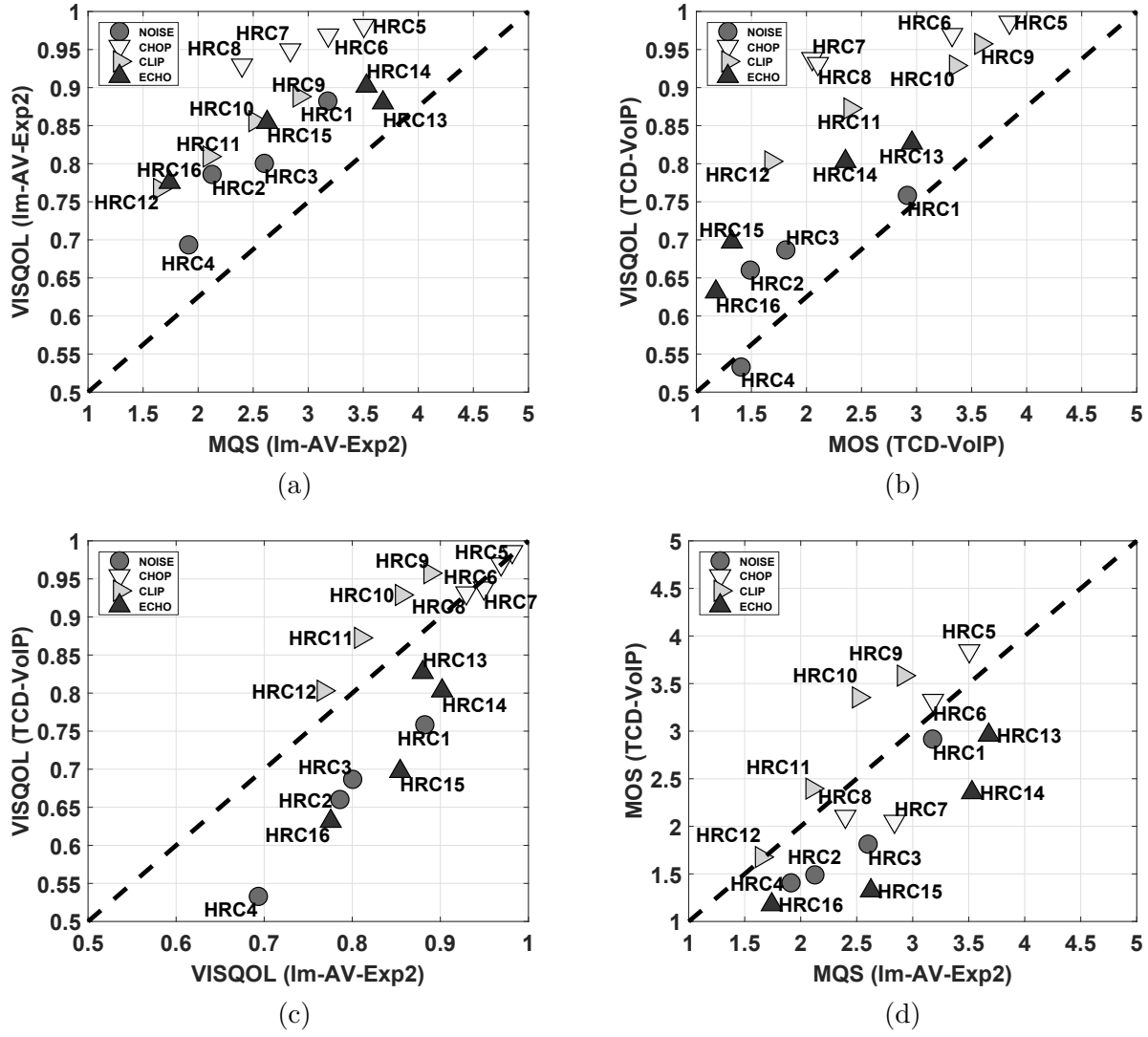



Figure 4.17: Subjective-Objective comparison for Im-AV-Exp2 and TCD-VoIP.

0.9 for both datasets). Regarding the Clip degradations, the VISQOL scores obtained for the TCD-VoIP dataset were higher than the VIQOL scores obtained for the Im-AV-Exp2 dataset (i.e. points are above the diagonal line). The VISQOL scores for Echo and Noise degradations, on the other hand, were smaller for the TCD-VoIP dataset than for the Im-AV-Exp2 dataset (i.e. points below the diagonal line). This result shows that, although the sample conditions for both datasets were generated using the same technique, the content had an influence on the perceived quality, causing a quality increase (Clip) or decrease (Chop and Noise) for speech content (TCD-VoIP) to general audio content (Im-AV-Exp2).

Figure 4.17 (d) depicts a scatter-plot of subjective scores for TCD-VOIP versus the subjective scores for Im-AV-Exp2. The comparison is made between the audio-only quality scores, MOS(TCD-VoIP), and the corresponding audio-visual scores, MQS(Im-AV-Exp2).

Notice that, although these quality scores come from different experiments with different content and different conditions, there is again a consistency between the audio-only and audio-visual scores for the Clip and Chop degradations, with only a few exceptions are far from the diagonal line (HRC7, HRC9, and HRC10). The subjective scores for the Noise degradations lie below the diagonal line, but not too far from it. It is interesting to note that, for the Echo degradations, the audio-only subjective scores are consistently higher than the audio-visual scores (i.e. points are below the diagonal line). This suggests that the video component has a more pronounced impact for Echo degradations, acting as a masking factor and producing higher quality scores. In other words, the Echo degradation had a smaller impact on the perceived overall quality of audio-visual stimuli than on the perceived audio quality of of audio-only stimuli. This result seems to be in agreement with previous studies [155] where participants rated echo distortions as imperceptible during video calls, i.e., in the presence of a visual component.

## 4.10 Subjective Experiment 3 (audiovisual)

The main goal of this work is to study the impact that combinations of audio and visual degradations have on the perceived quality of audio-visual signals. With this goal, we performed a psycho-physical experiment to estimate the overall quality of audio-visual sequences containing combinations of audio-only and video-only degradations. We used an immersive experimental methodology [72] to reduce user fatigue, produce a more realistic scenario and, as a consequence, obtain robust quality scores. Considering the limited number of databases that contain audio-visual content with realistic degradations and the associated quality scores, the second objective of this work is to build a large audio-visual database and make this database available for the researcher community. This section describes the motivation of this experiment, it lists the test conditions employed to build the third dataset of this work and performs an analysis of the collected data.

### 4.10.1 Test Conditions

A large stimuli pool was built by processing the original dataset. To generate the test stimuli pool, we introduced audio and video distortions in the audio and video components, respectively, of the original sequences. The video distortions were Bitrate compression, Packet-Loss, and Frame-Freezing. The video stimuli was compressed using H.264 and H.265 video codecs, with varying bitrates. With respect to Packet Loss and Frame-Freezing distortions, since these types of distortions do not occur simultaneously, the videos either contained one or another type of distortion. The Packet-loss distortions were generated by dropping packets from the bitstream at different rates (PLR), while

the Frame freezing distortions were generated by inserting pauses with different lengths. The test conditions were organized to produce a set of 16 Hypothetical Reference Circuits (HRCs). Table 4.11 shows the parameters and types of degradations of each HRC.

With respect to the audio component of the test stimuli, four (4) common streaming audio degradation types were introduced: Background noise, Chop, Clip, and Echo. These types of degradations, along with the insertion procedure, were inspired by the TCD-VoIP dataset [54]. The TCD-VoIP dataset includes some common degradations encountered in a voice over IP transmission. Degradations are considered as "platform-independent" as they are not influenced by the codec, hardware, or network in use. For this experiment, a sample of the test conditions used by the TCD-VoIP dataset was selected and inserted to the audio component of the original sequences. For each type of distortion (noise, chop, clip, and echo), two test conditions were selected and distributed along the 16 HRC arrangements. Additionally, 4 test conditions (ANC) were included as anchors. Table 4.11 shows the details of the HRCs and their corresponding parameters.

Altogether, 40 source stimuli were processed at 20 different test conditions (including 4 anchor conditions). This resulted in 800 different audio-visual sequences with different audio and video distortions. It is important to mention that, for each test session, the participant was presented with only 40 test stimuli of the 800 test sequences, as recommended by the immersive method.

### 4.10.2 Experimental Results

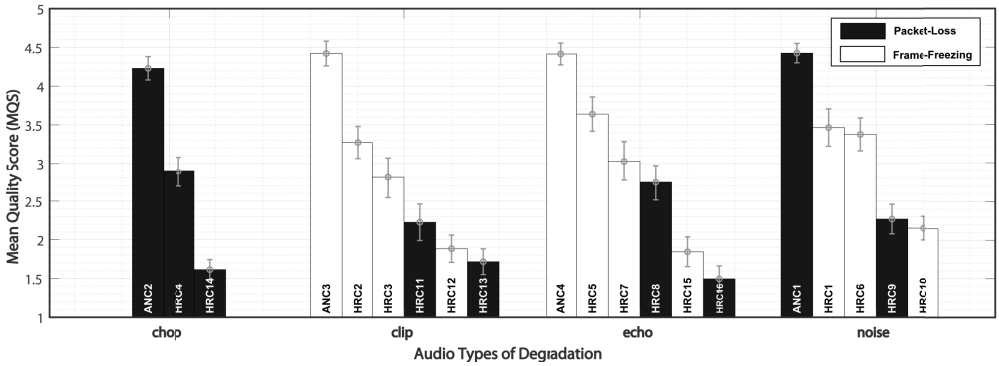
Results of the experiment are presented next. The participant's opinion with respect to the audio-visual quality of the sequences in Experiment 3 is organized in terms of the HRCs considering all audio and video degradations. Additionally, some objective quality metrics are used to compare the results and analyze the interaction between audio and video predicted quality.

#### Results

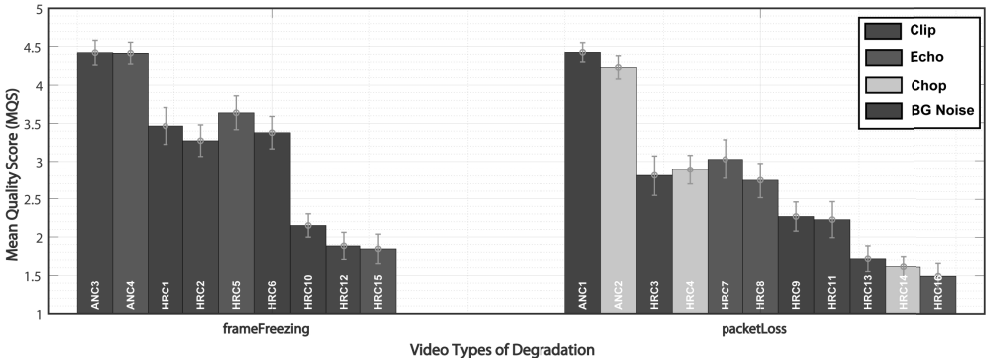
Figure 4.18 presents the MQS values collected from the subjective experiment. In Figure 4.18 (a) the MQS values are grouped according to the audio distortions (chop, clip, echo, and noise), meanwhile in Figure 4.18 (b) the values are grouped according to the video degradations (packet-loss and frame-freezing). It can be observed in this figure that most HRCs obtained quality scores equal or below 3.5, while the anchors sequences (ANC) obtained quality scores well above 4. Considering the different types of audio degradations, Clip degradations obtained slightly lower quality scores on average, while Echo test condition HRC16 received the lowest quality score. Additionally, by observing the gaps

Table 4.11: Coding parameters and types of degradations of the audio and video component of each HRC of the dataset.

| | Audio Component | | | | Video Component | | | |
|---|---|---|---|---|---|---|---|---|
| | **Noise** | **Chop** | **Clip** | **Echo** | **Video Codec** | **Bitrate (kbps)** | **PacketLoss** | **Freezing** |
| **HRC** | **Type, SNR (dB)** | **Period (s), Rate (chop/s), Mode** | **Multiplier** | **Alpha (%), Delay (ms), Feedback (%)** | | | **PLR** | **Pauses, Length (s)** |
| HRC1 | car, 15 | - | - | - | H.264 | 16,000 | - | 1, 2 |
| HRC2 | - | - | 11 | - | H.264 | 16,000 | - | 1, 2 |
| HRC3 | - | - | 11 | - | H.265 | 8,000 | 0.01 | - |
| HRC4 | - | 0.02, 2, zeros | - | - | H.265 | 80,00 | 0.01 | - |
| HRC5 | - | - | - | 0.3, 100, 0 | H.264 | 16,000 | - | 1, 2 |
| HRC6 | office, 10 | - | - | - | H.264 | 16,000 | - | 1, 2 |
| HRC7 | - | - | - | 0.3, 100, 0 | H.265 | 8,000 | 0.01 | - |
| HRC8 | - | - | - | 0.3, 100, 0 | H.264 | 2,000 | 0.05 | - |
| HRC9 | office, 10 | - | - | - | H.264 | 2,000 | 0.05 | - |
| HRC10 | office, 10 | - | - | - | H.264 | 800 | - | 3, 7 |
| HRC11 | - | - | 25 | - | H.264 | 2,000 | 0.05 | - |
| HRC12 | - | - | 25 | - | H.264 | 800 | - | 3, 7 |
| HRC13 | - | - | 25 | - | H.265 | 400 | 0.08 | - |
| HRC14 | - | 0.02, 5, zeros | - | - | H.265 | 400 | 0.08 | - |
| HRC15 | - | - | - | 0.3, 180, 0.8 | H.264 | 800 | - | 3, 7 |
| HRC16 | - | - | - | 0.3, 182, 0.8 | H.265 | 400 | 0.08 | - |
| ANC1 | - | - | - | - | H.264 | 64,000 | - | - |
| ANC2 | - | - | - | - | H.265 | 32,000 | - | - |
| ANC3 | - | - | - | - | H.264 | 64,000 | - | - |
| ANC4 | - | - | - | - | H.265 | 32,000 | - | - |




(a) HRCs grouped by audio degradations.




(b) HRCs grouped by video degradations.

Figure 4.18: Mean Quality Score (MQS) for the different combinations of audio and video degradations (Table 1 describes each HRC).

between the distortion levels for Clip and Echo distortions, we noticed that the differences between neighboring HRCs were roughly constant, while the differences between neighboring HRCs for Noise and Chop seemed more irregular. This might suggest that Noise and Chop degradations were more sensitive to variations, i.e., varying the distortion level for these distortions had a higher impact on the perceived quality.

In Figure 4.18 (b), where MQS scores were organized according to the different types of video degradations, we notice that there is a clear difference between the MQS values obtained for the Packet-loss and Frame-freezing distortions. On average, Frame-freezing distortions seemed to have a lower impact on the perceived quality than Packet-loss distortions. However, by observing the gaps between both types of distortions, variations of Frame-freezing distortion levels seemed to have a heavier impact on the perceived quality. In other words, varying the levels of distortion for Frame-freezing produced a more pronounced drop of quality, when compared to a variation in Packet-loss distortion.

For the case of audio degradations, no particular degradation was identified as having a determinant effect on the perceived quality. As already mentioned, for the case of video degradations, Packet-loss had a stronger influence on the perceived audio-visual

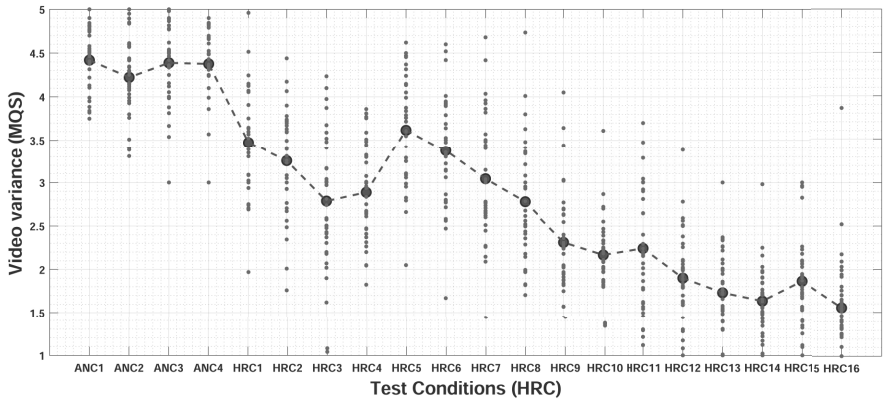



Figure 4.19: Mean Quality Score (MQS), and its respective spread of scores, for the different Hypothetical Reference Circuit (HRC) degradations.

quality. Therefore, in terms of combined degradations, audio degradations combined with Packet-loss had a stronger impact on the overall audio-visual quality.

Figure 4.16 presents the MQS values obtained for each of the HRCSs, along with the single user scores for experiment 3. It can be observed that for more 'degraded' HRC (see Table 4.11), the results are more consistent, i.e., the spread of points is smaller. But, for HRCs that received a MQS value around the center of the scale, the scores provided by participants varied more, resulting in a larger standard deviations around the average value.

### Objective Quality Comparison

We compare the subjective scores with the objective results gathered from one audio and one video quality metrics. Naturally, the subjective scores correspond to the overall audio-visual quality, while the quality scores predicted by the objective metrics represent the quality of a particular component (audio or video). Also, it is worth pointing out that the subjective scores are distributed on a five-point rating scale (ACR), while the scores predicted by the objective metrics do not have the same range, which might lead to scale calibration bias. Despite these issues, the comparison between subjective and objective scores can provide interesting insights concerning the predicted quality and their interaction with the overall audio-visual perceived quality.

The DIIVINE quality metric [108] was selected to predict the quality of the video component of the stimuli. The DIIVINE metric was originally developed as an image quality assessment metric, for this work, a video implementation was used by averaging the quality predictions for every frame in the video. Figure 4.20 depicts the scatter-plots of the subjective scores versus the corresponding DIIVINE scores, organized according to the types of degradation. In general terms, and independent if it is an audio or a

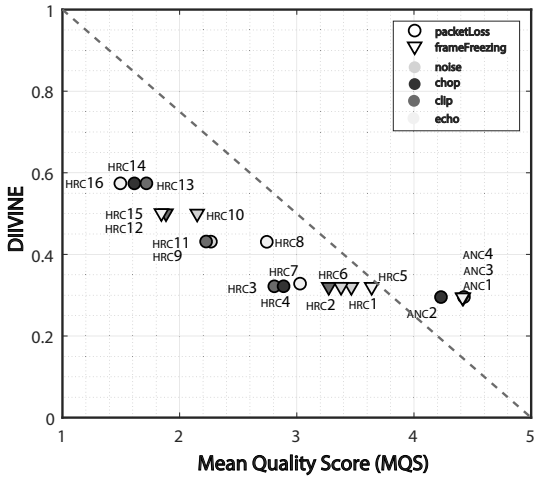



Figure 4.20: Scatter plot of audio-visual subjective scores (MQS) versus video objective scores (produced by DIIVINE).

video degradation, the scatter-plots presented a moderate negative correlation between the subjective audio-visual (MQS) and the DIIVINE scores. It seems that the DIIVINE metric tended to overestimate the video quality of sequences, since most points fall below the red line in the graph (this being interpreted as better quality). While MQS values occupied most of the rating scale (1 to 5), DIIVINE scores were concentrated on the middle of their scale (0 to 1). Despite this characteristic, DIIVINE scores varied along the MQS values, showing a good consistency.

Figure 4.20 shows that sequences affected by Packet-loss degradations (HRCs 13, 14, and 16) resulted in a lower quality, according to the DIIVINE metric. The same graph suggests that sequences with a Frame-freezing type of degradation (HRCs 1, 2, 5, and 6) were less affected in terms of quality. Naturally, regarding the audio distortions, no particular behavior was observed in terms of a higher or lower quality for a specific audio degradation. However, it can be observed that video degradations tend to group around similar conditions. This tendency is only broken for two cases that correspond to Noise and Chop audio degradations (HRCs 10 and 8), which suggests an influence of audio distortions on the perceived audio-visual quality.

VISQOLAudio was chosen as the audio quality metric [40]. Figure 4.21 depicts the scatter-plots of the subjective audio-visual quality scores (MQS) versus the VISQOLAudio scores, organized according to the audio and video types of degradation. In general terms, and considering that this comparison is made between audio and audio-visual scores, no particular pattern was observed. VISQOLAudio also seemed to overestimate the quality for most conditions (most marks fall above the red line), which is expected since only the audio component is being measured.

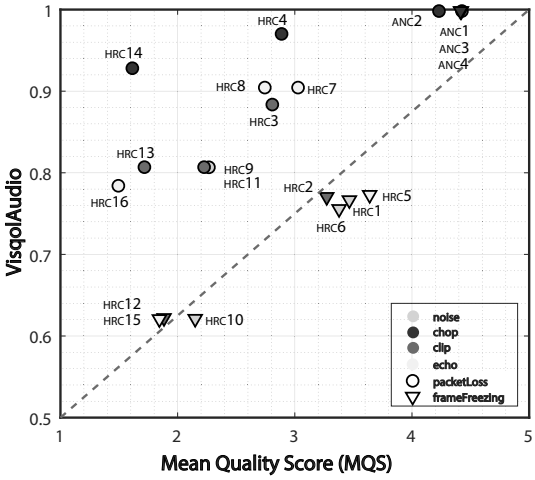



Figure 4.21: Scatter plot of audio-visual subjective scores (MQS) versus audio objective scores (produced by VISQOLAudio).

In Figure 4.21 it can be observed a clear difference between sequences affected by Frame-freezing and Packet-loss distortions. Again, similarly, video conditions tended to group around each other, but not as 'strongly' as it was seen in Figure 4.20. Regarding the type of audio degradations, Figure 4.21 shows that Chop sequences got higher quality scores.

Finally, both VISQOLAudio and DIIVINE scores were compared. Figure 4.22 depicts a scatter-plot of these scores, organized by the types of audio and video degradations. The graph shows a disperse negative relationship between both sets of scores. It can be observed that scores remained spread in the middle of the rating scale. It can be noticed that frame-freezing conditions (HRCs 10, 12, and 15) presented lower audio and video quality predictions.

## 4.11 General Discussion and Conclusions

In this Section, results from all three subjective experiments are compared. Equivalent test conditions are grouped in order to verify their results among all three experiments. The main objective of this section is to analyze the impact of the perceived quality by observing the results from different audio and video test conditions among all three experiments.

For this particular section, the labels assigned to the Hypothetical Reference Circuits (HRC) for experiments 1, 2 and 3, were redefined. This was done with the objective of comparing the HRCs from different experiments. Table 4.12 presents the parameters for the video component along with the video test condition labels. For the purpose

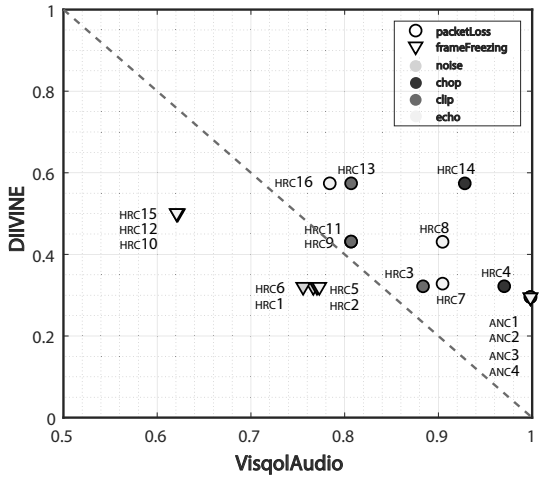



Figure 4.22: Scatter plot of audio objective scores (prduced by VISQOLAudio) versus video objective scores (produced by DIIVINE).

of comparing different databases, we use the term Video Test Condition (V-TC) that replaces the previously used HRC.

Similarly, Table 4.13 presents the parameters for the audio component along with the audio test condition labels. As in the video component, the term Audio Test condition replaced the previous term HRC.

### 4.11.1 Audio and Video Distortion Impact

Figure 4.23 compares the MQS and MCS responses collected from experiments 1 and 3 for the video test conditions. They were organized according to the type of distortion (packet loss and frame freezing). Figure 4.23 (a) depicts the results for the packet loss type of distortion. It can be observed that for the same video test conditions responses were lower

Table 4.12: Parameter details for the video test conditions.

| **Packet Loss** | | | |
|---|---|---|---|
| **Video Test Condition** | **Codec** | **Bitrate (kb/s)** | **PLR** |
| V-TC1 | H.264 | 500 | 10% |
| V-TC2 | H.265 | 400 | 8% |
| V-TC3 | H.264 | 2000 | 5% |
| V-TC4 | H.265 | 1000 | 3% |
| V-TC5 | H.265 | 8000 | 1% |
| V-TC0 | H.264 | 64000 | - |
| **Frame Freezing** | | | |
| **Video Test Condition** | **Codec** | **Bitrate (kb/s)** | **Freezing** |
| V-TC6 | H.265 | 200 | S5 |
| V-TC7 | H.264 | 800 | S4 |
| V-TC8 | H.265 | 1000 | S3 |
| V-TC9 | H.264 | 2000 | S2 |
| V-TC10 | H.264 | 16000 | S1 |
| V-TC0 | H.265 | 32000 | - |

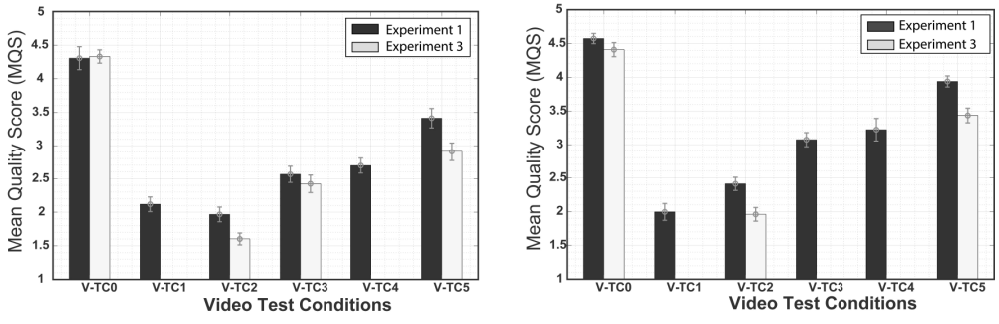
Table 4.13: Parameter details for the audio test conditions..

| **Background Noise** | | | |
|---|---|---|---|
| **Audio Test Condition** | **Noise** | **SNR (dB)** | |
| A-TC1 | car | 15 | |
| A-TC2 | babble | 10 | |
| A-TC3 | office | 10 | |
| A-TC4 | road | 5 | |
| A-TC0 | - | - | |
| **Chop** | | | |
| **Audio Test Condition** | **Period (s)** | **Rate (chops/s)** | **Mode** |
| A-TC5 | 0.02 | 1 | previous |
| A-TC6 | 0.02 | 2 | zeros |
| A-TC7 | 0.04 | 2 | previous |
| A-TC8 | 0.02 | 5 | zeros |
| A-TC0 | - | - | - |
| **Clipping** | | | |
| **Audio Test Condition** | | **Multiplier** | |
| A-TC9 | | 11 | |
| A-TC10 | | 15 | |
| A-TC11 | | 25 | |
| A-TC12 | | 55 | |
| A-TC0 | | - | |
| **Echo** | | | |
| **Audio Test Condition** | **Alpha (%)** | **Delay (ms)** | **Feedback (%)** |
| A-TC13 | 0.5 | 25 | 0 |
| A-TC14 | 0.3 | 100 | 0 |
| A-TC15 | 0.175 | 140 | 0.8 |
| A-TC16 | 0.3 | 180 | 0.8 |
| A-TC0 | - | - | - |

when the audio component was distorted (experiment 3). This impact is more pronounced for video test conditions V-TC2 and V-TC5. As for the frame freezing distortion, Figure 4.23 (b) presents a similar behavior. For the same test conditions, quality responses with audio distortions presented lower quality scores. These graphs confirm that there is a clear impact of the audio component in terms of the perceived quality. As observed, audio distortion affected both types of video distortion (packet loss and frame freezing) in the same manner. Moreover, given that the analysis is made based on a test condition configuration, it can be implied that this impact affected the overall quality regardless of the content. However, this last assumption should be reviewed with a larger number of test conditions and a deeper analysis of the content.

Regarding the participant's personal opinion about the content (MCS), Figures 4.23 (c) and (d) presented a similar behaviour compared to the quality scores (MQS). However, they are not statistically significant so no real conclusion can be made about these results.

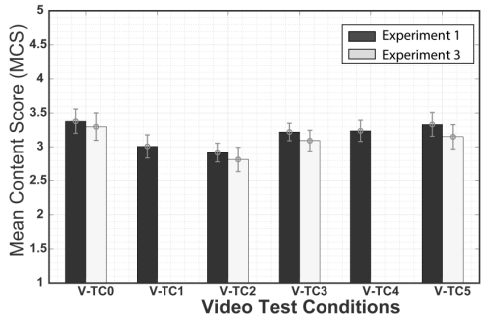
Figure 4.24 compares the MQS responses gathered from experiments 2 and 3 for the audio test conditions. As in the video analysis, results were organized according to the audio type of distortion: background noise, chop, clip and echo. Figure 4.24 (a) presents the results for the background noise type of distortion. Aside from the A-TC0 test condition, no particular difference can be spotted between results from experiment 2



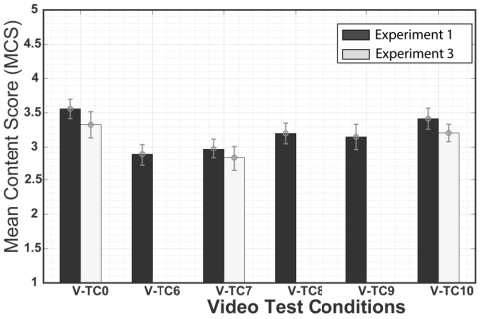

(a) MQS - Packet Loss




(b) MQS - Frame Freezing




(c) MCS - Packet Loss

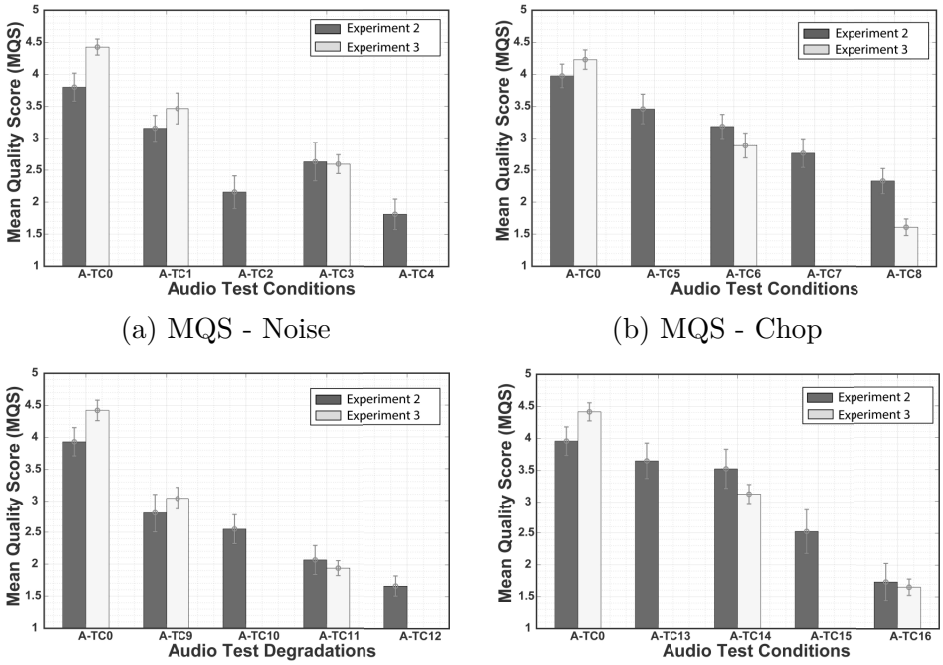



(d) MCS - Frame Freezing

Figure 4.23: *MQS* and *MCS* responses collected from experiments 1 and 3 for the video test conditions.

and 3. This result might suggest that the background noise distortion had an equivalent impact on the audiovisual quality when compared to the test conditions on experiment 3 that included both audio and video distortions. Similarly, Figure 4.24 (b) presents the results for the chop distortion. Results for the A-TC8 test condition shows a difference between results when the video component has been distorted. This might suggest that a chop type of distortion by itself doesn't have a strong impact on the overall quality. Figure 4.24 (c) depicts the results for the clipping type of distortion. Results from this figure are similar to the one saw on the background noise scenario. This might suggest that a clipping distortion levels the perceived quality of a test condition where the video and audio components have been distorted. Finally, Figure 4.24 (d) presents the results for the echo type of distortion. No particular behaviour can be spotted from this figure. Test condition A-TC14 suggests that the video distortion had a higher impact on the perceived quality. Meanwhile, test condition A-TC16 suggests that the echo distortion had an equivalent impact on the overall quality compared to the audio and video distortions combined. Overall, no particular conclusion can be made based on these results. Results showed that some types of audio distortion have a greater impact compared to others. More particularly, background noise and clipping levelled the quality impact that audio and video distortions had. Further analysis is needed in order to conclude this assumption.

A similar comparison is depicted in Figure 4.25 considering the MCS for experiments




(a) MQS - Noise

(b) MQS - Chop

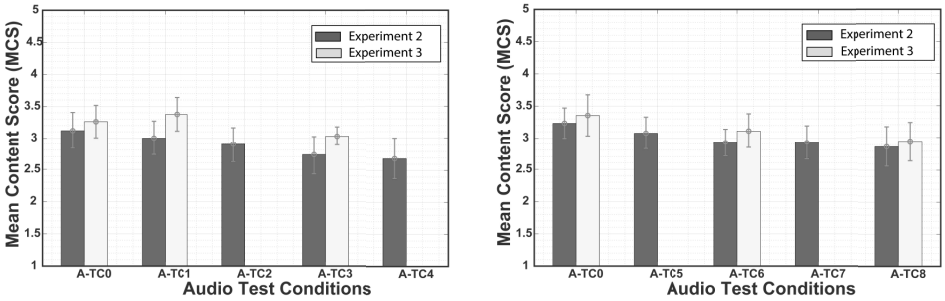
(c) MQS - Clip

(d) MQS - Echo

Figure 4.24: *MQS* responses collected from experiments 2 and 3 for the audio test conditions.

2 and 3. Similar behaviour was observed, however, since the differences are not statically significant, no conclusion can be made regarding these results. Additionally, Figure 4.26 presents scatter plots of the results from experiment 1, 2, and 3. From Figure 4.26 (a), we notice that there is a positive correlation between both sets of results. As it was observed in the previous analysis, scores from Experiment 1 (video only degradation) had higher quality responses when compared to the scores from Experiment 3 (audio and video distortions), which can be noticed by observing the values above the red line. This behavior was observed for both types of video distortions: packet loss and frame freezing.

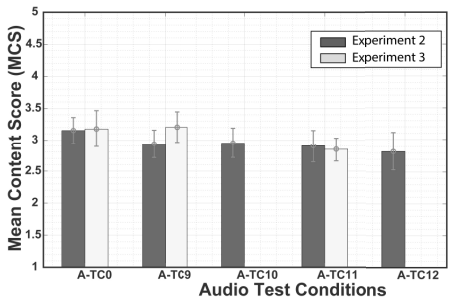
Figure 4.26 (b) depicts a scatter plot comparing scores from Experiment 2 and Experiment 3. From this figure, it can be noticed a subtle positive correlation between both score sets. As it was observed in the previous analysis, most of the test conditions obtained higher quality scores for the case of Experiment 2 (audio only distortion) as seen in the values above the red line. However, some test conditions (A-TC 9 and A-TC1) presented lower scores equivalent to the results obtained in experiment 3.



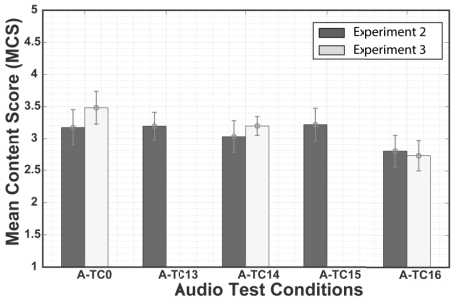

(a) MCS - Noise




(b) MCS - Chop

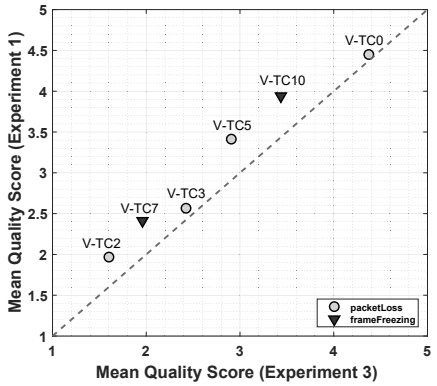



(c) MCS - Clip

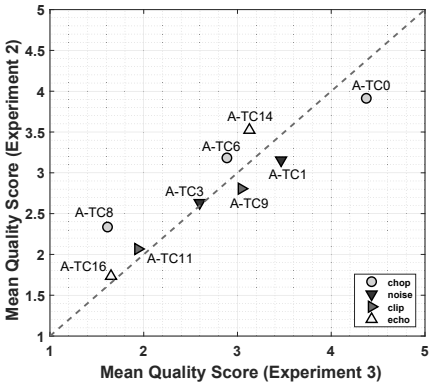



(d) MCS - Echo

Figure 4.25: $MCS$ responses collected from experiments 2 and 3 for the audio test conditions.




(a)




(b)

Figure 4.26: $MQS$ results from experiment 1 and 3.

# Chapter 5

# Deep Autoencoder model for audio-visual quality assessment

This Chapter presents the proposed No-Reference audiovisual quality model for objective evaluation of audiovisual quality. Along with the NR audiovisual quality model, two additional quality models are presented in this work: one NR audio quality model and one NR video quality model. The novelty of the proposed metrics lies in using an autoencoder approach to extract low-dimensional features from the audio and video components of the signal and then finding a mapping between those features and subjective scores using a classification function.

Figure 5.1 presents a block diagram of the proposed approach. The diagram depicts both training and testing phases of the three models, which uses a set of audiovisual sequences and the corresponding subjective quality scores gathered in psychophysical experiments. Sets of audio and video features, which have relevant audio and visual characteristics associated, are extracted from these signals in the first diagram block. Then, the set of features are used to train a network model in the second diagram block, which consists of two-layer blocks, namely an autoencoder and the classification layers. The output of this training phase is an Autoencoder Network that is able to predict the quality of a sequence. It is important to emphasize that the training of the audio model is made using only the signal audio features, while the video model is trained using the visual features. For the present work, the video set of features is a set of natural scene statistics (NSS) used in several image and video quality metrics [108, 107], plus the spatial and temporal information associated to the video sequence. As for the audio features, a spectrogram (2-D representation) is used as a feature source to describe the audio component of the signal. These sets of features, both audio and video, represents the input of the entire quality assessment model.

One important issue to consider when using ML paradigms is the training of the

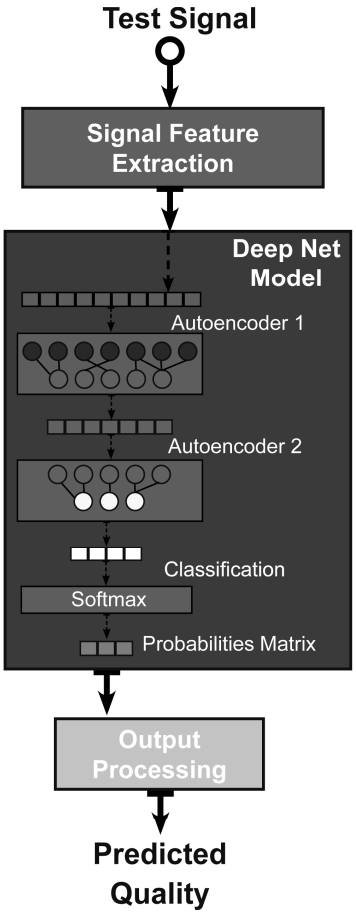



Figure 5.1: Simplified block diagram of the **Autoencoder Network** approach.

system and its prediction accuracy. A good composition of the training stimuli requires the construction of a sufficiently large and representative set of the audiovisual stimuli with a broad set of distortions. For such purpose, as described in Chapter 4, three large datasets were constructed and their corresponding quality scores were gathered on a set of subjective quality experiments. Altogether, the three datasets sum a total of two-thousand-three-hundred and twenty (2320) different audiovisual sequences which are used for the training and testing of the proposed models.

The proposed set of Auteoncoder Network models is tested against several FR and NR video, audio and speech quality metrics. Considering the reduced number of audiovisual quality metrics in the literature, this model constitutes an important contribution that serves as a reference to the development of new quality assessment methods.

The remainder of this Chapter is organized as follows. In Section 5.1, the visual and audio feature extraction procedures are presented. In Section 5.2, the Autoencoder Network approach is described along with the overall structure of the proposed quality metrics. Then, in Section 5.3, the quality metrics performance is presented. Finally, Section 5.4 presents the conclusions of this chapter.

## 5.1 Feature Extraction

### 5.1.1 Visual Features

In order to obtain a set of visual features that are able to describe the visual characteristics (and distortions) of the video sequence under analysis, the present work relied on two commonly used properties: 1) a set of natural scene statistics, and 2) spatial and temporal information. Altogether, they formed the set of features used as input for the video quality model. Next, some details about the extraction and the organization of the features are presented.

- *Natural Scene Statistics Features (f1 – f88)*

  Natural scene statistics are widely used to describe regularities (or irregularities) in a still image. Its usage has been extended to videos and they have become the base of several image and video quality metrics [108, 107]. Given its distortion-agnostic nature, these type of features can be employed to describe several types of visual distortions, including video coding, packet loss, and frame freezing [108, 105].

  In the present work, we used the feature extract function from the Diivine image quality metric implementation [108] to extract a total of eighty-eight (88) features. A detailed description of all 88 features can be found in Chapter 3. For each frame of the video under analysis, a set of 88 features is extracted. This resulted in an 88-by-$n$ matrix ($n$ being the number of video frames), that represents the NSS set of features.

- *Spatial and Temporal Features (f89 – f90)*

  In order to capture the spatial and temporal characteristics of the video sequence, we used the algorithm presented by Ostaszewska and Kloda [135] to compute the spatial and temporal information. These values describe the video behavior along the time and characterize some important visual distortions, more specifically frame freezing distortion. Again, spatial and temporal values are computed for each frame of the video ($n$) which results in a 2-by-$n$ matrix, that represents the spatial and temporal features.

Next, both sets of features are merged to form the *visual set of features* of a single video sequence, represented by a 90-by-$n$ matrix. Figure 5.2 depicts the visual set of features for a single video sequence.

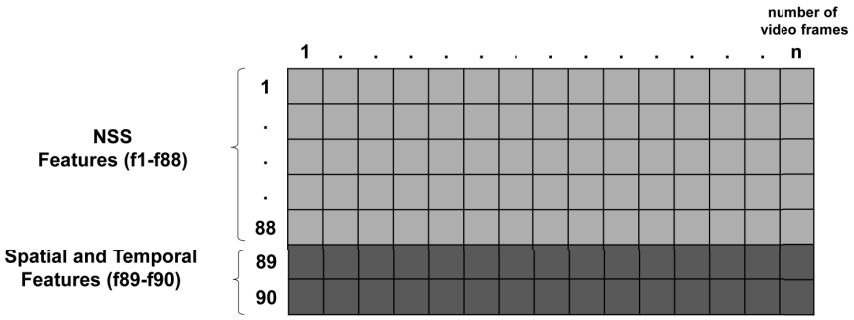



Figure 5.2: Visual Set of Features composed of NSS features and Spatial and Temporal features.

### 5.1.2 Audio Features

With the objective of getting a set of features capable of describing and characterizing audio distortions, a spectrogram representation is used as the feature source for this model. It is basically a time-frequency color intensity representation of the audio activity. Spectrograms have been used on several studies related to speech intelligibility and noise suppression with good results [25, 39, 55]. Some details about the spectrogram computing are presented next.

- *Spectrograms Features (f1 – f25)*

  The use of spectrograms as a descriptive source of the audio signal was inspired by the Visqol speech and audio metrics [39]. The spectrogram extraction function from the Visqol implementation is used to obtain a 25-by-$m$ matrix, where 25 represent the number of frequency bands and $m$ is the number of audio samples of the signal. Each column of the spectrogram provides a set of 25 descriptive values corresponding to each sample of the audio signal. Figure 5.3 depicts a sample of the spectrogram matrix extracted from the audio signal.

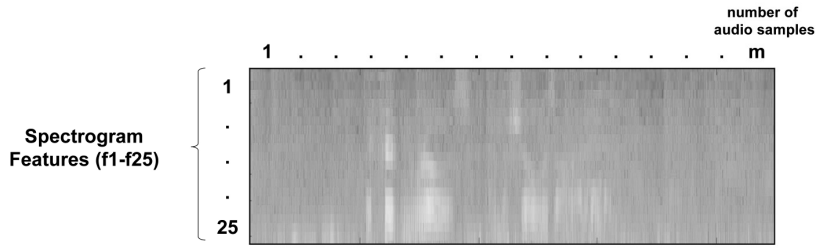



Figure 5.3: Sample of the Spectrogram Matrix extracted from the audio signal.

### 5.1.3 Audiovisual Features

In order to describe the characteristics and the distortions associated to the audio and video components of an audiovisual sequence, both audio and visual descriptive features used in the previously are merged to build one large set of audiovisual features. That is, the visual set of features, composed of the NSS features and the spatial and temporal features, and the audio set of features, represented by the spectrogram of the audio signal are grouped to produce an audiovisual set of descriptive features. Some details about the combination of these sets of features are presented next.

- *Audiovisual Features (f1 – f115)*

  In order to build the audiovisual set of features, the same extraction procedure described previously is followed. Once the visual features (90-by-$n$ matrix) and the audio features (25-by-$m$ matrix) are obtained, they are merged together to compute a total of 115 descriptive features. However, given that the number of video frames ($n$) and the number of audio samples ($m$) are not the same, a scaling process is required to perfectly match these two sets before merging them.

For the present work, the selected approach to uniformize the length of the two matrices is to replicate the values of the matrix that has the shorter length so it matches the other matrix. Since the number of frame videos ($n$) is smaller compared to the number of audio samples ($m$), values of the visual feature set are replicated to match the audio feature set. Figure 5.4 presents a graphic explanation of this scaling procedure. Once the length of both sets matches, they are merged to form a 115-by-$m$ matrix, denoted as the audiovisual features set, where 115 is the sum of the 90 visual features and the 25 audio features.

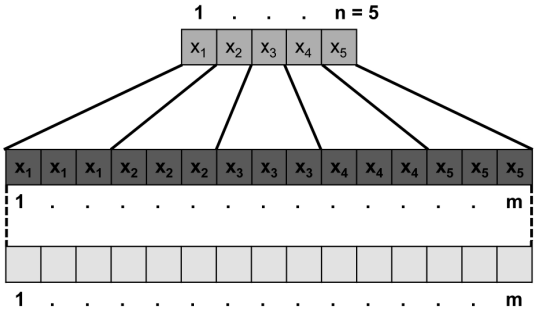



Figure 5.4: Simplified illustration presenting the scaling procedure to match the visual and audio feature matrices.

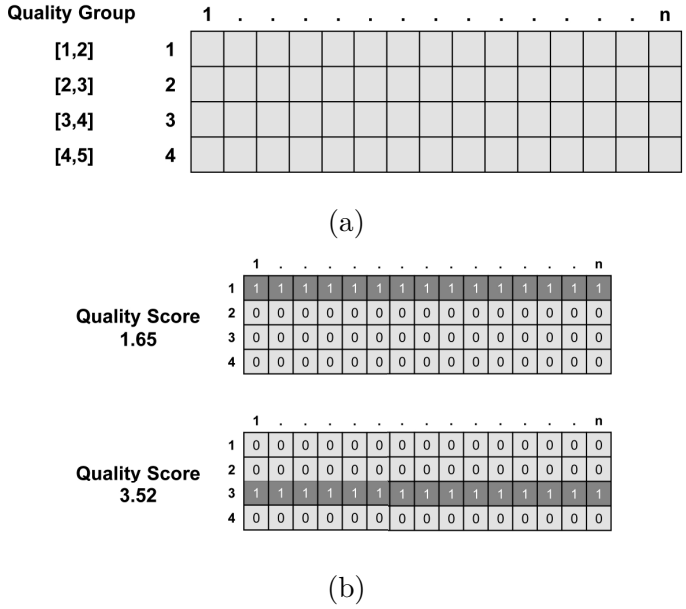



Figure 5.5: **(a)** Target Quality Group Matrix representing the 4 quality group intervals. **(b)** Sequence with subjective score of 1.65 is assigned the quality group 1, interval [1,2]. Sequence with subjective score of 3.52 is assigned the quality group 3, interval [3,4].

### 5.1.4 Training Input

Additionally, an extra target set is built using the subjective score associated with the signal under analysis. This set represents the target quality score that is going to be used during the training of the model, more specifically, during the classification phase. The target set is represented by a zeros and ones 4-by-$n$ matrix, where 4 represents the number of quality groups and $n$ is the number of frames of the video sequence. For the case of the training of the audio and audio-visual models, the value of $n$ is replaced by $m$, which corresponds to the number of audio samples of the signal. There are 4 quality groups, which denote the ranges of scores presented in an ACR quality scale. The target set is built by taking the subjective score associated with the video sequence and assign this value to its corresponding quality group. For example, a sequence that has a subjective score of 1.65 is assigned the quality group 1 since the score is in the interval $< 1, 2 >$, while a sequence with a subjective score of 3.52 is assigned to the quality group 3 since the score is in the interval $< 3, 4 >$. Then, the row corresponding to the quality group is set to one and the rest is set to zero. Figure 5.5 depicts some examples of this setup. Considering that each column represents a video frame (or an audio sample), this setup guarantees that each frame (or sample) has only one quality group associated. Later on, during the training of the model, this target set is used to map the corresponding quality group of each frame (sample) in the signal.

Finally, the descriptive feature and target sets of all training signals are concatenated

to build two large global sets, i.e., a global feature set and global target quality set. Figure 5.6 depicts an illustration of both global feature and target sets. The global features set is represented by a 90-by-$N$ matrix, where 90 denotes the number of visual features and $N$ represents the sum of the number of frames of all video sequences. Again, for the case of the training of the audio and audiovisual model, the value of $N$ is replaced by $M$, which denotes the sum of all audio (audio model) and audiovisual (audiovisual model) samples of each signal considered for the training. Meanwhile, the global target set is represented by a 4-by-$N$ matrix, where 4 represents the number of quality target groups ($M$ for the audio and audiovisual models). These two global sets served as input for the training of the model at different stages. The global feature set is passed to the autoencoder, meanwhile, the global target set is used during the classification phase.

## 5.2 Network Model

### 5.2.1 Model Training

Once the global sets are built, the model is trained using these elements as input. The training phase consists basically of two main layers: 1) the autoencoder layer, which receives the global feature set as input, and 2) the classification layer, which receives a low-dimensional set of features and the global target set as input. Finally, the trained models resulting from these two layers are stacked together and re-trained to form the resulting network model. Next, we detail these two layers.

- *Autoencoder Layer*

  This first layer has the objective of training a model to produce a low-dimensional representation of the input features. For this purpose, we train an autoencoder

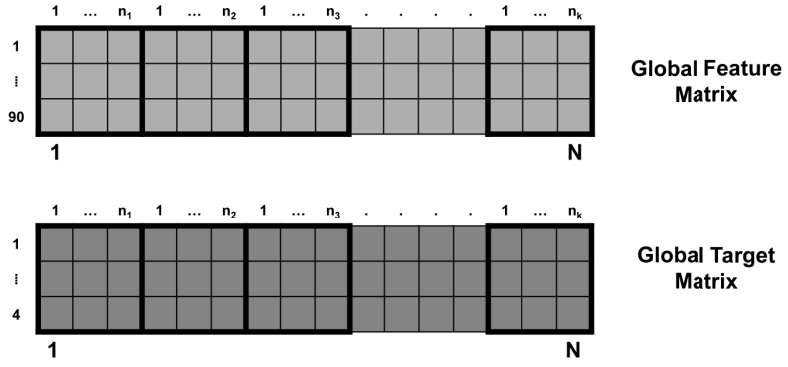



Figure 5.6: Feature and Target matrices concatenation to build the Global Feature Set and Global Target Quality Set.

network, which is formed by two sub-layers (two autoencoders). The ability of the autoencoder to find important properties during the training of the data is exploited and it is expected that this new low-dimensional feature representation is able to characterize the visual and audio distortions of the signal, as the interactions between both audio and video components.

The autoencoder receives the global feature set as input (see Table 5.1). Using this set of features as input the first autoencoder is trained using a different hidden layer size (see Layer size #1 from Table 5.1). This means that the output of this first training is a matrix, denoted as *Features 1*. Along with this new set of features, a trained autoencoder is also available, which is denoted as *Autoencoder 1*. The next autoencoder is trained using as input the *Features 1* set. This autoencoder uses a different hidden layer size(see Layer size #2 from Table 5.1) and the result is a matrix, denoted as *Features 2*. Also, a second trained autoencoder, denoted as *Autoencoder 2*, is produced after this second training stage. Table 5.1 depicts some additional parameters considered for the training of the model.

Overall, the output of this autoencoder layer is composed of: 1) two trained autoencoders (Autoencoder 1 and Autoencoder 2), and 2) two sets of features (Features 1 and Features 2). From this group of elements, only the *Features 2* set are used as input in the following classification layer. As for the rest of elements, they are used during the overall training of the network model. Figure 5.7 depicts a simplified diagram of the autoencoder layer.

- *Classification Layer*

  This layer has the objective to find a mapping between the input set of features and the subjective scores of the corresponding video sequences. In order to obtain this

Table 5.1: Training parameters for the Video, Audio, and Audiovisual Autoencoder Network Models ($N$ sum of number of frames of all videos, $M$ sum of number of all audio samples.

| Layer | Parameters | Video Model | Audio Model | Audiovisual Model |
|---|---|---|---|---|
| **Autoencoder Layer** | **Input** | 90-by-N matrix | 25-by-M matrix | 115-by-M matrix |
| | **Layer size #1** | 50 | 18 | 60 |
| | **Layer size #2** | 20 | 10 | 25 |
| | **Decoder transfer function** | Linear | Linear | Linear |
| | **L2 weigth regularization** | 0.001 | 0.001 | 0.001 |
| | **Sparsity Regularization** | 4 | 4 | 4 |
| | **Sparsity Proportion** | 0.05 | 0.05 | 0.05 |
| **Classification Layer** | **Input** | 20-by-N matrix | 10-by-M matrix | 25-by-M matrix |
| | | 4-by-N matrix | 4-by-M matrix | 4-by-M matrix |
| | **Loss Function** | Cross Entropy | Cross Entropy | Cross Entropy |
| **Additional Info** | **Training Set** | Experiment 1 | Experiment 2 | Experiment 3 |
| | **# sequences** | 720 | 800 | 800 |
| | **Method** | 10-fold CV | 10-fold CV | 10-fold CV |

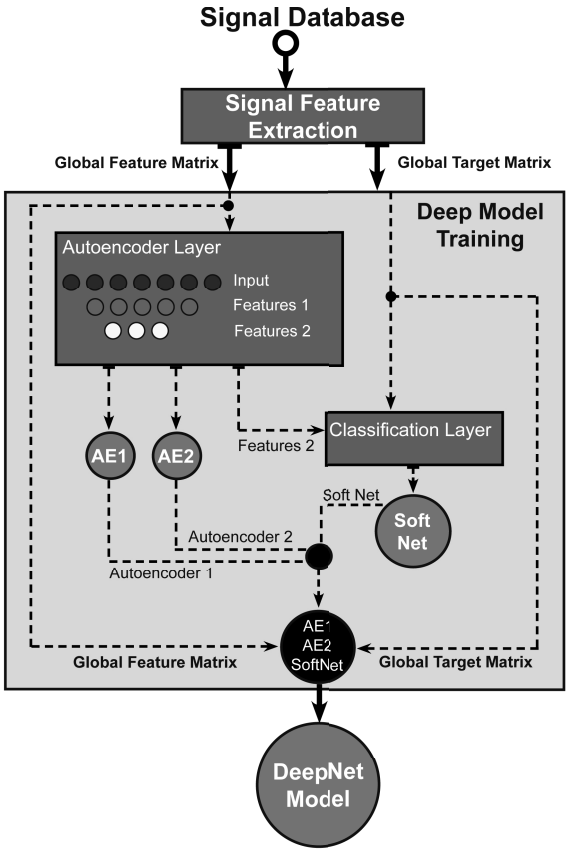



Figure 5.7: Detailed block diagram of the training phase of the **Autoencoder Network** approach.

mapping, a softmax function for classification is used to discover the quality group corresponding to the set of features.

The classification layer receives the *Features 2* set and the target set as input (see Table 5.1). The resulting classification function, denoted as Soft Net, is trained to map a set of features onto a probabilities matrix. The mapped matrix has values between 0 and 1 which represent the probability of a single video frame belonging to a quality group (highest probability value will be the target quality group). Figure 5.7 depicts a diagram of this classification layer.

Once the autoencoders (Autoencoder 1 and Autoencoder 2) and the classification function (Soft Net) are trained, they are stacked to form the model network, denoted as Autoencoder Network Model. Then, the autoencoder network is trained using the global feature and the global target sets. Figure 5.7 presents a diagram of this final training phase.

The resulting model is capable of predicting the corresponding quality group for every frame of the video sequence. An additional output processing phase is required in order to compute the overall video quality of a single video sequence.

## 5.3 Model Performance

In order to test the Autoencoder Network, it is first required the extraction of the descriptive features from the signal sequence under test. The global set of features is passed to the trained autoencoder network. The output is a matrix which contains the probability values in the interval [0, 1] of a frame belonging to a quality group. In order to estimate the overall video quality of the sequence, the probabilities output need to be processed.

Figure 5.8 presents a simplified illustration of the output processing stage. First, the maximum value and its corresponding row index are calculated for each column in the probabilities matrix. Then, a vector is built by adding the index and the max value for each column in the vector. In other words, for each column (representing a video frame or and audio sample) the corresponding quality group index is summed with the corresponding probability value resulting in a quality value in the interval [1, 5]. Finally, the quality scores of all frames of the video, or samples of the audio, are averaged and the overall signal quality score is computed.

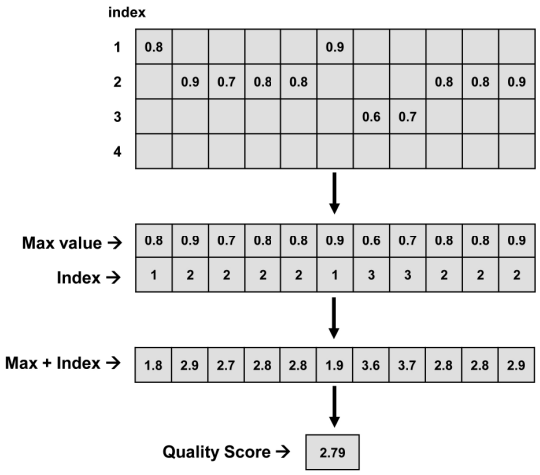



Figure 5.8: Simplified illustration of the output processing stage applied to the results of the **Autoencoder Network** model.

Three quality metrics were obtained based on the trained autoencoder network models: 1) video, 2) audio, and 3) audiovisual. As it was presented before (see Table 5.1), the autoencoder network models were trained and tested using sequences from all three experiments: Experiment 1, Experiment 2 and Experiment 3. A total of 720 (Experiment 1), 800 (Experiment 2), and 800 (Experiment 3) audiovisual sequences are employed along with a 10-fold cross-validation method to test the models results. These results are compared against a set of popular FR and NR video quality metrics from the literature. The FR video quality metrics considered are: SSIM (video adaptation) [92], and PSNR (video adaptation). The NR video metrics considered are: VIIDEO [107], DI-

IVINE (video adaptation) [108], BIQI (video adaptation) [156], NIQE (video adaptation) [157], and BRISQUE (video adaptation) [158].

As for the audio quality metric, results are compared against a set of popular FR and NR audio and speech quality metrics from the literature. The FR audio quality metrics considered are: VisqolAudio [40] and PEAQ [159], additionally, the speech metric Visqol [39] is also considered. Finally, the NR speech quality metric P.563 [114] is considered for this testing phase.

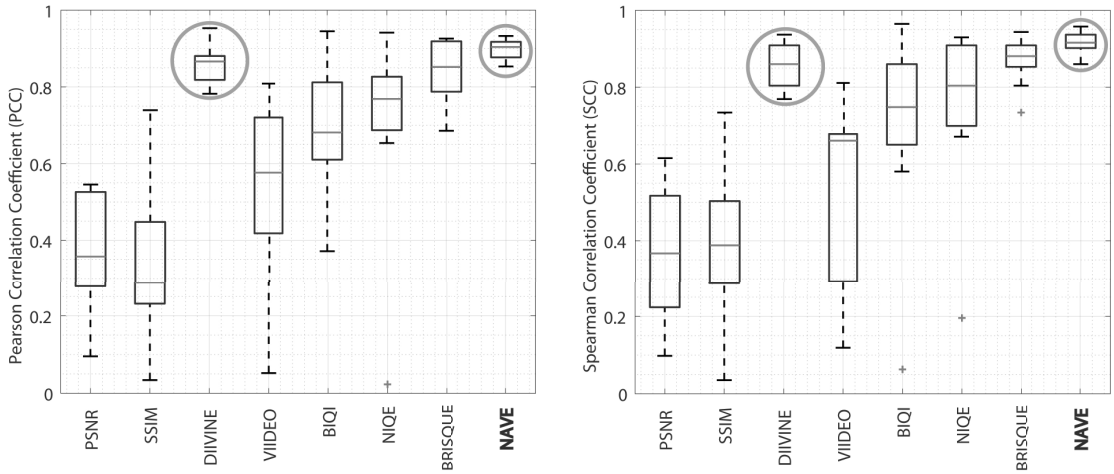
Finally, for the audiovisual metric, the same FR and NR video quality metrics were used for comparison. Similarly, the same set of audio quality metrics were used to verify its performance against the proposed audiovisual model. One last group of audiovisual combination models are considered for comparison: Linear, Minkowski, and Power audiovisual models. These models (introduced in Chapter 3), take as input the results from one video and one audio objective metrics: DIIVINE and P.563.

Table 5.2 presents the Pearson and Spearman correlation coefficients (PCC and SCC), along with the root mean square errors (RMSE) gathered from testing the FR and NR video quality metrics and the first proposed metric: the video autoencoder metric. The results are organized according to the video type of distortion for a better analysis. As can be observed, the proposed model has the best performance in the overall analysis achieving high correlation coefficients at a low error margin. Regarding the packetloss distortion, the proposed model also presents the best performance achieving correlation coefficients above 0.93. As for the frame freezing distortion, both DIIVINE and the proposed model presented the best performance. For a better visualization of the results, Figure 5.9 (a) and (b) shows bar plots of the average PCC and SCC values (over the 10 folds) for all metrics. Notice that, for all the metrics tested, the PCC and SCC results provided by the proposed video metric are the highest and have the smallest variation, which means that their results are very consistent.

Table 5.3 presents the Pearson and Spearman correlation coefficients, along with the root mean square errors gathered from testing the FR and NR audio and speech quality metrics and the proposed audio autoencoder model. The results are organized according to the four audio types of distortion for a better analysis. As can be observed, the proposed model has a fair performance in the overall analysis comparable to the standardized P.563 speech metric. Regarding the type of distortion, the proposed model presents a fair level of prediction for distortions like noise, clip, and echo. For a better visualization of the results, Figure 5.10 (a) and (b) depicts bar plots of the overall PCC and SCC values (over the 10 folds) for all metrics. Besides the high correlation values presented by the proposed metric, it can observed that results presented a small variation on both PCC and SCC coefficients. This shows that the results are very consistent compared to the rest of the

Table 5.2: Pearson and Spearman Correlation Coefficients (PCC and SCC), and Root Mean Square Error (RMSE) gathered from testing the FR and NR video quality metrics on the database from Experiment 1.

| **Type** | **Metric** | **Measure** | **Packet-Loss** | **Frame-Freezing** | **All** |
|---|---|---|---|---|---|
| Full-Reference | PSNR | PCC | 0.8352 | 0.7482 | 0.4508 |
| | | SCC | 0.8857 | 0.7714 | 0.4615 |
| | | RMSE | 8.1694 | 12.7864 | 10.7292 |
| | SSIM | PCC | 0.8886 | 0.2741 | 0.2423 |
| | | SCC | 0.9429 | 0.3714 | 0.2378 |
| | | RMSE | 2.8559 | 2.3673 | 2.6230 |
| No-Reference | DIIVINE | PCC | -0.9173 | -0.9101 | -0.8835 |
| | | SCC | -0.9429 | -0.8857 | -0.8951 |
| | | RMSE | 2.5274 | 2.8885 | 2.7139 |
| | VIIDEO | PCC | -0.6728 | -0.5962 | -0.6393 |
| | | SCC | -0.7714 | -0.4286 | -0.6923 |
| | | RMSE | 2.3137 | 2.6892 | 2.5084 |
| | BIQI | PCC | -0.8490 | -0.8597 | -0.8568 |
| | | SCC | -0.9429 | -0.8857 | -0.9161 |
| | | RMSE | 33.8984 | 31.3417 | 32.6451 |
| | NIQE | PCC | -0.7382 | -0.9204 | -0.8485 |
| | | SCC | -0.7714 | -0.8857 | -0.8811 |
| | | RMSE | 1.9239 | 1.7098 | 1.8200 |
| | BRISQUE | PCC | -0.8135 | -0.9254 | -0.8800 |
| | | SCC | -0.7714 | -0.9429 | -0.8741 |
| | | RMSE | 44.8406 | 41.5565 | 43.2298 |
| | **DAE-Video** | **PCC** | **0.9332** | **0.8959** | **0.8966** |
| | | **SCC** | **0.9429** | **0.9143** | **0.9175** |
| | | **RMSE** | **0.4281** | **0.4995** | **0.4681** |




(a) PCC (b) SCC

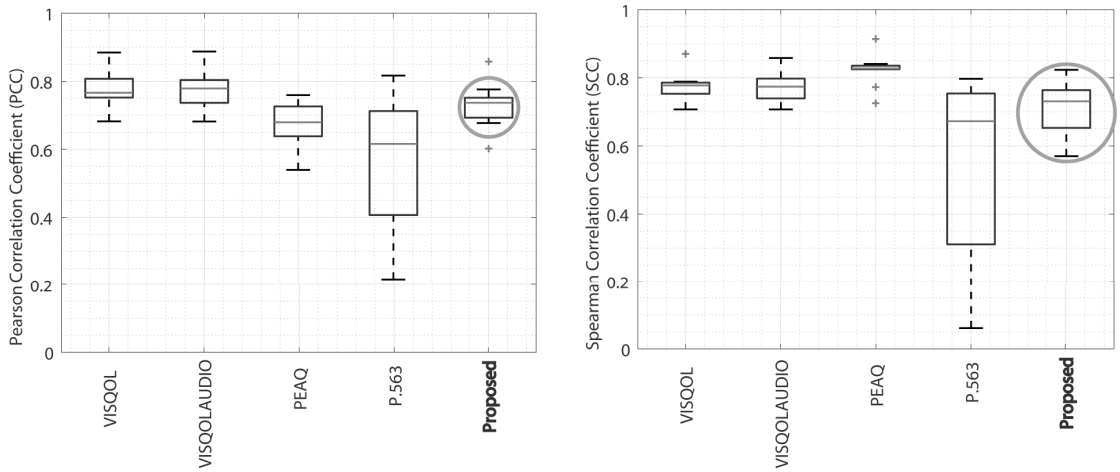
Figure 5.9: Box plot of the Pearson and Spearman Correlation Coeficients (PCC and SCC) gathered from testing the FR and NR video quality metrics on the database from Experiment 1.

literature metrics.

Table 5.4 presents the Pearson and Spearman correlation coefficients, along with the root mean square errors gathered from all video quality metrics under test. Similarly, Table 5.5 presents the same set of results gathered from the audio quality metrics under

Table 5.3: Pearson and Spearman Correlation Coefficients (PCC and SCC), and Root Mean Square Error (RMSE) gathered from testing the FR and NR audio quality metrics on the database from Experiment 2.

| **Type** | **Metric** | **Measure** | **Noise** | **Chop** | **Clip** | **Echo** | **All** |
|---|---|---|---|---|---|---|---|
| Full-Reference | VISQOL | PCC | 0.9851 | 0.9914 | 0.9939 | 0.9082 | 0.8416 |
| | | SCC | 1.0000 | 1.0000 | 1.0000 | 0.9000 | 0.8740 |
| | | RMSE | 2.0110 | 2.2403 | 1.8804 | 2.3340 | 2.1240 |
| | VISQOLAUDIO | PCC | 0.9820 | 0.9928 | 0.9993 | 0.8979 | 0.8541 |
| | | SCC | 1.0000 | 1.0000 | 1.0000 | 0.9000 | 0.8740 |
| | | RMSE | 1.9841 | 2.2387 | 1.8736 | 2.3177 | 2.1113 |
| | PEAQ | PCC | 0.8262 | 0.8841 | 0.8701 | 0.5915 | 0.7689 |
| | | SCC | 0.9000 | 1.0000 | 1.0000 | 1.0000 | 0.9011 |
| | | RMSE | 5.7925 | 5.7104 | 5.7503 | 6.1188 | 5.8452 |
| No-Reference | P.563 | PCC | 0.7626 | 0.8508 | 0.9886 | 0.9253 | 0.7486 |
| | | SCC | 0.8000 | 0.9000 | 1.0000 | 0.6000 | 0.6974 |
| | | RMSE | 0.7987 | 1.1537 | 0.8855 | 1.0952 | 0.9941 |
| | **DAE-Audio** | **PCC** | **0.8291** | **0.3632** | **0.9149** | **0.8711** | **0.7312** |
| | | **SCC** | **0.8200** | **0.2600** | **0.7700** | **0.7300** | **0.7082** |
| | | **RMSE** | **0.9725** | **1.0216** | **0.9497** | **1.1910** | **1.0502** |




(a) PCC (b) SCC

Figure 5.10: Box plot of the Pearson and Spearman Correlation Coeficients (PCC and SCC) gathered from testing the FR and NR video quality metrics on the database from Experiment 2.

test. Both tables present results organized according to the type of distortion of interest. From Table 5.4, it can be observed that the proposed audiovisual model presents a good performance, in comparison to the other video quality metrics. In general, the proposed model achieves a correlation above 0.88 at low error rates. Regarding the type of distortion, the model presented a better performance for frame freezing (0.91) compared to packetloss (0.86). As for Table 5.5, results show a clear advantage of the proposed model in comparison to the audio and speech quality metrics. This advantage was expected since audio and speech metrics use only the audio component of the sequence to predict the perceived quality. Regarding the audio distortions, the model presented a better performance for chop and echo distortions (0.92 and 0.90). As for the combination models, the proposed method also performs better and shows a clear advantage.

Table 5.4: Pearson and Spearman Correlation Coefficients (PCC and SCC), and Root Mean Square Error (RMSE) gathered from testing the FR and NR video quality metrics on the database from Experiment 3.

| **Type** | **Modality** | **Metric** | **Measure** | **Packet-Loss** | **Frame-Freezing** | **All** |
|---|---|---|---|---|---|---|
| Full-Reference | Video | Psnr | PCC | 0.8997 | 0.8629 | 0.7694 |
| | | | SCC | 0.9455 | 0.8833 | 0.7368 |
| | | | RMSE | 19.2054 | 16.5837 | 18.0728 |
| | Video | SSIM | PCC | 0.8563 | 0.3899 | 0.3620 |
| | | | SCC | 0.8500 | 0.3727 | 0.3579 |
| | | | RMSE | 2.7378 | 2.2027 | 2.4579 |
| No-Reference | Video | DIIVINE | PCC | -0.8071 | -0.8647 | -0.8344 |
| | | | SCC | -0.8182 | -0.5167 | -0.7519 |
| | | | RMSE | 2.4662 | 2.9484 | 2.6939 |
| | Video | VIIDEO | PCC | -0.7968 | -0.9883 | -0.8496 |
| | | | SCC | -0.6729 | -0.9234 | -0.7834 |
| | | | RMSE | 2.2337 | 2.6804 | 2.4449 |
| | Video | BIQI | PCC | -0.8575 | -0.9022 | -0.8310 |
| | | | SCC | -0.9382 | -0.6000 | -0.8799 |
| | | | RMSE | 34.8427 | 32.6918 | 33.8917 |
| | Video | NIQE | PCC | -0.7608 | -0.9332 | -0.8394 |
| | | | SCC | -0.7798 | -0.7289 | -0.7195 |
| | | | RMSE | 2.9388 | 2.4057 | 2.7119 |
| | Video | BRISQUE | PCC | -0.7094 | -0.9525 | -0.8395 |
| | | | SCC | -0.6360 | -0.9662 | -0.7728 |
| | | | RMSE | 45.1371 | 41.4226 | 43.5049 |
| | Audiovisual | Linear | PCC | 0.3919 | 0.5501 | 0.4431 |
| | | | SCC | 0.2455 | 0.6333 | 0.3368 |
| | | | RMSE | 10.5249 | 11.0035 | 10.7430 |
| | Audiovisual | Minkowski | PCC | 0.2912 | 0.4594 | 0.3422 |
| | | | SCC | 0.2091 | 0.6333 | 0.3143 |
| | | | RMSE | 1.9879 | 2.4289 | 2.1973 |
| | Audiovisual | Power | PCC | -0.6273 | -0.6938 | -0.6616 |
| | | | SCC | -0.6727 | -0.4333 | -0.6075 |
| | | | RMSE | 24.2614 | 23.7806 | 24.0462 |
| | Audiovisual | **DAE-AV** | **PCC** | **0.8638** | **0.9167** | **0.8819** |
| | | | **SCC** | **0.8773** | **0.9050** | **0.8904** |
| | | | **RMSE** | **0.5931** | **0.5718** | **0.5850** |

For a better visualization of the results, Figure 5.11 (a) and (b) depicts bar plots of the overall PCC and SCC values (over the 10 folds) for all metrics. Besides the high correlation values presented by the proposed metric, it can observed that results presented a small variation on both PCC and SCC coefficients. This shows that the results are very consistent compared to the rest of the literature metrics.

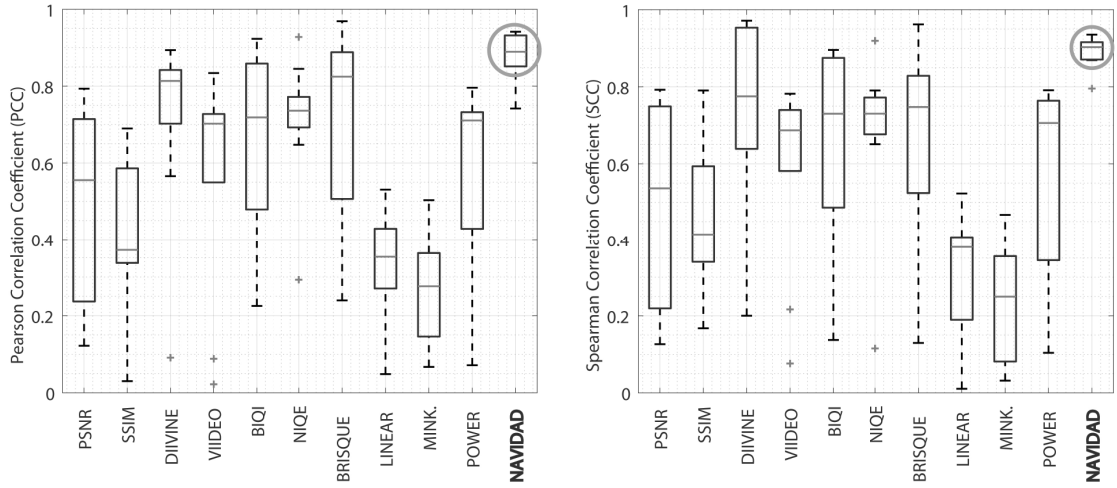
These results backup the use of the autoencoder network approach for the signal quality assessment. Further tests can be performed by using different types of training parameters, which might lead to better results. Seeing that, it is clear that this is still an open task which might lead to new quality assessment methods.

### 5.3.1 LiveNetflix-II Database Analysis

In order to validate the autoencoder network approach, the audiovisual quality model was tested on a second database (LiveNetflix-II Database), provided by the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin (UT Austin)

Table 5.5: Pearson and Spearman Correlation Coefficients (PCC and SCC), and Root Mean Square Error (RMSE) gathered from testing the FR and NR audio quality metrics on the database from Experiment 3.

| Type | Modality | Metric | Measure | Noise | Chop | Clip | Echo | All |
|---|---|---|---|---|---|---|---|---|
| Full-Reference | Audio | VISQOLAudio | PCC | 0.7945 | 0.9909 | 0.7429 | 0.6844 | 0.6008 |
| | | | SCC | 0.7000 | 1.0000 | 0.4928 | 0.5218 | 0.4781 |
| | | | RMSE | 2.4702 | 2.2047 | 2.0815 | 2.2300 | 2.2464 |
| | Speech | VISQOL | PCC | 0.6102 | 0.9915 | 0.5084 | 0.4963 | 0.4236 |
| | | | SCC | 0.7000 | 1.0000 | 0.4928 | 0.5218 | 0.4645 |
| | | | RMSE | 2.6143 | 2.2045 | 2.1639 | 2.3136 | 2.3341 |
| | Audio | PEAQ | PCC | 0.7573 | 0.9347 | 0.8261 | 0.7096 | 0.7689 |
| | | | SCC | 0.2000 | 1.0000 | 0.3189 | 0.3479 | 0.3437 |
| | | | RMSE | 6.3196 | 5.1643 | 5.9748 | 6.0418 | 5.9704 |
| No-Reference | Speech | P.563 | PCC | 0.7305 | 0.9964 | 0.9413 | 0.7752 | 0.7037 |
| | | | SCC | 0.8000 | 1.0000 | 0.8407 | 0.4638 | 0.6367 |
| | | | RMSE | 1.3415 | 1.3252 | 1.2310 | 1.2004 | 1.2650 |
| | Audiovisual | Linear | PCC | 0.4520 | 0.9649 | 0.7718 | 0.0409 | 0.4431 |
| | | | SCC | 0.6000 | 1.0000 | 0.3143 | -0.2571 | 0.3368 |
| | | | RMSE | 10.9449 | 10.7825 | 10.6525 | 10.6429 | 10.7430 |
| | Audiovisual | Minkowski | PCC | 0.3032 | 0.9109 | 0.6881 | -0.2842 | 0.3422 |
| | | | SCC | 0.6000 | 1.0000 | 0.1429 | -0.2571 | 0.3143 |
| | | | RMSE | 2.3585 | 2.2612 | 2.0770 | 2.1419 | 2.1973 |
| | Audiovisual | Power | PCC | -0.7187 | -0.6990 | -0.5271 | -0.8383 | -0.6616 |
| | | | SCC | -0.6000 | -0.5000 | -0.6000 | -0.7714 | -0.6075 |
| | | | RMSE | 23.7961 | 24.0376 | 24.2251 | 24.0783 | 24.0462 |
| | Audiovisual | **DAE-AV** | **PCC** | **0.8879** | **0.9252** | **0.8794** | **0.9044** | **0.8819** |
| | | | **SCC** | **0.9200** | **1.0000** | **0.8629** | **0.9086** | **0.8904** |
| | | | **RMSE** | **0.5764** | **0.6125** | **0.5406** | **0.6013** | **0.5850** |




(a) PCC (b) SCC

Figure 5.11: Box plot of the Pearson and Spearman Correlation Coeficients (PCC and SCC) gathered from testing the FR and NR video quality metrics, plus three Audiovisual combination models, on the database from Experiment 3.

[127]. This database is composed of four hundred and twenty (420) sequences with audio and video components at a Full HD resolution (1920 x 1080, 4:2:0, 24 fps). The videos were processed from 15 source sequences at 7 different network conditions and 4 bitrate adaptation strategies. No audio degradations were included. A total of 65 subjects rated the overall audiovisual quality of the sequences. Regarding the content, a diverse set of content material was used, such as action, documentary, video games, and sports. Figure

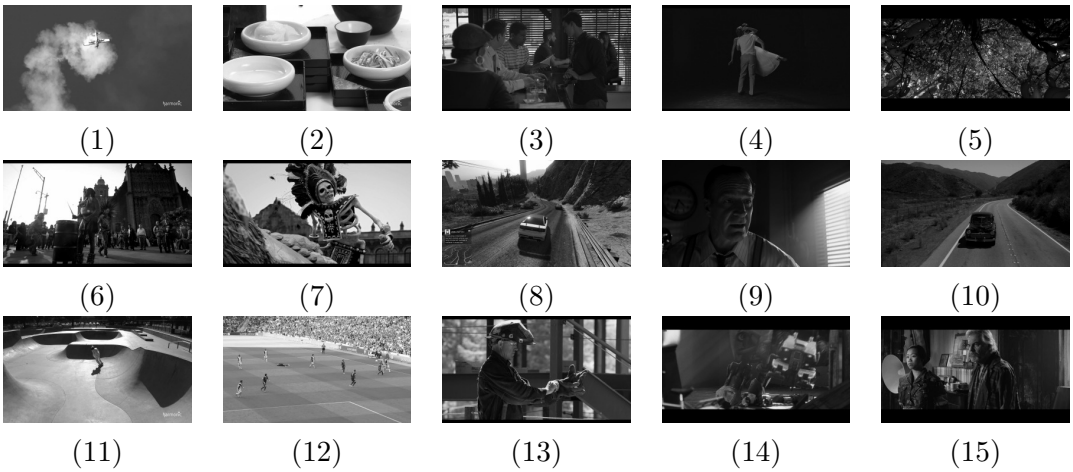



Figure 5.12: Sample frames of the original videos from the LiveNetflix-II database.

5.12 depicts a set of representative frames of the original videos from the LiveNetflix-II database.

The same FR and NR video quality metrics used for comparison previously were also tested on the LiveNetflix-II database. Table 5.6 presents the Pearson and Spearman correlation coefficients, along with the root mean square errors gathered from all quality metrics under test. Results show that the proposed method performs better than the other audio and video quality metrics achieving correlation coefficients above 0.85. These results prove that the proposed model responds well and is able to produce accurate predictions on a external database. Figures 5.13 (a) and (b) present the bar plots for the average PCC and SCC values (over the 10 folds) for the tested metrics. As with the test database (Experiment 3), results show that the proposed metric's correlation values varied very little across the simulations, which shows the consistency of the metric. We believe the proposed metric can be used in real-time streaming environments, specially in cases where audio distortions are expected to happen.

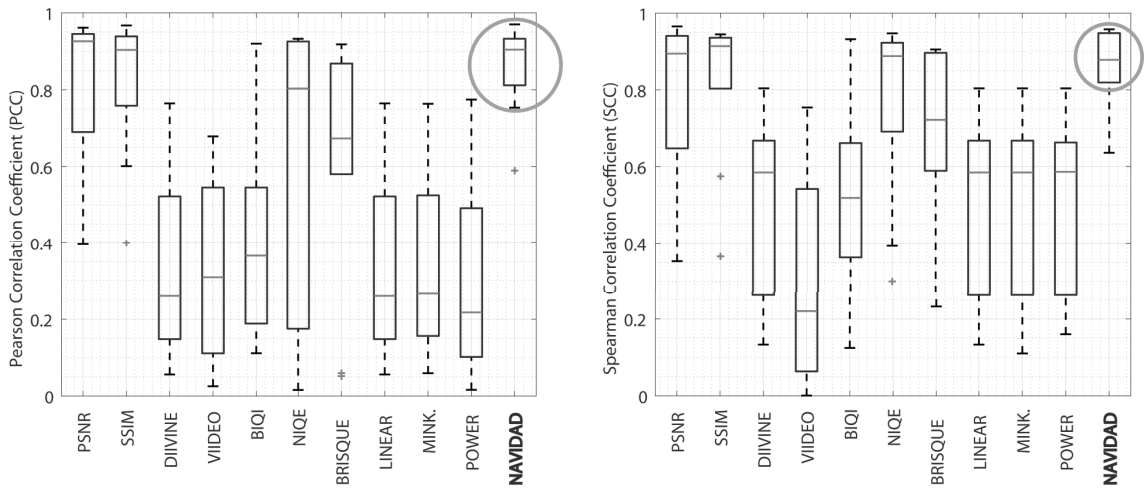
## 5.4 Discussion and Conclusions

Overall, this Chapter presented a set of three NR quality models: 1) a video quality model, 2) an audio quality model, and 3) an audiovisual quality model. These models were built following an autoencoder network approach, and they used the audiovisual material, along with their subjective scores, used in the subjective experiments presented in Chapter 4. The models were trained using a two-layer autoencoder plus a classification function.

In general, the proposed quality models presented good results when predicting the perceived quality of signals. More particularly, the NR video and audiovisual quality

Table 5.6: Pearson and Spearman Correlation Coefficients (PCC and SCC), and Root Mean Square Error (RMSE) gathered from testing the FR and NR audio quality metrics, plus three Audiovisual combination models, on the external database LiveNetflix-II.

| Type | Metric | Measure | All |
|---|---|---|---|
| Full-Reference | PSNR | PCC | 0.6981 |
| | | SCC | 0.6911 |
| | | RMSE | 32.2445 |
| | SSIM | PCC | 0.7333 |
| | | SCC | 0.7123 |
| | | RMSE | 2.3024 |
| No-Reference | DIIVINE | PCC | -0.8364 |
| | | SCC | -0.8106 |
| | | RMSE | 2.6126 |
| | VIIDEO | PCC | -0.6598 |
| | | SCC | -0.7153 |
| | | RMSE | 2.5265 |
| | BIQI | PCC | -0.4263 |
| | | SCC | -0.4724 |
| | | RMSE | 38.3084 |
| | NIQE | PCC | -0.7550 |
| | | SCC | -0.7701 |
| | | RMSE | 3.8324 |
| | BRISQUE | PCC | -0.7271 |
| | | SCC | -0.7115 |
| | | RMSE | 56.2907 |
| | **DAE-AV** | **PCC** | **0.8611** |
| | | **SCC** | **0.8599** |
| | | **RMSE** | **0.5929** |




(a) PCC (b) SCC

Figure 5.13: Box plot of the Pearson and Spearman Correlation Coeficients (PCC and SCC) gathered from testing the FR and NR audio quality metrics, plus three Audiovisual combination models, on the external database LiveNetflix-II.

models showed better performance when they were compared to the FR and NR video quality metrics and some audio-visual methods found in the literature. As for the audio quality model, its results were comparable to the standardized speech NR quality metric. These results proved the value and capacity of the proposed models to predict the quality of signals. For the particular cases of the audio and the audiovisual NR quality models,

they represent an important contribution to the area of audio-visual quality assessment.

# Chapter 6

# Conclusions

In this work, our goal was to investigate how to estimate the audiovisual quality using a no reference autoencoder network approach. Inspired by previous works in the area [25], our proposal used a set of audio and video feature descriptors as input to estimate the overall audiovisual quality. These sets of features were passed on to a two-layer autoencoder that produced a set of features of low dimension. Then, a classification function mapped these features into subjective scores. As a final stage, the output of the model was processed to compute the overall audiovisual quality.

For these experiments, a large group of audiovisual sequences were processed to add different types of video and audio distortions, such as video coding, packet loss, and frame freezing (visual component), and background noise, clip, echo, and chop (audio component). This resulted in a test pool of 720 (Experiment 1), 800 (Experiment 2), and 800 (Experiment 3) audiovisual sequences with their corresponding subjective scores. The experiment results helped us analyze the level of impact certain artifacts have on the perceived quality of the signal and the interaction between audio and video distortions.

Based on the proposed approach, three different NR models were presented: a NR video quality model, a NR audio quality model, and a NR audiovisual quality model, this was focus of the present work. The training of the models was described and their corresponding results were compared against several FR and NR metrics from the literature. Finally, the NR audiovisual quality model was tested on an external audiovisual database.

## 6.1 Summary of the Contributions

The main contributions of this work are:

- Generation and publication of three large databases of audio-visual stimuli, containing different audio and video distortions, and their corresponding subjective data. These databases can be used for:

  - Comparison of audio-visual techniques,
  - Training new ML-based quality assessment techniques,
  - Exploring how humans perceive different types of artifacts, like for example coding, packet loss and frame freezing artifacts.

- Development of a NR video quality assessment model based on an Autoencoder Network approach. The model is able to predict, at a fair level of accuracy, the perceived video quality for a variety of common visual distortions.

- Development of a NR audio quality assessment model based on an Autoencoder Network approach. This model has a significant potential to produce better results and it represents an important contribution given its non-intrusive nature.

- Development of a NR audiovisual quality assessment model based on an Autoencoder Network approach. This model is able to predict, at a good level of accuracy, the perceived audiovisual quality over a variety of common audio and video distortions. This model is the main focus of this work, representing an important contribution given the reduced number of models available in the current literature. Additionally, because of its non-intrusive nature, it serves as a base to the development of better and more complex audiovisual quality assessment tools.

## 6.2 Future Work

Some activities for future work include:

- NR Video Quality Model
  - Refining the training parameters to achieve the best possible performance of the model.
  - Searching and testing additional visual features with the objective of increasing the set of descriptive features of the model.
  - Training the model on different video databases in order to verify its performance on different content and visual distortions.

- NR Audio Quality Model
  - Refining the training parameters to achieve the best possible performance of the model.

  - Searching and testing additional audio features with the objective of increasing the set of descriptive features of the model.
  - Training the model on different audio (speech) databases in order to verify its performance on different content and audio distortions.
  - Revise the model performance using a content classification module. More specifically, by classifying and training individual models for specific content like music, speech, and environmental sound.

- NR Audiovisual Quality Model

  - Refining the training parameters to achieve the best possible performance of the model.
  - Searching and testing additional audio and visual features with the objective of increasing the set of descriptive features of the model.
  - Training the model on different audiovisual databases in order to verify its performance on different content and audiovisual distortions.
  - Explore the adaptation of the proposed approach with the objective of dealing with the audio and video synchronization problem.

# Bibliography


[1] Jari Korhonen. Audiovisual quality assessment in communications applications: current status, trends and challenges. *Signal Processing*, pages 6–9, 2010. 4

[2] ITU Recommendation BT.500-8. *Methodology for subjective assessment of the quality of television pictures*. 1998. 4

[3] European broadcasting union. 4

[4] *Visual quality assessment algorithms: what does the future hold?*, volume 51, February 2011. 5, 6

[5] *Digital images and human vision*. MIT Press, 1993. 5

[6] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam. Objective video quality assessment methods: A classification, review,and performance comparison. 57(2):165–182, 2011. 5

[7] *Perceptual visual quality metrics: A survey*, volume 22, 2011. 5, 6

[8] S. Daly. The visible differences predictor: an algorithm for the assessment of image fidelity. In Andrew B. Watson, editor, *Digital Images and Human Vision*, pages 179–206. MIT Press, Cambridge, Massachusetts, 1993. 5

[9] M.H. Pinson and S. Wolf. An objective method for combining multiple subjective data sets. In *Proc. SPIE Conference on Visual Communications and Image Processing*, volume 5150, pages 583–92, Lugano, Switzerland, 2003. 5, 56

[10] Z Wang, L Lu, and A. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Comm.*, vol19:121–132, 2004. 5, 42

[11] Stefan Winkler and Praveen Mohandas. The evolution of video quality measurement: From psnr to hybrid metrics. *IEEE Transactions on Broadcasting*, 54(3):660–668, 2008. 5

[12] E. Grineko, K. Glasman, and A. Belozertsev. Content-adaptive bitrate reduction in mobile multimedia applications. pages 176–180, Sept 2014. 5

[13] K. Soh and S. Iah. Subjectively assessing method for audiovisual quality using equivalent signal-to-noise ratio conversion. *Trans. Inst. Electron., Inform. Commun. Eng. A*, 11:1305–1313, 2001. 6

[14] David S Hands. A Basic Multimedia Quality Model. *Multimedia, IEEE Transactions on*, 6(6):806–816, 2004. 6, 35, 43, 49, 56

[15] 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*, pages 1–14, 2011. 6, 35, 43, 49, 56

[16] K. Yamagishi and S. Gao. Light-weight audiovisual quality assessment of mobile video: Itu-t rec. p.1201.1. In *Multimedia Signal Processing (MMSP), IEEE 15th International Workshop on*, pages 464–469, Sept 2013. 6

[17] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE transactions on broadcasting*, 57(2):165, 2011. 6

[18] Margaret Pinson, William Ingram, and Arthur Webster. Audiovisual quality components. *IEEE Signal Processing Magazine*, 6(28):60–67, 2011. 6

[19] MN Garcia and A Raake. Impairment-factor-based audio-visual quality model for iptv. In *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, pages 1–6. IEEE, 2009. 6

[20] Helard A Becerra Martinez and Mylene CQ Farias. Combining audio and video metrics to assess audio-visual quality. *Multimedia Tools and Applications*, pages 1–20, 2018. 6

[21] Mikołaj Leszczuk, Mateusz Hanusiak, Ignacio Blanco, Andrzej Dziech, Jan Derkacz, Emmanuel Wyckens, and Silvio Borer. Key indicators for monitoring of audiovisual quality. In *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*, pages 2301–2305. IEEE, 2014. 6, 19, 21, 50

[22] Junyong You, Ulrich Reiter, Miska M Hannuksela, Moncef Gabbouj, and Andrew Perkis. Perceptual-based quality assessment for audio–visual services: A survey. *Signal Processing: Image Communication*, 25(7):482–501, 2010. 6, 12

[23] Benjamin Belmudez and Sebastian Möller. Audiovisual quality integration for interactive communications. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):24, 2013. 6

[24] Zahid Akhtar and Tiago H Falk. Audio-visual multimedia quality assessment: A comprehensive survey. *IEEE Access*, 5:21090–21117, 2017. 6, 35, 50

[25] Meet H Soni and Hemant A Patil. Novel deep autoencoder features for non-intrusive speech quality assessment. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 2315–2319. IEEE, 2016. 7, 8, 28, 29, 102, 117

[26] Gemma Calvert, Charles Spence, Barry E Stein, et al. *The handbook of multisensory processes*. MIT press, 2004. 10

[27] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment. Technical report, VQEG, 2003. 10

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

[29] Rafael C Gonzalez, Richard E Woods, and Prentice Hall. *Digital Image Processing (2nd Edition)*. Tom Robbins, 1987.

[30] David Alleysson, Sabine Susstrunk, and Jeanny Hérault. Linear demosaicing inspired by the human visual system. *IEEE Transactions on Image Processing*, 14(4):439–449, 2005.

[31] Felipe Viegas Rodrigues. Fisiologia sensorial. 5:24–32, 2010. 11

[32] Juan Pedro Lopez Velasco. Video quality assessment. In Teodora Smiljanic, editor, *Video Compression, Edited by Amal Punchihewa*, pages 129–154. InTech, amal punch edition, 2012. 11, 59

[33] James E Birren, Gene D Cohen, R Bruce Sloane, Barry D Lebowitz, Donna E Deutchman, May Wykle, and Nancy R Hooyman. *Handbook of mental health and aging*. Academic Press, 2013. 12

[34] Ferreira. Olivia Dayse Leite Freire. Rosalia Carmen de Lima Santos. Natanael Antonio dos Gadelha. Maria Jose Nunes, Andrade. Michael Jackson Oliveira. Sensibilidade ao contraste acromatico para grades senoidais verticais em adolescentes e adultos. *Psicologia: teoria e pratica*, 12:59 – 70, 00 2010. 12

[35] Yoshiharu Soeta and Yoichi Ando. *Neurally based measurement and evaluation of environmental noise*. Springer, 2015. 12, 21

[36] Ian McLoughlin. *Applied speech and audio processing: with Matlab examples*. Cambridge University Press, 2009. 12, 14

[37] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2013. 13

[38] Andrew Hines and Naomi Harte. Speech intelligibility prediction using a neurogram similarity index measure. *Speech Communication*, 54(2):306–320, 2012. 14

[39] Andrew Hines, Jan Skoglund, Anil Kokaram, and Naomi Harte. Visqol: the virtual speech quality objective listener. In *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, pages 1–4. VDE, 2012. 14, 47, 102, 109

[40] Andrew Hines, Eoin Gillen, Damien Kelly, Jan Skoglund, Anil Kokaram, and Naomi Harte. Visqolaudio: An objective audio quality metric for low bitrate codecs. *The Journal of the Acoustical Society of America*, 137(6):EL449–EL455, 2015. 14, 85, 92, 109

[41] R. E. Bosi, M. Goldberg. *Introduction to Digital Audio Coding and Standards*. Springer International Series in Engineering and Computer Science, 2003. 16

[42] Hong Ren Wu and Kamisetty Ramamohan Rao. *Digital video image quality and perceptual coding*. CRC press, 2017. 17

[43] Oliver Rose. Statistical properties of mpeg video traffic and their impact on traffic modeling in atm systems. In *Local Computer Networks, 1995., Proceedings. 20th Conference on*, pages 397–406. IEEE, 1995. 17

[44] Lajos Hanzo, Peter Cherriman, and Jurgen Streit. *Video compression and communications: from basics to H. 261, H. 263, H. 264, MPEG4 for DVB and HSDPA-style adaptive turbo-transceivers*. John Wiley & Sons, 2007. 18

[45] 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*, 16(1):134–140, 2006. 18

[46] Martin Řeřábek and Touradj Ebrahimi. Comparison of compression efficiency between hevc/h. 265 and vp9 based on subjective assessments. In *Applications Of Digital Image Processing Xxxvii*, volume 9217, page 92170U. International Society for Optics and Photonics, 2014. 18

[47] Grzegorz Pastuszak and Andrzej Abramowski. Algorithm and architecture design of the h. 265/hevc intra encoder. *IEEE Trans. Circuits Syst. Video Techn.*, 26(1):210–222, 2016. 18

[48] Andreas Unterweger. Compression artifacts in modern video coding and state-of-the-art means of compensation. *Multimedia Networking and Coding*, page 28, 2012. 19

[49] Mu Mu, Piotr Romaniak, Andreas Mauthe, Mikołaj Leszczuk, Lucjan Janowski, and Eduardo Cerqueira. Framework for the integrated video quality assessment. *Multimedia Tools and Applications*, 61(3):787–817, 2012. 19

[50] Eduardo Cerqueira, Lucjan Janowski, Mikołaj Leszczuk, Zdzisław Papir, and Piotr Romaniak. Video artifacts assessment for live mobile streaming applications. In *Future Multimedia Networking*, pages 242–247. Springer, 2009. 19

[51] Marie-Neige Garcia, Dominika Dytko, and Alexander Raake. Quality impact due to initial loading, stalling, and video bitrate in progressive download video services. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 129–134. IEEE, 2014. 21, 41, 60, 71

[52] Tilman Liebchen and Yuriy A Reznik. Mpeg-4 als: An emerging standard for lossless audio coding. In *Data Compression Conference, 2004. Proceedings. DCC 2004*, pages 439–448. IEEE, 2004. 21

[53] K Brandenburg and H Popp. Mpeg layer-3. *EBU Technical review*, pages 1–15, 2000. 22

[54] Naomi Harte, Eoin Gillen, and Andrew Hines. Tcd-voip, a research database of degraded speech for assessing quality in voip applications. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, pages 1–6. IEEE, 2015. 22, 23, 51, 55, 62, 79, 82, 83, 84, 88

[55] Andrew Hines. *Predicting Speech Intelligibility*. Citeseer, 2012. 22, 102

[56] Doh-Suk Kim and Ahmed Tarraf. Enhanced perceptual model for non-intrusive speech quality assessment. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006. 24

[57] Bruno Defraene, Toon van Waterschoot, Hans Joachim Ferreau, Moritz Diehl, and Marc Moonen. Perception-based clipping of audio signals. In *Signal Processing Conference, 2010 18th European*, pages 517–521. IEEE, 2010. 24

[58] MM Sondhi. An adaptive echo canceller. *Bell System Technical Journal*, 46(3):497–511, 1967. 24

[59] ITUT Rec. G. 107-the e model, a computational model for use in transmission planning. *International Telecommunication Union*, 8:20–21, 2003. 24, 34

[60] Mohammed Alreshoodi and John Woods. Survey on qoe\qos correlation models for multimedia services. *arXiv preprint arXiv:1306.0221*, 2013. 25

[61] Paolo Gastaldo and Judith A Redi. Machine learning solutions for objective visual quality assessment. In *6th international workshop on video processing and quality metrics for consumer electronics, VPQM*, volume 12, 2012. 25, 35

[62] Tom M Mitchell et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997. 26

[63] Sara Landset, Taghi M Khoshgoftaar, Aaron N Richter, and Tawfiq Hasanin. A survey of open source tools for machine learning with big data in the hadoop ecosystem. *Journal of Big Data*, 2(1):24, 2015. 26

[64] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 26, 27, 28, 29

[65] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012. 28

[66] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 28

[67] Anthony Stark and Kuldip Paliwal. Mmse estimation of log-filterbank energies for robust speech recognition. *Speech Communication*, 53(3):403–416, 2011. 29

[68] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997. 31

[69] ITUTJ ITU. Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference. 34, 42

[70] ITU Telecommunication Standardization Sector. Objective perceptual multimedia video quality measurement in the presence of a full reference. *ITU-T Recommendation J*, 247, 2008. 34

[71] ITU-R. Recommendation bs.1387 : Method for objective measurements of perceived audio quality. Technical report, 1998. 34, 46

[72] M Pinson, Marc Sullivan, and Andrew Catellier. A new method for immersive audiovisual subjective testing. In *Proceedings of the 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2014. 34, 35, 36, 38, 40, 66, 68, 87

[73] Stefan Winkler and Christof Faller. Perceived audiovisual quality of low-bitrate multimedia content. *Multimedia, IEEE Transactions on*, 8(5):973–980, 2006. 35, 43, 49

[74] H. Becerra Martinez and M.C.Q. Farias. A no-reference audio-visual video quality metric. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 2125–2129, Sept 2014. 35, 43, 49

[75] ITU-R. Recommendation bt.500-8: Methodology for subjective assessment of the quality of television pictures. Technical report, 1998. 36, 66

[76] F Kozamernik, V Steinmann, P Sunna, and E Wyckens. Samviq—a new ebu methodology for video quality evaluations in multimedia. *Motion Imaging Journal, SMPTE*, 114(4):152–160, 2005. 36

[77] ITU-T. Recommendation p.1301 : Subjective quality evaluation of audio and audiovisual multiparty telemeetings. Technical report, 2013. 36

[78] ITU-R. Recommendation p.800 : Methods for subjective determination of transmission quality. Technical report, 1996. 36

[79] EBU. Tech review 274: Subjective assessment of audio quality. Technical report, 1998. 36

[80] ITU-R. Recommendation P.911 : Subjective audiovisual quality assessment methods for multimedia applications. 1998. 36, 70

[81] ITU-T. P.913 : Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment. Technical report, 2014. 36

[82] Margaret H Pinson, Marcus Barkowsky, and Patrick Le Callet. Selecting scenes for 2d and 3d subjective video quality tests. *EURASIP Journal on Image and Video Processing*, 2013(1):1–12, 2013. 37, 39, 40, 77

[83] Nicolas Staelens, Stefaan Moens, Wendy Van den Broeck, Ilse Marien, Brecht Vermeulen, Peter Lambert, Rik Van de Walle, and Piet Demeester. Assessing quality of experience of iptv and video on demand services in real-life environments. *Broadcasting, IEEE Transactions on*, 56(4):458–466, 2010. 37

[84] *ITU-T Recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures*, 1998. 37, 70

[85] Adam Borowiak, Ulrich Reiter, and U Peter Svensson. Quality evaluation of long duration audiovisual content. In *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*, pages 337–341. IEEE, 2012. 37, 56

[86] Frank E. Beerends, John G.; De Caluwe. The influence of video quality on perceived audio quality and vice versa. *J. Audio Eng. Soc*, 47(5):355–362, 1999. 38

[87] Werner Robitza, Marie Neige Garcia, and Alexander Raake. At home in the lab: Assessing audiovisual quality of http-based adaptive streaming with an immersive test paradigm. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, pages 1–6. IEEE, 2015. 41

[88] Nicolas Staelens, Paulien Coppens, Niels Van Kets, Glenn Van Wallendaef, Wendy Van den Broeck, Jan De Cock, and Filip De Turek. On the impact of video stalling and video quality in the case of camera switching during adaptive streaming of sports content. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, pages 1–6. IEEE, 2015. 41, 56

[89] Deepti Ghadiyaram and Alan C Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision*, 17(1):32–32, 2017. 41

[90] Mona Hakami and W Bastiaan Kleijn. Machine learning based non-intrusive quality estimation with an augmented feature set. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5105–5109. IEEE, 2017. 41

[91] Damon M Chandler and Sheila S Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *Image Processing, IEEE Transactions on*, 16(9):2284–2298, 2007. 42

[92] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402. Ieee, 2003. 42, 108

[93] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: a feature similarity index for image quality assessment. *Image Processing, IEEE Transactions on*, 20(8):2378–2386, 2011. 42

[94] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *Image Processing, IEEE Transactions on*, 20(5):1185–1198, 2011. 42

[95] Hamid Rahim Sheikh and Alan C Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2):430–444, 2006. 42

[96] Kalpana Seshadrinathan and Alan Conrad Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on*, 19(2):335–350, 2010. 42

[97] Zhou Wang and Eero P Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Electronic Imaging 2005*, pages 149–159. International Society for Optics and Photonics, 2005. 42, 43

[98] Judith A Redi, Paolo Gastaldo, Ingrid Heynderickx, and Rodolfo Zunino. Color distribution information for the reduced-reference assessment of perceived image quality. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(12):1757–1769, 2010. 42, 43

[99] M.C.Q. Farias and S.K. Mitra. No-reference video quality metric based on artifact measurements. *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, 3(2):III – 141–4, 2005. 43

[100] Jorge E Caviedes. No-reference quality metric for degraded and enhanced video. In Touradj Ebrahimi and Thomas Sikora, editors, *Proceedings of SPIE*, volume 5150, pages 621–632. SPIE, 2003. 43

[101] Rajiv Soundararajan and Alan C Bovik. Rred indices: Reduced reference entropic differencing for image quality assessment. *Image Processing, IEEE Transactions on*, 21(2):517–526, 2012. 42

[102] Ravi Soundararajan and Alan C Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. *Circuits and Systems for Video Technology, IEEE Transactions on*, 23(4):684–694, 2013. 42

[103] Zhou Wang, Alan C Bovik, and BL Evan. Blind measurement of blocking artifacts in images. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 3, pages 981–984. Ieee, 2000. 43

[104] Michael Yuen and HR Wu. A survey of hybrid mc/dpcm/dct video coding distortions. *Signal processing*, 70(3):247–278, 1998. 43

[105] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 723–727. IEEE, 2011. 43, 101

[106] A. Mittal, A.K. Moorthy, and A.C. Bovik. No-reference image quality assessment in the spatial domain. *Image Processing, IEEE Trans. on*, 21(12):4695–4708, Dec 2012. 43

[107] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *Image Processing, IEEE Transactions on*, 25(1):289–300, 2016. 44, 78, 99, 101, 108

[108] Yi Zhang, Anush K Moorthy, Damon M Chandler, and Alan C Bovik. C-diivine: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes. *Signal Processing: Image Communication*, 29(7):725–747, 2014. 44, 45, 91, 99, 101, 109

[109] Eero P Simoncelli, William T Freeman, Edward H Adelson, and David J Heeger. Shiftable multiscale transforms. *IEEE transactions on Information Theory*, 38(2):587–607, 1992. 44

[110] Martin J Wainwright and Odelia Schwartz. 10 natural image statistics and divisive. *Probabilistic models of the brain: Perception and neural function*, page 203, 2002. 44

[111] Karnran Sharifi and Alberto Leon-Garcia. Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(1):52–56, 1995. 45

[112] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 2, pages 749–752. IEEE, 2001. 46

[113] John G Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment. *Journal of the Audio Engineering Society*, 61(6):366–384, 2013. 46

[114] L. Malfait, J. Berger, and M. Kastner. The itu-t standard for single-ended speech quality assessment. Technical Report 6, 2006. 47, 109

[115] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015. 48, 85

[116] Kazuhisa Yamagishi and Shan Gao. Light-weight audiovisual quality assessment of mobile video: Itu-t rec. p. 1201.1. In *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, pages 464–469. IEEE, 2013. 49

[117] New Recommendations CCITT. Q1200—q series: Intelligent network recommendation. Technical report, Technical report, CCITT, COM XI. 51

[118] DJ MEARS. Nbc time/frequency module subjective tests: overall results. *ISO/IEC JTC1/SC29/WG11/N0973*, 1995. 51

[119] Zhonghua Li, Ju-Chiang Wang, Jingli Cai, Zhiyan Duan, Hsin-Min Wang, and Ye Wang. Non-reference audio quality assessment for online live music recordings. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 63–72. ACM, 2013. 51

[120] Johannes A Louw, Avashlin Moodley, and Avashna Govender. The speect text-to-speech entry for the blizzard challenge 2016. In *Proceedings of The Blizzard Challenge 2016 Workshop*, 2016. 51

[121] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on image processing*, 19(6):1427–1441, 2010. 51

[122] Video Quality Experts Group et al. Report on the validation of video quality models for high definition video content. *http://www. its. bldrdoc. gov/media/4212/vqeg_hdtv_final_report_version_2. 0. zip*, 2010. 51, 52

[123] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086, 2016. 51

[124] Margaret H Pinson, Christian Schmidmer, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Phillip Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram. Subjective and objective evaluation of an audiovisual subjective dataset for research and development. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 30–31. IEEE, 2013. 51, 52

[125] Helard Becerra Martinez and Mylène CQ Farias. Full-reference audio-visual video quality metric. *Journal of Electronic Imaging*, 23(6):061108–061108, 2014. 51, 52, 56

[126] Edip Demirbilek and Jean-Charles Grégoire. Towards reduced reference parametric models for estimating audiovisual quality in multimedia services. *arXiv preprint arXiv:1604.07211*, 2016. 51

[127] Christos G Bampis, Zhi Li, Ioannis Katsavounidis, Te-Yuan Huang, Chaitanya Ekanadham, and Alan C Bovik. Towards perceptually optimized end-to-end adaptive video streaming. *arXiv preprint arXiv:1808.03898*, 2018. 51, 52, 56, 113

[128] Nicolas Staelens, Brecht Vermeulen, Stefaan Moens, Jean-François Macq, Peter Lambert, Rik Van de Walle, and Piet Demeester. Assessing the influence of packet loss and frame freezes on the perceptual quality of full length movies. In *4th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2009)*, 2009. 55

[129] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo De Veciana. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):652–671, 2012. 55

[130] Tiago H Falk and Wai-Yip Chan. Performance study of objective speech quality measurement for modern wireless-voip communications. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009:12, 2009. 55

[131] Takeshi Yamada, Masakazu Kumakura, and Nobuhiko Kitawaki. Subjective and objective quality assessment of noise reduced speech signals. In *Nonlinear Signal and Image Processing, 2005. NSIP 2005. Abstracts. IEEE-Eurasip*, page 28. IEEE, 2005. 55

[132] Dorothea Wendt, Torsten Dau, and Jens Hjortkjær. Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology*, 7:345, 2016. 55

[133] Margaret H Pinson, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Christian Schmidmer, Phillip Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram. The influence of subjects and environment on audiovisual subjective tests: An international study. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):640–651, 2012. 56

[134] VQEG. Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, Phase I. Technical report, 2008. 57

[135] A. Ostaszewska and R. Kloda. Quantifying the amount of spatial and temporal information in video test sequences. In *Recent Advances in Mechatronics, Springer*, pages 11–15. 2007. 57, 101

[136] 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*, pages 90–93, 2007. 57

[137] ITU-T. H.264 : Advanced video coding for generic audiovisual services. Technical report, 2003. 60

[138] ITU-T. H.265 : High efficiency video coding. Technical report, 2013. 60

[139] Michael Horowitz, Faouzi Kossentini, Nader Mahdi, Shilin Xu, Hsan Guermazi, Hassene Tmar, Bin Li, Gary J Sullivan, and Jizheng Xu. Informal subjective quality comparison of video compression performance of the hevc and h. 264/mpeg-4 avc standards for low-delay applications. In *SPIE Optical Engineering+ Applications*, pages 84990W–84990W. International Society for Optics and Photonics, 2012. 60, 75, 76

[140] Alexis Michael Tourapis, Karsten Sühring, and Gary Sullivan. H. 264/mpeg-4 avc reference software manual. *Geneva, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q*, 6, 2007. 61

[141] Frank Bossen, Davin Flynn, and Karsten Suhring. Hm software manual. *JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, AHG chairs*, 2014. 61

[142] Jill M Boyce and Robert D Gaglianello. Packet loss effects on mpeg video sent over the public internet. In *Proceedings of the sixth ACM international conference on Multimedia*, pages 181–190. ACM, 1998. 61

[143] Judith Redi, Ingrid Heynderickx, Bruno Macchiavello, and Max Farias. On the impact of packet-loss impairments on visual attention mechanisms. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 1107–1110. IEEE, 2013. 61

[144] Pedro Garcia Freitas, Judith A Redi, Mylene CQ Farias, and Alexandre F Silva. Video quality ruler: A new experimental methodology for assessing video quality. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, pages 1–6. IEEE, 2015. 61

[145] Stephan Wenger. H. 264/avc over ip. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(7):645–656, 2003. 61

[146] Peter Kabal. Tsp speech database. *McGill University, Database Version*, 1(0):09–02, 2002. 63

[147] J Martin Bland and Douglas G Altman. Statistics notes: Cronbach's alpha. *Bmj*, 314(7080):572, 1997. 69

[148] Jose M Cortina. What is coefficient alpha? an examination of theory and applications. *Journal of applied psychology*, 78(1):98, 1993. 69

[149] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951. 69

[150] J-R Ohm, Gary J Sullivan, Holger Schwarz, Thiow Keng Tan, and Thomas Wiegand. Comparison of the coding efficiency of video coding standards—including high efficiency video coding (hevc). *Circuits and Systems for Video Technology, IEEE Transactions on*, 22(12):1669–1684, 2012. 75, 76

[151] Basak Oztas, Mahsa T Pourazad, Panos Nasiopoulos, and Victor Leung. A study on the hevc performance over lossy networks. In *Electronics, Circuits and Systems (ICECS), 2012 19th IEEE International Conference on*, pages 785–788. IEEE, 2012. 75

[152] Pablo Pinol, Abel Torres, Oscar Lopez, Manuel Martinez, and Manuel P Malumbres. Evaluating hevc video delivery in vanet scenarios. In *Wireless Days (WD), 2013 IFIP*, pages 1–6. IEEE, 2013. 75

[153] Tobias Hoßfeld, Sebastian Egger, Raimund Schatz, Markus Fiedler, Kathrin Masuch, and Charlott Lorentzen. Initial delay vs. interruptions: between the devil and the deep blue sea. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 1–6. IEEE, 2012. 76

[154] Philip Kortum and Marc Sullivan. The effect of content desirability on subjective video quality ratings. *Human factors: the journal of the human factors and ergonomics society*, 52(1):105–118, 2010. 77

[155] Maria-Dolores Cano and Fernando Cerdan. Subjective QoE analysis of VoIP applications in a wireless campus environment. *Telecommunication Systems*, 49(1):5–15, 2012. 87

[156] AK Moorthy and AC Bovik. A modular framework for constructing blind universal quality indices. *IEEE Signal Processing Letters*, 17, 2009. 109

[157] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a" completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013. 109

[158] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 109

[159] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes. Peaq-the itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29, 2000. 109

[160] Helard Becerra Martinez and Mylène C Farias. Full-reference audio-visual video quality metric. *Journal of Electronic Imaging*, 23(6):061108, 2014.

# Anexo I

# Representative Frames of Source Videos


Figure I.1: Sample frames of original videos used in Experiment 1.


Figure I.2: Sample frames of original videos used in Experiment 2.


Figure I.3: Sample frames of original videos used in Experiment 3.

# Anexo II

# Source Stimuli, Content Description

Table II.1: Video content description

| Sequence | Sequence | Length | Video Content | Audio Content | Content Id |
|---|---|---|---|---|---|
| v01 | Guy Sleeping | 00:56 | Random People | Music | 1 |
| v02 | Flamenco (Seq1) | 00:33 | People dancing | Music | 2 |
| v03 | Big Buck Bunny (Seq1) | 00:37 | Computer graphics | Music, surround sound | 3 |
| v04 | Big Buck Bunny (Seq2) | 00:38 | Computer graphics | Music, surround sound | 3 |
| v05 | Elephant (Seq1) | 01:08 | Computer graphics | Music, speech, surround sound | 4 |
| v06 | Elephant (Seq2) | 00:42 | Computer graphics | Speech , surround sound | 4 |
| v07 | France Tourism (Seq1) | 00:39 | Random people, landscape | surround sound | 7 |
| v08 | WomanDay (Seq1) | 00:34 | Slow motion scenes | Soft music | 18 |
| v09 | Taiwan (Seq1) | 00:34 | Landscape, fast motion | Soft music | 22 |
| v10 | Barca vs Atletic (Seq1) | 00:39 | Sports (Soccer match) | Speech (narrative, background noise) | 20 |
| v11 | FootMusic (Seq1) | 00:33 | Rock band playing | rock music | 17 |
| v12 | Atlanta Betline (Seq1) | 00:43 | Landscape, people talking | Speech | 15 |
| v13 | Taiwan (Seq2) | 00:34 | Landscape night, fireworks | rock music | 22 |
| v14 | Netflix El Fuente (Seq1) | 00:35 | Random people, landscape | Music | 24 |
| v15 | Box interview NTIA (Seq1) | 00:31 | Boxing, people talking | Speech, surround sound | 6 |
| v16 | Honey Bees (Seq1) | 00:38 | Bees in nature | Music | 11 |
| v17 | Barca vs Atletic (Seq2) | 00:38 | Sports (Soccer match) | Speech (narrative, background noise) | 20 |
| v18 | WomanDay (Seq2) | 00:34 | Slow motion scenes | Soft music | 18 |
| v19 | France Tourism (Seq2) | 00:33 | Random people, landscape | surround sound | 7 |
| v20 | Kenpo Strikes NTIA | 00:31 | Sports (kempo performance) | minimal surround sound | 8 |
| v21 | Box interview NTIA (Seq2) | 00:33 | Boxing, people talking | Speech, surround sound | 6 |
| v22 | Taipei Fireworks (Seq1) | 00:55 | Landscape night | Soft music | 21 |
| v23 | WomanDay (Seq3) | 00:34 | Slow motion scenes | Soft music | 18 |
| v24 | Taiwan (Seq2) | 00:34 | Landscape, fast motion | Soft music | 22 |
| v25 | Old Town Car NTIA | 00:22 | People, car | car sound, speech | 9 |
| v26 | Netflix El Fuente (Seq2) | 00:35 | Random people, landscape | Music | 24 |
| v27 | Barca vs Atletic (Seq3) | 00:50 | Sports (Football match) | Speech (narrative, background noise) | 20 |
| v28 | NTIA Violin (Seq1) | 00:30 | Violin performance | Speech, violin music | 23 |
| v29 | Netflix El Fuente (Seq3) | 00:40 | Random people, landscape | Music | 24 |
| v30 | Atlanta Betline (Seq2) | 00:54 | Landscape, people talking | Speech | 15 |
| v31 | Puppies (Seq1) | 00:34 | Puppies | Music, surround sound | 10 |
| v32 | Taiwan (Seq3) | 00:34 | Landscape, fast motion | Soft music | 22 |
| v33 | Big Green Rabbit | 00:30 | Computer graphics | Music | 13 |
| v34 | Movie Trailer Sintel | 00:35 | Computer graphics | Speech, music, surround sound | 5 |
| v35 | Honey Bees (Seq2) | 00:42 | Bees in nature | Music | 11 |
| v36 | Atlanta Betline (Seq3) | 00:54 | Landscape, people talking | Speech | 15 |
| v37 | Atlanta Betline (Seq4) | 00:38 | Landscape, people talking | Speech | 15 |
| v38 | Netflix El Fuente (Seq4) | 00:40 | Random people, landscape | Music | 24 |
| v39 | Landscape Fast | 00:37 | Landscape, fast motion | Music | 12 |
| v40 | Barca vs Atletic (Seq4) | 00:40 | Sports (Soccer match) | Speech (narrative, background noise) | 20 |
| v41 | FoxBird | 00:30 | Computer graphics | Speech, music | 16 |
| v42 | Fishing Florida (Seq1) | 00:33 | Random fishing scenes | Music | 14 |
| v43 | Kenpo NTIA | 00:28 | Sports (kempo performance) | minimal surround sound | 8 |
| v44 | Taipei Fireworks (Seq2) | 00:51 | Landscape night, fireworks | rock music | 21 |
| v45 | WomanDay (Seq4) | 00:34 | Slow motion scenes | Soft music | 18 |
| v46 | Netflix El Fuente (Seq5) | 00:33 | Random people, landscape | Music | 24 |
| v47 | NTIA Violin (Seq2) | 00:33 | Violin performance | Speech, violin music | 23 |
| v48 | Puppies (Seq2) | 00:47 | Puppies | Music, surround sound | 10 |
| v49 | Netflix El Fuente (Seq6) | 00:33 | Random people, landscape | Speech, music | 24 |
| v50 | Barca vs Atletic (Seq4) | 00:40 | Sports (Soccer match) | Speech (narrative, background noise) | 20 |
| v51 | Netflix El Fuente (Seq7) | 00:49 | Random people, landscape | Music | 24 |
| v52 | Food | 00:37 | Dishes and people | Soft music | 19 |
| v53 | France Tourism (Seq3) | 00:33 | Random people, landscape | surround sound | 7 |
| v54 | Netflix El Fuente (Seq8) | 00:34 | Random people, landscape | Speech, music | 24 |
| v55 | FootMusic (Seq2) | 00:33 | Rock band playing | rock music | 17 |
| v56 | Fishing Florida (Seq2) | 00:33 | Underwater scenes | Music, speech, surround sound | 14 |
| v57 | Big Buck Bunny (Seq3) | 00:36 | Computer graphics | Music, surround sound | 3 |
| v58 | Box interview (Seq3) | 00:33 | Boxing, people talking | Speech, surround sound | 6 |
| v59 | Elephant (Seq3) | 00:40 | Computer graphics | Music, speech, surround sound | 4 |
| v60 | Flamenco (Seq2) | 00:33 | People dancing | Music | 2 |

# Anexo III

# Encoder parameters: AVC - HEVC

Table III.1: Encoder parameters - AVC

| Input YUV file | : VideoFile.yuv |
|---|---|
| Output H.264 bitstream | : CodedFile.264 |
| Output YUV file | : YuvFile.yuv |
| YUV Format | : YUV 4:2:0 |
| Frames to be encoded | : 0 |
| Freq. for encoded bitstream | : 30 |
| PicInterlace / MbInterlace | : 0/0 |
| Transform8x8Mode | : 1 |
| ME Metric for Refinement Level 0 | : SAD |
| ME Metric for Refinement Level 1 | : SAD |
| ME Metric for Refinement Level 2 | : Hadamard SAD |
| Mode Decision Metric | : Hadamard SAD |
| Motion Estimation for components | : Y |
| Image format | : 1280x720 (1280x720) |
| Error robustness | : On |
| Search range | : 32 |
| Total number of references | : 1 |
| References for P slices | : 1 |
| References for B slices (L0, L1) | : 1, 1 |
| Sequence type | : IPPP (QP: I 6, P 6) |
| Entropy coding method | : CAVLC |
| Profile/Level IDC | : (100,40) |
| Motion Estimation Scheme | : EPZS |
| EPZS Pattern | : Large Diamond |
| EPZS Dual Pattern | : Extended Diamond |
| EPZS Fixed Predictors | : All P |
| EPZS Aggressive Predictors | : Disabled |
| EPZS Temporal Predictors | : Enabled |
| EPZS Spatial Predictors | : Enabled |
| EPZS Threshold Multipliers | : (1 0 1) |
| EPZS Subpel ME | : Basic |
| EPZS Subpel ME BiPred | : Basic |
| Search range restrictions | : none |
| RD-optimized mode decision | : used |
| Data Partitioning Mode | : 1 partition |
| Output File Format | : H.264/AVC Annex B Byte Stream Format |

Table III.2: Encoder parameters - HEVC

| Input File | : VideoFile.yuv |
|---|---|
| Bitstream File | : CodedFile.265 |
| Reconstruction File | : YuvFile.yuv |
| Real Format | : 1280x720 30Hz |
| Internal Format | : 1280x720 30Hz |
| Frame/Field | : Frame based coding |
| Frame index | : 0 (0 frames) |
| CU size / depth | : 64 / 4 |
| RQT trans. size (min / max) | : 4 / 32 |
| Max RQT depth inter | : 3 |
| Max RQT depth intra | : 3 |
| Min PCM size | : 8 |
| Motion search range | : 32 |
| Intra period | : 16 |
| Decoding refresh type | : 0 |
| QP | : 32 |
| Max dQP signaling depth | : 0 |
| Cb QP Offset | : 0 |
| Cr QP Offset | : 0 |
| QP adaptation | : 0 (range=0) |
| GOP size | : 4 |
| Internal bit depth | : (Y:8, C:8) |
| PCM sample bit depth | : (Y:8, C:8) |
| RateControl | : 1 |
| TargetBitrate | : 204800 |
| KeepHierarchicalBit | : 2 |
| LCULevelRC | : 1 |
| UseLCUSeparateModel | : 1 |
| InitialQP | : 0 |
| ForceIntraQP | : 0 |
| Max Num Merge Candidates | : 5 |