ROBUST QUALITY METRICS FOR ASSESSING MULTIMODAL DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BARIŞ KONUK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

MARCH 2015

Approval of the Thesis:

ROBUST QUALITY METRICS FOR ASSESSING MULTIMODAL DATA

submitted by **BARIŞ KONUK** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Gönül Turhan Sayan
Head of Department, **Electrical and Electronics Engineering** _____

Prof. Dr. Gözde Bozdağı Akar
Supervisor, **Electrical and Electronics Eng. Dept., METU** _____

Examining Committee Members:

Prof. Dr. Tolga Çiloğlu
Electrical and Electronics Engineering Dept., METU _____

Prof. Dr. Gözde Bozdağı Akar
Electrical and Electronics Engineering Dept., METU _____

Assoc. Prof. Dr. İlkay Ulusoy Parnas
Electrical and Electronics Engineering Dept., METU _____

Assoc. Prof. Dr. Banu Günel
Graduate School of Enformatics, METU _____

Assist. Prof. Dr. Gökçe Nur Yılmaz
Electrical and Electronics Eng. Dept., Kırıkkale University _____

Date: 12.03.2015

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Barış Konuk

Signature :

ABSTRACT

ROBUST QUALITY METRICS FOR ASSESSING MULTIMODAL DATA

Konuk, Barış

Ph. D., Department of Electrical and Electronics Engineering

Supervisor: Prof. Dr. Gözde Bozdağı Akar

March 2015, 124 pages

In this thesis work; a novel, robust, objective, no-reference video quality assessment (VQA) metric, namely Spatio-Temporal Network aware Video Quality Metric (STN-VQM), has been proposed for estimating perceived video quality under compression and transmission distortions. STN-VQM uses parameters reflecting the spatiotemporal characteristics of the video such as spatial complexity and motion. STN-VQM also utilizes parameters representing distortions due to compression and transmission such as bit rate and packet loss ratio. STN-VQM has been trained on the Laboratory of Image and Video Engineering (LIVE) VQA database, owned by University of Texas at Austin, and evaluated on LIVE, Ecole Politechnique Federale de Lausanne (EPFL)- Politecnico di Milano (PoliMI) and Instituto de Telecomunicacoes, Instituto Superior Tecnico (IT-IST) VQA databases and also on video streams in University of Plymouth audiovisual quality assessment (AVQA) database. STN-VQM is proven to predict perceived video quality accurately on these databases, which span a wide range of video contents, video codecs, spatial resolutions, bit rates, frame rates, packet losses etc. Comparison to the existing state-of-the-art VQA metrics indicates that the STN-VQM provides promising results. Moreover, a novel, objective, no-reference audio quality assessment (AQA) metric has been introduced in order to predict perceived audio quality under compression and transmission distortions. Proposed AQA metric appraises perceived audio quality based on parameters such as sampling frequency, bit rate and packet loss

ratio. Proposed AQA metric has been trained and evaluated on two different AQA databases. The AQA metric is shown to appraise perceived audio quality reliably on these AQA databases, which have different audio encoding types. Finally, an objective, no-reference AVQA metric (namely, Direct AudioVisual Quality Assessment – DAVQA) has been obtained by applying the classical approach in the literature, i.e., by combining perceived video quality estimate, perceived audio quality estimate and their product. Moreover, a novel video classification method which classifies videos according to their spatio-temporal characteristics has been developed. Using this spatio-temporal based video classification method, a novel, content-dependent AVQA algorithm (namely Content Dependent AudioVisual Quality Assessment – CDAVQA) has been designed. The CDAVQA model is shown to be more accurate than the DAVQA model on the audiovisual data in the University of Plymouth AVQA database.

Keywords: Quality of Experience, No-reference objective video quality assessment, No-reference objective audio quality assessment, Spatio-temporal characteristics based video classification, No-reference objective audiovisual quality assessment

ÖZ

ÇOK KİPLİ VERİ DEĞERLENDİRME İÇİN DAYANIKLI NİTELİK ÖLÇÜTLERİ

Konuk, Barış
Doktora, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. Gözde Bozdağı Akar

Mart 2015, 124 sayfa

Bu tez çalışmasında; kodlama ve iletim tabanlı bozulmalar karşısında algılanan video kalitesini kestirmek için yeni, gürbüz, nesnel referanssız, "Uzam-Zamansal Ağ farkında Video Kalite Metriği" (UZA-VKM) isimli bir video kalite değerlendirme (VKD) metriği önerilmiştir. UZA-VKM, uzamsal karmaşıklık ve hareket gibi videonun uzam-zamansal karakteristiğini yansıtan parametreler kullanır. UZA-VKM, aynı zamanda bit hızı ve paket kayıp oranı gibi kodlama ve iletim tabanlı bozulmaları temsil eden parametreler kullanır. UZA-VKM, Austin'deki Texas Üniversitesi'ne ait Görüntü ve Video Mühendisliği Laboratuarı (LIVE) VKD veri tabanında eğitilmiştir. UZA-VKM; LIVE, Ecole Politechnique Federale de Lausanne (EPFL)- Politecnico di Milano (PoliMI) ve Instituto de Telecomunicacoes, Instituto Superior Tecnico (IT-IST) VKD veri tabanları ile Plymouth Üniversitesi Odyovizüel kalite değerlendirme (OVKD) veri tabanındaki videolarda değerlendirilmiştir. UZA-VKM'nin algılanan video kalitesini; çeşitli video içeriği, video kodlayıcı, uzamsal çözünürlük, bit hızı, çerçeve hızı, paket kaybı vb. içeren bu veri tabanlarında doğrulukla tahmin ettiği kanıtlanmıştır. Mevcut en gelişkin VKD metriklerle karşılaştırma, UZA-VKM'nin umut verici sonuçlar sağladığını işaret etmektedir. Buna ek olarak; kodlama ve iletim tabanlı bozulmalar karşısında algılanan ses kalitesini tahmin etmek amacıyla yeni, nesnel referanssız ses kalite değerlendirme (SKD) metriği geliştirilmiştir. Önerilen SKD metriği, algılanan ses kalitesine örnekleme frekansı, bit hızı ve paket kayıp

oranı gibi parametrelere dayanarak değer biçer. Önerilen SKD metriği iki farklı SKD veri tabanında eğitilmiş ve değerlendirilmiştir. SKD metriğin, farklı ses kodlama çeşitlerine sahip bu SKD veri tabanlarında algılanan ses kalitesine güvenilir bir şekilde değer biçtiği gösterilmiştir. Son olarak, literatürdeki klasik yaklaşımı uygulayarak (algılanan video kalite kestirimi, algılanan ses kalite kestirimi ve bu ifadelerin çarpımlarını birleştirerek) nesnel referanssız, "Doğrudan OdyoVizüel Kalite Değerlendirme" (DOVKD) isimli bir OVKD metrik elde edilmiştir. Ayrıca, videoları uzam-zamansal karakteristiklerine göre sınıflandıran yeni bir video sınıflandırma yöntemi tanıtılmıştır. Bu uzam-zamansal tabanlı video sınıflandırma yöntemi kullanılarak, "İçerik Bağımlı OdyoVizüel Kalite Değerlendirme" (İBOVKD) isimli yeni, içerik-bağımlı bir OVKD algoritması tasarlanmıştır. Plymouth Üniversitesi OVKD veri tabanındaki odyovizüel veri için, İBOVKD modelinin DOVKD modelinden daha doğru olduğu gösterilmiştir.

Anahtar Kelimeler: Deneyimleme kalitesi, Referanssız nesnel video kalitesi değerlendirme, Referanssız nesnel ses kalitesi değerlendirme, Uzam-zamansal karakteristik tabanlı video sınıflandırma, Referanssız nesnel odyovizüel kalite değerlendirme

To İrem

ACKNOWLEDGMENTS

The author would like to express his sincere appreciation to his supervisor, Prof. Dr. Gözde BOZDAĞI AKAR for her valuable supervision, support and tolerance throughout the development and improvement of this thesis.

The author is grateful to Prof. Dr. Tolga ÇİLOĞLU, Assoc. Prof. Dr. İlkay ULUSOY PARNAS, Assist. Prof. Dr. Gökçe NUR YILMAZ, and Assoc. Prof. Dr. Banu GÜNEL for their support and advices.

The author would like to express his gratings to Emin ZERMAN for his help throughout the thesis.

The author is also grateful to ASELSAN Inc. for the resources and facilities that I use throughout the thesis.

Thanks a lot to all my friends for their great encouragement and their valuable help to accomplish this work.

The author would like to express his deepest gratitude to his parents, and his brother for their invaluable love, support, encouragement and advices.

Finally, the author feels grateful to his wife İrem for her patience and understanding. Her encouragement, support and endless love have always been the greatest assets of his life.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AQA	Audio Quality Assesment
AVQA	Audiovisual Quality Assessment
AVC	Advanced Video Coding
CDAVQA	Content Dependent Audiovisual Quality Assessment
CIF	Common Intermediate Format
DAVQA	Direct Audiovisual Quality Assessment
DCT	Discrete Cosine Transform
DMOS	Difference Mean Opinion Score
EPFL	Ecole Politechnique Federale de Lausanne
FR	Full-Reference
GOP	Group of Picture
HVS	Human Visual System
IT-IST	Instituto de Telecomunicacoes, Instituto Superior Technico
ITU	International Telecommunication Union
LIVE	Laboratory of Image and Video Engineering
MB	Macroblock
MOS	Mean Opinion Score
MSE	Mean Square Error
MV	Motion Vector
NR	No-Reference
QoE	Quality of Experience
QA	Quality Assessment
PCC	Pearson Correlation Coefficient
PEAQ	Perceptual Evaluation of Audio Quality
PESQ	Perceptual Evaluation of Speech Quality
PLR	Packet Loss Rate
PoliMI	Politecnico di Milano
PSNR	Peak Signal-to-Noise Ratio
RR	Reduced-Reference

RTP	Real-Time Transfer Protocol
SI	Spatial Perceptual Information
SROCC	Spearman Rank Order Correlation Coefficient
SSIM	Structural Similarity Index
STN-VQM	Spatio-Temporal Network aware Video Quality Metric
TI	Temporal Perceptual Information
VQA	Video Quality Assessment

CHAPTER 1

INTRODUCTION

For media delivery industry, guarantee of user experience is among the most targeted factors for many media service presented to consumers. Therefore instead of Quality of Service (QoS), the concept of Quality of Experience (QoE) has been the focused concern for media delivery industry. QoE refers to "the overall acceptability of an application or service, as perceived subjectively by the end user" [1]. Media delivery industry considers end-user QoE monitoring as either "critical" or "very important" to their video initiatives, and meanwhile the top issue reported from industry is that current QoE assessment solutions deployed today are not accurate enough and too costly to measure end user experience.

As mentioned, end user experience has become a popular research area with the increasing demand in delivery of multimedia over wired and/or wireless networks. This increasing demand along with the advent of more efficient video technologies resulted in a requirement for methods for assessing perceived video quality, which refers to end user experience for video. This requirement for the development of VQA models has become obvious after realizing that well-known and widely-used metrics such as peak signal-to-noise ratio (PSNR) has only an approximate relationship with the perceived video quality. This approximate relationship is a result of the fact that PSNR performs only a pixel-by-pixel comparison without considering what these pixels represent to human visual system (HVS). Moreover, full-reference VQA metrics such as PSNR require the reference video to be available in unimpaired and uncompressed form in order to appraise the video quality. Obviously, this requirement limits the application area of full-reference metrics in practical scenarios like video streaming. The difficulty in obtaining an objective, reference free video quality metric should be clear considering the fact that the designed metric has to take possible distortions occurring both in compression and transmission phases; including blur, motion jerkiness, blockiness, green block, frame

freeze, packet loss, re-buffering; into account. After a comprehensive survey of available video quality metrics, we started to design an objective, no-reference video quality assessment (VQA) model, which aims accurate estimation of perceived video quality considering distortions in both compression and transmission phases.

The term perceived audiovisual quality refers to end user experience for videos with audio. In fact, audio accompanying video is known to have an important impact on the perceived audiovisual quality. Hence, perceived audio quality should also be considered while estimating perceived audiovisual quality. There are many audio quality assessment (AQA) algorithms trying to estimate perceived audio quality. Among standardized AQA methods, Perceptual Evaluation of Speech Quality (PESQ) is known to produce inaccurate perceived audio quality estimates under special circumstances. Similarly, another well-known standard, Perceptual Evaluation of Audio Quality (PEAQ), which has been developed to appraise wideband audio signal's perceived quality, may fail to estimate perceived audio quality accurately in some conditions. Actually, it is said that some signal degradations may not affect audibility. Noting that the mentioned AQA methods require the reference audio signal to be available in order to decide the perceived audio quality, it should be obvious that designing a no-reference objective audio quality assessment model is a challenging task.

The general approach in appraising perceived audiovisual quality is predicting video quality and audio quality separately and then combining these quality terms with a linear combination of them and their product. However, the impacts of video quality and audio quality on perceived audiovisual quality may differ in different video contents. To illustrate, video quality may be more dominant in a soccer video whereas audio quality may be more dominant in a news video. Hence, estimating audiovisual quality even in the case that perceived audio and video quality terms are known is also a challenging task.

In addition to designing an objective, no-reference VQA model, we also tried to propose a solution for both of these challenging tasks, namely developing an

objective, no-reference AQA model and predicting perceived audiovisual quality based on video and audio quality estimations.

1.1. Major Contribution of the Thesis

In this thesis, a novel, objective no-reference VQA algorithm has been proposed. The proposed algorithm is structured on spatiotemporal characteristics of the video being analyzed, bit rate, and packet loss information. The proposed metric has been trained on the LIVE VQA database and evaluated on LIVE, EPFL-PoliMI and IT-IST VQA databases. These VQA databases contain a wide range of video contents, spatial resolutions, bit rates, frame rates and packet losses. The proposed VQA algorithm is proven to estimate perceived video quality in a robust and accurate way.

In addition, an objective no-reference AQA metric based on sampling frequency, encoding bit rate, signal-to-noise ratio and packet loss has been proposed. This metric has been trained and evaluated on two different AQA databases with different audio encoding types. The AQA metric provides promising results on these AQA databases.

Finally, objective no-reference VQA and AQA models are combined in order to obtain the perceived audiovisual quality estimate. There are two approaches followed while estimating the perceived audiovisual quality. In the first approach, the audiovisual quality is directly obtained by using video quality estimate, audio quality estimate and their product, as in the literature. The second approach focuses on spatiotemporal characteristics of the video and classifies audiovisual data according to these characteristics. For that purpose, a new classification method which is robust to distortions in both compression and transmission is proposed. Since the second approach considers video characteristics while estimating perceived audiovisual quality, it provides more accurate results than the first approach.

1.2. Scope of the Thesis

This thesis is organized as follows:

In Chapter 2, an overview of QoE concept is introduced. In addition, the classification of visual quality metrics according to the availability of the reference and employed methodologies are discussed. Moreover, challenges in the VQA model design are introduced. Furthermore, approaches in the full-reference and no-reference AQA are presented. Finally, existing audiovisual quality assessment (AVQA) models in the literature are described.

In Chapter 3, the proposed VQA model, namely Spatio-Temporal Network aware Video Quality Metric (STN-VQM), is presented. Moreover, the proposed AQA model is described. Finally, the direct and content dependent AVQA algorithms are presented.

In Chapter 4, evaluation of the quality assessment models has been described. In addition, results of the designed VQA algorithm on different VQA databases are provided. Furthermore, results of the proposed AQA model on two AQA databases are given. Finally, results of the direct and content dependent AVQA models are presented.

In Chapter 5, we present a brief summary of the thesis. Moreover, we conclude the thesis and we also mention some of the future works to be done in order to improve the performance of the proposed AVQA model.

CHAPTER 2

AUDIOVISUAL QUALITY ASSESSMENT

2.1. Video Quality Assessment

2.1.1. Perceptual-based Video Quality Metrics

Accurate prediction of perceptual video quality is getting more important with the advances in multimedia applications. With these advances, multimedia services are desired to be offered to the end users in a way that the perceived quality of multimedia services satisfies end users. Hence, the new notion Quality of Experience (QoE) has started to be the focus instead of traditional Quality of Service criteria. As the name implies, QoE is driven by end user experience such as expectation and preferences of the end user [1].

QoE has started to be used in order to characterize the application- and user-oriented quality of video and multimedia services. QoE consists of many different aspects, among which the video quality is expected to be the most important [2]. Nevertheless, impacts of these aspects make QoE a rather complicated concept. Some of the factors affecting QoE and the ways these factors contribute are listed below [2]–[4]:

- Individual preferences of the viewer on programs he desires to watch determine the attention level.
- Quality expectations of the viewer are different in different display devices' properties such as size, resolution, brightness (small/large screen CRT/LCD televisions, cinema or mobile devices).
- Technical knowledge of the viewer determines the focus of attention and quality expectations.
- Viewing environment and interaction with the display device directly affects QoE.

- If the quality and synchronization of the accompanying audio is not satisfactory, QoE diminishes significantly.

As it is seen, these factors form a multidimensional problem space in which some of these factors are very subjective. Hence, the accurate estimation of quality of digital video systems is a very complicated problem. Moreover, optimization of the perceived quality cannot be accomplished unless the perceived quality is accurately predicted. Most of the VQA models try to take a small subset of the factors above into account. They mainly try to measure the visual fidelity of the video in terms of the compression and transmission degradations. Even in this well-defined case, there are challenging issues:

- Modeling all possible distortions in video systems is a very complex problem. There are many components in video systems such as capture and display devices, codecs, streamers, routers, switches and lots of different algorithms inserted in various devices. All of these components somehow process the video. Hence, they may have an impact on the video quality. It may also be difficult to minimize and isolate influences of external factors, such as viewing environment, individual preferences and video content, which are not easily modeled in the metrics.
- Modeling visual perception is even more complex. This is due to the fact that, we perceive video quality after processing video in HVS, which is not identified clearly. In fact, only a small part of processing visual stimuli and generating visual perception in brain have been well understood. This fact indicates that perceived quality prediction requires multi disciplinary collaboration not limited to but including signal processing, cognition, and psychology. However, even well understood parts of this process are not perfectly modeled. This is because; VQA metrics focus only on some simple psychophysical mechanisms which can be modeled in a computable approach easily. To illustrate, the contrast sensitivity function which is widely used is based on spatiotemporal frequency. Nevertheless, the contrast sensitivity of human vision can be affected by many more factors including eye movement, which is the visual attention mechanism. Visual attention mechanism, which

controls the vision behavior while perceiving, is known to be an important part of the perception system. Current VQA models usually utilize a simple combination of attention map obtained with computable attention models and quality map produced in VQA metrics in order to obtain the attention-based visual quality. Hence, deep understanding of visual attention mechanism should be integrated to VQA models to be developed [5].

Moreover, designing a VQA model which can appraise a wide range of visual stimuli is even harder. To illustrate, a VQA metric designed for standard definition video signals may not be that successful in estimating perceived quality of high definition video signals. Most importantly, with the advances in multimedia applications, VQA metrics' aim is becoming satisfaction of different users instead of providing the content itself.

The most reliable and most accurate solution to this complex problem is performing subjective VQA tests in which the ground truth values are obtained by human subjects. International Telecommunication Union (ITU) has standardized a lot of subjective VQA methodologies for different application scenarios [6]. Although it is stated that the subjective VQA is the best solution, it is always time- and money-consuming and can only be used as an offline solution. These disadvantages in subjective VQA methods led to a lot of research efforts for developing reliable objective VQA metrics that can automatically appraise perceived quality reliably and in a short time [7]. Objective VQA metrics aim estimating the perceived quality as much as possible correlatively with the ground truth values, which are subjective quality results in this case.

Objective VQA models can be classified into three categories according to availability of the reference video. These categories are full-reference (FR), reduced-reference (RR) and no-reference (NR) approaches. As the name implies, FR VQA metrics have full access to the reference video, whereas, the RR VQA metrics do have access to a bunch of features extracted from the reference video. These features are expected to represent the quality characteristics of the reference signal. Obviously, these features are assumed not to be distorted while transferring them.

The same quality features are extracted from the distorted signal at the receiver side. Then the RR VQA metrics decide on quality using features extracted from the reference signal and the distorted signal. Some RR metrics also propose a solution to the problem that the quality features extracted from the reference signal may be distorted during transmission. A solution to this problem is based on embedding pseudo watermark into the signal and measuring the transmission distortion on the embedded watermark assuming that the distortion on the watermark can approximate the signal distortion. NR VQA metric is known to be the most complicated VQA type, since there is no information about the reference signal, i.e, neither the reference signal nor the quality features extracted from the reference signal. Therefore, NR VQA metrics usually try to appraise the quality degradation caused by compression and transmission distortions. For example, they may focus on blockiness which commonly appears in block-based compression schemes and transmission in error prone networks. They may also detect blurring artifacts occurring due to lossy quantization [8]. NR VQA metrics are generally trained based on subjective tests, since quality estimates of NR VQA metrics may be very difficult to interpret due to unavailability of the reference signal, which may be used as a benchmark of the highest quality. Machine learning algorithms have been widely used in order to train NR VQA models.

RR and NR VQA metrics have a much wider application area in real life applications, since they do not require the reference video, which is not available in most of the applications. However, it should be obvious that RR and NR VQA metrics have a significant disadvantage since they do not have access to the reference video, which is available in FR VQA metrics even in unimpaired and uncompressed form in some cases. Although RR VQA metrics can access some features extracted from the reference video and transferred by a secure channel, NR VQA models do not have any information about the reference video. Hence, it should be clear that designing reliable NR VQA metrics is the most challenging one among VQA model designs. Moreover, it is usually expected that NR metrics perform worse than both FR and RR metrics, since they have the least prior information. Thus, the development of reliable NR VQA models requires much more research effort than that of RR and FR VQA models.

As stated, FR metrics decide on perceived quality by comparing the perceived difference between the reference video and the distorted video. In order to do this comparison, these metrics try to model HVS. Upto now, researchers focused mostly on FR metrics and a lot of advanced FR VQA models have been introduced. Some of these FR VQA metrics are even standardized by VQEG and ITU [9]. Two well-known and widely used FR VQA metrics are mean square error (MSE) and peak signal-to-noise ratio (PSNR). However, these metrics are known not to estimate perceived visual quality correlatively with the subjective tests. The reason is that these metrics do not consider video as a visual data and the attributes of the HVS are not considered in their design [10]. Figure 2-1 illustrates two images with different perceived quality and identical PSNR values. It should be clear that the influences of distortion type and visual content on perceived visual quality must be considered while estimating perceived quality [11]. Therefore, there are many VQA metrics which perform better than both MSE and PSNR [7], since these VQA metrics are trying to simulate the viewers' visual perception mechanism. Nevertheless, HVS is not sufficiently understood as stated. Hence, simulating HVS is a very challenging task and has not been perfectly accomplished. As a result, current VQA metrics are still long away from being widely accepted and universally recognized.

Figure 2-1: Images with different perceived quality and identical PSNR [11].

Additionally, VQA metrics can also be classified into two categories based on the utilized methodologies. These categories are psychophysical and engineering approaches. In psychophysical approach, psychological quality perception process in the human brain is tried to be simulated in order to assess the perceived visual quality. For example, contrast sensitivity is a commonly utilized perceptive characteristic in order to model the visual sensitivities to various spatial and temporal frequencies along with the attentive information [12]. There are also other important factors contributing to visual quality perception such as masking effect and color perception. The strategies to imitate HVS result in the fact that psychophysical approach based VQA metrics provide high correlation with subjective tests. However, the same strategy causes these VQA models to be computationally complex. Due to the computational complexity of psychological methods, there are also engineering approaches which are based on the extraction of certain features and/or artifacts. Engineering approach based VQA methods try to estimate perceived quality based on extracted features and detected artifacts. These VQA metrics may also take the HVS into account. However, they are mainly based on visual content and degradation analysis rather than fundamental principles of HVS [13].

2.1.2. Objective Video Quality Assessment Literature

In literature, even though several FR and RR metrics are developed to measure video quality, the studies on the NR metrics are rather limited. There are basically two different non-hybrid approaches used for the NR quality metrics: artifact and network quality based. In the artifact based methods, quality is described as a function of artifacts such as blurriness, blockiness, jerkiness etc. Various methods are used to parameterize these artifacts in this metric type. In the network quality based metric type, the quality is related with parameters of compression, transmission etc.

Blocking and blurriness were used as video quality degradation metrics in early video quality researches, and also they are still used in some of recent researches. In order to acquire a score that would reflect the opinion score, the values acquired from

blocking and blurriness metrics are generally pooled and weighted using different weighting or voting algorithms. The NR metric proposed in [14] extracts blurring, blockiness, and noisiness artifacts from videos with different bit rates and computes their weighted sum as the quality score whereas the work proposed in [15] uses a different weighting policy depending on the region of interest (ROI) found by analyzing the motion characteristics of the given video. In [16] and [17], the basic principle is to extract feature maps and to feed the features extracted from the feature maps to a neural network. In [16], the feature maps are found by analyzing the DCT coefficients and kurtosis, smoothness, sharpness, and blockiness are extracted as features and temporally pooled. In [17], the image has been decomposed by Laplacian pyramids and entropy ratio, energy ratio, kurtosis ratio, MSSIM, and smoothness are taken as features.

There are also some NR VQA metrics that try to find a quality score by estimating an FR VQA metric. This FR VQA metric is generally Peak Signal to Noise Ratio (PSNR) [18], [19], [20], and in some cases it is Mean Square Error (MSE) [21], [22]. In [18], Eden proposed an NR metric that estimates the PSNR values for the Advanced Video Coding (H.264/AVC) encoded video sequences using Laplace density function. Even though good PSNR results can be observed for the I-frames, the same is not applicable for the P and B-frames in this study. Brandao proposed a similar work estimating PSNR and also Mean Opinion Score (MOS) values by the local errors in the Discrete Cosine Transform (DCT) coefficients [19], and extended that work by evaluating the developed VQA metric in a set of subjective tests [20]. The assessment results present that the developed NR metric outperforms the previous works which also rely on PSNR estimates. In [21], Naccari et al. proposed an NR VQA metric (NORM) that estimates the MSE at the Macroblock (MB) level from the packet loss ratio and utilizes that estimate in a reduced reference SSIM to obtain quality scores. Valenzise et al. extended that metric by estimating the used parameters, motion vector, prediction residual and lost MBs, without the access to the bitstream [22]. However, these two metrics lack good correlation with the subjective results and Human Visual System (HVS) due to the use of the MSE in the VQA.

Finding visibility of packet losses has also been a different endeavor for NR VQA metric research. In [23], Staelens et al. used the classification of the visibility of the packet losses based on decision tree classification for the video quality estimation. In [24], Argyropoulos et al. proposed an NR metric that considers the effect of the visibility of the packet losses in the Standard Definition (SD) and High Definition (HD) H.264/AVC sequences utilizing Support Vector Machine (SVM) classifier. This work is extended by changing the calculation scheme to Support Vector Regression (SVR) [25]. In addition to that, [21] and [22] can also be considered to find visibility of packet losses due to their calculations of lost MBs. An FR metric considering Packet Loss Rate (PLR) is utilized to develop an NR metric by training and optimization in [26]. The proposed NR metric considers PLR as well as interval between Instantaneous Decoder Refreshment (IDR) frames. As the authors suggest the proposed metric can be improved considering more parameters for the quality evaluation.

Bitstream based NR VQA metrics generally use different parameters such as bit count, quantization parameter (QP), motion vector (MV) information, frame types etc. In [27], Keimel et al. proposed an NR metric that does not require decoding the bitstream, instead it extracts features such as its per slice, average Quantization Parameter (QP) per slice, etc to estimate video quality. A subjective test is conducted and comparison with MOS scores has been made in this work. In [24] and [25], similar features are used such as number of impaired pictures, MV data, MSE estimates, and lost MB counts. Lin et al. also utilized features such as QP, MV data, and bit allocation parameters for the NR VQA metric proposed in [28].

Considering the studies discussed above, it can be clearly stated that the NR metrics reported in the literature mostly consider the visual distortion parameters such as blurring, blocking, etc. in the VQA. Moreover, the temporal relationship between frames has seldom been addressed in these metrics. In addition, only packet loss network quality parameter is utilized. A few studies focusing on using hybrid approaches for the quality estimation also exist in literature. In [29], 42 different video quality parameters are extracted from the video bit stream and Symbolic Regression, which is a machine learning technique, is exploited to determine the

most important ones among these parameters while devising an NR metric for estimating video quality of the HD H.264/AVC video sequences. The performance evaluation results point out that only a few video properties have influence on the VQA. Liu et al. proposed an NR real-time video quality monitoring metric which is called as G.1070E, which enhances the ITU-T Recommendation G.1070 [30]. The original metric G.1070 [31] uses three parameters of a video bitstream; namely bit rate, frame rate, and packet loss rate, and this metric is offline. The enhanced metric G.1070E extracts those parameters in an N frame window and also normalizes bit rate by the frame complexity and supplies those parameters to the trained G.1070 metric. Implementing a windowed approach, G.1070E can estimate quality in a real-time and online manner. In [32], Zhao et al. proposed a hybrid NR video quality indication framework for video transmission of H.264 encoded bitstreams over LTE networks. This method extracts key features at different levels (packet/frame/image level) in LTE networks after obtaining PCAP file from an LTE network node as input. Some of these key features are packet loss and packet size (packet level); frame error and frame duration (frame level); and blockiness and blur (image level). They claim that the same method is also applicable in wireline packet based video transmission networks. Even though hybrid methods provide better results compared to the non-hybrid ones, the performance results of these algorithms are shown on a limited data set.

2.2. Audio Quality Assessment

Until 1990's researchers preferred performing subjective tests in a well-controlled listening environment, such as Recommendation ITU-R BS.1116 [33], in order to assess the quality in speech and audio communications. Although subjective tests are known to reflect the perceived quality very well, they are expensive and time-consuming. Therefore, the development of objective audio quality assessment algorithms became necessary.

2.2.1. Objective Audio Quality Assessment

In the beginning, objective AQA metrics such as Signal-to-Noise-Ratio and Total-Harmonic-Distortion are used in order to assess the perceived audio quality. However, these metrics are shown to be unsuccessful in assessing the perceived quality. These metrics are especially unsuccessful when assessing the performance of non-linear and non-stationary modern codec [34]. The requirement for developing objective AQA algorithms estimating perceived quality consistent with subjective evaluations has led to several ITU standards. Among these standards, BS.1387 - Perceptual Evaluation of Audio Quality (PEAQ) is the ITU standard for audio quality [35]. PEAQ has been the only available standardized AQA model [36]. Similarly, ITU-T P.862 - Perceptual Evaluation of Speech Quality (PESQ) [37] is the corresponding speech quality assessment model. PESQ is known to estimate quality of narrow-band speech subject to various signal degradation types such as coding distortions, environmental noise and packet losses. However, PESQ is known to have certain limitations. To elaborate some of these limitations, PESQ cannot estimate the perceived quality of the speech degraded by talker echo, conversation delay, side tone and noise suppression algorithms. Due to these drawbacks, researchers proposed objective AQA models based on the classic model of PEAQ and improved by better psychoacoustic models or cognitive algorithms. Nonetheless PEAQ also has certain limitations. As a consequence, more practical and accurate objective AQA models have been proposed.

Objective AQA algorithms can be divided into two categories as full-reference AQA algorithms and no-reference AQA algorithms. As the name applies, full-reference AQA algorithms have access to the reference signal. They usually try to extract some key features from both the reference signal and the degraded signal and they measure the distortion between the reference signal and degraded version. As it is in the video case, the reference signal may, and actually will, not always be available. Hence, there is an obvious need for the development of no-reference AQA algorithms. No-reference AQA models can be further divided into two categories as signal-based models and parametric models. Signal-based no-reference AQA models try to estimate the perceived audio quality based on processing the degraded audio signal.

Parametric models, on the other hand, try to estimate the perceived quality by using the underlying transport and terminal properties including noise, speech levels and echo [38], VoIP network characteristics [39] [40], or cellular radio receptions [41]. Moreover, parametric no-reference AQA models are utilized as a network planning tool in order to estimate the perceived quality based on tabulated values of bit rate, codec type and packet loss statistics [42]. It is here worth noting that no no-reference AQA algorithm has yet been standardized by the ITU although some no-reference speech quality assessment algorithms such as ITU P.563 [43] have been developed. Hence, the study on no-reference AQA has been an open research area.

2.2.1.1. Full-reference Objective Audio Quality Assessment

Objective AQA models have some limitations. To elaborate these limitations, these models have been trained on subjective data covering only a set of distortion conditions. Moreover, voting errors in subjective test directly affect the success of the developed objective AQA model. In spite of these limitations, researchers have been trying to develop objective AQA algorithms for many years. In 1979, Schroeder et al [44] proposed the noise loudness concept and used a simple masking method to estimate the audibility of coding noise in a speech coder. By applying perceptual methods to speech codecs, he optimized codecs in terms of minimum audibility instead of mean squared error [34]. In 1987, Brandenburg [45] introduced Noise to Mask Ratio (NMR) concept, which added a simple perceptual masking model in order to measure the coding noise level with the reference signal. It is worth noting that it was the first real-time hardware implementation among the audio quality methods [35]. Although, there are other methods based on waveform difference measure [46] [47], they are not successful in estimating perceived quality. The reason is that large waveform differences such as waveform inversion or phase distortion may correspond to a little or no audible distortion. In 1985, Karjalainen proposed Auditory Spectral Difference (ASD) model [48]. In this model, there is a transformation from the time domain to a time frequency domain based on psychophysical frequency and loudness. This approach has become very successful and used in ITU-R BS.1387, ITU-T P.861 and ITU-T P.862. In the early 1990s,

Wang et al [49] proposed Bark spectral distance (BSD), a similar approach to the ASD. This model is based on calculating the mean squared Euclidean distance on a Sone scale in the Bark bands. Nonetheless, temporal masking has not been taken into account in this model.

Several models designed at the beginning of 1990s were submitted to a contest organized by ITU-R. Among these AQA algorithms, Beerends and Stemerdink's (1992) Perceptual Audio Quality Measure (PAQM) was very successful [50]. To increase the accuracy in estimations of PAQM in perceived quality, PAQM was integrated with NMR and the other submitted models. Then it was adapted into an AQA model for speech coder evaluation known as Perceptual Speech Quality Measure (PSQM), later adopted as PESQ. PSQM focused on noise during speech rather than noise in silent periods. It also used asymmetry weighting which models the increase in disturbance when uncorrelated, new time frequency components are added to the signal rather than attenuating or deleting components. In 1998, six candidates were received to publish the document ITU-R BS.1387, with audited revision ITU-R BS.1387-1 (PEAQ). Nevertheless, PEAQ seemed to be insufficient when the correlation between subjective and objective scores was computed. This fact was more obvious when PEAQ is applied in case of signals with large impairment resulting from low bitrate coding [51] and maximum of two channels [52]. It is noted that advanced version of PEAQ was not good at estimating the low bitrate scalable audio quality. Creusere et al. proposed the addition of the Energy Equalization parameter in the Advanced ITU metric and showed that the resulting performance is better than both the basic ITU metric and the Energy Equalization Approach [53].

For audio, Charles D. Creusere estimated the audio quality as a function of time. He assessed the audio quality based on a subset of PEAQ features, the structural similarity measure and the segmental SNR [54]. He also studied dynamic subjective testing methodology due to lack of suitable temporal subjective scores. Another AQA model [55] utilized fuzzy logic and it has been implemented with modifications in the cognitive stage of PEAQ.

2.2.1.2. No-reference Objective Audio Quality Assessment

Liang and Kubichek proposed the first no-reference signal-based AQA model [56] in 1994. The idea behind this approach was estimating the dissimilarity between the degraded signal and some ideal speech signals space. In this model, they first trained reference centroids from the perceptual linear prediction (PLP) coefficients [57] of undistorted speech signals. Then they used the time-averaged Euclidean distance between degraded PLP coefficients and the nearest reference centroid as a speech quality distortion. Talwar et al. introduced a Hidden Markov Model based approach [58]. Recently, Falk et al. introduced the idea to determine the deviation of distorted speech from the statistical model obtained via training undistorted speech. In this algorithm, PLP feature vectors for undistorted speech are modeled by Gaussian mixture models. Degraded speech signals were also utilized to acquire the multivariate adaptive regression splines to map the Gaussian mixture model output to absolute category rating listening quality [59] [60].

It was also assumed that biological human speech production systems cannot produce most speech quality distortions caused by telecommunication networks' speech processing systems because the human vocal tract has a limited motor mechanism. Gray devised a model in which a vocal tract model sensitive to distortions in telecommunication networks is parameterized [61]. Hekstra and Beerends introduced a model to estimate the speech quality. This model was obtained by integrating the PESQ model and a speech production model in order to exploit signal segments that can't be output of human vocal tracts [62]. Kim proposed a no-reference AQA algorithm, namely auditory model for nonintrusive quality estimation (ANIQUE), in which he did not use the speech production model directly [63]. In this approach, both peripheral and central levels of auditory signal processing are modeled to extract the perceptual modulation spectrum. Then this perceptual modulation spectrum is related to the speech production systems' mechanical limitation in order to measure the naturalness level in speech signals [64].

From 2002 to 2004, ITU-T organized a competition to standardize a no-reference signal-based AQA model. The scope of the model contained subjective tests including acoustic inputs with a wider range, broader noise types, people talking in noisy environments and network measurements. It can be said that the scope was slightly broader than ITU-T P.862 in terms of network conditions. One of the two proposals submitted was the ANIQUE model [63]. The other proposal, namely single-ended assessment model (SEAM), was a combined model based on three previous models proposed by Gray [61] and Hekstra and Beerends [62]. ANIQUE was narrowly beaten by SEAM, which was adopted as ITU-T P.563 in 2004 [43] [65]. ITU-T P.563 measures unnaturalness of speech and vocal tract, additive noise, time clipping, mutes and interruptions. Using these measurements, the intermediate speech quality estimate is computed for each degradation class. Then these intermediate speech qualities are linearly combined with 11 additional features to obtain the perceived speech quality estimate.

The conversational quality in traditional telecommunications networks subject to coding distortions (especially networks employing μ-law or A-law coding at 64 kbps) or minimal channel errors is mostly determined by round-trip delay, noise, talker echo from analog connections and changes to the speech level. There are two standardized models utilizing these features in order to estimate conversational quality; namely the call clarity index [38] and the E-model [42].

The E-model tries to estimate the conversational quality based on distortions due to low bit rate encoding, background noise, loudness, and network parameters such as packet loss and delay regarding their effect on the conversational quality. In other words, the E-model focuses on three main impairments: impairments which occur almost at the same time with the voice signal, impairments caused by delay and impairments caused by low bit rate codec and errors such as packet losses. The E-model is based on several simplifying assumptions such as order independence and linearity. Nonetheless, these assumptions are invalid in certain conditions. Hence, it is suggested that the E-model is used as a network planning tool rather than a perceived conversational quality estimation tool.

2.3. Audiovisual Quality Assessment

Audiovisual quality assessment (AVQA) is a relatively new and under-explored research area. As expected, many subjective tests have been performed in order to assess perceived audiovisual quality according to listening and viewing conditions defined in ITU-T Rec. P.800 [67], P.910 [6] and ITU-R BT-500 [66]. In addition to performed subjective tests, there are also studies investigating the impact of factors such as number of subjects, individual differences, environmental conditions and also the effect of distortion variations over time on subjective tests. Borowiak et al. [68] performed subjective tests and investigated the relation between perceived quality and degradation variations over time. They presented a long-term subjective methodology for AVQA of long duration content. By doing so, they tried to obtain an understanding of AVQA of long duration audiovisual content, which they claim to be very useful for real-life applications. Pinson et al. [69] tried to find the most suitable way to perform wide-range subjective testing for AVQA. In this study, six laboratories from 4 countries conducted a systematic audiovisual subjective testing, in which the wide range audiovisual data and the scale were kept constant. The only variable in these subjective tests were the test environments. They noted that the number of subjects and individual differences were important. They also added that other environmental factors do not have significant effect on perceived audiovisual quality. Finally, they conclude with the claim their laboratory environments represent user environment quite well.

Based on subjective tests, there are many AVQA models utilizing independently computed video quality and audio quality terms. Most of these AVQA models in the literature appraise audiovisual quality as a linear combination of video quality, audio quality and their product. In [70], audiovisual quality at very low bit rates is investigated. For this purpose, Winkler et al. conducted subjective tests based on the audiovisual data content and encoding parameters, which are typically used in mobile applications. They analyzed subjective tests' results thoroughly and considered audio-video bit budget allocation. Moreover, they proposed AVQA models based on the linear combination of video quality, audio quality and their product. Based on subjective tests, they concluded that both video quality and the

audio quality have significant impact on audiovisual quality. Nevertheless, product term is said to be more correlated to the audiovisual quality. Pinson et al. [71] also performed subjective tests and inspected many subjective test based AVQA models [72]-[81] obtained as a linear combination of video quality, audio quality and their product. They claim that the most significant contribution to overall audiovisual quality comes from the product term. They also note that audiovisual synchronization errors such as lip synchronization should be considered in AVQA models.

There are also research efforts modeling audiovisual quality as a function of video quality, audio quality and their product while considering the content of the audiovisual material. Thang et al. performed subjective test and obtained a graph-based formulated AVQA model using these subjective tests [82]. In their model, they tried to consider not only video quality and audio quality, but also the contribution of the relation between them. They applied multiple regressions for three different video contents and they claim that they take contextual factors such as usage conditions and user preferences into account. Yaodu et al. [83] conducted subjective tests in order to develop an AVQA model. Their AVQA model uses linear combination of video quality, audio quality and their product. However, utilized coefficients are different for different video contents. The results indicate both spatial and temporal impairments degrade the audio and video quality. Moreover, it is shown that the distortion type has no impact on the quality integration with interaction. However, the cross model interaction depends on the distortion type and the audiovisual content. You et al. [84] described a relative multimodal complexity analysis to compute the fusion parameter to be used in objective AVQA algorithms. They utilized Video Quality Metric (VQM) [13] as an objective full-reference VQA model and they used Perceived Evaluation of Audio Quality (PEAQ) [35] as an objective full-reference AQA metric. Based on experimental results, they conclude with the claim that content adaptive fusion parameter can result in more accurate AVQA model than the fusion parameter obtained based on subjective tests. Ries et al. [85] proposed an objective full-reference AVQA model, in which the audio content is considered, for low bit rate videos. The VQA model is based on gain and loss in the spatial activity and the change in the orientation of the spatial activity.

They divided AQA into two categories, namely speech quality and music quality. For speech quality, codec dependent auditory distance is employed. Integrated frequency distance and two other disturbance indicators have been used in order to evaluate music quality. Finally, they estimated audiovisual quality based on a second-order polynomial of two variables, namely video quality and audio quality. The coefficients of the mentioned polynomial are selected accordingly for three different contents.

There are also some studies, in which audiovisual quality is modeled in different ways. In [86], a second-degree polynomial, whose variables are video quality and audio quality, are proposed in order to assess audiovisual quality. Peregudov et al. performed subjective tests and they claimed that video quality can be expressed by an exponential curve, where the variable is video bit rate. Similarly, they claimed that the perceived audio quality may be computed by exponential curve, where the variable is audio bit rate. Having obtained the coefficients of the second-degree polynomial, they add that their model may be utilized in order to allocate bit rate efficiently in portable and mobile multimedia services. In [87], subjective tests are performed in order to see the impact of video quality and audio quality on audiovisual quality. Results indicate that both video quality and audio quality contribute to audiovisual quality. Goudarzi et al. claimed that the sum of video quality and audio quality has high correlation with the audiovisual quality. In addition, they proposed an objective AVQA algorithm based on packet error rate and frame rate. Moreover, they evaluated the performance of the AVQA method based on full-reference objective VQA metric Peak Signal to Noise Ratio (PSNR) and full-reference objective AQA metric Perceptual Evaluation of Speech quality (PESQ) and additional network/application parameters. In [88], Ries et al. conducted subjective tests in which the audiovisual material is selected from three different content types (soccer, video call and video clip). They proposed a no-reference objective AVQA model utilizing ensemble based model, whose inputs are five video features and two audio features. The proposed AVQA metric is said to be content adaptive since audiovisual materials are treated differently according to the content of the audio part, that is, the audio is speech or non-speech audio.

Finally, there are two standardized no-reference objective AVQA models, one for low resolution applications and the other one for high resolution applications. ITU - Telecommunication Standardization Sector Study Group 12 has studied the parametric non-intrusive assessment of audiovisual media streaming quality (P.NAMS) for lower resolution application area and the P.NAMS was standardized as ITU-T Recommendation P.1201.1 in October 2012 [89]. The P.1201.1 model can be utilized to predict audio, video, and audiovisual quality for mobile audiovisual media streaming services employing packet headers. The VQA part of the model is based on parameters such as video codec type, video resolution, bit rate, frame rate, packet-loss events and number of re-buffering events. The AQA part of the models considers parameters such as audio codec, bit rate, packet loss and re-buffering events. Then audiovisual quality is obtained by a relatively complex model, in which video quality and audio quality are utilized. Nevertheless, audiovisual data content is not considered while obtaining audiovisual quality from video quality and audio quality. The study in [90] is the winner model of the P.NAMS competition for the higher resolution application area. Garcia et al. conducted subjective tests covering degradations such as video and audio compression artifacts, packet loss resulting in audio frame loss, slicing and freezing (with frame skipping). They claim that the model can cope with random and burst packet losses. AQA part of the model considers degradations due to both compression and transmission. VQA part of the model takes both compression and transmission distortions along with the video content complexity into account. They obtained all coefficients utilized in the model by least-square curve fitting process. Finally, they obtained audiovisual quality as a linear combination of compression and transmission degradations of both video and audio and their product without considering audiovisual content.

CHAPTER 3

PROPOSED AUDIOVISUAL QUALITY ASSESSMENT MODEL

There are many factors affecting end user quality [91], [94], [70], [85] such as mutual interaction between video quality and audio quality [79], [92], audiovisual content, compression and transmission distortions [93]. This mutual interaction is more obvious in some cases. To illustrate, the video may freeze when an anchorman continues speaking. Most probably, this will not significantly disturb viewers, since the speech of the anchorman, which is the most important component for this particular audiovisual content, will still be available. Moreover, perceived audiovisual quality of a video clip is significantly different when the scene freezes but the song continues and when the scene freezes and there is no song involved. In consequence, video and audio not only form the multimedia stream, but they together determine the perceived audiovisual quality. In other words, perceived audiovisual quality is a combination of perceived video quality and perceived audio quality [91], [94], [85], [95]. This mutual interaction results in perceived audiovisual quality differences for audiovisual material with a dominant voice such as news, teleconference and eventually video clips [85], [91]. Finally, AVQA models considering audiovisual content are known to perform better than general AVQA models [70], [91]. Therefore, we focus on the design of an AVQA model taking audiovisual content into account by considering video content features.

3.1. Video Quality Assessment Model

The perceptual video quality is a very subjective concept for each unique observer. The perception of a viewer is based on the HVS, and there are different artifacts affecting perceived video quality related to the HVS. It is known that the perceived video quality is affected by the high frequency spatial and temporal characteristics of the presented image or video when it is considered in the HVS. Therefore, having an

understanding of the spatial and temporal complexity of the given video may be a good starting point in order to estimate perceived quality. Another factor affecting perceived video quality can be considered as the network. Effects of network disturbances are most commonly modeled by bit rate and packet loss ratio. The bit rate is responsible for reflecting distortions introduced during the compression, whereas, the packet loss ratio reflects distortions caused by the transmission. In order to take quality loss due to network disturbances into account, we find the bit rate and the packet loss ratio after determining spatiotemporal characteristics of the given video.

The block diagram of the proposed VQA model, namely Spatio-Temporal Network aware Video Quality Metric (STN-VQM), is depicted in Figure 3-1. STN-VQM consists of three main blocks; namely the feature extractor block, the feature integrator block and the quality estimator block.

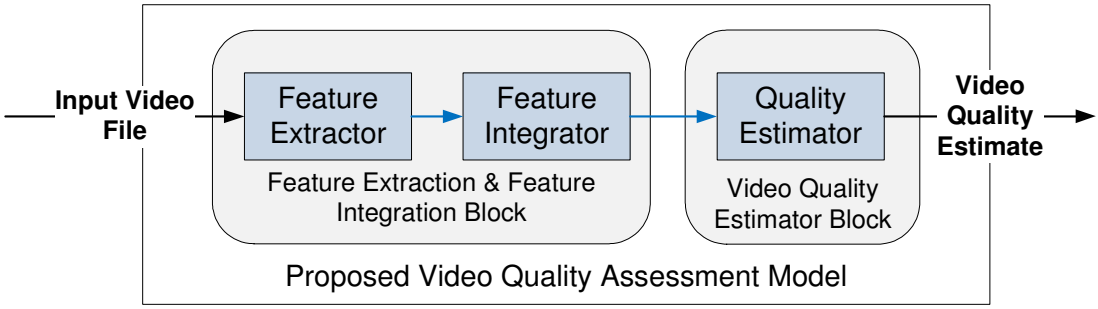

Figure 3-1: Block diagram of the proposed VQA model, STN-VQM.

3.1.1. Feature Extractor Block

The first block in the STN-VQM is the feature extractor block. The input of this block is the video bitstream whose perceived quality is desired to be assessed. This block is responsible for extracting five features as shown in Figure 3-2. These

features, which are detailed in the following subsections, are related to the spatiotemporal complexity and network disturbances.

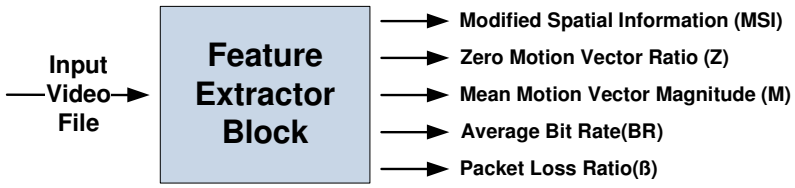

Figure 3-2: The feature extractor block.

3.1.1.1. Spatial Complexity Analysis

Although there are different methods in order to compute spatial complexity, we employed spatial perceptual information measurement (SI) index, endorsed by International Telecommunication Union (ITU) Recommendation P.910 [6] as a measure of the spatial complexity. SI is defined as the maximum value of the standard deviation in spatial extent of Sobel-filtered frames at time n, $\{F_n\}$:

$$SI = max_{time}\left\{std_{space}\{Sobel(F_n)\}\right\} \tag{3-1}$$

This definition may hinder SI value to correctly represent the spatial complexity of a video sequence, because peaks that may occur due to a scene cut and/or an erroneous frame may be the maximum value of the standard deviations over all video frames. Therefore, we replace SI with Modified Spatial Information (MSI), which we define as the average value of the standard deviation in spatial extent of Sobel filtered frames at time n, $\{F_n\}$:

$$MSI = avg_{time}\left\{std_{space}\{Sobel(F_n)\}\right\} \tag{3-2}$$

The plot of subjective DMOS against Modified Spatial Information (MSI) is shown in Figure 3-3. As the compression amount in video increases, blocking and other artifacts increase. The increase in spatial distortion results in a decrease in perceived video quality. Hence, as subjective DMOS increases, MSI slightly decreases for all video contents.

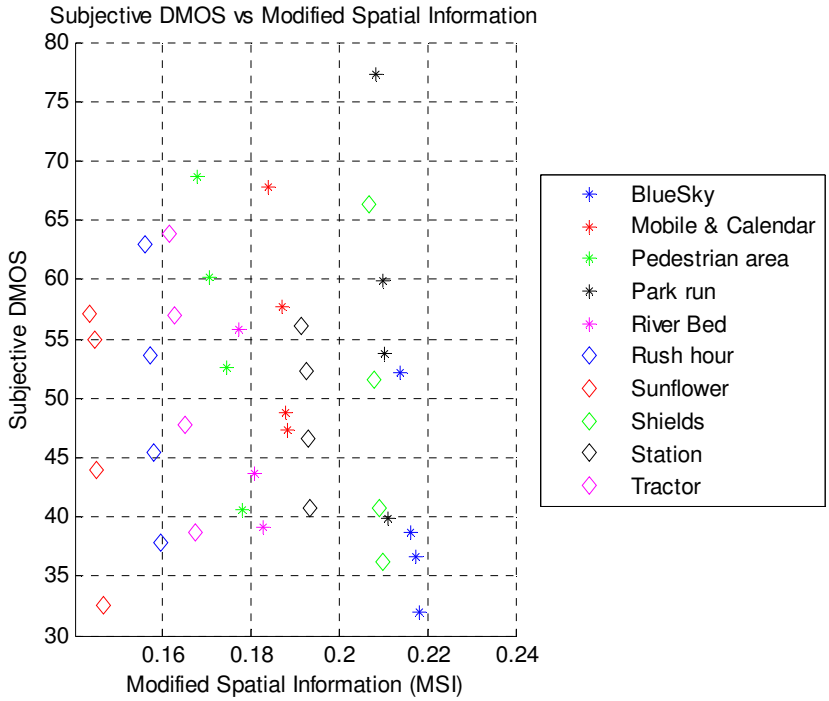

Figure 3-3: Scatter plot of subjective DMOS against Modified Spatial Information.

3.1.1.2. Temporal Complexity Analysis

Having obtained a parameter giving idea about the spatial complexity of a given video, the next step is finding parameters providing information about the temporal complexity. Hence, we focus on motion features of the video sequences. Actually, Temporal Perceptual Information (TI) may have been used while obtaining temporal complexity [6]. However, we consider that utilizing motion vectors (MV) as the

fundamental structuring element is more appropriate due to the fact that humans perceive video through segments and objects, not through pixels.

The first temporal complexity related feature is Zero MV Ratio (Z), which is defined as the percentage of zero MVs between two consecutive frames averaged over all frames in the video. Zero MV ratio for frame n, Z_n, is calculated as the percentage of MVs with the value 0 to all MVs in the frame n:

$$Z_n = \frac{count_n(MV = 0)}{count_n(MV)} \tag{3-3}$$

where $count_n$ (MV=0) represents the number of 4x4 macroblocks (MB) for which the corresponding MV equals 0 in frame n, whereas $count_n$ (MV) represents the total number of 4x4 MBs in frame n. In the H.264 standard, the MBs can have different sizes between 16x16 and 4x4. In order to avoid the unequal weighting due to the different size of MBs, all MBs other than 4x4 MBs are divided into 4x4 MBs by copying their MV displacement values to the corresponding 4x4 MBs. On the other hand, this operation is not performed on MPEG-2 encoded videos, since all MPEG-2 MBs are of equal size (16x16) and this situation does not change the result in (3-3) due to the division operation. Then Z is obtained by averaging Z_n values over all frames:

$$Z = \frac{1}{N}\sum_{n=1}^{N} Z_n \tag{3-4}$$

where N is the total number of the frames in the given video.

The plot of subjective DMOS against Zero Motion Vector Ratio (Z) is provided in Figure 3-4. As the compression amount in video increases, new distortions such as blocking are introduced. Therefore, the motion cannot be estimated as accurate as before and Z starts to increase. The increase in temporal distortion results in a decrease in perceived video quality. As a result, as the compression amount increases, both subjective DMOS and Z increase for all video contents.

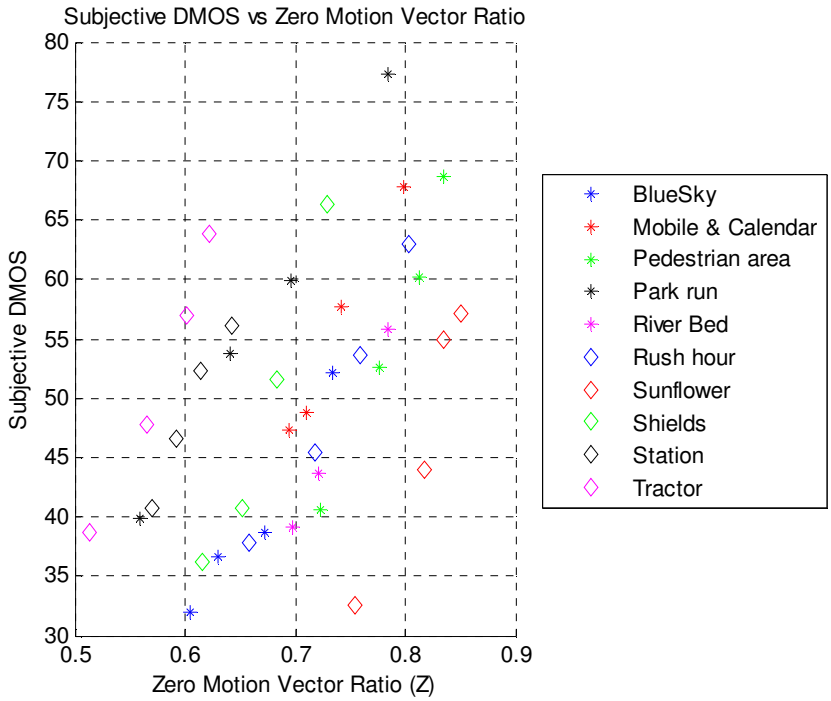

Figure 3-4: Scatter plot of subjective DMOS against Zero Motion Vector Ratio (Z).

Z helps us in estimating the proportion of the still regions in the video pictures [88]. High Z values indicate that the video is a very static sequence in which some small local movements may exist. In this case, viewer attention is expected to be on these small local movements. On the other hand, small Z values indicate uniform global movement. This global movement may be accompanied by lots of local movements. In this case, it may be hard to guess where viewers focus since it may be different for different videos. This particular MV feature makes it possible for discriminating between still sequences and frames with high amount of motion. Nonetheless, it does not distinguish between slowly and rapidly changing video sequences. In order to make this distinction, we utilize Mean MV magnitude, M, as the second temporal complexity related feature. Mean MV magnitude for frame n, M_n, is calculated by averaging normalized non-zero MV magnitudes:

$$M_n = \frac{1}{K_n} \sum_{i=1}^{K_n} \frac{mag(non-zero\ MV(i))}{w * h} \tag{3-5}$$

where K_n represents the number of non-zero MVs in frame n and w and h are the width and height of the screen in pixels, respectively. The division in (3-5) is a normalization procedure to avoid erroneous results that would arise due to different video resolutions. Then M is found by averaging M_n values over all frames, N, as shown below:

$$M = \frac{1}{N} \sum_{i=1}^{N} M_i \tag{3-6}$$

Since both Z and M are MV based features, we had to decode videos and extract MV of encoded videos. This is accomplished by modifying The Joint Model 12.3 (JM) reference software [96] for the H.264 coding standard and MPEG-2 reference software [97] for the MPEG-2 coding standard.

The plot of subjective DMOS against Mean Motion Vector Magnitude (M) is given in Figure 3-5. As illustrated in the figure, M is almost constant for all video contents and it is also independent from the compression amount. As a consequence, M is used in order to classify videos with different level of distortions according to the video content. The decrease in M in River Bed and Pedestrian area video contents is due to the fact that small local movements in these video contents cannot be accurately estimated as the compression amount increases.

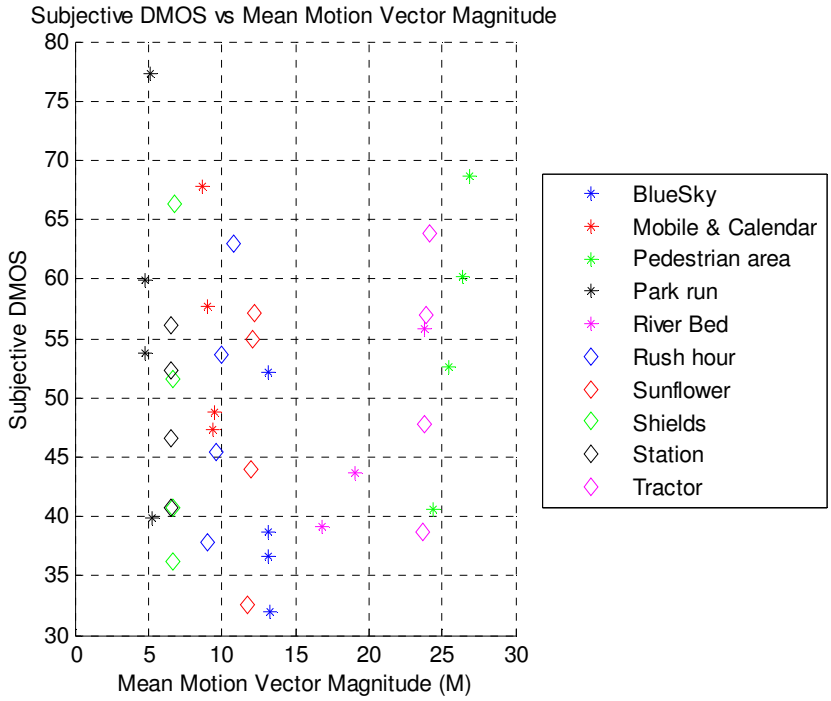

Figure 3-5: Scatter plot of subjective DMOS against Mean Motion Vector Magnitude (M).

3.1.1.3. Compression and Transmission Degradation Analysis

Compression amount and packet losses occurring in transmission are said to be significant factors affecting loss in perceived quality. Compression amount can be inferred from average bit rate parameter, BR, since bit rate is usually used as a limit during compression. Furthermore, higher bit rate values are expected to result in higher perceived quality avoiding different artifacts such as blurriness, blockiness, jerkiness etc. that may occur when lower bit rate values are utilized during compression. It is here worth noting that decrease in bit rate results in different perceived quality losses for different videos, since the videos have different spatiotemporal complexities. BR is calculated by dividing the video payload to the video duration as follows:

$$BR = \frac{Video\ Payload}{Video\ Duration\ in\ Seconds} \tag{3-7}$$

The plot of subjective DMOS against Bit Rate (BR) is given in Figure 3-6. As expected, reduction in bit rate results in new spatiotemporal artifacts which cause perceived quality to decrease. Hence, DMOS increases when bit rate decreases as illustrated in the figure.

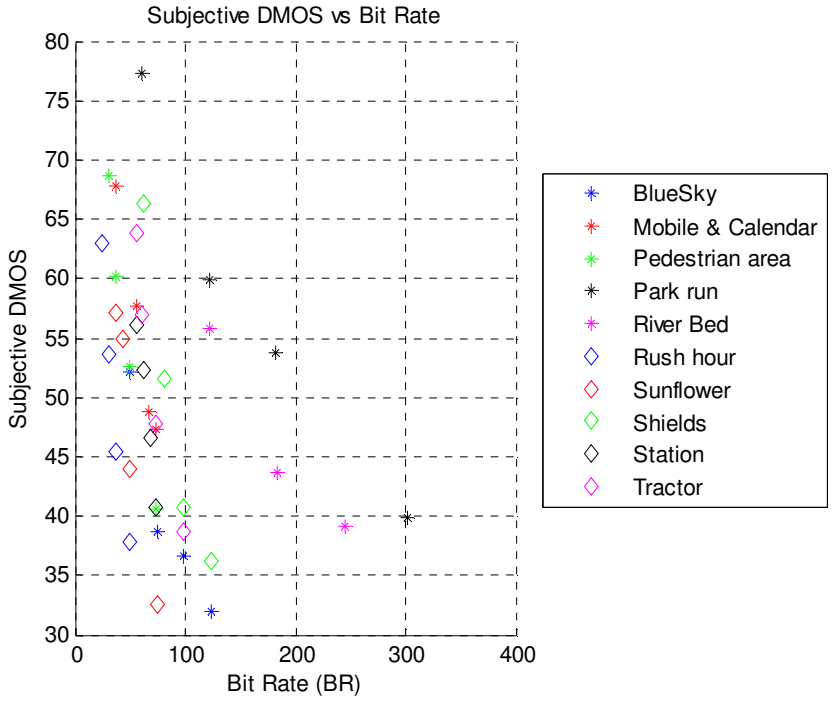

Figure 3-6: Scatter plot of subjective DMOS against Bit Rate (BR).

Having considered effects of compression amount on perceived quality, the next step is taking packet losses into account. Although there are different error concealment algorithms providing solutions to packet losses, packet losses are still important factors affecting perceived quality. Hence, packet losses should not be ignored in an accurate VQA model. We define the last feature of STN-VQM, the packet loss ratio

(PLR), β, as the percentage of the number of lost Real-Time Transfer Protocol (RTP) packets to the total number of RTP packets as:

$$\beta = \frac{Number\ of\ lost\ RTP\ Packets}{Number\ of\ total\ RTP\ Packets} * 100 \quad (3\text{-}8)$$

These five features, MSI, Z, M, BR and β, are the outputs of the feature extractor block. They are also inputs of the feature integrator block, which will be detailed below.

3.1.2. Feature Integrator Block

The second block in STN-VQM, namely the feature integrator block, is responsible for determining the parameters that are used in the quality estimator block. The outputs of this block are spatial distortion S, temporal distortion T and packet loss ratio β as illustrated in Figure 3-7.

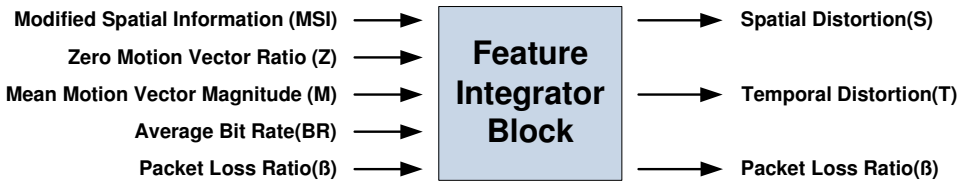

Figure 3-7: The feature integrator block.

First output of the feature integrator block, S, is computed as the ratio of MSI to BR as follows:

$$S = \frac{MSI}{BR} \quad (3\text{-}9)$$

In order to understand why S represents spatial distortion, consider K videos v_1, v_2, v_3, …, v_K encoded with the same average bit rate. Further assume that these videos have similar motion characteristics. Without loss of generality, let v_1 have the highest spatial complexity value MSI_1 and let v_K have the lowest MSI value MSI_K. Then v_1 has the highest S value and v_K has the lowest S value, since $MSI_1 > MSI_K$ and all videos are encoded with the same average bit rate. Among these videos encoded with the same average bit rate, v_1 is expected to be the most spatially distorted video, since it carries relatively more spatial information. Similarly, v_K is expected to be the least spatially distorted video, since it carries relatively less spatial information. Hence, S represents the spatial distortion amount in this case.

Consider again K videos v_1, v_2, v_3, …, v_K. This time, assume that these videos have identical spatial complexity values and similar motion characteristics. Without loss of generality, let v_1 be encoded with the highest average bit rate BR_1 and let v_K be encoded with the lowest average bit rate BR_K. Then v_1 has the lowest S value and v_K has the highest S value, since $BR_1 > BR_K$ and all videos have the same spatial complexity. Among these videos having identical spatial complexities, v_1 is expected to be the least spatially distorted video, since it is encoded with the highest bit rate Similarly, v_K is expected to be the highest spatially distorted video, since it is encoded with the lowest bit rate. Hence, S represents the spatial distortion amount also in this case.

The plot of subjective DMOS against Spatial Distortion (S) is given in Figure 3-8. It is clear in the figure that as S increases DMOS increases for all video contents. Hence, S is proven to be a distortion term for perceived video quality by experimental data.

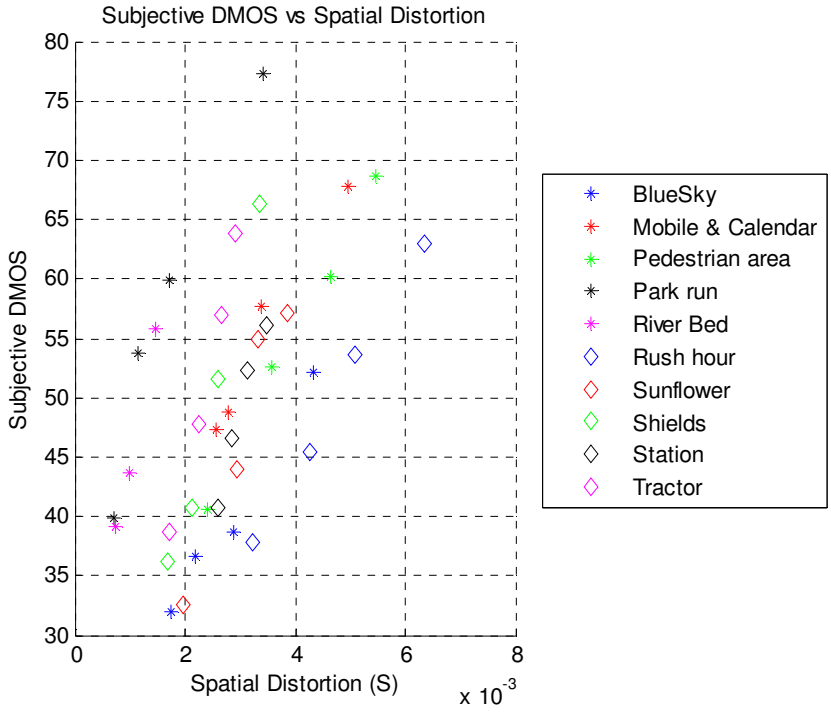

Figure 3-8: Scatter plot of subjective DMOS against Spatial Distortion (S).

Similar to the definition of spatial distortion, we define the temporal distortion, T, which is the second output of the feature integrator block, as the ratio of the amount of motion, represented by the multiplicative term (1-Z)*M, to average bit rate, BR:

$$T = \frac{(1 - Z) * M}{BR} \tag{3-10}$$

The term in the numerator can be considered as the amount of motion, since (1-Z) is the moving region proportion in a frame and M is the mean MV magnitude of this moving region. Hence the product of (1-Z) and M in (3-10) represents the amount of motion. Therefore, the multiplicative term in the numerator represent the temporal complexity of the video being analyzed.

Similar arguments why S represents spatial distortion also apply for the representation of the temporal distortion by T. As a result, T given in (3-10) is used as the temporal distortion.

The plot of subjective DMOS against Temporal Distortion (T) is given in Figure 3-9. As seen in Figure 3-9, DMOS increases as the temporal distortion increases for all video contents. However, the rate of increases in DMOS is different for different video contents.

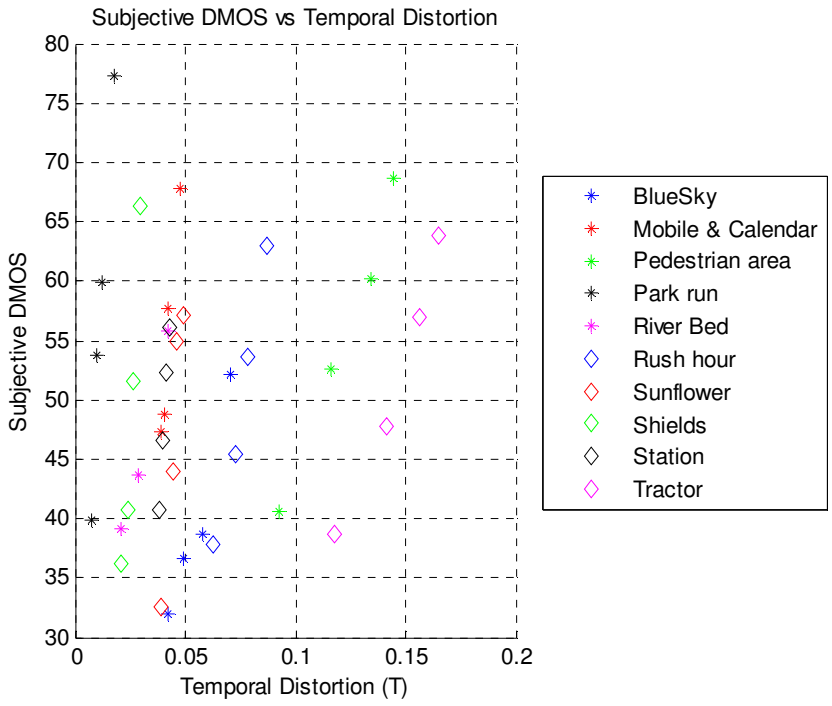

Figure 3-9: Scatter plot of subjective DMOS against Temporal Distortion (T).

It is here worth noting that the variance of T is much larger than that of S in all natural videos. Therefore, we define a dynamic upper limit for T which is computed based on the value of S as follows:

$$T = \begin{cases} T & ,if\ T < (100 * S) \\ 100 * S & ,if\ T \geq (100 * S) \end{cases} \quad (3\text{-}11)$$

Finally, the last output of the feature integrator block is β, which is also input of the same block, i.e., it is left unchanged.

3.1.3. Video Quality Estimator Block

The final block in STN-VQM is the video quality estimator block. This block, as the name implies, is responsible for estimating the perceived video quality based on S, T and β, which are inputs of this block as depicted in Figure 3-10.

Figure 3-10: The video quality estimator block.

We have employed the Laboratory of Image and Video Engineering (LIVE) VQA database in order to obtain the functional form of STN-VQM. The LIVE VQA database is a publicly available VQA database owned by University of Texas at Austin [98], [99]. In LIVE VQA database, there are 10 different reference videos with various video contents (one frame for each video content is illustrated in Figure 3-11) and 15 distorted videos for each of these reference videos. Therefore, there are 150 distorted videos, which are distorted by 4 different distortion processes; namely, H.264 compression, MPEG-2 compression and simulated transmission of H.264 compressed bitstreams through error prone IP and wireless networks.

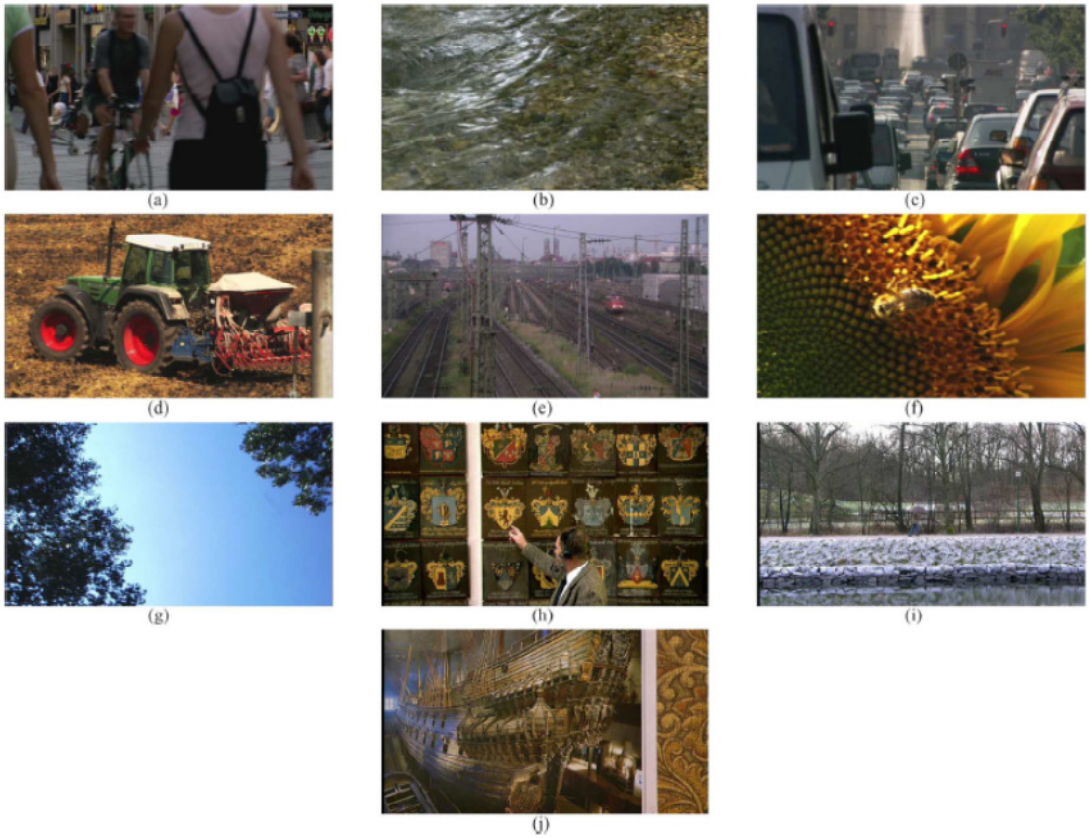

Figure 3-11: One frame from each of the 10 video contents. a) Pedestrian Area, b) River Bed, c) Rush Hour, d) Tractor, e) Station, f) Sunflower, g) Blue Sky, h) Shield, i) Park Run, j) Mobile & Calendar [99].

Blue sky video has duration of 8.68 seconds, whereas remaining 9 videos have duration of 10 seconds. Park Run, Shields and Mobile & Calendar have a frame rate of 50 frames per second and the remaining seven videos have a frame rate of 25 frames per second. Brief information about these video contents is given in Table 3-1:

Table 3-1: Brief information about 10 video contents in LIVE VQA database

Video Name	***Camera Motion***	***Video Content***
Blue Sky	Circular camera motion	Trees and blue sky
River Bed	No camera motion	River bed with some pebbles and wavy water
Pedestrian Area	No camera motion	People walking and cycling at a street intersection
Tractor	Camera pan	A green tractor moving across Fields
Sunflower	No camera motion	A bee flying over a sunflower closely
Rush Hour	No camera motion	Rush hour traffic
Station	No camera motion	A railway track, a train, and people walking across the track
Park Run	Camera pan	A person running across a park
Shields	First camera pan, then still camera, then zoom in	A person with headphone moving across a display while pointing at it
Mobile & Calendar	Camera pan	A horizontally moving toy train with a vertically moving calendar in the background

We strongly believe that distortions introduced in compression and distortions resulting from during transmission do not affect each other, i.e., they are considered to be independent. As a result, these distortions are treated separately.

First, we focused on obtaining an expression which tries to estimate the perceived quality reduction caused by compression distortions only. This expression totally ignores packet losses occurring during transmission and treats all videos as if there is no packet loss. In order to obtain this expression, we randomly selected 10 H.264 compressed videos in the LIVE VQA as the training videos. Using these training videos and the Curve Fitting Toolbox of Matlab, we fit a second degree polynomial function whose inputs are the spatial distortion, S, and temporal distortion, T. The

output of the mentioned polynomial is the DMOS estimate of the particular video bitstream, denoted by $DMOS_{initial}$:

$$DMOS_{initial}(S,T) = a + bS + cT + dS^2 + eST + fT^2 \quad (3\text{-}12)$$

Table 3-2: Coefficients for compression distortion

Coefficient Name	***Value***
a	45.6
b	8.2×10^3
c	-590
d	3.97×10^5
e	-5.04×10^4
f	4.2×10^3

The coefficients of the expression in (3-12) are provided in Table 3-2. The initial functional form of STN-VQM given in (3-12) does not correctly express perceived quality reduction caused by compression distortions. The reason is that the expression in (3-12) is a quadratic equation in both S and T. In order to elaborate the problem originating from the quadratic structure of (3-12), assume there are K different videos (v_1, v_2, …, v_K) with identical spatial distortion values ($S = S_o$) and with different temporal distortion values (T_1, T_2, …; T_K). Without loss of generality, further assume that $T_1 < T_2 < \ldots < T_{K-1} < T_K$. It should be clear that, v_1 is expected to have the smallest DMOS estimate since these K videos have the same spatial distortion S_o and v_1 has the smallest temporal distortion T_1. Consequently, v1 is expected to be perceived as the highest quality among these K videos. We can replace S by S_o in (3-12) since each video v_1, v_2, …, v_K is assumed to have the same S value, S_o. Then the expression (3-12) simplifies to:

$$DMOS_{initial}(S_o,T) = fT^2 + \gamma(S_o)T + \lambda(S_o) \quad (3\text{-}13)$$

where $\gamma(S_o)$ and $\lambda(S_o)$ are given in (3-14) and (3-15), respectively:

$$\gamma(S_o) = c + eS_o \tag{3-14}$$

$$\lambda(S_o) = a + bS_o + dS_o{}^2 \tag{3-15}$$

Clearly, expression in (3-13) can be replaced with (3-16):

$$DMOS_{initial}(S_o, T) = \left(\sqrt{f}T + \frac{\gamma(S_o)}{2\sqrt{f}}\right)^2 + \left(\lambda(S_o) - \frac{\gamma(S_o)^2}{4f}\right) \tag{3-16}$$

Obviously, the expression in (3-16) is minimum when:

$$T = T_{\min}(S_o) = \frac{-\gamma(S_o)}{2f} = -\frac{c + eS_o}{2f} \tag{3-17}$$

Reconsider the videos $v_1, v_2, \ldots, v_K$. Now assume that v_i $(1 < i \leq K)$ has a temporal distortion value $T_i \leq T_{min}$. Since $T_1 < T_i$, $T_1 < T_{min}$. Therefore, we know that v_1 has identical spatial distortion with v_i and it has less temporal distortion than v_i. Then, $DMOS_{initial}$ estimate of v_1 should be lower than that of v_i, since v_1 has less distortion than v_i and v_1 is expected to be perceived as higher quality. Nonetheless, if the expression in (3-12) is used without any modification, $DMOS_{initial}$ estimate of v_1 becomes higher than that of v_i. The reason is that the expression in (3-12) suffers from its quadratic structure and it should be corrected. This is depicted in Figure 3-12.

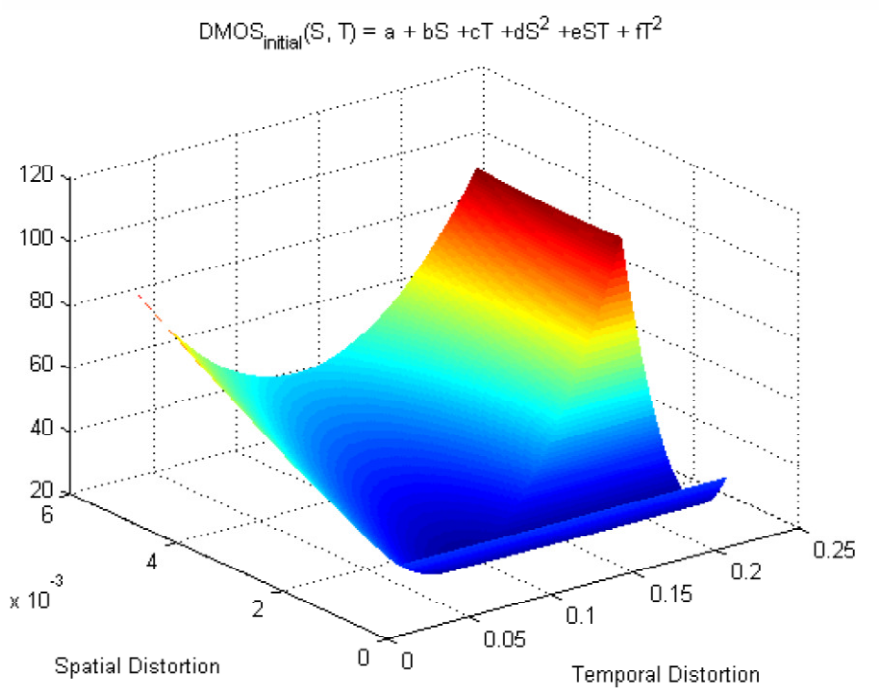

Figure 3-12: The characteristics of $DMOS_{initial}$ with respect to spatial and temporal distortion in three dimensions.

It may be easier to visualize the mentioned problem in two dimensions. Figure 3-13 shows the characteristics of $DMOS_{initial}$ when the upper limit in (4-11) is not applied to T. Figure 3-14 illustrates the two-dimensional projection of Figure 3-12.

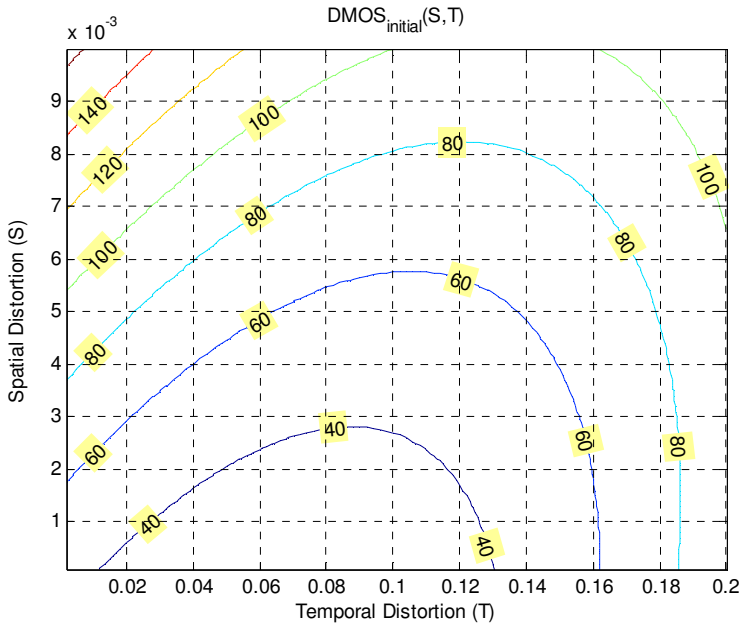

Figure 3-13: The characteristics of $DMOS_{initial}$ with respect to spatial and temporal distortion in two dimensions when T is not upper limited as in (3-11)

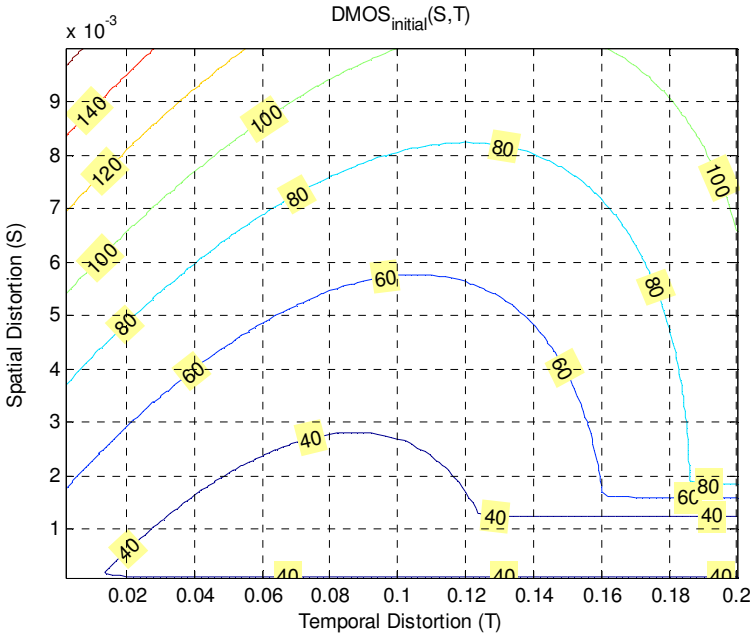

Figure 3-14: The characteristics of $DMOS_{initial}$ with respect to spatial and temporal distortion in two dimensions.

In order to correct the misevaluation detailed above, we first insert T_{min} in (3-12) instead of T_1. Nevertheless, this is not enough, since $DMOS_{initial}$ estimate of v_1 and v_i become equal. However, we know that v_1 is less distorted than v_i. Hence, $DMOS_{initial}$ of v_1 should be less than that of v_i. This is accomplished by multiplying $DMOS_{initial}$ estimate with a correction function, $h_{CRF}(T, T_{min})$, satisfying the following conditions:

$$i.\, h_{CRF}(T, T_{min}) = 1 \quad , if\ T = T_{min} \tag{3-18}$$

$$ii.\, 0 \leq h_{CRF}(T_j, T_{min}) < h_{CRF}(T_m, T_{min}) < 1 \ , if\ 0 \leq T_j < T_m < T_{min} \tag{3-19}$$

Conditions above state that the correction function, $h_{CRF}(T, T_{min})$, should be a monotonic decreasing function when $T < T_{min}$. It is clear that the expression in (3-20) satisfies both conditions when $\alpha > 0$:

$$h_{CRF}(T, T_{min}) = \left(\frac{T}{T_{min}}\right)^{\alpha} \tag{3-20}$$

We have observed the correction function gives best results when $\alpha = 0.05$. The plot of the correction function is shown in Figure 3-15.

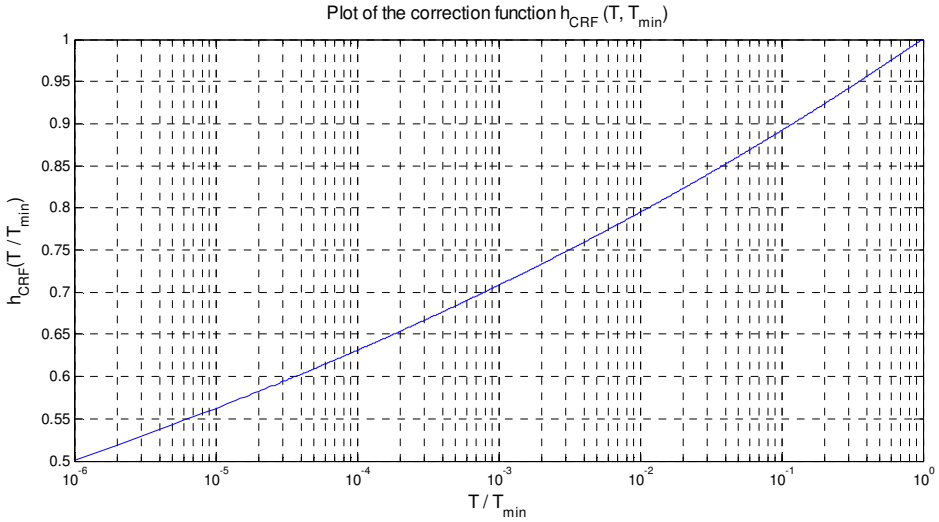

Figure 3-15: Plot of the correction function $h_{CRF}(T, T_{min})$.

It is here worth noting that S has a much smaller variance than T. Therefore, there is no need to define a similar correction function for S.

Then the DMOS estimate of H.264 compressed bitstreams can be found as:

$$DMOS_{H264}(S,T) = \begin{cases} DMOS_{initial}(S,T), & ,if\ T > T_{\min}(S) \\ DMOS_{initial}(S,T_{min}(S)) * h_{CRF}\left(T,T_{min}(S)\right), & if\ T \le T_{\min}(S) \end{cases} \quad (3\text{-}21)$$

The problem originating from the quadratic structure of (3-12) has been solved in (3-21) as depicted in Figure 3-16 and Figure 3-17.

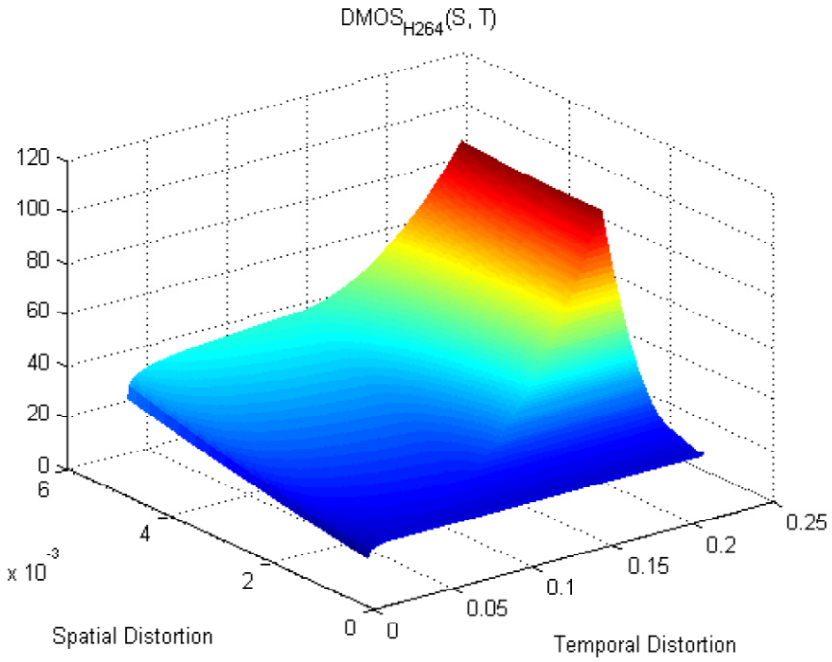

Figure 3-16: The characteristics of $DMOS_{H264}$ with respect to spatial and temporal distortion in three dimensions.

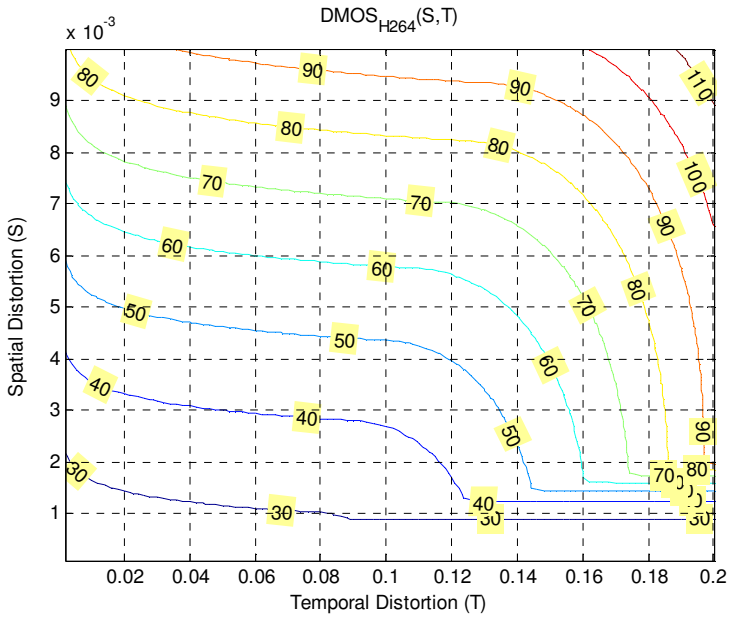

Figure 3-17: The characteristics of $DMOS_{H264}$ with respect to spatial and temporal distortion in two dimensions.

We have observed that DMOS estimates obtained by using the expression in (3-21) along with the coefficients in Table 3-2 resulted in an offset for MPEG-2 compressed videos. In order to remove this offset in DMOS estimates of MPEG-2 compressed videos, a second training step has been performed. In this second training step, we randomly selected 10 MPEG-2 compressed videos in the LIVE VQA database. After training, we have inserted a first order polynomial multiplicative term to $DMOS_{H264}$ in order to have $DMOS_{comp}$ expression, which is the final functional form of STN-VQM regarding only compression distortions. After inserting this multiplicative term, we obtained the VQA metric estimating DMOS accurately for videos with Z values below 0.01, which is the case for MPEG-2 compressed videos in the LIVE VQA database. The functional form of $DMOS_{comp}$ is computed as follows:

$$DMOS_{comp}(S,T) = \begin{cases} 0.97 * DMOS_{H264}(S,T) - 5.18, & if\ Z < 0.01 \\ DMOS_{H264}(S,T) & ,if\ Z \geq 0.01 \end{cases} \quad (3\text{-}22)$$

Having determined $DMOS_{comp}$, we have completely modeled STN-VQM for compression distortions. In order to take transmission distortions into consideration, we concentrated on finding an expression reflecting the perceived quality reduction due to transmission distortions. As it is seen from the block diagram of the video quality estimator block (Figure 3-10), the packet loss ratio parameter, β, is the parameter utilized to model transmission distortions' effect on perceived quality.

As mentioned before, compression and transmission distortions are considered to be independent. Therefore, the final form of STN-VQM, DMOS, is modeled as a multiplication of two independent functions, one of which considers only compression distortions ($DMOS_{comp}$) and the other one considers only transmission distortions (h_{TR}) as shown below:

$$DMOS(S,T,\beta) = DMOS_{comp}(S,T) * h_{TR}(\beta) \quad (3\text{-}23)$$

In order to finalize the design of STN-VQM, we have to obtain the functional form of $h_{TR}(\beta)$. To accomplish this, a third training step has been performed. In this third training step, 10 IP and 10 wireless network distorted videos have been utilized. Since IP and wireless networks have different network characteristics, IP and wireless network distorted videos have been treated separately. In this training step, we extracted the relation between the ratio of subjective DMOS to $DMOS_{comp}(S, T)$ and the packet loss ratio, β. This relation is illustrated in Figure 3-18.

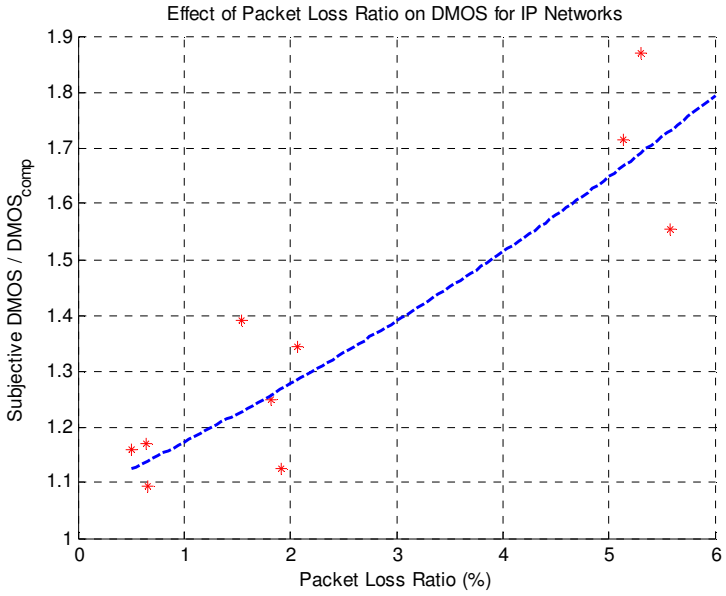

Figure 3-18: The relation between the ratio of subjective DMOS to $DMOS_{comp}$ and the packet loss ratio, β, for IP network distorted training videos in LIVE VQA database.

We have observed that there is no scene cut in reference and distorted video sequences of all video contents in LIVE VQA database. As a result, we ignored the locations of network losses (whether the packet loss is in I or P frames/slices) while estimating perceived quality for videos in LIVE VQA database since we expect all packet losses to have a similar effect on perceived quality. However, this assumption may be objected because the frame type in which the packet loss occurred is

expected to be significant in the perceived quality. Particularly, we consider that the location of network distortion is critical for estimating perceived quality of videos having scene cuts, which is not the case for videos in LIVE VQA database.

We utilized the Curve Fitting Toolbox of Matlab in order to obtain $h_{TR}(\beta)$. The functional form of the expression modeling the effect of distortions occurring during transmission, h_{TR}, to the expression modeling the effect of compression distortions while considering the video content, $DMOS_{comp}$, is the same for both IP and wireless network distorted videos. However, the coefficients of h_{TR} in (3-24), given in Table 3-3, are different for the IP and wireless network distorted bitstreams since these networks have different characteristics. Hence, appropriate coefficients should be used while evaluating h_{TR} according to the employed network type (IP or wireless). Finally, it is here worth noting that h_{TR} will be equal to unity for all videos which are transmitted without packet loss. Obviously, this statement also covers the H.264 and MPEG-2 compressed videos in LIVE VQA Database, since their packet loss ratio is zero.

$$h_{TR}(\beta) = \begin{cases} m * exp(n * \beta), if\ \beta > 0 \\ 1 \qquad\qquad\qquad , if\ \beta = 0 \end{cases} \tag{3-24}$$

Hence, the final functional form of the STN-VQM, considering both compression and transmission distortions has been obtained.

Table 3-3: Coefficients for transmission distortion

Distortion Type	***Coefficient Name***	***Value***
IP Network	m	1.38
IP Network	n	0.05
Wireless Network	m	1.08
Wireless Network	n	0.09

3.2. Audio Quality Assessment Model

No-reference audio quality assessment model is a research area that has not been studied much. Most of the researchers working in this area concentrate on developing full-reference AQA models. In full-reference AQA models, encoding bit rate and sampling frequency are two important parameters that have been used either directly or indirectly in the literature.

We first concentrated on these two parameters and inspected the relationship between the perceived audio quality and these parameters. We employed the AQA database detailed in [70]. The audio contents in the AQA database are given in Table 3-4.

Table 3-4: Audio content in the AQA database

Name	***Audio Content***	***Duration (seconds)***
Building	Orchestral background music	7.48
Conversation	Male and female voices	8.36
Football	Crowd cheering and chanting; female commentator	7.60
Music video	Rock music with vocals	8.08
Trailer 1	Theme music and voice-over	8.84
Trailer 2	Theme music and voice-over	8.08

Figure 3-19 shows the scatter plots of perceived audio quality against the sampling frequency of the corresponding audio signal. It is seen from Figure 3-19 that the perceived audio quality increases as the sampling frequency of the audio signal increases. However, the perceived quality starts to saturate after a certain value of the sampling frequency.

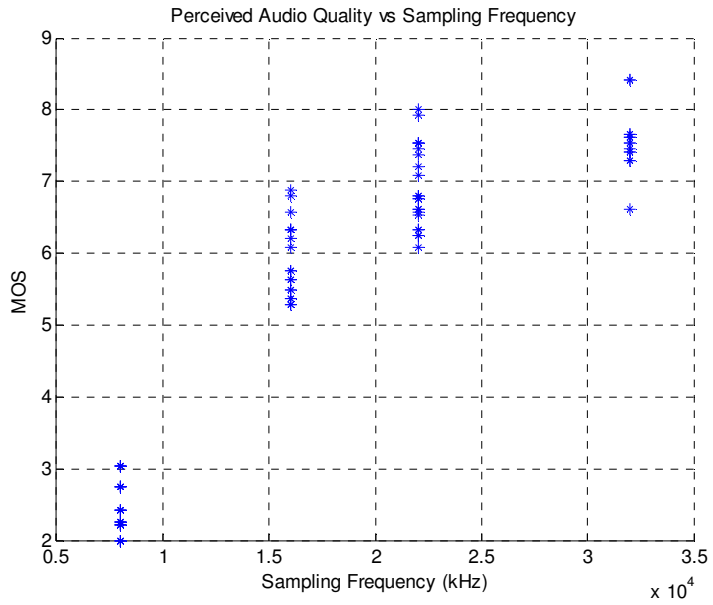

Figure 3-19: Perceived audio quality against sampling frequency.

Figure 3-20 shows the scatter plots of perceived audio quality against the encoding bitrate of the corresponding audio signal. It is depicted in Figure 3-20 that the perceived audio quality increases as the encoding bit rate of the audio signal increases.

Based on these observations, we decided to use the sampling frequency and the encoding bit rate in our AQA model. We also added another parameter signal-to-noise ratio, which is commonly used in full-reference AQA models. Using these three parameters, we obtained the functional form of the AQA metric (considering compression distortions only), $MOS_{audio,comp}$, by performing a curve-fitting procedure on Matlab. It is worth noting that the AQA metric whose functional form is given below discards distortions occurring in the transmission:

$$MOS_{audio,comp} = (a * ((b * f_s)^c) + d) * (e * BR + f) * (g * SNR + h) \quad (3\text{-}25)$$

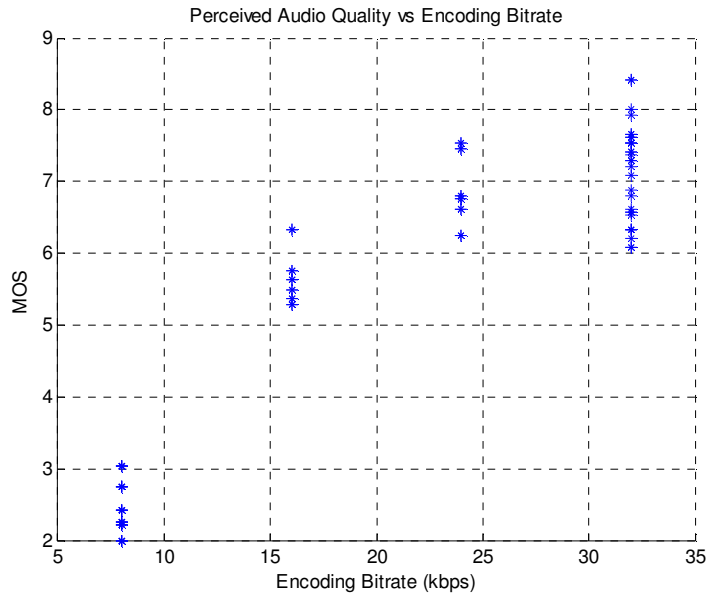

Figure 3-20: Perceived audio quality against encoding bit rate.

where f_s, BR and SNR represent the sampling rate in kHz, encoding bit rate in kbps and the signal-to-noise ratio of the audio signal in dB, respectively.

The coefficients of the expression in (3-25) are provided in Table 3-5:

Table 3-5: Coefficients for compression distortion

Coefficient Name	***Value***
a	-1325
b	15.63
c	-1.10
d	9.04
e	0.005
f	0.86
g	0.003
h	0.8

As mentioned, the AQA metric in (3-25) does not consider packet losses that may occur during transmission. The reason is that the employed AQA database in [70] does not contain packet losses. In order to take packet losses into account, we had to train the AQA metric on audio data of the AVQA database provided by University of Plymouth [87]. Audio files in this database are encoded with G.711μlaw voice codec. There are packet losses which occur in the wireless segment of the network using a Gilbert-Elliot model with packet error rates, 0.01, 0.05, 0.1, 0.15 and 0.20.

Network trunks and low-power edge devices are known to carry very large number of active calls. Therefore, real-time speech quality monitoring models are trying to compute degradation amount from RTP transport instead of processing each speech waveform [34]. We followed a similar procedure and we investigated the relationship between the perceived audio quality and packet loss rate obtained from RTP transport as illustrated in Figure 3-21.

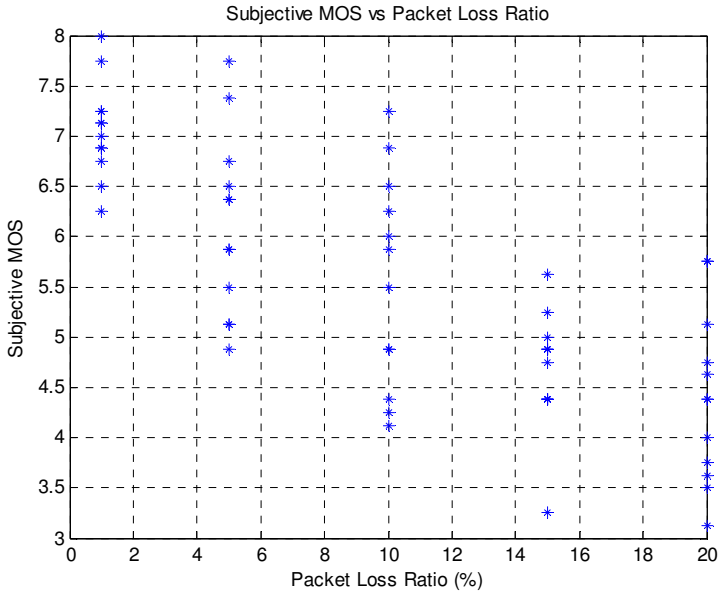

Figure 3-21: Perceived audio quality against packet loss ratio.

As expected, subjective MOS decreases as the packet loss ratio increases. Similar to the VQA case, we inserted a multiplicative exponential term in order to consider transmission distortions. The inserted function, $g_{TR}(\beta)$, is given below:

$$g_{TR}(\beta) = \begin{cases} 0.66 * exp(-0.026 * \beta), if\ \beta > 0 \\ 1 \qquad\qquad\qquad\qquad\quad , if\ \beta = 0 \end{cases} \tag{3-26}$$

Plot of g_{TR} against packet loss rate is depicted in Figure 3-22.

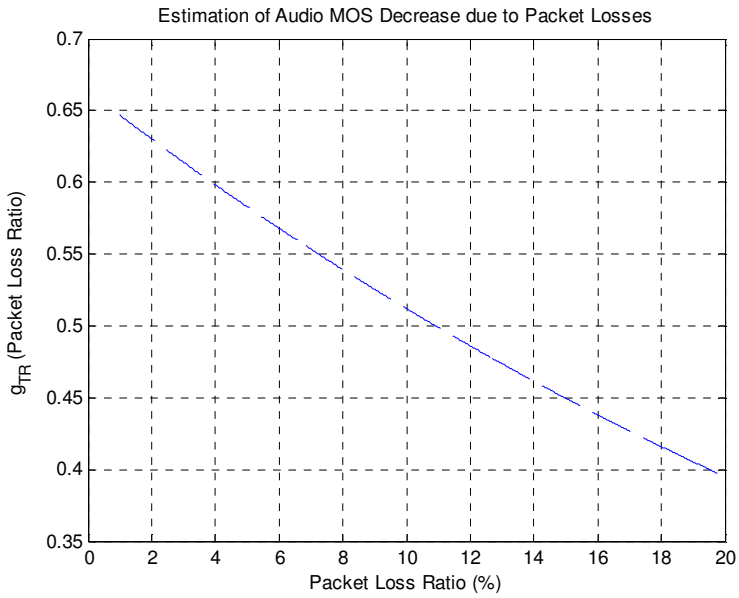

Figure 3-22: The multiplicative exponential term, $g_{TR}(\beta)$

Hence, the final form of the AQA metric is obtained as:

$$MOS_{audio} = MOS_{audio,comp} * g_{TR}(\beta) \tag{3-27}$$

3.3. Audiovisual Quality Assessment Model

Audiovisual quality assessment is generally considered as a combination of video and audio quality assessment. First we followed a similar approach, which we named

as Direct Audiovisual Quality Assessment (DAVQA) Model, to the studies in the literature. In this approach, which is detailed in 3.3.1, we have not considered video and audio characteristics while assessing audiovisual quality. However, it is stated in [88] that video content is very important for perceived audiovisual quality. For certain audiovisual contents such as news and video call, the dominance of audio quality on audiovisual quality is observed. Moreover, as the picture content of the video gets more complex and/or the motion in the video increases, video quality is said to have more effect on perceived audiovisual quality. Inferring from these observations, we proposed a second approach, which we named as Content Dependent Audiovisual Quality Assessment (CDAVQA) Model, based on estimating audiovisual quality according to the video content. Details of CDAVQA are provided in 3.3.2.

It is here worth mentioning about the AVQA database used in this study. We utilized the AVQA database of University of Plymouth. This AVQA database consists of 60 audiovisual samples with 6 different video contents. These 6 video contents are representative of video call conditions with different spatial complexities and low motion at QCIF spatial resolution (176x144). Figure 3-23 and Table 3-6 illustrate the content of the audiovisual material in the AVQA database. Audiovisual data in the AVQA database are 7 to 14 seconds long. The source audio material was 16-bit PCM mono sampled at 8 kHz. Video files are encoded with H.263 with frame rates either 8 or 15 and audio files are encoded with G.711μlaw voice codec as mentioned. These encoders are said to be selected because of their low complexity and popularity among video-conferencing applications and SIP clients such as x-lite and IMS-communicator. There are packet losses which occur in the wireless segment of the network using a Gilbert-Elliot model with packet error rates, 0.01, 0.05, 0.1, 0.15 and 0.20 [87].

Table 3-6: Audiovisual data content in University of Plymouth AVQA database [87]

Video Name	***Video***	***Audio***
Lecture	Male speaker, head and shoulder, light background	Normal male voice
Job Interview	Female speaker, head and shoulder, light background	High-pitched female voice
CBS News	News speaker, colored background	Normal female voice
Newspart	TV presenter, hand movements, colored background	Fast speaking male voice
Gold	Lecturer behind a desk, reading from a note	Normal male voice
Conversation	A male and a female in front of camera, male speaking, hand movement	Normal male voice

Figure 3-23: One frame from each of the 6 video contents.

Plots of subjective audiovisual MOS with respect to subjective video MOS and subjective audio MOS in the AVQA database of University of Plymouth are given in Figure 3-24 and Figure 3-25.

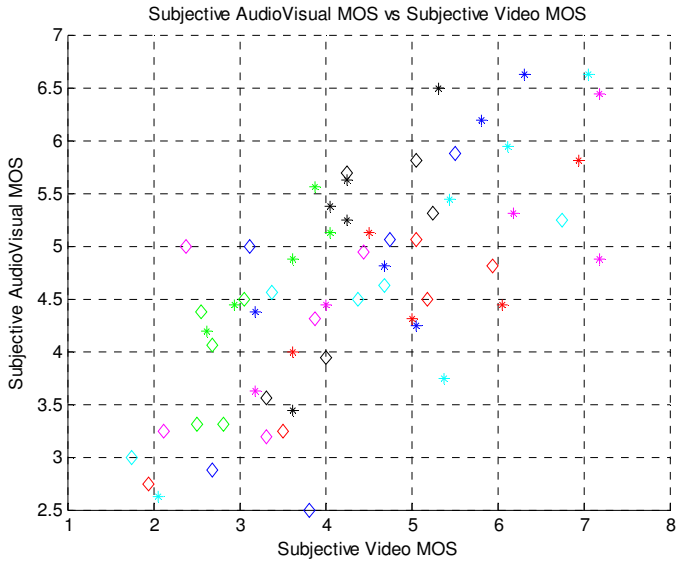

Figure 3-24: Subjective audiovisual quality against subjective video quality.

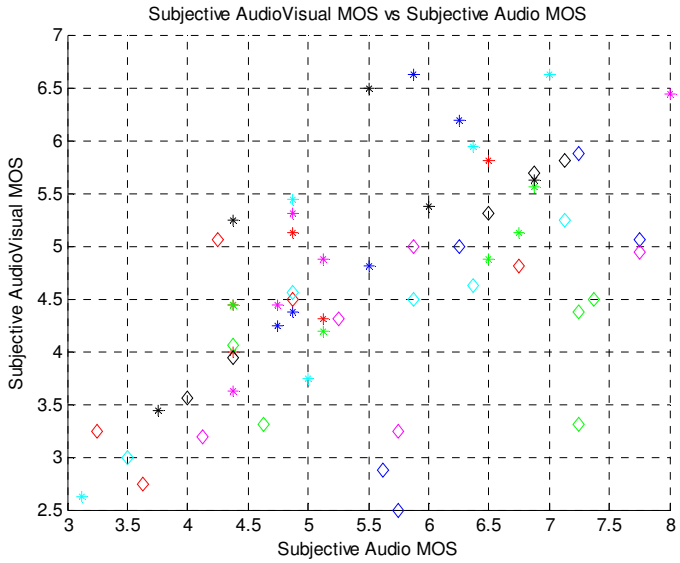

Figure 3-25: Subjective audiovisual quality against subjective audio quality.

As seen from the figures above, audiovisual quality increases as video quality increases. Although there is a similar tendency in the audiovisual quality with respect to audio quality, this tendency is less clear in this case. In other words, there are more outliers in the audiovisual quality versus audio quality graph, meaning that audio quality is less dominant than video quality in determining audiovisual quality.

3.3.1. Direct Audiovisual Quality Assessment (DAVQA) Model

In this approach, we modeled audiovisual quality estimate, MOS_{AV}, as a linear combination of video quality estimate, MOS_V, and audio quality estimate, MOS_A, and their product, $MOS_V.MOS_A$ as shown below:

$$MOS_{AV} = a + b * MOS_V + c * MOS_A + d * MOS_V * MOS_A \quad (3\text{-}28)$$

where video quality estimate, MOS_V, is obtained via subtracting $DMOS_V$ from 100 to convert DMOS to MOS and then normalizing to quality range 0-8 as below:

$$MOS_V = 0.08 * (100 - DMOS_v) \quad (3\text{-}29)$$

In order to obtain coefficients given in (3-28), we performed training using both subjective video MOS with subjective audio MOS and objective video MOS with objective audio MOS along with the subjective audiovisual MOS. The coefficients of the expression in (3-28) obtained using subjective video MOS with subjective audio MOS are provided in Table 3-7. The coefficients of the same expression obtained using objective video MOS with objective audio MOS are given in Table 3-8.

Table 3-7: Coefficients for DAVQA when subjective video MOS and subjective audio MOS are utilized

Coefficient Name	**a**	**b**	**b**	**d**
Value	1.1111	0.4156	0.3109	0

Table 3-8: Coefficients for DAVQA when objective video MOS and objective audio MOS are utilized

Coefficient Name	**a**	**b**	**b**	**d**
Value	2.216	0.4965	0	0.1358

Having obtained coefficients, we checked whether these coefficients are meaningful by checking audiovisual MOS estimates for different video and audio MOS values. Figure 3-26 illustrates the audiovisual MOS estimates for various video and audio MOS values when the coefficients in Table 3-7 are utilized. Similarly, Figure 3-27 illustrates the audiovisual MOS estimates for various video and audio MOS values when the coefficients in Table 3-8 are utilized.

As seen from figures below, as video MOS increases, audiovisual MOS increases in constant audio MOS case. Similarly, as audio MOS increases, audiovisual MOS increases in constant video MOS case. These results indicate that (3-28) along with coefficients given in both Table 3-7 and Table 3-8 provide reasonable audiovisual quality estimates.

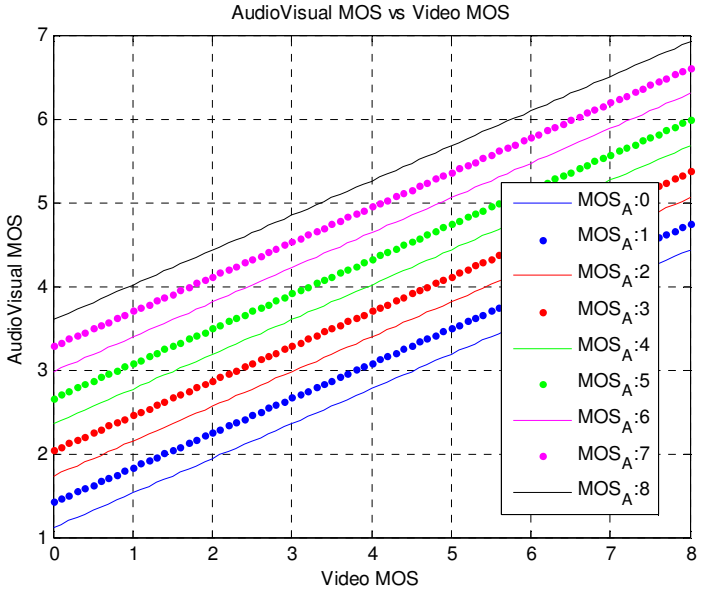

Figure 3-26: Reasonability check for coefficients in Table 3-7.

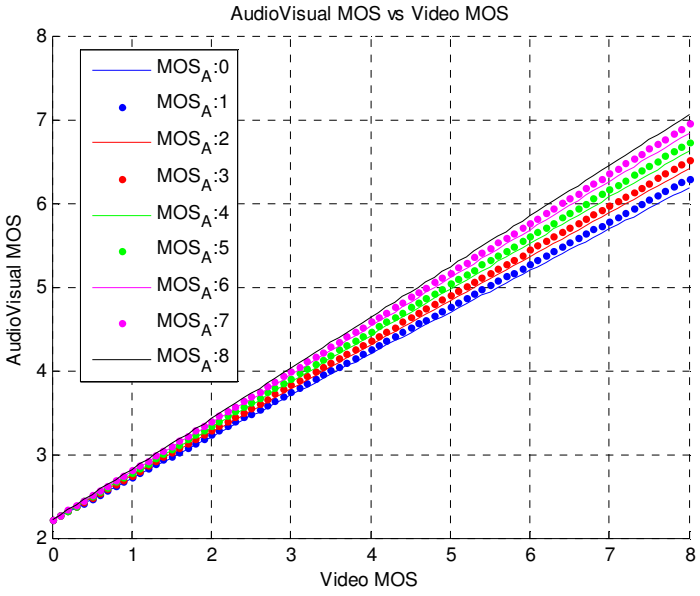

Figure 3-27: Reasonability check for coefficients in Table 3-8.

3.3.2. Content Dependent Audiovisual Quality Assessment (CDAVQA) Model

In this approach, we also use (3-28) and (3-29). However, we use multiple coefficient sets instead of a single coefficient set. As a starting point, we classify audiovisual data in the database so that videos with similar characteristics are in the same class. Then we train each class in order to determine the most appropriate coefficient set for each class. Obviously, each video with the same content should be classified as a member of the same class. Therefore, each class can be composed of a single video content. However, it is better to classify videos with similar characteristics in the same class in order to decrease the number of total classes, especially in real life applications. It has been stated that, the dominance of video quality and audio quality on audiovisual quality are different for different video contents. In this approach, coefficient sets for each class can be selected independently in order to estimate audiovisual quality in a more accurate way based on the video characteristics.

Videos in VQA and AVQA databases are subject to various compression and transmission distortions. Features that will be utilized to classify videos should be robust to these distortions. In other words, selected features should not change significantly with distortion amount. We considered to group videos according to their spatiotemporal characteristics. Having inspected features detailed in 3.1.1, we decided to use MSI and M for the mentioned classification. Figure 3-28 shows MSI with respect to M for H.264 compressed videos in the LIVE VQA database belonging to 10 different video contents.

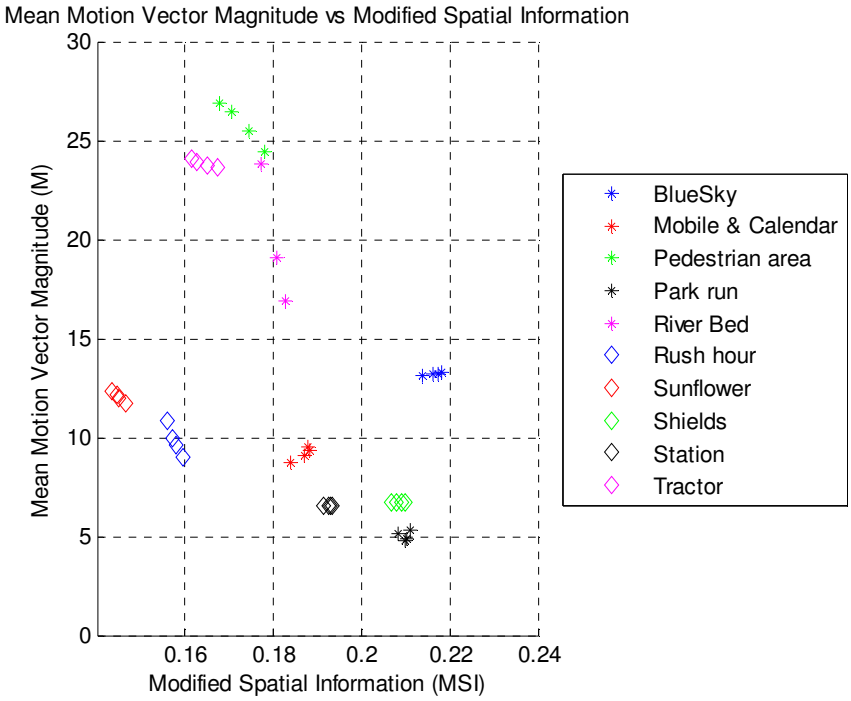

Figure 3-28: Scatter plot of Mean Motion Vector Magnitude against Modified Spatial Information in the LIVE VQA Database.

As it is seen from the figure above, distorted videos of same content are very near to each other in the MSI-M plane. Hence, MSI and M seem to be good at reflecting the spatiotemporal characteristics of videos. However, it must be noted that videos of same content in the LIVE VQA database have identical frame rates. Since we develop an AVQA model that can assess audiovisual quality of audiovisual data which have videos with various frame rates, we should take the frame rate into account. Therefore, we define Mean Motion Vector Magnitude per second (Mps) as follows:

$$Mps = M * fr \tag{3-30}$$

where fr represent the frame rate in frame per seconds.

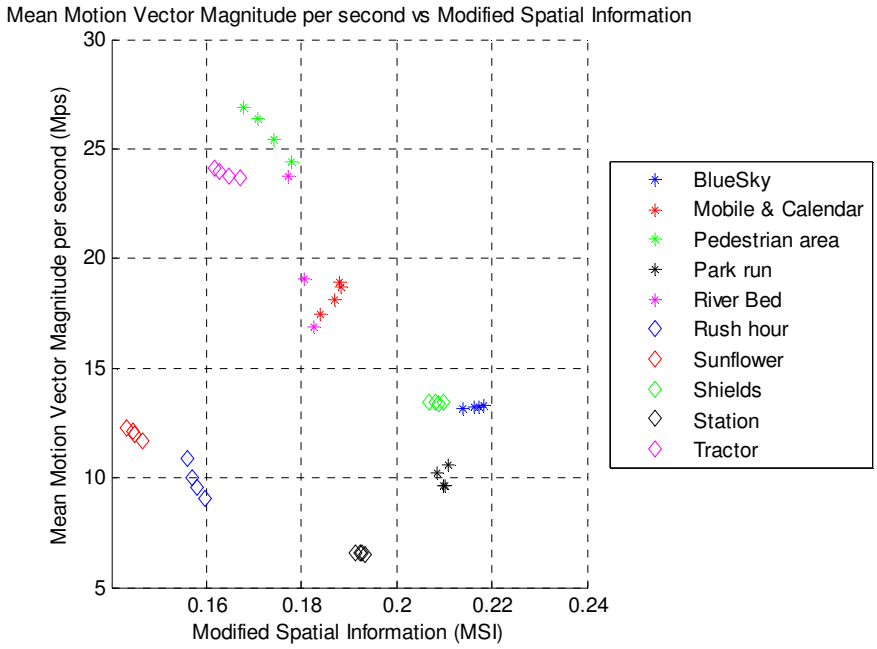

Figure 3-29: Scatter plot of Mean Motion Vector Magnitude per second against Modified Spatial Information in the LIVE VQA Database.

Figure 3-29 illustrates MSI with respect to Mps for the same videos in Figure 3-28. Then we checked whether MSI and Mps can classify 6 different video contents in the AVQA database of University of Plymouth. This AVQA database includes videos of same video content with different frame rates. It also contains videos with transmission distortions. Figure 3-29 and Figure 3-30 prove that MSI and Mps can successfully group videos of same video content with different frame rates, distorted by compression and/or transmission.

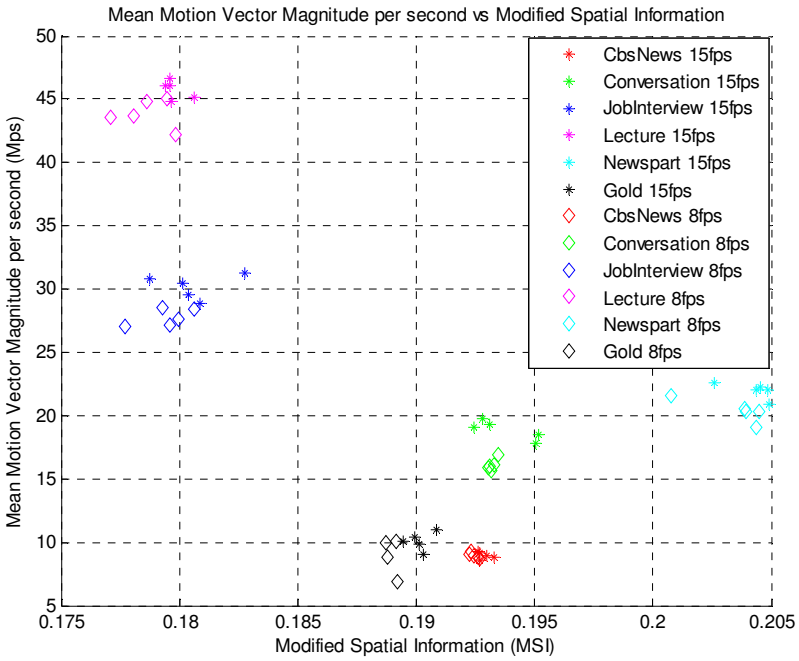

Figure 3-30: Scatter plot of Mean Motion Vector Magnitude per second with respect to Modified Spatial Information in the University of Plymouth AVQA Database.

Having classified videos in the University of Plymouth AVQA database according to their spatiotemporal characteristics, we use (3-28) and (3-29). We divide audiovisual data in the database, whose spatiotemporal characteristics are given in Figure 3-30, into 4 classes. Class 1 consists of “Lecture”, Class 2 contains “Job Interview”, Class 3 includes “Gold”, “Cbs News” and “Conversation” and Class 4 contains “Newspart” audiovisual data. The classification of the audiovisual data according to spatiotemporal characteristics is illustrated in Figure 3-31. For each class, we performed training and obtained 4 coefficient sets (that is a, b, c and d in 3-28) for both subjective video MOS with subjective audio MOS and objective video MOS with objective audio MOS. For each class, the coefficients of the expression in (3-28) obtained using subjective video MOS with subjective audio MOS are provided in Table 3-9. For each class, the coefficients of the same expression obtained using objective video MOS with objective audio MOS are given in Table 3-10. (Coefficients for class 1 and class 2 are identical in this case).

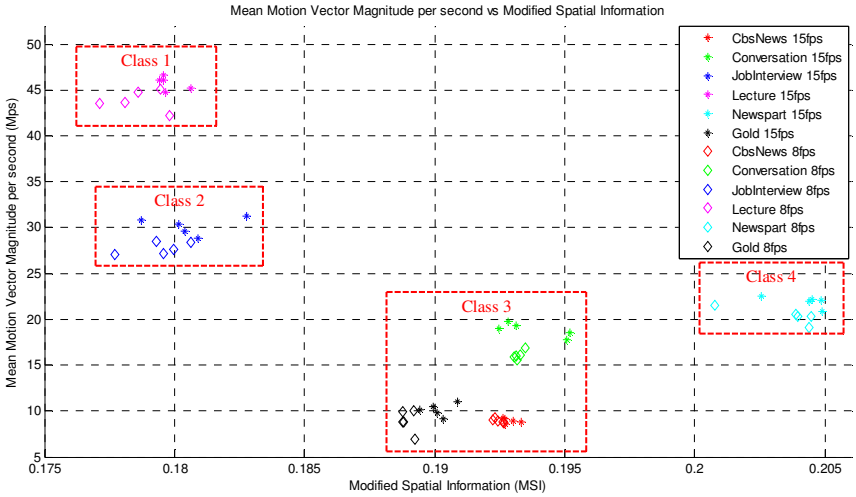

Figure 3-31: Classification of the audiovisual data according to spatiotemporal characteristics in the University of Plymouth AVQA database.

Table 3-9: Coefficients for all classes in CDAVQA for audiovisual data when subjective video MOS and subjective audio MOS are utilized.

Class	**a**	**b**	**c**	**d**
Class 1	1.221	0.260	0.419	0
Class 2	1.279	0.416	0	5.91×10^{-2}
Class 3	1.286	0.390	0.294	6.18×10^{-3}
Class 4	1.178	0.406	0.285	0

Table 3-10: Coefficients for all classes in CDAVQA for audiovisual data when objective video MOS and objective audio MOS are utilized

Class	**a**	**b**	**c**	**d**
Class 1	2.367	0.343	0	0.040
Class 2				
Class 3	2.273	0.534	8.24×10^{-5}	4.79×10^{-5}
Class 4	1.75	0.532	1.79×10^{-5}	0.028

CHAPTER 4

EVALUATION OF THE PROPOSED QUALITY ASSESSMENT MODELS

4.1. Evaluation of the Quality Assessment Models

Quality assessment (QA) models should be thoroughly analyzed and tested in order to discover potential improvements and identify specific failure cases so that the developed QA models will eventually reach to robust performance. Insufficient evaluation of QA metrics may cause false performance claims and unavoidable failure of the QA metric. All QA metrics had better been evaluated as described in the following subsections.

4.1.1. Assumption and Operation Verification

Assumptions made throughout the development of the QA model should be either verified or refuted. If a QA metric is developed in order to detect signal model violations, it should be evaluated on a wide range of undistorted inputs. Similarly, if a QA metric is designed in order to detect a particular artifact, it should not accidentally measure other artifacts [100]. Moreover, it should not detect that particular artifact in undistorted inputs. Individual measurements' monotonicity is another critical issue that must be verified. To illustrate, blockiness is expected to increase with increase in quantization amount. The pooling step should also be tested in order to validate proper operation in case of multiple artifacts. NR QA metrics relying on restricted inputs such as only test video or only bitstream parameters should be evaluated in order to see how the performance is limited by input constraints.

QA metrics may also be evaluated under synthetic inputs. Synthetic inputs enable us to test QA metrics under specific conditions. While composing synthetic inputs, new

artifacts can be added, current artifacts may be temporally and spatially distributed, or amplified. Inputs violating or exactly matching the assumed statistical signal models in the design can clearly indicate whether the quality metric predicts desired statistical quantities sufficiently.

4.1.2. Classical Numerical Measures

There are classical measures in order to evaluate the performance of QA metrics. Almost all of these measures first try to quantify differences between the subjective quality (Q_{subj}) and the predicted objective quality (Q_{obj}). There should be no doubt that the ground truth data utilized in training the QA metrics should be excluded in the test set. Pearson Correlation Coefficient (PCC), outlier ratio, and root mean square error (RMSE) measure the performance of the QA metrics based on how well the metric estimates individual subjective quality on an absolute scale (prediction accuracy and prediction consistency). Spearman Rank Order Correlation Coefficient (SROCC), on the other hand, measures how well the QA metric maintains the scores' relative ranking (prediction monotonicity). These four measures are the most widely utilized quantities [101]. In order to determine specific failure cases of the designed QA metric, these measures can also be calculated on special subsets (subsets with/without specific artifacts) of the test set.

4.1.3. Resolving Power and Classification Errors

A QA metric's accuracy can also be measured in terms of subjective quality difference between pairs v_1 and v_2, $\Delta Q_{subj} = Q_{subj}(v_1) - Q_{subj}(v_2)$. Assuming that subjective quality, $Q_{subj}(.)$, and objective quality, $Q_{obj}(.)$, are in the same scale, Brill et al. [102] introduced the resolving power of a QA metric, which measures a confidence in the predicted quality difference between pairs v_1 and v_2, $\Delta Q_{obj} = Q_{obj}(v_1) - Q_{obj}(v_2)$. The resolving power gives an understanding of whether ΔQ_{obj} is reasonable. It is worth here noting that the resolving power depends on subjective data. Hence, a QA metric may have different resolving powers on different datasets.

Classification errors take place if subjective quality difference, ΔQ_{subj}, and estimated quality difference, $\Delta Q_{obj,}$ disagree for two different sources, in one of following ways [102], [103]:

- when $|\Delta Q_{subj}| > \Delta$ but $|\Delta Q_{obj}| < \Delta$ (false tie)
- when $|\Delta Q_{subj}| < \Delta$ but $|\Delta Q_{obj}| > \Delta$ (false difference)
- when $Q_{subj}(v_1) > Q_{subj}(v_2)$ but $Q_{obj}(v_1) < Q_{obj}(v_2)$ (false ranking)

where the threshold Δ may depend on application. Δ can also be associated with the minimum desired quality difference, which is usually the Just Noticeable Difference.

4.1.4. Application-specific Evaluation

Finally, the QA model should be tested in the specific application it was developed for. To illustrate, a model developed for optimization of an algorithm should be inserted in the loop of the real algorithm in order to validate that the algorithm generates better outputs in this case than outputs produced when the model is not in the loop. Similarly, a model developed for troubleshooting should be evaluated with a multi-component system prototype where different components' failures are possible. Obviously, these verifications need a subjective testing.

Throughout this thesis, performance evaluation of the proposed QA models have been performed using PCC and SROCC, mentioned in 4.1.2. The reason why outlier ratio and RMSE are not utilized is that most of the QA models to which we compare our QA models do provide only PCC and SROCC. Therefore, providing outlier ratio and RMSE does not give any idea about the performance of the proposed QA models since it is not possible to compare to other QA models.

4.2. Results of the Video Quality Assessment Model

Three different VQA databases, LIVE [98], [99], EPFL-PoliMI [104], [105], and IT-IST [106], are selected to compare the performance of STN-VQM with the existing

VQA metrics. These VQA databases have been selected for comparison because they are publicly available, widely accepted and consist of videos with different properties such as content, spatial resolution, bit rate, frame rate, packet loss etc. Therefore, we are strongly convinced that these VQA databases provide us a thorough test environment.

Among these databases, we used only LIVE VQA database for training purpose. The remaining databases are utilized only for the evaluation of STN-VQM. The results of STN-VQM on these VQA databases are compared with those of the well-known FR and NR VQA metrics. While comparing STN-VQM to other VQA metrics, numerical measures, PCC and SROCC are employed in order to calculate the correlation between the subjective Mean Opinion Scores (MOS) and the perceived quality estimates of STN-VQM.

The advantages and performance of the STN-VQM over the other VQA algorithms will be described separately in the following subsections. It is here worth noting that STN-VQM considers spatial resolution, temporal information, bit rate, and PLR which are believed to be very significant for the perceived quality of a video sequence.

4.2.1. Results on LIVE VQA Database

As mentioned, we trained STN-VQM on the LIVE VQA database. To remember, the LIVE VQA database consists of 10 video contents distorted by 4 different processes. These processes are H.264 compression (4 videos), MPEG-2 compression (4 videos), ethernet packet losses (4 videos) and wireless bit losses (3 videos). The LIVE VQA database covers a large number of videos of the whole video space due to the variety in both video characteristics and distortion types [99].

The experimental setup is designed according ITU-T Recommendations ITU-R BT-500 [66]. During subjective tests, discrete 5-point scale (1 refers to “bad” and 5 refers to “excellent”), is employed. 38 students at the University of Texas at Austin

participated in the subjective test. 9 out of the 38 subjects were rejected due to their consistently pessimistic or optimistic quality judgments. In the subjective evaluation, Single Stimulus method, in which the test video is presented alone without being paired with the reference video, is adopted. A short training session, in which subjects familiarized with the user interface and the range of visual quality they could expect in the study, preceded the actual test. Videos in the training session are different from the videos in the study. Nevertheless, they have the same distortion types with the videos in the study.

The testing method on LIVE VQA database is cross-validation. We randomly selected 10 videos from each distortions type for the training process. Since there are 4 different distortion processes, we utilized 40 distorted video bitstreams for training among 150 distorted bitstreams. Remaining bitstreams are used in the evaluation step. Figure 4-1 shows the scatter plots of the subjective DMOS (y-axis) against the DMOS estimates of the STN-VQM (x-axis) on LIVE VQA database.

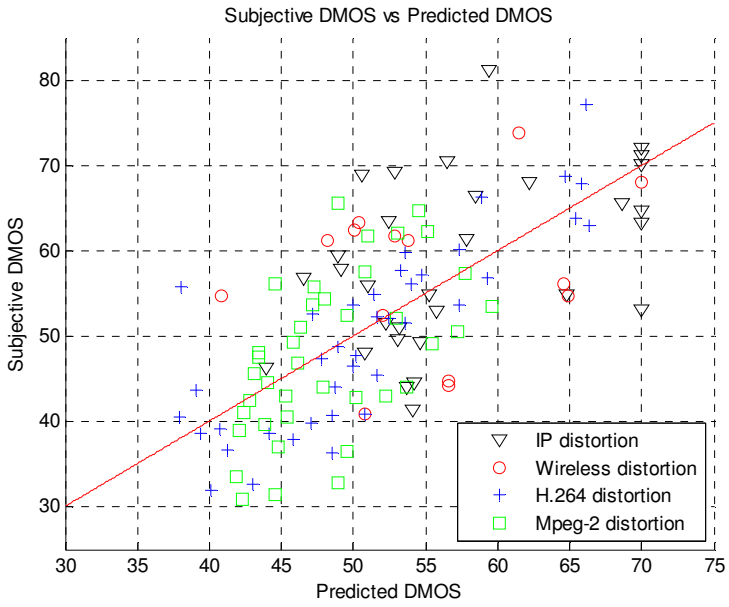

Figure 4-1: Scatter plot of subjective DMOS against predicted DMOS by the STN-VQM.

We compare the performance of the STN-VQM, to the FR metrics such as PSNR, VSNR, SW-SSIM, MS-SSIM, VQM and MOVIE. The comparison of STN-VQM to mentioned FR VQA metrics on both the H.264 compressed bitstreams and all bitstreams of the LIVE VQA database are presented in Table 4-1. We also compare the performance of the STN-VQM to the NR metrics such as MLSP moj, C-VQA, LapPyr, DVQPM, Zero-shot prediction, and Video-BLIINDS. The performance results of STN-VQM and mentioned NR VQA metrics on both the H.264 compressed and all bitstreams of the LIVE VQA database are presented in Table 4-2. It is worth noting that MLSP moj, C-VQA, LapPyr, and DVQPM are designed to be used only on H.264 compressed bitstreams.

Table 4-1: Comparison of STN-VQM to FR VQA metrics on LIVE video quality database

Method	***Type***	***H.264***	***All data***	***H.264***	***All data***
		PCC		**SROCC**	
PSNR	FR	0.4385	0.4035	0.4296	0.3684
VSNR [107]	FR	0.6216	0.6896	0.646	0.6755
SW-SSIM [108]	FR	0.7206	0.5962	0.7086	0.5849
MS-SSIM [109]	FR	0.6919	0.7441	0.7051	0.7361
VQM [13]	FR	0.6459	0.7236	0.652	0.7026
MOVIE [110]	FR	0.7902	0.8116	0.7664	0.789
STN-VQM	NR	0.8122	0.6730	0.8026	0.6697

Table 4-2: Comparison of STN-VQM to NR VQA metrics on LIVE video quality database

Method	*Type*	*H.264*	*All data*	*H.264*	*All data*
		PCC		**SROCC**	
MLSP moj [111]	NR	0.524	-	0.563	-
C-VQA [112]	NR	0.7927	-	0.7720	-
LapPyr [17]	NR	0.911	-	0.940	-
DVQPM [113]	NR	0.967	-	0.963	-
Zero-shot Pred. [114]	NR	0.778	0.62	0.777	0.604
Video-BLIINDS [115]	NR	0.893	0.881	0.839	0.759
STN-VQM	NR	0.8122	0.6730	0.8026	0.6697

We analyze the results given in Table 4-1 and Table 4-2 in two different cases, considering only H.264 compressed bitstreams and considering all bitstreams in the LIVE VQA database.

In the H.264 compressed bitstreams case, STN-VQM outperforms all FR VQA metrics as seen from Table 4-1. STN-VQM also performs better than MLSP moj, C-VQA and Zero-shot prediction NR metrics as shown in Table 4-2. Although LapPyr and DVQPM seem to be more accurate, these algorithms utilize leave-one-out strategy while evaluating their algorithm. In the leave-one-out strategy, all videos except a test video are used for training and validation; the remaining sequence is used for the testing. This procedure is repeated for all videos and the obtained results are averaged over all videos. Hence, the performance results of the VQA models obtained via the leave-one-out strategy usually yield higher scores. Similarly, Video-BLIINDS yields higher correlation results; nevertheless, the results of Video-BLIINDS algorithm are obtained using all possible combinations of %80 train and %20 test splits. In addition, Video-BLIINDS does not consider bit rate and PLR features together with the spatiotemporal information of a video sequence while estimating perceived quality. Actually, it only uses spatial and temporal features of a video sequence in the DCT domain. However, STN-VQM appraises the quality in a hybrid way by combining the spatiotemporal information, bit-rate, and PLR, which

are all important parameters for the HVS judgment of a video sequence. To sum up, comparing STN-VQM to LapPyr, DVQPM, and Video-BLIINDS is not fair since our testing procedure uses only 40 of 150 distorted bitstream for training and the remaining videos are used for testing.

When all bitstreams in the LIVE VQA database considered, STN-VQM provides competitive results even though the MOVIE and Video-BLIINDS outperform all algorithms in FR and NR cases, respectively. The arguments above regarding the difference in test procedure and features considered for the Video-BLIINDS are also valid in this case. Noting that MOVIE is an FR algorithm, STN-VQM gives quite promising results since most of the NR metrics in Table 4-2 are designed to estimate only the perceived quality of H.264 compressed bitstreams. In order to ease to interpret Table 4-1 and Table 4-2, correlation results of VQA models in Table 4-2 tables are shaded with light gray if the corresponding VQA model has a higher correlation than STN-VQM. Similarly, correlation results of VQA models in both tables are shaded with dark gray if the corresponding VQA model has a lower correlation than STN-VQM.

4.2.2. Results on PoliMI-EPFL VQA Database

We utilized the subjective data collected at two universities in different countries: Politecnico di Milano (PoliMI) – Italy, and Ecole Politechnique Federale de Lausanne (EPFL) – Switzerland. This publicly available VQA database includes subjective scores, relative to quality assessment of 156 video streams encoded with H.264/AVC and corrupted by simulating packet losses over an error-prone network. The video Group of Picture (GOP) structure in this database is IBBP.

78 of these 156 video streams are at Common Intermediate Format (CIF) spatial resolution (352x288 pixels). There are 6 different video contents, namely Foreman, Hall, Mobile, Mother, News and Paris, at 6 different packet loss rates, ranging from 0.1% to 10% (One frame for all video contents is depicted in Figure 4-2). The 6

packet loss free sequences were also included in the test material, thus finally 78 sequences were rated by each subject.

Figure 4-2: One frame from each of the 6 video contents at CIF resolution. a) Foreman, b) Hall, c) Mobile, d) Mother, e) News, f) Paris [116].

Remaining 78 video streams are at 4CIF spatial resolution (704x576 pixels). Similarly, there are six different video sequences corresponding to 6 different video contents, namely Ice, Harbour, Soccer, CrowdRun, DucksTakeoff and ParkJoy, at different packet loss rates [0.1%, 0.4%, 1%, 3%, 5%, 10%]. One frame for all video contents is depicted in Figure 4-3 .

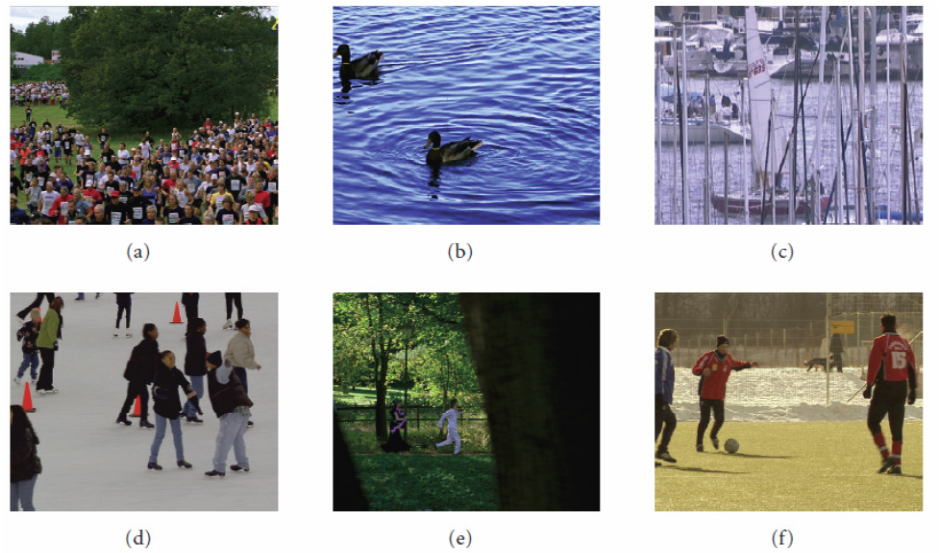

Figure 4-3: One frame from each of the 6 video contents at 4CIF spatial resolution. a) CrowdRun, b) DucksTakeoff, c) Harbour, d) Ice, e) ParkJoy, f) Soccer [116].

These sequences at both CIF and 4CIF spatial resolution are said to be selected since they are claimed to be representative of different levels of spatial and temporal complexity, as computed by means of the Spatial Information (SI) and Temporal Information (TI) indexes (Figure 4-4).

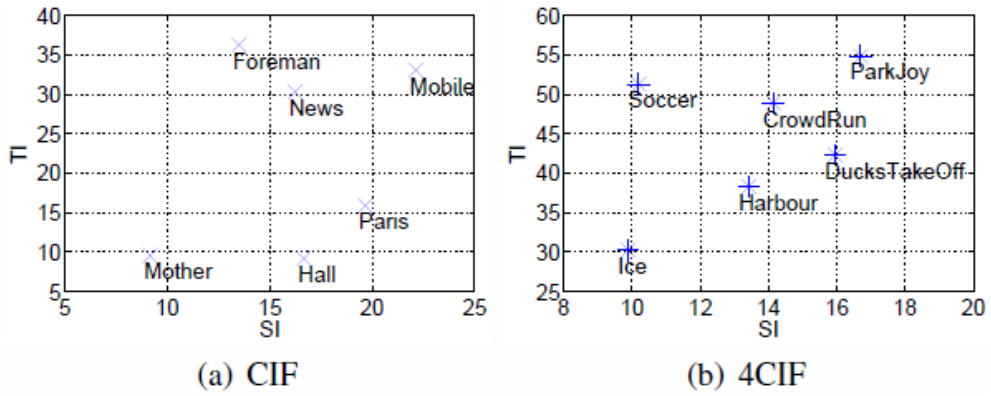

Figure 4-4: Spatial Information (SI) and Temporal Information (TI) indexes computed on the luminance component of the CIF and 4CIF video sequences [105].

The compressed bitstreams were obtained using the H.264/AVC High Profile with the encoding parameters below:

Table 4-3: H.264/AVC encoding parameters in PoliMI-EPFL VQA database [116]

Reference Software	JM 14.2
Profile	High
Number of frames	298
Chroma format	4:2:0
GOP size	16
GOP structure	IBBPBBPBBPBBPBB ...
Number of reference frames	5
Slice Mode	Fixed number of MBs
Rate Control	Disabled, fixed quantization parameter
MB partitioning for motion estimation	Enabled
Motion estimation algorithm	Enhanced Predictive Zonal Search
Early skip detection	Enabled
Selective intra mod decision	Enabled

It is worth noting that each frame is split in 18 slices and each slice consists of a full row of MBs. In the NAL, the bitstreams were formatted for IP networks. Each packet just contains the single slice information. Hence, a packet loss means a loss of a full slice. The error concealment method employed in EPFL-PoliMI VQA database is frame-copy error concealment method. At first all videos in this database were utilized in the evaluation of STN-VQM, i.e., none of these 156 video streams have been used for training. Then we also performed training on EPFL-PoliMI VQA database.

The experimental setup is designed according to ITU-T Recommendations ITU-R BT-500 [66]. During subjective tests, 5-point ITU continuous scale ([0-1] refers to

"bad" and (4-5] refers to "excellent"), is employed. The number of subjects participating in the test was as follows: 23 for CIF and 21 for 4CIF at PoliMI and 17 for CIF and 19 for 4CIF at EPFL. Subjects' ages range from 24 to 40 years. Some of the subjects were familiar with image and video processing. In the subjective evaluation, Single Stimulus method, in which the test video is presented alone without being paired with the reference video, is adopted. First, written instructions detailing the procedure were provided to subjects. Then a short training session, in which subjects familiarized with the user interface and assessment procedure, preceded the actual test.

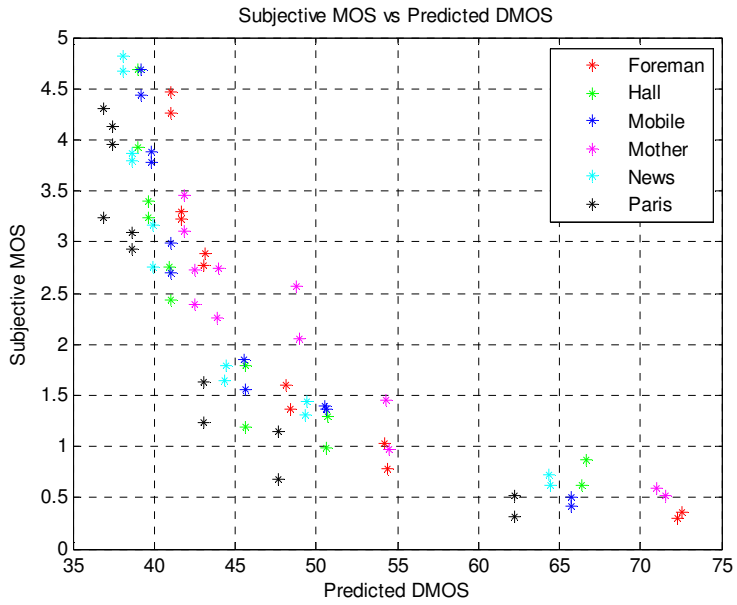

Figure 4-5: Scatter plot of subjective MOS against predicted DMOS by the STN-VQM for video streams at CIF spatial resolution without training.

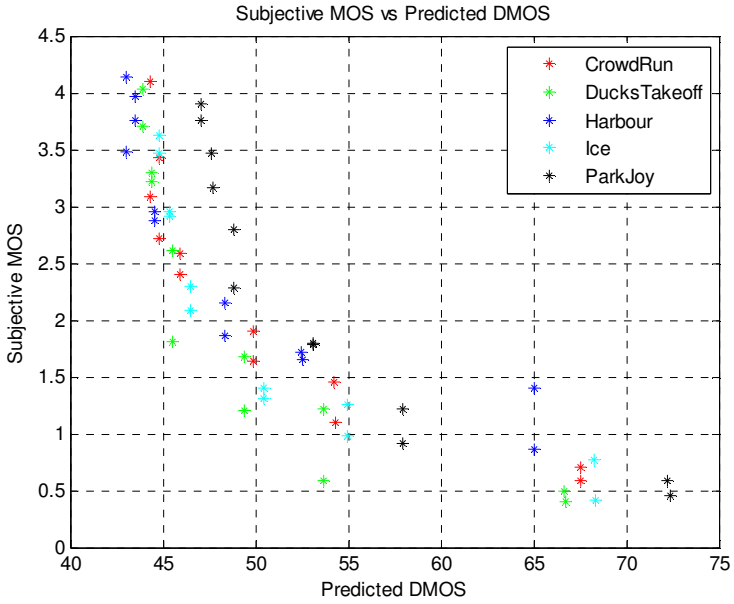

Figure 4-6: Scatter plot of subjective MOS against predicted DMOS by the STN-VQM for video streams at 4CIF spatial resolution without training.

Figure 4-5 shows the scatter plots of the subjective MOS against the DMOS estimates of STN-VQM on the PoliMI-EPFL VQA database at CIF spatial resolution without training. Figure 4-6 shows the scatter plots of the subjective MOS against the DMOS estimates of STN-VQM on the PoliMI-EPFL VQA database at 4CIF spatial resolution without training.

We compare the performance of STN-VQM with the well-known FR metrics (i.e., PSNR, VSNR, SSIM, MS-SSIM, VQM, MOVIE and VIF) and the NR VQA methods: Genetic Programming-based Symbolic Regression (GP metric) and Video-BLIINDS. The performance comparison of STN-VQM and mentioned FR VQA metrics on the PoliMI-EPFL VQA database is presented in Table 4-4. The performance comparison of STN-VQM and mentioned NR VQA metrics on the PoliMI-EPFL VQA database is given in Table 4-5. Similar to Table 4-1 and Table 4-2, in order to ease to interpret Table 4-4 and Table 4-5, correlation results of VQA models in both tables are shaded with light gray if the corresponding VQA model has a higher correlation than STN-VQM. Similarly, correlation results of VQA models in

Table 4-4 tables are shaded with dark gray if the corresponding VQA model has a lower correlation than STN-VQM.

As can be seen in Table 4-4, STN-VQM is validated on EPFL-PoliMI VQA databases at two different spatial resolutions. In general, the FR metrics are expected to perform better than the NR VQA models since they have full access to the reference video. Considering the FR methods, STN-VQM outperforms the VQA metrics such as PSNR, VQM, SSIM and VIF even though it is an NR metric. Nevertheless, MS-SSIM and MOVIE perform better than the STN-VQM.

Table 4-4: Comparison of STN-VQM to FR VQA metrics on PoliMI-EPFL video quality database without training.

Method	***Type***	***PCC***	***SROCC***
PSNR	FR	0.793	0.800
VSNR [107]	FR	0.894	0.895
MS-SSIM [109]	FR	0.915	0.922
VQM [13]	FR	0.843	0.838
SSIM	FR	0.678	0.677
VIF	FR	0.749	0.740
MOVIE [110]	FR	0.930	0.920
STN-VQM	NR	0.848	0.906

Table 4-5: Comparison of STN-VQM to NR VQA metrics on PoliMI-EPFL video quality database without training.

Method	***Type***	***PCC***	***SROCC***
GP metric [29]	NR	0.882	0.883
Video-BLIINDS [115]	NR	0.752	0.807
G.1070 [31]	NR	0.910	0.890
G.1070E [30]	NR	0.930	0.920
STN-VQM	NR	0.848	0.906

Considering the NR VQA metrics, STN-VQM has similar accuracy with the GP metric as shown in Table 4-5. This result is impressive noting that the GP metric uses %60 of all videos in EPFL-PoliMI VQA database for training and STN-VQM uses all videos in the same database for testing. Since there are limited number of NR VQA metrics utilizing the EPFL-PoliMI VQA database, Table 4-5 consists of the results of the Video-BLIINDS, G.1070 and G.1070E. Similar to the difference in testing procedure on LIVE VQA database, the testing procedure of Video-BLIINDS is based on averaging all possible combinations of %80 train and %20 test splits of only 4CIF videos in the EPFL-PoliMI VQA database. G.1070 and G.1070E VQA models also use only 4CIF videos in the EPFL-PoliMI VQA database. Moreover, the training procedures of these methods are not stated clearly. Although it may not be fair to compare STN-VQM with these algorithms due to these differences, we feel it is necessary to include them in Table 4-5.

We also performed training in EPFL-PoliMI VQA database. Since the main distortion in this database is network distortion, the focus of the training step was finding better coefficients for m and n in h_{TR} in (4-24). This time, we directly trained according to MOS instead of DMOS and obtained MOS estimates in 0-5 range.

Table 4-6 shows the new coefficients for transmission distortion after training video streams in EPFL-PoliMI VQA database at CIF spatial resolution. Figure 4-7 depicts the scatter plot of subjective MOS against predicted MOS by the STN-VQM after

training video streams in EPFL-PoliMI VQA database at CIF spatial resolution. The resulting PCC and SROCC are 0.94 and 0.92, respectively. These correlation coefficients are higher than those obtained in the same database without training, as expected.

Table 4-6: New coefficients for transmission distortion after training video streams in EPFL-PoliMI VQA database at CIF spatial resolution

Coefficient Name	***Value***
m	0.14
n	-0,27

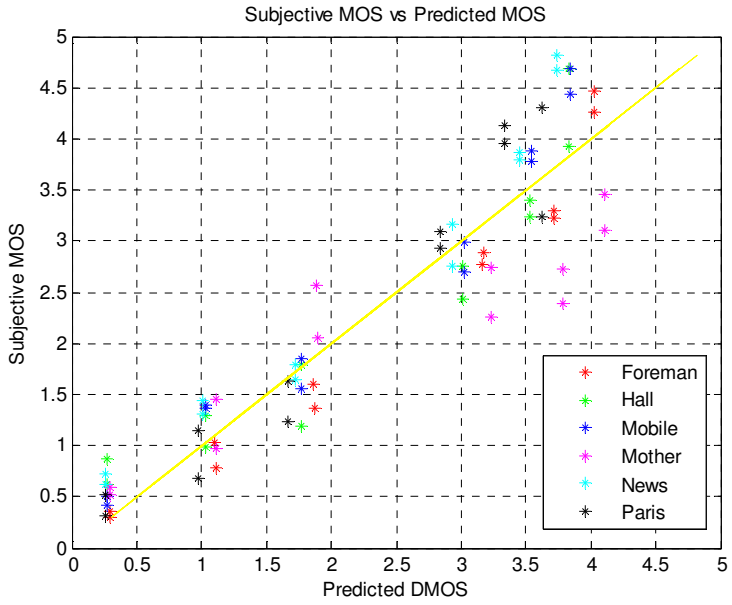

Figure 4-7: Scatter plot of subjective MOS against predicted MOS by the STN-VQM after training video streams in EPFL-PoliMI VQA database at CIF spatial resolution.

Table 4-7 depicts the new coefficients for transmission distortion after training video streams in EPFL-PoliMI VQA database at 4CIF spatial resolution. Figure 4-8 shows

the scatter plot of subjective MOS against predicted MOS by the STN-VQM after training video streams in EPFL-PoliMI VQA database at 4CIF spatial resolution. The resulting PCC and SROCC are 0.93 and 0.92, respectively. These correlation coefficients are also higher than those obtained in the same database without training, as expected.

Table 4-7: New coefficients for transmission distortion after training video streams in EPFL-PoliMI VQA database at 4CIF spatial resolution.

Coefficient Name	***Value***
m	0.14
n	-0.24

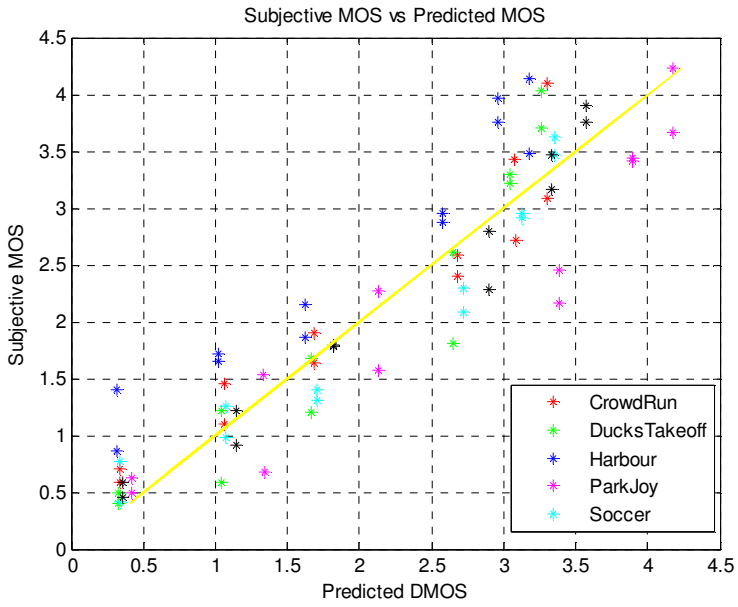

Figure 4-8: Scatter plot of subjective MOS against predicted MOS by the STN-VQM after training video streams in EPFL-PoliMI VQA database at 4CIF spatial resolution.

4.2.3. Results on IT-IST VQA Database

We also evaluated STN-VQM on subjective data collected by the Image Group of Instituto de Telecomunicacoes, Instituto Superior Tecnico (IT-IST) [106]. There are video streams with various video contents at CIF spatial resolution with corresponding subjective scores in IT-IST VQA database (One frame of all video content is illustrated in Figure 4-9). These video streams are encoded with H.264/AVC at various bit rates ranging from 32 to 2048 kbit/s (see Table 4-8). There is no packet loss in the utilized IT-IST VQA database. The videos in the IT-IST VQA database have a GOP structure IBBP.

Figure 4-9: One frame from each of the video contents in IT-IST VQA database [117].

The experimental setup is designed according to ITU-T Recommendations ITU-R BT-500 [66]. During subjective tests, discrete 5-point scale (1 refers to “very

annoying" and 5 refers to "imperceptible"), is employed. 22 subjects, most of whom were students, participated in the subjective test. In the subjective evaluation, Double Stimulus Impairment Scale method, in which the test video is presented to subjects after presenting the reference video, is adopted. Before subjective tests, Snellen Eye Chart and Ishihara's plates are utilized in order to screen subjects for visual acuity and color blindness.

Table 4-8: Encoding bit rates of the vdeo streams using H.264/AVC

Video Sequence	***Encoding Bit Rates (kbps)***
City	128, 200, 256, 612
Costguard	64, 100, 128, 200, 256, 512
Container	64, 128, 256, 512
Crew	128, 200, 400, 1024
Football	256; 400, 512, 750, 1024, 2048
Foreman	64, 128, 256, 512
Mobile	64, 128, 200, 256, 400, 512
Silent	64, 200, 400, 1024
Stephan	128, 200, 256, 400, 512, 1024
Table	64, 128, 256, 512
Tempete	128, 200, 400, 750

At first, none of the videos in this database are employed for training, i.e., all videos in this database are used to test the STN-VQM. Then we also performed training in IT-IST VQA database.

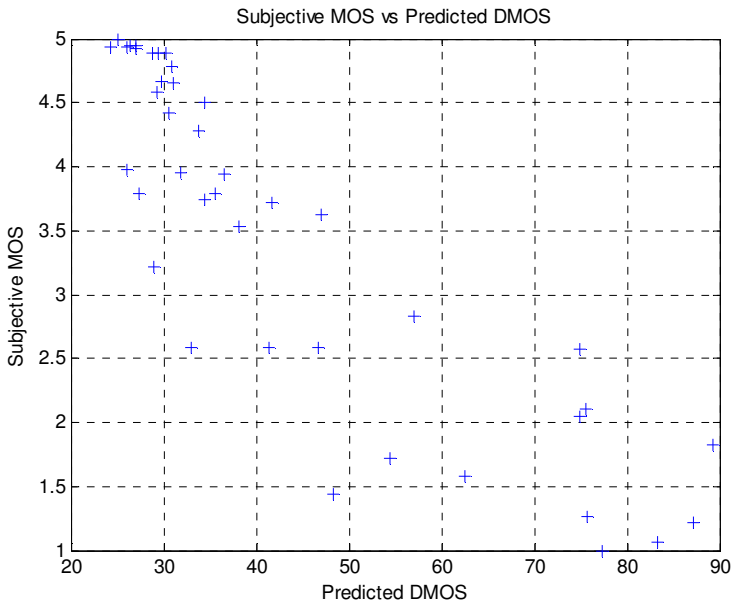

Figure 4-10: Scatter plot of subjective MOS against predicted DMOS by the STN-VQM for video streams in IT-IST VQA database without training.

Figure 4-10 depicts the scatter plot of the subjective MOS against the DMOS estimates of STN-VQM on the IT-IST VQA database without training. The performance of STN-VQM is compared to two NR VQA algorithms, namely the ITU-T Recommendation G.1070 [31] and enhanced G.1070 system, referred as G.1070E [30]. Comparison of the results of the STN-VQM on the IT-IST VQA database with the results of the G.1070 and G.1070E on the same database is provided in Table 4-9.

Table 4-9: Comparison of STN-VQM to NR VQA metrics on IT-IST VQA database

Method	***PCC***	***SROCC***
G.1070 [31]	0.71	0.81
G.1070E [30]	0.91	0.94
STN-VQM	0.87	0.90

Results on Table 4-9 states that STN-VQM is also validated on the IT-IST VQA database. Clearly, STN-VQM outperforms the ITU-T Recommendation G.1070. STN-VQM and G.1070E seem to have similar accuracy. It is here worth restating that videos in IT-IST VQA database are used only for testing, i.e., none of the video streams in this database are used for training. Furthermore, the G.1070 and G.1070E models do not consider spatial resolution feature which can yield blurring artifact influencing the HVS while evaluating the quality of a video sequence. However, STN-VQM uses the spatial resolution, temporal information, bit rate, and PLR parameters which are all important for the quality assessment in a hybrid manner.

We also performed training in IT-IST VQA database. Since the main distortion in this database is compression distortion, the focus of the training step was finding better coefficients for a through f in $DMOS_{initial}(S, T)$ in (3-12). This time, we directly trained according to MOS instead of DMOS and obtained MOS estimates in 0-5 range.

Table 4-10 shows the new coefficients for compression distortion after training video streams in IT-IST VQA database. Figure 4-11 depicts the scatter plot of subjective MOS against predicted MOS by the STN-VQM after training video streams in IT-IST VQA database. The resulting PCC and SROCC was 0.89 and 0.88 respectively. We obtained higher prediction accuracy (indicated by higher value of PCC) after training as expected. In terms of prediction monotonicity, the performance is almost the same.

Table 4-10: New coefficients for compression distortion after training video streams in IT-IST VQA database

Coefficient Name	*Value*
a	5.65
b	-17.6
c	-480.6
d	57.7
e	-914.1
f	3.23×10^4

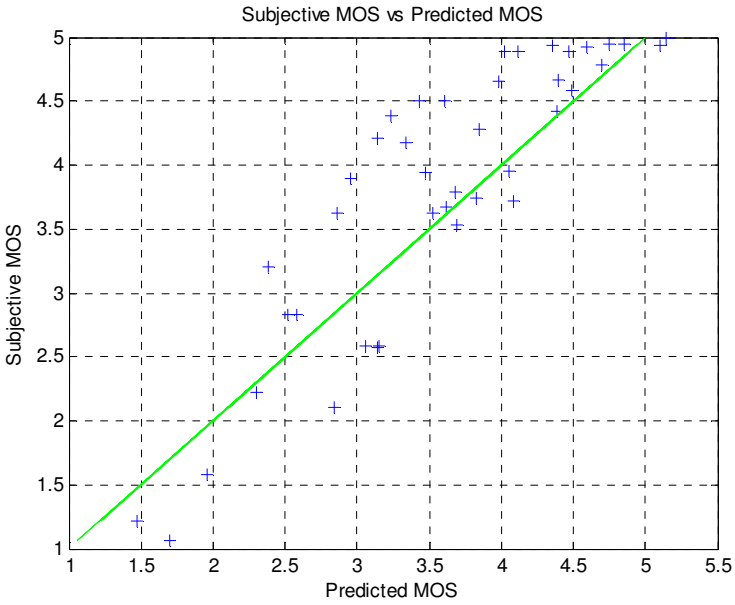

Figure 4-11: Scatter plot of subjective MOS against predicted MOS by the STN-VQM after training video streams in IT-IST VQA database.

4.3. Audio Quality Assessment Results

The AQA database in [70] consists of audio coded with coding standard MPEG-4 AAC-LC (low complexity). QuickTime Pro was used for encoding, with the

"recommended" sampling rate for each target bit rate. Audio test conditions are detailed in Table 4-11 [70]. The audio source material was 16-bit PCM stereo sampled at 48 kHz. The audio sampling rate reduction was carried out internally by the encoder.

The experimental setup is designed according to ITU-T Recommendations P.911 [118]. Absolute Category Rating, where the test clips are viewed one at a time and evaluated independently on a discrete 11-level scale (0 refers to "bad" and 10 refers to "excellent"), is employed in subjective testing. Six female and 18 male subjects, whose ages range from 25 to 36 years, participated in the test. One of the subjects was familiar with the audio processing. High-quality headphones (Sennheiser HD 600) were connected to an external digital-to-analog converter (Emagic EMI A26) for the audio playback. First, written instructions detailing the procedure were provided to subjects. Then a short training session, in which subjects adjusted headphone volumes, preceded the actual test.

Table 4-11: Test conditions of the audio quality assessment database

Condition	***Channels***	***Sampling Rate (kHz)***	***Bit Rate (kbps)***
1	Mono	8	8
2	Mono	16	16
3	Mono	22	24
4	Mono	32	32
5	Mono	22	32
6	Stereo	22	32
7	Stereo	16	32

Figure 4-12 shows the scatter plots of the subjective MOS against the MOS estimates of the proposed AQA metric on the AQA database. Proposed AQA metric has a PCC value of 0.978 and SROCC value of 0.912. Noting the fact that proposed AQA model is a NR model, these results are quite promising.

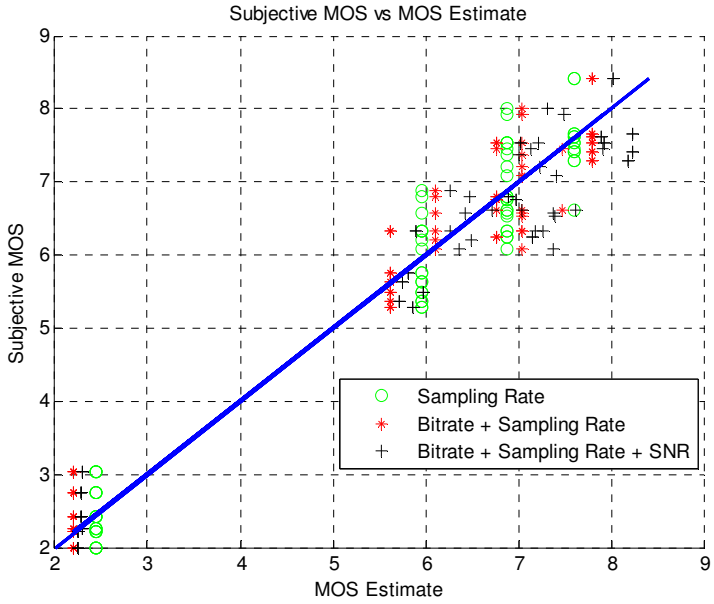

Figure 4-12: Scatter plot of subjective MOS against predicted MOS by the proposed AQA metric.

4.4. Audiovisual Quality Assessment Results

First, we evaluated STN-VQM on video data in the AVQA database provided by University of Plymouth [87]. This AVQA database consists of 60 audiovisual samples. There are 6 different video contents and video files are encoded with H.263 with frame rates either 8 or 15. It is here worth restating that there are packet losses which occur in the wireless segment of the network using a Gilbert-Elliot model with packet error rates, 0.01, 0.05, 0.1, 0.15 and 0.20.

The experimental setup is designed according to ITU-T Recommendations P.910 [6], P.911 [118] and P.800 [67]. Absolute Category Rating, where the test clips are viewed one at a time and evaluated independently on a discrete 9-level scale (0 refers to “bad” and 8 refers to “excellent”), is employed in subjective testing. 22 female and 26 male paid subjects, whose ages range from 18 to 40 years, participated in the test. The subjective test and subjects were divided into three parts. 8 male and 8 female participated in video-only part, 9 male and 7 female participated in audio-only part

and 9 male and 7 female participated in audiovisual part. Subjects were either staff or student at the University of Plymouth. Subjective tests are conducted through a website, where the samples are presented to subjects and their corresponding scores are acquired. Full instructions detailing the test procedure were given in the website. It is here worth noting that a warm-up page is presented at the preceding the actual test.

Figure 4-13 shows the scatter plots of the video MOS against the video MOS estimates of the STN-VQM on Plymouth AVQA database. STN-VQM has a PCC value of -0.81 and SROCC value of -0.81. Negative correlations are due to the fact that STN-VQM outputs DMOS values instead of MOS values. It is here worth stating that none of the videos in this database are employed for training, i.e., all videos in this database are used to test the STN-VQM.

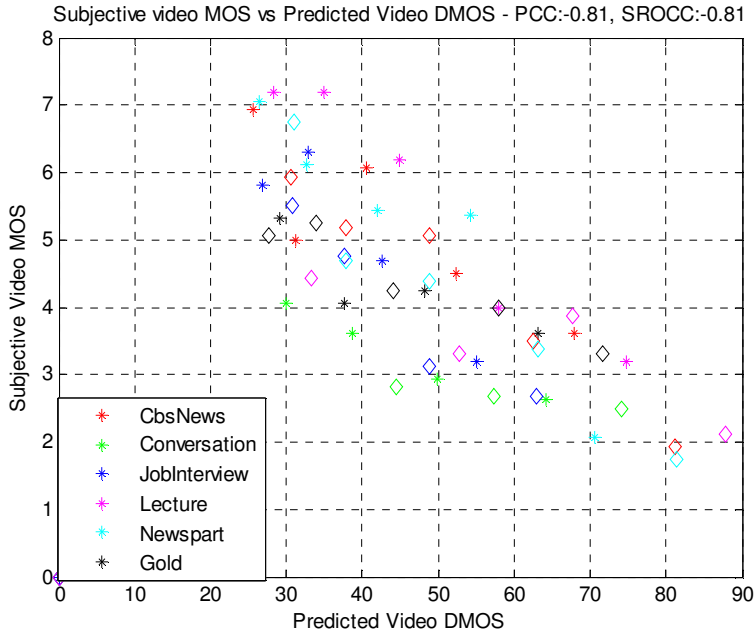

Figure 4-13: Scatter plot of subjective video MOS against predicted video DMOS by the STN-VQM for video streams in Plymouth AVQA database.

Figure 4-14 shows the scatter plots of the subjective audio MOS against the audio MOS estimates of the proposed AQA metric on Plymouth AVQA database. Proposed AQA metric has a PCC value of 0.74 and SROCC value of 0.74. Keeping in mind that proposed AQA model is a NR model, these results are said to be quite satisfactory.

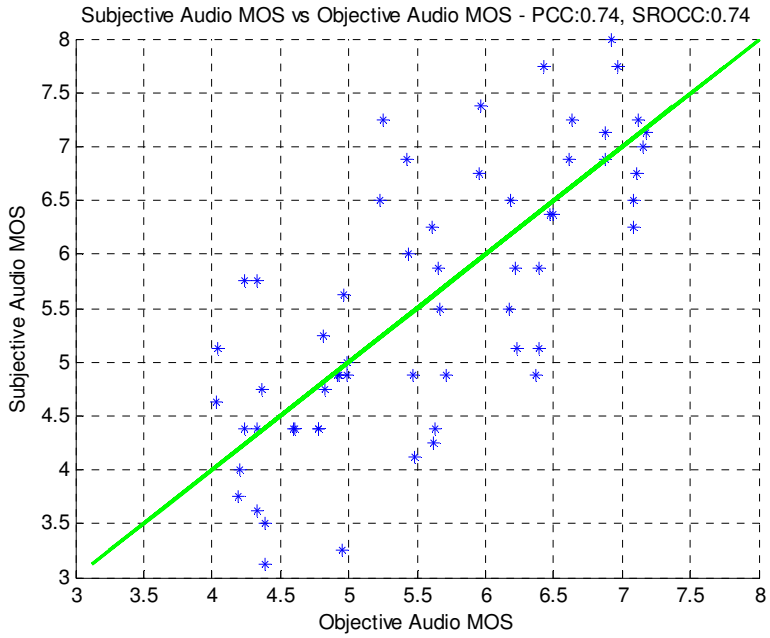

Figure 4-14: Scatter plot of subjective audio MOS against predicted audio MOS by the AQA algorithm for audio streams in Plymouth AVQA database.

4.4.1. Results of DAVQA Model

Figure 4-15 shows the scatter plots of the audiovisual MOS against the audiovisual MOS computed by DAVQA model using subjective video MOS and subjective audio MOS in Plymouth AVQA database.

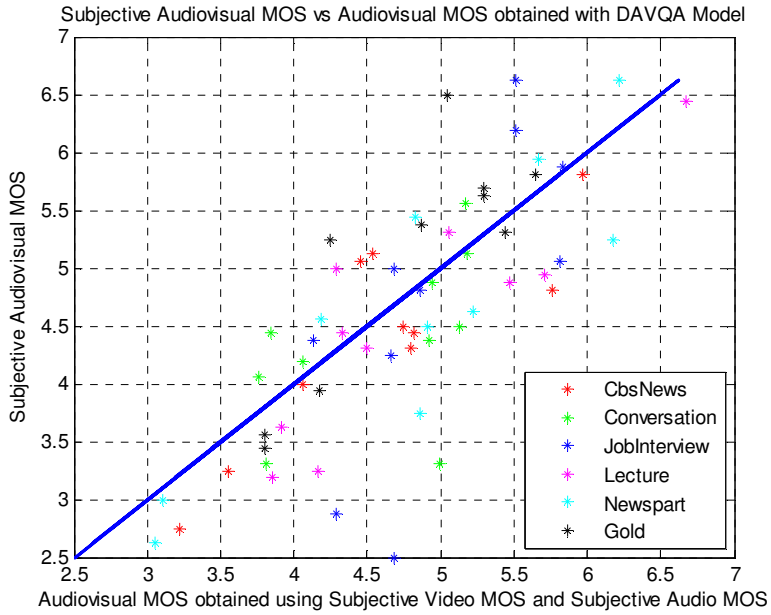

Figure 4-15: Scatter plot of subjective audiovisual MOS against audiovisual MOS obtained by DAVQA model using subjective video MOS and subjective audio MOS in Plymouth AVQA database.

Figure 4-16 shows the scatter plots of the audiovisual MOS against the predicted audiovisual MOS obtained by DAVQA model using objective video DMOS estimated with STN-VQM and objective audio MOS estimated with the proposed AQA algorithm in Plymouth AVQA database.

Figure 4-15 and Figure 4-16 show that audiovisual MOS obtained by DAVQA model provides high correlation with subjective audiovisual MOS.

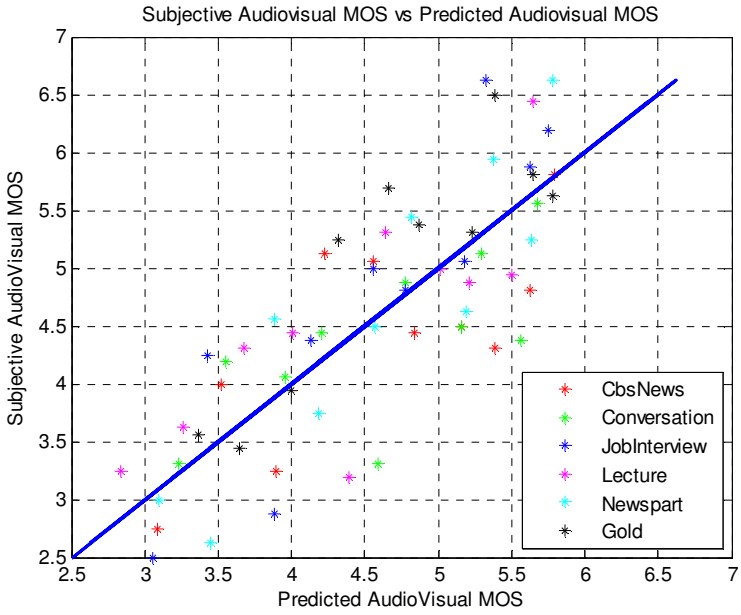

Figure 4-16: Scatter plot of subjective audiovisual MOS against predicted audiovisual MOS by DAVQA model in Plymouth AVQA database.

4.4.2. Results of CDAVQA Model

Figure 4-17 shows the scatter plots of the audiovisual MOS against the audiovisual MOS computed by CDAVQA model using subjective video MOS and subjective audio MOS in Plymouth AVQA database.

Figure 4-17 depicts that audiovisual MOS obtained by CDAVQA model provides high correlation with subjective audiovisual MOS. The correlation coefficients obtained using subjective scores with DAVQA and CDAVQA models are presented in Table 4-12. It is obvious that CDAVQA performs better than DAVQA in this case. Figure 4-18, Figure 4-19, Figure 4-20 and Figure 4-21 illustrate the correlation of each class when subjective video MOS and subjective audio MOS are employed. The correlation coefficients for each class are presented in Table 4-13.

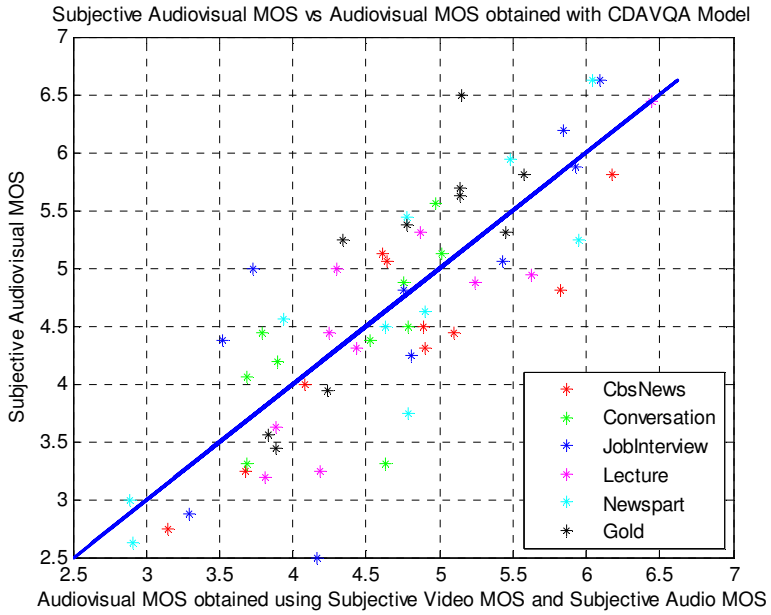

Figure 4-17: Scatter plot of subjective audiovisual MOS against audiovisual MOS obtained by CDAVQA model using subjective video MOS and subjective audio MOS in Plymouth AVQA database.

Table 4-12: Performance comparison of DAVQA and CDAVQA models when subjective video and subjective audio scores are utilized

Model	***Audiovisual Data***	***PCC***	***SROCC***
DAVQA	Subjective video MOS / Subjective Audio MOS	0.775	0.781
CDAVQA	Subjective video MOS / Subjective Audio MOS	0.813	0.802

As it can be seen in Table 4-13, correlation coefficients of Class 1 and Class 4 are higher than correlation coefficients of Class 2 and Class 3. Nevertheless, the correlation coefficients of Class 3, which consists of three video contents, are quite satisfactory.

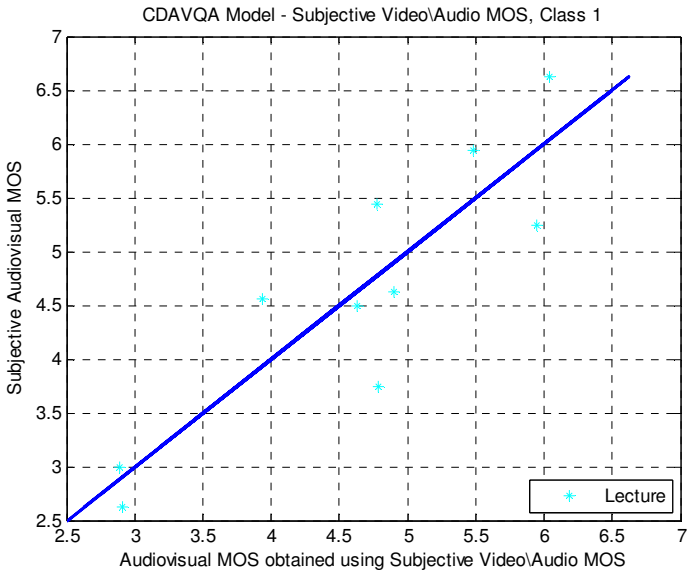

Figure 4-18: Scatter plot of subjective audiovisual MOS against audiovisual MOS obtained by CDAVQA model using subjective video MOS and subjective audio MOS for videos classified as Class 1 in Plymouth AVQA database.

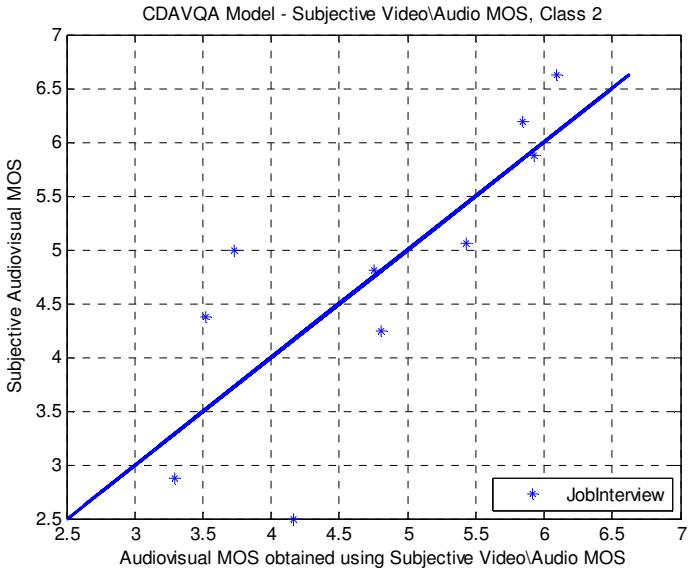

Figure 4-19: Scatter plot of subjective audiovisual MOS against audiovisual MOS obtained by CDAVQA model using subjective video MOS and subjective audio MOS for videos classified as Class 2 in Plymouth AVQA database.

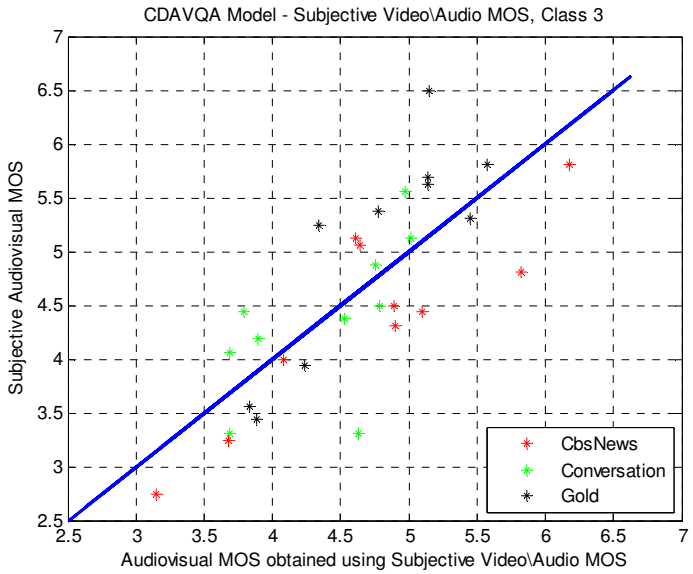

Figure 4-20: Scatter plot of subjective audiovisual MOS against audiovisual MOS obtained by CDAVQA model using subjective video MOS and subjective audio MOS for videos classified as Class 3 in Plymouth AVQA database.

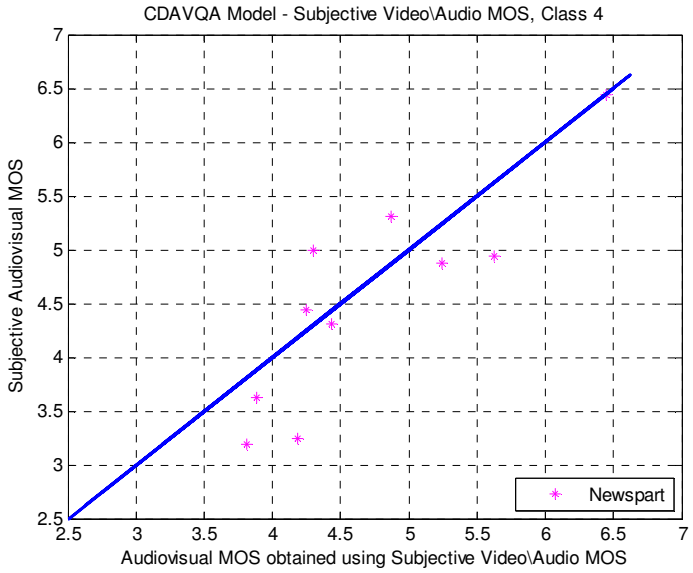

Figure 4-21: Scatter plot of subjective audiovisual MOS against audiovisual MOS obtained by CDAVQA model using subjective video MOS and subjective audio MOS for videos classified as Class 4 in Plymouth AVQA database.

Table 4-13: Performance of CDAVQA model for each class when subjective video and subjective audio scores are utilized

Class	*Audiovisual Data*	*PCC*	*SROCC*
Class 1	Subjective video MOS / Subjective Audio MOS	0.883	0.812
Class 2	Subjective video MOS / Subjective Audio MOS	0.786	0.794
Class 3	Subjective video MOS / Subjective Audio MOS	0.774	0.807
Class 4	Subjective video MOS / Subjective Audio MOS	0.859	0.830

Figure 4-22 shows the scatter plots of the audiovisual MOS against the predicted audiovisual MOS obtained by CDAVQA model using objective video DMOS estimated with STN-VQM and objective audio MOS estimated with the proposed AQA algorithm in Plymouth AVQA database.

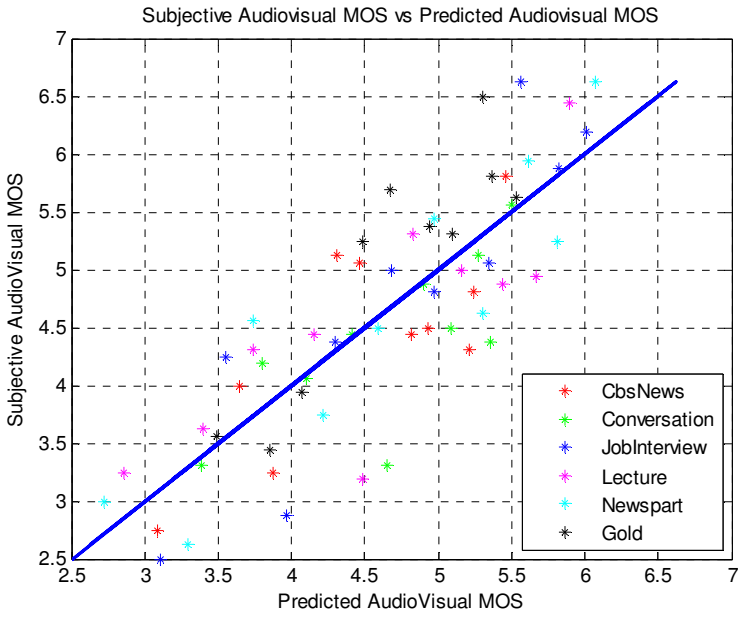

Figure 4-22: Scatter plot of subjective audiovisual MOS against predicted audiovisual MOS by CDAVQA model in Plymouth AVQA database.

Moreover, we investigated the performance of DAVQA and CDAVQA models when objective scores obtained by STN-VQM and our AQA model are utilized. We also compared their performance to full-reference objective VQA metric, PSNR, and full-reference objective AQA metric, PESQ, and their combination on the University of Plymouth AVQA database.

Table 4-14: Performance comparison of DAVQA and CDAVQA models when objective video and objective audio scores are utilized

Model	***Audiovisual Data***	***PCC***
PESQ	Objective Audio MOS	0.679
PSNR	Objective Video MOS	0.678
PESQ + PSNR	Objective Video/Audio MOS	0.814
PESQ.PSNR	Objective Video/Audio MOS	0.789
PESQ + PSNR + PESQ.PSNR	Objective Video/Audio MOS	0.788
DAVQA	Objective Video/Audio MOS	0.802
CDAVQA	Objective Video/Audio MOS	0.829

As can be observed from the results in Table 4-14, it is impressive that CDAVQA outperforms other models in the same table. Moreover, both PSNR and PESQ are full-reference metrics and have full-access to the reference audiovisual data whereas CDAVQA is a no-reference AVQA model.

Figure 4-23, Figure 4-24, Figure 4-25 and Figure 4-26 illustrate the correlation of each class when objective video DMOS estimated with STN-VQM and objective audio MOS estimated with the proposed AQA algorithm are employed. The correlation coefficients for each class are presented in Table 4-15.

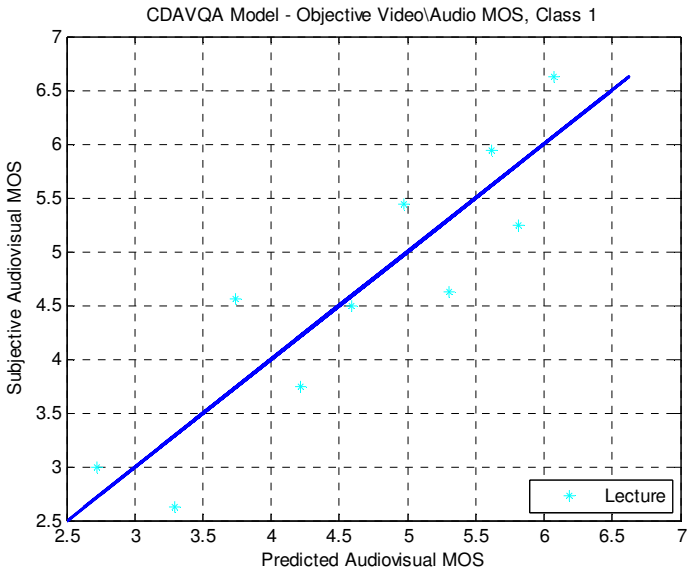

Figure 4-23: Scatter plot of subjective audiovisual MOS against predicted audiovisual MOS by CDAVQA model for videos classified as Class 1 in Plymouth AVQA database.

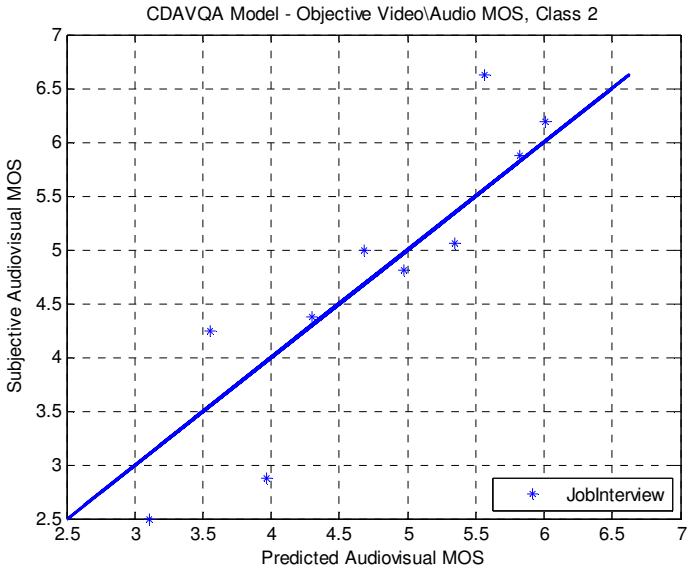

Figure 4-24: Scatter plot of subjective audiovisual MOS against predicted audiovisual MOS by CDAVQA model for videos classified as Class 2 in Plymouth AVQA database.

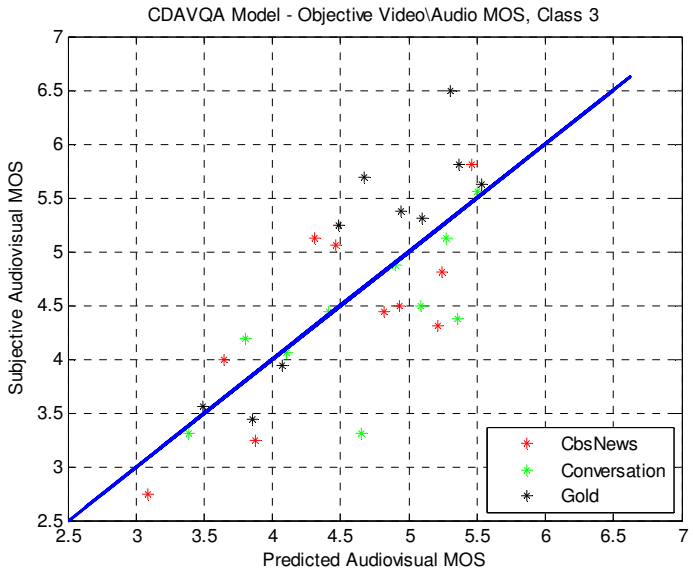

Figure 4-25: Scatter plot of subjective audiovisual MOS against predicted audiovisual MOS by CDAVQA model for videos classified as Class 3 in Plymouth AVQA database.

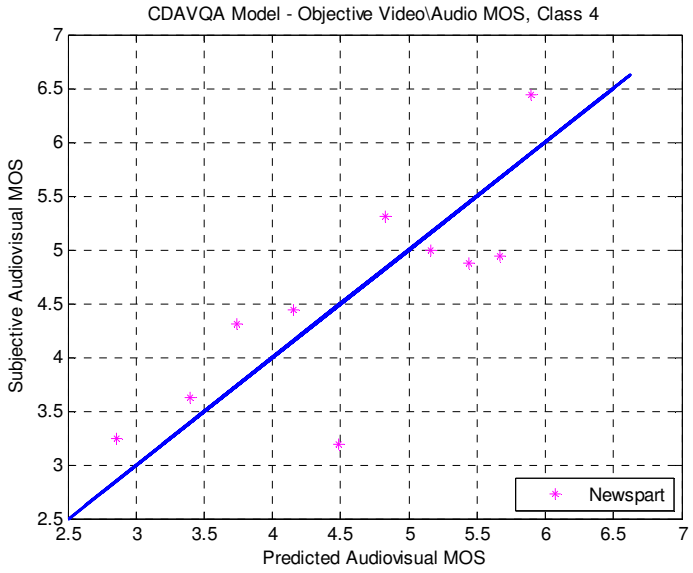

Figure 4-26: Scatter plot of subjective audiovisual MOS against predicted audiovisual MOS by CDAVQA model for videos classified as Class 4 in Plymouth AVQA database.

Table 4-15: Performance of CDAVQA model for each class when objective video and objective audio scores are utilized

Class	***Audiovisual Data***	***PCC***	***SROCC***
Class 1	Objective video MOS / Objective Audio MOS	0.895	0.891
Class 2	Objective video MOS / Objective Audio MOS	0.901	0.939
Class 3	Objective video MOS / Objective Audio MOS	0.768	0.762
Class 4	Objective video MOS / Objective Audio MOS	0.800	0.770

As it can be seen in Table 4-15, correlation coefficients of Class 1 and Class 2 are higher than correlation coefficients of Class 3 and Class 4. However, the correlation coefficients of Class 3, which contains three video contents, are quite satisfactory. Moreover, it can be deduced that if we had classified video contents into more than 4 classes, the performance of CDAVQA would have been higher.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1. Summary of the Thesis

In this thesis, a novel, spatiotemporal, bit rate, and packet loss structured, objective NR VQA metric, STN-VQM, has been proposed. STN-VQM is designed to predict perceived video quality degraded by compression and transmission distortions. STN-VQM is developed for videos of various contents with different spatial resolutions, bit rates, frame rates etc. assuming that if there is a packet loss, it occurs randomly. STN-VQM has been trained on the LIVE VQA database and evaluated on LIVE, EPFL-PoliMI and IT-IST VQA databases. STN-VQM has also been evaluated on University of Plymouth AVQA database. STN-VQM is shown to produce robust and accurate estimates for DMOS on these VQA databases, which span a wide range of video contents, spatial resolutions, bit rates, frame rates, packet losses etc. Comparison to the existing state-of-the-art FR and NR VQA metrics indicates that the STN-VQM provides promising results.

In addition, a novel sampling rate, bit rate and packet loss structured, objective NR AQA metric has been introduced. This AQA metric is developed for audio signals with different sampling rates, bit rates etc. assuming that if there is a packet loss, it occurs randomly. This AQA metric has been trained and evaluated on two different AQA databases. The AQA metric provides accurate estimates on these AQA databases, which have different audio encoding types.

Moreover, designing an objective NR AVQA metric to be used in order to predict audiovisual quality degradations due to compression and transmission distortions in multimedia streaming applications is aimed. This AVQA metric is expected to be robust to changes in video characteristics such as video content, codec, bit rate, frame rate, spatial resolution etc. It is also expected to be robust to changes in audio

codec, sampling rate and bit rate. Two different approaches are followed in order to obtain the AVQA model. First, we followed the classical approach in the literature, which we name as DAVQA, and obtained audiovisual quality as a combination of perceived video quality estimate, perceived audio quality estimate and their product. In the second approach (which we propose and name as CDAVQA), we again obtained perceived audiovisual quality estimate as a linear combination of perceived video quality estimate, perceived audio quality estimate and their product. However, we used different coefficient set for each class, which is obtained by a video spatiotemporal characteristics based classification algorithm proposed by us. This classification algorithm is shown to classify videos subject to compression and/or transmission distortions successfully. Results of DAVQA and CDAVQA are compared using both subjective and objective video and audio MOS on University of Plymouth AVQA database. Based on this comparison, CDAVDA is proven to be more accurate than DAVQA.

Nevertheless, there are some issues not addressed in our AVQA model, CDAVQA. To begin with, although STN-VQM, proposed AQA model and CDAVQA are evaluated on many different quality assessment databases, they are still based on limited data. Proposed quality assessment models provide promising results for various content audiovisual materials with different video characteristics such as video content, spatial resolution, bit rate, frame rate, packet loss ratio etc. and with different audio characteristics such as audio content, sampling rate, bit rate and packet loss ratio. Although mentioned audiovisual characteristics cover a wide range of audiovisual material, it may not span the audiovisual material space, i.e., there may be some other factors contributing to the perceived audiovisual quality.

In addition, burst packet losses are not addressed in the proposed quality assessment models. Moreover, effects of freezing and re-buffering on perceived audiovisual quality have not been considered in the proposed AVQA model. Another important quality degrading factor is the synchronization problems in the video and accompanying audio. This factor also has not been taken into account in the proposed AVQA model. The proposed AVQA model may also fail in assessing the audiovisual quality of some specific synthetic audiovisual material. To illustrate,

assume that a synthetic audiovisual data is obtained by combining a high quality video and a high quality audio, which belongs to another video content. Although, this content mismatch may significantly disturb viewers, the proposed AVQA model will most probably decide the audiovisual data as high quality since it does not check whether video and audio contents are compatible. This fact indicates that perceived audiovisual quality estimation is still an open research are in both subjective and objective domains.

5.2. Conclusions

With the advances in multimedia applications, end user satisfaction becomes an important issue. Among many subjective and objective factors, perceived video quality is believed to be the most important factor contributing to end user satisfaction. Hence, there is an obvious need in accurate perceived quality prediction. However, it is not trivial to estimate perceived video quality since there are many different components in video systems contributing to perceived quality such as capture and display devices, codecs, routers etc. Moreover, modeling human visual system (HVS) is very complicated. The best solution to VQA is conducting subjective tests. However, subjective tests are time- and cost-expensive, therefore, they are not applicable in online applications. Classical approaches such as PSNR and MSE are known to fail in appraising perceived quality since these methods treat data without considering what the data represent visually. As a result, development of accurate VQA models is necessary. Preferably, these models should not need the reference video signal in order to be used in multimedia streaming applications.

In order to assess video quality accurately, perceived video quality degradation sources such as compression and transmission should be analyzed. Compression is known to result in spatial and temporal distortion, whereas the most disturbing transmission distortion is believed to be packet losses. Moreover, compression and transmission distortions are expected to affect perceived video quality in a different way for different video contents. Since HVS is not perfectly explored, it is not possible to perfectly imitate HVS. In this thesis, we show that STN-VQM is an

accurate VQA model capable of estimating spatial and temporal degradations caused by compression and transmission distortions.

It is known that perceived audio quality also contributes to the perceived audiovisual quality. Hence, in order to assess audiovisual quality, audio quality should also be accurately estimated. Similar to the video case, perceived audio quality is subject to distortions occurring in compression and transmission. In this thesis, we show that sampling rate and bit rate are strong indicators of compression distortion. Moreover, we show that packet loss ratio can be used in order to model the impact of transmission distortions on perceived audio quality.

Finally, perceived audiovisual quality is generally predicted using perceived video and audio quality estimates. The most common approach is using linear combination of video quality, audio quality and their product. Although, this approach is proven to produce audiovisual quality estimates correlated with subjective tests, there is a strong consensus on the fact that audiovisual content should be considered while evaluating perceived audiovisual quality estimate. The reason is that the dominance of video quality on audiovisual quality is known to depend on the content. It is evident that the audio quality is more important in multimedia applications such as video call, news, interview etc. In this thesis, we proposed a video spatiotemporal characteristics based content classification algorithm which is shown to be robust to both compression and transmission distortions. We also proved that audiovisual quality estimates become more accurate when the audiovisual content classification is taken into account.

CDAVQA model can be used in different ways. To illustrate, results of CDAVQA model can be used in order to control audiovisual data transfer rate in multimedia streaming applications. To illustrate, assume that a highly compressed audiovisual data is transferred over a network with certain bandwidth. In this case, packet losses may seldom occur since the size of the audiovisual data is relatively low due to high compression. However, when the compression amount decreases, the frequency of packet losses may increase since the size of the audiovisual data to be transferred increases. This increase in packet loss ratio may result in lower perceived quality. As

it is seen, even higher compression may cause higher perceived audiovisual quality when the network characteristics are not considered. In such situations, if the packet loss ratios may somehow be accurately estimated before transferring different amount of audiovisual data over the same network, the optimum compression amount resulting in the highest audiovisual perceived quality may be selected.

Another scenario in which CDAVQA is employed may be the following. Results of CDAVQA model can also be used in order to decide optimal tradeoff between video and audio bit rate budget allocation for audiovisual material transfer over a network with certain bandwidth to achieve the maximum perceived audiovisual quality. Some studies indicate that for complex video contents, an increase in audio bit rate should be preferred to the same amount of increase in video bit rate for higher perceived audiovisual quality [70]. This may not seem reasonable at first. Nonetheless, the reason is that an increase in video bit rate may improve perceived video quality negligibly whereas the same amount of increase in audio bit rate improves perceived audio quality significantly.

5.3. Future Work

As a future work, STN-VQM and the proposed AQA model may be improved by means of adding new features in order to cope with packet losses which are not random. Furthermore, freezing, re-buffering and video-audio synchronization are important issues that must be considered. Moreover, algorithms capable of deciding whether video and audio contents are compatible may be inserted into the CDAVQA model. Obviously, the effect of these distortion types on the multimodal quality is required to be examined. For this purpose, subjective tests investigating the impact of mentioned factors on perceived audiovisual quality may be conducted. In addition, performance of CDAVQA may be improved by extending video content classification algorithm by adding specific audio characteristics.

REFERENCES

[1] ITU-T P.10/G.100, Vocabulary for performance and quality of service: *Amendment 1New Appendix I – Definition of Quality of Experience (QoE)*, International Telecommunication Union, Jan. 2007.

[2] A. J. Ahumada, Jr. and C. H. Null, "Image quality: A multidimensional problem," *in Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 141–148.

[3] S. A. Klein, "Image quality and image compression: A psychophysicist's viewpoint," *in Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 73–88.

[4] S. Jumisko-Pyykkö, J. Häkkinen, and G. Nyman, "Experienced quality factors—Qualitative evaluation approach to audiovisual quality," *in Proc. SPIE Multimedia on Mobile Devices*, San Jose, CA, January 28–31, 2007, vol. 6507.

[5] J. You, A. Perkis, M. Gabbouj, and M. M. Hannuksela, "Perceptual quality assessment based on visual attention analysis", *in Proc. ACM Int. Conf. Multimedia (MM)*, Beijing, China, Oct. 2009, pp. 561-564.

[6] ITU-T Rec. P.910, Subjective video quality assessment methods for multimedia applications, *International Telecommunication Union*, Geneva, Switzerland, 1999.

[7] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based objective quality metrics for audio-visual services - A survey," in *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 482-501, 2010.

[8] S. S. Hemami, and A. R. Reibman, "No reference image and video quality estimation: Applications and human motivated design," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 469-481, 2010.

[9] ITU-T J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference," *International Telecommunication Union*, Aug. 2008.

[10] B. Girod, "What's wrong with mean-square error," *in Digital Images and Human Vision*, A. B. Watson, Cambridge, MA: MIT Press, pp. 207-220, 1993.

[11] S. Winkler, and P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 660-668, Sep. 2008.

[12] S. Daly, "The visible differences predictor: An algorithm for the Assessment of Image Fidelity", *in Digital Images and Human Vision*, A. B. Watson, Cambridge, MA: MIT Press, pp. 179-206, 1993.

[13] M. H. Pinson, and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312-322, Sep. 2004.

[14] M. C. Farias, S. K. Mitra, "No-reference video quality metric based on artifact measurements", *in Image Processing (ICIP), International Conference on*, Vol. 3, IEEE, 2005.

[15] X. Lin, X. Tian, Y. Chen, "No-reference video quality assessment based on region of interest", *in Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on*, IEEE, 2012, pp. 1924-1927.

[16] K. Zhu, V. Asari, D. Saupe, "No-reference quality assessment of H.264/AVC encoded video based on natural scene features", *in SPIE Defense, Security, and Sensing, International Society for Optics and Photonics*, 2013, pp. 875505-875505.

[17] K. Zhu, K. Hirakawa, V. Asari, D. Saupe, “A no-reference video quality assessment based on laplacian pyramids”, *in Image Processing (ICIP), 2013 20th IEEE International Conference on*, 2013, pp. 49-53.

[18] A. Eden, “No-reference estimation of the coding PSNR for H. 264-coded sequences”, *Consumer Electronics, IEEE Transactions on 53 (2),* 2007, pp. 667-674.

[19] T. Brandao, M. P. Queluz, “No-reference image quality assessment based on DCT domain statistics”, *Signal Processing 88 (4)*, 2008. pp. 822-833.

[20] T. Brandao, M. P. Queluz, “No-reference quality assessment of H. 264/AVC encoded video”, *Circuits and Systems for Video Technology, IEEE Transactions on 20 (11),* 2010, pp. 1437-1447.

[21] M. Naccari, M. Tagliasacchi, S. Tubaro, “No-reference video quality monitoring for H. 264/AVC coded video”, *Multimedia, IEEE Transactions on 11 (5),* 2009, pp. 932-946.

[22] G. Valenzise, S. Magni, M. Tagliasacchi, S. Tubaro, “Estimating channel-induced distortion in H.264/AVC video without bitstream information”, *in Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, IEEE, 2010, pp. 100-105.

[23] N. Staelens, N. Vercammen, Y. Dhondt, B. Vermeulen, P. Lambert, R. Van de Walle, P. Demeester, Viqid: “A no-reference bit stream-based visual quality impairment detector”, *in Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, IEEE, 2010, pp. 206-211.

[24] S. Argyropoulos, A. Raake, M.-N. Garcia, P. List, “No-reference bit stream model for video quality assessment of H.264/AVC video based on packet loss visibility”, *in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, IEEE, 2011, pp. 1169-1172.

[25] S. Argyropoulos, A. Raake, M. Garcia, P. List, “No-reference video quality assessment for SD and HD H.264/AVC sequences based on continuous estimates of packet loss visibility”, *in Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, IEEE, 2011, pp. 31-36.

[26] I. Sedano, K. Brunnstrom, M. Kihl, A. Aurelius, “Full-reference video quality metric assisted the development of no-reference bitstream video quality metrics for real-time network monitoring”, *EURASIP Journal on Image and Video Proce*ssing 2014 (1), 2014, pp. 1-15.

[27] C. Keimel, M. Klimpke, J. Habigt, K. Diepold, “No-reference video quality metric for HDTV based on H.264/AVC bitstream features, i*n Image Processing (ICIP), International Conference on*, IEEE, 2011, pp. 3325-3328.

[28] X. Lin, H. Ma, L. Luo, Y. Chen, “No-reference video quality assessment in the compressed domain”, *Consumer Electronics, IEEE Transactions on 58* (2, 2012, pp. 505-512.

[29] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, P. Demeester, “Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression”, *Circuits and Systems for Video Technology, IEEE Transactions on* 23 (8), 2013, pp. 1322-1333.

[30] T. Liu, N. Narvekar, 523 B. Wang, R. Ding, D. Zou, G. Cash, S. Bhagavathy, J. Bloom, “Real-time video quality monitoring”, *EURASIP Journal on Advances in Signal Processing 2011 (1),* 2011, pp. 1-18.

[31] Recommendation ITU-T G.1070, Opinion model for video-telephony applications (2007).

[32] S. Zhao, H. Jiang, Q. Cai, S. Sherif, A. Tarraf, “Hybrid framework for no-reference video quality indication over LTE networks,” *in Wireless and Optical Communications Conference, IEEE*, 2014

[33] "ITU-R Recommendation BS.1116-1 methods for subjective assessment of small impairments in audio system including multichannel sound systems," 1997.

[34] A. W. Rix, J. G. Beerends, Kim Doh-Suk, P. Kroon, and O. Ghitza, "objective assessment of speech and audio quality - technology and applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, nol. 14, pp. 1890-1901, 2006.

[35] "ITU-R recommendation BS.1387-1.Methods for Objective Measurements of Perceived Audio Quality," 2001

[36] D. Campbell, E. Jones, and M. Glavin, "Audio quality assessment techniques-A review, and recent developments," *Signal Processing*, Vol. 89, pp. 1489-1500, Aug 2009.

[37] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. 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*, 2001 vol. 2, pp. 749-752.

[38] “Analysis and interpretation of INMD voice-service measurements,” 2000, ITU-T P.562.

[39] A. Clark, “Description of VQMON algorithm,” 2003, ITU-T del. cont. COM12-D105.

[40] S. Broom, “VoIP quality assessment: taking account of the edge-device,” *in IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1977–1983, No. 2006.

[41] M. Werner, T. Junge, and P. Vary, “Quality control for AMR speech channels in GSM networks,” *in Proc. IEEE ICASSP*, 2004, pp. 1076–1079.

[42] “The E-model, a computational model for use in transmission planning,” 2002, ITU-T G.107.

[43] L. Malfait, J. Berger, and M. Kastner, "P.563 The ITU-T standard for single-ended speech quality assessment," *in Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 14, pp. 1924-1934, 2006.

[44] Schroeder M R, Atal Bs, and Hall J.L., "Objective measure of certain speech signal degradations based on masking properties of human auditory perception," *Frontiers of Speech Communication Research*, 1979.

[45] Brandenburg and Karlheinz, "Evaluation of Quality for Audio Encoding at Low Bit Rates," *Audio Engineering Society Convention 82*, p. 2433, 1987.

[46] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, “Objective measures of speech quality,” Englewood Cliffs, NJ: Prentice-Hall, 1988.

[47] “Objective quality measurement of telephone-band (300–3400 Hz) speech codecs,” 1998, ITU-T P.861.

[48] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio systems," *in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, 1985 Vol. 10, pp. 608-611.

[49] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *Selected Areas in Communications, IEEE Journal on*, Vol. 10, pp. 819-829, 1992.

[50] Beerends, John G, Stemerdink, and Jan A., "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc*, Vol. 40, pp. 963-978, 1992.

[51] C. D. Creusere, K. D. Kallakuri, and R. Vanam, "An Objective Metric of Human Subjective Audio Quality Optimized for a Wide Range of Audio Fidelities," *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 16, pp. 129-136, 2008.

[52] U. Engelke and H. J. Zepernick, "Perceptual-based Quality Metrics for Image and Video Services: A Survey," *in Next Generation Internet Networks, 3rd EuroNGI Conference on*, 2007 pp. 190-197.

[53] R. Vanam and C. D. Creusere, "Evaluating low bitrate scalable audio quality using advanced version of PEAQ and energy equalization approach," *in Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005 Vol. 3, pp. 189- 192 Vol. 3.

[54] C. D. Creusere and J. C. Hardin, "Assessing the Quality of Audio Containing Temporally Varying Distortions," *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 19, pp. 711-720, 2011.

[55] L. Abanto, G. Kemper, and J. Telles, "A novel fuzzy logic-based metric for audio quality assessment: Objective audio quality assessment," *in Telecommunications (CONATEL), 2011 2nd National Conference on*, 2011 pp. 1-10.

[56] J. Liang and R. Kubichek, “Output-based objective speech quality,” *in Proc. IEEE Vehicular Technol. Conf.*, Stockholm, Sweden, 1994, pp. 1719–1723.

[57] H. Hermansky, “Perceptual linear prediction (PLP) analysis of speech,” *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[58] G. Talwar and R. Kubichek, “Output based speech quality measurement using hidden Markov models,” *in Proc. Int. Signal Process. Conf.*, Dallas, TX, 2003.

[59] T. H. Falk, Q. Xu, and W.-Y. Chan, “Non-intrusive GMM-based speech quality measurement,” *in Proc. IEEE ICASSP*, Philadelphia, PA, 2005, pp. 125–128.

[60] T. H. Falk and W.-Y. Chan, “Single-ended speech quality measurement using machine learning methods,” *in IEEE Trans. Audio, Speech, Lang. Process*., vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[61] P. Gray, M. P. Hollier, and R. E. Massara, “Non-intrusive speech quality assessment using vocal tract models,” *Inst. Elect. Eng. Proc. Vis. Image Sig. Process.*, vol. 147, no. 6, pp. 493–501, 2000.

[62] J. G. Beerends, P. Gray, A. P. Hekstra, and M. P. Hollier, “Call for proposals for a single-ended speech quality measurement method for non-intrusive measurements on live voice traffic,” 2000, ITU-T cont. COM12-C11.

[63] D.-S. Kim, “ANIQUE: An auditory model for single-ended speech quality estimation,” *in IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 821–831, Sep. 2005.

[64] D.-S. Kim, “A cue for objective speech quality estimation in temporal envelope representations,” *in IEEE Signal Process. Lett.*, vol. 11, pp.849–852, 2004.

[65] “Single-ended method for objective speech quality assessment in narrow-band telephony applications,” 2004, ITU-T P.563.

[66] ITU-T Rec. ITU-R BT-500, Methodology for the subjective assessment of the quality of the television pictures, *International Telecommunication Union*, March 2000.

[67] ITU-T Rec. P.800, Methods for subjective determination of transmission quality, *International Telecommunication Union*, 2001.

[68] A. Borowiak, U. Reiter, U. P. Svensson, "Quality evaluation of long duration audiovisual content," *in IEEE Consumer Communications and Networking Conference*, pp. 337-341, 2009.

[69] M. H. Pinson, L. Janowski, R. Pépion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. L. Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective tests: An international study," *in IEEE Journal of Selected Topics in Signal Processing*, Vol. 6, No. 6, October 2012.

[70] S. Winkler, C. Faller, "Perceived Audiovisual Quality of Low-Bitrate Multimedia Content", *IEEE Trans. Multimedia*, Vol. 8, no. 5, pp. 973- 980, 2006.

[71] M. H. Pinson, W. Ingram, and A. Webster, "Audiovisual quality components," i*n IEEE Signal Processing Magazine*, November 2011.

[72] ANSI-Accredited Committee T1 Contribution, "Report on an experimental combined audio/video subjective test method," *Bellcore, T1A1.5/93-104*, Red Bank, New Jersey, July 22, 1993.

[73] ANSI-Accredited Committee T1 Contribution, "Report on extension of combined audio/video quality model," *Bellcore, T1A1.5/94-141*, Red Bank, New Jersey, July 22, 1993.

[74] ANSI-Accredited Committee T1 Contribution, "Combined A/V model with multiple audio and video impairments," *Bellcore, T1A1.5/94-124*, Red Bank, New Jersey, Apr. 10, 1995.

[75] C. Jones and D. Atkinson, "Development of opinion-based audiovisual quality models for desktop video-teleconferencing," *in Proc. Rec. 6th IEEE Int. Workshop Quality of Service*, Napa, CA, 1998.

[76] "Study of the influence of experimental context on the relationship between audio, video, and audiovisual subjective qualities," *ITU-T Contribution COM12- 61-E*, France Telecom/CNET, France, Sept. 1998.

[77] "Relations between audio, video, and audiovisual quality," KPN Research, The Netherlands, *ITU-T Contribution COM 12-19-E*, Feb. 1998.

[78] D.S.Hands, "A Basic Multimedia Quality Model", *in IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806- 816, Dec. 2004.

[79] J.G. Beerends and F.E. de Caluwe, "The Influence of Video Quality on Perceived Audio Quality and Vice-Versa", *in J. Audio Engineering Society*, vol. 47, no. 5, pp. 355-362, May 1999.

[80] M. N. Garcia and A. Raake, "Impairment-factor based audio-visual quality model for IPTV," *in Proc. Int. Workshop Quality of Multimedia Experience*, pp. 1–6, 2009.

[81] M. McFarland, M. Pinson, C. Ford, A. Webster, W. Ingram, S. Hanes, and K. Anderson, "Relating audio and video quality using CIF video," *NTIA TM-10-472*, September 2009.

[82] T. C. Thang, J. W. Kang, Y.M. Ro, "Graph-based perceptual quality model for audiovisual contents," *in Multimedia and Expo, International Conference on,* 2007.

[83] W. Yaodu, X. Xiang, K. Jingming, “A study on the impact of spatial and temporal degradations on audiovisual quality integration,” *in Proceedings of IC-NIDC*, pp. 887-892, 2009.

[84] J. You, J. Korhonen, and U. Reiter, “Audiovisual quality fusion based on relative multimodal complexity,” *in Image Processing (ICIP), International Conference on*, pp. 3337-3340, IEEE, 2011.

[85] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, M. Rupp,“Audivisual Quality Estimation for Mobile Streaming Services,” *in Proc. of 2nd Int. Symp. on Wireless Communications (ISWCS)*, pp. 173–177, Siena, Italy, Sep. 2005.

[86] A. Peregudov, E. Grinenko, K. Glasman, A. Belozertsev, “An Audiovisual quality model of compressed television materials for portable and mobile multimedia applications,” *in Consumer Electronics, IEEE International Symposium on,* 2010.

[87] M. Goudarzi, L. Sun , E. Ifeachor, “Audiovisual Quality Estimation for Video Calls in Wireless Applications”, *in Proc. IEEE GlobalTelecommunications Conference*, pp. 1-5, 2010.

[88] M. Ries, B. Gardlo, “Audiovisual quality estimation for mobile video services”, *Selected Areas in Communications, IEEE Journal on 28 (3)*, pp. 501-509, 2010.

[89] K. Yamagishi and S. Gao, “Light-weight audiovisual quality assessment of mobile video: ITU-T Rec. P.1201.1,” *in Multimedia Signal Processing, International Workshop on*, pp. 464-469, 2013.

[90] M. N. Garcia, P. List, S. Argyropoulos, D. Lindegren, M. Pettersson, B. Feiten, J. Gustafsson, A. Raake, “Parametric model for audiovisual quality assessment in IPTV: ITU-T Rec. P.1201.2,” *in Multimedia Signal Processing, International Workshop on*, pp. 482-487, 2013.

[91] M. Ries, ”Video Quality Estimation for Mobile VideoStreaming,” Doctoral thesis, INTHFT, Vienna Universityof Technology, Vienna, Austria, Oct. 2008, available at http://publik.tuwien.ac.at/files/PubDat 170043.pdf.

[92] M.P. Hollier, A.N. Rimmel, D.S. Hands, and R.M. Volcker, ”*Multimodal Perception”, in BT Technology Journal*, 17 (1), pp. 35-46, Jan. 1999.

[93] M. Rupp, “Video and Multimedia Transmissions over Cellular Networks”, Vienna: Willey, 2009.

[94] S. Tasaka, Y. Ishibashi, “Mutually Compensatory Property of Multimedia QoS,” *in Proc. of IEEE International Conference on Communications 2002*, vol. 2, pp. 1105–1111, NY, USA, 2002.

[95] C. Jones, D.J. Atkinson, “Development of Opinion-Based Audiovisual Quality Models for Desktop Video-Teleconferencing,” *6th IEEE International Workshop on Quality of Service*, Napa, CA, USA, May, 1998.

[96] Joint Video Team, “H.264/AVC software coordination,” *http://iphome.hhi.de/suehring/tml*, 2007, Date of last access: 10 March 2015.

[97] International Organization for Standardization, “MPEG-2 standard,” *http://standards.iso.org/ittf/PubliclyAvailableStandards*, 2005, Date of last access: 10 March 2015.

[98] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, “A subjective study to evaluate video quality assessment algorithms,” *in IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics*, 2010, pp. 75270H-75270H.

[99] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[100] M.C.Q. Farias, S.K. Mitra, "A Methodology for Designing No-Reference Video Quality Metrics", *in VPQM*, 2009.

[101] Video Quality Experts Group (VQEG), "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment", phase II, August 2003.

[102] M.H. Brill, J. Lubin, P. Costa, S. Wolf, J. Pearson, "Accuracy and Cross-Calibration of Video-Quality Metrics: New Methods from ATIS/ T1A1", *Signal Processing Image Commun*., vol. 19, pp 101–107, 2004.

[103] ATIS Technical Report T1.TR.77-2002, "Data and Sample Program Code to be Used with the Method Specified in T1.TR.72-2001 for the Calculation of Resolving Power of the Video Quality Metrics" in T1.TR.74-2001 and T1.TR.75-2001, January 2002.

[104] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, T. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel", *in Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, 2009, pp. 204-209.

[105] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics", *in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 2430-2433.

[106] Instituto Superior Tecnico of Instituto de Telecomunicacoes dataset, *http://amalia.img.lx.it.pt/~tgsb/H264_test,* Date of last access: 10 March 2015.

[107] D.M. Chandler and S.S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing 16(9)*, pp.2284-2298, 2007.

[108] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Amer. A*, vol. 24, no. 12, pp. B61–B69, 2007.

[109] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," in Signals, Systems and Computers", 2003. *Conference Record of the Thirty- Seventh Asilomar Conference on*, vol. 2, pp. 1398 – 1402, 2003.

[110] K. Seshadrinathan and A.C. Bovik, "Motion tuned spatiotemporal quality assessment of natural videos," *Image Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 335 –350, 2010.

[111] M. Dimitrievski, Z. Ivanovski, "Fusion of local degradation features for no-reference video quality assessment", *in: Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, 2012, pp. 1-6.

[112] X. Lin, H. Ma, L. Luo and Y. Chen, "No-reference video quality assessment in the compressed domain," in *Consumer Electronics, IEEE Transactions on*, vol. 58, no. 2, pp. 505-512, 2012.

[113] K. Zhu, V. Asari, D. Saupe, "No-reference quality assessment of H.264/AVC encoded video based on natural scene features", *Proc. SPIE 8755*, 2013, 875505-875505-11.

[114] A. Mittal, M. A. Saad, A. C. Bovik, "Zero shot prediction of video quality using intrinsic video statistics", *Proc. SPIE 9014*, 2014, 90140R-90140R-13.

[115] M. Saad, A. Bovik, C. Charrier, "Blind prediction of natural video quality," Image Processing, *IEEE .Transactions on 23 (3)*, 2014, 1352-1365.

[116] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, T. Ebrahimi, “Subjective Quality Assessment of H.264/AVC Video Streaming with Packet Losses”, *in EURASIP Journal on Image and Video Processing,*Volume 2011.

[117] M. F. C. Chin, “Video quality evaluation in IP networks”, M.Sc Thesis, April 2012.

[118] ITU-T Rec. P.911, Subjective audiovisual quality assessment methods for multimedia applications, *International Telecommunication Union*, Geneva, Switzerland, 1998.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Konuk, Barış

Nationality: Turkish (TC)

Date and Place of Birth: 5 December 1982, Ayvalık

Marital Status: Married

Phone: +90 536 610 02 06

Fax: +90 312 592 10 43

email: bkonuk@aselsan.com.tr

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Electrical and Electronics Eng.	2007
BS	METU Electrical and Electronics Eng.	2004
High School	Çankaya Atatürk Lisesi, Ankara	2000

WORK EXPERIENCE

Year	Place	Enrollment
2004-present	ASELSAN Inc. /TURKEY	Software Engineer

FOREIGN LANGUAGES

Advanced English, Intermediate German, Basic Spanish

PUBLICATIONS

Journal Papers

1. B. Konuk, E. Zerman, G. Nur, G. B. Akar, "A Spatio-Temporal Network Aware No-Reference Video Quality Prediction Metric," C*onsumer Electronics, IEEE Transactions on*. (under review)

Conference Papers

1. B. Konuk, E. Zerman, G. Nur, G. B. Akar, "Content Aware Audiovisual Quality Assessment", submitted to *22nd IEEE International Conference on Image Processing (ICIP)*, Quebec City, September 2015.

2. E. Zerman, B. Konuk, G. Nur, G. B. Akar, "A Parametric Video Quality Model Based on Source and Network Characteristics", *21st IEEE International Conference on Image Processing (ICIP)*, Paris, October 2014.

3. E. Zerman, G. B. Akar, B. Konuk, G. Nur "Referanssız Video Kalite Değerlendirme Metrikleri Üzerine Karşılaştırmalı Bir Çalışma – A Comparative Study on No-Reference Video Quality Assessment Metrics", *22nd IEEE Signal Processing and Communication Applications Conference*, Trabzon, April 2014.

4. B. Konuk, E. Zerman, G. Nur, G. B. Akar, "A Spatiotemporal No-Reference Video Quality Assessment Model", 2*0th IEEE International Conference on Image Processing (ICIP)*, Melbourne, September 2013.

5. B. Konuk, S. E. Hakyemez, "İki-Kanallı Zaman Serpiştirmeli Analog-Sayısal Çeviricilerde Uyumsuzluk Hataları için Denkleştirme Yöntemi – Equalization Method for Mismatch Errors in Two-Channel Time-interleaved Analog-to-Digital Converters", *21st IEEE Signal Processing and Communication Applications Conference*, Girne, April 2013.

6. B. Konuk, E. Zerman, G. Nur, G. B. Akar, "Kodlama Tabanlı Bozulmalar Üzerine Uzam-Zamansal Referanssız Video Kalite Değerlendirme Modeli – Spatiotemporal No-Reference Video Quality Assessment Model on Distortions Based on Encoding", *21st IEEE Signal Processing and Communication Applications Conference*, Girne, April 2013.

7. B. Konuk, E. Ener, M. Dursun, "Sayısal-Analog Çevirici Performans Ölçümü ve Test Altyapısı – Performance Evaluation and Test Infrastructure of Digital-Analog Converters", *21st IEEE Signal Processing and Communication Applications Conference*, Girne, April 2013.

8. B. Konuk, S. E. Hakyemez, "Calibration and Equalization Methods for Mismatch Errors in a High Frequency Two-channel Time-interleaved ADC", *in Proceedings 55th IEEE Int.Midwest Symposium on Circuits and Systems (MWSCAS)*, Boise, ID, August 2012.

9. B. Konuk, S. E. Hakyemez, "Analog-Sayısal Çeviricilerin Performans Analizi – Performance Analysis of Analog-to-Digital Converters", *20th IEEE Signal Processing and Communication Applications Conference*, Fethiye, April 2012.

GRANTS/AWARDS

2000-2004 Assoc. Prof. Dr. Bülent Kerim Altay Award (Fall 2000, Fall 2002, Spring 2002).
EE 213 Best Ten Projects Award (Spring 2001).
Nortel Networks Netaş Inc. Senior Year Program Scholar (2003-2004).

2005-2006 The Scientific and Technological Research Council of Turkey (TUBITAK) Graduate Program Scholar.