NANYANG TECHNOLOGICAL UNIVERSITY

Learning Based Signal Quality Assessment for Multimedia Communications

A thesis submitted to
School of Computer Engineering
Nanyang Technological University

by

Manish Narwaria

in Partial Fulfillment of the Requirement for the Degree of
Doctor of Philosophy (Ph.D)

2012

Abstract

Multimedia contents (including image/video, speech, audio, graphic and so on) can be affected by a wide variety of distortions during the process of acquisition, compression, processing, transmission, and reproduction which generally leads to loss of perceptual quality. As a result, signal quality assessment is an important component in today's multimedia communication systems. In this thesis, perceptual quality assessment algorithms are proposed for three important types of multimedia signals, namely image, video, and speech. This involves two crucial stages: (a) feature extraction/detection, and (b) feature pooling.

The first stage calls for investigation and analysis into appropriate and effective signal features to extract meaningful information and provide a compact representation of the signal with the regard of quality. This is crucial because the selected features form the basis of the resultant quality metric. In this thesis, we discuss and provide detailed analysis of features based on Singular Value Decomposition, 2D mel-cepstrum and phase of Fourier Transform for visual quality assessment. We analyse the advantages and disadvantages of these features with regards to prediction accuracy and complexity. We also investigate into mel filter bank energies as features for evaluating quality of noise-suppressed speech and provide justification for their effectiveness via theoretical and experimental analysis.

On the other hand, the second stage requires the determination of appropriate weights

for fusing the features into a single score that can accurately reflect the human judgement of perceptual quality. We tackle this by using machine learning techniques which have been successfully employed in numerous research areas (for example in computer vision tasks such as object localization/tracking/recognition) but have not been adequately addressed in the literature within the realm of objective quality evaluation. Their major advantage is the introduction of a more systematic pooling methodology thereby avoiding unrealistic assumptions imposed in existing pooling methods. In this thesis, we demonstrate that machine learning can be effective in quality assessment if proper signal features are detected. We also provide insights into machine learning based feature pooling by analyzing the system trained on subjective scores which quantify human perception.

The proposed algorithms have been validated on a large number of subjectively rated databases which are publicly available. We have performed careful experimental analysis (including within database and cross database tests) and demonstrated that the proposed schemes overall perform better than several relevant methods. The better alignment with human perception confirms the effectiveness of the algorithms proposed in this thesis.

Acknowledgements

First, I would like to express my sincere gratitude to my PhD. supervisor Dr. Weisi Lin, whose expertise and professional attitude have greatly influenced my academic career. I thank him for putting faith in my abilities to pursue independent research. I also acknowledge the funding received from the Singapore Ministry of Education Academic Research Fund (AcRF, Tier I) throughout my studies. I am grateful for the mentoring I received from Dr. C. Santosh Kumar, Dr. K. Narayanankutty and Dr. K.P. Soman during my undergraduate studies.

I thank Dr. Ian Vince McLoughlin, Dr. Sabu Emmanuel and Dr. Liang-Tien Chia for their suggestions and insightful advice on my research work. Sincere thanks are also expressed to Dr. A. Enis Cetin for fruitful collaborations and discussions. I also appreciate Dr. Philip Loizou for providing the speech database that was used for experiments. I also wish to acknowledge all the researchers who made their databases publicly available which greatly facilitated the experimental verifications reported in this thesis.

I express my appreciation to the Centre for Multimedia and Network Technology (CeMNet) at the School of Computer Engineering for providing excellent computing facilities for carrying out my research. I wish to thank my teammates for their support and encouragement: Liu Anmin, Manoranjan Paul, Chenwei Deng, Zhouye Gu, Yuming Fang, Lu Dong, Huan Yang, Nevrez Imamoglu. I also express my gratitude towards the lab staff for being extremely helpful.

I would not have been here today if it were not for the love and care of my family. I would like to greatly acknowledge my parents who have always supported me. Graduate studies would not have been fun if it were not for all my friends - who always supported and helped me in academic and non-academic matters. You know who you are; Thanks!

I would also like to thank in advance the thesis examiners for accepting to be part of the committee, and for their comments and suggestions to improve this thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

ACR	Absolute Category Rating
ANSI	American National Standards Institute
CSF	Contrast Sensitivity Function
CV	Cross Validation
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DMOS	Differential Mean Opinion Score
DPS	Direct Phase Similarity
DSCQS	Double-Stimulus Continuous Quality-Scale
DSIS	Double Stimulus Impairment Scale
DWT	Discrete Wavelet Transform
ETSI	European Telecommunications Standards Institute
FBEs	Filter Bank Energies
FR	Full Reference
FT	Fourier Transform
GMMs	Gaussian Mixture Models
HVS	Human Visual System
IFC	Information Fidelity Criterion
IQA	Image Quality Assessment
ITU-T	International Telecommunication Union-Telecommunication Standardization
JND	Just Noticeable Difference
JP2K	JPEG 2000
JPEG	Joint Photographic Experts Group
LAR	Locally Adaptive Resolution
MMSE	Minimum Mean Square Error
MNB	Measuring Normalizing Block
MOS	Mean Opinion Score
MOVIE	Motion based Video Integrity Evaluation Index
MPEG	Motion Picture Expert Group
MPQM	Moving Pictures Quality Metric
MSE	Mean Squared Error
NR	No Reference
PC	Pair Comparison
PDM	Perceptual Distortion Metric
PESQ	Perceptual Estimation of Speech Quality
POLQA	Perceptual Objective Listening Quality Assessment
PSNR	Peak Signal to Noise Ratio
PVQM	Perceptual Video Quality Measure
QoE	Quality of Experience
QoS	Quality of Service

RMSE	Root Mean Squared Error
RR	Reduced Reference
SNR	Signal to Noise Ratio
SS	Single Stimulus
SSACR	Single Stimulus Absolute Category Rating
SSCQE	Single Stimulus Continuous Quality Evaluation
SSIM	Structural Similarity Index Measure
SVD	Singular Value Decomposition
SVR	Support Vector Regression
SVs	Support Vectors
VA	Visual Attention
VDP	Visual Differences Predictor
VIF	Visual Information Fidelity
VoIP	Voice Over Internet Protocol
VQA	Video Quality Assessment
VQEG	Video Quality Experts Group
VQM	Video Quality Metric
VSNR	Visual Signal to Noise Ratio
VSSIM	Video Structural Similarity Measure
WGN	White Gaussian Noise
WIQ	Wireless Imaging Quality

Chapter 1

Introduction

1.1 Background and Motivation

The explosion in the number of computers and digital systems connected by networks such as the Internet has brought a flow of instant information into a large and increasing number of homes and businesses. Most of the information is in the form of digital multimedia signals as intuitive and faithful depiction of things in life and work. As a result, products (e.g. phone cameras) and services (e.g. windows media players, YouTube) based upon multimedia signals have grown at an explosive rate. Where low-cost telephony is concerned, VoIP has been gaining grounds rapidly. Recent technologies include cable VoIP, mobile VoIP (also known as wireless-VoIP), as well as conventional VoIP, where service providers, such as Skype and Vodafone, have gained wide popularity.

An important issue in multimedia communications is that of ensuring proper delivery/transmission of the multimedia contents from the producer to the consumer. However, the nature of transmission channels (e.g. lossy transmission networks) and the constraints arising out of limited resources (for instance, this prompts the need for compression) usually lead to loss of perceptual quality. This in turn will lower the

satisfaction and enjoyment level of the viewers/consumers for whom these multimedia contents are meant. Therefore signal quality assessment plays an important role in multimedia content delivery. Subjective viewing tests are the most reliable way of assessing perceptual quality. However they are time-consuming, cumbersome, expensive, and tend to be non-repeatable. As a result, they cannot be easily and routinely performed for many scenarios involving in-service or real-time applications (e.g. on-line monitoring of video quality in TV broadcasting). Therefore, objective quality assessment using computational models is an important part of today's multimedia communication systems. We consider three practical scenarios to demonstrate how quality metrics are useful.

First, consider the case of signal compression. In general, multimedia signals in uncompressed formats require excessive storage capacity and a huge transmission bit rate. For example, a single digital television signal in Consultative Committee of International Radio 601 format [1] requires a transmission rate of 216 Mega-bits per second. This is unacceptably high in bit rate for most practical purposes. Thus there is a need to reduce the data rate via coding, before digital television and video can be fed into the storage systems and communication networks. While coding ensures efficiency in terms of the information required to be transmitted, on the downside it will degrade the quality of the received/decoded signals as perfect signal reconstruction is usually not possible at the decoder side. To ensure a trade-off between the coding efficiency and perceptual quality, a quality metric forms an invaluable tool. Currently, the PSNR is widely used as the optimization criterion in video coding algorithms.

Second, we consider the area of information hiding [225], [229] where secret messages are embedded into images so that an unauthorized user cannot detect the hidden

messages. Because such an embedding process will degrade image quality, an image quality assessment (IQA) metric can help in guiding the optimization process between the desired quality and the strength of the message to be embedded.

Thirdly, regarding speech, additive noise is one of the most common factors that affects speech quality. Hence noise-suppression is employed frequently. For e.g., in mobile voice communication devices [105] which are used in an environment with a high level of ambient noise or in pay phones located in noisy environment (e.g., airports, busy street). Further, to benchmark the performance of different speech enhancement approaches [104], [224], a metric to assess the impact of the noise-suppression on the perceptual quality is necessary [4], [106].

Due to the widespread applications, a number of quality assessment algorithms for image, video and speech signals have been proposed and used over the past years. The Mean Squared Error/Peak Signal to Noise Ratio (MSE/PSNR) continues to enjoy wide acceptability as a quality metric due to its mathematical simplicity and ease of implementation. However, it is well known that the PSNR may not be always in accordance with the Human Visual System (HVS)'s perception [14], [230]. Consider the two images shown in Figure 1.1 which have the same PSNR (25.24 dB). Clearly, (a) looks much better than (b) to the human eye and this highlights the limitations of simple pixel based quality measures like PSNR.

To overcome the shortcomings of PSNR, many visual quality metrics have been proposed in the literature. However, no single visual quality assessment algorithm can perform best in all test cases (i.e. distorted images from multiple databases with varying image and distortion types/levels). Furthermore, some algorithms perform better for near-threshold distortions while others are good for supra-threshold distortions.

(a)

(b)

Figure 1.1 Two images with the same PSNR (25.24 dB).

For example VIF metric [44] usually has excellent performance for supra-threshold distortions while VSNR [45] is more suitable for near-threshold distortions. Computational complexity is another issue which plays a key role in the practical use of a visual quality metric. For example, although the metric MAD [118] achieves good prediction accuracy it has relatively higher computational burden. Computational costs in particular can be the major factor in determining the suitability of a VQA method for practical deployment. Lastly, a scalable algorithm i.e. whose performance can gracefully adjust according to the reference image information will be more useful. Most of the

existing FR algorithms are not scalable.

For speech quality assessment, the ITU has released P.862 PESQ [3] as the current standard for intrusive (it will be soon replaced by P.863 or POLQA [4]) and P.563 [5] as the current standard for nonintrusive speech quality assessment. However quality assessment of noise-suppressed speech is challenging for intrusive (i.e. FR) metrics due to the following reasons:

- They assume that the reference signal is of perfect quality.
- The test (i.e. processed) signal is of quality no better than the input (i.e. reference) signal.

The above assumptions are violated in case of noise-suppression: test signal is usually has higher quality, the reference signal (i.e. noisy signal) is of not perfect quality. Therefore nonintrusive assessment is the obvious alternative. However, it is found that the current ITU standard for nonintrusive speech quality assessment (P.563) is not accurate [9] in estimating quality of noise-suppressed speech. Therefore, more research effort is needed to develop a stand alone metric for assessing the perceptual effects of noise-suppression.

In summary, although the existing quality metrics have been found to be useful in many applications, they suffer from drawbacks and there still exists room for further improvement which can be explored in order to make the related products and services more effective, as well as enabling new functionalities. In this thesis, we attempt to address some of limitations of the existing visual and speech quality metrics.

1.2 Objective and Scope of This Work

The objective of this study is to develop new methods for quality evaluation of image, video and speech. To that end, we focus on the two crucial aspects in quality metric design, namely, feature detection and feature pooling. Both aspects are not straightforward given the complexities and intricacies involved in the way humans perceive signal quality. Moreover, the human brain comprises of sophisticated mechanisms which work in conjunction (rather than independently) to produce perception and our current knowledge of these is limited. In other words, direct and complete modeling of human perception is difficult. Therefore for a more effective and feasible solution, it is beneficial to exploit the relevant high and low level properties of the human perception system by employing signal processing techniques and fusing the resultant features via data-driven methods.

1.3 Thesis Contributions

As mentioned, quality evaluation can be modeled as a two-stage (for feature detection and feature pooling) process. The key contributions in this work are towards these two stages, and are briefly summarized as follows:

a) For the first stage (i.e., signal feature detection), we first investigate into features for visual signals (image and video), based on Singular Value Decomposition (SVD), the phase of Fourier Transform (FT) and the two-dimensional (2D) mel-cepstrum. We provide analysis and justification for their use in assessing visual quality. In particular, these features are effective as they account for relevant high level properties of the HVS (like sensitivity to *structural* changes). Furthermore, based on theoretical and experimental analysis, mel filterbank energies (FBEs) are employed

as features for evaluating the quality of noise-suppressed speech. They can capture the effects of noise injection and suppression reasonably well and can be exploited to quantify the effects of noise-suppression on speech quality. Our contributions to feature detection in signal quality evaluation are original.

b) To address the second stage (i.e., feature pooling), we employ machine learning based feature pooling because it is more systematic and the required weights are determined via training with substantial *ground truth* (i.e. subjective scores). As a result, it helps in avoiding unrealistic assumptions currently imposed in the existing feature pooling methods. It is therefore an attractive alternative to bridge the gap between the psychophysical ground truth and the realistic engineering solution. We believe it is beneficial due to following reasons: (1) the actual feature pooling mechanisms in the HVS are not well understood and quite complex to be implemented; (2) the training process uses the subjective viewing/listening results as the target scores, and as a result, we can expect to mimic the human perception indirectly, given a sufficiently large training set; (3) in-depth analysis of the model developed as result of training provides insights into how the system predicts quality. Our attempts are among the early ones to exploit machine learning in signal quality evaluation.

1.4 Organization of the Thesis

This thesis has been divided into 9 chapters as outlined as follows. Chapter 1 (this chapter) gives a brief introduction about the thesis, including the background and motivation, objective and scope, thesis contributions and thesis organization.

Chapter 2 describes the major related existing work and algorithms for assessing the

perceptual quality of image, video and speech. We survey the state-of-the-art quality assessment methods and outline their advantages and shortcomings. More specific literature survey to each proposed technique in this thesis will be further introduced whenever appropriate in Chapters 3-8.

Chapter 3 discusses the benefits of using SVD for visual quality assessment. With SVD, one can account for the structural changes better and hence achieve more accurate quality prediction.

Chapter 4 investigates feature pooling based on machine learning technique. Such pooling technique is more systematic as compared to existing methods which tend to be somewhat ad-hoc. It is also more convincing since the required weights are determined via proper training with ground truth (i.e. subjective scores).

Chapter 5 describes our new visual quality assessment metric using 2D mel-cepstrum. The relevant and useful properties of 2D mel-cepstrum for visual quality assessment are discussed and exploited for more efficient quality prediction.

Chapter 6 focuses on developing a FT based scalable quality measurement algorithm for image and video. The proposed metric accounts for the masking effects and unequal sensitivity of the HVS to changes in different frequency components.

In Chapter 7, a low complexity but effective approach for VQA is introduced and described. It uses the variation of quality along the temporal axis as a measure of the temporal quality.

In Chapter 8, we propose a new method for quality assessment of noise-suppressed speech. This approach uses mel FBEs as the speech features. We exploit their sensitivity to noise injection and suppression and provide theoretical analysis as ground for their use in the said task.

Lastly, Chapter 9 closes the thesis with a summary of the main research work performed and directions for further studies.

The Appendix provides the details of the subjectively rated image and video databases used in the experimental verifications.

Chapter 2

Literature Survey

In this chapter, we give a brief overview of the major relevant existing work in image, video and speech quality assessment. We also introduce the two-stage procedure for developing quality metrics and provide brief description of the developments regarding these two stages. In addition, we outline the disadvantages of the existing methods in order to provide the motivation for the remaining thesis. For better organization, we have divided this chapter into 3 separate sections: one for image, one for video and the last one for speech quality assessment (we also discuss the specific issues in quality assessment of noise-suppressed speech).

2.1 Image Quality Assessment

The rapid proliferation of digital imaging and communications technologies has given rise to a growing number of applications which yield images. In many cases, the end-user receives a distorted version of the original digital image (e.g., due to lossy compression, digital watermarking, packet loss), and it is, therefore, necessary to quantify the visual impact of the distortion by way of quality evaluation. Generally speaking, an image quality metric has three kinds of applications: First, it can be used to monitor image quality for quality control. For example, an image and video acquisition system can use a

quality metric to monitor and automatically adjust itself to obtain the best quality image and video data. A network video server can use it to examine the quality of the digital video transmitted on the network and control video streaming. Second, it can be employed to benchmark image processing systems and algorithms. Suppose we need to select one algorithm for a specific task, then a quality metric can help to evaluate which of them provides the best quality images. Third, it can be embedded into an image processing system to optimize the algorithms and the parameter settings.

Image quality can be measured in two different ways. The first, known as subjective quality assessment, consists of the use of human observers who should score image quality during experiments. The second one is called objective quality assessment which means the use of a computational model to predict image quality.

2.1.1 Subjective Image Quality Assessment

Subjective viewing tests are performed for various model design, tuning and verification. The ITU has standardized methods to conduct subjective viewing tests [8], [10]-[13], [109], [161]. This promotes acceptance and facilitates sharing of data among various laboratories, researchers and users. The standardized test methods are briefly described below:

a) *Double Stimulus Impairment Scale (DSIS) Method:* The reference (unimpaired image) is displayed before the test stimulus (impaired image), and each subject rates the test stimulus keeping in mind of the reference. The Mean Opinion Score (MOS) consists of a five-level impairment scale: 5- imperceptible, 4- perceptible but not annoying, 3- slightly annoying, 2- annoying, 1- very annoying. The DSIS is usually used in evaluating clearly visible impairments.

b) *Double-Stimulus Continuous Quality-Scale (DSCQS) Method:* The subject rates each

of the reference and impaired images separately without prior knowledge of which image is the impaired one. Each subject is provided with a vertical scale on which he/she marks the scores, which are normalized to a range of 0 to 100. Then, the difference between the results of the reference and the impaired ones are calculated as Difference Mean Opinion Scores (DMOS). This method is often used when the quality of test and reference stimuli are rather similar.

c) *Single Stimulus (SS) Method:* A single image or sequence of images is presented. The evaluation is based on either categorical or numerical scale. For the former, an image or image sequence is assigned to one set of categories that are defined in semantic terms; *e.g.*, excellent, good, fair, poor, and bad for image quality and imperceptible, perceptible (but not annoying), slightly annoying, annoying, and very annoying for image impairment. For the latter (numerical scale), a value is used to describe each shown image or image sequence.

Even though subjective assessment remains as the most accurate way of assessing image quality, it suffers from various drawbacks that limit its applicability. Firstly, it is time-consuming, laborious and expensive, since the resultant MOSs need to be obtained by many observers through repeated viewing sessions. Moreover, incorporation of subjective viewing tests is not feasible for on-line visual signal manipulations (such as encoding, transmission, relaying, etc.). Secondly, even in situations where human examiners are allowed (*e.g.*, visual inspection in a factory environment) and the manpower cost is not a problem, the assessment results still depend upon viewers' physical conditions, emotional states, personal experience, and the context of preceding display [17]-[19].

As a result of these limitations, it is necessary to build computational models to predict the evaluation of an average observer in a consistent and objective manner.

2.1.2 Objective Image Quality Assessment

The simplest and most widely used objective quality metrics are the MSE and PSNR; however, they can be poor predictor of visual quality [14], [230], especially when the noise is not additive. The major reason for the overall poor performance of MSE (or PSNR) is its assignment of equal importance to all the changes in a visual signal (image or video) regardless of their perceptual significance. Objective evaluation of picture quality in line with the human perception is a difficult task [15]-[19] due to the complex, multi-disciplinary nature of the problem (related to physiology, psychology, vision research, and computer science) and the limited understanding of the HVS mechanisms. There has not been a clear-cut and general scheme so far which can account for all the related characteristics of the HVS (please refer to [16]-[19] for recent reviews).

Objective IQA algorithms can be classified into 3 categories based on the amount of information used for predicting quality: (1) Full reference (FR) metrics which use complete reference image information, (2) Reduced reference (RR) metrics which use only partial information from the reference image and (3) No reference (NR) metrics which do not use any reference image information. FR metrics are generally more accurate while NR metrics although less accurate and usually distortion specific can be used when the reference image is not available. RR algorithms are essentially a trade-off between these two because only partial information of the reference image is required.

With regards to developing an IQA algorithm, it can be handled [16]-[19] by two broad approaches: i) the vision modeling approach and ii) the signal processing based or engineering approach. These two approaches and their advantages/disadvantages are discussed next.

2.1.2.1 Vision Modeling Approach

The vision modeling approach, as the name implies, is based on modeling various components of the HVS. The HVS-based metrics aim to simulate the processes of the HVS from the eye to the visual cortex. These metrics are intuitive and appealing since they attempt to account for the properties of the HVS relevant to perceptual quality assessment. The first image and video quality metrics were developed by Mannos et al. [20] and Lukas et al. [21]. Later the well-known HVS-based metrics are the Visual Differences Predictor (VDP) [22], the Sarnoff JND metric [23], Moving Pictures Quality Metric (MPQM) [24], and Perceptual Distortion Metric (PDM) [25].

Although the HVS-based metrics are attractive in theory, they may suffer from some drawbacks. The HVS comprises of many complex processes which work in conjunction rather than independently, to produce visual perception. However, the HVS-based metrics generally utilize results from psychophysical experiments which are typically designed to explore a single dimension of the HVS at a time. In addition, these experiments usually use simple patterns such as spots, bars, and sinusoidal gratings which are much simpler than those occurring in real images. For instance, psychophysical experiments characterize the masking phenomenon of the HVS by superposing a few simple patterns. In essence, these metrics suffer from drawbacks which mainly stem from the use of simplified models describing the HVS. Moreover these metrics generally depend on the modeling of the HVS characteristics which are not yet fully understood. While our knowledge about the HVS has been improving over the years, we are still far from a complete understanding of the HVS and its intricate mechanisms.

Furthermore, due to the complex and highly non-linear nature of the HVS, these

metrics can be complicated and time-consuming to be used in practice. The complexity of these models usually leads to high computational cost and memory requirement, even for images of a moderate size. In addition, the psychophysical experiments that underlie many error sensitivity models are specifically designed to estimate the threshold at which a stimulus is just barely visible. These measured threshold values are then used to define visual error sensitivity measures, such as the CSF and various masking effects. However, very few psychophysical studies indicate whether such near-threshold models can be generalized to characterize perceptual distortions significantly larger than threshold levels, as is the case in a majority of image processing situations. As it turns out, many of the IQA metrics based on vision modeling approach are less effective for suprathreshold distortions [16]-[19]. Owing to these limitations, the second approach namely the engineering approach has gained popularity during recent years and is described next.

2.1.2.2 The Engineering Approach

The engineering approach is based primarily on the extraction and analysis of certain features or artifacts in the video. These can be either structural image elements such as contours, or specific distortions that are introduced by a particular processing step, compression technology or transmission link, such as blocking artifacts. These metrics look at how pronounced these features are in the image/video to estimate overall quality. This does not necessarily mean that such metrics disregard human vision, as they often consider psychophysical effects as well, but image content and distortion analysis is the conceptual basis for their design rather than fundamental vision modeling. The metrics developed with this approach attempt to quantify visual quality based on the premise that a high-quality image is one whose structural content [2] most closely matches that of the reference image.

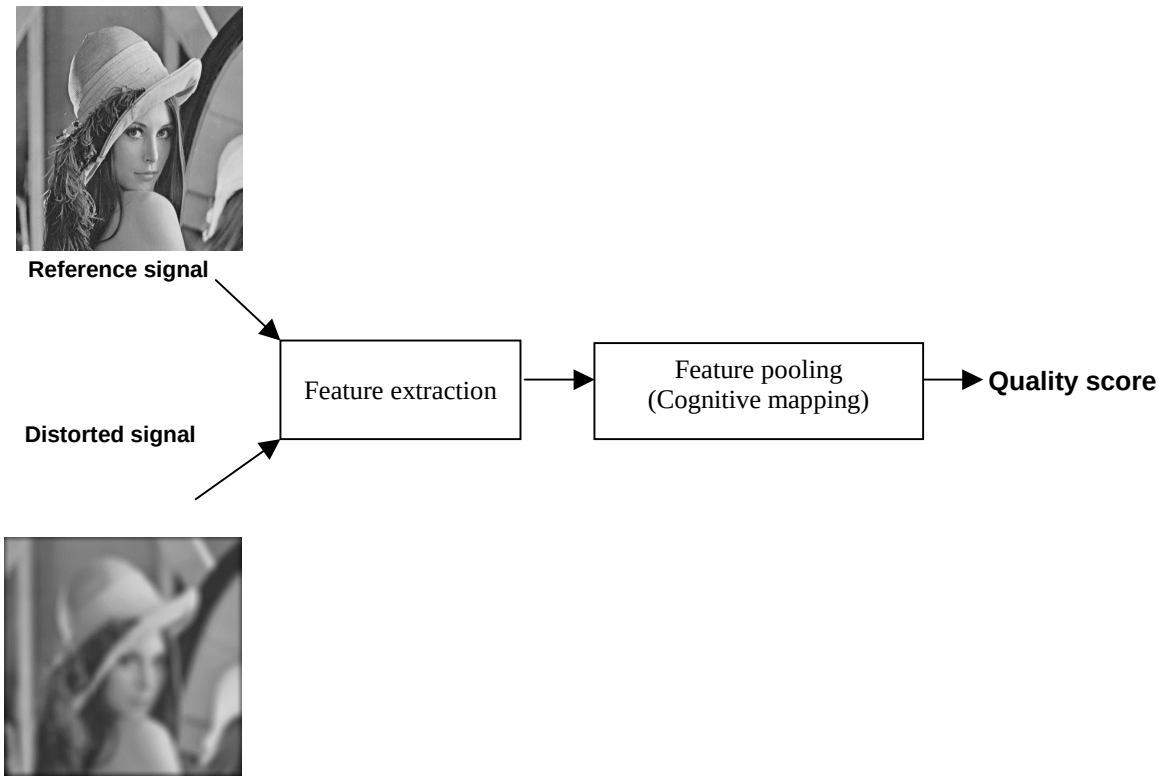


Figure 2.1: Representative diagram of the engineering approach to developing IQA/VQA algorithms (FR and RR)

These metrics do not use low-level properties of vision, but instead operate based on hypotheses of what the HVS attempts to achieve when shown a distorted image.

With the engineering approach, assessment of image quality can be considered as a two-stage process: (a) feature detection, (b) feature pooling. A representative diagram is shown in Figure 2.1. As for the first stage, the selected features have to be an effective representative of visual quality variations, while the second stage determines the relationship among different features and the perceived visual quality. Pertaining to the issue of feature extraction, in [26] a study was conducted to investigate and analyze the distortion criteria of the human viewing. It was found that for pictures where distortion is greater at edges, the MSE (or one of its relatives) is less satisfactory. Similar results have been also demonstrated in [27]-[28], where distortions at the edges have been

differentiated. As a result, image structure needs to be accounted for effective visual quality assessment. Therefore, during the recent years there has been a growing interest to take image structure into account for picture quality evaluation, because structural properties play a big role in the human perception [15], [29]-[31] as well as image content recognition [32]. Evaluating the loss of structure is therefore expected to give good estimate of visual quality degradation.

A well known FR metric is the SSIM [2], [33], which is mainly based on the idea of equating the perceived image distortion to the measurement of structural distortion. In SSIM, the mean of quality scores of individual image blocks gives the overall image quality score. Some other IQA metrics [34]-[51], [62]-[63], [236], [239]-[242] have also been proposed. The metric known as MSVD [37] evaluates quality of each image block based on the error in singular values.

A NR metric using the SVD of local image gradients has been proposed in [40] and used for proper selection of the parameters of image denoising algorithms. The authors in [41] proposed an SVD based method in which the difference between the reflection coefficients (obtained by projecting the two images onto the right singular vectors) of the original and distorted images are used for quality assessment. The method proposed in [42] also projects the distorted image on the singular vectors of the original image and uses a referee matrix of the distorted image to assess quality. In [36], the Harris response has been used to describe the geometric structure on a pixel-by-pixel basis with the overall quality score being computed by a simple averaging over all the image pixels. The algorithm proposed in [43] extracts structural features from the images. The overall quality score is then computed as a weighted sum of the features where the weights have been determined by subjective experiments.

Another FR IQA scheme known as the VIF index [44] has also been developed. It

equates perceptual quality to the amount of information regarding the reference image that can be extracted from the distorted image (images are modeled using Gaussian scale mixtures to measure the amount of image information). The VSNR proposed in [45] deals with both detectability of distortions (low level property of vision) and structural degradation based on the global precedence (mid-level visual property). Several other transform based FR methods have also been proposed in literature in various domains: frequency domain transforms like DCT and wavelets [39], discrete orthogonal transforms [46], contourlet transform [34], [47], wave atoms transform [62], Riesz transform [63], etc. The base idea of these methods is to compare the transformed image signal components (of reference and distorted images) because the transformation will usually help in better representation of the image signal. Another class of FR schemes employs image gradient ([48]-[49] for instance) to quantify image quality and is based on the idea that edges play an important role in perceiving image *structure*. The authors in [50]-[51] have explored the combination of multi-scale SSIM, VIF and R-SVD [41] algorithms to assess image quality.

As for the issue of feature pooling (also known as error pooling), literature survey shows that scant research effort has been directed to develop effective cognitive models to map the features into a quality score, with the major reason being the complexity and limited knowledge about the HVS. Researchers have employed techniques like simple summation based fusion, Minkowski combination, linear (i.e. weighted) combination, etc. to fuse the visual features into a quality score, and some examples have been already given above. These pooling techniques, however, impose constraints on the relationship between the features and the quality score. A simple summation or averaging of features implicitly constraints the relationship to be linear. Problems also arise with the simple average approach when the distortion is highly non-uniform over the image space. For

example, when only a small region in an image is corrupted with extremely annoying artifacts, but all other regions have high quality, human subjects tend to pay more attention to the low quality region and give an overall quality score lower than the average of the quality/distortion map. A weighted summation requires the determination of appropriate weighting coefficients and there is no general method available for this. Subjective experiments may be used to compute the weights [43] but such a method is less consistent and unsuitable for real-time applications. The use of Minkowski summation for spatial pooling of the features/errors implicitly assumes that errors at different locations are statistically independent. In addition, there is no systematic method to determine the proper/optimal value of the Minkowski summation exponent and is generally determined experimentally. Another method has been developed [52], [235] which involves weighting quality scores as a monotonic function of quality. The weights are determined by local image content, assuming the image source to be a local Gaussian model and the visual channel to be an additive Gaussian model. However, there is lack of convincing ground for these assumptions.

Recently, two pooling strategies have been proposed [53] for the SSIM. Instead of using a simple mean as the overall quality score, these approaches attempt to weigh the quality scores of different blocks based on visual importance. The first strategy is based on the idea that lower quality regions in images attract more attention than the ones with higher quality; the second strategy uses VA to provide weighting [54]-[58] which is based on the idea that certain regions attract more human attention than the others. The strategy of feature pooling using VA while intuitive may suffer from drawbacks due to the fact that it is not always easy to find regions that attract visual attention. Furthermore, improvement in quality prediction by using VA is not yet clearly established and still open to further investigations [54], [58]. One reason for this is that the perception in still

images varies with allowed observation time [17]. That is, if an observer has time long enough to perceive an image, every point of the image can become the attention center eventually. This may render direct VA based pooling less effective for perceptual IQA.

In summary, the existing feature pooling techniques tend to suffer from one or more drawbacks and there is a need for a more systematic and effective feature pooling strategy. This is one of the objectives of this study.

2.2 Video Quality Assessment

Like IQA, the most accurate approach to VQA is subjective assessment. The prominent subjective tests used for video from ITU-R Rec. BT.500-11 [13] and ITU-T Rec.P.910 [12] are:

a) *Double Stimulus Continuous Quality Scale* (*DSCQS*) [ITU-R Rec. BT.500-11] - In this test, the reference and processed video sequences are presented twice to the evaluators in alternating fashion, with randomly chosen order (Example: reference, degraded, reference, degraded). At the end of the screening, the evaluators are asked to rate the video quality on a continuous quality scale of 0–100 (with 0 being *Bad* and 100 *Excellent*). Multiple pairs of reference and processed video sequences and of rather short durations (around 10 seconds) are used. The evaluators are not told which video sequence is the reference and which is the processed.

b) *Double Stimulus Impairment Scale* (*DSIS*) [ITU-R Rec. BT.500-11] - Unlike the DSCQS, in the DSIS, the evaluators are aware of the presentation sequence, and each sequence is showed only once. The reference video sequence is shown first followed by the processed video sequence. The evaluators rate the sequences on a discrete five-level scale ranging from *very annoying* to *imperceptible* after watching the video

sequences. ITU-T Rec.P.910 has an identical method called Degradation Category Rating (DCR).

c) *Single Stimulus Continuous Quality Evaluation* (*SSCQE*) [ITU-R Rec. BT.500-11] - As the name suggests, the evaluators are only shown the processed video sequence, usually of long duration (typically 20–30 minutes). The evaluators rate the instantaneous perceived quality on the DSCQS scale of *bad* to *excellent* using a slider.

d) *Absolute Category Rating* (*ACR*) [ITU-T Rec.P.910] - This is also a single stimulus method similar to SSCQE with only the processed video being shown to the evaluators. The evaluators provide one rating for the overall video quality using a discrete five-level scale ranging from *Bad* to *Excellent*.

e) *Pair Comparison* (*PC*) [ITU-T Rec.P.910] - In this method, test clips from the same scene but under varying conditions, are paired in all possible combinations and screened to the evaluators for preference judgment about each pair.

Due to the previously mentioned drawbacks associated with subjective tests, objective VQA has attracted significant research attention in recent years [64]-[72], [74]-[83]. A straightforward and convenient approach to VQA is to use an IQA method on a frame-by-frame basis. The global quality score is then usually determined by simple average or Minkowski summation. Indeed, the widely used FR quality assessment metric the PSNR is applied on frame-by-frame basis. However, as already pointed out, it can be a poor predictor of visual quality [14], [230] especially when the distortion is non-additive in nature.

Unlike images, video signals carry information over spatial as well as the temporal domain. Therefore, the frame-level averaging of spatial quality alone may be insufficient since the temporal factors crucial to VQA are disregarded. As has been noted by many researchers, considering quality along the temporal axis is an important factor for VQA

[64]-[71], [74]-[79], [237]-[238]. Use of temporal factors for VQA has therefore been explored in the existing works. Like IQA, VQA methods are also usually either vision model based or signal processing based. Popular HVS-based VQA algorithms include the MPQM [24], PDM [25] and the Sarnoff JND vision model [23]. All of these explore the temporal dimension for VQA. As already highlighted in Section 2.1.2.1, the HVS based methods are less effective [16]-[19], [30], [66], [84]-[85] for supra-threshold distortions, i.e., the case when artifacts in the video sequences are clearly visible. Due to this limitation, another class of VQA algorithms have been explored which directly attempt to account for features associated with loss of visual quality (like blur, blockiness etc.). Such algorithms have been found to be better for supra-threshold distortions and as a result have received more research attention in recent years and we mention some of them below.

The PVQM proposed in [74] combines 3 factors which can characterize quality namely 'edginess' of the luminance, the normalized color error and the temporal decorrelation. The authors in [79] proposed a VQA algorithm called VQM. Due to its excellent performance in the VQEG Phase II validation tests, the VQM was adopted as a national standard by the American National standards Institute ANSI. In [67], temporal distortions such as mosquito noise were modeled as a temporal evolution of a spatial distortion in a scene, and visual attention mechanism was used for VQA. In [68], the method TetraVQM was proposed where motion estimation algorithm was used to take into account temporal errors. In this method, a degradation duration map is generated for each frame by analyzing the motion trajectory, and serves as a weighting matrix for spatial pooling.

The well-known IQA scheme SSIM has also been extended for VQA. In [77], SSIM was employed for VQA with the use of a weighting scheme that took into account

motion information using a block motion estimation algorithm. The method described in [78] also used SSIM with an alternate weighting scheme based on human perception of motion information. The scheme presented in [75] uses SSIM scores between the motion compensated blocks as a measure of the temporal distortions and this scheme is referred to as Motion Compensated SSIM or MC-SSIM. The VSSIM algorithm has also been proposed [76] which extends SSIM for VQA by estimating an optical flow field and calculating similarities along the trajectories. The authors in [82] extended the VIF [44] (an IQA scheme) criterion for VQA by using temporal derivatives. Another method known as MOVIE [66] uses Gabor filter to decompose the reference and distorted videos into spatio-temporal bandpass channels. Motion information is computed from the reference video sequence in optical flow fields. The set of Gabor filters used to compute the spatial quality is also used to calculate optical flow from the reference video.

Although significant research effort has been spent in recent years for objective VQA, we are still far from a practically useful VQA scheme. The reason is that most VQA schemes have very high complexity making them unsuitable for practical applications (this is in fact the reason why PSNR continues to enjoy wide popularity despite its limitations [16]-[19]). So a low complexity VQA method which can accurately mimic HVS's perception would be invaluable.

2.3 Speech Quality Assessment

The rapid increase in usage of speech processing algorithms in multi-media and telecommunications applications raises the need for speech quality evaluation. Accurate and reliable assessment of speech quality is thus becoming vital for the satisfaction of the end-user or customer of the deployed speech processing systems (e.g., cell phone, speech

synthesis system, etc.). Like visual signal cases, assessment of speech quality can be done using subjective listening tests or using objective quality measures. Subjective evaluation involves comparisons of original and processed speech signals by a group of listeners who are asked to rate the quality of speech along a pre-determined scale. Objective evaluation involves a mathematical comparison of the original and processed speech signals. Objective measures quantify quality by measuring the numerical "distance" between the original and processed signals. Objective speech quality assessment can be done intrusively (i.e. FR) or nonintrusively (i.e. NR). Contrary to the area of visual quality assessment, research into FR speech quality assessment has resulted in methods with acceptable prediction performance. The ITU-T released P.862 PESQ [3] as the standard for intrusive speech quality assessment. Because PESQ cannot tackle noise-suppression case, it will be soon replaced by P.863 POLQA [4]. However, nonintrusive (i.e. NR) quality assessment of speech quality has been a more active research area. Although P.563 [5] has been standardized by the ITU-T as the standard for nonintrusive speech quality assessment, it does not perform well specifically for noise-suppression induced distortions. Therefore, in Section 2.3.2, we first analyze the specific issues pertaining to quality assessment of noise-suppressed speech and propose new solution to tackle this in Chapter 8. As a result, our proposed method can fill the gap in current nonintrusive speech quality assessment.

2.3.1 **Objective Quality Assessment**

Early intrusive methods (similar to the FR case of IQA) include SNR and segmental SNR [86]. More sophisticated measures (e.g., [87]) were proposed once low bitrate speech coders, which may not preserve the original signal waveform, were introduced.

More recently, quality measurement research has focused on algorithms that exploit models of human auditory perception. Representative algorithms include bark spectral density (BSD) [88], perceptual speech quality measure (PSQM) [89], measuring normalized blocks (MNB) [90]-[91], and statistical model-based quality measurement [92]-[93]. The ITU-T P.862 standard, also known as PESQ represents the current state-of-the-art double-ended algorithm [3]. Recent research, however, has suggested decreased PESQ performance for VoIP communications and algorithm sensitivity to connection parameters such as speech codec, packet size, packet loss rate, and packet loss pattern (e.g., [94]-[95]).

The main limitation of intrusive metrics is the requirement of the reference signal. In many practical situations like wireless communications, voice over IP and other in-service networks requiring speech quality monitoring, an intrusive approach is not applicable because the input speech signal is unavailable. In such cases a non-intrusive (similar to the NR case of IQA) measurement which depends only on the altered speech signal is desirable.

Non-intrusive evaluation, which is also termed as no-reference, single-ended or output based evaluation, is a more challenging problem since the measurement of speech quality has to be performed with only the output speech signal of the system under test, without using the original signal as a reference. As opposed to intrusive measurement, non-intrusive measurement is a more recent research field. An early attempt towards nonintrusive speech quality measurement, based on the spectrogram of the perceived signal, is presented in [97]. In [98] vector quantization codebooks were replaced by Gaussian mixture probability models to improve quality measurement performance. Other proposed schemes have made use of vocal tract models [100] for single-ended quality measurement. The method described in [99] uses Gaussian Mixture Models

(GMMs) to create an artificial reference model to compare the degraded speech; while in [102], speech quality is predicted by Bayesian inference, and MMSE estimation, based on a trained set of GMMs. A perceptually motivated speech quality assessment algorithm based on temporal envelope representation of speech is presented in [101]. In [103] a low complexity, non-intrusive speech quality assessment scheme has been proposed based on features computed from commonly used speech coding parameters (e.g. spectral dynamics).

2.3.2 Specific Issues in Quality Assessment of Noise Suppressed Speech

Literature survey shows that the problem of assessing quality of speech corrupted due to codecs and transmission networks has received more research attention [3], [5], [97]-[103] over the past years. Most of the existing speech quality metrics were developed for evaluating the distortions introduced by speech codecs and/or communication channels. As mentioned, the ITU-T has also released P.862 PESQ [3] as the current standard for intrusive (it will be soon replaced by P.863 POLQA [4]) and P.563 [5] as the current standard for nonintrusive quality assessment. By contrast, scant research has been done [104] to develop algorithms (intrusive or nonintrusive) that assess quality of noise-suppressed speech. Furthermore, only a few studies have examined the correlation between objective measures and the subjective quality of noise-suppressed speech. The reader is referred to [104] which reports the performance of several intrusive methods for assessing quality of noise-suppressed speech.

A metric for assessing quality of noise-suppressed speech is important since distortion of speech signals with additive noise is one of the most common factors that affect

speech quality and hence, noise-suppression is employed frequently. With advances in speech communication technology, noise suppression has become essential for applications such as human-machine interfaces like automatic speech recognition, hearing aids, video conferencing, and voice-controlled systems. Often mobile voice communication devices employ noise-suppression since they are used in an environment with a high level of ambient noise. Some other examples where severe speech quality degradation occurs due to noise include air-ground communication systems in which aircraft cockpit noise corrupts the pilot's speech, military voice communication systems which operates in noisy environment created by fighter aircrafts, machine guns, tanks, etc., in-car communication systems which suffers from car noise, teleconferencing systems where multi-talker babble noise comes into play. In all these scenarios, noise-suppression is employed and therefore, a metric to assess the impact of the noise-suppression scheme(s) with regards to quality is necessary. Furthermore, such a metric will also be useful in benchmarking the performance of speech enhancement approaches which are widely used [104]. In addition, many current speech coders use noise suppression algorithms and thus, it is crucial to assess the impact of noise-suppression on the perceived quality. In general, noise suppression has applications [105] in virtually all fields of communications (channel equalization, radar signal processing, etc.) and other fields (pattern analysis, data forecasting, and so on). Due to its practical significance, quality assessment of noise-suppressed is an important research problem but less investigated.

Quality assessment of noise-suppressed speech is a bi-dimensional problem (includes both signal and noise distortion components) as outlined by the ITU-T Recommendation P.835 [106]. This is because the goal of noise suppression is to reduce the noise or the background component without adversely affecting the speech or signal component of

the waveform. This is difficult to be realized in practice especially for higher levels of noise suppression. In such cases, there is an increasing degradation in the quality of the speech or signal component as more of the noise or background component is suppressed. This can lead to a situation where although the noise or background component has been reduced the speech signal component has been degraded. In such scenarios subjects can often become confused [106] as to what they should be responding to in their ratings of the overall quality of the waveform: while the background may have been improved because there is less noise present in the waveform, the speech signal may have been degraded in the process. This is the reason why P.835 was standardized and is based on triple notations consisting of signal quality rating, background quality rating and the overall quality rating. It aims to reduce the listener's uncertainty by requiring him to successively attend to and rate the waveform on: the speech signal, the background noise, and the overall effect: speech + background. The ETSI has also released ETSI EG 202 396-2[1] [107] which describes a recording and reproduction setup for realistic simulation of background noise scenarios in lab-type environments for the performance evaluation of terminals and communication systems.

As mentioned, most existing metrics were developed for the purpose of evaluating the distortions introduced by speech codecs and/or communication channels. Therefore, their use in assessing perceptual quality of noise-suppressed speech is not obvious. The ITU-T Recommendation P.862.3 [96] states that “the use of the current intrusive algorithm P.862 (PESQ) with systems that include noise suppression algorithms is not recommended.” The reason for this is that most existing intrusive algorithms (including PESQ) assume that the input is undistorted and that the processed signal (i.e. signal

[1] [Online] Available at: http://www.etsi.org

under test) is of quality no better than the input. Clearly these assumptions are violated in case of noise-suppression or speech enhancement. It is due to this that the ITU-T is in the process of introducing a new standard namely ITU-T Recommendation P.863 or POLQA [4] which will replace PESQ. The scope of POLQA which is an intrusive metric includes quality assessment of noise suppressed signals. Another intrusive method described in ETSI EG 202 396-31 [108] objectively determines the signal or speech quality, the background noise quality and combines the two via regression to obtain the overall quality.

It may be pointed that the metrics discussed here determine speech quality intrusively, i.e., requiring a reference signal which may not be always available. Usually when noise suppression algorithms are used, only the noise corrupted signal is available, along with its enhanced counterpart. In such cases, since the original clean signal is unavailable, intrusive methods cannot be used. By contrast, nonintrusive algorithms will naturally avoid these problems since they do not require a reference signal. However, it has been found that the current nonintrusive ‘state-of-art’ algorithm ITU-T P.563 [5] yields low correlation with subjective quality of noise-suppressed speech [9]. As a result of these limitations, there is need for research into methods for quality assessment of noise-suppressed speech.

Chapter 3

Visual Quality Assessment Using Singular Value Decomposition

3.1 Introduction

The SVD is a useful and widely used tool of linear algebra for matrix factorization. It has been used in many applications such as matrix approximation, noise reduction (by using truncated SVD), data compression, analysis of numerical algorithms, computing pseudo inverse, to list a few (the reader is referred to [110] for a tutorial on SVD). The SVD can be employed for image processing tasks by assuming images as 2D matrices (i.e. 2D collection of pixel values). In this chapter, a new SVD based visual quality assessment algorithm is presented. The SVD of a matrix yields singular vectors (these define a set of basis images) and values (these correspond to the weights assigned to the basis images). The singular vectors and values can be respectively used to characterize structural and luminance changes in the visual signal (as elaborated in Section 3.2). The presented scheme uses changes in singular vectors and values as a more comprehensive measure of the structural and luminance changes respectively. Accordingly, it achieves significantly better performance than the existing SVD based method [37] which

employs only singular values. In addition, since structural changes have larger impact on the perceived visual quality, singular vectors alone can also be used for quality measurement with small prediction accuracy loss (to be demonstrated in this chapter).

The remainder of this chapter is organized as follows. Section 3.2 discusses the theoretical and experimental aspects of characterization of images by SVD. Section 3.3 describes the proposed method. In Section 3.4, we describe the image and video databases used in this thesis. The performance of the proposed method is evaluated in Section 3.5 while Section 3.6 concludes the chapter.

3.2 Feature Extraction with SVD

Visual features must be extracted effectively for objective perceptual quality assessment. Various transforms like DFT, DCT, DWT, contourlet transforms etc. can be used. In general, any 2-D transform decomposes the image into several basis images weighted by transformation coefficients. Visual quality can be assessed by measuring the changes in transformation coefficients [34]-[37], [39]. For example, in [39], image quality was predicted by computing the difference between frequency-domain coefficients of the original and distorted images. For the frequency-domain transforms like DFT and DCT, the basis images (accounting for image structure) are same for all the images, so the changes in visual signal can be captured only by the transformation coefficients. On the contrary, the basis images for SVD are unique for each image, and are expected to be able to represent the structure of an individual image better. Hence any change caused in the image structure is reflected in the individualized basis images with SVD. Due to this, SVD is more advantageous for capturing structural components in the visual signal. As stated before, effective differentiation of structural changes is the

prerequisite for its deserved treatments in visual quality evaluation, in order to remedy the mistake in MSE/PSNR and other existing metrics.

The SVD [110] of an image matrix $\boldsymbol{X}$ (size $r \times c$) yields the left singular vector matrix $\boldsymbol{U}$, the right singular vector matrix $\boldsymbol{V}$ and the diagonal matrix of singular values $\boldsymbol{\sigma}$:

$$\boldsymbol{X} = \boldsymbol{U}\ \boldsymbol{\sigma}\ \boldsymbol{V}^T \tag{3.1}$$

such that

$$\boldsymbol{U} = [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \ldots \boldsymbol{u}_r]$$

$$\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots \boldsymbol{v}_c]$$

$$\boldsymbol{\sigma} = diag(\sigma_1, \sigma_2, \ldots \sigma_t)$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are column vectors while σ_k is a singular value (i = 1,2,…,r, j = 1,2,…c, k = 1,2,…t, and $t = min(r,c)$). The singular values appear in descending order, i.e. $\sigma_1 > \sigma_2 > \ldots \sigma_t$.

3.2.1 Analysis of singular vectors

Any row of $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{p}_i = \sum_k u_{ik} \sigma_k \boldsymbol{v}_k^T \tag{3.2}$$

Similarly, any column of $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{q}_j = \sum_k \boldsymbol{u}_k \sigma_k v_{jk} \tag{3.3}$$

Therefore, $\boldsymbol{p}_i$ is a linear combination of the right singular vectors $\boldsymbol{v}_k$ and $\boldsymbol{q}_j$ is a linear combination of the left singular vector $\boldsymbol{u}_k$.

The matrix $\boldsymbol{U}\boldsymbol{V}^T$ can be interpreted as the ensemble of the basis images while the singular values $\boldsymbol{\sigma}$ are the weights assigned to these basis images. The image structure can be therefore be represented as

$$\boldsymbol{X}_z = \sum_{i=1}^{z} \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{3.4}$$

where z ($z \leq t$) is the number of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ pairs used.

Each basis image (i.e. $\boldsymbol{u}_i \boldsymbol{v}_i^T$) specifies a layer of the image geometry and the sum of these layers denotes the complete image structure. The first a few singular vector pairs of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ account for the major image structure while the subsequent pairs account for the finer details in the image. We illustrate this point through an example shown in Figure 3.1 where the image size is 512 × 512 and thus, t = 512. We can see that the first 20 pairs of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ (i.e. z = 20 in Eq. (3.4)) capture the major image structure and the subsequent pairs of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ signify the finer details in image structure. As an increasing number of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ pairs are used, the finer image structural details appear. $\boldsymbol{U}$ and $\boldsymbol{V}$ can, therefore, be used to represent the structural elements in images.

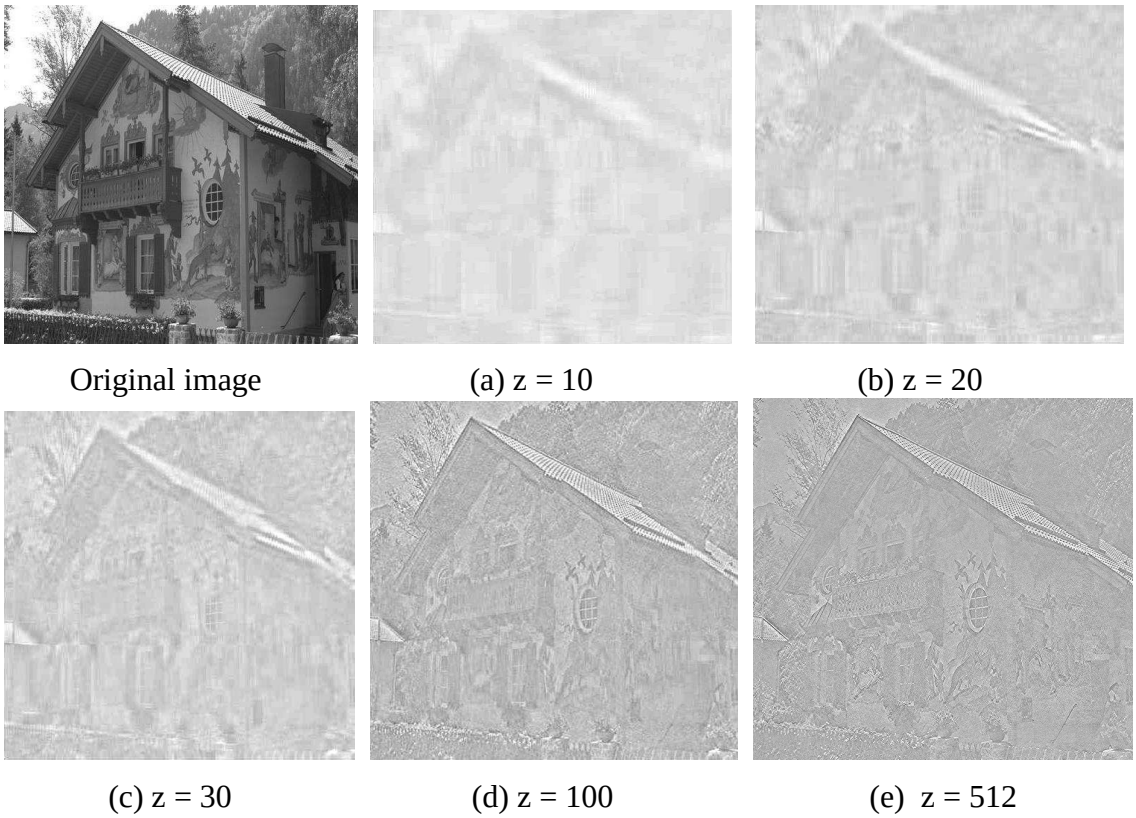
Because $\boldsymbol{V}$ is square, and also row-orthogonal, we can write the SVD of $\boldsymbol{X}$ as

$$\boldsymbol{X}_{i,j} = \sum_k u_{ik} \sigma_k v_{jk}^T = \sum_k c_{ik} v_{jk}^T \tag{3.5}$$

We can compare this with the DFT, which decomposes the original data into an orthogonal basis that can be expressed as follows

$$\boldsymbol{X}_{i,j} = \sum_k b_{ik} e^{\boldsymbol{i} 2\pi jk/r} \tag{3.6}$$

We can see from Eqs. (3.5) and (3.6) that SVD is similar to the DFT in the sense that the cyclical term $e^{\mathbf{i}2\pi jk/r}$ is replaced by the normalized vector term v_{jk}^T. Although the coefficient matrix $\boldsymbol{C} = \{ c_{ik} \}$ of SVD is orthogonal (since $\boldsymbol{U}$ is orthogonal), the coefficient matrix $\boldsymbol{G} = \{ b_{ik} \}$ of the DFT is not orthogonal in general. Nevertheless this demonstrates that the SVD is similar to the DFT, where the basis images are determined in a very specific way from image data rather than being given at the outset as for the DFT.



Original image (a) z = 10 (b) z = 20

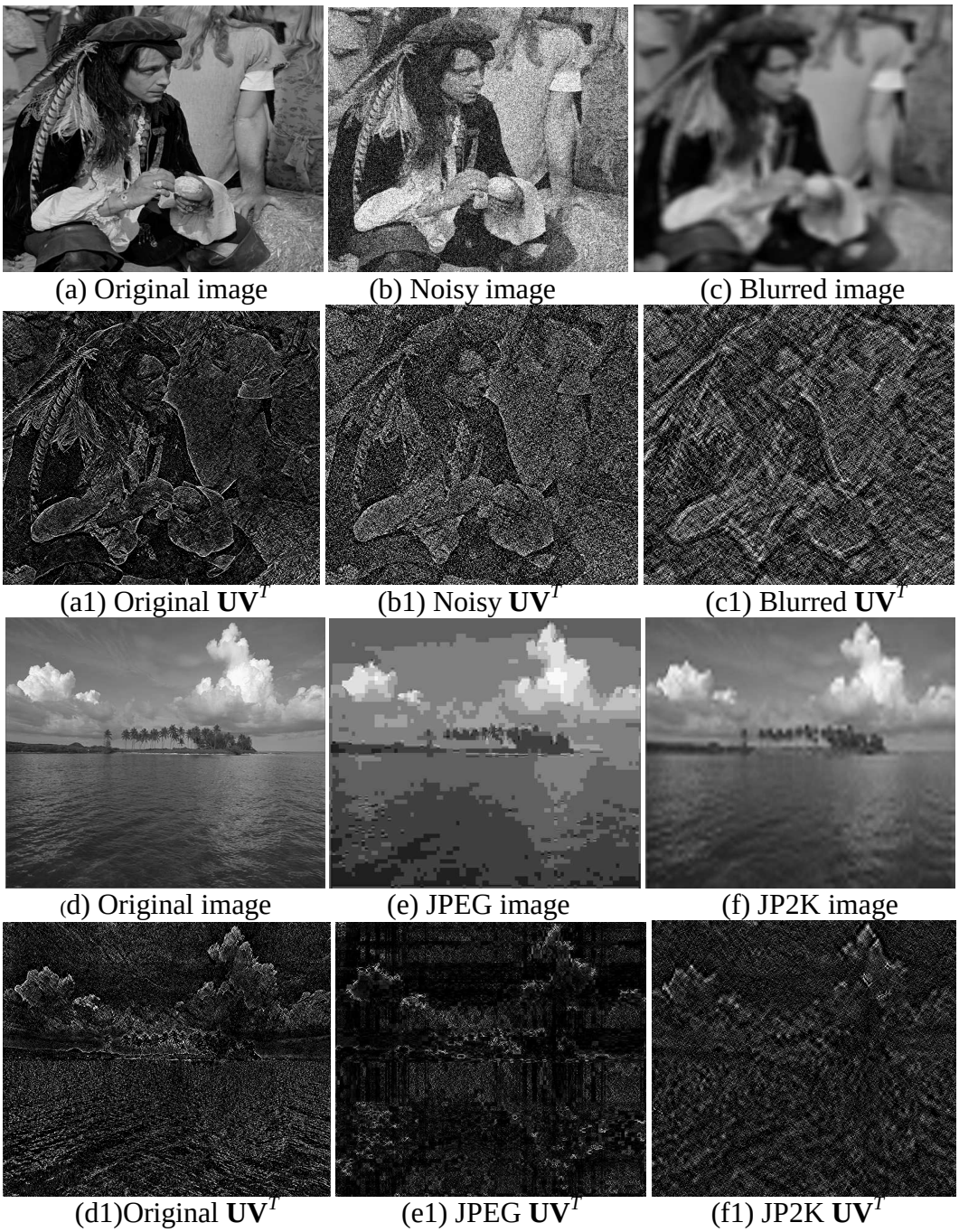
(c) z = 30 (d) z = 100 (e) z = 512

Figure 3.1: $\boldsymbol{X}_z$ as defined by Eq. (3.4) for different z values

In view of the analogy between SVD and DFT, the first a few singular vectors denote the low frequency components of the image while the subsequent vectors account for the higher frequency, as can be seen from Figure 3.1. We can see that using the first 10 or 20 vectors, mainly the low frequency components are visible. The high frequency components appear as the number of vectors is increased. The major advantage of using SVD in comparison with DFT is that the basis images adaptively defined in Eq. (3.4) leads to the possibility of representing the image structure better.

We would also like to point out that the image structure represented by Eq. (3.4) is different from the one in SSIM. The image structure defined by us in Eq. (3.4) is more intuitive and with a physical meaning in that it relates to edges and other salient parts of the image (for example with small value of z in Eq. (3.4) one can only the basic structure

and finer details appear with increasing *z*). On the other hand, the term *structure* in SSIM refers to the fact that pixels in natural images are correlated or there exists statistical similarity between nearby pixels. Accordingly in SSIM the *structure* has been defined as the correlation coefficient between the reference and distorted image patches (i.e. a higher value indicates lower damage to structure). In essence, SSIM works on the idea that any distortion will disturb this correlation or in other words damage the *structural* relationship between neighbouring pixels. Thus, the term structure in SSIM has more to do with statistical properties rather than the actual image structure (which comprises of say edges). As a result of its definition, the SSIM structure comes into picture only when one compares two image patches (so for a single image patch structure is not defined). On the other hand, the SVD based structure in Eq. (3.4) is defined for each image (or image patch) individually. From the perturbation analysis theory [111]-[112], $\boldsymbol{U}$ and $\boldsymbol{V}$ are found to be sensitive to perturbations. Therefore, any changes introduced in the image (due to distortion) affect the singular vectors significantly. The sensitivity of singular vectors can be exploited to assess the visual quality since the changes in visual quality are characterized by structural changes. For example, blur affects image structure by damaging edges and high frequency regions. The commonly-used JPEG image compression scheme damages structure by introducing blockiness; JPEG-2000, which is a more recent compression standard based on the wavelet transform, makes images blurry along the edges and in high frequency areas. As shown in Figure 3.2, different types of distortions (added noise, blurring, and JPEG/JPEG-2000 compression) affect the structure of visual signal represented by $\boldsymbol{U}$ and $\boldsymbol{V}$. Since the changes in adaptively determined **U** and **V** account for such structural changes, they provide an effective basis for assessing visual quality.



(a) Original image (b) Noisy image (c) Blurred image

(a1) Original $\mathbf{UV}^T$ (b1) Noisy $\mathbf{UV}^T$ (c1) Blurred $\mathbf{UV}^T$

(d) Original image (e) JPEG image (f) JP2K image

(d1)Original $\mathbf{UV}^T$ (e1) JPEG $\mathbf{UV}^T$ (f1) JP2K $\mathbf{UV}^T$

Figure 3.2: Structure denoted by the singular vectors i.e. $\mathbf{UV}^T$ in images

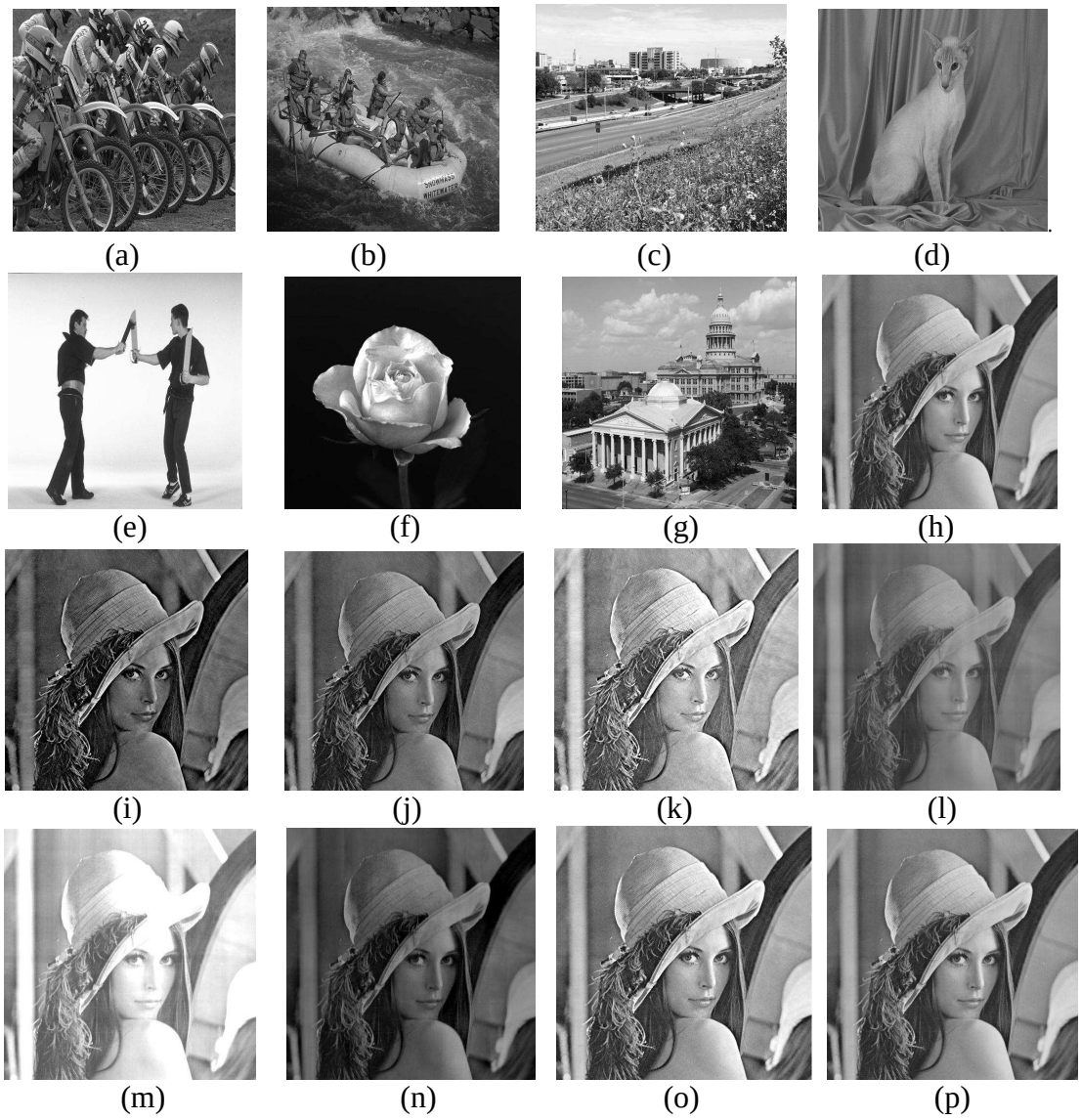


Figure 3.3: Effect of changing **σ** in images

Images (i) to (o) are constructed from **U**, **V** of original 'Lena' image (h) and the **σ** matrices of images from (a) to (g) respectively. Image (p) is constructed from **U**, **V** of original 'Lena' image (h) and the average of **σ** matrices of images from (a) to (g)

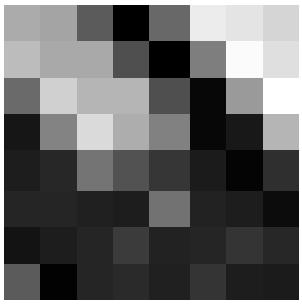
3.2.2 Further analysis with singular values

The $\boldsymbol{\sigma}$ values are mainly related to the luminance changes in images, as shown in Figure 3.3: (a) to (h) show eight test images; (i) to (o) show the 'Lena' image (h) constructed with its own $\boldsymbol{U}$ and $\boldsymbol{V}$, but using the $\boldsymbol{\sigma}$ matrices of the other images (a) to (g)

respectively; in (p), we also show the 'Lena' image constructed with its own $\boldsymbol{U}$ and $\boldsymbol{V}$ and the average $\boldsymbol{\sigma}$ of images (a) to (g).

We can observe the luminance changes in the reconstructed 'Lena' images (i) to (p). A closer examination of Figure 3.3 reveals that the images (e) and (f) are with much brighter and much darker luminance respectively, compared to other images. The corresponding luminance changes can be seen in (m) and (n) which are formed from the $\boldsymbol{\sigma}$ matrices of images (e) and (f) respectively.

In the MSVD metric [37], $\boldsymbol{\sigma}$ was used on the basis that it denotes the activity level in an image block. Activity level is defined as the luminance variation in pixels of an image block. A high activity level represents roughness or strong texture. Similarly, a low activity level corresponds to smoothness or weak texture. Due to its ability to characterize luminance changes, $\boldsymbol{\sigma}$ has also been used for image texture classification [113]. To illustrate this point further, we show two 8 × 8 blocks taken from '*bikes*' image (shown in Figure 3.3 (a)) of the LIVE image database [115], one with a larger pixel intensity variation (denoted by $\mathbf{B}_H$) and the other with a smaller variation in pixel intensities (denoted by $\mathbf{B}_L$).



$\mathbf{B}_H$

$\mathbf{B}_L$

The singular values of $\mathbf{B}_H$ and $\mathbf{B}_L$ are as follows:

diag ($\boldsymbol{\sigma}_H$) = [478.75, 129.22, 64.71,40.68, 26.4, 15.42, 4.84, 1.05],

diag ($\boldsymbol{\sigma}_L$) = [791.68, 10.42, 4.25, 2.17, 0.69, 0 ,0 , 0]

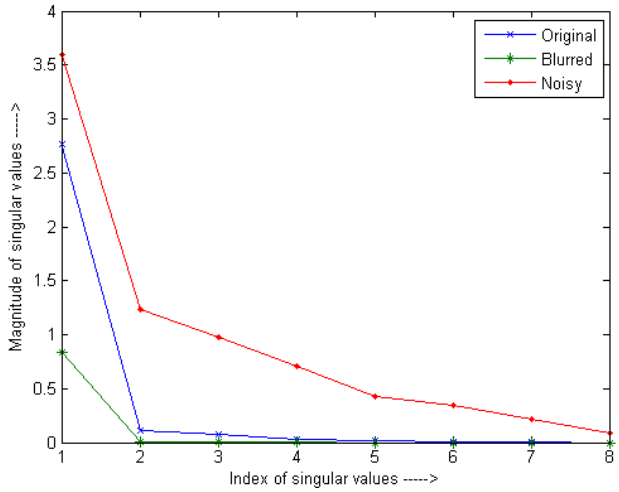


Figure 3.4: Behavior of singular values for noise and blur distortion

The ratio of the largest to the second largest singular value can be used to indicate the activity level [37]. In this example, this ratio is 3.70 for $\mathbf{B}_H$ (with high pixel variation) and it is 75.97 for $\mathbf{B}_L$ (with low pixel variation). The extreme case of $\mathbf{B}_L$ is a block in which all the pixel values are equal to, say q (i.e. no variation in pixel intensity); for such a block, the first singular value will be 8 × q and the rest will be all zero. In this case, the said ratio is infinite, indicating no variation in pixel luminance. Different types of distortions bring about different changes in image luminance (with the related textural changes) which are captured reasonably well by the changes in singular values.

As mentioned in Section 3.2.1, singular values are the weights for the basis images which can also be related to the changes in the frequency components of the image. Consider an 8 × 8 block of the "*rapids*' image (shown in Figure 3.3 (b)) of the LIVE image database. This block is then distorted by noise and blur. We show the singular values of the original, noisy and blurred blocks in Figure 3.4. One can notice that $\boldsymbol{\sigma}$ of the noisy block has higher values than that of the original block and they decay slower. We can interpret $\boldsymbol{\sigma}$ to denote the effect of change in frequency because the noise increases the frequency and this is captured in the increased $\boldsymbol{\sigma}$ values. On the other hand,

blur reduces the frequency and the reader will notice that $\boldsymbol{\sigma}$ of blurred block have lower values as compared to the original block and it decays very fast implying loss of frequency. In view of these, $\boldsymbol{\sigma}$ also reflects the changes induced in images due to the distortion and thus provide useful information to characterize the quality. This is the reason why $\boldsymbol{\sigma}$ can be used for quality evaluation [37].

3.2.3 The Proposed SVD-Based Metric

Based on the analysis and reasoning in the previous section, we observe the following:

- $\boldsymbol{U}$ and $\boldsymbol{V}$ denote the "*building blocks*" (or basis images) while $\boldsymbol{\sigma}$ determines how much (i.e. the weight) of each basis image is needed for image formation.
- In general any distortion will affect $\boldsymbol{U}$ and $\boldsymbol{V}$ and $\boldsymbol{\sigma}$. This has been illustrated by visual examples shown in Figures 3.2 and 3.4. $\boldsymbol{U}$ and $\boldsymbol{V}$ can capture structural changes better. However, for the special case of luminance change only (for example multiplying the image by a constant), $\boldsymbol{U}$ and $\boldsymbol{V}$ remain unchanged (this is similar to gradient of the image which is not affected due to simple luminance changes) but $\boldsymbol{\sigma}$ can reflect such luminance changes.
- The changes in $\boldsymbol{U}$ and $\boldsymbol{V}$ are more important because they can account for the major factor in quality degradation. On the other hand, changes in $\boldsymbol{\sigma}$ can be used to characterize luminance changes. As a result, changes in $\boldsymbol{U}$ and $\boldsymbol{V}$ and $\boldsymbol{\sigma}$ together provide a more comprehensive basis for visual quality assessment.

In summary of the analysis in the section above, the SVD transform has two major advantages over the other transforms for visual quality evaluation: (a) the adaptively derived singular vectors allow better representation of image structure, (b) the separation of structure and luminance components enables more effective differentiation of their

effects on perceptual quality (while in other transforms, all changes are reflected in the transform coefficients). We now describe the proposed scheme.

We decompose the original image (or video frame) $\boldsymbol{X}$ of the original video using (3.1) and the distorted image (or video frame) $\boldsymbol{X}^{(d)}$ as

$$\boldsymbol{X}^{(d)} = \boldsymbol{U}^{(d)} \boldsymbol{\sigma}^{(d)} \boldsymbol{V}^{(d)T}$$

where $\boldsymbol{U}^{(d)}$, $\boldsymbol{V}^{(d)}$ and $\boldsymbol{\sigma}^{(d)}$ denote the left, right singular vectors and singular value matrices respectively for $\boldsymbol{X}^{(d)}$ .We then measure the change in singular vectors as

$$\alpha_j = \boldsymbol{u}_j . \boldsymbol{u}_j^{(d)} \tag{3.7}$$

$$\beta_j = \boldsymbol{v}_j . \boldsymbol{v}_j^{(d)} \tag{3.8}$$

where α_j (j = 1 to t) represents the dot product between the unperturbed (i.e. reference) and the perturbed (i.e. distorted) j^{th} left singular vectors ($\boldsymbol{u}_j$ and $\boldsymbol{u}_j^{(d)}$) and β_j denotes that for the right singular vectors ($\boldsymbol{v}_j$ and $\boldsymbol{v}_j^{(d)}$).

To illustrate the meaning of Eq. (3.7) (and also for Eq. (3.8)), we take a further look at the dot product between two vectors $\boldsymbol{u}_j$ and $\boldsymbol{u}_j^{(d)}$ (angle between them is $\theta_{\boldsymbol{u}}$) which is defined as

$$\boldsymbol{u}_j . \boldsymbol{u}_j^{(d)} = |\boldsymbol{u}_j| \, |\boldsymbol{u}_j^{(d)}| \, cos(\theta_{\mathbf{u}}) \tag{3.9}$$

In the case of singular vectors, the magnitude of each vector is unity, i.e. $|\boldsymbol{u}_j| = |\boldsymbol{u}_j^{(d)}| = 1$. Thus the dot product between the unperturbed and the perturbed singular vectors (as given by Eqs. (3.7) and (3.8)) directly measures the cosine of the angle between the two singular vectors, and $-1 \leq \alpha_j, \beta_j \leq 1$.

We then define the feature vector $\boldsymbol{\Gamma}_j$ for representing the change in $\boldsymbol{U}$ and $\boldsymbol{V}$ as follows

$$\boldsymbol{\Gamma}_j = |\alpha_j + \beta_j| \tag{3.10}$$

Note that in the above Eq. we could also use $|\alpha_j| + |\beta_j|$ instead of $|\alpha_j + \beta_j|$. However, we found that the two yield largely similar results. The reason is that there are many α_j and β_j which are positive (i.e. $0 \leq \alpha_j, \beta_j \leq 1$) in which case $|\alpha_j| + |\beta_j| = |\alpha_j + \beta_j|$. We can see that Eq. (3.10) defines a *t*-dimensional vector $\boldsymbol{\Gamma}_j = \{\gamma_j\}$ (j = 1 to t). We then use Minkowski summation and logarithmic scale to obtain the quality score

$$Q_S = log\left(1 + \left(\left(\sum_{j=1}^{t} \gamma_j^p\right)\right)^{\frac{1}{p}}\right) \tag{3.11}$$

where p (> 1) is the pooling exponent. A larger p puts more emphasis on large γ_j values. We used p = 2 for the experiments. Q_S defined above can be used to quantify the structural modifications. For measuring the change in singular values, we use the existing MSVD metric in which the difference between the singular values of the reference and distorted image blocks is computed. We denote the overall quality based on change in singular values as Q_L

$$Q_L = \frac{\sum_{j=1}^{B} |D_j - D_{mid}|}{N_{total}} \tag{3.12}$$

where for each j^{th} block we calculate

$$D_j = \sqrt{\sum_{i=1}^{B} \{\sigma_i - \sigma_i^{(p)}\}^2} \tag{3.13}$$

where B defines the block size (in an image with size of $r \times c$), $N_{total} = r/B \times c/B$, and D_{mid} represents the midpoint of the sorted D_j's.

We now combine Q_S and Q_L to obtain the overall quality score. If a linear combination is used, the overall composite quality metric Q_C can be defined as

$$Q_C = Q_S - \mu\ Q_L \tag{3.14}$$

where μ is a user-defined parameter (we used $\mu = 5$) to cater for the different valuation between the two. The negative sign has been introduced in the definition of Q_C to accommodate the opposite trend of change in Q_S and Q_L (note that higher Q_S means better visual quality while higher Q_L implies lower quality). Note that when $\mu = 0$, we have $Q_C = Q_S$ i.e. the contribution from singular values is ignored and quality prediction depends only on the degradation of singular vectors. As μ value will be increased, obviously the contribution from Q_L will increase. We also note that the existing MSVD method is a special case of Eq. (3.14) when the value of μ is made significantly large (due to this there will be no contribution from Q_S to the overall quality). The composite metric Q_C accounts for the changes in singular vectors and values through a simple linear combination. The database description and the experimental results for Q_S, Q_L and Q_C are presented in the next section. We also compare the proposed scheme with three other schemes, namely SSIM [2], VSNR [45], IFC [114] and MSVD [37].

3.3 Experimental Results

3.3.1 Performance Evaluation

Following the Video Quality Experts Group (VQEG) validation methodology [81], a nonlinear mapping between the objective model outputs and the subjective quality ratings was also employed. This is to remove any nonlinearity due to the subjective rating process and to facilitate the comparison of the metrics in a common analysis space. For the experimental results reported in this chapter, we fitted the objective scores to

subjective scores via a four-parameter cubic polynomial $a_1x^3+a_2x^2+a_3x+a_4$ where a_1, a_2, a_3 and a_4 are determined by using the subjective scores and the objective outputs. We used three databases namely LIVE, TID and Toyama for evaluating and comparing the prediction accuracy. The reader is referred to the Appendix for a description of the databases. The Pearson correlation coefficient (C_P) , Spearman correlation coefficient (C_S) and Root Mean Square Error (RMSE) are shown in Figure 3.5 for SSIM [2], IFC [114], Q_L (i.e. MSVD metric [37]), VSNR [45], Q_S and Q_C. Further Table 3.3 reports the C_P for individual distortion types. For the codes of SSIM, IFC and VSNR, we have used the publicly accessible Matlab package that implements a variety of visual quality assessment algorithms [170]; they are the original codes provided by the IQA algorithm designers. The MSVD method was implemented by us.

We can see that Q_C generally performs better than the other metrics. We also observe that the prediction performance of Q_S and Q_C is close (Q_C being slightly better). The reason is that structural changes account for the major factor affecting visual quality.

To assess the statistical significance of each metric's performance relative to the other metrics, an F-test was performed on the prediction residuals between the objective predictions (after non-linear mapping) and the subjective scores. Obviously smaller the residuals, the better the metric is. The test is based on an assumption of Gaussianity of the residual differences. Suppose $Q_{proposed}$ denotes the proposed metric and X denotes the other metrics to be compared.

Table 3.1: Implications of different ranges of F values.

$F > F_{critical}$	$1 < F < F_{critical}$	$1/F_{critical} < F < 1$	$F < 1/F_{critical}$
X has significantly larger residuals than $Q_{proposed}$, so $Q_{proposed}$ is statistically better than X.	Although $Q_{proposed}$ performs better than X since F>1, both $Q_{proposed}$ and X are statistically indistinguishable.	Although X performs better than $Q_{proposed}$ since F<1, both $Q_{proposed}$ and X are statistically indistinguishable.	X has significantly smaller residuals than $Q_{proposed}$, so $Q_{proposed}$ is statistically worse than X.

Table 3.2: F-statistics of different metrics with respect to the proposed method

Database/ Metric	LIVE	Toyama	TID
SSIM	**1.28**	**1.75**	1.07
IFC	0.91	**1.42**	**1.69**
MSVD (Q_L)	**1.55**	**1.61**	**1.50**
VSNR	1.09	1.17	**1.48**
Q_S	1.12	1.14	1.11
Q_C	1	1	1
$F_{critical}$	1.18	1.41	1.12
1/ $F_{critical}$	0.84	0.70	0.89

Table 3.3: Performance comparison on individual distortion types

Metric	JP2K	JPEG	White noise	Blurring	Fastfading
SSIM	0.956	0.943	0.970	0.945	0.948
IFC	0.957	0.932	0.976	0.969	0.963
VSNR	0.953	0.943	0.978	0.934	0.902
MSVD (Q_L)	0.941	0.926	0.979	0.919	0.891
Q_C	0.952	0.951	0.984	0.962	0.949

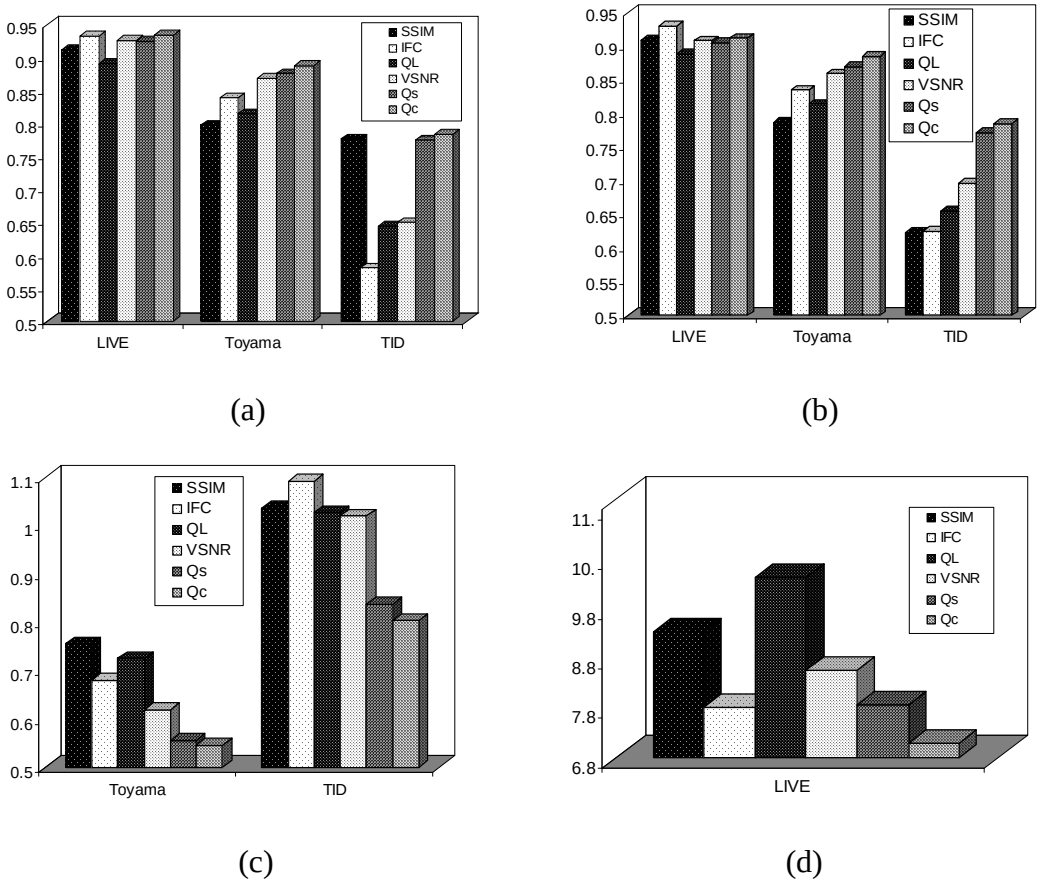


Figure 3.5: Performance comparison on 3 image databases

(a) Comparison of Pearson correlation coefficients, (b) Comparison of Spearman correlation coefficients (c) Comparison of Root Mean Square Error for Toyama and TID databases (d) Comparison of Root Mean Square Error for LIVE database

Let σ_X^2 and $\sigma_{Q_{proposed}}^2$ denote the variances of the residuals from metrics X and $Q_{proposed}$ respectively, and then the F-statistic with respect to metric $Q_{proposed}$ is given by $F = \sigma_X^2 / \sigma_{Q_{proposed}}^2$. The F value is then compared with the critical F-statistic (denoted as $F_{critical}$) which is computed based on the number of residuals and the desired confidence level, to judge if $Q_{proposed}$ and X are statistically indistinguishable. Table 3.1 lists the

implications of different ranges of F values. For the experimental results, we have used a 99% confidence level for the calculation of $F_{critical}$ values. The values of $F > F_{critical}$ indicated by boldfaced letters in Table 3.2 denote that the proposed Q_C has significantly smaller residuals than the corresponding metric and so Q_C performs statistically better than that metric. Since the MSVD metric (i.e. Q_L) uses only **σ**, it is worth pointing out that Q_C is statistically better than MSVD with all databases; this demonstrates the effectiveness of incorporating **U** and **V.** We can also see that there is big margin for F to be compared with $1/F_{critical}$ (even for the three cases in Table 3.2, in which $F < F_{critical}$). The experimental results and the related statistical analysis therefore confirm that the use of **U** and **V** along with **σ** improves the quality prediction performance significantly.

3.4 Concluding Remarks

In this chapter, we have investigated SVD based features for visual quality assessment. The major advantage of SVD over other transforms is that the adaptively determined singular vectors lead to a better structural representation of the signal while singular values can be used to measure luminance changes. The proposed method has been tested using three databases and has been found to be better or very competitive with the existing popular and relevant methods like SSIM, VSNR, IFC and MSVD.

Although the linear combination as the fusion rule in this chapter is computationally simpler, it may not be optimal. Further, both Q_S and Q_L themselves are computed using simple Minkowski summation (this has its own drawbacks as discussed in Section 2.1.2.2). The pooling exponent p in Eqs. (3.11) and (3.13) has been set to 2 but this again may be sub-optimal. The parameter μ in Eq. (3.14) is also determined empirically. We believe that these issues can be tackled better by employing machine learning based

feature pooling. Therefore more sophisticated method for feature pooling using machine learning will be explored in the next chapter (Chapter 4).

Chapter 4

Machine Learning Based Visual Feature Pooling

4.1 Introduction

As pointed out in Chapter 2, there are two important issues in objective quality assessment: (1) feature extraction for representing the visual signal appropriately, (2) pooling of the features for the result to be consistent with the HVS's perception of visual quality. In the previous chapter (Chapter 3), we have discussed SVD based features for assessing image and video quality. However, the pooling was still done using existing simplistic methods which may impose undesirable constraints on the relationship between feature changes and the visual quality.

As mentioned in Section 2.1.2.2, the existing pooling techniques implicitly make assumptions on the relative importance of distortion statistics, and there is lack of convincing ground for these assumptions. Even the more recent methods such as those based on VA have their drawbacks as detailed in Section 2.1.2.2. Overall, feature pooling is done largely using ad-hoc methods and therefore calls for further investigation and analysis. Appropriate feature pooling is an essential step for perceptual quality

assessment but there is lack of physiological and psychological knowledge for the convincing modeling (the psychophysical studies that have been conducted in the related field are for a single or at most two visual stimuli (e.g., in frequency, orientation, etc.), while real-world images are with many stimuli simultaneously).

Therefore, we propose to use machine learning to tackle the complex issue of feature pooling. It is an attractive alternative for feature pooling because such an approach is general, more systematic and reasonable, and the related model parameters (weights) are estimated via training from the sufficient, available data (i.e., the substantial ground truth). Given the strong theoretical foundations and proven success of machine learning techniques in numerous applications (such as face detection [157], handwriting/signature verification [158], video surveillance [159], robot tutoring [160], speech quality assessment [7] and so on), we believe that it can be exploited for perceptual quality assessment.

In contrast to the existing pooling methods, a machine learning technique in visual quality evaluation helps in avoiding assumptions on the relative significance and relationship of different distortion statistics (i.e. feature changes). There has been some early work in applying machine learning techniques for visual quality evaluation. In [162]-[163], objective VQA using Neural Networks (NNs) has been reported while the use of NNs has been demonstrated in [164]-[165] for image cases. Overall, machine learning in visual quality evaluation remains as a largely uninvestigated area.

The rest of this chapter is organized as follows. In Section 4.2, we first describe SVD based feature preparation for quality assessment and introduce the notations. A brief overview of SVR which we have employed for feature pooling is provided in Section 4.3. Extensive experimental results and related analysis are then reported in Section 4.4. We

give the concluding remarks in Section 4.5.

4.2 SVD Based Feature Preparation

Features can be detected globally with large blocks or locally with small blocks. We found that global SVD gives better prediction performance than local SVD. One reason for this is that when small blocks are employed in SVD based feature detection they are assumed to be completely independent which may not always be true. A global SVD, on the other hand, can tackle the interaction/dependencies between the blocks better. Furthermore, local SVD is also disadvantageous when used with machine learning: it will mean much larger number of features. For instance, for block size of 16 × 16 (image size 512 × 512) one would need 32768 dimensional vector (16384 features each for singular vectors and values). A large feature vector may contain redundant information which leads to performance degradation. Therefore, a global SVD is more effective for our purpose. However with the global SVD approach the feature vector dimension will depend on the image size and this will result in feature vectors of unequal dimensions for the images in different databases. This will lead to mismatch in the dimension of the training and test feature vectors in case of cross database evaluation (i.e. training with images from one database and test set comes from the other databases as detailed later in Section 4.5).

There are two ways to tackle the aforementioned problems. The first way to make the feature vector dimension equal for all images is to resize them to a common size. This is a straightforward solution but such an approach may introduce or remove some distortions and so the original subjective scores may not be valid. To tackle the drawback associated with image resizing, we use an approach in between. We divide an image into

blocks of size $B \times B$ and compute the SVD for each block. Then we use the average of the feature values of these blocks to define the final feature vector which will be $2B$ dimensional (B features for singular vectors and B features for the singular values). The only requirement is that image size should be greater than or equal to $B \times B$. For images with smaller size, we must use smaller block size and proceed in a similar way. We now outline the feature detection procedure.

First, the original and distorted images are divided into non-overlapping blocks of size $B \times B$. Let us denote the k^{th} (the total number of blocks is denoted as N_{block}) block in the original image as $\boldsymbol{A}_k$ and that in the distorted image as $\boldsymbol{A}_k^{(d)}$. We then obtain the respective singular values and singular vectors by applying SVD. The change in singular vectors is measured as:

$$\alpha_{jk} = \boldsymbol{u}_{jk} \cdot \boldsymbol{u}_{jk}^{(d)} \tag{4.1}$$

$$\beta_{jk} = \boldsymbol{v}_{jk} \cdot \boldsymbol{v}_{jk}^{(d)} \tag{4.2}$$

where α_{jk} (j = 1 to B and k = 1 to N_{block}) represents the dot product between the unperturbed and the perturbed j^{th} left singular vectors ($\boldsymbol{u}_j$ and $\boldsymbol{u}_j^{(d)}$) and β_{jk} denotes that for the right singular vectors ($\boldsymbol{v}_j$ and $\boldsymbol{v}_j^{(d)}$) of the k^{th} block. The reader can notice that Eqs. (4.1) and (4.2) are the same as Eqs. (3.7) and (3.8) respectively but the former use the additional subscript k to indicate the k^{th} block.

Note that $-1 \leq \alpha_{jk}, \beta_{jk} \leq 1$. We then define the feature vector $\boldsymbol{\Gamma}_k$ for the k^{th} block for representing the change in $\boldsymbol{U}$ and $\boldsymbol{V}$ as follows, after the absolute-valuation (for the reason explained at the end of this section) and normalization (for the values to range between 0 and 1):

$$\Gamma_k = \frac{|\alpha_{jk}|+|\beta_{jk}|}{2} \qquad (j = 1 \text{ to } B) \tag{4.3}$$

To measure the change in singular values (let σ_k and $\sigma_k^{(d)}$ denote the original and distorted singular value matrices), we let $s = diag(\sigma)$ and $s_k^{(d)} = diag(\sigma_k^{(d)})$. We then define the feature vector for representing the change in singular values as

$$\tau_k = \left(s_k - s_k^{(d)}\right)^2 \tag{4.4}$$

The length of Γ_k and τ_k will be B. From Eq. (4.4), it is easy to see that all the elements of τ_k are greater than or equal to 0. It is found that for natural images the dynamic range of τ_k is very large. Therefore, we divide each element in τ_k by the maximum value in τ_k for normalization to the range [0, 1], and define the resultant vector λ_k as

$$\lambda_k = \tau_k / max(\tau_k) \tag{4.5}$$

The feature vector for the k^{th} block is then defined as

$$x_k = \left\{ \Gamma_k, \lambda_k \right\} \tag{4.6}$$

It follows that vector x_k will be of length $2B$. The final feature vector for the image is then obtained by averaging out the features over all the blocks

$$x = \frac{1}{N_{block}} \sum_{k=1}^{N_{block}} x_k \tag{4.7}$$

We found that the prediction errors were reduced significantly when we used the absolute values of α_{jk} and β_{jk} in Eq. (4.3) (instead of using the actual α_{jk} and β_{jk}), with the explanation as follows. By definition, $-1 \le \alpha_{jk}, \beta_{jk} \le 1$ and so $(\alpha_{jk} + \beta_{jk})$ can be positive or negative. Thus, two coefficients next to each other can be of similar magnitude but opposite sign to cause a large swing in the input data. This may affect the generalization performance of a machine learning algorithm. Therefore, we have used the absolute

values as the feature input for the machine learning stage. A similar conclusion can be found in [233] which discusses the application of SVR for image coding when the absolute magnitudes of DCT coefficients were used as the input to the SVR. The reader may also note that Eq. (4.3) is also slightly different from Eq. (3.10) but the two yield largely similar results as explained in Section 3.2.3 of the previous chapter.

In our case, we used a block size of 128 (i.e. $B = 128$) and thus the feature vector for an image will be 256 dimensional. We also experimented with smaller block sizes 64 × 64, 32 × 32, etc., but the prediction performance especially for cross database evaluation is better at a bigger block size. There are two reasons for this observation:

(a) As already mentioned, smaller blocks may not take into account the dependencies or interactions among them because features are extracted for each block independent of other blocks.

(b) With smaller block size say 8 × 8 there will be 16 features for each block and there will be a total of 4096 blocks. It is quite possible that in such a case the useful information about change in quality may be suppressed due to averaging over a large number of blocks. In fact we use a machine learning technique in the first place to avoid such direct averaging/pooling methods. Nevertheless, with a larger block size such as 128 × 128, the average of features is computed over fewer blocks and therefore more reasonable for the purpose.

The reader will note that the chosen block size of 128 × 128 can handle almost all the existing image and video resolutions. For example, the typical resolution for DVD, miniDV and Digital8 is 720 × 480 while newer technologies use higher resolutions (for instance, Blue ray uses 1280 × 720, 2K digital cinema uses 2048 × 1080, and so on); the other commonly used video resolution are CIF (352 × 288), QCIF (176 × 144), 4 CIF

(704 × 576), QVGA (320 × 240), VGA (640 × 480), XVGA (1024 × 768), DVD NTSC (720 × 480), DVD Pal (720 × 576), HDTV 720p (1280 × 720), etc. Note that for image sizes which are not multiples of the chosen block size, one can use overlapping blocks (or zero padding) to compute the averaged feature vector as outlined. We found that overlapped blocks (or zero padding) do not have any significant effect on the prediction accuracy.

4.3 Feature Pooling using SVR

Our aim is to represent the quality score Q as a function of the proposed feature vector $\boldsymbol{x}$:

$$Q = f(\boldsymbol{x}) \tag{4.8}$$

where f is a function relating the elements of $\boldsymbol{x}$ to the final quality score and is difficult to be determined *apriori*. To estimate f we use a machine learning approach. In this work, we use SVR to map the high dimensional feature vector into a perceptual quality score, by estimating the underlying complex relationship among the changes in $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{\sigma}$ and the perceptual quality score. Although other choices of machine learning techniques are possible, we have used SVR because it is popular and well established. Furthermore, with SVR one can obtain the SVs which are critical datapoints for the SVR learning; their analysis can provide additional insights about the learning problem in hand, as will be shown later in the chapter.

The goal of SVR is to find f, based on training samples. Suppose that $\boldsymbol{x}_i$ is the feature vector of the i^{th} image in the training image set (i = 1, 2… l; l is the number of training images). In the ε-SV regression [166]-[167] the goal is to find a function $f(\boldsymbol{x}_i)$ that has the deviation of at most ε from the targets s_i (being the corresponding subjective quality

score) for all the training data, and at the same time is as flat as possible [167]. The function to be learned is $f(\boldsymbol{x}) = \boldsymbol{W}^T\varphi(\boldsymbol{x}) + b$; where $\varphi(\boldsymbol{x})$ is a non-linear function of $\boldsymbol{x}$, $\boldsymbol{W}$ is the weight vector and b is the bias term. We find the unknowns $\boldsymbol{W}$ and b from the training data such that the error

$$|s_i - f(\boldsymbol{x}_i)| \leq \varepsilon \tag{4.9}$$

for the i^{th} training sample $\{\boldsymbol{x}_i, s_i\}$. In SVR, a kernel function $\phi(\boldsymbol{x})$ is employed to map the data into a higher dimensional space. We solve the following optimization problem

$$\min_{W,b,\xi,\xi^*} \frac{1}{2}\boldsymbol{W}^T\boldsymbol{W} + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \tag{4.10}$$

$$\text{subject to} \quad \begin{cases} s_i - \left(\mathrm{W}^T\phi(\mathrm{x}_i) + b\right) \leq \varepsilon + \xi_i \\ \left(\mathrm{W}^T\phi(\mathrm{x}_i) + b\right) - s_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

where ξ_i is the upper training error (ξ_i^* is the lower training error) ε being a threshold; $\frac{1}{2}\boldsymbol{W}^T\boldsymbol{W}$ is the regularization term to smooth the function $\boldsymbol{W}^T\varphi(\boldsymbol{x}) + b$ in order to avoid overfitting; $C > 0$, being the penalty parameter of the error term. Eq. (4.10) can be solved using the dual formulation to obtain the solution $(\boldsymbol{W}, b)$.

It has been shown in [166] that

$$\boldsymbol{W} = \sum_{i=1}^{n_{sv}}(\eta_i^* - \eta_i)\ \varphi(\boldsymbol{x}_i) \tag{4.11}$$

where η_i^* and η_i ($0 \leq \eta_i^*$, $\eta_i \leq C$) are the Lagrange multipliers used in the Lagrange function optimization, C is the trade off error parameter and n_{sv} is the number of SVs. For data points for which inequality (4.9) is satisfied, i.e. the points which lie within the

ε tube, the corresponding η_i^* and η_i will be zero so that the Karush Kuhn Tucker (KKT) conditions are satisfied [166]. The samples that come with nonvanishing coefficients (i.e. non zero η_i^* and η_i) are SVs, and the weight vector $\boldsymbol{W}$ is defined only by the SVs (not all training data). The function to be learned then becomes

$$f(\boldsymbol{x}) = \boldsymbol{W}^T \varphi(\boldsymbol{x}) + b = \sum_{i=1}^{n_{sv}} (\eta_i^* - \eta_i)\ \varphi(\boldsymbol{x}_i)^T \varphi(\boldsymbol{x}) + b$$

$$= \sum_{i=1}^{n_{sv}} (\eta_i^* - \eta_i)\, \mathrm{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b \qquad (4.12)$$

where $\mathrm{K}(\boldsymbol{x}_i, \boldsymbol{x}) = \varphi(\boldsymbol{x}_i)^T \varphi(\boldsymbol{x})$ is the kernel function. In SVR, the actual learning is based only on the critical points (i.e., the SVs). In this chapter, we have used the Radial Basis Function (RBF) as the kernel which is of the form $\mathrm{K}(\boldsymbol{x}_i, \boldsymbol{x}) = \exp(-\rho \|\boldsymbol{x}_i - \boldsymbol{x}\|^2)$ where ρ is a positive parameter controlling the radius.

4.4 Performance Evaluation

For the image case we used LIVE, TID, Toyama, IVC, watermarked image database, WIQ, CSIQ and A57 while for video case we used EPFL and LIVE video databases (Refer to the Appendix for details). Most of the existing visual quality metrics work only with the luminance component of the image/video. Therefore, all experimental results reported in this dissertation are for the luminance component only (because the luminance component plays a more significant role in human visual perception than color components).We now outline the training and test procedure.

4.4.1 Test procedure

We evaluate the performance of the proposed scheme in two different ways. Firstly, we

have employed the *k*-fold CV strategy [168] for each database separately: the data was split into *k* chunks, one chunk was used for test, and the remaining (*k*-1) chunks were used for training. The experiment was repeated with each of the *k* chunks used for testing. The average accuracy of the tests over the *k* chunks was taken as the performance measure. The splitting of the data into *k* chunks was done carefully so that the image contents present in 1 chunk did not appear in any of the remaining chunks (and this chunk is used as the test set). One image content is defined as all the distorted versions of an original image. As an example, consider the CSIQ database which consists of 30 original images. In this case, the first chunk included all the distorted versions of the first 3 original images. The second chunk consisted of distorted versions of the next 3 original images and so on. Thus, for the CSIQ database there were a total of 10 chunks each of which comprised different image contents. In the same way, the Toyama database (with 14 original images) was split into 7 chunks with each chunk comprising of 2 image contents. The LIVE database with 29 original images was split into 10 chunks with the first 9 chunks consisting of 3 image contents each while the last chunk included 2 image contents. Similar splitting procedure was followed for the other databases as well. In this way, it was ensured that images appearing in the test set are not present in the training set. In this chapter, the symbol Q_{full} has been used for the proposed method to indicate the results for *k*-fold CV.

Since the proposed metric involves training, we need to further examine the feasibility and robustness of such machine learning based system to untrained image and distortion types. To that end, we use the cross database validation: one database is used for training and others are used for validation. In this thesis, we use the notation $Q_{database}$ to denote training with a particular database. So Q_{CSIQ}, Q_{LIVE} and Q_{TID} denote that training is done

with the CSIQ, LIVE and TID databases respectively, and similar notation has been followed for other databases as well. However, some databases have a few images in common, e.g., LIVE, TID and Toyama. Therefore, we have reported the cross database evaluation results for the cases when none of the images in the training set has appeared in the test set. This is again to ensure that the system is trained and tested on entirely different sets of images. We have also used the symbol Q_{vector} to denote the metric that uses only singular vectors as the features (introduced in [179]) and we will compare its performance with Q_{full} which uses both singular vectors and values as a more comprehensive method. Note that for Q_{vector} we have reported only the best results among those obtained on training with different databases.

A 5-parameter logistic mapping between the objective outputs (x) and the subjective scores was employed before performance comparison. The used logistic function has the following form

$$x_m = \beta_1 \left(0.5 - \frac{1}{1+exp(\beta_2 (x-\beta_3))} \right) + \beta_4 x + \beta_5$$

The parameters $\beta_{1\sim5}$ are determined by minimizing the sum of squared differences between the mapped scores x_m and the subjective scores. The experimental results are reported in terms of C_P, C_S and RMSE, (similar to Section 3.3.2) between the subjective score and the objective prediction (after logistic transformation). A better quality metric has higher C_P, C_S and lower RMSE.

We have also compared the performance of the proposed Q_{full} (with k-fold CV) with the following existing visual quality estimators: PSNR, SSIM [2], MSVD [37], VSNR [45], IFC [114], VIF [44], Q_{vector} and the method proposed in the previous chapter (we refer to it as Ref. [73] in the figures for the experimental results). The publicly available LibSVM

software package [171] was used to implement the SVR algorithm. As we already mentioned in Section 3.3.1 we have used the publicly accessible Matlab package [170] to obtain the codes for SSIM, VSNR, IFC and VIF while the MSVD method was implemented by us. Similar to the previous chapter, we also carried out statistical analysis to examine the statistical reliability of results obtained.

4.4.2 Visual quality prediction test

To demonstrate that the proposed method properly accounts for the distortion in different image areas, we show 4 images in Figure 4.1. First we consider the "hat" part and the "shoulder" part of the "Lena" image as indicated by the boxes in Figure 4.1 (a) and (b). Note that the amount of noise in the two blocks in Figure 4.1 (a) and (b) is the same. Because the effect of white noise is uniformly distributed, it can be observed that it does not cause too much damage to the edge in the "hat" and the *structure* is largely preserved.

As a result, noise in the "hat" part is less annoying. Of course, there will be loss of visual quality. On the other hand, the reader will notice that the shoulder in "Lena" image is smooth due to which the added noise is clearly visible and therefore more annoying to the human eye. This leads to a higher level of annoyance in the shoulder as compared to the hat in spite of the same amount of distortion introduced in the two portions. We have indicated the objective quality scores from PSNR and the proposed Q_{TID} (which means that TID database is used as the training set). Note that Q_{TID} will predict scores in the form of MOS because the training database (TID) comprises of MOS. Therefore, a higher Q_{TID} means better quality.

(a) PSNR = 32.0705, Q_{TID} = 3.4677

(b) PSNR = 32.1293, Q_{TID} = 3.3173

(c) PSNR = 33.8977, Q_{TID} = 4.7885

(d) PSNR = 33.1732, Q_{TID} = 5.4665

Figure 4.1: Perceptual effect of noise in different image areas
(a) White noise distorted hat part, (b) White noise distorted shoulder part, (c) White noise distorted building part and (d) White noise distorted plants part. The objective predictions from PSNR and Q_{TID} have been indicated below each image. For reference Q_{TID} = 5.7966 for the image with no distortions. The images have been cropped for better visibility.

One can see that PSNR predicts higher score for the image which has noise in the

smooth part (more annoying) as compared to the other image which is not consistent with HVS. In contrast, Q_{TID} predicts lower score for the image with distortion in the smooth part (shoulder) and higher score for the other image. Next, we consider the images shown in Figure 4.1 (c) and (d). We have indicated two portions in this image by boxes. We added the same amount of WGN to these two portions. As can be seen, the distortion in the "building" part is more visible and thus more annoying to the human eye. On the other hand, the area with "plants" is textured and can tolerate such distortions [27]. In fact, the white noise in that part cannot be easily noticed by the human eye. It can be noted that PSNR gives higher score for image in Figure 4.1 (c) while lower score to the image in (d) where most of the noise is not visible due to masking. We have already mentioned that this happens because PSNR assigns equal importance to all the errors independent of their perceptual impact. On the other hand, the proposed approach is able to capture the effect of noise masked due to texture and assigns higher score to Figure 4.1 (d) and lower score to the image in Figure 4.1 (c). It may be mentioned that in the four images shown in Figure 4.1, the distortion (in this case white noise) has been added in different parts of the image. In (a) mainly the edge part is distorted, in image (b) smooth portion has been corrupted, in image (c) a visually more salient region has been distorted and in image (d) textured portion is distorted.

It may be further mentioned that according to Q_{TID} scores given in Figure 4.1, noise in smooth portion causes largest perceptual annoyance (Q_{TID} is smallest) followed by noise in edge regions while the perceived loss of visual quality is the least in the textured region (Q_{TID} is the largest). This confirms that the perceptual impact of distortion in different portions is reasonably well handled by the proposed scheme. This demonstrates the effectiveness of the proposed SVD based features and their proper pooling via SVR.

The foregoing discussion and analysis was meant to provide a visual illustration of the effectiveness of the proposed method and how it can handle distortions according to their perceptual significance. In the following sections we provide the test results using a large number of images and distortion types for a more thorough and comprehensive metric validation.

4.4.3 Performance evaluation on image databases

In Figure 4.2, the results for the proposed Q_{full} and other existing metrics are presented. The C_P values of different metrics are shown in Figure 4.2 (a), where we can see that Q_{full} performs well in general. We can also see that the existing metrics do not perform well for all the test databases. For example, we note that the performance of PSNR, VIF, VSNR, SSIM, MSVD and IFC is worse on WIQ database since these metrics generally perform better for images containing single artifact in image [44]. As aforementioned, the images in the WIQ database can contain more than one artifact (like blocking and ringing together in a same image) due to the complex nature of a wireless communication link. Similarly, for the A57 database, the performance of PSNR, VIF, SSIM, MSVD and IFC is relatively poor. As can be also seen, VSNR which performs well for A57 database does not perform as well on the other databases. The IFC metric performs well for LIVE and IVC databases but its prediction performance is worse on the remaining databases. By contrast to the existing metrics, the performance of Q_{full} is more consistent across all the databases and generally better than all the other metrics being compared. Recall that for Q_{full} none of the images in the training set appear in the test set. Therefore, the proposed metric exhibits robustness and training with specific image contents is not necessary.

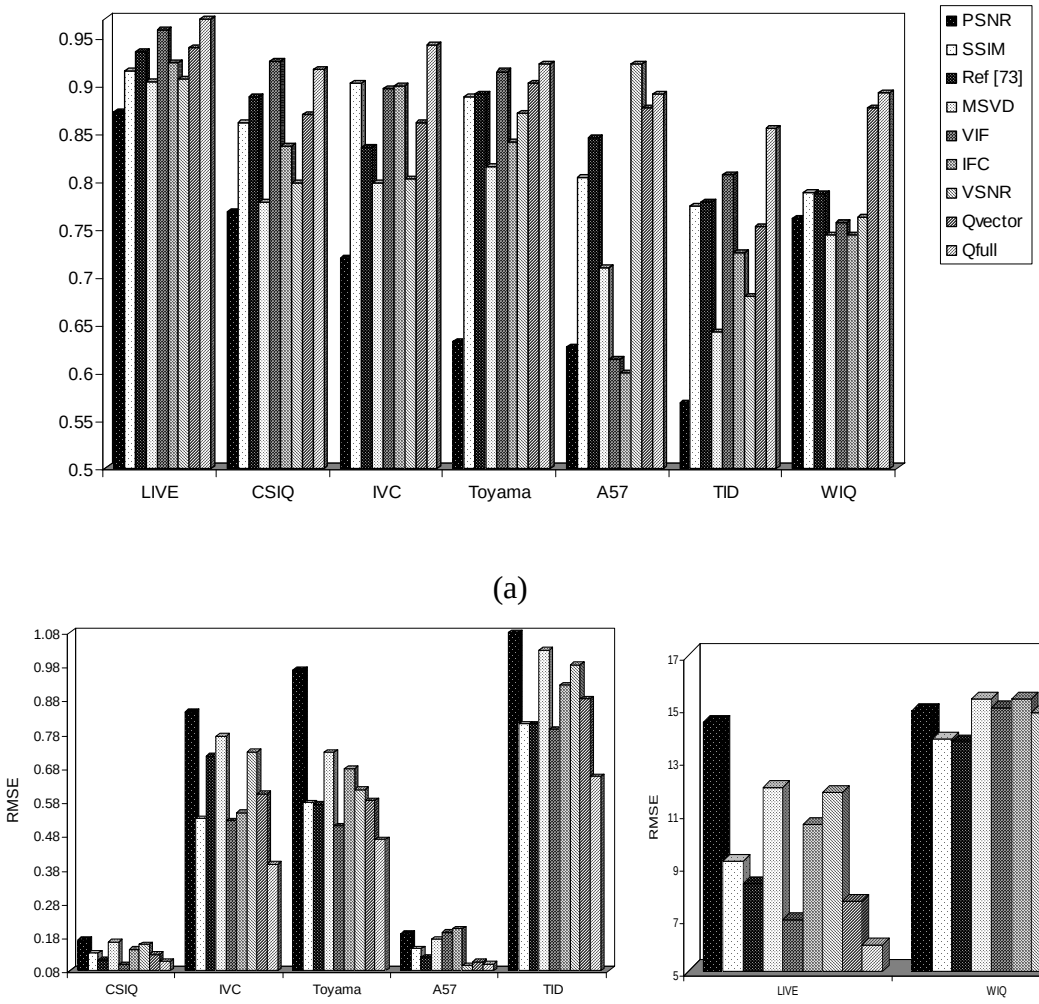
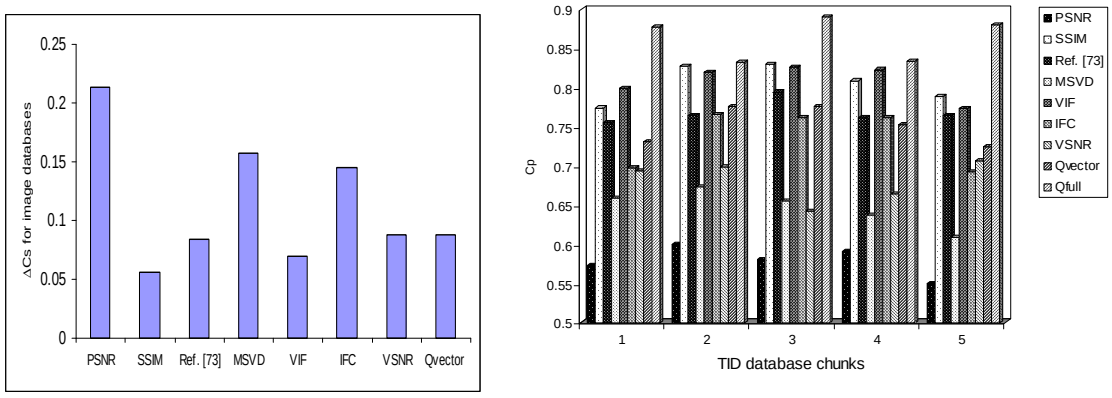


Figure 4.2: Performance comparison on 7 image databases
(a) C_P comparison on different image databases, (b) RMSE for CSIQ, IVC, A57 and TID databases and (c) RMSE for LIVE and WIQ databases

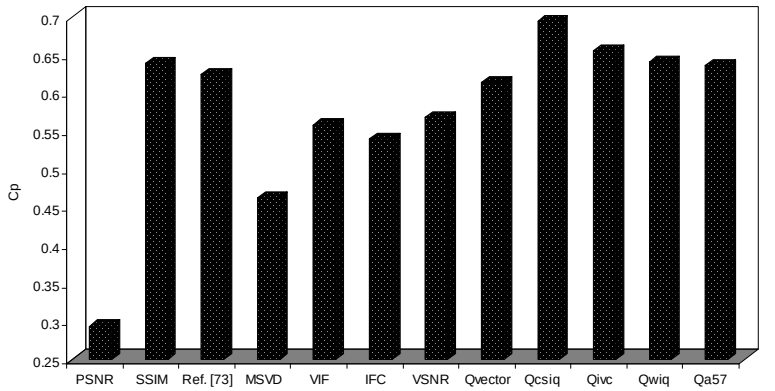
It has been also found that, in general, the prediction performance of the proposed scheme is consistent over all the test chunks. We illustrate this through the performance on the TID database which consists of 25 original images. Following the splitting procedure detailed in Section 4.4.1, we obtained 5 chunks each with 340 images. The C_P values of different metrics for each TID test chunk are shown in Figure 4.3 (b).



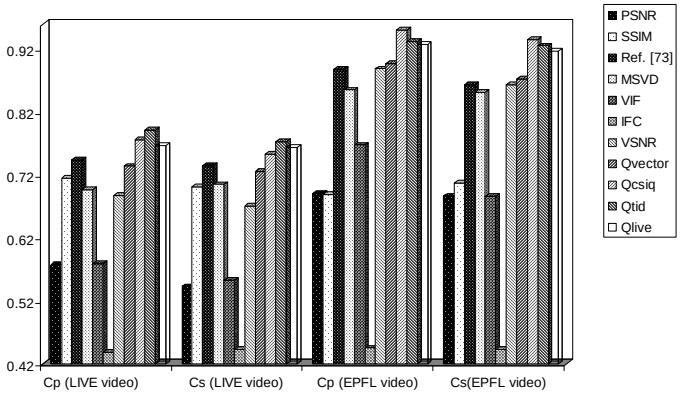
(a) (b)

Figure 4.3: Performance comparison

(a) Average ΔC_S values over the 7 image databases for different metrics (b) C_P values for 5 chunks of TID database



(a)



(b)

Figure 4.4: Performance comparison for 5 distortion types and video databases

(a) C_P values for the 500 images from TID database with 5 distortion types (see text for further explanation), (b) C_P and C_S values for LIVE and EPFL video databases

The proposed system performs well consistently for all the test chunks and is better than the other metrics. The consistency in prediction performance was similarly observed for all the other databases. This indicates that the proposed system performs well across varied images and distortions and does not show any dependency on any specific image/distortion content. We further present the results of the F-test in Figure 4.5. According to Table 3.1, the points which lie above the $F_{critical}$ boundary denote the cases for which the proposed scheme is better and also statistically distinguishable than the existing metric under comparison. We can see from the figure that a large number of points (about 70% of them) are above the $F_{critical}$ boundary, indicating that the proposed Q_{full} is statistically better in comparison with the other metrics. The points which lie between the $F_{critical}$ curve and the line $F = 1$ (i.e. $1 < F < F_{critical}$) denote that the cases for which the proposed Q_{full} is still better than the corresponding metric since $F>1$, but statistically indistinguishable.

We note that only two points (2.8% of the cases) fall below the $F = 1$ boundary. In these two cases, the proposed scheme performs worse than the corresponding metric since $F < 1$, and is statistically indistinguishable from those two metrics (VIF and VSNR for CSIQ and A57, respectively). There is no single case for which $F < 1/F_{critical}$, i.e., the proposed method has not been statistically worse than any existing metric with any database under comparison.

Since C_P and C_S exhibit similar trends, we only show the average difference in C_S values (with respect to Q_{full}) over the 7 image databases for the 8 existing metrics in Figure 4.3 (a). As can be seen, all the ΔC_S are positive, indicating the better performance of Q_{full}. Figure 4.2 (b) and (c) also indicate that Q_{full} outperforms other metrics in terms of RMSE.

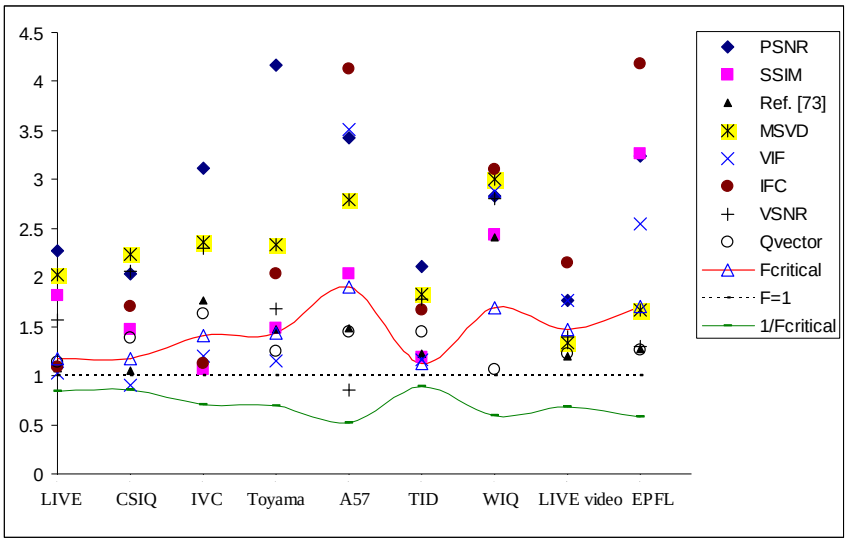


Figure 4.5: F-test plot for different image and video databases (the points above the Fcritical boundary denote the cases for the proposed scheme to be statistically better than the corresponding metric

4.4.4 Cross-database validation

For the cross database evaluation, we selected the 3 biggest image databases available, namely TID (1700 images), CSIQ (866 images) and LIVE (779 images), for training. As can be seen from Table 4.1 with different databases as training and test sets, the proposed method performs well across all the databases similar to the *k*-fold CV tests. We have reported only the C_P values in Table 4.1 since C_S and RMSE show similar results as C_P. We can also see from Table 4.1 that Q_{CSIQ} gives C_P value of 0.7550 for TID database which is comparable to other metrics like SSIM and VIF and better than PSNR, VSNR, IFC and MSVD. This is significant since in this case, the training set (866 images) is only about a half of the size of the test set (1700 images of different visual contents). The proposed metric also performs better than all the other metrics as indicated by higher C_P values achieved by Q_{CSIQ}, Q_{LIVE} and Q_{TID} for the WIQ database.

Table 4.1: C_P values for cross-database validation

Test database/ Metric	LIVE	CSIQ	IVC	Toyama	A57	TID	WIQ
Q_{CSIQ}	0.9086	--	0.8828	0.8327	0.8843	0.7550	0.7764
Q_{LIVE}	--	0.8581	0.8877	--	0.8807	--	0.7314
Q_{TID}	--	0.8831	0.8755	--	0.8854	--	0.7580
$Q_{watermark}$	0.9004	0.8267	--	0.8782	0.8064	0.7219	0.7202
Q_{vector}	--	0.8525	0.7884	--	0.8223	--	0.7573

Overall, we can see from Table 4.1 that the proposed metric is consistent and gives good prediction performance for the cross database evaluation. We have also shown the scatter plot for LIVE image database with Q_{CSIQ} as the objective metric in Figure 4.6. The data points corresponding to the 5 types of distortions present in this database are highlighted using different notations/colors. As can be seen, the plot is compact around the logistic fitting curve and shows low scattering around it. Therefore the prediction performance of the proposed metric is good for all the distortions as none of the data points scatter too much around the logistic fitting curve. Note that a large scatter would imply poorer performance.

As mentioned Section 4.4, we also use the image database in which images are distorted due to watermarking (please refer to the Appendix for details). This type of distortion is different from other commonly occurring distortions (like JPEG, blur, white noise distortion etc.) due to the specific processing that images undergo. We used this database only as a training set to further confirm the robustness of the proposed scheme to new and untrained distortions. Similar to the previous notations, $Q_{watermark}$ denotes the training with watermarked image database. However, out of 5 we only used 3 original images and their distorted versions as the training set. This again ensures that images used for training are excluded from the test sets. Note that we excluded two images

namely '*monarch*' and '*rapids*' which are present in many other databases. As can be seen from Table 4.1, $Q_{watermark}$ performs quite well. This further confirms that quality degradation due to different distortion types can be assessed by exploiting the underlying common patterns characterized by the *structure* loss. Note that for $Q_{watermark}$ the training set which consists of 126 images and their corresponding subjective scores is relatively small as compared to test databases like TID, LIVE and CSIQ. Thus, the results obtained for $Q_{watermark}$ are significant because the system is trained on a completely different distortion to those in the test databases.

It is also worth pointing out that the subjective quality score range is different for all the databases. For instance, LIVE includes subjective scores as DMOS in the range 0-100 while TID gives subjective results in the form of MOS in the range 0-9. The IVC database consists of MOS in the range 0-5 while the CSIQ database reports DMOS in the range 0-1. The A57 database includes subjective scores as DMOS in the range 0-1. Thus, Q_{TID}, before the logistic fitting, gave C_P value of -0.8755 for the CSIQ database, -0.7656 for the WIQ database and -0.8752 for the A57 database. All the resulting correlations here are negative due to the fact that the system was trained with MOS while it was tested with DMOS, which has opposite range of valuation in quality specification.

Another aspect of note is the robustness to untrained distortions. For the cross database tests, since the training and test sets come from different databases, many of the distortion types appearing in the test set are not represented in the training set. The good performance of Q_{CSIQ}, Q_{LIVE}, Q_{TID} and $Q_{watermark}$ for WIQ database shows the robustness of the proposed method to complex distortions which are not present in the training set.

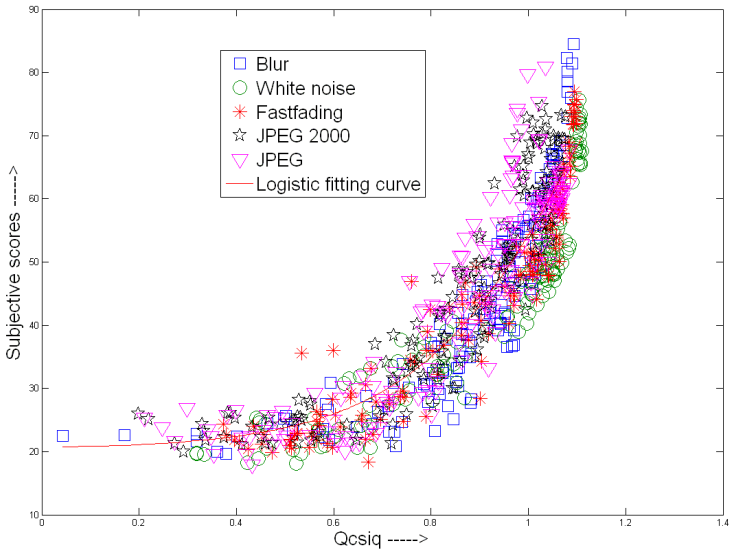


Figure 4.6: Scatter plot for the LIVE image database with Q_{CSIQ} as the objective metric

Similarly, many of the distortion types present in the TID database do not occur in the CSIQ database and hence the C_P value of 0.7550 given by Q_{CSIQ} is noteworthy. Similar observations hold for Q_{LIVE}, Q_{TID} and $Q_{watermark}$. In order to further test the robustness to untrained distortions, we tested the images from the TID database which were distorted by 5 types (from the total of 17 types) of distortions: image denoising, non eccentricity pattern noise, local block-wise distortions of different intensity, mean shift (intensity shift) and contrast change. These 5 distortions were chosen since they do not appear in any other database and also form a challenging set of distorted images to be assessed for visual quality. For example, consider the case of denoised images. The PSNR for a denoised image is generally higher than that of the original noisy image but at the same time, the denoised image may visually look worse than the corresponding original noisy image [120]. Hence, quality assessment of such images is not straightforward. The next distortion type considered is the local block-wise distortions of different intensity. For the first level of distortion, 16 image blocks (block size is 32 × 32) were distorted in each

image, for the second level of distortion 8 blocks were distorted, for the third level of distortion 4 blocks and for the fourth level 2 blocks were distorted. Recall that for TID database, the first distortion level corresponds to the highest PSNR while the fourth level of distortion corresponds to the lowest PSNR. It has been found that [120] an image in which two blocks have been corrupted (i.e. the fourth distortion level) is perceived as having a better visual quality (although it has smaller PSNR) than the image in which 16 blocks have been corrupted (i.e. the first distortion level). This suggests that a lower amount of distortion spread over a larger area is likely to cause more quality degradation than a higher amount of distortion spread over a smaller area. Therefore, quality assessment of such images can be tricky for the metrics. Likewise, contrast and intensity changes (up to a certain level) generally do not affect the visual quality substantially (in spite of the presence of pixel errors) although the PSNR may change considerably.

Hence, images with these 5 distortion types are indeed challenging for metrics. We have tested these 500 images (100 images for each of the 5 distortion types) with the training sets being CSIQ (Q_{CSIQ}), IVC (Q_{IVC}), WIQ (Q_{WIQ}) and A57 databases (Q_{A57}). By training with these databases, it is ensured that the training and test images are different. We have also computed the results for the other metrics for comparison. We can see from Figure 4.4 (a) that metrics like VIF, VSNR, PSNR and MSVD do not perform well ($C_P < 0.6$ for these metrics), while the proposed scheme performs better than the other metrics. It may be noted that Q_{CSIQ}, Q_{IVC}, Q_{WIQ}, and Q_{A57} all perform quite well. This result is significant since the IVC, WIQ and A57 databases contain significantly less number of images than the number of test images.

4.4.5 Performance evaluation on video databases

The performance of the proposed method has been evaluated on the video databases using the cross database evaluation. The trained system is used to predict the quality score of each individual frame. The same procedure was also adopted for evaluating the other metrics. In this study, the overall quality score of the video is determined as the average of the scores all the frames in the video. We present the C_P and C_S values of different metrics for the two video databases in Figure 4.4 (b). As can be seen, Q_{CSIQ}, Q_{LIVE} and Q_{TID} all perform better than the existing metrics under comparison. One can also note that the C_P and C_S values lead to similar conclusion regarding metric performance. The RMSE values (not shown here to save space) were also found to be consistent with C_P and C_S. Since the training is done with image databases only, the good performance of our method is indicative of its generalization ability to new visual/distortion content. The F-test results for video have also been indicated in Figure 4.5. For the video databases, the F values were calculated against residuals of Q_{TID}.

The two video databases used in this study (LIVE and EPFL) represent different visual contents since they use different original video sequences and thus provide diverse visual contents for testing the robustness of the proposed algorithm. Interestingly, we can see from Figure 4.4 (b) that all the metrics give relatively better performance for EPFL video database than for the LIVE video database. One reason for this is that LIVE video database includes 4 distortion types as compared to the EPFL video database in which the sequences are impaired only by packet loss. Another reason is that in the LIVE video database the distortion strength has been adjusted perceptually [172]. As an example of the perceptual adjustment, consider four labels for visual quality (“Excellent”, “Good”, “Fair” and “Poor”) and one reference video sequence ‘*Tractor*’ from the LIVE video

database. Four MPEG-2 compressed versions of '*Tractor*' are chosen to approximately match the four labels for visual quality. Similar procedure is applied to select H.264 compressed, wireless and IP distorted versions. The "Excellent" MPEG-2 video and "Excellent" H.264 video are designed to have the approximately same visual quality and similar perceptual adjustment has been made for other distortion categories and quality labels. On the other hand, for EPFL database, the packet loss rates have been fixed apriori. It has been argued [172] that adjusting the distortion strength perceptually, as done for LIVE video database, is far more effective towards challenging and distinguishing the performance of visual quality metrics than, for instance, fixing the compression rates/packet loss rates across sequences. Due to these two reasons, LIVE video database is more challenging for visual quality metrics.

The adopted procedure of assessing video quality by using the average quality scores of frames takes into account the spatial information in the video but the temporal information is disregarded in this case. Nonetheless, in this part of the work, our aim is to demonstrate the performance of the proposed system to untrained visual/distortion contents.

4.4.6 Computational Complexity

In this section, we provide an indication of the execution time of different metrics i.e. time required for predicting quality of an image. We measured the average execution time required per image in the A57 database (image resolution is 512 × 512) on a PC with 2.40 GHz Intel Core2 CPU and 2 GB of RAM. Table 4.2 shows the average time required per image (in seconds), with all the codes implemented in Matlab.

Table 4.2: Average execution time for different metrics (in sec.).

Metrics	SSIM	MSVD	VIF	VSNR	PSNR	IFC	Ref. [73]	Q_{vector}	Proposed
Time	0.0454	0.6036	3.4829	0.4452	0.0037	4.4490	5.0333	5.1723	1.03

We note that the proposed method is computationally more expensive than metrics like PSNR and SSIM due to the fact that SVD is computationally intensive. The exact SVD of a $r \times c$ matrix has time complexity $O(\min\{rc^2, r^2c\})$. However, the computational cost and time are reduced due to the fact that we use block based SVD (although block size is large but still smaller than the full image). Furthermore, many fast and efficient implementations of SVD are available which can lead to decrease in SVD computation. Training the SVR is of higher computational requirement but the model training can be done off-line.

To give more precise estimates of the time required for training and testing, we present an example below with TID as the training database and A57 being the test database. First we extract the features for the images in the TID database for training the system, and the time taken is about 1306 sec (totally there are 1700 images in TID database) which means about 0.7687 sec. per image (note that image size is 512 × 384 in TID database). Next we train to obtain the model Q_{TID} by training with the features extracted. It took about 2.5776 sec. to obtain the trained model Q_{TID}. So the total time for developing Q_{TID} is approximately 1309 seconds. This of course can be developed off-line. Note that training time is directly proportional to the number of training samples used. For testing, the time required for feature extraction per image is about 1 sec. per image (note that image size is 512 × 512 in A57 database) as measured from the 54 images of the A57 database (it took 53.7765 sec. for extracting the feature vectors of the 54 images in the database). The time required for the prediction of quality (after extracting the

features) using Q_{TID} is negligible (only about 0.03 sec. per image). Because the prediction model (in this example Q_{TID}) is developed off-line, it takes approximately 1 (feature extraction) + 0.03 (for prediction) = 1.03 seconds to predict the quality of a 512 × 512 image. The proposed method is however has lower complexity than more sophisticated metrics like VIF and IFC which employ wavelet decomposition.

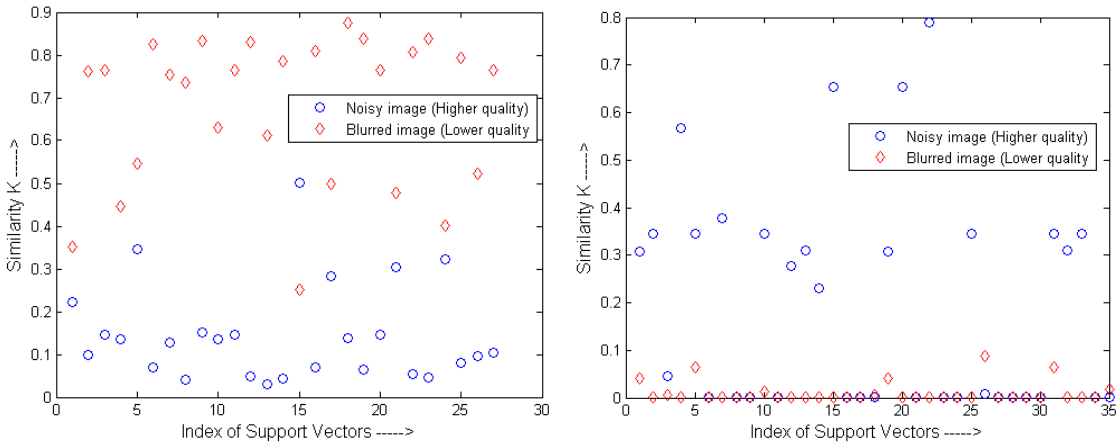
4.4.7 Further observations

As aforementioned, SVs are the samples for which inequality (4.9) is not satisfied, i.e., they lie outside the ε-tube. They are the critical datapoints which can be considered as the representative of the whole training set. In our experiments, we observed that the SVR algorithm tends to select the images which either have near-threshold distortions (i.e. low distortion level) or images with much higher distortion levels as the SVs. For example, consider the CSIQ database for which DMOS is in the range [0, 1]: a DMOS close to 0 implies low distortion while that close to 1 means high distortion as perceived by the subjects. We have found that samples which were chosen as the SVs for the CSIQ database corresponded to either DMOS less than 0.056 or DMOS greater than 0.846. Similarly for the other databases, the selected SVs corresponded to either relatively low or high distortion levels. This appears to be a reasonable and intuitive selection of SVs for visual quality assessment, because images with very low and very high distortions are the representative of the overall visual quality range variations. The significance of this can be explained based in the fact that the term $K(\boldsymbol{x}_i,\boldsymbol{x})$ represents the similarity between the SVs $\boldsymbol{x}_i$ and the test image $\boldsymbol{x}$. Obviously if the test image is of higher quality, it will yield greater kernel similarity value (i.e. $K(\boldsymbol{x}_i,\boldsymbol{x})$ will be bigger) with the SVs which represent higher quality signal. On the other hand, it will have lower similarity ((i.e.

$K(\boldsymbol{x}_i,\boldsymbol{x})$ will be smaller) with the SVs representing low quality signals.

To illustrate this point further we considered two distorted images: (a) image with white Gaussian noise, (b) Blurred image. The noisy image was of higher visual quality than the blurred image. We denote the feature vector of noisy image as $\boldsymbol{x}_n$ while $\boldsymbol{x}_b$ denotes that for the blurred image. We then computed the kernel similarity scores $K(\boldsymbol{x}_i,\boldsymbol{x}_n)$ and $K(\boldsymbol{x}_i,\boldsymbol{x}_b)$ by measuring their distances from the SVs $\boldsymbol{x}_i$. Note that $K(\boldsymbol{x}_i,\boldsymbol{x}_n)$ and $K(\boldsymbol{x}_i,\boldsymbol{x}_b)$ will be n_{sv} (the number of SVs) dimensional vectors and their elements denote the similarity scores of the respective image feature vectors with the SVs (0 indicates no similarity and 1 means complete similarity). We show the kernel similarity of the feature vectors for noisy and blurred images in Figure 4.7 where the plot in (a) are the similarity scores with the SVs corresponding lower quality images (MOS < 2) while the plot in (b) shows the similarity with the SVs corresponding higher quality images (MOS > 6.5). We chose MOS < 2 and MOS > 6.5 because in TID database 0 < MOS < 9 with 0 denoting worse quality and 9 indicating best quality.

One can observe from Figure 4.7 that the noisy image tends to have higher similarity with SVs corresponding to higher quality images and lower similarity with SVs corresponding to lower quality images. On the other hand, blurred image shows the opposite trend. Examination of the corresponding scaling factors $(\eta_i^* - \eta_i)$ (see Eq. (4.12)) reveals that they are generally large and positive for the SVs corresponding to higher quality images. In contrast, they are either small or negative for the SVs corresponding to the lower quality images. Because the final quality score is a summation (as given by Eq. (4.12)) of the similarity scores scaled by the corresponding factor (the bias is same), this results in a higher quality score for noisy image and lower quality score for the blurred image.



(a) (b)

Figure 4.7: Plot of kernel similarity scores

(a) Kernel similarity scores of the noisy and blurred images with the SVs corresponding to lower quality images (MOS<2), (b) Kernel similarity scores of the noisy and blurred images with the SVs corresponding to lower higher images (MOS>6.5)

Therefore the selection of SVs provides an insightful explanation of how the system predicts quality. This also highlights the effectiveness of the proposed SVD features since they enable proper selection of the SVs by allowing adequate distinction between images of different perceived qualities. We also observed that the number of SVs was much smaller compared to the number of training samples. This is advantageous from point of view of computational requirement. The number of SVs was found to decrease with increasing ε value which is expected since more samples fall within the ε-tube, and the associated performance changes were graceful. For example, the experiments with the LIVE image database show that the number of SVs decreases from 295 for which $C_P = 0.9677$ to as low as 54 (i.e., only 7% of datapoints) for which $C_P = 0.9579$.

We have a few additional remarks for the feature selection. First of all, the smaller number of SVs as a result of SVR training indicates the efficiency of the proposed feature selection and SVR formulation. Secondly, as we know, metrics MSVD, Q_{vector}

and Q_{full} use singular values, singular vectors and their combination, respectively; as demonstrated consistently in Figures 4.2-4.5, the performance of these three increases in the aforementioned order with each performance assessment criteria, namely, C_P, C_s, RMSE, and F-test. It can be concluded that as analyzed and expected, singular vectors and singular values together provide a more comprehensive basis for visual quality assessment.

4.5 Concluding Remarks

In this chapter, to tackle effective feature fusion in visual quality evaluation, we have proposed an SVR based metric which operates with SVD based features as the input (we have demonstrated the effectiveness of SVD even with a simpler fusion rule in the previous chapter). The feature selection based on comprehensive SVD analysis is novel, since adaptively determined singular vectors allow the capturing of structural information for each image (or a frame in video) and the separation of luminance and structural information enables their differentiation toward the assessment of perceptual quality.

We have used SVR to result in a model for combining the SVD features to predict the perceptual quality score. Note that we also adopted the following modifications (as compared to the previous chapter) in implementing our method: (1) block-based feature extraction so that feature vector does not depend on image size, (2) weight of SVD values for normalization in the range [0,1] and (3) different summation method in comparison to Eq. (3.10) for reasons already given in Section 4.2. With the proposed model, we have avoided apriori assumptions on the distortion statistics (as an important advantage over the existing pooling methods) and exploited the underlying common

patterns associated with visual quality degradation characterized by structural and luminance/textural changes (that is, training with specific visual and/or distortion content is not necessary). Each high dimensional feature vector was mapped into a perceptual quality score which is better aligned with the subjective viewing ground truth.

We have devoted a significant portion of this chapter for the experimental results and the related analysis to provide thorough and convincing ground for the proposed scheme. The proposed scheme is found to be consistently better in its prediction accuracy than the eight existing metrics across all the ten public databases which span a wide variety of visual and distortion content. It performs well for visual and distortion content which do not appear in the training set (within a same database and also across different databases). The robustness to untrained images and distortions is crucial since in practice the visual and distortion contents are generally unknown. The chapter also provides more insights regarding the SVs and their role in visual quality prediction. Finally, as expected, the experimental results in this chapter demonstrate improvement in prediction accuracies as compared to non machine learning based feature pooling (employed in Chapter 3).

Chapter 5

Visual Quality Assessment with 2D Mel-cepstrum

5.1 Introduction

In this chapter we present a new method based on the engineering approach to evaluate visual quality objectively. To that end, we use visual features based on 2D mel-cepstrum and machine learning for feature pooling. The 2D mel-cepstrum features are derived from 2D cepstrum which has been used in the past in image registration and filtering applications [234]. We first investigate and provide justification for the use of the said features in assessing visual quality. It is shown that they can be exploited to capture the loss of important structural information which in turn is used to quantify the loss of visual quality. Furthermore, these features also account for the supra-threshold [193] effect that plays a role in visual quality assessment (as further elaborated in Section 5.2.1).

Similar to the previous chapter (Chapter 4), we use SVR for feature pooling due to the advantages outlined earlier. Extensive experiments conducted using six publicly available image databases (totally 3,211 images with diverse distortions) and one video

database (with 78 video sequences) demonstrate the effectiveness and efficiency of the proposed metric, in comparison with seven relevant existing metrics. We also compare the proposed algorithm with the SVD-based method proposed in the previous chapter. It is found to be overall slightly better with regards to prediction accuracy and as an added advantage the new method takes less time for predicting the quality of an image (or video). As a result, the scheme presented in this chapter is more efficient than many existing schemes as well as our SVD-based scheme.

The remainder of this chapter is organized as follows. Section 5.2 discusses the proposed visual quality metric detailing the feature extraction and pooling procedure with proper analysis and justification. Substantial experimental results and the related analysis are presented in Section 5.3. Section 5.4 provides a comparison with SVD-based algorithm (proposed in the previous chapter). Finally, Section 5.5 gives the concluding remarks.

5.2 New Visual Quality Metric using 2D Mel-cepstrum

In this section, we describe the details of the proposed metric whose block diagram is shown in Figure 5.1. The first step is to extract the 2D mel-cepstral features from both the reference and distorted images. Then, the difference vector between the two feature vectors is computed to measure the loss of structural information. Finally, machine learning is used to fuse the elements of the difference vector.

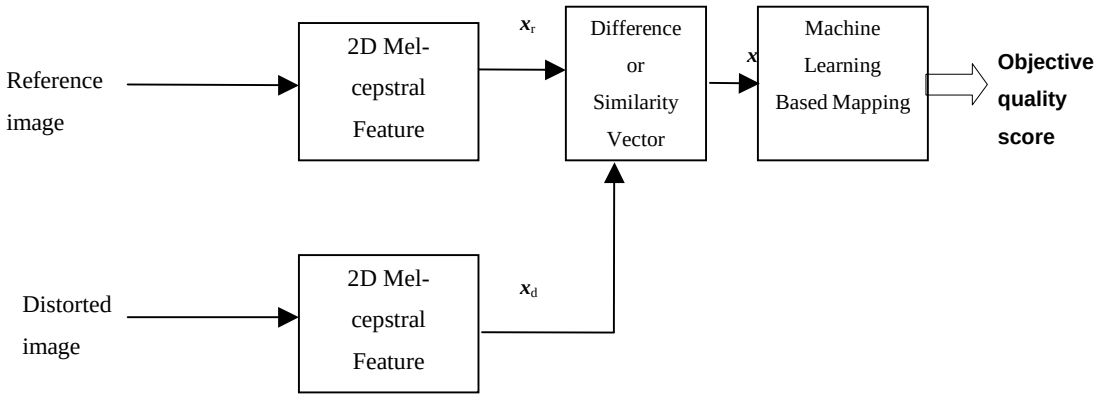


Figure 5.1: Block Diagram of the proposed scheme

5.2.1 **Feature extraction based on 2D mel-cepstrum**

An error (or distortion) in different contexts may not have the same perceptual impact on quality. For example, low pass filtering (i.e. blurring) has lesser effect on the smooth areas in an image while it has a higher impact on edges. Due to this, it is important to distinguish/differentiate error in different image components. This is the reason why the PSNR (or related metrics like MSE) is less effective: it does not separate/differentiate the signal components because it assigns equal weights to all the pixel errors irrespective of their perceptual impact. Therefore, the motivation behind feature extraction is to separate/differentiate the image signal into its components since their contribution to the perceived quality is different. This is a crucial step towards more effective quality assessment because the separation of the components will then allow us to treat (i.e., weigh) them appropriately according to their perceptual significance. In this chapter, we use the mel-cepstral analysis for images to extract meaningful components from the image signal.

Mel-cepstral analysis is one of the most successful and widely used feature extraction

techniques in speech processing applications including speech and sound recognition [173]. Inspired by its success in various areas of audio/speech processing, we propose its exploitation to assess the quality of images objectively. The 2D mel-cepstrum has been proposed recently [151] and the proposed scheme is the first attempt in the existing literature to explore the 2D mel-cepstrum for visual quality assessment.

The 2D cepstrum $\hat{c}(p,q)$ of a 2D image $y(n_1,n_2)$ is defined as

$$\hat{c}(p,q) = F_2^{-1}(log(|Y(u,v)|^2)) \tag{5.1}$$

where (p,q) denote 2D cepstral quefrency [174] coordinates, $Y(u,v)$ is the 2D DFT of the image $y(n_1,n_2)$ (size N by N) and defined as

$$Y(u,v) = \frac{1}{N}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} y(n_1,n_2) e^{-j2\pi\left(\frac{un_1+vn_2}{N}\right)} \tag{5.2}$$

and F_2^{-1} denotes the 2D Inverse DFT given by

$$F_2^{-1} = \frac{1}{N}\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} Y(u,v) e^{j2\pi\left(\frac{un_1+vn_2}{N}\right)} \tag{5.3}$$

Direct use of frequency coefficients as features will be less effective in determining the visual quality, due to the following two reasons. Firstly, energy of natural images drops at high frequencies (i.e. natural images have stronger low frequency components as compared to high frequency ones). Due to this, the effect of high frequency components is suppressed as the bigger values of low frequency coefficients will tend to dominate. Furthermore, the number of coefficients for the whole image is very large (equal to image size). Instead of direct use of frequency coefficients, we use 2D mel-cepstrum in which non-uniform weighting is employed to group the frequency coefficients.

Specifically, in 2D mel-cepstrum the DFT domain data are divided into non-uniformly in a logarithmic manner and the energy of each bin is computed as

$$G(m,n) = \sum_{k,l \in B(m,n)} w(k,l)\ Y(k,l) \tag{5.4}$$

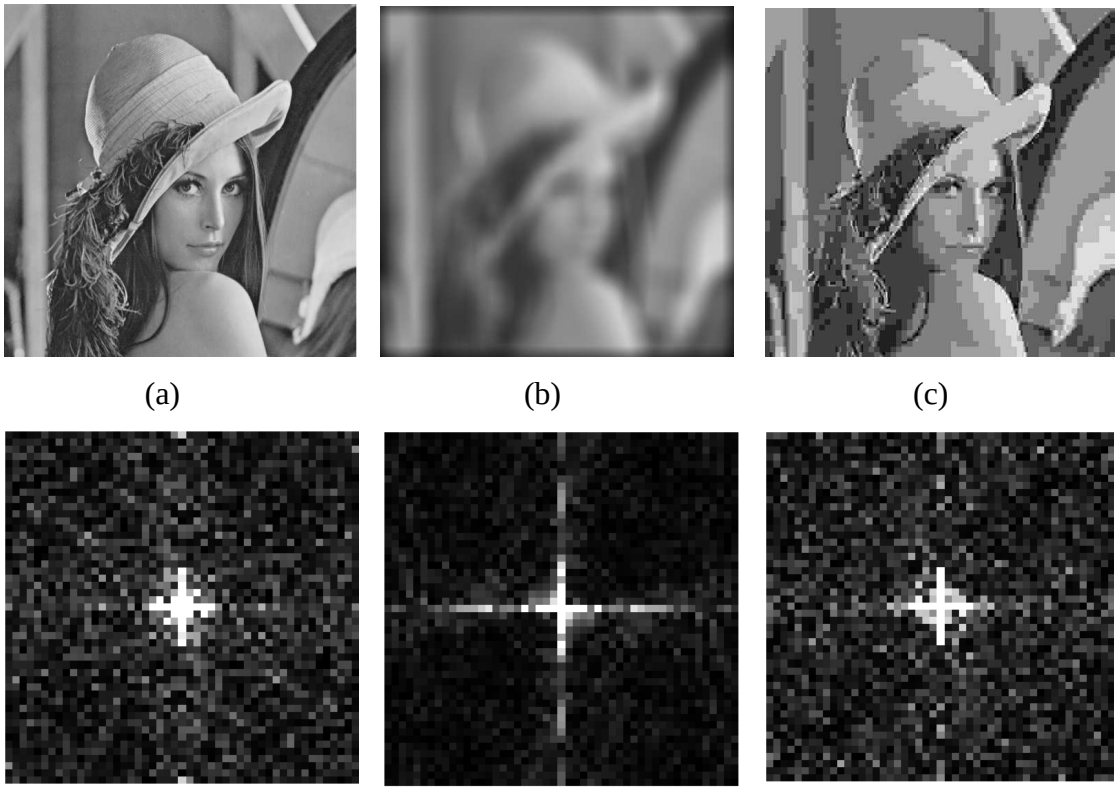
where $B(m,n)$ is the $(m,n)^{th}$ cell of the logarithmic grid corresponding to weight $w(k,l)$ (bigger weight is assigned to high-frequency coefficients) [151].

This approach is similar to the mel-cepstrum computation in speech processing where the weights are assigned using a mel scale in accordance with the perception of the human ear. Like speech signals, most natural images contain more low frequency information. Therefore, as mentioned, there is more signal energy at low-frequencies compared to high frequencies. So non-uniform weighting is employed to emphasize high frequencies. Finally, the 2D mel frequency cepstral coefficients $\hat{c}(p,q)$ are computed using the inverse DFT (IDFT) as

$$\hat{c}(p,q) = F_2^{-1}(log(|G(m,n)|^2)) \tag{5.5}$$

Note that in Eq. (5.5) we have used the absolute value of the bin energy $G(m,n)$ (i.e. magnitude) and discarded phase for reasons given later in Section 5.3.4.

We now analyze why the 2D mel-cepstrum features form a good image representation for visual quality assessment. Psychovisual studies have shown that edges, texture and smooth components in images have different influence on quality. Apart from the distortion in smooth areas (mainly the low frequencies), the HVS is also sensitive to image areas containing edges [175]-[176] and image structure in general (these usually correspond to higher frequency components). Further, image content recognition is widely believed to rely on the perception of image details, such as sharp edges, which are conveyed by higher spatial frequencies [177]-[178].



(a) (b) (c)

(d) (e) (f)

Figure 5.2: Effect of distortions on 2D mel-cepstrum: (a) Original Lena image, (b) blurred image, (c) JPEG compressed image, (d) 2D mel-cepstrum of (a), (e) 2D mel-cepstrum of (b) and (f) 2D mel-cepstrum of (c). White indicates a value of 1 (the highest strength) while black corresponds to 0 (zero strength).

Therefore edges and other higher frequency components also play a role [180]-[182] in quality evaluation. For instance, the SSIM metric has been improved [183] by incorporating edge information. Some other IQA metrics based on edge information can be found in [48]-[49], [184]-[186]. Recently image contours/edges have also been explored for image utility assessment [187] which is related to IQA. As outlined, the 2D mel-cepstrum uses unequal weights for different frequency. As a result, high frequency components can be further emphasized. This is also the reason why the 2D mel-cepstrum

representation is suitable for face recognition [151] (since it highlights edges and other facial features in a face image).

To give an illustration, we show the original 'Lena' image, it's blurred version and JPEG compressed version in Figure 5.2 (a), (b) and (c) respectively. The corresponding 2D mel-cepstrum of the images is shown below the respective images. We observe that blurring mainly damages the high frequency components. This can be visualized through its 2D mel-cepstrum where the strength of high frequency components is reduced. We can also see that the strength of lower frequency components is increased since blur makes the image more uniform. In the extreme case, if all pixels have the same value then we will see only one white spot exactly in the centre of the 2D mel-cepstrum (i.e. the DC component). The case of JPEG compression is different in that it causes blockiness and can introduce false structure or edges in the image. This can again be captured from the 2D mel-cepstrum features because the strength/magnitude of frequency components changes due to JPEG compression. Therefore, of the difference between the 2D mel-cepstrum features of the reference and the distorted images is expected to give a good indication of change in image spatial content (or structural change).

To summarize, the followings are the major advantages of the 2D mel-cepstrum which can be exploited for visual quality assessment:

- Because it is possible to emphasize the high frequency components apart from retaining the lower frequency ones, a more informative and comprehensive representation can be obtained. Specifically, it provides more details about features like edges and contours which are important for the HVS' perception of visual quality. Therefore, it is more effective as compared to other transforms since more

discriminatory and meaningful image signal components can be extracted.

- The resultant 2D mel-cepstrum sequence computed using the IDFT has smaller dimensions than the original image. It can therefore be viewed as a *perceptually* motivated dimension reduction tool which can preserve image structure. That is, it can be considered as a good trade-off between retaining important image information and achieving dimensionality reduction. For an N by N image, using the 2D mel-cepstrum we can obtain the dimension reduced data M by M with $M < N$.

- We obtain decorrelated features, so the redundant information is discarded. This results in a more compact numerical representation of the visual signal to characterize its quality. Thus, the advantage of 2D mel-cepstrum features is that they produce representations that lie in an orthogonal space (due to using IDFT as shown in Eq. (5.5)).

- Another advantage of 2D mel-cepstral features is that small change in the features corresponds to small change in perceptual quality and vice-versa. This property is especially crucial for quality prediction of images with near threshold (i.e. just noticeable) distortions as will be demonstrated later in Section 5.3.1.

- The reader will notice from Eq. (5.5) that 2D mel-cepstrum involves the logarithms of the squared bin energies denoted by $|G(m,n)|^2$. This reduces the dynamic range of the values and is consistent with the so-called suprathreshold or the saturation effect. This means that the ability to perceive variations in the distortion level decreases as the degree of distortion increases [44]-[45], [188]-[189]. The logarithm operation essentially accomplishes this desirable property as elaborated later in Eq. (5.7) and illustrated graphically in Figure 5.3.

- The 2D mel-cepstrum is also associated with clearer physical meaning because it essentially works in the Fourier (frequency) domain which is a well established method for image analysis. However, in the Fourier or DCT domain one usually discards the higher frequency components (for example JPEG compression) in order to achieve dimension reduction. By contrast in 2D mel-cepstrum, the high frequency DFT and DCT coefficients are not discarded in an ad-hoc manner. Instead the high frequency component cells of the 2D DFT grid are multiplied with higher weights as compared to the low frequency component bins in the grid, thus resulting in more suitable image representation for quality assessment.

Let $\boldsymbol{x}_r$ and $\boldsymbol{x}_d$ denote the 2D mel-cepstral features of the reference and distorted images respectively. The vectors $\boldsymbol{x}_r$ and $\boldsymbol{x}_d$ can be thought to represent the *timbral texture space* [190] of the two image signals and we use them to quantify perceived similarity between them. This is similar at the conceptual level to tasks, like computing music similarity [191], genre classification [192], etc. in the field of audio/speech processing. Because our aim is to compute quality of the distorted image with respect to the reference image, we use the absolute difference between the two feature vectors for computing quality of the distorted image and define

$$\boldsymbol{x} = | \boldsymbol{x}_r - \boldsymbol{x}_d | \tag{5.6}$$

We can see that the elements of $\boldsymbol{x}$ represent the absolute difference between the 2D mel-cepstrum coefficients of the reference and distorted images. This lends $\boldsymbol{x}$ a better physical meaning since its elements can be thought as the change in frequency components of the reference image due to distortion, i.e., it accounts for the loss of spatial information.

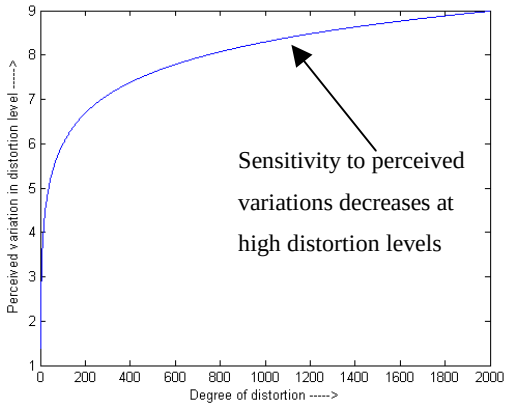


Figure 5.3: Illustration of the suprathreshold or saturation effect

As aforesaid, suprathreshold effect implies that the same amount of distortion becomes perceptually less significant as the overall distortion level increases. Researchers have previously modeled suprathreshold effect using visual impairment scales [193] that map error strength measures through concave nonlinearities. The definition of feature vector in Eq. (5.6) accounts for this effect and can be explained as follows. Eq. (5.6) can be written as

$$\begin{aligned} \boldsymbol{x} &= |\,\boldsymbol{x}_{\mathrm{r}} - \boldsymbol{x}_{\mathrm{d}}\,| \\ &= |F_2^{-1}(log(|G_r(m,n)|^2)) - F_2^{-1}(log(|G_d(m,n)|^2))| \\ &= |F_2^{-1}\left[log\left\{ \frac{|G_r(m,n)|^2}{|G_d(m,n)|^2} \right\} \right]| \end{aligned} \tag{5.7}$$

where $G_r(m,n)$ and $G_d(m,n)$ denote the bin energies from reference and distorted images respectively. We can observe from Eq. (5.7) that the ratio of the absolute bin energies can be regarded as the distortion measure on which suprathreshold function (logarithm) has been applied.

For a simple intuitive explanation, consider the two quantities

$\log\left\{\frac{60}{40}\right\}=0.4055$ and $\log\left\{\frac{1020}{1000}\right\}=0.0198$. As we can see, the difference between the numerator and denominator in the two cases is the same (it is 20). However, the perceived change is smaller in the second case. This saturation effect is visually exemplified in Figure 5.3.

As mentioned before, the feature vector $\boldsymbol{x}$ is effective in characterizing the loss of image structure. To illustrate this point further, we show 7 images in Figure 5.4. In this Figure, image (a) is the original image taken from the LIVE image database [115], while the images (b), (c) and (d) have been obtained by blurring the original image with increasing blur levels; images (e), (f) and (g) have been generated by JPEG compression of the original image with increasing compression levels. As can be seen, the increasing blurring reduces the high frequency content of the original image and destroys its spatial information. Similarly in JPEG compression the high-frequency components are largely removed owing to non-uniform quantization and this leads to loss of visual quality as shown in the second row of Figure 5.4. We also computed the difference vector $\boldsymbol{x}$ for each distorted image with respect to the original image; next, we obtained the sum of the elements of the respective feature vector for each image and the same has been indicated below each respective image. We find that the sum is large for the heavily blurred image (Figure 5.4 (d)), i.e., indicating higher loss of spatial information, while it is small for the less blurred image. A similar trend can be seen for the JPEG distorted images. That is, we get an indication of the loss of spatial information due to artifacts like blur and JPEG compression which can damage image structure. Of course, a simple summation of the elements of feature vector alone will be insufficient for determining overall quality especially in case of complex and diverse distortion types.

(a)

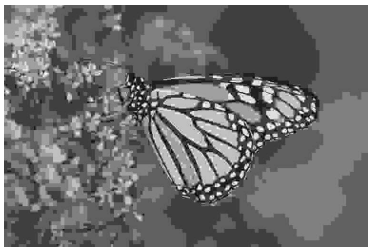
(b)
11.0165

(c)
19.8260

(d)
21.634

(e)
6.4457

(f)
8.8806



(g)
18.4224

Figure 5.4: Indication of the amount of spatial information lost
(a) Original image, (b) low blurring, (c) medium blurring, (d) high blurring, (e) low JPEG compression level, (f) medium JPEG compression level and (g) high JPEG compression level. The number below each respective image denotes the sum of the elements of the feature vector defined in Eq. (5.6). A bigger number denotes more loss of spatial information i.e. higher distortion.

Nevertheless, this analysis indicates that the feature vector $\boldsymbol{x}$ defined in Eq. (5.6) can be expected to be effective for assessing the extent of structure damage or the change in image spatial information because of the external perturbation (distortion). Furthermore, $\boldsymbol{x}$ can be used to assess quality independent of the distortion or image content and the

reason is as follows. Different types of distortions affect visual quality in a largely similar fashion: by introducing structural changes (or change in spatial contents) that lead to different extents of perceived quality degradation. That is, even though $\boldsymbol{x}$ does not take into account the effects of different distortions explicitly, the perceptual annoyance introduced by them is expected to be captured reasonably well. Due to the existence of the underlying common patterns associated with quality degradation, machine learning can be exploited to develop a general model by learning through examples. Like Chapter 4, we have used SVR to fuse the feature vector $\boldsymbol{x}$ into a perceptual quality score.

5.3 Experimental Results and Discussions

We used the same notations as Chapter 4 i.e. symbol Q to indicate the results for the k-fold CV tests, and the symbol $Q_{database}$ to denote the algorithm trained using that database. For instance, Q_{CSIQ} means that CSIQ database has been used for training (refer to the Appendix for database details).

5.3.1 Performance evaluation

The results for the k fold CV tests (denoted by Q) for the individual image databases are presented in Table 5.1. We can see that the proposed Q performs better than the other schemes. Recall that for Q we made sure that the images used for training did not appear in the test set. It was also found that in general, the proposed scheme performed well for individual distortion types. Furthermore, for an overall comparative performance, the averaged results over the 5 image databases are given in Table 5.2.

Table 5.1: Experimental results for the image databases. The 3 best metrics have been highlighted in boldface.

Criteria	Metric	LIVE	A57	WIQ	IVC	TID
C_P	SSIM	0.9473	0.8033	0.7876	**0.9018**	0.7756
	MSVD	0.8880	0.7099	0.7433	0.7975	0.6423
	VIF	**0.9655**	0.6139	0.7559	**0.8966**	**0.8049**
	VSNR	0.9520	**0.9210**	0.7623	0.8025	0.6820
	PSNR	0.9124	0.6273	0.7601	0.7196	0.5677
	PSNR-HVS-M	0.9432	0.8896	0.8191	0.8902	0.5784
	Ref. [46]	0.9253	0.6799	--	0.8776	--
	Q	**0.9684**	**0.9021**	**0.9048**	**0.9511**	**0.8092**
	Q_{TID}	0.9519	**0.9019**	**0.8489**	0.8772	--
	Q_{LIVE}	--	0.8944	0.8473	0.8784	0.7859
	Q_{IVC}	**0.9554**	0.9008	0.8472	--	0.7840
	$Q_{watermark}$	0.9552	0.9011	**0.8480**	0.8794	**0.7881**
C_S	SSIM	**0.9500**	0.8103	0.7261	**0.9017**	**0.7792**
	MSVD	0.9102	0.6485	0.6362	0.7734	0.6520
	VIF	**0.9735**	0.6223	0.6918	**0.8964**	0.7491
	VSNR	0.9400	**0.9355**	0.6558	0.7993	0.7000
	PSNR	0.9056	0.6189	0.6257	0.6885	0.5773
	PSNR-HVS-M	0.9372	0.8962	0.7568	0.8832	0.5952
	Ref. [46]	0.9216	0.7255	--	0.8952	0.6740
	Q	**0.9599**	**0.8586**	0.8064	**0.9171**	**0.7848**
	Q_{TID}	0.9383	**0.8561**	**0.8410**	0.8677	--
	Q_{LIVE}	--	0.8532	**0.8396**	0.8690	0.7732
	Q_{IVC}	0.9442	0.8496	**0.8420**	--	0.7645
	$Q_{watermark}$	0.9433	0.8551	0.8389	0.8688	**0.7690**
RMSE	SSIM	8.0553	0.1914	13.8160	**0.5303**	0.8511
	MSVD	10.6315	0.1731	15.3228	0.7739	1.0285
	VIF	**6.0174**	0.1940	14.9964	**0.5239**	**0.7945**
	VSNR	7.0804	**0.0957**	14.8864	0.7269	0.9851
	PSNR	9.0864	0.6189	14.8856	0.8460	1.1047
	PSNR-HVS-M	8.0564	0.1156	13.1412	0.5550	**1.0947**
	Ref. [46]	--	--	--	--	--
	Q	**5.5731**	**0.0988**	**7.6384**	**0.3649**	**0.7930**
	Q_{TID}	7.0830	**0.1062**	**12.1058**	0.5849	--
	Q_{LIVE}	--	0.1099	**12.1305**	0.5823	0.8296
	Q_{IVC}	**6.8303**	0.1068	12.1688	--	0.8331
	$Q_{watermark}$	6.8430	0.1066	12.1658	0.5800	**0.8261**

Table 5.2: Average performance of different algorithms over 5 images databases. The 3 best metrics have been highlighted in boldface.

Type of Average	Criteria	SSIM	MSVD	VIF	VSNR	PSNR	PSNR-HVS-M	Q	$Q_{watermark}$
Direct Averaging	C_P	**0.8431**	0.7562	0.8074	0.8240	0.7174	0.8241	**0.9071**	**0.8744**
	C_S	**0.8335**	0.7241	0.7866	0.8061	0.6832	0.8145	**0.8654**	**0.8550**
	RMSE	4.6888	5.5839	**4.4988**	4.7547	5.3083	4.5926	**2.8936**	**4.1043**
Weighted Averaging	C_P	**0.8404**	0.7362	**0.8584**	0.7842	0.6961	0.7290	**0.8743**	0.8520
	C_S	**0.8418**	0.7435	0.8279	0.7877	0.6936	0.7369	**0.8522**	**0.8356**
	RMSE	3.5225	4.5175	**2.8547**	3.3182	4.0592	3.6430	**2.5008**	**3.0691**

Table 5.3: C_P values for the 4 distortion levels in TID database. Level 1 indicates the lowest distortion while Level 4 corresponds to the highest distortion. The 3 best metrics have been highlighted in boldface.

Metric	Level 1	Level 2	Level 3	Level 4
SSIM	**0.7564**	0.6102	0.6326	0.6766
MSVD	0.4811	0.5844	0.3869	0.6050
VIF	0.5355	0.5197	**0.8146**	**0.8851**
VSNR	0.6180	**0.6402**	0.4687	0.6492
PSNR	0.5742	0.3241	0.3601	0.3601
PSNR-HVS-M	0.4232	0.5036	0.4657	0.5114
Q	**0.7649**	**0.6464**	**0.6882**	**0.7655**
$Q_{watermark}$	**0.7579**	**0.6376**	**0.6723**	**0.7401**

We computed the average values for two cases. In the first case, the correlation scores were directly averaged, while in the second case, a weighted average was computed with the weights depending on the number of distorted images in each database (similar to [235]).

Table 5.4: Experimental results for EPFL video database. The 3 best metrics have been highlighted in boldface.

Criteria/ Metric	C_P	C_S	RMSE
SSIM	0.6878	0.7080	0.9790
MSVD	0.8554	0.8508	0.6987
VIF	0.7519	0.7524	0.8892
VSNR	0.8838	0.8631	0.6310
PSNR	0.6910	0.6869	0.9750
PSNR-HVS-M	0.8865	0.8760	0.6240
Q_{TID}	0.9390	0.9293	0.4640
Q_{LIVE}	**0.9426**	**0.9321**	**0.4502**
Q_{IVC}	**0.9411**	**0.9311**	**0.4562**
$Q_{watermark}$	**0.9394**	**0.9304**	**0.4626**

We can observe from Table 5.2 that the proposed metric gives better overall performance in both averaging cases for the three evaluation criteria.

As mentioned in Section 4.4.4, we can again see from Table 5.1 that some existing metrics are less consistent since they do not perform well for all the databases. For instance, VSNR does well on A57 but its performance is relatively low for other databases; VIF performs well on 3 databases but performs rather poorly on A57. By contrast, the proposed scheme is more consistent in its performance. To gain more insights into such behaviour of quality metrics, we perform additional analysis using the TID database. In our opinion, the variation in performance of quality metrics over the different databases is partly because of the distortion levels. For instance, A57 database mainly contains images with near-threshold distortions i.e. image quality degradation is just noticeable. On the other hand, databases like LIVE and IVC consist of images with supra-threshold distortions i.e. image quality degradation could be severe and more noticeable to the human eye. We conducted further tests to verify this. We observed the

performance of different metrics for the 4 distortion levels of the TID database. The first level (Level 1) denotes just noticeable or near threshold distortion while the fourth level (Level 4) indicates higher distortion levels.

With a total of 1700 distorted images and 4 distortion levels, there are 425 images for each distortion level. Table 5.3 presents the C_P values for the prediction performance of different metrics on the 4 distortion levels. The C_S and RMSE values are not presented here since they lead to similar conclusion as C_P values. We can see that MSVD, VIF, VSNR and PSNR-HVS-M perform relatively better for the fourth distortion level (i.e. higher amount of distortion) while they are relatively poor for lower distortion levels. Also we find that there is large variation in prediction accuracies for MSVD, VIF and PSNR-HVS-M as we go from Level 1 to Level 4. On the other hand, SSIM, VSNR and *Q* are more consistent for the 4 levels with *Q* being better than the two. Therefore, *Q*, in general, not only performs better for each distortion level but is also more stable and consistent for the 4 levels. We believe this to be a reason for the better performance of the proposed metric for all the databases. That is, it achieves a better trade-off for the performance on near-threshold and supra-threshold distortions.

5.3.2 Cross Database Validation

We further present the results for the cross-database evaluation in Table 5.1 where Q_{TID}, Q_{LIVE}, Q_{IVC} and $Q_{watermark}$ denote that training is done with TID, LIVE, IVC and watermarked image databases respectively while the remaining databases form the test sets. Since the training and testing sets come from different databases, the cross database evaluation helps to evaluate the robustness of the proposed scheme to untrained data. We can again see that the proposed scheme performs quite well with all the 3 test criteria (C_P,

C_S and RMSE). It is also worth pointing out that Q_{IVC} achieves good results for the TID database since in this case the training set size (185 images) is relatively smaller than the test set (1700 images). Similar comments can also be made for $Q_{watermark}$ where training set consists of 210 images.

As the last test in cross database evaluation, we test the performance of the proposed scheme for a video database. The trained system is used to predict the quality score of each individual frame and the overall quality score of the video is determined as the average of the scores all the frames in the video. The same procedure was also adopted for evaluating the other metrics. We present the results in Table 5.4. We can see that Q_{TID}, Q_{IVC}, Q_{LIVE} and $Q_{watermark}$ all perform better than the existing metrics under comparison. Note that the videos in this database have been distorted due to H.264/AVC compression which is obviously not presented in the image databases. Since the training is done with image databases, the good performance of the proposed metric is again indicative of its generalization ability to new visual/distortion content.

The better performance of the proposed metric for this video database is also important since H.264/AVC is a recent video coding standard which is fast gaining industry appreciation. Although VQA may also involve temporal factors for quality estimation, the aforesaid procedure of using the average of frame level quality as the overall video quality score is still a popular and widely used [17]. Further, similar to the previous chapter, we used the video databases primarily to evaluate the proposed metric's performance to untrained contents.

Table 5.5: Average execution time for different metrics (in sec.).

Metrics	SSIM	MSVD	VIF	VSNR	PSNR	PSNR-HVS-M	Proposed
Time (sec.)	**0.0454**	0.6036	3.4829	0.4452	**0.0037**	2.5586	**0.3268**

5.3.3 Metric Efficiency Evaluation

An important criterion to judge the performance of a visual quality metric is its efficiency in terms of computational time required. The practical utility of a metric will reduce significantly if it is slow and computationally expensive. In this section, we compare the efficiency (i.e., computational complexity) of different metrics. We measured the average execution time required per image in the A57 database (image resolution is 512 × 512) on a PC with 2.40 GHz Intel Core2 CPU and 2 GB of RAM. Table 5.5 shows the average time required per image (in seconds), with all the codes implemented in Matlab. We can see that the proposed metric takes less time than all the existing metrics except PSNR and SSIM. This is because the feature extraction stage in the proposed metric takes the advantage of the Fast FT (FFT) algorithm during the DFT computation. Note that DFT normally requires $O(N^2)$ operations to process N samples but for FFT this number is only $O(N \log(N))$. The proposed metric is therefore reasonably efficient in terms of execution time required (in addition to better prediction accuracy) and more suitable for real time applications.

5.3.4 Further Discussion

We have three points which deserve further discussion and are explained in what follows. The first point is regarding the use of multiple databases (throughout the thesis). It ensures that the proposed system is tested for its robustness to a wide variety of image

and distortion contents on which the proposed system is not trained. Besides, it also helps in more comprehensive metric testing since a metric performing well for one database may not do well on another. In addition, it facilitates the cross database evaluation which provides a strong and convincing demonstration of the proposed system's ability to predict the quality for untrained data. It may be mentioned here that for the cross database evaluation, we did not do any parameter optimization towards the test database. For instance consider $Q_{watermark}$. In this case, once we learn the model using all the images and associated subjective scores of the watermarked image database, we use the same model for testing LIVE, A57, TID, WIQ, IVC and EPFL (video database) databases. That is, we used the same kernel function namely RBF and the other parameters (i.e., radius of Gaussian function, the trade-off error parameter and regression tube width) were all kept constant when testing other image databases. Similar comments can be made for Q_{TID}, Q_{LIVE}, and Q_{IVC}. The performance improves further if we train a model specifically for each test database separately. It is also worth pointing out that the proposed metric is pretty robust to the different SVR parameters in that small changes in them do not cause large change in the prediction performance.

Secondly, as demonstrated the proposed scheme is more consistent and stable in its performance across multiple databases than the existing metrics. This highlights that the selected features based on the 2D mel-cepstrum are effective. In addition, they convey a clearer physical meaning. The exploitation of 2D mel-cepstral features for quality assessment is novel and interesting since originally mel-cepstrum analysis was formulated for speech/audio signals. Since audio and visual signals have certain similarity as natural signals, it is not surprising that a similar approach can be used for analyzing them. The theory of natural signal statistics [228] also confirms that natural

signals (including images and sounds) share statistical properties (for instance natural signals are highly structured). These features are also of interest for pattern recognition applications since they allow representing the spectra by points in a multidimensional vector space.

Lastly, the reader will recall that we used only the magnitude of the bin energy $G(m,n)$ in Eq. (5.5). Note that $G(m,n)$ will be a complex number in general which we denote as $Ae^{j\alpha}$ with magnitude A and phase α. The 2D mel-cepstrum computation involves the logarithm of $G(m,n)$, so we have $\log((G(m,n)) = \log(Ae^{j\alpha}) = \log(A) + j\alpha$. Now both A and α should be continuous functions for them to have a valid FT. However, since $\alpha \in [-\pi,\pi]$ we must first unwrap the phase so that it becomes continuous. The major problem is that unwrapping the phase in 2-D is difficult [152] due to two reasons. First, a typical image may contain thousands of individual phase wraps. Some of these wraps are genuine, while others may be false and are caused by the presence of noise and sometimes by the phase extraction algorithm itself. The process of differentiating between genuine and false phase wraps is extremely difficult and this adds complexity to the phase unwrapping problem. A second reason that complicates the phase unwrapping problem is its accumulative nature. The image is processed sequentially on a pixel-by-pixel basis. If a single genuine phase wrap between two neighboring pixels is missed due to noise, or a false wrap appears in the phase map, an error occurs in unwrapping both pixels. This kind of error then propagates throughout the rest of the image. In addition, phase unwrapping will be a computationally expensive step and potentially a major bottleneck in the use of the proposed metric for real-time applications.

Although the 2D mel-cepstrum representation is beneficial due to being perceptually relevant (as explained in Section 5.2.1), we cannot use the phase information for reasons

given above. However, phase has been known to convey signal information. There have been several studies examining the role played by phase and magnitude in images (see for instance [124]-[125]). The phase of FT corresponds to the relative locations of events such as lines and edges while magnitude determines the strength of such features. In addition, numerous studies have concluded that phase generally contains more image information and the so called phase dominance in images has been long established. The early study in [126] highlighted the importance of phase in image processing filters. The work of Oppenheim and Lim [127] demonstrated that exchanging the magnitude or phase spectrum of two images tends to produce a hybrid image more closely resembling the image that contributed the phase spectrum. This means that phase conveys more crucial information than magnitude. Further statistical evidence in favor of this has been presented in [128] where it has shown that random re-assignment of phase has severer effect on image quality as compared to random re-assignment of the magnitude. The image denoising method proposed in [129] also relies on preserving the perceptually important phase information in the signal. The study in [130] concludes that while both phase and magnitude convey information regarding the signal, it is the phase information that provides more significant details. The authors in [131] have justified that edges can be detected more efficiently at points of maximum phase congruency. A recent work reported in [150] employed Fourier analysis for the task of ranking data and it was concluded that the phase is much more important to matching the appearance of the data than the magnitude. In addition, psychophysical studies [132]-[133] also provide evidence in favour of the importance of phase to understanding scenes in images. These studies primarily examine the effect of phase and magnitude manipulations on the interpretability of images. The studies in [134]-[135] explore the relative importance of

spectral amplitude and phase errors on reconstructed images in terms of the expected MSE in the image. Ref. [136] investigates the human visual sensitivity to phase perturbations (namely phase quantization and randomization) by examining the global image statistics (skewness and kurtosis). In light of the importance of phase, we believe that it would be interesting to investigate its effectiveness for visual quality assessment (this is done in the next chapter).

5.4 Comparison with SVD based algorithm

In Chapter 4, we have proposed an SVD based metric. For comparison, we present the results (C_P values) for the methods based on 2D mel-cepstrum and SVD in Figure 5.5. Note that both employ SVR based feature pooling and so this comparison helps evaluate the prediction accuracy of the features used in the respective schemes. As can be seen, the method based on 2D mel-cepstrum is slightly better than the one based on SVD, and the more important point is regarding computational complexity. As pointed out previously, SVD based metric is computationally more expensive than some of the existing schemes. This is because of the fact that SVD is computationally more challenging especially for larger image blocks. Note that SVD of a $N \times N$ matrix has time complexity $O(N^3)$ i.e. increases cubically (exponent 3) with increasing N. On the other hand, feature extraction based on 2D mel-cepstrum is much more efficient with complexity $O(N \log(N))$ (due to the use of FFT) as mentioned in Section 5.3.3.

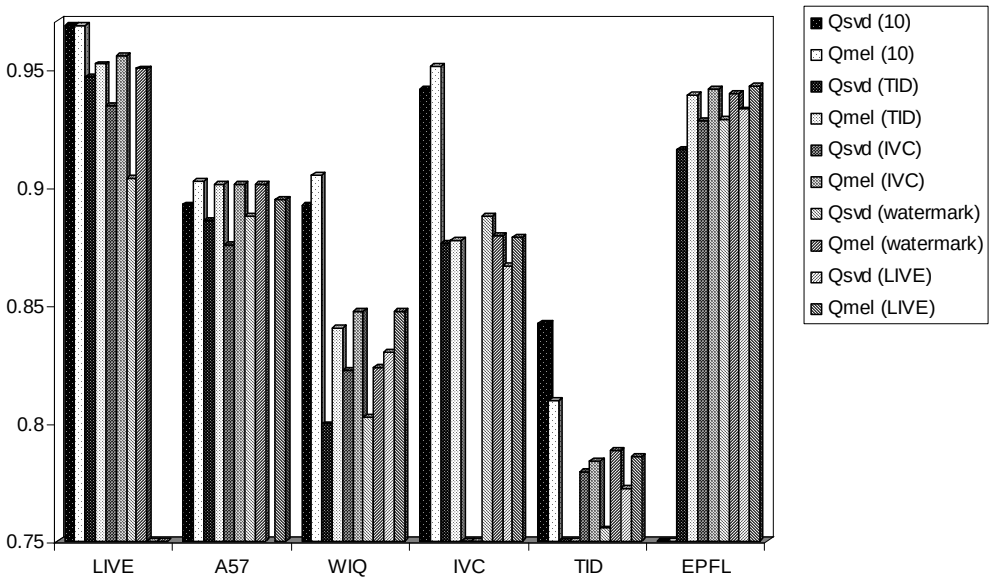


Figure 5.5: Performance comparison for SVD and 2D mel-cepstrum based methods (C_P values). The parenthesis in the legend with 10, TID, IVC and LIVE are respectively used to indicate 10 fold CV results and respective the training databases.

Regarding the computation time for an image of size 512 × 512, the 2D mel-cepstrum based method takes about 0.32 seconds (as indicated in Table 5.5) while SVD based scheme requires about 1.03 seconds (refer to Table 4.2). Therefore, the method based on 2D mel-cepstrum is faster and more efficient.

5.5 Concluding Remarks

In this chapter, we have explored the 2D mel-cepstrum features and provided analysis and justification for their use in visual quality evaluation. We employed machine learning based (SVR) feature pooling because of its advantages. A thorough experimental validation using 7 independent and publicly available image/video databases with diverse distortion types provides strong ground for the usefulness of the proposed metric. The experimental results confirm the effectiveness of the proposed

feature selection and pooling method towards more effective and consistent quality valuation. We have also compared the performance of the proposed metric with seven relevant existing metrics and shown that the proposed metric performs consistently better across all the databases.

However as pointed out in Section 5.3.4, phase is not readily available for the 2D mel-cepstrum representation. In order to overcome this problem and exploit phase (which is known to convey important signal information) for more effective visual quality evaluation, we develop a new algorithm in the next chapter. Further, the proposed method can be extended to video quality assessment in the following ways:

- Using the simple frame level averaged scores. This is probably the most straightforward method to extend 2D mel-cepstrum for video quality assessment. This of course has its drawbacks but incorporating better pooling schemes (like worst case pooling) will help over some of those limitations.
- We could use motion estimation based analysis to account for the temporal information. For instance assigning lower weighting in case of a large global motion. This is based on the idea that in case of large motion the distortion is not as apparent as it is in case of still images or slowly moving video. Another possible approach to incorporating motion information is to use the similarity between motion vectors as the temporal factor. One can also use 2D mel-cepstrum to estimate the quality between motion compensated blocks as the measure of temporal quality (as in [75]).

Chapter 6

Fourier Transform Based Scalable Visual Quality Measure

6.1 Introduction

As mentioned in Section 2.1.2, objective visual quality evaluation algorithms can be classified into 3 categories based on the amount of information used for predicting quality: (1) FR, (2) RR and (3) NR. FR algorithms are generally more accurate while NR methods although less accurate and usually distortion specific can be used when the reference image is not available. RR algorithms are essentially a trade-off between these two because only partial information of the reference image is required.

As pointed out in Section 5.3.4, it is difficult to obtain the phase using the 2D mel-cepstrum representation. In this chapter we present a new visual quality assessment algorithm based on the FT (which is generic transform and the phase is readily available for further processing). The base idea is to compare the phase and magnitude of the reference and distorted images to compute the quality score. However, it is well known that the HVS' sensitivity to different frequency components is not the same. We accommodate this fact via a simple yet effective strategy of non-uniform binning of the

frequency components. This process also leads to reduced space representation of the image thereby enabling RR prospects of the proposed scheme. We then employ linear regression to integrate the effects of the changes in phase and magnitude to evaluate the overall quality. Lastly, using the fact that phase usually conveys more information than magnitude, we use only phase for RR quality assessment. This provides the crucial advantage of further reduction in the amount of reference image information required. The proposed method is therefore further scalable for RR scenarios. Extensive experiments show that the proposed method is overall better than the relevant existing FR and RR algorithms. There is a graceful degradation in prediction performance as the amount of reference image information (for the RR case) is reduced thereby confirming its scalability prospects.

Because the phase can capture perceptually important features (such as edges and contours) it has been used in many image processing applications. For example, phase has been used in measuring image sharpness [137], image registration [138]-[140], palmprint recognition [141], visual saliency detection [142] and face recognition [143]-[144], to list a few. There also exist a few works (e.g. [145]-[146]) which have exploited phase for IQA. Even though several studies have pointed out (as mentioned here and also in Section 5.3.4 of the previous chapter) that phase plays a bigger role, the magnitude information cannot be completely ignored. This is obvious because both phase and magnitude are required for perfect image reconstruction. In this chapter, we propose a new scheme for visual quality assessment by utilizing the phase and magnitude of the FT. The proposed method is different from existing works based on phase in the following ways:

1. We take into account the human sensitivity to different frequency components: in

general, the HVS can tolerate more error in higher frequency components while the distortion in lower frequency components has a larger impact on the visual quality. We achieve this via binning of the higher frequency components leading to reduced space. This provides the additional advantage of scalability associated with the proposed scheme: only a fraction of the total amount of information is required from the reference image to determine the quality.

2. We employ a regression based method for combining the quality scores from phase and magnitude changes leading to more convincing integration of the two.
3. A thorough set of experimental results is presented which provide evidence in favour of the proposed scheme. To that end, we have used a total of 9 publicly available subjectively rated databases: 7 image databases (with a total of 3832 distorted images having diverse distortions) and 2 video databases (totally 228 distorted videos). We also compare the performance of the proposed scheme with several existing and well known FR quality assessment methods.
4. We also explore and demonstrate the scalability of the proposed method by using only the phase information. This helps in significantly reducing the amount of reference information needed and renders scalability of the proposed method.

The remaining sections are organized as follows. Section 6.2 introduces the relevant work pertaining to the phase and magnitude of the FT. We then describe the proposed scheme with proper analysis and reasoning. Extensive experimental validation and related analysis is then reported in Section 6.3 while concluding remarks are given in Section 6.4.

6.2 The Proposed Method Using Phase and Magnitude of 2D DFT

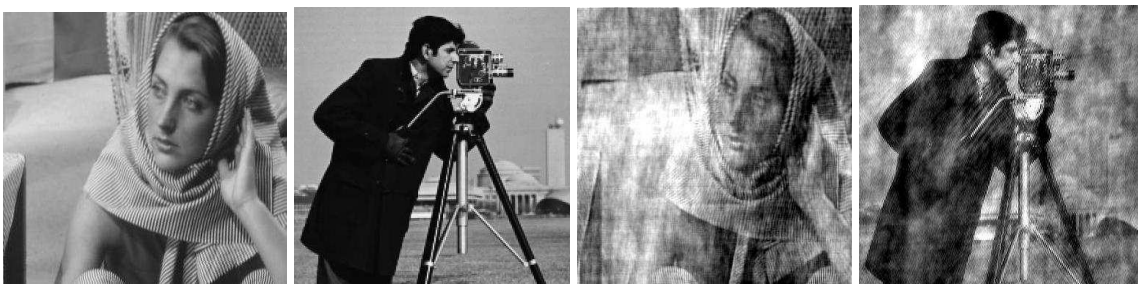
6.2.1 Phase and Magnitude characterization

In this section, we present visual examples to illustrate the roles that phase and magnitude play. First, we consider the 'Boat' image shown in Figure 6.1 (a) and obtain its phase and magnitude spectra. Similar to [137], we distorted the magnitude spectra by adding a random shift $\alpha.S$ where α is a constant and S is made of i.i.d. random variables uniformly distributed on $(-\pi,\pi)$. The reconstructed image from distorted magnitude and original phase is shown in Figure 6.1 (b). Next, we distorted the phase with the same energy in a similar way and the reconstructed image from original magnitude and distorted phase is shown in Figure 6.1 (c). As can be observed, for a same distortion, phase has a bigger impact on the image structure and hence its quality. However, it is important to note that magnitude distortion still provides some information regarding quality loss.

The second visual example is shown in Figure 6.2 which involves image reconstruction using interchanged phase and magnitude. The two original images are shown in Figure 6.2 (a) ('Tiffany') and (b) ('Cameraman'). By interchanging the phase and magnitude, we obtained the two images in (c) and (d) as {phase of (a) and magnitude of (b)} and {phase of (b) and magnitude of (a)}, respectively. It can be noticed that phase conveys more information regarding the perceived image content.

(a) (b) (c)

Figure 6.1: Effect of random phase and magnitude perturbations
(a) Original image, (b) image reconstructed with original phase and distorted magnitude and (c) image reconstructed with distorted phase and original magnitude.



(a) (b) (c) (d)

Figure 6.2: Interchanging phase and magnitudes in images
(a) Original 'Tiffany' image, (b) Original 'Cameraman' image, (c) image constructed from phase of (a) and magnitude of (b) and (d) image constructed from phase of (b) and magnitude of (a).

The Fourier phase determines the locations of perceptually-significant features such as edges, and has a bigger contribution than magnitude in determining the appearance.

It has been pointed out [142], [147] that the magnitude spectrum specifies how much of each sinusoidal component is presented and the phase information specifies where each of the sinusoidal components resides within the image. The authors in [142] used 1D signals and demonstrated that when the waveform is a positive or negative pulse, its

phase-only reconstruction contains the largest spikes at the jump edge of the input pulse. This is because many varying sinusoidal components locate there. On the other hand, when the input is a single sinusoidal component of constant frequency, there is no distinct spike in the reconstruction. Thus, phase of the signal carries information regarding edges and other salient parts. This is also true for a 2D signal (image/video frame), and due to this phase has been used [142], [148] to obtain the image saliency map and also in edge detection [149]. We conducted experiments in which we reconstructed the image using constant magnitude and original phase (and vice-versa i.e. original magnitude and a constant phase). Figure 6.3 shows three images and their reconstructed versions. The first row of Figure 6.3 shows (a) original, (b) blurred and (c) JPEG compressed images. The second row shows the images reconstructed from their respective phases but constant magnitude spectra, while the third row in Figure 6.3 shows the images reconstructed using their respective magnitude but a constant phase. As can be seen, the images in Figure 6.3 (d), (e) and (f) capture the most important features such as edges and contours. One can also notice the damage that is caused to the image *structure* due to blurring and JPEG compression. Therefore, phase similarity (or difference) between the reference and a distorted image is expected to give a reasonable estimate of *structural* degradation (provided that signal contents are properly discriminated as will be explained in Section 6.2.2). On the other hand, the images in Figure 6.3 (g), (h) and (i) convey less information although some changes can be noticed due to the blurring and JPEG distortions.

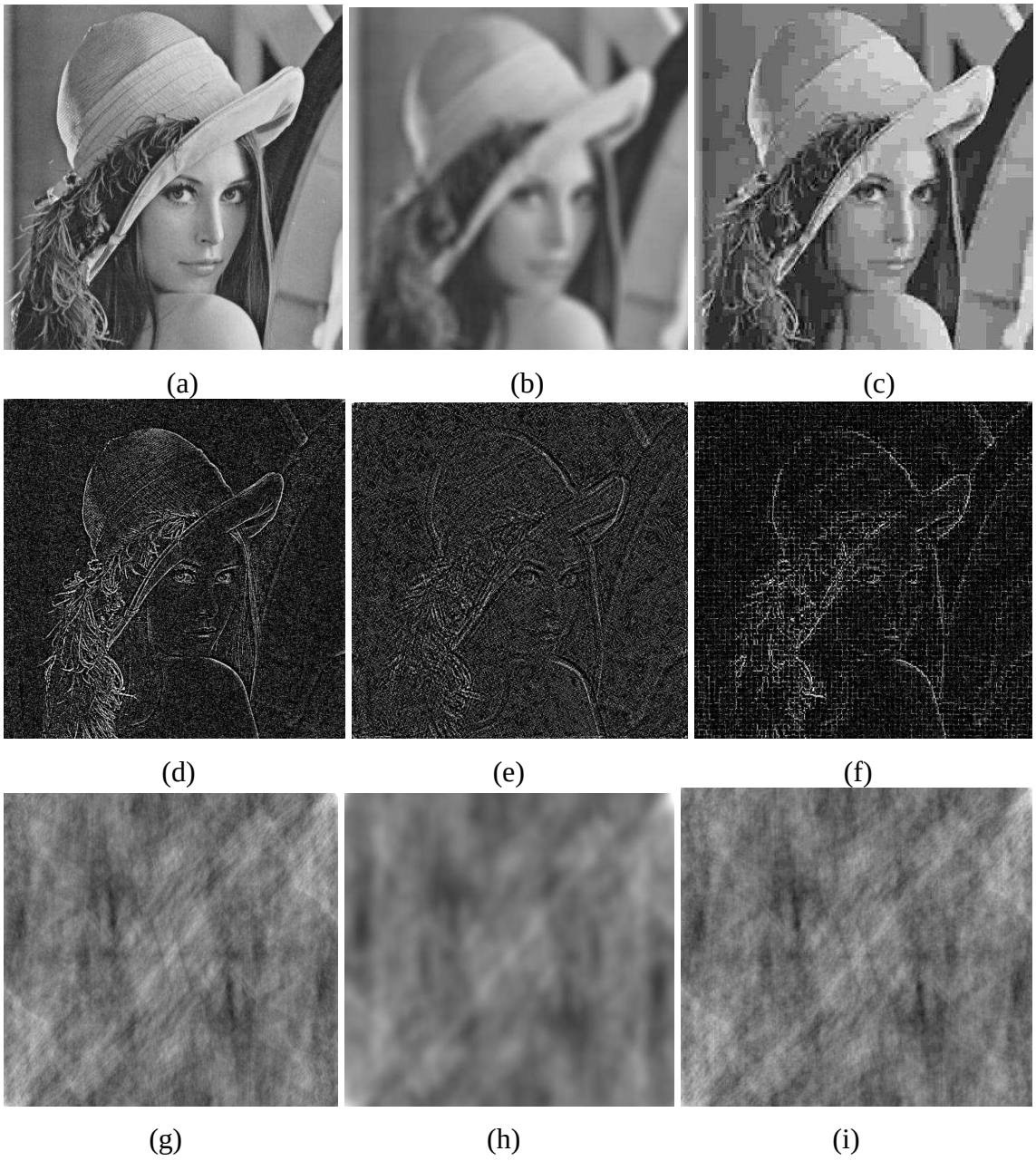


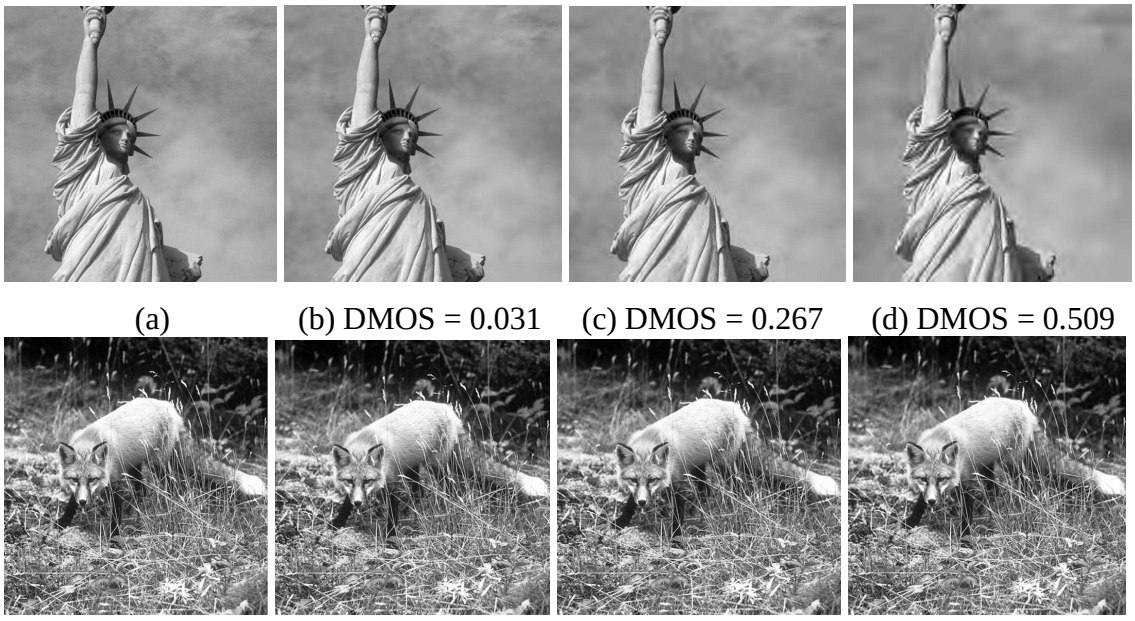
Figure 6.3: Image reconstruction with constant phase (or magnitude). (a) Original image, (b) Blurred image, (c) JPEG compressed image, (d) image constructed using constant magnitude and phase of (a), (e) image constructed using constant magnitude and phase of (b), (f) image constructed using constant magnitude and phase of (c), (g) image constructed using constant phase and magnitude of (a), (h) constructed using constant phase and magnitude of (b) and (f) constructed using constant phase and magnitude of (c).

6.2.2 Non-uniform binning of 2D DFT coefficients for visual quality assessment

Non-uniform binning of frequency coefficients has been explored previously [151] for face recognition. In this chapter, we provide analysis and justification for binning the frequency coefficients towards more accurate and efficient (in terms of RR prospects) visual quality assessment.

As mentioned in the introduction of this chapter, we propose the use of phase and magnitude to compute the image quality. A natural (and intuitive) way of determining the quality of distorted images (compared with the reference image) is to measure the similarity (or difference) between the phases and magnitude of the reference and distorted images. Conceptually, this would be similar to MSE (or PSNR) which directly computes the difference between the pixels of the reference and distorted images. However, like MSE, such an approach would be less effective because it does not account for the HVS' characteristics and signal contents. In other words, it fails to consider the unequal sensitivity of the HVS to distortions in different frequency components.

It is known that natural images are characterized by a fair amount of redundancy. A common characteristic of most images is that the neighboring pixels are correlated. Exploiting this, there have been image compression techniques aiming to reduce the number of bits needed to represent an image by removing the spatial and spectral redundancies. It can therefore be argued that for perceptual quality assessment, it would be more effective to focus on changes/distortion in perceptually important components. For example, it is well known that textured regions can usually tolerate more distortion (error) than smooth regions.



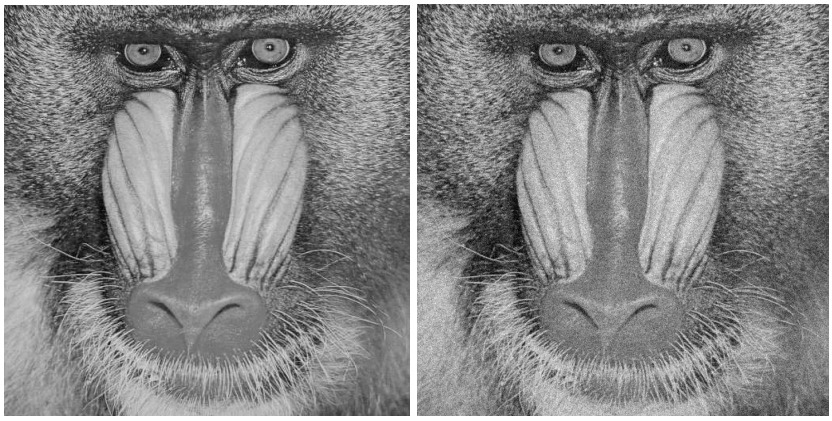
(a) (b) DMOS = 0.031 (c) DMOS = 0.267 (d) DMOS = 0.509

(e) (f) DMOS = 0.015 (g) DMOS = 0.056 (h) DMOS = 0.179

Figure 6.4: Effect of distortion on image with relatively more smooth areas and more textured areas (these images are from CSIQ image database).

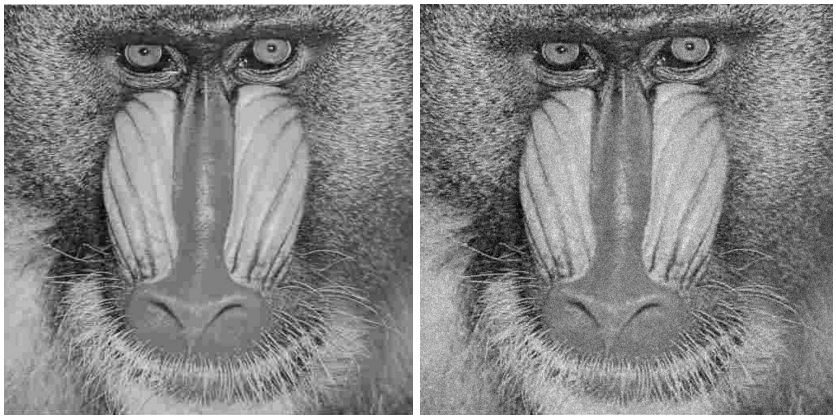
First Row: (a) Original image 'lady_liberty', (b) JPEG 2000 distorted image for the first distortion level, (c) JPEG 2000 distorted image for the second distortion level and (d) JPEG 2000 distorted image for the third distortion level. Second Row: (a) Original image 'foxy', (b) JPEG 2000 distorted image for the first distortion level, (c) JPEG 2000 distorted image for the second distortion level and (d) JPEG 2000 distorted image for the third distortion level. The respective subjective scores in the form of Difference MOSs (DMOS) have been indicated below each distorted image.

This is because the distortion in textured regions is usually masked (texture masking). Masking effect refers to the reduction of the visibility of image distortion due to the presence of the original content in the reference image. Stated differently, the JND in textured regions is higher than that in smoother areas of the image [27]. Psycho-visual experiments have also shown that the HVS has reduced sensitivity for patterns with high spatial frequencies and therefore their distortion/perturbation is less annoying.



(a) (b)

SSIM =0.8124, PSNR = 21.5090, Q_{Phase} = 0.9495, Q_{mag} = 0.9776



(c) (d)

SSIM =0.7948, PSNR = 23.1508, Q_{Phase} = 0.9511, Q_{mag} = 0.9801 SSIM = 0.6785, PSNR = 18, 5705, Q_{Phase} = 0.9292, Q_{mag} = 0.9629

Figure 6.5: Illustration of masking effect due to high texture.
(a) Original 'baboon' image, (b) Noisy image, (c) JPEG compressed image and (d) JPEG image in (c) distorted by the same amount of Gaussian noise that was used to obtain (b). Objective predictions from SSIM, PSNR and proposed method are also indicated below each image.

The reduced sensitivity of the HVS to higher frequencies has therefore been used in JPEG compression where the higher frequency signals are largely discarded using non-uniform quantization.

As an illustration, we have shown 4 images in the first row of Figure 6.4 (image (a) is the original image which is relatively smooth) and 4 images in the second row (image (e) is the original image which has more texture). These images have been taken from the

CSIQ database [118]. The two original images were compressed by JP2K technique to obtain the distorted images shown in (b), (c), (d) and (f), (g), (h) respectively. The compression level along each row increases from left to right and equally for both images. We have also indicated the respective subjective scores (in the form of DMOS with a lower value indicating better subjective quality) for the distorted images. As can be seen, for a same compression, the DMOS scores are higher for the images in the first row as compared to those in the second row. This means that the textured image can tolerate more distortion than the relatively smooth image.

Another example is shown in Figure 6.5. As can be observed, the 'Baboon' image is highly textured and the increased amount of distortion (or error) does not necessarily imply the same loss of perceived quality. This is because a large amount of distortion is masked because of texture and its visibility is reduced.

Based on the foregoing discussion and analysis, it is evident that for assessing visual quality, the unequal sensitivity of the HVS to distortions should be taken into consideration. We divide the spectrum into non-uniform bins such that the bin size is bigger for higher frequency and smaller at lower frequency. A representative diagram is shown in Figure 6.6 where each red dot represents a DFT coefficient. In this figure, the DC component at the centre is indicated by a bigger red dot and higher frequency components lie away from the centre as indicated. Next, we obtain the average of frequency coefficients in each bin. With the said procedure, we obtain a reduced space representation of the spectrum in which the higher frequency components are represented by the average of the components in each bin. This can also be taken as a special case of down sampling wherein the frequency components in each bin are represented by just one sample (the average).

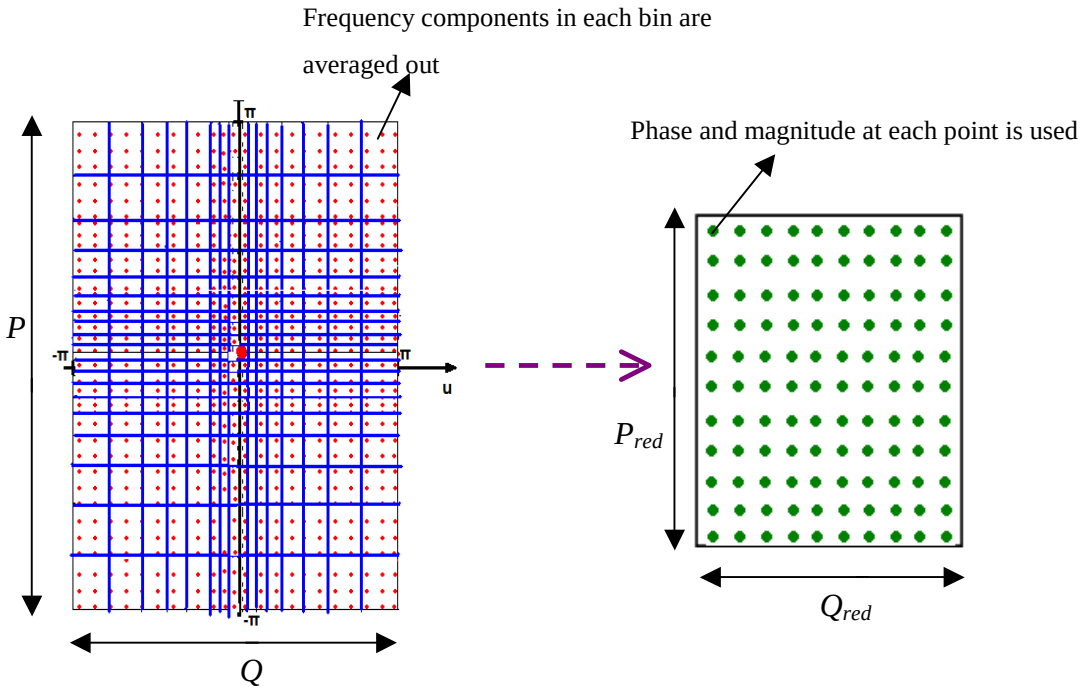


Figure 6.6: A representative diagram of the non-uniform binning of the DFT coefficients. Notice that the bin sizes are bigger for higher frequency components.

The smaller bin size for lower frequency means that we analyze them at a finer resolution. On the other hand, the higher frequency components are analyzed at coarser resolution because more components are averaged out due to larger bin size. As an illustration of the effectiveness of this, we have indicated in Figure 6.5 the objective quality scores from SSIM, PSNR and the proposed Q_{Phase} and Q_{mag} (defined later in Eqs. (6.7) and (6.8) respectively). Note that SSIM, Q_{Phase} and Q_{mag} predict scores in the range [0, 1] with 1 denoting best quality and 0 indicating worst quality. The reader will notice that the images shown in Figure 6.5 (b), (c) and (d) are of similar visual quality as the original image (a). Therefore, objective predictions should be close to 1 in case of SSIM, Q_{Phase} and Q_{mag} and a large number in case of PSNR. However, both PSNR and SSIM tend to overestimate the error. On the other hand, Q_{Phase} and Q_{mag} are better (both predict scores

closer to 1) because of higher emphasis on the distortion in lower frequency components. The details of the proposed method are given next.

The 2D DFT $Y(u,v)$ of the image $y(n_1,n_2)$ (size N by N) is defined as:

$$Y(u,v)=\frac{1}{N}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1}y(n_1,n_2)e^{-j2\pi\left(\frac{un_1+vn_2}{N}\right)} \tag{6.1}$$

where n_1 and n_2 denote the spatial coordinates, and u and v are the frequency coordinates. $Y(u,v)$ is in general a complex number consisting of the real and imaginary parts. Using Euler's formula we can express $Y(u,v)$ as

$$Y(u,v)=|Y(u,v)|\ e^{j\phi(u,v)} \tag{6.2}$$

where $|Y(u,v)|$ represents the magnitude and $\varphi(u,v)$ denotes the phase such that

$$|Y(u,v)|=\sqrt{(re(Y(u,v))^2+(im(Y(u,v))^2} \tag{6.3}$$

and

$$\phi(u,v)=arctan\left(\frac{im(Y(u,v))}{re(Y(u,v))}\right) \tag{6.4}$$

with $re(.)$ and $im(.)$ denoting real and imaginary parts respectively.

We now describe the procedure of determining the quality of a distorted image compared with the reference image. We first divide the image into non-overlapping blocks of size $P \times Q$. Next, we obtain the 2D DFT coefficients of each block. We then group the DFT coefficients via non-uniform binning as demonstrated in Figure 6.6. In this chapter, we assumed square blocks, i.e. $P = Q$. Finally, the phase and magnitude are extracted (for both reference and distorted images) from the reduced space representation and their similarity scores is computed.

Let $P_{ij}^{(ref)}$, $M_{ij}^{(ref)}$ respectively denote the phase and magnitude at the i^{th} point (as

illustrated in Figure 6.6 there will be totally $P_{red} \times Q_{red}$ such points) in j^{th} block of the reference image while $P_{ij}^{(dis)}$, $M_{ij}^{(dis)}$ denote that for the distorted image. The phase and magnitude similarity scores for j^{th} block are then obtained as

$$q_{phase}^{(j)} = \frac{1}{N_{red}} \sum_{i=1}^{N_{red}} \frac{2P_{ij}^{(ref)} P_{ij}^{(dis)} + C}{\left(P_{ij}^{(ref)}\right)^2 + \left(P_{ij}^{(dis)}\right)^2 + C} \tag{6.5}$$

$$q_{mag}^{(j)} = \frac{1}{N_{red}} \sum_{i=1}^{N_{red}} \frac{2M_{ij}^{(ref)} M_{ij}^{(dis)} + C}{\left(M_{ij}^{(ref)}\right)^2 + \left(M_{ij}^{(dis)}\right)^2 + C} \tag{6.6}$$

where C is a constant used to avoid division by zero and $N_{red} = P_{red} \times Q_{red}$. The phase (and magnitude) similarity score for the whole image is obtained by averaging the scores over all the image blocks (let N_{block} be the number of blocks) to obtain the two overall scores as

$$Q_{Phase} = \frac{1}{N_{block}} \sum_{j=1}^{N_{block}} q_{phase}^{(j)} \tag{6.7}$$

$$Q_{mag} = \frac{1}{N_{block}} \sum_{j=1}^{N_{block}} q_{magnitude}^{(j)} \tag{6.8}$$

Note that $0 \leq Q_{Phase}, Q_{mag} \leq 1$ with 0 indicating no similarity (worst quality) and 1 implying perfect similarity (highest quality).

As mentioned in the introduction of this chapter, a few existing works use the phase information directly and one such method has been proposed in [146]. We denote it as DPS (direct phase similarity) in the remaining sections. In this method, image quality is computed by measuring the Pearson correlation coefficient between the phase of the reference and distorted images. Therefore, DPS scores are in the range [0, 1] with 0 denoting worst quality and 1 denoting perfect quality. To demonstrate the effectiveness of the proposed method (denoted as $Q_{combined}$ and defined in Eq. (6.9)) in comparison to

DPS, we show 4 distorted images in Figure 6.7 which are taken from A57 database [119]. The subjective scores in the form of DMOS (smaller means higher image quality) have also been indicated below each image. As can be seen, DPS scores are not consistent with subjective opinions while the scores from proposed $Q_{combined}$ are closer to subjective viewing results. Note that $Q_{combined}$ scores are in the form of DMOS (in the range 0 to 100) due to training with a database with DMOS (refer to Section 6.2.4 for details). We have also indicated the PSNR values for these images for comparison. Both DPS and PSNR do not explicitly account for the fact that HVS' sensitivity to error in different frequency components is not the same. As opposed to this, the proposed method is more sensitive to error (distortion) in lower frequency. At the same time the error in higher frequency is not simply ignored but analyzed such that its overall impact is lower. It may be pointed out that Figure 6.7 is meant to provide a simple visual illustration and in Section 6.3 we will present more comprehensive experimental results to show the effectiveness of proposed scheme in comparison to DPS and other IQA algorithms.

6.2.3 Reduced-space representation of image

As pointed out earlier, with the procedure summarized in Figure 6.6, we obtain a reduced-space representation of the image. That is, an image block of $P \times Q$ pixels will be represented by ($2 \times P_{red} \times Q_{red}$) coefficients where $P_{red} < P$, $Q_{red} < Q$, and the factor of two is because we need both phase and magnitude at each point. We can further reduce the information required by using the fact that for real sequences the 2D DFT is symmetric. Because the image $y(n_1,n_2)$ is real valued, $Y(u,v)$ exhibits complex conjugate symmetry, i.e. $Y(u,v)=Y^*(-u,-v)$. Therefore, the same magnitude information will be repeated because $|Y(u,v)|=|Y(-u,-v)|$.

(a) DMOS = 0.5250
DPS = 0.7735, PSNR = 28.9830,
$Q_{combined}$ = 41.5661

(b) DMOS = 0.1080
DPS = 0.7572, PSNR = 29.0030,
$Q_{combined}$ = 32.4690

(c) DMOS = 0.1870
DPS = 0.6726, PSNR = 28.8700,
$Q_{combined}$ = 35.5580

(d) DMOS = 0.1370
DPS = 0.7044, PSNR = 29.0560,
$Q_{combined}$ = 35.0577

Figure 6.7: Visual quality prediction by DPS, PSNR and proposed method. The subjective scores, DPS, PSNR and proposed method's scores are indicated below each image (images and the corresponding subjective scores are from A57 database).

On the other hand, for phase we have $\phi(-u,-v)=-\phi(u,v)$, i.e., phase differs only by the sign and so discarding the phase with negative sign does not have any impact. By this we do not imply that the symmetric phase does not provide any information; merely, for our purpose in this chapter, it is not useful because we only need the similarity between the phases of reference and distorted images.

In this chapter we used a block size of 128, i.e., $P = Q = 128$ while $P_{red} = Q_{red} = 31$. This means that in this case we need only 11.73% of the information from the reference image to compute the quality. Due to exploiting the symmetry, we further reduce the information required to only about 6% of the information from the reference image to

compute Q_{Phase} and Q_{mag} .

6.2.4 Combining Q_{Phase} and Q_{mag} via linear regression

Q_{Phase} and Q_{mag} need to be integrated into an overall quality score. To that end, we employed linear regression to obtain overall quality score. Assuming that $Q_{combined}$ denotes the overall quality score, we can express the solution as

$$Q_{combined} = w_1\ Q_{Phase} + w_2\ Q_{mag} + b \tag{6.9}$$

where w_1 and w_2 are the respective weights for Q_{Phase} and Q_{mag} while b is the intercept (a constant). Because phase plays a more crucial role in determining the change in the image *structure*, it is expected to have a larger impact on the overall quality and so $|w_1|>|w_2|$. However, to determine the exact contribution (weights) of each term, it will be more convincing to obtain them via training instead of ad-hoc selection. Let $\{X_1, X_2, ...X_l\}$ and $\{y_1, y_2, ...y_l\}$ denote the training set. Here, each $X_i = \left[(Q_{Phase})_i, (Q_{magnitude})_i\right]$ represents the 2-dimensional vector consisting of the phase and magnitude similarity scores and each y_i is the associated subjective score (i.e. target value) for the i^{th} image. Given the training data $(X_1, y_1), ...(X_l, y_l)$, we find the weight vector $W = (w_1, w_2)$ and b by solving the following optimization problem

$$\min_{W,b} \sum_{i=1}^{l} \left(y_i - \left(W^T X_i + b\right)\right)^2$$

Therefore, $W^T X_i + b$ approximates the training data by minimizing the sum of squared errors. We used the A57 database for training and as a result of which $w_1 = -10.57$, $w_2 = -5.59$ and b = 16.14. As expected, Q_{Phase} has a larger impact (contribution) than Q_{mag}. It is also easy to see that with this set of w_1, w_2 and b, $Q_{combined}$ will be close to zero for the

perfect quality image (because $Q_{Phase} = Q_{mag} = 1$) and increase as image quality decreases. Hence $Q_{combined}$ will predict DMOS because the training database A57 provides DMOS as the subjective scores.

6.3 Experimental Results and Analysis

In this section, we present the experimental results to assess the prediction performance of the proposed scheme. Note that all the results in this chapter are for the luminance component of the image. We also include the results for PSNR, MSSIM[2] [153], VSNR [45], VIF [44], PSNR-HVS-M [154] and DPS [146] all of which are FR schemes.

6.3.1 Performance Assessment Criteria

As mentioned in Section 6.2.4, we used A57 database (please refer to the Appendix for database details) for training with the remaining image and video databases as test sets. Note that none of the images in A57 database appear in the remaining databases. For reporting the results for A57 database, we used WIQ database as the training set. Thus, training and test contents do not overlap and there is no parameter *optimization* towards any of the test databases. Furthermore, we also use two video databases to investigate the effectiveness of the "learned" relationship in predicting quality of individual video frame. In this chapter, the overall video quality is obtained by simply averaging out the quality scores over all the frames. As already stated in Chapters 4 and 5, we realize that simple averaging does not account for the temporal factor that has been shown to play a crucial role in VQA (we will tackle this aspect in Chapter 7). However, in this chapter, our primary aim is to examine the performance of the trained weights to new and untrained

[2] We include results only for MSSIM because it is usually better than the single scale SSIM.

contents.

Like Section 4.4.1, we employed a 5-parameter logistic mapping between the objective outputs and the subjective scores. The prediction performance is compared using C_P, C_S, Kendall rank correlation coefficient C_K and RMSE, between the subjective score and the objective prediction (after logistic transformation). A better quality metric has higher C_P, C_K, C_S and lower RMSE.

6.3.2 Performance Comparison

Table 6.1 presents the comparative results for the 7 image databases while the results for the 2 video databases are presented in Table 6.2. One can observe that the proposed method performs better (in many cases) or is very competitive when compared to the FR schemes. This is significant given that $Q_{combined}$ requires only about 6% (of the total number of pixels) of the reference information in contrast to the FR schemes (which need the complete reference information). One can also note from that the proposed scheme is more consistent in its performance across different databases.

In addition to the overall performance, the proposed method in general performed well for individual distortion types. As an example, we have presented the C_P values in Table 6.3 for some typical distortion types including JPEG, JPEG 2000, additive white noise and fast fading.

6.3.3 Scalability and Further Reduction in Required Reference Information

We can further reduce the required reference information with small loss in prediction accuracy.

Table 6.1: Performance comparison of the proposed method with FR methods for image databases.

Database	Criteria	PSNR	PSNR-HVS-M	MS-SSIM	VSNR	VIF	DPS	$Q_{combined}$	$Q^{(1)}$	$Q^{(2)}_{Phase}$	$Q^{(3)}_{Phase}$
LIVE	C_S	0.8756	0.9295	0.9513	0.9280	0.9632	0.9292	0.9563	0.9479	0.9454	0.9287
	C_K	0.6865	0.7659	0.8044	0.7625	0.8270	0.7571	0.8190	0.7992	0.7932	0.7664
	C_P	0.8723	0.9251	0.9409	0.9237	0.9598	0.9246	0.9537	0.9450	0.9423	0.9228
	RMSE	13.3597	10.3722	9.2593	10.4694	7.6670	10.4058	8.2193	8.9325	9.1485	10.5488
TID	C_S	0.5794	0.6128	0.8542	0.7049	0.7496	0.7059	0.8338	0.8210	0.7804	0.7847
	C_K	0.4210	0.4764	0.6568	0.5345	0.5863	0.5189	0.6425	0.6259	0.5869	0.5907
	C_P	0.5726	0.6051	0.8451	0.6823	0.8090	0.7549	0.8441	0.8302	0.8053	0.8023
	RMSE	1.1003	1.0685	0.7173	0.9810	0.7888	0.8801	0.7195	0.7482	0.7957	0.8010
Toyama	C_S	0.6132	0.8480	0.8874	0.8608	0.9077	0.9203	0.9148	0.9001	0.9029	0.8590
	C_K	0.4443	0.6568	0.7029	0.6745	0.7315	0.7541	0.7384	0.7171	0.7224	0.6696
	C_P	0.6353	0.8580	0.8922	0.8704	0.9138	0.9264	0.9184	0.9061	0.9084	0.8604
	RMSE	0.9664	0.6428	0.5652	0.6160	0.5084	0.4711	0.4951	0.5295	0.5213	0.6378
A57[3]	C_S	0.6189	0.8962	0.8414	0.9355	0.6223	0.4443	0.8937	0.8802	0.9181	0.8697
	C_K	0.4309	0.7261	0.6478	0.8031	0.4589	0.3148	0.7191	0.7051	0.7443	0.6939
	C_P	0.6347	0.8749	0.8575	0.9497	0.6157	0.4745	0.9147	0.9093	0.9294	0.9053
	RMSE	0.1899	0.1190	0.1264	0.0769	0.1937	0.2163	0.0993	0.1023	0.0907	0.1044
IVC	C_S	0.6884	0.8832	0.8980	0.7993	0.8964	0.8819	0.8943	0.8905	0.8960	0.7881
	C_K	0.5218	0.6935	0.7203	0.6053	0.7158	0.6853	0.7114	0.7033	0.7150	0.5823
	C_P	0.7196	0.8905	0.9108	0.8034	0.9028	0.8941	0.9046	0.9003	0.9046	0.7935
	RMSE	0.8460	0.5544	0.5029	0.7255	0.5239	0.5456	0.5192	0.5304	0.5190	0.7414
CSIQ	C_S	0.8005	0.8179	0.9133	0.8104	0.9195	0.7831	0.9237	0.9344	0.8082	0.8197
	C_K	0.5984	0.6430	0.7393	0.6237	0.7537	0.5951	0.7619	0.7773	0.6472	0.6247
	C_P	0.7998	0.8137	0.8990	0.7993	0.9277	0.8376	0.9171	0.9336	0.8815	0.8687
	RMSE	0.1576	0.1526	0.1150	0.1578	0.0980	0.1434	0.1047	0.0940	0.1240	0.1295
WIQ	C_S	0.6257	0.7261	0.7360	0.6558	0.6918	0.6631	0.8418	0.8360	0.8271	0.7518
	C_K	0.4626	0.5569	0.5645	0.4873	0.5246	0.5069	0.6519	0.6500	0.6386	0.5575
	C_P	0.7549	0.7632	0.7761	0.7625	0.7333	0.7352	0.8547	0.8511	0.8481	0.7766
	RMSE	15.0235	14.8022	14.4442	14.8199	15.5734	15.5267	11.8914	12.0274	12.1378	14.4301

[3]Results are reported with WIQ as the training database.

Table 6.2: Performance comparison of the proposed method with FR methods for video databases.

Database	Criteria	PSNR	PSNR-HVS-M	MS-SSIM	VSNR	VIF	DPS	$Q_{combined}$	$Q^{(1)}$	$Q^{(2)}_{Phase}$	$Q^{(3)}_{Phase}$
LIVE video	C_S	0.5431	0.6889	0.7389	0.6710	0.5662	0.3654	0.7481	0.7487	0.7393	0.7397
	C_K	0.3818	0.5179	0.5579	0.4977	0.3948	0.2561	0.5561	0.5581	0.5484	0.5492
	C_P	0.5583	0.6947	0.7447	0.6878	0.5875	0.4379	0.7619	0.7623	0.7611	0.7590
	RMSE	9.1072	7.9262	7.3262	7.9687	8.8833	9.8689	7.1104	7.0844	7.1568	7.2270
EPFL video	C_S	0.6869	0.8760	0.9220	0.8631	0.6866	0.7206	0.9301	0.9268	0.9187	0.9117
	C_K	0.5058	0.6754	0.7642	0.6757	0.5178	0.5385	0.7749	0.7669	0.7590	0.7349
	C_P	0.6907	0.8865	0.9499	0.8890	0.7681	0.7224	0.9438	0.9422	0.9356	0.9289
	RMSE	0.9753	0.6240	0.4216	0.6176	0.8636	0.9326	0.4458	0.4520	0.4711	0.4995

To that end, we first obtain the averaged phase and magnitude over all the image blocks as

$$P_i^{(ref)} = \frac{1}{N_{block}} \sum_{j=1}^{N_{block}} P_{ij}^{(ref)} \tag{6.10}$$

$$P_i^{(dis)} = \frac{1}{N_{block}} \sum_{j=1}^{N_{block}} P_{ij}^{(dis)} \tag{6.11}$$

$$M_i^{(ref)} = \frac{1}{N_{block}} \sum_{j=1}^{N_{block}} M_{ij}^{(ref)} \tag{6.12}$$

$$M_i^{(dis)} = \frac{1}{N_{block}} \sum_{j=1}^{N_{block}} M_{ij}^{(dis)} \tag{6.13}$$

We then calculate phase and magnitude similarities $Q^{(1)}_{Phase}$ and $Q^{(1)}_{mag}$ as

$$Q_{phase}^{(1)} = \frac{1}{N_{red}} \sum_{i=1}^{N_{red}} \frac{2P_i^{(ref)} P_i^{(dis)} + C}{\left(P_i^{(ref)}\right)^2 + \left(P_i^{(dis)}\right)^2 + C} \tag{6.14}$$

$$Q_{mag}^{(1)} = \frac{1}{N_{red}} \sum_{i=1}^{N_{red}} \frac{2M_i^{(ref)} M_i^{(dis)} + C}{\left(M_i^{(ref)}\right)^2 + \left(M_i^{(dis)}\right)^2 + C} \tag{6.15}$$

The overall quality $Q^{(1)}$ is then determined via linear regression based combination of $Q_{Phase}^{(1)}$ and $Q_{mag}^{(1)}$. Note that $Q^{(1)}$ is different from $Q_{combined}$ which uses phase and magnitude similarity between the individual image blocks. Expectedly, for $Q^{(1)}$ there will be some loss of prediction accuracy because of the averaging indicated by Eqs. (6.10) ~ (6. 13) but importantly there is a further reduction in the amount of reference information. For instance in the TID database, the image resolution is 512 × 384. In this case, the required reference information will be only about $\frac{1}{200}$ of the image size, and this is quite a significant reduction. The experimental results for $Q^{(1)}$ have been presented in Tables 6.1, 6.2, 6.3 and 6.5. One can notice that $Q^{(1)}$ performs quite well and is very competitive with FR schemes.

We have already mentioned that phase information is generally more crucial than magnitude. To verify this further, as an example we present the individual results for $Q_{Phase}^{(1)}$, $Q_{mag}^{(1)}$ and $Q^{(1)}$ separately in Table 6.5 (RMSE is omitted as it leads to similar conclusions as from other criteria). One can see that $Q_{Phase}^{(1)}$ gives higher correlation with the subjective scores across all the databases. Nevertheless, we note that $Q_{mag}^{(1)}$ also plays a role. It is therefore not surprising that $Q^{(1)}$ (which is a linear combination of $Q_{Phase}^{(1)}$ and $Q_{mag}^{(1)}$) achieves the best results for each database.

Table 6.3: Performance comparison for typical distortion types.

Distortion Type	Database	PSNR	PSNR-HVS-M	MS-SSIM	VSNR	VIF	DPS	$Q_{combined}$	$Q^{(1)}$	$Q^{(2)}_{Phase}$	$Q^{(3)}_{Phase}$
JPEG	LIVE	0.8897	0.9485	0.9812	0.9735	0.9859	0.9742	0.9773	0.9704	0.9690	0.9559
	TID	0.8703	0.9720	0.9607	0.9379	0.9547	0.9308	0.9469	0.9282	0.9231	0.9308
	CSIQ	0.8788	0.9576	0.9815	0.9487	0.9882	0.9695	0.9771	0.9704	0.9686	0.9644
JPEG 2000	LIVE	0.8997	0.9200	0.9706	0.9641	0.9760	0.9583	0.9637	0.9546	0.9478	0.9321
	TID	0.8672	0.9669	0.9753	0.9531	0.9730	0.9629	0.9700	0.9604	0.9600	0.9510
	CSIQ	0.9463	0.9680	0.9785	0.9561	0.9776	0.9618	0.9722	0.9654	0.9627	0.9502
Blur	LIVE	0.7835	0.8869	0.9591	0.9369	0.9740	0.9412	0.9737	0.9627	0.9560	0.9310
	TID	0.8736	0.9143	0.9512	0.9277	0.9401	0.8857	0.9413	0.9229	0.9149	0.8968
	CSIQ	0.9081	0.9553	0.9669	0.9342	0.9717	0.9427	0.9728	0.9678	0.9606	0.9510
Additive white noise	LIVE	0.9857	0.9865	0.9725	0.9816	0.9841	0.9757	0.9847	0.9710	0.9692	0.9563
	TID	0.9341	0.9363	0.8021	0.7577	0.8725	0.6750	0.7144	0.6510	0.6827	0.5944
	CSIQ	0.8978	0.9433	0.9465	0.9260	0.9606	0.8703	0.9212	0.9004	0.9101	0.8720
Fastfading	LIVE	0.8897	0.9093	0.9284	0.9055	0.9613	0.9488	0.9431	0.9461	0.9380	0.9087
	TID	0.8536	0.9257	0.8386	0.7797	0.8372	0.7359	0.8369	0.7769	0.7796	0.7690

Table 6.4: Comparison of C_P values achieved by phase and magnitude.

Database/ Algorithm	LIVE	TID	Toyama	IVC	CSIQ
$Q^{(2)}_{mag}$	0.8810	0.6893	0.8831	0.8469	0.7598
$Q^{(2)}_{Phase}$	0.9423	0.8053	0.9084	0.9046	0.8815

Table 6.5: Results for phase and magnitude scores separately.

Criteria/	C_P			C_S			C_K		
Database	$Q^{(1)}_{Phase}$	$Q^{(1)}_{magnitude}$	$Q^{(1)}$	$Q^{(1)}_{Phase}$	$Q^{(1)}_{magnitude}$	$Q^{(1)}$	$Q^{(1)}_{Phase}$	$Q^{(1)}_{magnitude}$	$Q^{(1)}$
LIVE	0.9413	0.8803	0.9450	0.9475	0.8798	0.9479	0.7980	0.6998	0.7992
TID	0.8135	0.6853	0.8302	0.7883	0.7050	0.8210	0.5960	0.5384	0.6259
Toyama	0.8992	0.8789	0.9061	0.8874	0.8748	0.9001	0.7011	0.6893	0.7171
A57	0.8996	0.7470	0.9093	0.8840	0.7393	0.8802	0.7051	0.5694	0.7051
IVC	0.8927	0.8389	0.9003	0.8820	0.8315	0.8905	0.6955	0.6342	0.7033
CSIQ	0.8811	0.7549	0.9336	0.8067	0.7686	0.9344	0.6480	0.5894	0.7773
WIQ	0.8470	0.8342	0.8511	0.8284	0.8179	0.8360	0.6418	0.6297	0.6500

Table 6.6: Performance comparison of the proposed method Ref. [155] for LIVE database.

Criteria	Algorithm	All data	JP2(1)	JP2(2)	JPG(1)	JPG(2)	Noise	Blur	Fastfading
C_P	DNT [155]	0.8930	0.9115	0.9422	0.8501	0.9354	0.9401	0.8773	0.9243
	$Q^{(4)}$	0.9009	0.9031	0.9362	0.8850	0.9623	0.9512	0.8931	0.8897
C_S	DNT [155]	0.9093	0.9081	0.9239	0.8389	0.8734	0.9316	0.8608	0.9237
	$Q^{(4)}$	0.9031	0.9140	0.9225	0.8915	0.8831	0.9435	0.8858	0.8888

Table 6.7: Performance comparison of the proposed method with RR SSIM [156].

Criteria	Algorithm	LIVE	TID	Toyama	IVC	A57	CSIQ
C_P	RR SSIM [156]	0.9194	0.7231	0.8051	0.8177	0.7044	0.8426
	$Q^{(5)}$	0.8968	0.7682	0.8134	0.7400	0.8036	0.8576
C_S	RR SSIM [156]	0.9129	0.7210	0.8003	0.8154	0.7301	0.8527
	$Q^{(5)}$	0.9073	0.7547	0.8067	0.7356	0.7973	0.7917
C_K	RR SSIM [156]	0.7349	0.5236	0.6090	0.6164	0.5345	0.6540
	$Q^{(5)}$	0.7334	0.5611	0.6108	0.5355	0.6198	0.6211
RMSE	RR SSIM [156]	11.3026	0.9270	0.7423	0.7014	0.1744	0.1413
	$Q^{(5)}$	12.0863	0.8592	0.7279	0.8195	0.1456	0.1345

Even though the improvement in some cases (over $Q_{Phase}^{(1)}$) is small, the consistency in improvement for all the databases indicates that both play a role in overall quality score determination.

As another example of reduction in the required number of coefficients from the reference image, we develop another algorithm following the same procedure as outlined for obtaining $Q_{Phase}^{(1)}$ and $Q_{mag}^{(1)}$. The only difference is that for this case we use P_{red} = 31 and Q_{red} = 25 (instead of P_{red} = 31 and Q_{red} = 31). Let $Q_{Phase}^{(2)}$ and $Q_{mag}^{(2)}$ respectively denote the phase and magnitude similarities for this case. Note that we need only 400 phase coefficients from the reference image for computing $Q_{Phase}^{(2)}$ (similarly we require 400

magnitude coefficients from the reference image to calculate $Q_{mag}^{(2)}$). We have mentioned in the introduction that the phase conveys more information. To verify that Table 6.4 indicates the correlation values (only C_P values are shown) achieved by $Q_{Phase}^{(2)}$ and $Q_{mag}^{(2)}$ on the 5 biggest image databases. As expected, $Q_{Phase}^{(2)}$ performs better than $Q_{mag}^{(2)}$ and can alone be used as a quality estimator for want for information reduction from the reference image. Thus, $Q_{Phase}^{(2)}$ is effective for reducing the reference information on one hand and achieving reasonably high prediction accuracy on the other. The prediction accuracy of $Q_{Phase}^{(2)}$ is also reported in Tables 6.1~ 6.3. As expected, it is slightly worse than Q and $Q^{(1)}$ but still achieves reasonably good overall performance in spite of the fact that it needs only 400 coefficients from the reference image. For image with size 512 × 384, this amounts to using only about $\frac{1}{490}$ of the total reference information. This is a significant reduction in reference information requirement.

To further demonstrate the effectiveness of the reduced-space representation and its potential for scalability, we used $P_{red} = Q_{red} = 15$. Following the similar procedure as outlined for $Q_{Phase}^{(1)}$, we arrive at $Q_{Phase}^{(3)}$. Note that we again use only the phase information (i.e. 120 phase coefficients and this is about $\frac{1}{1640}$ of the image size). The results for $Q_{Phase}^{(3)}$ are also given in Tables 6.1~ 6.3. While $Q_{Phase}^{(3)}$ gives lower correlations as compared to other schemes including $Q_{combined}$, the performance drop is within a reasonable range. Of course the most crucial advantage of $Q_{Phase}^{(3)}$ is with regards to its requirement of the reference information. The performance of $Q^{(1)}, Q_{Phase}^{(2)}, Q_{Phase}^{(3)}$ on individual distortion types presented in Table 6.3 again indicates the scalability in the proposed method i.e. the degradation in prediction performance is graceful with reduction in reference information.

Finally we compare the performance of the proposed scheme with two recent RR schemes which we denote as DNT [155] (it is based on divisive normalization transform) and RR SSIM [156]. DNT and RR SSIM respectively require 48 and 36 coefficients from the reference image. We first use $P_{red} = Q_{red} = 15$. Next we use the average (or sum) of the coefficients in every 2 × 2 window. This means we denote the 4 coefficients in every 2 × 2 window by a single sample thus reducing the number of coefficients further. Obviously the averaging in 2 × 2 window will result in loss of prediction accuracy but this is done only to make the required number of coefficients in our scheme the same as those in DNT and RR SSIM (so that the comparison is fair). We use the symbols $Q^{(4)}$ and $Q^{(5)}$ to denote the proposed the scheme requiring 48 and 36 coefficients respectively (we use only the phase). Note that both $Q^{(4)}$ and $Q^{(5)}$ use only the phase information. The prediction performance of $Q^{(4)}$ and DNT for LIVE image database are presented in Table 6.6 (C_P and C_S values are presented). It may be pointed out that DNT requires training and its authors have reported the experimental results for two training cases: (a) training with LIVE database, (b) training with A57 database. For fair comparison with the proposed scheme, we have included the results derived from [155] with A57 database as the training set and LIVE image database as the test set. We have also presented the results for the individual distortion types present in the LIVE image database. We find that $Q^{(4)}$ which requires no training performs well and is overall better. The results for $Q^{(5)}$ and RR SSIM are presented in Table 6.7. For RR SSIM, we have reported the results as provided by its authors for 6 image databases. We can see that $Q^{(5)}$ performs better than RR SSIM for A57, CSIQ and TID databases and achieves competitive performance on LIVE and Toyama databases. It is also fair to mention here that RR SSIM also

employs training (which was done using images from LIVE image database) for finding optimal value of the slope parameter (we refer the reader to [156] for details). On the other hand, $Q^{(5)}$ (also, $Q_{Phase}^{(2)}$, $Q_{Phase}^{(3)}$ and $Q^{(4)}$) do not require any training because all of them do not use any magnitude information and hence there is no regression required.

In summary, we have presented 6 results namely $Q_{combined}$, $Q^{(1)}$, $Q_{Phase}^{(2)}$, $Q_{Phase}^{(3)}$, $Q^{(4)}$ and $Q^{(5)}$ which respectively require approximately $\frac{1}{17}$, $\frac{1}{200}$, $\frac{1}{490}$, $\frac{1}{1640}$, $\frac{1}{4096}$ and $\frac{1}{5460}$ of the reference information (for image resolution of 512 × 384). These algorithms perform well and are usually better or very competitive with FR schemes (and the two RR schemes). Importantly, the degradation in prediction performance is graceful with the reduction in reference information across all the databases. This enables scalability of the proposed method which is a crucial advantage. The good overall prediction performance on the 9 subjectively rated databases is also indicative of the robustness to diverse image and distortion contents.

6.3.4 Further Discussion

We have shown the effectiveness of the proposed method with regards to its prediction accuracy and scalability. As stated before, these are achieved as a result of accounting for the unequal sensitivity of the HVS to changes/distortions in different frequency components. To examine the impact of unequal emphasis on different frequency components as done in the proposed method, we have also presented the results for DPS in Tables 6.1 and 6.2. The following observations can be made from Tables 6.1 and 6.2:

1. As already mentioned, we can regard DPS as similar to PSNR (or MSE) because it uses each phase point for computing image quality while PSNR uses each pixel. However, DPS performs better than PSNR for most databases. This suggests that

phase conveys more precise information regarding *structural* changes than pixel.

2. The proposed $Q_{combined}$, $Q^{(1)}$, $Q_{Phase}^{(2)}$, $Q_{Phase}^{(3)}$ are overall better and more consistent than DPS across databases. As mentioned before, DPS is just the DPS measure, so the results clearly demonstrate the positive impact of using the reduced-space representation which leads to objective predictions that are better aligned with HVS' perception. It also confirms that discrimination of signal contents is an important aspect towards more accurate quality prediction.
3. A closer look at Tables 6.1 and 6.2 reveals that DPS actually performs quite well for LIVE, Toyama and IVC image databases while its performance is relatively poor on TID, A57, CSIQ, WIQ and the two video databases. This can be explained by considering the distortion levels in the databases. In LIVE, Toyama and IVC image databases, the distortion levels are relatively higher (i.e. suprathreshold) and more clearly visible. As a result of higher amounts of distortion, any change in the visual signal usually corresponds to a similar magnitude of the reduction in visual quality and hence the prediction accuracy of DPS is reasonable. In contrast to this, the distortion levels in databases such as TID, A57, CSIQ and WIQ are lower and many images are with near-threshold distortions (just noticeable). In this case, the change in the signal due to the distortion may not necessarily imply the same loss of visual quality (for example, as shown in Figure 6.5, the effect of distortion is masked). Hence DPS is overall less effective while the proposed method tackles this much better as already explained.

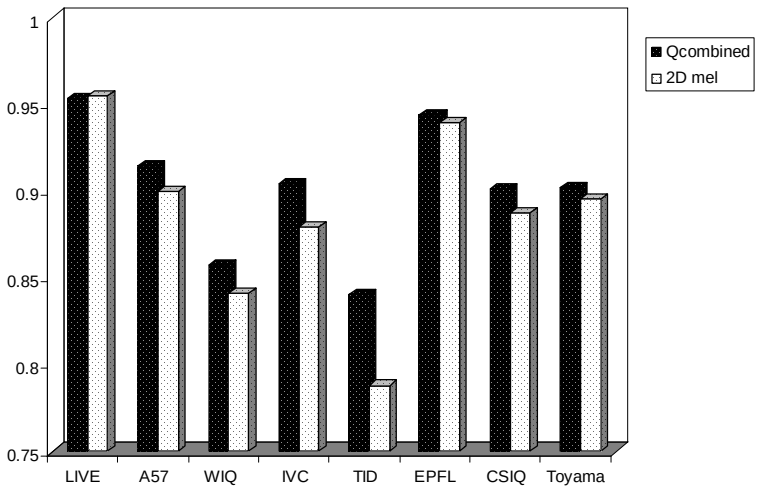


Figure 6.8: Performance comparison of 2D mel-cepstrum based method and $Q_{combined}$

In order to compare the prediction performance of 2D mel-cepstrum based method (proposed in the previous chapter) which does not use phase (for reasons discussed in Section 5.3.4) and the one proposed in this chapter, we present the C_P values for 8 databases in Figure 6.8. Note that we used the watermarked image database for training both the metrics. One can notice that $Q_{combined}$ gives a better overall performance which confirms that phase indeed is more effective.

The reader will also recall that we used a block size of 128, i.e., $P = Q = 128$. We then used different values of P_{red} and Q_{red} to obtain the reduced-space representation and thus obtained a group of algorithms which are suitable for reduced-reference scenarios. We also experimented with smaller block sizes. It was found that performance usually degraded with smaller block sizes and there are two possible reasons for this observation. Firstly, when smaller blocks are employed, they are assumed to be independent which may not always be true. A more global Fourier analysis, on the other hand, can tackle the interaction/dependencies between the blocks better. Secondly, with decreasing block size,

the number of blocks will obviously increase. It is quite possible that in such case the useful information about change in quality may be suppressed due to averaging over a large number of blocks. Further, we found that using overlapping blocks lead to similar prediction performances but with increased metric execution time.

6.4 Concluding Remarks

Phase has been known to convey more useful information (as compared to the magnitude) regarding important features like edges or contours. In this chapter, we first employed the phase and magnitude together as a comprehensive way to compute visual quality. We obtained an effective reduced space representation of the image (or video frame) by non-uniform binning of the high frequency components. This is based on the fact that the human eye can tolerate more error (distortion) in high frequencies (such as texture) and error in smooth (low frequency) area is more annoying. The proposed method can achieve better performance than many FR schemes in spite of using much less reference information. In addition, since phase is more important with regards to image *structure*, we further explored the scalability of the proposed method by using only the phase of the reduced-space representation. A thorough experimental verification of the effectiveness of the proposed method was done using 9 publicly available image and video databases. We presented the experimental results for the 6 algorithms developed such that they require decreasing amount of reference information. Each of the 6 algorithms performs well considering the reduced amount of reference information.

Chapter 7

Low-Complexity Video Quality Assessment Using Temporal Quality Variations

7.1 Introduction

Objective VQA is a challenging problem and there are three important issues that arise: (1) the temporal factors apart from the spatial ones also need to be considered, (2) the contribution of each factor (spatial and temporal) and their interaction to the overall video quality need to be determined, and (3) the computational complexity of the resultant method. In this chapter, we seek to tackle the first issue by utilizing the worst case pooling strategy and the variations of spatial quality along the temporal axis with proper analysis and justification. The second issue is addressed by the use of machine learning; as emphasized in the thesis, we believe this to be more convincing since the relationship between the factors and the overall quality is derived via training with substantial ground truth (i.e. subjective scores). Experiments conducted using publicly available video databases show the effectiveness of the proposed FR algorithm in

comparison to the relevant existing VQA schemes. Similar to Chapters 4 ~ 6 of this thesis, focus has again been placed on demonstrating the robustness of the proposed method to new and untrained data. To that end, extensive cross-database tests have been carried out to provide a proper perspective of the performance of proposed scheme as compared to other VQA methods.

The third issue regarding the computational costs also plays a key role in determining the feasibility of a VQA scheme for practical deployment given the large amount of data that needs to be processed/analyzed in real-time. A limitation of many existing VQA algorithms is their higher computational complexity. In contrast, the proposed scheme is more efficient due to its low complexity (as further explained in Section 7.3.3).

The remainder of this chapter is organized as follows. Section 7.2 describes the details of the proposed method. We outline the process of calculating the spatial and the temporal scores. We also discuss their combination into an overall score via machine learning. In Section 7.3 we present the experimental results using video sequences from three publicly available databases. We use a total of 260 video sequences encompassing a wide variety of vide contents and distortion types. Finally, Section 7.4 draws conclusions.

7.2 The Proposed VQA Algorithm

We can consider objective VQA as a two stage process: (a) computing the spatial and temporal factors, (b) pooling the two factors into an overall quality score. A block diagram of the proposed method is shown in Figure 7.1.

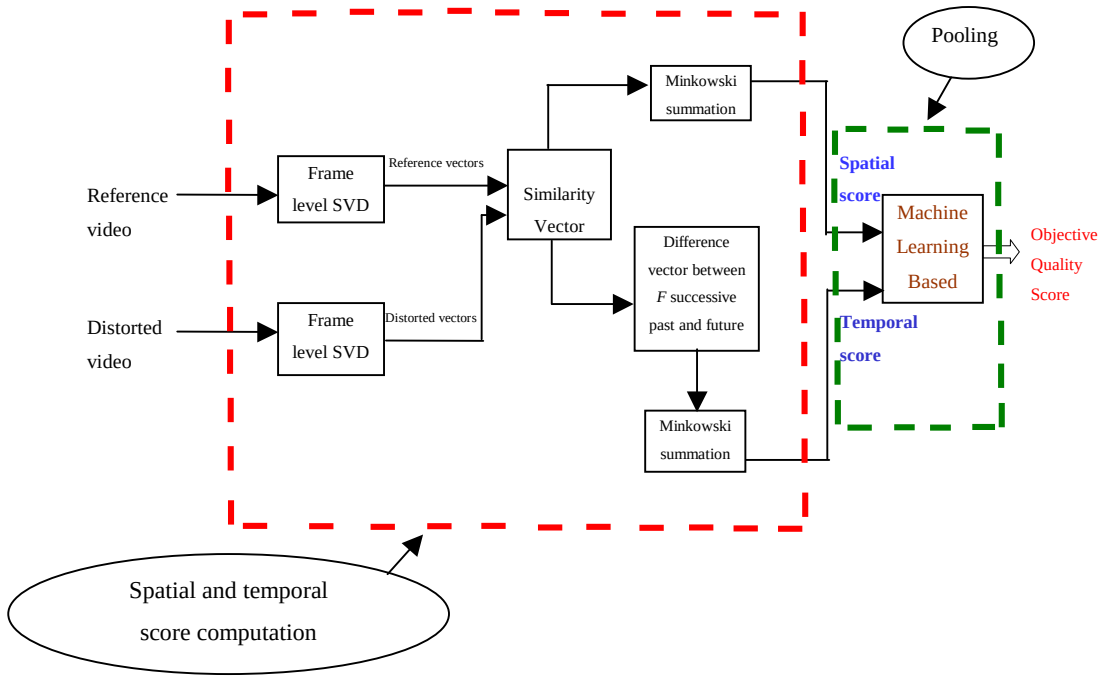


Figure 7.1: Block diagram of the proposed VQA scheme

The first block namely spatial and temporal score computation uses the SVD based metric (presented in Chapter 3) for computing the spatial scores. The temporal scores are computed from the difference of features between the video frames. Finally, the two scores are combined via SVR. Further details about the proposed method are described in subsequent sections.

7.2.1 Spatial Quality Measure

We use the SVD based method presented in Chapter 3 for the assessment of spatial quality. The reader will recall that in this method, the dot product between singular vectors of the reference and distorted images is utilized for quality assessment. We chose the SVD based method due to 2 reasons: (a) it gives reasonably good accuracy in predicting the spatial quality, (b) we obtain a low dimension (as compared to frame size) feature vector (which characterizes the quality) for each video frame which is used to

compute the quality variations in time as elaborated in Section 7.2.2.

We now briefly describe the procedure to compute the spatial quality of each video frame. Following similar notations and procedure as in Chapter 3, we write the SVD of a video frame $\boldsymbol{A}$ (with dimensions $r \times c$) as

$$\boldsymbol{A} = \boldsymbol{U}\ \boldsymbol{\sigma}\ \boldsymbol{V}^T \tag{7.1}$$

where $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{\sigma}$ represent the left singular vector matrix, the right singular vector matrix, and the diagonal matrix of singular values**.**

Let k denote the frame index such that k = 1 to N_f (assuming there are N_f frames in the video sequence). We decompose the frame $\boldsymbol{A}$ of the original video using Eq. (7.1) and the corresponding frame $\boldsymbol{A}^{(d)}$ of the distorted video as

$$\boldsymbol{A}^{(d)} = \boldsymbol{U}^{(d)}\ \boldsymbol{\sigma}^{(d)}\ \boldsymbol{V}^{(d)T}$$

where $\boldsymbol{U}^{(d)}$, $\boldsymbol{V}^{(d)}$ and $\boldsymbol{\sigma}^{(d)}$ denote the left, right singular vectors and singular value matrices respectively for $\boldsymbol{A}^{(d)}$. We then measure the change in singular vectors using dot products as

$$\alpha_{jk} = \boldsymbol{u}_{jk} \cdot \boldsymbol{u}_{jk}^{(d)} \tag{7.2}$$

$$\beta_{jk} = \boldsymbol{v}_{jk} \cdot \boldsymbol{v}_{jk}^{(d)} \tag{7.3}$$

where α_{jk} (j = 1 to t and k = 1 to N_f) represents the dot product between the unperturbed (i.e. original) and the perturbed (i.e. distorted) j^{th} left singular vectors ($\boldsymbol{u}_j$ and $\boldsymbol{u}_j^{(d)}$) and β_{jk} denotes that for the right singular vectors ($\boldsymbol{v}_j$ and $\boldsymbol{v}_j^{(d)}$) of the k^{th} video frame.

Similar to Eq. (3.10) in Chapter 3, we define the feature vector $\boldsymbol{\Gamma}_{jk}$ for the k^{th} frame to represent the change in the singular vectors as follows

$$\boldsymbol{\Gamma}_{jk} = |\alpha_{jk} + \beta_{jk}| \tag{7.4}$$

Note that Eq. (7.4) defines a t-dimensional vector for the k^{th} video frame with

elements γ_j (j = 1 to t). Same as Eq. (3.11), we use a Minkowski summation (with pooling exponent as 2) and logarithmic scale to obtain the spatial quality of the k^{th} video frame as

$$Q_k^{(spatial)} = log\left(1+\left(\left(\sum_{j=1}^{t}\gamma_j^2\right)\right)^{\frac{1}{2}}\right) \tag{7.5}$$

We also normalized $Q_k^{(spatial)}$ in the range [0, 1] with 1 denoting perfect quality and 0 denoting worst quality.

Now one can use a simple average of the spatial qualities of all the frames as the overall spatial quality measure. Therefore, the overall spatial quality score S_{avg} is obtained as

$$S_{avg} = \frac{1}{N_f}\sum_{k=1}^{N_f} Q_k^{(spatial)} \tag{7.6}$$

It follows that S_{avg} also lies in the range [0, 1] with S_{svg} = 1 denoting the perfect spatial quality while S_{svg} = 0 denoting the worst quality (as compared to the original video). Further, for simplicity in notation, we drop the sub-script j from $\boldsymbol{\Gamma}_{jk}$ and define

$$\boldsymbol{c}_k = \boldsymbol{\Gamma}_k \text{ (with } k = 1 \text{ to } N_f) \tag{7.7}$$

Here $\boldsymbol{c}_k$ will be a t-dimensional vector (since j = 1 to t) associated with frame k and can be used to compute the spatial quality of frame k by Minkowski summation of its elements and the use of logarithmic scale. In the next section, we will use $\boldsymbol{c}_k$ to compute the changes in spatial quality over time.

7.2.2 Temporal Quality Measure

The spatial quality S_{svg} defined in Eq. (7.6) alone may be deficient in predicting the overall quality.

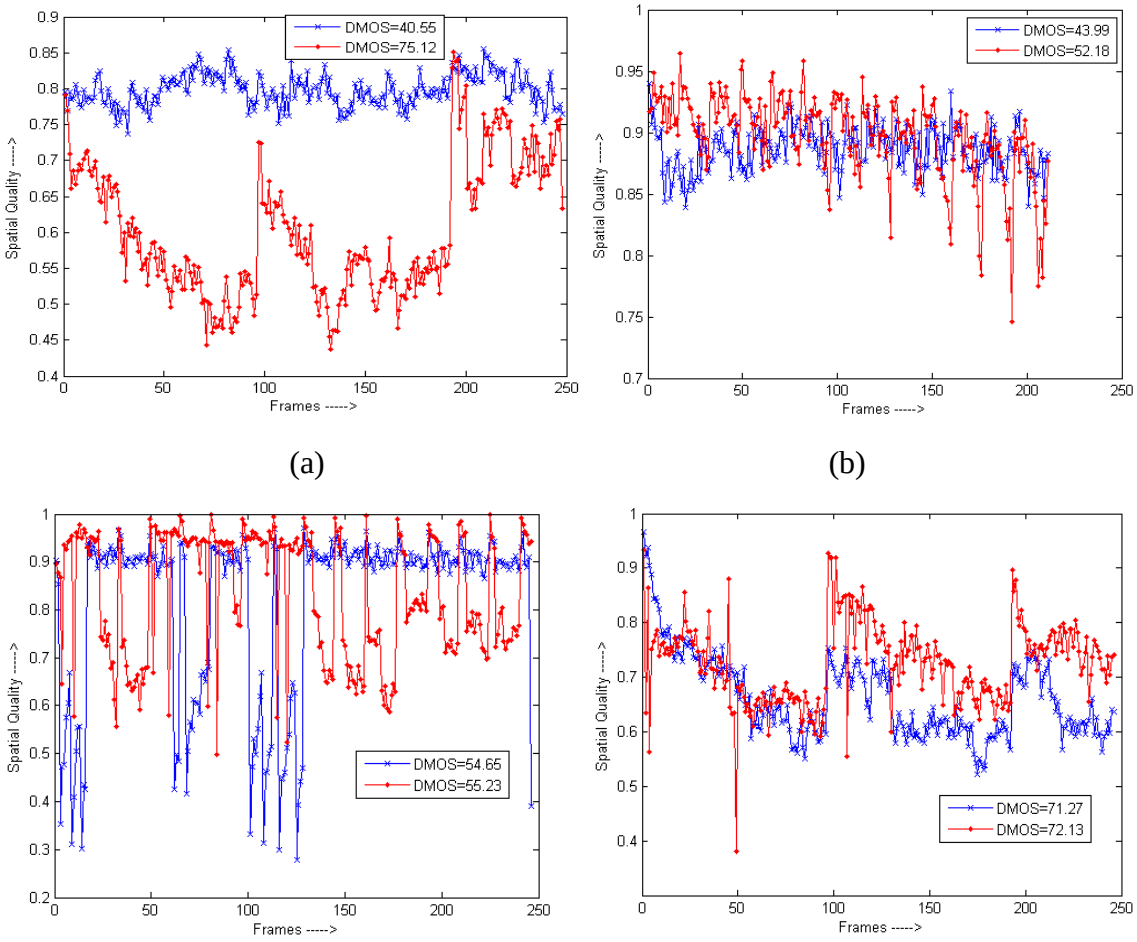


Figure 7.2: Plots of spatial quality of frames for videos with different DMOS's.

We begin with the observation that a better quality video tends to have smaller spatial quality variation over time as compared to a video with poorer quality. In other words, in general, a better video will be characterized by a more constant quality over time while a poorer quality video exhibits larger variation, i.e. more fluctuations. As an example to illustrate this point, we show the spatial quality (computed using Eq. (7.5)) of each frame for videos with different DMOSs in Figure 7.2. These videos have been taken from the LIVE video database [80] which has also been used as the main test database (details will be described later in Section 7.3). The four sub-plots shown in Figure 7.2 are from videos such that they are representative of different video contents and distortions. Note that a

lower DMOS implies a higher quality while higher DMOS indicates poorer video quality. We make the following observations from Figure 7.2.

- As shown in Figure 7.2 (a), a lower quality video (DMOS = 75.12) has frames with lower spatial quality as compared to a video with DMOS = 40.55. Here, the average spatial quality will be lower for video with DMOS = 75.12 while the same will be higher for the video with DMOS = 40.55. In such cases, on the expected lines, the average spatial quality score will be reasonably effective in predicting the overall video quality. This is the reason why averaged spatial quality can still be used as an approximation of the overall video quality.
- In Figure 7.2 (b), the reader will notice that while the frame level average quality is similar for both the videos they have different DMOS's as indicated. In fact a higher averaged spatial quality corresponded to the video which actually had a lower overall perceived quality (i.e., higher DMOS). For that reason simple averaging will be less meaningful to determine the overall video quality and less accurate.

From the two points mentioned above we can conclude the following. S_{svg} alone can still be used for overall VQA as explained in point (a) above. However, it can be ineffective in the cases when a similar proportion of high and low quality frames appear in the video sequence as demonstrated in point (b). The reason is that it gives equal weight (i.e. importance) to all the frames irrespective of their perceptual impact (this is similar to MSE/PSNR). One approach to tackle this is to use the worst case pooling strategy (or percentile pooling) [53], [194]. In this, instead of averaging over all the frames, one uses only the lowest (i.e. worst quality) H % quality scores. Therefore, the

overall spatial quality score S in this case is thus obtained as

$$S = \frac{1}{N_H} \sum_{k \in \mathbf{H}} Q_k^{(spatial)} \tag{7.8}$$

Here, $\boldsymbol{H}$ denotes the set with lowest H % quality scores and N_H is the number of elements in $\boldsymbol{H}$. S computed above is expected to be more effective because it is known that lower quality frames usually have larger impact on the overall perceived quality. Hence we expect S to tackle the deficiency of S_{svg} mentioned in point (b) above. However, there is another factor which can impact the overall perception of the video quality namely the temporal variations/fluctuation of spatial quality. We show two cases in Figure 7.2 (c) and (d) for which S will overestimate or underestimate the error. From Figure 7.2 (c) one can see that the subjective score for the two videos is nearly the same. However, the worst case pooling strategy will predict lower score (we found that S = 0.7274, note that bigger value of S implies higher quality) for the video with DMOS = 54.65 since the low quality frames in this video have lower quality than the low quality frames in the video with DMOS = 55.23. It is also easy to see that a higher score (it was found that S = 0.7856) will be assigned to the video with DMOS = 55.23. Another example is shown in Figure 7.2 (d) where S = 0.6626 for the video with DMOS = 71.27 and S = 0.7160 for the video with DMOS = 72.13. Here again the two videos have nearly the same subjective score (in fact the video shown in red in Figure 7.2 (d) has slightly lower subjective visual quality) but S scores cannot capture this and assigns higher score to the video with lower subjective quality. This happens because the video with DMOS = 72.13 (red) has relatively more fluctuations than the one with DMOS = 71.60 (blue). This can lower the satisfaction level of the viewers despite the fact that many frames of the video with DMOS = 72.13 have higher quality than the one with DMOS = 71.60. In

other words, S can overestimate (as in Figure 7.2 (c)) or underestimate (as in Figure 7.2 (d)) the error. This happens because the worst case pooling ignores the temporal impact of poor quality frames: the occurrence poor quality frames at regular intervals is usually more annoying. Therefore S alone can be inadequate in capturing the effect of quality variations which play a role in the overall subjective viewing experience. This idea is in agreement with the results reported in [69]. The authors in [69] proposed a subjective quality assessment method known as Mean Time Between Failures (MTBF) in which the viewer continuously indicates the presence of perceptual artifacts (like blockiness, blurriness) in the video sequence by using a buzzer. The viewer is allowed to keep the buzzer pressed if the entire stretch of the video sequence looks bad. The idea behind this methodology is that, the viewer intuitively tends to give feedback intermittently, with a frequency correlating with how bad the video looks. In essence, the MTBF attempts to take into account the variations of quality along the time axis. In addition, it has been reported in [71] that a poorer quality video tends to have larger difference between frames (indicated by a larger standard deviation of the differences between pixel values at the same location in space at successive frames) in comparison to a higher quality video. Therefore, variation of quality in time is an important factor in VQA [69], [71]-[72], [169] and we use it to adjust the value of S computed from worst case pooling. We proceed as follows to account for it.

Using $\boldsymbol{c}_k$ defined in Eq. (7.7), we define

$$\boldsymbol{d}_k = \frac{1}{4F}\sum_{z=1}^{F}\left(|\boldsymbol{c}_k - \boldsymbol{c}_{k-z}| + |\boldsymbol{c}_k - \boldsymbol{c}_{k+z}|\right) \tag{7.9}$$

where $\boldsymbol{c}_k$, $\boldsymbol{c}_{k-z}$ and $\boldsymbol{c}_{k+z}$ are respectively the feature vectors of frames k, k-z and k+z; F specifies the number of frames (on each side of the k^{th} frame) to consider. So, the frame

index k will be from F+1 to N_f -F (instead of 1 to N_f). In this chapter, we used F = 2, i.e. two frames on either side of the current frame k to compute $\boldsymbol{d}_k$, which denotes the change or the variation in the elements of vector $\boldsymbol{c}_k$ of frame k with respect to the neighboring F frames. Therefore, the elements of $\boldsymbol{d}_k$ can be thought as the indicator of the change in spatial quality. In other words, $\boldsymbol{d}_k$ accounts for the variance in the spatial quality which is perceptually relevant as we have already pointed out. Further, it is easy to see that all the elements of $\boldsymbol{d}_k$ will be zero if the neighboring F frames have the same quality. This means that in such case the instantaneous video quality would depend only on the spatial factor because there are no temporal fluctuations at that instant.

Like $\boldsymbol{c}_k$, $\boldsymbol{d}_k$ will also be a t-dimensional vector. We then sum the elements of the vector $\boldsymbol{d}_k$ using Minkowski summation and then use logarithmic scale (the same procedure was followed to compute spatial quality of each frame) to obtain the temporal score $Q_k^{(temporal)}$ for each frame k. Finally, we obtain the overall temporal score T for the video as

$$T = \frac{1}{(N_f - 2F)} \sum_{k=F+1}^{N_f - F} Q_k^{(temporal)} \tag{7.10}$$

As seen from Eq. (7.10), there will be N_f -2F frames (instead of N_f) since the first and last F frames are left out due to insufficient number of past and future frames respectively. We can consider the temporal quality T as a factor which accounts for the effect that the qualities of the nearby frames have on the current frame. That is the per-frame distortion is not the perceived distortion at the specified time point [83] because the perceived distortion is also affected by the distortion in the nearby frames.

7.2.3 Overall Video Quality Prediction

We have two factors namely S (spatial) and T (temporal) contributing to the overall

video quality. However, there is evidence that the human perception is also affected by the interaction between S and T [65], [79]. This is because the interaction term can be thought as the overlap between the two factors and it represents the adjustment for the combined effect in perception [195]. Therefore, a simple linear combination of S and T or their interaction (multiplicative) alone may be not suffice for effective quality prediction. This has also been noted by the authors in [65] where the spatial and temporal factors were combined using different exponent parameters.

In this chapter, we used SVR for combining the two factors. Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ... \boldsymbol{x}_l\}$ and $\{y_1, y_2, ... y_l\}$ denote the training set. Here, each $\boldsymbol{x}_i = (S_i, T_i, S_i T_i)$ represents the 3-dimensional vector consisting of the spatial, temporal and the interaction (i.e. multiplicative) term and each y_i is the associated subjective score (i.e. target value) for the i^{th} video. Given the training data $(\boldsymbol{x}_1, y_1), ... (\boldsymbol{x}_l, y_l)$, we find the weight vector $\boldsymbol{W} = (w_1, w_2, w_3)$ and the bias (constant) b (refer to Section 4.3 for details).

7.3 Experimental Results and Analysis

We used video sequences from three publicly available video databases (details are provided in the Appendix). In total, we have used 260 distorted video sequences from the three databases: 150, 78 and 32 video sequences (with their associated subjective viewing scores) respectively from LIVE, EPFL and TUL video databases. It may also be pointed out that many of video sequences used in these databases are different and this allows us to evaluate the performance of different video quality metrics on wider video contents. In addition, as mentioned, the distortion types occurring in these databases are due to video processing/coding techniques like H.264/AVC which are fast gaining

industry appreciation. This therefore helps us to evaluate the performance of different VQA algorithms in predicting quality of video sequences corrupted/distorted by 'state-of-the-art' coding/processing techniques. We now describe the partitioning of the training and test sets.

First, we will report the results for the LIVE video database using 10 fold CV test. To this end, the data is split into 10 chunks, one chunk is used for testing and the remaining 9 chunks are used for training. The experiment is repeated with each of the 10 chunks used for testing. The average accuracy of the tests over the 10 chunks is taken as the performance measurement. The splitting of the data into 10 chunks was done such that the video contents presented in one chunk do not appear in any of the remaining chunk (and this chunk is used as the test set). Similar to the definition in Section 4.4.1, one video content is defined as all the distorted versions of an original video sequence. We use Q_{10} to denote the results for the 10 fold CV. This test allows us to judge the performance of the proposed scheme to untrained video content(s).

Next we use cross-database validation: training set comes from one video database while the test set is from another video database. Here, we first used the EPFL database for training while LIVE video database forms the test set. This case is denoted by the symbol Q_{EPFL} (which means training is done with EPFL database). This test is meaningful on two counts: (a) the video sequences in the two databases are all different, (b) three (out of 4) distortion types do not appear in the training set. Secondly, we used the LIVE video database after excluding the videos with H.264 distortion as the training set while the test set is comprised the 78 video sequences from the EPFL database. We excluded the sequences with H.264 distortion from the training set since the test set contains sequences that have been distorted due to H.264/AVC compression. Once again this

ensures proper and fair metric verification due to the design of the training and test sets. We denote this test case as Q_{LIVE} (which means training is done with LIVE database after excluding the sequences distorted due to H.264 compression). Lastly, the smallest video database i.e. TUL is used for training while the remaining two databases form the test set. This test case is denoted by Q_{TUL} (i.e. training database is TUL). Although EPFL and TUL databases share two reference sequences namely 'foreman' and 'mobile' nevertheless the remaining 6 (out of 8) reference sequences in TUL database are different. In addition, the distortion levels in the two databases are different. Further, none of the video sequences in LIVE database is present in EPFL and TUL databases. Like in previous cases, this test case also helps to ensure that training and test sets come from distinct video content(s) and/or distortion types.

In all we present the results for 4 different types of training and test sets i.e. Q_{10}, Q_{EPFL}, Q_{LIVE} and Q_{TUL}. The reader will appreciate the fact that all the four cases help in better and more comprehensive metric verification. In summary, the described test methodologies are quite effective and help ensure that there is no *parameter tuning* or *optimization* towards the test set. In other words, the problem of overfitting (wherein a trained system performs well only for training data and poorly on new unseen data) is alleviated.

The experimental results are reported in terms of two criteria which are commonly used for performance comparison namely: C_P (for prediction accuracy) and Spearman rank order correlation coefficient C_S (for monotonicity), between the subjective score and the objective prediction. A better quality metric will have higher C_P and C_S. The 95% confidence intervals (CI) have also been used to indicate the statistical significance of the results. A 4-parameter monotonic logistic mapping between the objective outputs and the

subjective quality ratings was also employed, for reasons already explained in Section 3.3.1.

7.3.1 Performance Comparison

We first compare the performance of the proposed algorithm with the relevant existing metrics with LIVE as the test database. The other algorithms being compared include the widely used PSNR, Speed-SSIM [78], V-VIF [82], the VQM [79], the algorithm based on ABT-JND model [196], MC-SSIM [75] and MOVIE [66]. The results for all these algorithms except MC-SSIM and ABT-JND have been obtained from [66] while the results for these two have been cited from their respective papers. We first present C_P and C_S values for full LIVE database (150 distorted videos) in Figure 7.3 (a), and (b) respectively. One can observe that the proposed scheme denoted by Q_{10}, Q_{EPFL} and Q_{TUL} performs well and is better than other algorithms. We further present the results for the performance on individual distortions[4] in Figure 7.3 (c) and (d). We can see that the proposed scheme again performs well overall better than other schemes. In addition, the 95% CIs are smaller for the proposed scheme.

We further present the results of the proposed algorithm for the EPFL database denoted as Q_{LIVE} (recall that this means training with LIVE database after excluding the sequences distorted due to H.264 compression) and Q_{TUL} (which implies training with TUL database) in Figure 7.4 (a). We also show the results for MOVIE, VQM and PSNR. We are unable to report results for other schemes such as Speed-SSIM, V-VIF as their codes are not publicly available. Moreover, the primary aim of reporting the results for EPFL database is to examine the performance on untrained data.

[4] Note that the results of individual distortion types for MC-SSIM and the algorithm based on ABT-JND model are not plotted as they are not available in their respective references [75], [196].

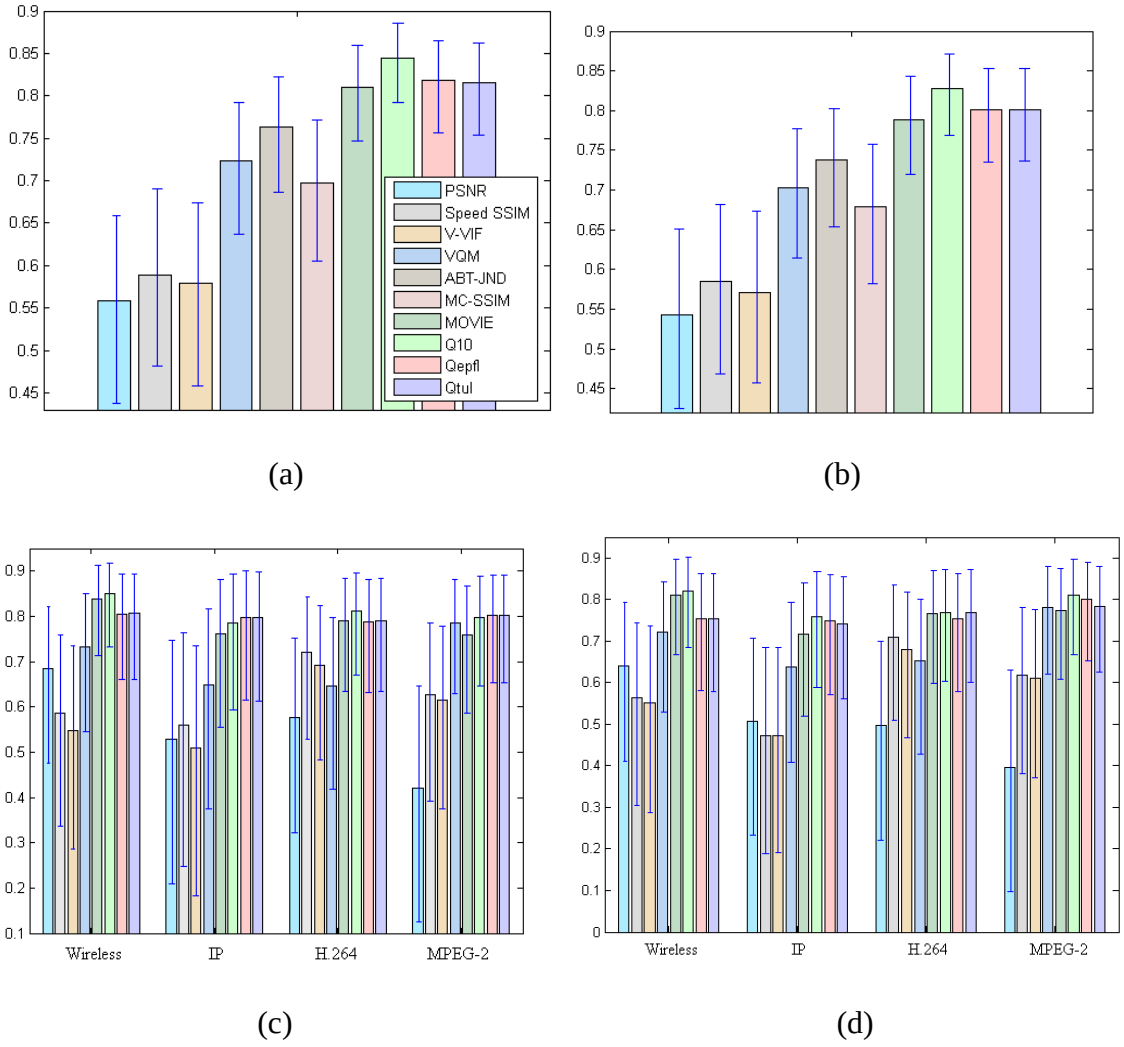
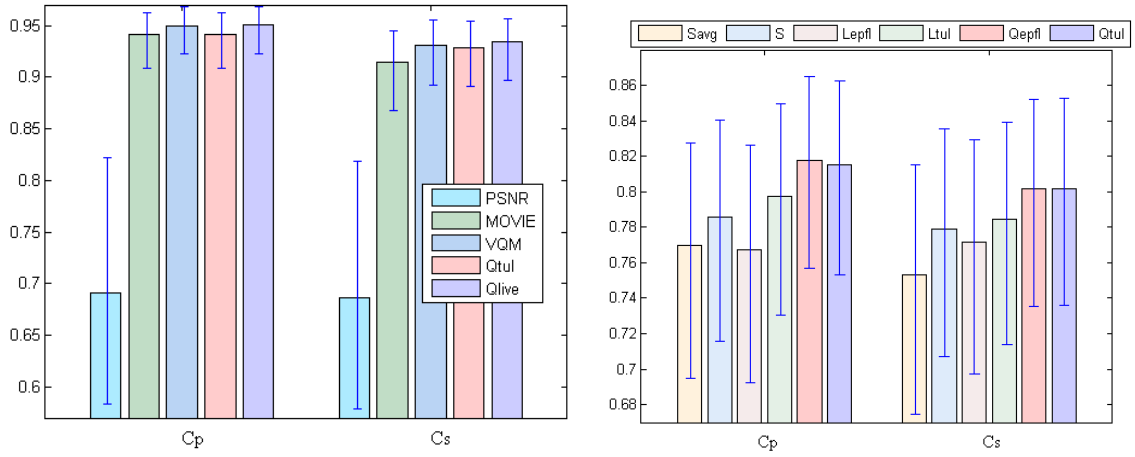


Figure 7.3: Performance comparison for LIVE video database with 150 distorted videos (a) C_P comparison for the full test database, (b) C_S comparison for the full test database, (c) C_P comparison for individual distortion types, (c) C_S comparison for individual distortion types. For (a) and (b), the bars from left to right are for PSNR, Speed-SSIM, V-VIF, VQM, ABT-JND, MC-SSIM, MOVIE, Q_{10}, Q_{EPFL} and Q_{TUL}. For (c) and (d), the bars from left to right are for PSNR, Speed-SSIM, V-VIF, VQM, MOVIE, Q_{10}, Q_{EPFL} and Q_{TUL}. The error bars denote the 95% CIs for C_P and C_S.



(a) (b)

Figure 7.4: Performance comparison for the EPFL database (totally 78 distorted videos) (a) C_P and C_S; the bars from left to right are for PSNR, MOVIE, VQM, Q_{TUL} and Q_{LIVE} (b) C_P and C_S comparison for LIVE video database; the bars from left to right are for S_{avg}, S, L_{EPFL}, L_{TUL}, Q_{EPFL} and Q_{TUL}. The error bars denote the 95% CIs for C_P and C_S.

Table 7.1: Performance comparison on HD video database

	PSNR	VQM	Q_{TUL}	Q_{EPFL}	Q_{10}
C_P	0.687	0.672	0.7923	0.8023	0.8456
C_S	0.694	0.685	0.7880	0.7925	0.8344
RMSE	0.759	0.767	0.6388	0.6012	0.5822

Nevertheless, we still report the results for PSNR (widely used for VQA) and MOVIE (code is publicly available at [80]) which is the best performer (excluding the proposed metric) for LIVE database. One can observe that the C_P values for Q_{TUL} and MOVIE are nearly the same but Q_{TUL} has higher C_S. Q_{LIVE} on the other hand outperforms MOVIE both in terms of C_P and C_S. This again demonstrates the robustness to untrained video contents and/or distortion types. As expected, PSNR's performance is much worse than MOVIE, VQM, Q_{LIVE} and Q_{TUL}. Finally, Table 7.1 reports the C_P, C_S and RMSE values for

a high definition (HD) video database (resolution being 1920 by 1088) with totally 128 distorted videos along with the subjective scores [243]. In addition, we also report the performance of PSNR and VQM for this database. Since the source codes for other metrics such as Speed SSIM, V-VIF etc. are not available, those results are not computed. Furthermore, we cannot report the performance of the existing metric MOVIE as its code runs out of memory (this is probably due to high resolution of the HD video sequences present in this new database) which again confirms its high computational requirements and that a low complexity method like ours is more desirable. We can see that the proposed metric (denoted by Q_{TUL}, Q_{EPFL} and Q_{10}) outperforms PSNR and VQM (both are widely used in video quality assessment) by a significant margin.

7.3.2 Further Discussion

The proposed method relies on quality variations along time axis for computing the temporal factor. As can be noted from Eq. (7.9) we used the absolute difference between the feature vectors of past and future frames. We therefore experimented by replacing the SVD based method used in this chapter with SSIM, VIF and the methods proposed in Chapters 5 and 6. We however found that while we achieve reasonably good spatial quality prediction, the temporal quality evaluation is not as effective. The reason for this could be the large number of features whose difference is taken for temporal quality determination. For example, SSIM provides a distortion map which is equal to the frame size i.e. the feature vectors $\boldsymbol{c}_k$, $\boldsymbol{c}_{k-z}$ and $\boldsymbol{c}_{k+z}$ (used in Eq. (7.9)) in this case are very high dimensional (equal to the total number of pixels in the frame). In contrast to this, feature vectors from the SVD based method are relatively low dimensional (as compared to the frame size).

We have already mentioned in Section 7.2.2 that even the simple averaging out of frame level qualities is reasonably effective for VQA for cases such as those shown in Figure 7.2 (a). We have also discussed that it would fail in cases such as those plotted in Figure 7.2 (b), and therefore a worst case pooling would be more effective. Finally, we have pointed out the limitation of the worst case pooling strategy (i.e. overestimation or underestimation of error/quality) and proposed the use of temporal quality variations to remedy. It is therefore informative to point out the positive impacts of each component in the proposed scheme. To that end, we have shown the C_P and C_S values for LIVE video database for the cases of simple averaging (denoted by S_{svg} and defined in Eq. (7.6)), worst case pooling (denoted by S and defined in Eq. (7.8)) and the results after considering the temporal variations (denoted by Q_{EPFL} and Q_{TUL}) in Figure 7.4 (b). One can clearly see the increase in prediction accuracy from S_{svg} to Q_{EPFL} (or Q_{TUL}). These highlight the positive impact of the using worst case pooling, quality variations and their combination via SVR.

The last point is regarding the use of SVR. As mentioned, we believe that SVR based non-linear combination is better than a linear one. To verify this, we have also shown the results for linear combination case denoted as L_{EPFL} (i.e. training database is EPFL) and L_{TUL} (i.e. training database is TUL) in Figure 7.4 (b). It can be observed that Q_{EPFL} and Q_{TUL} perform better than both L_{EPFL} and L_{TUL}. The reason for better performance of using SVR is that it allows more flexibility for combining the different factors via the use of kernels. On the other hand, the linear combination is less effective as it constraints the relationship to be linear and hence less effective in adjusting the effects (i.e. the weights) of each contributing factor. We can also say that the linear combination is just a special

case for SVR which it can handle by using a linear kernel. As a result, SVR is a better and more powerful tool for feature combination.

We also have similar observation as Chapter 4 regarding the SVs obtained as result of training. For example consider Q_{EPFL} for which the MOS data from the EPFL database is in the range 0 < MOS < 5 (higher implies better quality).We found that, in general, the samples which were chosen as the SVs corresponded to samples with either very low subjective quality score (MOS < 2) or very high quality scores (MOS > 4). This is a reasonable and intuitive selection of SVs since videos with very low or very high quality are the representative of the overall quality range. Obviously if the test signal is of higher quality, it will have greater similarity (i.e. bigger $K(\boldsymbol{x}_i,\boldsymbol{x})$) with the SVs that represent higher quality signal. On the other hand, it will have low similarity ((i.e. $K(\boldsymbol{x}_i,\boldsymbol{x})$ will be smaller) with the SVs corresponding to low quality signals. In essence, SVR predicts quality by determining how "similar" the test signal is with the chosen SVs. The final quality score is just a summation (scaled by appropriate SV coefficients) of such similarity scores with respect to each SV.

7.3.3 Computational Complexity Versus Prediction Accuracy

Even though quite a lot of research effort has been spent on developing VQA algorithms, PSNR is still popular and used widely. The obvious reasons are its low computational complexity and ease of implementation. Therefore, for a VQA algorithm to be practically deployable, its complexity is as important (perhaps more in some situations like encoding) as its prediction accuracy. In this respect, it is worth pointing out that the proposed scheme is computationally much more efficient as compared to

other VQA algorithms that employ motion information from motion estimation (ME) or optical flow. In fact, the complexity of the proposed scheme is only slightly higher than using an IQA algorithm on frame-by-frame basis.

We also note that MOVIE is a reasonably good VQA scheme. However, its major bottleneck stems from the high computational costs as it employs three-dimensional optical flow computation. Regarding computational complexity, for a video sequence with 250 frames (resolution being 768 × 432) the proposed algorithm (assuming training is done offline which will be the case more often than not) requires approximately 1.75 minutes (104.4 seconds) for predicting its quality. On the other hand, a C++ language implemented MOVIE needs approximately 100 minutes. Clearly, the processing time and the related computational effort for MOVIE are too high for practical deployment. Therefore, not only does the proposed scheme perform overall slightly better than MOVIE (the performance is also better or very competitive for individual distortion types), it is much more efficient.

As stated in Section 4.4.7, the computational complexity for frame level SVD in the proposed scheme (assuming frame size $r \times c$) is $O(\min\{rc^2, r^2c\})$. Percentile pooling can be performed with a worst-case complexity of $O(rc\ log(rc))$. The training of the SVR required in the proposed scheme can be done offline and hence does not incur any computational overhead for real time implementation. The overall complexity of our scheme is of course more than IQA metrics like PSNR and SSIM (applied on frame-by-frame basis) but they are less effective for VQA as already pointed out in the chapter. The complexity of our algorithm is also less than VQM for which it is $O((rc)^2)$. Also note that the prediction accuracy of VQM is lower than the proposed method. MC-SSIM which uses SSIM is also computationally more demanding as it utilizes ME and clearly

does not perform as well for LIVE database. Similar remarks can be made for ABT-JND model based algorithm. Many other existing VQA schemes also resort to ME for incorporating motion information which is usually the major factor contributing to the increased computational burden. On the contrary, the proposed SVD based scheme is much simpler since we do not use ME/optical flow fields. Instead, we exploit some basic temporal characteristics that affect video quality. Furthermore, the proposed method benefits from non-linear training based methodology via the use of SVR (with subjective scores as the *ground truth*). These enable efficiency and good prediction accuracy of the proposed scheme. In other words, the proposed scheme achieves a better trade-off between prediction accuracy on one hand the complexity on the other.

7.4 Concluding Remarks

In this chapter, we have first argued and shown that a simple averaging procedure for quality of different frames alone is inadequate for VQA due to the higher impact of poor quality frames. We then employed the worst case pooling strategy to tackle the shortcomings of simple averaging. Next, we have explained and analyzed the drawback of worst case pooling and explored the use of temporal quality fluctuations as an important factor towards effective VQA. Furthermore, the issue of establishing non-linear relationship between the different factors has been tackled with the use of machine learning. Since the individual contribution of each factor to the overall video quality can be non-linear and difficult to be determined *apriori*, the use of machine learning to determine the weights/parameters is more convincing and meaningful than ad-hoc methods.

The proposed metric has been validated using three public video databases (totally 260

distorted videos). It is found to perform better than the relevant existing metrics in terms of agreement with the subjective scores. We have also shown the robustness of the proposed scheme with regards to untrained video content and/or distortion types by way of cross-database validation. The performance of various components of the proposed algorithm has also been shown to assess the impact of each stage. The most crucial advantage of the proposed method is its efficiency as it has lower computational complexity and achieves good prediction accuracy. We have also presented analysis to show that the proposed scheme achieves a better trade-off in complexity and prediction accuracy.

Chapter 8

Nonintrusive Quality Assessment of Noise Suppressed Speech

8.1 Introduction

As mentioned in Chapter 2, perceptual quality assessment of noise-suppressed speech has received less attention in comparison to that of speech distorted by codecs/communication channels. We have also detailed the specific issues concerning quality assessment of noise-suppressed speech in Section 2.3.1. Recognizing this ITU-T has recently approved POLQA, P.863 [4] as the new standard for intrusive speech quality assessment which will also cater to noise-suppression scenarios. However POLQA is still an intrusive metric (i.e. requiring both reference and processed speech files) and so it cannot be employed when the reference signal is unavailable (this can occur in many practical situations, as already elaborated in Chapter 2). In this chapter, we develop a nonintrusive scheme for assessing the quality of noise-suppressed speech.

The remainder of this chapter is organized as follows. Section 8.2 describes the proposed scheme based on mel FBEs and SVR, with reasoning and justification. Experimental results and comparisons are presented in Section 8.3 using two third party

databases, while Section 8.4 gives the performance evaluation with subset of features and presents further discussion. The last section presents the concluding remarks.

8.2 The Proposed Speech Quality Evaluation Scheme

Like visual quality assessment, the task of assessing speech quality can also be considered as a two-step process. In the first step, features are selected/extracted from the speech signal to provide a compact representation of the signal with the regard of quality. The second stage comprises a "cognitive mapping" to fuse the extracted features into a quality score. In this section, we provide the details of the proposed scheme. We first describe the detection of features, and then discuss the feature mapping procedures.

8.2.1 Feature Selection for Quality Assessment of Noise Suppressed Speech

Feature selection/extraction is the process of computing a compact numerical representation that can be used to characterize the speech signal for quality evaluation purposes. Different speech features have been used for quality assessment. In [103], spectral flatness, spectral dynamics, spectral centroid, speech variance, pitch period and excitation variance have been used. Perceptual linear prediction (PLP) cepstral coefficients have been used in [7], [9]. Ref [101] takes into account the temporal discontinuity in the signal (since it usually has negative impact on perceived quality) and adjusts the quality scores accordingly.

For evaluating the perceptual quality of speech affected by noise suppression, we look

for features that can represent the variations in speech quality due to varying noise conditions (i.e. determine the impact of different noise-suppression schemes). In this chapter, we propose the use of mel FBEs as the speech features. This is because they are sensitive to noise, and can capture the effects of noise addition and noise suppression reasonably well. Since FBEs have been successfully used in enhancing speech quality [197]-[202] we believe that they can be effective and be exploited for quality assessment of noise-suppressed speech. The method described in [203]-[204] uses the log MMSE estimator of the FBEs to obtain enhanced FBEs. Let c_y and c_s denote the Mel-FBEs for the noisy and clean speech signal, respectively. The enhanced FBEs can then be estimated using the log-MMSE estimator as

$$\hat{c}_s(b) = \exp(\mathrm{E}\{\log c_s(b) | c_y(b)\}) \tag{8.1}$$

where E{.} denotes the expectation operator, and b is the Mel filter bank channel index. One of the solutions for Eq. (8.1) is

$$\begin{aligned} \hat{c}_s(b) &= \exp(\mathrm{E}\{\log c_s(b) | c_y(b)\}) \\ &= G\big(\xi(b), v(b)\big)\, c_y(b) \end{aligned} \tag{8.2}$$

where the gain is given as

$$G\big(\xi(b), v(b)\big) = \frac{\xi(b)}{1+\xi(b)} \exp\left\{ \frac{1}{2} \int_{v(b)}^{\infty} \frac{e^{-t}}{t} dt \right\}$$

and $v(b)$ is defined by the adjusted apriori SNR $\xi(b)$ for each filter bank and the adjusted aposteriori SNR $\gamma(b)$ such that

$$v(b) = \frac{\xi(b)}{1+\xi(b)} \gamma(b)$$

where

$$\xi(b) \equiv \frac{\sigma_s^2(b)}{\sigma_z^2(b)} \quad , \quad \gamma(b) \equiv \frac{e_y^2(b)}{\sigma_z^2(b)}$$

such that

$$\sigma_z^2 \cong \sigma_d^2 + 2\frac{\sum_k \psi_b^2(k)}{\left(\sum_k \psi_b(k)\right)^2}\sqrt{\frac{\sigma_s^2(b)}{\sigma_d^2(b)}}\sigma_d^2$$

with $\sigma_s^2(b), \sigma_d^2(b)$ denoting the variance of clean speech *s* and additive noise *d* and $\psi_b(k)$ being the Mel band-pass filter.

We can observe from Eq. (8.2) that the enhanced FBEs $\hat{c}_s(b)$ are affected by the gain due to the enhancement. Eq. (8.2) is a special case of enhancement of FBEs using the MMSE approach. The approach in [205] also utilizes the MMSE estimation of FBEs for speech enhancement to achieve more robust speech recognition performance. Other methods, such as spectral subtraction, have also been recently used [199]-[200] for the enhancement of FBEs. The method reported in [202] uses a Wiener filter (which is derived using visual features) for estimating the enhanced FBEs. The approach described in [206] uses a channel attention matrix to obtain weighted FBEs such that the less corrupted channel is given more attention to improve recognition performance. A similar scheme reported in [207] uses a top-down multiplicative attention filter for enhancing FBEs of noisy speech. FBEs can also be used for speech enhancement using a statistical framework. For example, a GMM in the log FBE domain can be used [202]. Furthermore, the FBEs have also been used [208]-[209] in subband adaptive speech filtering techniques for improving speech recognition performance in reverberant environments. Recently, FBEs have also been used to obtain more robust features for overlapping speech recognition (i.e. recognizing speech from multiple distant microphones (multi-channel) for multiparty meetings where more than one speaker can be active at the same time). The basic idea [210]-[211] to achieve this is to find a mapping (by a neural

network or some regression analysis) between the log FBEs of signals from distant microphones and the log FBEs of clean signal. We therefore expect that the FBEs provide a reasonably effective and discriminative representation space of the speech signal towards differentiating the effects of noise injection and noise-suppression (i.e. speech enhancement) and hence assess the quality.

With *s* denoting clean speech which has been corrupted by additive *d* the noisy signal *y* is represented as*:*

$$y(n) = s(n) + d(n) \tag{8.3}$$

with *n* being the time-sample index. We may write Eq. (8.3) in the frequency domain as

$$Y_w(k) = S_w(k) + D_w(k) \tag{8.4}$$

where $Y_w(k)$, $S_w(k)$ and $D_w(k)$ respectively denote the DFT of noisy speech signal *y*, clean speech signal *s* and the noise signal *d* with frame index *w*, while *k* is the frequency index.

The aim of speech enhancement is to obtain an estimate of the underlying clean speech signal from the noisy signal. We denote the complex gain (in frequency domain) due to speech enhancement as $H_w(k)$. Then, the estimate $\hat{S}_w(k)$ of the clean signal $S_w(k)$ can be written as

$$\hat{S}_w(k) = H_w(k).\, Y_w(k) \tag{8.5}$$

We can regard Eq. (8.5) to be a general expression for speech enhancement where the complex gain $H_w(k)$ is different for different speech enhancement schemes [213].

The enhanced speech signal $\hat{s}(n)$ (sampling frequency being 8 kHz) is segmented into 50% overlapping frames of 20 ms in length, with a frame rate of 100 Hz. Each individual frame $\hat{s}_w(n)$ is Hamming windowed and transformed to frequency domain by applying an *N*-point FFT. We denote the resulting amplitude spectrum as $|\hat{S}_w(k)|$ $(1 \leq k \leq N)$. Note that

the phase is discarded and only the magnitude of the spectrum used. We used a 512-point DFT (N = 512) and thus obtain 512 frequency coefficients for each windowed speech frame. In the human ear basilar membrane, there are more receptors for frequencies between 0 to 1 kHz and their number decreases rapidly thereafter. To emulate this, mel-filter banks are used. The mel-filter bank [214] consists of overlapping triangular filters with cutoff frequencies determined by the centre frequencies of the two adjacent filters. These filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale. Mel-filter banks have been shown to be reasonably successful in mimicking the non-linear frequency selectivity of the human ear, as demonstrated by the success of MFCCs in speech recognition [173].

Using Eq. (8.5), the FBE from the b^{th} Mel band-pass filter $\psi_b(k)$ can be written as

$$\text{FBE}_b^w = \sum_{k=1}^{N} |\hat{S}_w(k)|^2 \psi_b(k) = \sum_{k=1}^{N} |H_w(k)|^2 |Y_w(k)|^2 \psi_b(k) \tag{8.6}$$

where $1 \leq b \leq M$ (M is the number of Mel-scaled triangular band-pass filters). We can see that each FBE is computed as a linear combination of the energy in a particular subset of DFT subbands. With the use of the mel-scale, lower frequency filtering is at a higher resolution while higher frequency filtering is at coarser resolution. The advantage of using the mel band-pass filters is thus twofold: (a) perceptually important frequencies are enhanced; (b) they help to reduce the feature dimensions (from N to M).

We can observe from Eq. (8.6) that noise-suppression will have impact on the FBEs due to the gain $H_w(k)$. As mentioned, the gain due to different speech enhancement algorithms will be different and thus, FBEs can be used to characterize the effects of noise-suppression. Even though it is observed that noise (and noise suppression) affects FBEs, our aim is to assess whether such changes are efficient and parameterizable for the

purpose of quality assessment. In this chapter, we have used thirteen linearly spaced and twenty seven log spaced triangular filters for grouping the FFT bins and thus, M = 40. The lowest frequency was chosen to be 133.33 Hz, and a linear spacing of 66.66 Hz and log spacing of 1.049 were used. Because the speech signals are sampled at 8 kHz such parameter settings ensure that the filter bandwidth is up to Nyquist frequency of 4 kHz. Thus, 40 FBEs are obtained as the local (i.e. per frame) features.

8.2.2 Further Analysis for Detected Features

Since speech signals carry information through time-domain variation, FBE amplitudes at any given moment will be less meaningful than frame-to-frame variation. In tasks such as speech recognition [173], generally each frame is analyzed for its acoustic content since the goal is to determine the basic units (phonemes) which are used to find the possible underlying word sequence. By contrast, for speech quality assessment, it is necessary to determine a single score for the entire signal. Hence, speech quality is not predicted directly from the per-frame vector, but from its global statistical properties, characterized by the mean and variance of the per-frame features. In addition to the first and second order moments i.e. mean and variance respectively, higher order moments may be used, like skewness and kurtosis as in [103]. However, our experiments show that the higher order moments do not improve the prediction accuracy significantly and also increase the feature vector dimensions. A similar conclusion has been reported in [215] where it was found that lower order moments (mean and variance) are more important than higher order ones (skewness and kurtosis). Therefore, we have used only the mean and variance of log FBEs of all the frames in order to obtain an 80-dimensional (i.e. 2M dimensional) global feature vector to characterize the entire speech signal.

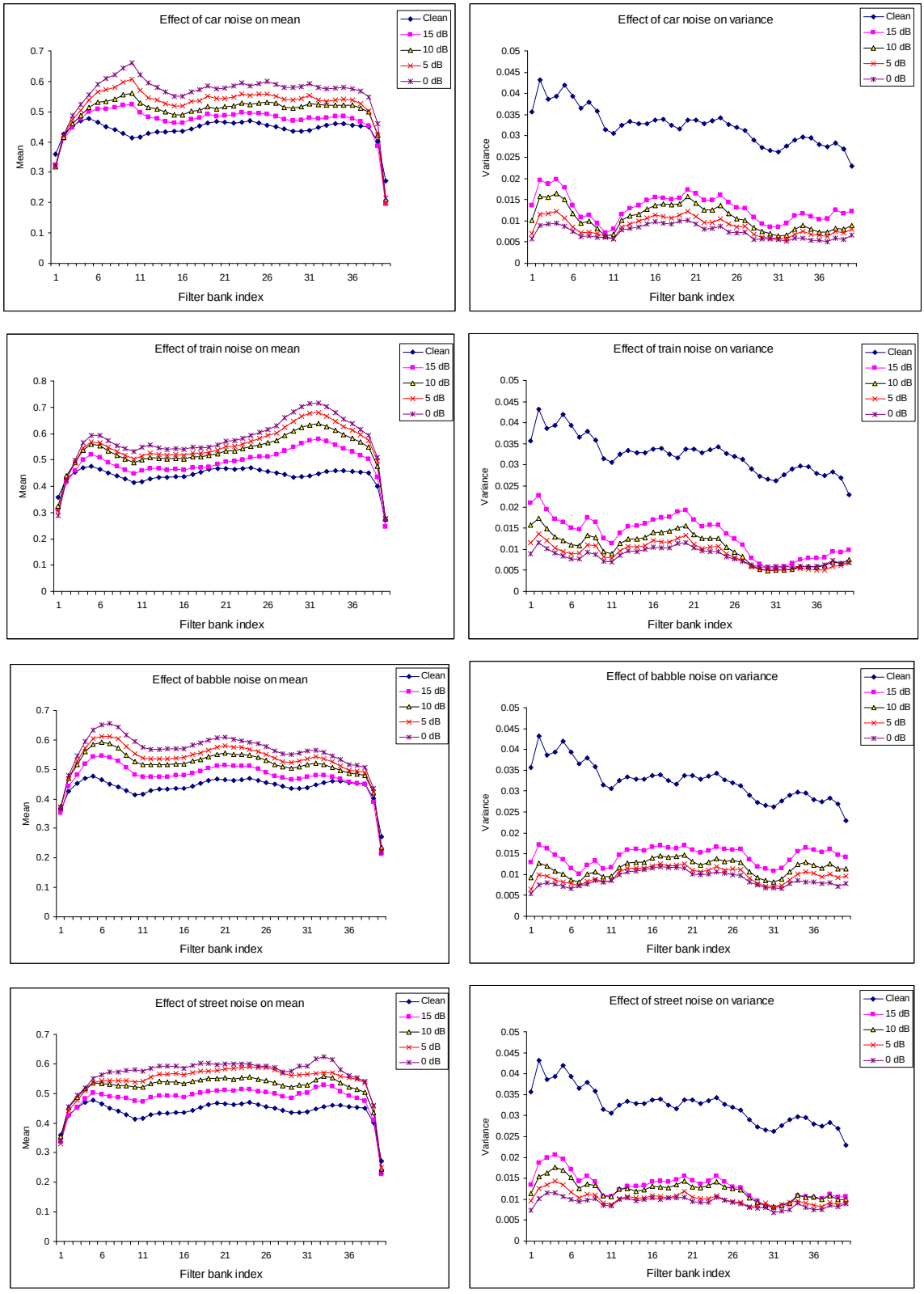


Figure 8.1: The effect of different levels of noise on mean and variance of FBEs

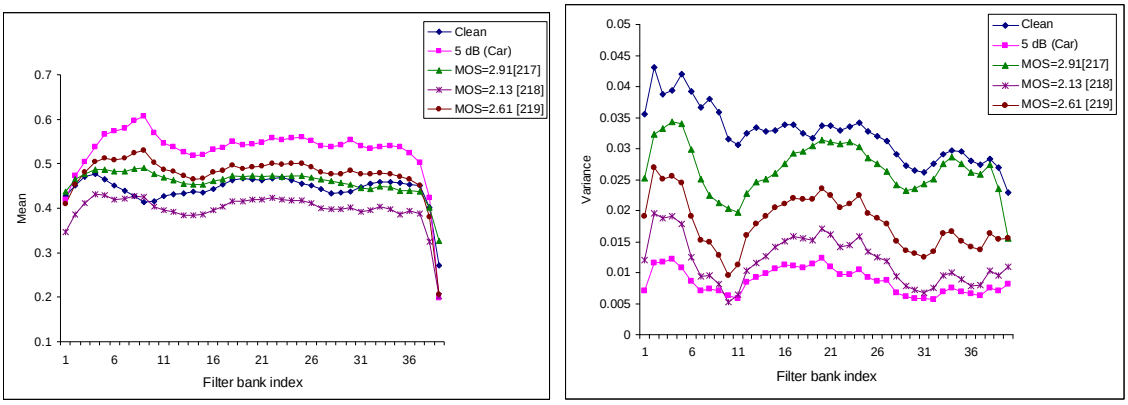


Figure 8.2: Effect of noise-suppression on the mean and variance of FBEs. References [217], [218] and [219] in the legend refer to the papers which describe the specific noise suppression algorithm whose results are plotted.

The complete feature vector $\boldsymbol{x}$ for the speech signal is therefore represented as

$$\boldsymbol{x} = (m_1, m_2 \ldots m_{40}, v_1, v_2 \ldots v_{40})^T \tag{8.7}$$

where m_i is the mean and v_i denotes the variance of the i^{th} log FBE of all the frames. After computing the feature vector defined in Eq. (8.7) we also normalized in the interval [0, 1] before using machine learning in order to avoid the domination of attributes in greater numeric ranges over those in smaller numeric ranges.

As mentioned previously, FBEs are sensitive to noise. In Figure 8.1, we show the effects of 4 different levels of noise (at 0 dB, 5dB, 10 dB and 15 dB) on the mean and variance of log FBEs of speech signals (these have been taken from the speech database used in our experiments). We observe that the mean value generally increases while the variance generally decreases with increasing noise level for all the 4 noise types. Furthermore, FBEs are affected also by different noise suppression algorithms as indicated by Eq. (8.6). As an example, Figure 8.2 shows the effect of noise-suppression on the mean and variance of log FBEs for a speech signal. The subjective scores (MOS)

have also been indicated; the higher the MOS score, the better the subjective quality. In addition, the corresponding noise-suppression schemes have been indicated by their respective references [217]-[219] in the legend of Figure 8.2. The plots in Figures 8.1 and 8.2 are condition averaged plots over 16 speech files. We find that the mean and variances of log FBEs of the speech signal with higher subjective scores (MOSs) are closer to those of the clean signal. We thus observe (refer to Figures 8.1 and 8.2) and infer the following about the changes in FBEs due to noise injection and noise-suppression:

1. Non-linear changes/distortions occur in the feature space.
2. The distortion of the features will also transform the probability distributions. The probability density function (pdf) is expected to be affected and can be characterized by the displacement of mean and variance. In fact, due to the distortion of the features, the pdfs representing clean speech cannot appropriately represent noisy or the noise-suppressed speech. This mismatch leads to increased error rates in typical speech recognition systems [216], [220].
3. The mean values generally increase with noise injection and decrease with noise-suppression. On the other hand, the variance follows a trend opposite to that of the mean. The decrease in variance due to noise injection is expected since the increasing noise tends to bring the value of FBEs closer to each other thereby reducing their variance. This indicates a reduction in the discrimination capabilities of the FBEs. This again explains why the performance of MFCC-based speech recognition systems performing well under clean speech conditions degrades in noisy conditions.
4. The mean and variance provide reasonable distinction between signals of

different qualities. For example, in Figure 8.2, the means and variances of FBEs of the signal with higher MOS (MOS = 2.91) tend to be closer to those of the clean speech. This will help the machine learning algorithm to distinguish the different signals better.

As mentioned, the effects of noise injection and suppression can be complicated and non-linear. It is therefore difficult to establish an *apriori* relation between the changes in mel FBEs and the perceptual quality.

8.2.3 **Feature Mapping**

As mentioned before, the aim of the feature mapping stage is to obtain a single number which denotes the perceived quality of the speech signal. For this, simple techniques like summation, averaging, Minkowski summation, etc. can be used. However, these techniques are generally inadequate due to their inherent limitations as mentioned in Section 2.1.2.2. In our opinion, features may jointly affect the human auditory system's perception of quality; possibly non-linear relationships and partly unknown mechanisms make the task of feature mapping complicated. It is due to this fact that alternative techniques have been used during the past. In [221], a Neurofuzzy inference system has been used while Bayesian modeling has been utilized in [102], [222]. GMMs have also been exploited [7], [99] for feature mapping. Another technique known as multivariate adaptive regression splines has also been explored [7], [104] for feature mapping.

Similar to the previous chapters, we advocate the use of the kernel based method (i.e. SVR) for feature mapping. The major advantage of a kernel based method is: if a problem is non-linear, then instead of trying to fit a non-linear model, one can map the problem from the input space to a new (higher-dimensional) space (called the feature

space) by a nonlinear transformation using suitably chosen basis functions, and use a linear model in the feature space.

8.3 Overall Experimental Results and Discussion

In this section, we present the experimental results with respect to quality prediction accuracy. We also compare the proposed scheme (denoted as Q) with ITU-T P.563 which is the current standard for nonintrusive speech quality assessment. In addition, wherever possible we report the relevant results derived directly from Refs. [7] and [9] as they also used the same speech database for the experiments.

8.3.1 Database description

As mentioned, we use a third-party database which has been developed by employing 13 different noise-suppression schemes on the speech files present in the NOIZEUS database which is a publicly available[5] noisy speech corpus. NOIZEUS database contains 30 IEEE sentences produced by three male and three female speakers, and was corrupted by eight different real-world noises at different SNRs. The noise was taken from the AURORA database [223] and includes suburban train noise, multi-talker babble, car, exhibition hall, restaurant, street, airport and train-station noise. The sentences were recorded in a sound-proof booth using Tucker Davis Technologies recording equipment. The IEEE database was used as it contains phonetically-balanced sentences with relatively low word-context predictability. The thirty sentences were selected from the database so as to include all phonemes of spoken American English. The sentences were originally sampled at 25 kHz and downsampled to 8 kHz.

[5] [Online] Available: http://www.utdallas.edu/~loizou/speech/noizeus/

The developers of the NOIZEUS database subsequently used it [104], [224] in a comprehensive subjective evaluation of 13 speech enhancement algorithms encompassing four different classes of algorithms: spectral subtractive, subspace, statistical-model-based and Wiener-filtering type algorithms. The enhanced speech files were sent to Dynastat, Inc. (Austin, TX) for subjective evaluation using the recently standardized methodology for evaluating noise suppression algorithms based on ITU-T P.835 [106]. It may be mentioned that to reduce the costs of subjective tests, they used 16 clean files (out of 30) corrupted by 4 types of noise (babble, car, street, and train) at two SNR levels (5 and 10 dB). This results in a total of 1792 samples including the unprocessed speech files (16 clean sentences x 4 types of noise x 2 SNR levels x 14 processing algorithms (inclusive of unprocessed noisy speech)). A complete description of the noise-suppression algorithms and the noise-suppressed speech database can be found in [104], [224].

The subjective ratings in the database are available along three quality scales namely signal quality rating (*SIG*), background noise quality rating (*BAK*) and the overall quality rating (*OVRL*) in accordance with P.835. Although there are 1792 speech files available in the database, for comparison between objective and subjective scores a usual way is to compare the per-condition MOS with the per-condition average objective score [106]. The 13 different speech enhancement algorithms were used for processing noisy speech files, and by including the unprocessed noisy speech files also, we get a total of 14 algorithms. Thus, for the per-condition analysis, we obtain a total of 112 (14 algorithms x 2 SNR levels x 4 noise types) objective scores and subjective ratings for comparison.

8.3.2 Evaluation Criteria

In [5], it is suggested that offsets and non-linearities between the scales of objective scores and subjective MOSs be eliminated by applying a 3rd order monotonic function to map the objective scores onto the subjective scale. Following this, we used the 3rd order polynomial to map the objective scores and subjective MOSs. The experimental results are reported in terms C_P, C_S and RMSE between the subjective MOSs and the objective scores (after 3rd order polynomial mapping). A better quality metric will have higher C_P and C_S values and lower RMSE. In addition, we have also employed confidence intervals (for C_P and C_S values) since they can be used to indicate the reliability of an estimate.

8.3.3 Test Results for overall quality assessment

Since the NOIZEUS based database is comprehensive with totally 1792 noise suppressed speech files and their associated subjective quality scores, we test the performance of the proposed metric Q by partitioning the database in different ways to obtain the training and test sets. This is to test the robustness of Q to varied conditions. For the first set of experiments described in this Section, we have used the *OVRL* scores as the ground truth for the SVR algorithm.

First we used 10 fold CV, for which the data is split into 10 chunks, one chunk is used for testing and the remaining 9 chunks are used for training. The experiment is repeated with each of the 10 chunks used for testing. The average of the accuracy of the tests over the 10 chunks is taken as the performance measure. For a visual comparison, we show the scatter plots for Q and P.563 in Figure 8.3. For an ideal metric, all the points would lie on the 45^0 line and the better metric will show less scatter around this line. We can see that Q scatters less around the 45^0 line as compared to the P.563 points.

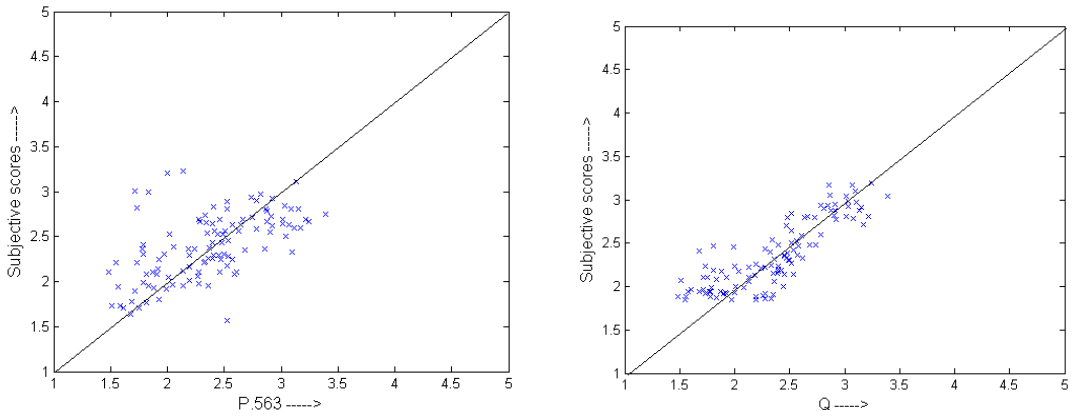
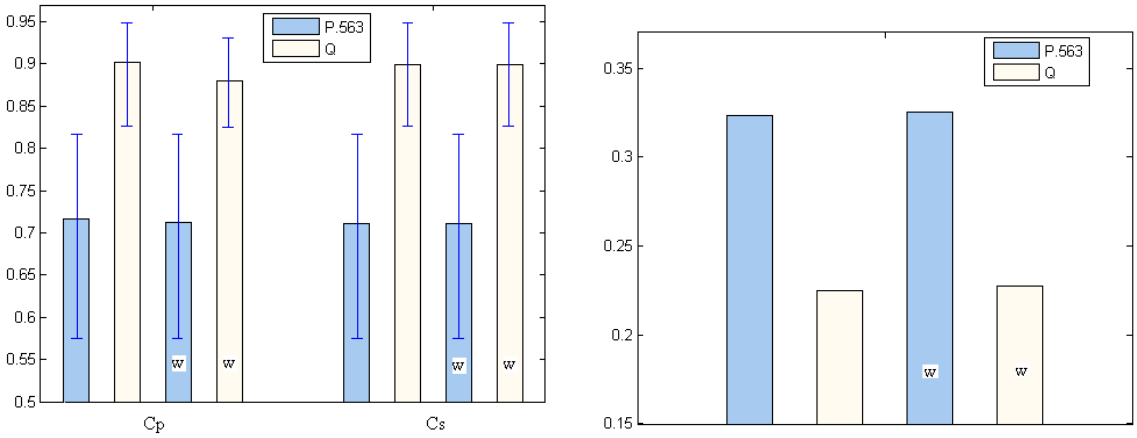


Figure 8.3: Scatter plots of subjective scores versus objective quality scores of P.563 and the proposed Q



(a) (b)

Figure 8.4: Results for proposed Q and P.563 for the full database.
(a) Comparison of C_P and C_S with the error bars denoting 99% confidence interval, (b) Comparison of RMSE values. The bars with 'w' inside are the results without the 3rd order polynomial mapping

For a quantitative comparison, the C_P, C_S and RMSE (with and without the polynomial mapping) are shown in Figure 8.4. We have also included the 99% confidence intervals (denoted by error bars) in Figure 8.4 for C_P and C_S. A smaller confidence interval is associated with higher consistency. We can see that the proposed scheme performs much better than P.563 and achieves significantly higher C_P and C_S with smaller confidence intervals; the results in RMSE show the similar advantages of the proposed scheme. It

may be stated that the current intrusive standard PESQ achieved a correlation of 0.89 for the overall quality prediction as reported by the authors in [104]. They also modified PESQ by employing a training procedure to determine the parameters and the prediction accuracy for *OVRL* scores was found to increase to 0.92 (for the test set).

It is also fair to mention here that P.563 does not use training while the proposed metric uses training. Due to this, it is crucial that the metrics which employ training be tested for their robustness to varying training and test contents. To that end, we select the training and test sets according to the noise sources in a similar way as in [7]. First, speech files are separated according to noise levels: speech files with SNR = 10 dB are used for training while speech files with SNR = 5 dB are used for testing. The results for this case are presented as the first set (Test 1) in Figure 8.5. Secondly, speech signals are separated according to noise sources. Signals corrupted by street and train noise are used for training, and signals corrupted by babble and car noise are left for testing. The results for this case are presented as the second set (Test 2) in Figure 8.5. Lastly, speech files are separated according to noise suppression algorithms. For training, noisy signals processed by spectral subtractive and subspace algorithms are used; noisy signals processed by statistical-model based and Wiener algorithms are left for testing. The results for this case are presented as the third set (Test 3) in Figure 8.5 and we have also included the results without the polynomial mapping for Q and P.563. We can see that Q is reasonably robust to untrained test conditions and again performs much better than P.563 in terms of C_P, C_S and RMSE. We have also presented the results for the double-ended metric (intrusive) proposed by Falk and Chan in [7]. Since they used the same database and data partitioning as ours, the results can be compared. One can observe from Figure 8.5 that Q is very competitive with the method proposed in [7].

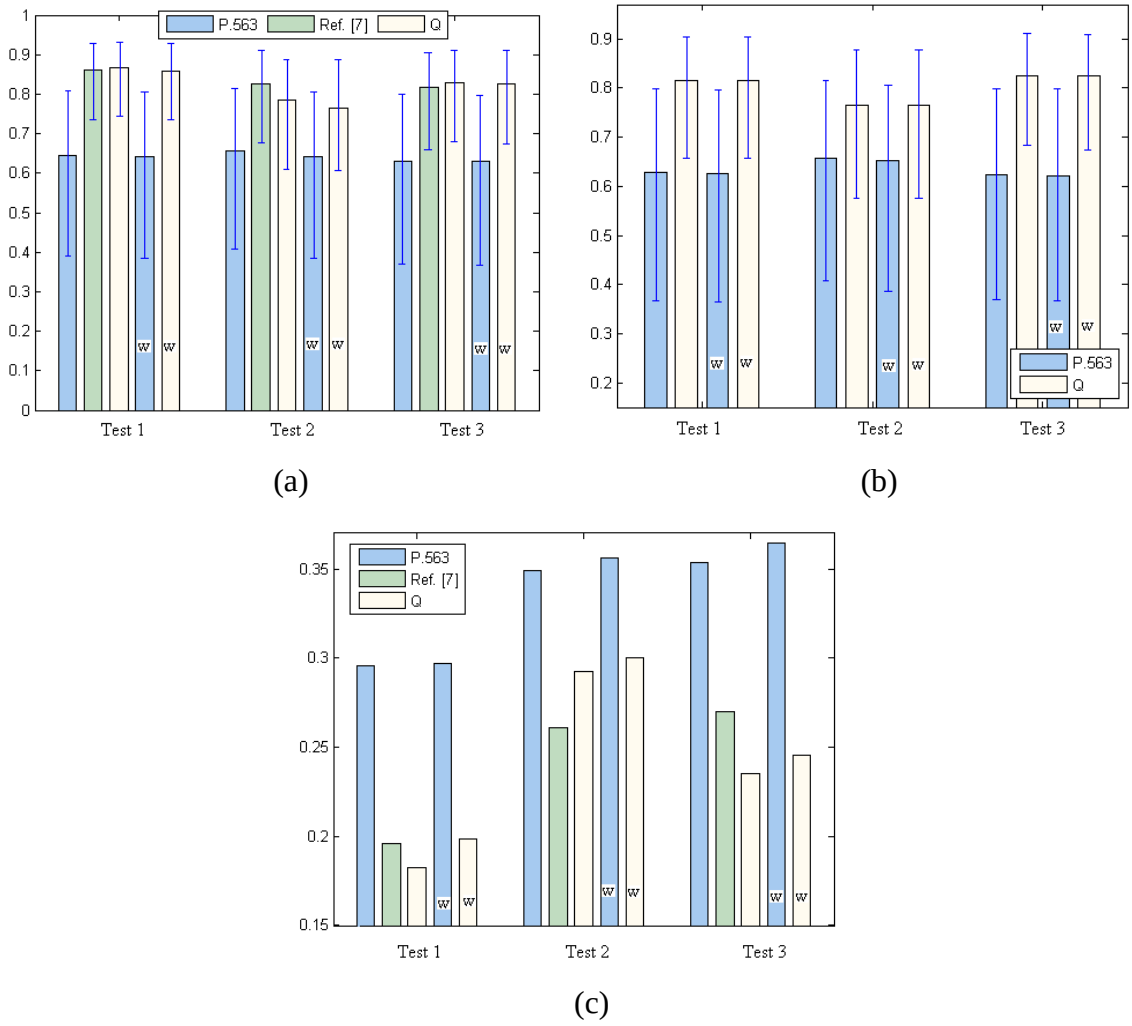


Figure 8.5: Results for the proposed Q with different splitting of data into training and test sets. Results for P.563 are also indicated. (a) Comparison of C_P, (b) Comparison of C_S, (c) Comparison of RMSE values (99% confidence interval bars also indicated for C_P and C_S; refer to text for details about Test 1, Test 2 and Test 3). The bars with 'w' inside are the results without the 3rd order polynomial mapping.

This is significant due to the fact that the method in [7] is intrusive while our method is non-intrusive.

As mentioned, there are 16 different sentences used in the database. Thus, there are 16 different contents available. We conducted tests in which we trained Q on 10 contents (i.e. 1120 sentences) while the remaining 6 contents (i.e. 672 sentences) were used for

testing. The C_P values for Q and P.563 for this test case are respectively 0.8096 and 0.7036 (RMSE values were 0.0654 and 0.1167 respectively). On swapping the training and test sets (which means training with 3/8 of the data and testing the remaining), Q gave C_P = 0.7902 which suggests that Q is fairly robust to untrained contents. The database used in this study uses 4 talkers (two males and two females) for the subjective evaluation. To see how the proposed system performs for untrained talkers, we split the data into 4 chunks with each chunk containing utterances from 1 talker. We then used a 4 fold-CV test i.e. training with data from 3 talkers and testing the data from the 4^{th} one. In this way, the system is tested for its robustness to each untrained talker. The average accuracy over the 4 test chunks was C_P = 0.8541, C_S = 0.8113 and RMSE = 0.1865. For comparison, we also computed the 4 fold CV results for the entire data (with random partitioning into 4 equal chunks) and found C_P = 0.8536, C_S = 0.8267 and RMSE= 0.1745. Thus, the two results obtained via different data partitioning (with the size of training and test sets being equal in both the cases) are quite close. This suggests that the proposed system performs well for untrained (unknown) talkers. There are other possibilities like training with only male talkers and testing the data from female talkers and vice versa. However, we found that the prediction accuracy were similar to the aforesaid 4 fold CV test. For that reason we do not include those results in this thesis.

We also tested our metric on a database with noise suppressed speech reported in [226] and we provide a brief description of this database. A sentence spoken by a male English speaker was corrupted using three background noise environments (car, factory, and train noises) at two levels of SNR (5 dB and 10 dB). The files were processed using eight speech enhancement algorithms. A total of 48 processed files were presented to 16 listeners for evaluation. Hence, each subject was required to rate the signals 144 times.

There are a total of 54 speech files (48 noise suppressed + 6 noisy). We refer the reader to [226] for details regarding the noise suppression schemes used in this database. This database is much smaller (in terms of the number of speech files and test conditions) than the first database used in this chapter. We used it to examine how our method performs given that the training data comes from the first database. For the proposed metric, we obtained $C_P = 0.6906$, $C_S = 0.6979$ and RMSE = 0.3603 while for P.563 $C_P = 0.5343$, $C_S = 0.5387$ and RMSE = 0.4243. So the proposed metric performs significantly better than P.563. One can however observe that the performance of proposed scheme as well as P.563 is relatively lower on this database. This could be possibly due to two reasons:

1. The subjective tests for this dataset may not have been performed in strictly controlled environment. As mentioned by its authors/developers, the subjective test was undertaken only to complement their objective evaluation tests. There was no calibration done for the headphone set. Additionally, the room conditions were not carefully controlled and other factors like external noises may not have been eliminated completely. In contrast to this, the subjective assessment tests for the first database (based on NOIZEUS) were done under more carefully controlled environment.
2. The new dataset uses only one clean speech file spoken by only one speaker (which was corrupted by 3 noise types at 2 SNR levels; these were processed by 8 speech enhancement algorithms to result in totally 48 processed speech files). Therefore, the content in the new dataset is quite limited. On the other hand, the first database uses 16 clean speech files (totally there are 1792 processed speech files). It is possible that more sentences in the new dataset might have given a clearer indication of the performance of the proposed scheme as well as P.563.

Nevertheless, the proposed scheme still outperforms P.563 by a relatively large margin for this database. Importantly, the training database is distinct from the test database as already pointed out.

In summary, the proposed scheme exhibits better performance in overall quality prediction for noise-suppressed speech in various test conditions.

8.3.4 Test Results for Signal and Noise Quality Assessment

As stated in Section 2.3.1, evaluating noise suppressed signals involves rating the signal quality (*SIG*), the background noise quality (*BAK*), and the overall quality (*OVRL*). It will be of further interest to devise an algorithm which is also capable of estimating the signal distortion and background distortion levels. Such estimates will provide more insights than merely predicting the overall quality. These can be useful in analyzing the performance of noise-suppression scheme(s) and to know how a particular scheme affects the noise corrupted signal. In the previous section, we presented the experimental results for the overall quality estimation. To evaluate how the proposed metric performs with regards to the prediction of *SIG* and *BAK* scores, we tested it by using *SIG* and *BAK* scores as the ground truth for training the SVR. One modification that we employ for predicting the *SIG* scores is the use of Voice Activity Detection (VAD). By using VAD, the signal is separated into active and inactive frames. We found that the prediction accuracy for *SIG* scores increased on using VAD (we employed the VAD from adaptive multi-rate (AMR) speech codec [227]).

The results for *SIG* and *BAK* prediction accuracies are presented in Figure 8.6 (a) and (b) respectively. We have also included the results for the method proposed in [9].

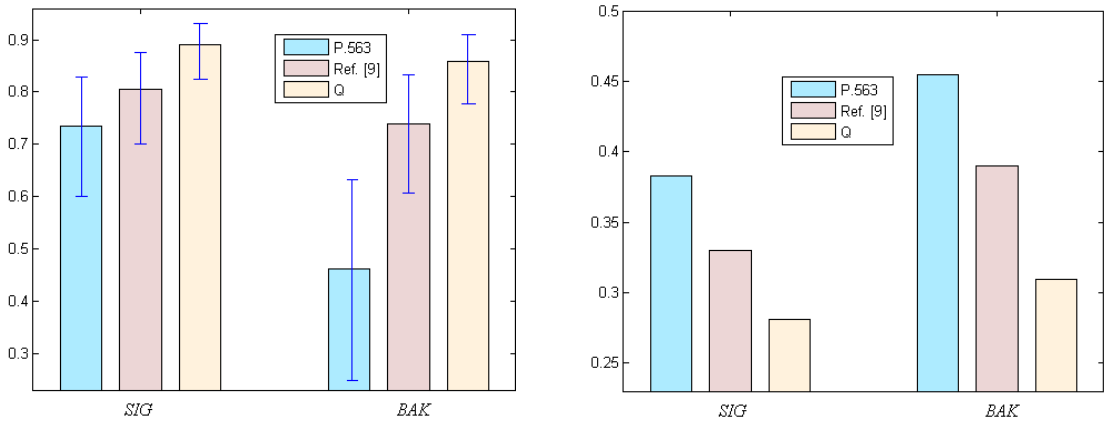


Figure 8.6: Results for 10 fold CV test for *SIG* and *BAK* scores (a) Comparison of C_P, (b) Comparison of RMSE (99% confidence interval bars are indicated in (a))

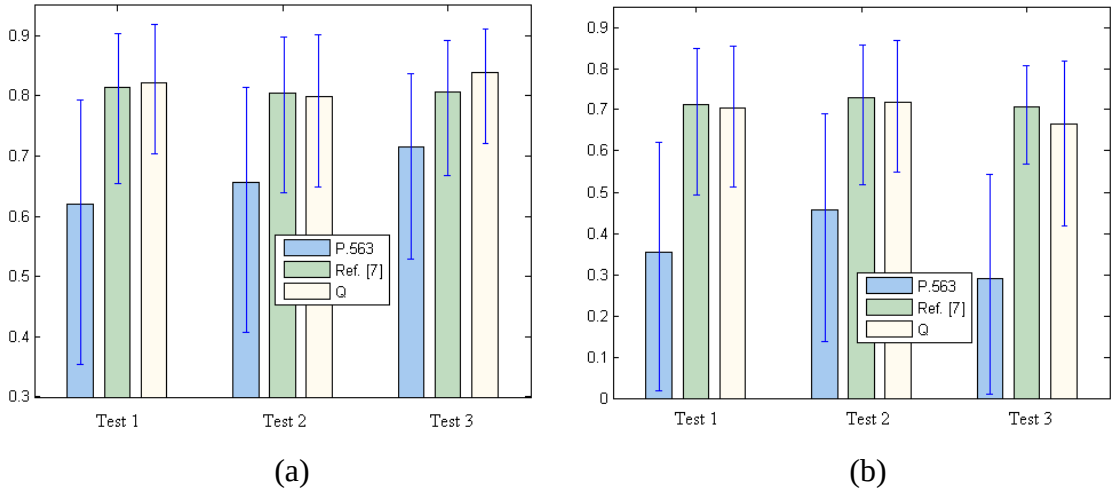


Figure 8.7: Results for proposed Q, P.563 and the method proposed in Ref. [7] with different splitting of data into training and test sets for *SIG* and *BAK* scores. (a) Comparison of C_P for *SIG* scores, (b) Comparison of C_P for *BAK* scores (99% confidence interval bars are also indicated)

Note that Ref. [9] has also reported the 10 fold CV results for *SIG* and *BAK* scores for the same database that we used in this chapter and so we have taken them directly from [9]. These results can be directly compared because [9] also employed the 3rd order polynomial mapping. We have omitted C_S values because they show a similar trend as C_P and RMSE. Also we do not include the results without the 3rd order polynomial

mapping as it does not have large effect on the prediction accuracies. In Figure 8.7, we further show the results for the 3 types of data partitioning (i.e. Test 1, Test 2 and Test 3 as discussed in the previous section). We include only the C_P values as C_S and RMSE values exhibit similar trends as C_P. It is informative to point out that PESQ achieves 0.81 and 0.76 in correlation [104] for *SIG* and *BAK* scores respectively. We find that the proposed scheme performs better than P.563 and the method proposed in [9] in both the cases (for *SIG* and *BAK* scores) for the 10 fold CV test. It achieves higher C_P and lower RMSE, indicating better alignment with the subjective viewing scores. Likewise, we can see from Figure 8.7, it gives significantly higher prediction accuracies for Test 1, Test 2 and Test 3 as compared to P.563. Our method also performs competitively (even slightly better in some cases) with the one proposed in Ref. [7]. This is significant given that the method in Ref. [7] uses the noisy and noise-suppressed signal for quality prediction, while the proposed metric being non-intrusive uses only the noise suppressed signal. The reader will also observe that *SIG* prediction accuracy is usually better than *BAK* score prediction accuracy. This is not surprising given that the proposed metric uses signal features only and we do not employ any additional features/parameters related to background noise. Of course the signal features (mean and variance of log FBEs in our metric) can indirectly account for the noise distortion to some extent. Another observation is that the prediction accuracies obtained are quite close for *SIG* and *OVRL* cases with $C_P = 0.9002$ and $C_P = 0.8968$ respectively for the 10 fold CV case. This trend is also observed in other test cases. A similar conclusion was also arrived at in [104] where it was found that intrusive algorithms like PESQ, LLR (Log-likelihood ratio) measure, frequency-weighted segmental SNR etc. predict signal and overall quality with similar accuracies but are less accurate in predicting background noise quality. This can

be explained from the observation [104], [224] that listeners are more sensitive to signal distortion than background distortion when making judgments on overall quality. This suggests that signal quality has more effect on the subjects when they judge the overall quality as compared to the noise distortion.

To further confirm this, we computed the correlation between the three quality scores and found that it was 0.5818 between *SIG* and *BAK*, 0.7793 between *OVRL* and *BAK* and 0.9505 between *SIG* and *OVRL* scores. It is clear that signal quality and overall quality follow more similar trend resulting in higher correlation between the two. Furthermore the regression analysis presented in [224] also confirms that listeners seem to place more emphasis on the distortion imparted on the speech signal itself rather than on the background noise, when making judgments of overall quality.

8.4 Performance Evaluation with Subset of Features

8.4.1 Prediction performance with reduced features

The feature vector proposed in this work is a 2*M* dimensional vector defined in Eq. (8.7). With $M = 40$, there are 80 features per signal which is relatively high. We evaluated the performance of our method using a smaller number of features and this can be done in two ways. First, we experimented with only the mean as the feature and that will give a 40 dimensional feature vector. Likewise using only the variance will also result in a 40-dimensional vector. Finally, we used mean and variance together (i.e. 80 features). In Figure 8.8 (a) we show the C_P values for different training and test sets discussed earlier.

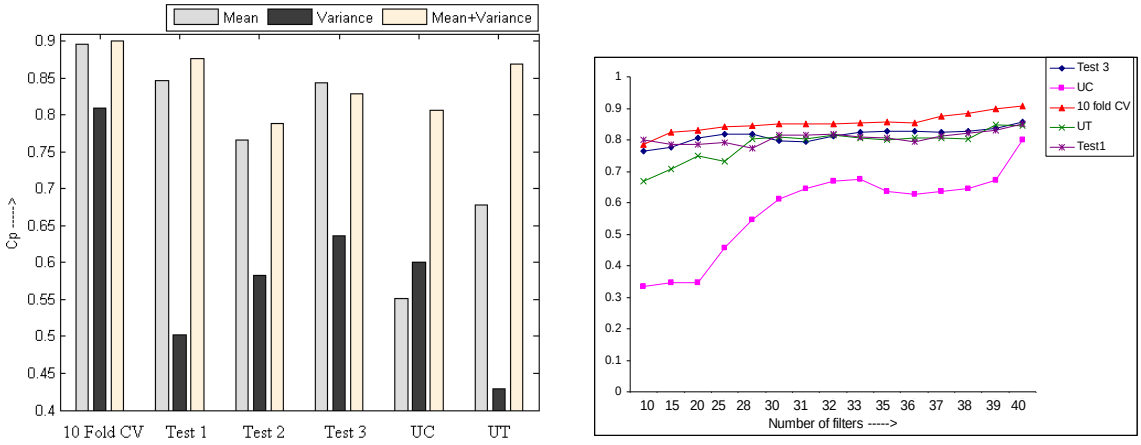


Figure 8.8: (a) Comparison of C_P for the different types of data partitioning. UC = Untrained Content, UT = Untrained Talker, (b) Performance variation (C_P values) with the total number of filters

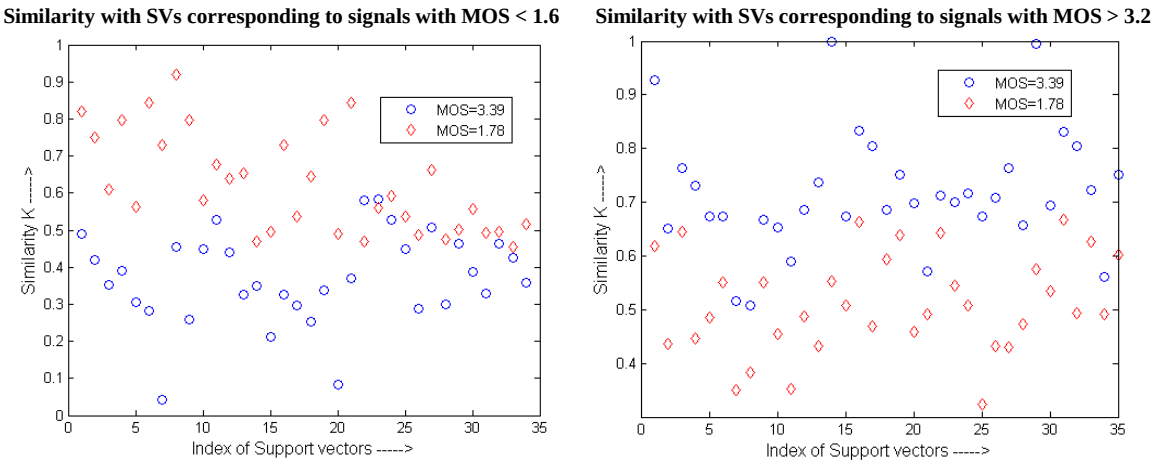
In this figure, the test with untrained contents is denoted with UC and UT denotes the test for untrained talker. The results show that variance alone gives the least accuracy, suggesting that it has the smallest overall contribution to the prediction performance. One can observe that mean alone in general gives reasonable accuracy and therefore can be useful if a lower dimensional feature vector is required, albeit with lower prediction accuracy. However, we find that mean and variance together overall gives the best performance. Although this implies a larger feature vector but is desirable for more robust and consistent performance.

The second way is to use only a subset of the filters to reduce the feature dimension. We show the variation of C_P values for the overall quality prediction with a different number of filters in Figure 8.8 (b). We include the results for Test 3, 10 fold CV, untrained contents and untrained talker for illustration. Other tests and *SIG* and *BAK* scores largely follow similar trends. The reader will observe that the best prediction accuracy is obtained at 40 filters. For the other tests also, a smaller number of filters

generally results in worse performance. Although a smaller number of filters may give good prediction performance for some tests, 40 filters (13 linear and 27 log spaced) achieves good performance in all the tests, and this suggests that the full filters lead to good discrimination between signals of varying qualities. Also even though the log FBEs are correlated and contain some redundant information, such redundancy is useful because the non-linear effects of noise-suppression can have larger impact on some FBEs while the same could be smaller on others. This is confirmed by the fact that the best performance in all the test conditions is achieved using all the 80 features.

8.4.2 Analysis of SVs

As already pointed out in Chapter 4, SVs are the data points which are relatively difficult for the SVR algorithm to fit within the ε-tube. We found that, in general, the samples which were chosen as the SVs had either very low quality score (MOS < 1.7) or very high quality scores (MOS > 3.2). This is a reasonable and intuitive selection of SVs since speech signals with very low or very high quality are the representative of the overall quality range. Similar to Chapter 5, we show an example to illustrate this point further. Consider two noise-suppressed speech files: one with MOS = 3.39 and the other with MOS = 1.78. We computed the kernel similarity scores $K(\boldsymbol{x}_i,\boldsymbol{x})$ by measuring their distances from the SVs $\boldsymbol{x}_i$ (0 indicates no similarity and 1 means completely similar). In Figure 8.9, we show the kernel similarity of the feature vectors for the two speech signals where the plot in (a) is the similarity scores with the SVs corresponding to relatively lower quality signals (MOS<1.6) while the plot in (b) shows the similarity with the SVs corresponding to relatively higher quality signals (MOS>3.2).



(a) (b)

Figure 8.9: Plots of kernel similarity scores

(a) Similarity scores of the signals with MOS = 3.39 (higher quality) and 1.78 (lower quality) with the SVs corresponding to lower quality signals (MOS<1.6), (b) Similarity scores for the two signals with the SVs corresponding to lower higher quality signals (MOS>3.2)

One can observe that the feature vector of the signal with MOS = 3.39 tends to have higher similarity with SVs corresponding to higher quality signals and lower similarity with SVs corresponding to lower quality signals. On the other hand, we observe opposite trend for the signal with MOS = 1.78. Examination of the corresponding scaling factors $(\eta_i^* - \eta_i)$ reveals that they are generally bigger and positive (which lead to higher score) for the SVs representing higher quality speech. In contrast, they are either small or negative (which lead to lower score) for the SVs corresponding to the lower quality speech. Because the final quality score is a summation of the "similarity" scores scaled by $(\eta_i^* - \eta_i)$, this results in assignment of higher score for signal with actual MOS = 3.39 and lower score for the signal with MOS = 1.78.

The number of SVs has a direct effect on the computational complexity of the learning algorithm since the weight vector is defined in their terms. In our experiments, we found

that the number of SVs were significantly smaller than the number of training points. For instance, for the 10 fold CV tests, on an average, 15 to 30% of the total number of training samples were chosen as the SVs. We found that the number of SVs decreased rapidly with increasing value of ε. This is due to the fact that more samples fall inside the ε-tube thereby reducing their number. It was also found that even when the number of SVs was made to as low as 5% (by increasing ε value) of the total number of training samples, the average prediction accuracy for the 10 fold CV test was $C_P = 0.8423$ which is comparable to the case when 15 to 30% datapoints are chosen as SVs (as shown in Figure 8.4 (a), the average $C_P = 0.8968$ in this case). This shows that the proposed SVR based scheme is efficient since the majority of the training examples can be safely ignored. In essence, the SVR focuses upon the small subset of examples that are important to predict the quality. We further noticed that as the value of C increases, the number of SVs also increases. This can be explained because C is the penalty for errors and is used to weigh the outliers. Obviously, as C is increased, the system tends puts a larger weight on the outliers.

8.4.3 Further Discussion

It may be pointed out that each data partitioning used in this chapter is meaningful since we exclude data from a certain category (for instance untrained SNR level, untrained noise suppression scheme, untrained content etc.) from the training set and test the excluded data. As mentioned, with SVR being a kernel method, it is more powerful and suitable for feature mapping in speech quality assessment. We also experimented with other less powerful techniques like multivariate adaptive regression splines but the performance was less satisfactory and also required larger processing time. With regards

to the use of different kernels, we observed that higher performance was achieved by using nonlinear kernels (like RBF, polynomial) as compared to the linear kernel, and this is expected due to the relationship of the input (FBEs) and output (quality score) variables. To be more precise, a major advantage of SVR that comes into picture due to the use of non-linear kernels is that the aforesaid "similarity" is measured in a new transformed space via the use of non-linear kernels. This enables the SVR algorithm to better distinguish/differentiate between signals of different qualities which may otherwise not be easily distinguishable in the original feature space. Due to this reason, the Gaussian kernel (or the RBF kernel) gives better results than the linear kernel. We also found that the sigmoid kernel gave the worst performance (correlation of 0.7256 for 10 fold CV test). This may be partly due to the fact that the sigmoid kernel may not be positive definite for certain situations and being positive definite is a requirement for a valid kernel function [166]-[167] and hence, once it ceases to be a valid kernel, the SVR doesn't perform well. Overall, the RBF kernel performs the best in terms of prediction accuracy and processing time required.

8.5 Concluding remarks

Nonintrusive speech quality assessment is a challenging problem since the measurement of quality has to be performed with only the output speech signal of the system under test, without using the original signal as reference. Furthermore, evaluating quality of noise-suppressed speech is an important but less investigated topic. In this chapter, we have presented a new method for nonintrusive quality assessment of noise-suppressed speech, by using mel-filter bank energies as features to capture signal variations, and SVR for feature mapping. We showed that noise injection and

suppression affects the FBEs and such changes (represented by the mean and variances) are also effective and parameterizable to assess quality. The advantage of SVR over other pooling methods comes due to the use of kernels which are advantageous for non-linear mapping problems. We have also given additional insights about quality prediction using SVR by analyzing the SVs obtained as a result of training. The proposed method has been validated on two speech databases with different contents and conditions. It performs better than P.563 and achieves higher correlation with the ground truth (i.e. subjective scores).

Chapter 9

Summary and Future Work

9.1 Summary

Signal quality assessment is either an important module in many multimedia processing systems (e.g. a visual quality metric can be embedded into a video encoder) or can be used as a standalone quality estimator. This thesis has presented new approaches for visual and speech quality assessment. To that end, we have investigated into the two aspects of developing a quality metric: feature detection and feature pooling. We have provided analysis and justification for the use of different signal features towards quality assessment. Further, we used machine learning for more systematic pooling of the features into a perceptual score.

In particular, we have developed and validated four new visual quality assessment schemes based on SVD, 2D mel-cepstrum and FT (we further used only the phase for scalability). The first two are FR methods which can be used to compute the visual quality when the reference signal is available. Our contribution lies in the analysis of the respective visual features which results in objective quality prediction that is better aligned with HVS's perception. The third method is attractive owing to its scalability: the degradation in its prediction performance is within reasonable limits with decreasing

reference signal information. As a result, it can also be employed for effective RR quality assessment. We further extended our SVD based method for VQA. This was achieved by using quality variation with time as the temporal factor. The advantage of this method lies in its lower complexity. We also proposed a new method for nonintrusive speech quality assessment based on mel FBEs and provided the appropriate theoretical and experimental analysis.

In summary, the methods proposed in this thesis were aimed at addressing some of the limitations of existing methods mentioned in Section 1.1. First, as demonstrated, the proposed algorithms give better prediction accuracy on a large number of images/videos (nearly 4000 distorted images) and distortion types/levels (more than 20 distortion types). Secondly, these algorithms perform more consistently for both near and suprathreshold distortions. Third, the proposed VQA algorithm has much lower computational overhead (as compared to many existing VQA schemes) in addition to objective predictions that are better aligned with the subjective opinion. Fourth, our method based on the phase of FT is scalable and thus more useful than FR or RR only schemes. Lastly, our speech quality assessment scheme being nonintrusive avoids the limitations of intrusive methods as mentioned in Section 1.1. This method was found to perform better than the current ITU standard P.563.

It is again worth emphasizing that all the methods developed in this thesis have been validated via comprehensive experimental analysis. This involved the usage of a large number of publicly and well accepted image, video and speech quality databases. Care was also taken in choosing the training and test sets so as to verify the prediction performance on untrained data (this was achieved via extensive CV and cross-database testing). The large scale use of subjectively rated databases in this thesis provides

convincing ground for the effectiveness of the methods proposed.

The major technical contributions of this thesis have been highlighted in Section 1.3, and in this section, we will give a summary of the actual research performed with links back into the preceding chapters.

9.1.1 Feature detection

Feature detection is an important step in developing a quality metric. Appropriate feature selection is a crucial step because the detected features would form the base of the resulting algorithm. Feature detection for quality assessment serves two purposes: (i) perceptually meaningful information is extracted, and (ii) it leads to dimension reduction i.e. the signal can be represented more compactly towards assessing its quality.

In Chapter 3, we explored SVD based features for visual quality assessment. There are two major advantages associated with these features: (a) the adaptively derived singular vectors allow better representation of image structure, (b) the separation of structure and luminance components enables more effective differentiation of their effects on perceptual quality. We also explained the process of computing visual quality using SVD based features using the necessary mathematical equations in Section 3.2.3. Additionally, we carried out analysis using F-tests to ascertain the statistical reliability of the results obtained. In addition, Chapter 7 explored the use of SVD based features for computing the temporal quality score towards VQA. This is based on the idea that larger fluctuations over time impact human judgement of video quality.

We then explored visual features based on the 2D mel-cepstrum. They are effective because the frequency information can be represented more compactly by using non-uniform weights. Another advantage is that they can account for the reduced sensitivity

at high levels of distortion, i.e. saturation effect (illustrated graphically in Figure 5.3). Lastly, in Section 5.4 we showed that the features based on 2D mel-cepstrum can be extracted more efficiently than the SVD based features and perform better with regards to the prediction accuracy.

Chapter 6 investigated into visual features based on the phase of FT. Although phase has been known to convey perceptually relevant image information, it has not been explored in details for visual quality assessment. We have mentioned in Sections 6.1 and 6.2 that a few existing works (such as [146]) have employed direct comparison between the phases of reference and distorted images. Such an approach however suffers from the limitation of ignoring the fact that not all changes in the visual signal have the same impact on quality. To tackle that, instead of using the similarity between the phase of reference and distorted images, we employed non-uniform binning of the frequency coefficients prior to phase extraction. This is based on the fact that error in lower frequency usually has larger impact on the visual quality. An important and unique advantage of the phase based algorithm lies in its scalability: the degradation in prediction performance is graceful with decreasing amount of reference image information. In contrast, to our knowledge none of the existing methods are scalable i.e. either they are FR or RR but not both.

We would also like to add some comments regarding the use of various transforms. In general, any 2-D transform decomposes the image into several basis images weighted by transformation coefficients. The SVD and DFT (2D mel-cepstrum is based on DFT but with additional processing) are based on orthogonal transforms. In that sense these transforms are related to each other. However, for the DFT, the basis vectors are fixed to be vectors based on trigonometric functions. In contrast, the basis vectors in SVD are

data dependent. These vectors are computed from the data to achieve optimality in reduce approximation error. But this also implies that we need to store the basis vectors in addition to the SVD coefficients if we want to reconstruct the time series. A particular drawback of DFT is that the basis vectors of DFT do not have compact support. This makes it very hard for DFT to approximate time series having short term bursts or jumps. SVD on the other hand deals with the problem of discontinuity in the time series data more gracefully. If a short term bursts or jumps are observed at the same location of most time series, it will be reflected by the basis vectors of SVD at that location. These are some of the key differences between SVD and DFT. With regards to the use of different transforms in this thesis, we have two conclusions for visual quality assessment:

- The basis vectors convey a more precise information regarding structural changes and hence should be more effective. Indeed we have discussed and demonstrated in Chapters 3 and 4, the effectiveness of the use of basis images (i.e. singular vectors) out of SVD in visual quality assessment. Further, Chapter 6 also uses phase for quality assessment. The 2D mel-cepstrum based transform is however inspired from speech processing (the use of Mel frequency cepstral coefficients) and is based on Fourier transform coefficients (as discussed in Section 5.3.4, in this case the phase cannot be used directly). As also pointed out in Section 3.2 of the thesis, there are several visual quality metrics which attempt to quantify visual quality by measuring the changes in transformation coefficients but these ignore the basis vectors which convey more precise information to evaluate quality objectively.
- We have argued and shown that the changes in basis images as well as the transformation coefficients should be used for the best results (for example

using singular vectors and values together as in Chapter 4 or using both phase and magnitude as in chapter 6).

The visual quality metrics based on SVD and 2D mel-cepstrum are FR methods. These can for example be used to adjust the parameters of image/video processing techniques in order to maximize visual quality or to reach a given quality in applications like image/video coding. These can also in general be employed for on-line monitoring of video quality in TV broadcasting, image/video compression (embedding the metric into say H.264/AVC), mobile communication systems (where speech suffers from noise and it is necessary to evaluate the impact of noise-suppression) and so on. The scalable metric developed in Chapter 6 for instance can also be used in scenarios with limited bandwidth. The low complexity of the method developed in Chapter 7 will be handy for it to be used in video processing tasks like compression, transmission etc.

Lastly, mel FBE based features were exploited in Chapter 8 to objectively evaluate the quality of noise-suppressed speech. We carried out both theoretical and experimental analysis (in Section 8.2.1) to show that noise injection and suppression affects mel FBEs. We further argued that mean and variance of mel FBEs can be used as global speech features for assessing quality. This is because speech signals carry information through time-domain variation; so FBE amplitudes at any given moment will be less meaningful than frame-to-frame variation. In other words, the distortion of FBEs will affect their pdfs which can be characterized by the displacement of mean and variance.

9.1.2 **Feature pooling**

Feature pooling is the second stage in quality metric design. We have first reviewed the limitations of the existing pooling schemes in Chapter 2 (Section 2.1.2.2). The major

problem with existing pooling schemes is that some of them (like simple averaging, Minkowski summation) tend to be over simplistic and ad-hoc. Likewise others such as those based on VA may be limited due to the fact that it is not easy to find regions of attention in an arbitrary image (not surprisingly VA is an active research area). Therefore, we have explored machine learning in order to pool/fuse the features more systematically. Although other techniques can be employed, we have used SVR in this thesis due to two reasons: (a) SVR is a well known kernel method and has been used widely in many other applications, (b) it employs a kernel for the non-linear mapping of the input data into higher dimensions thereby enabling it to achieve better distinction of different quality signals.

The pooling of SVD based features was discussed in Chapter 4 while Chapter 5 employed SVR for fusing the 2D mel cepstrum based features. SVR was also employed for combining the spatial and temporal factors for VQA in Chapter 7. Lastly, SVR was also used for pooling the mel FBEs for speech quality assessment. To provide convincing ground for the use of machine learning based pooling, thorough experimental analysis was done using a large number of image, video and speech signals. Because of the requirement of training, these experiments were carefully designed in order to show robustness to untrained data. To this end, extensive cross database validation results have also been reported. This is meaningful since content and distortion types vary across databases and thus help in proper metric verification.

One important point regarding the use of machine learning is that it can end up being a *black box* solution. In this thesis, we have further investigated the model obtained as result of training by analyzing the SVs, and noticed that the function $f(\boldsymbol{x})$ in Eq. (4.12) is a linear combination of Gaussian functions scaled by a factor of $(\eta_i^* - \eta_i)$. Hence, by

using SVR, we attempt to approximate the desired mapping function via a combination of Gaussian functions. In fact, the kernel function $K(\boldsymbol{x_i}, \boldsymbol{x})$ can be interpreted as the distance (or measure of similarity) between the i^{th} SV $\boldsymbol{x_i}$ and the test vector $\boldsymbol{x}$ in the transformed space. We can interpret $K(\boldsymbol{x_i}, \boldsymbol{x})$ as the cosine of the angle between the two Gaussian functions centered on $\boldsymbol{x}_i$ and $\boldsymbol{x}$. It is also easy to see from Eq. (4.12) that the predicted value is a weighted sum of the distances (or "similarities") between all the SVs and test vector $\boldsymbol{x}$. Due to this, SVs are the critical points with regards to SVR learning and their analysis can help in obtaining additional insights into the way the trained system predicts quality. As explained in Sections 4.4.7, 7.3.2 and 8.4.2, we found that the majority of the chosen SVs corresponded to data points with either very low or very high quality scores. This is intuitive because such data points cover the entire quality range and quality of the test signal can be determined by linear combination of the kernel similarity scores (after being scaled by an appropriate scaling factor).

9.2 Future work

This thesis has examined signal quality assessment by exploiting signal processing and machine learning approaches. We believe that there are several interesting avenues for further research to extend the current work.

The schemes developed for visual quality assessment (in Chapters 3-7) have not factored in color distortion. In the current studies, only the luminance information was utilized for quality computation. Therefore, investigation into how color distortion is perceived and interpreted by the HVS would be an interesting future work. A straightforward approach is to calculate the quality score for each color component (e.g., in RGB or HSV space) of the image/video frame, and then obtain the final quality score

through appropriate integration. However, such an approach is rather simplistic, and therefore, the color aspects deserve more careful and dedicated investigations.

Another aspect that could be an interesting future avenue is that of temporal quality computation for VQA. In Chapter 7, we employed the variation of quality as the temporal factor. However, more sophisticated models for computing the temporal distortion could be employed for better prediction accuracy (although this may contribute to increased computational complexity). An evidence of this is the existing VQA scheme MOVIE which is reasonably accurate but with very high computational requirements. Therefore, it would be a challenge to develop models for computing motion related distortion, in order to be both effective and computationally appealing.

We have developed a nonintrusive scheme for assessing quality of noise-suppressed speech in Chapter 8. It is worth mentioning that the current ITU standard for nonintrusive quality assessment P.563 works well for distortions due to codecs and communication channels but less accurate for noise-suppressed speech. It would be therefore interesting to embed our method with P.563 so that it can cater to speech suppression scenarios. In this way, we can enhance the capacity and scope of P.563.

Lastly, investigation into joint audiovisual quality assessment would be another attractive future direction. This is because more often than not humans perceive the ‘overall’ quality of the multimedia content rather than separate assessment of say video and audio. For example, a movie clip with very high video quality may not be enjoyable if sound quality is poor, and vice-versa. We believe that the problem of joint audiovisual quality assessment can be tackled using two broad approaches: (a) two-stage fusion (TSF) and (b) one-stage fusion (OSF).

In case of TSF approach, the first stage involves the fusion/pooling of the respective

audio and video features into overall audio and video quality scores respectively. At this stage, one treats audio and VQA as separate components. In the second stage, the two quality scores are fused/combined to obtain the overall audiovisual quality score. The problem of joint quality assessment for the TSF approach can be formulated as follows. If $v_1, v_2, ..., v_N$ denote the detected visual features (from video), the perceptual visual quality can be represented as:

$$Q_v = f_v(v_1, v_2, ..., v_N) \tag{9.1}$$

where the mapping function $f_v(.)$ can be determined via a *top-down*, *bottom-up*, or *hybrid* approach. Alternatively, machine learning techniques could also provide a solution in establishing the proper mapping function $f_v(.)$. Likewise, let $a_1, a_2, ..., a_N$ denote the audio features, the perceptual audio quality can be represented as:

$$Q_a = f_a(a_1, a_2, ..., a_N) \tag{9.2}$$

where $f_a(.)$ is the mapping function for audio/speech features.

The joint audiovisual quality model is then expressed as

$$Q_{av}^{(TSF)} = f_{av}(Q_a, Q_v, I_{av}) \tag{9.3}$$

where f_{av} is the required mapping function and I_{av} accounts for the interaction between the audio and video in terms of quality evaluation.

With the OSF approach, the overall quality impression is a result of analyzing all the relevant factors together, i.e. audio and visual features are tackled simultaneously. In this case, the problem of joint quality assessment can be formulated as

$$Q_{av}^{(OSF)} = g_{av}(v_1, v_2, ..., v_N, a_1, a_2, ..., a_N) \tag{9.4}$$

where g_{av} is the mapping function. One can see that in the above formulation, both audio and video features are considered jointly.

There are three issues that need further research: (a) the types of audio and visual features to be used, (b) how to combine them effectively especially when their interactions also need to be accounted for, and (c) to study which of the two (TSF or OSF) is the better integration methodology. Another crucial aspect of joint audiovisual quality assessment is setting up of subjectively rated databases to enable investigation of the problem. To the best of our knowledge currently there are no publicly available databases for the said task. The amount of effort, cost and expertise required in setting up such databases has also hindered progress in joint quality assessment. It will therefore be interesting to work towards the mentioned aspects to advance research in joint audiovisual quality assessment.

Appendix

Visual Database Description

In this thesis, we have used a total of 8 publicly available image databases and 3 video databases (all of them have been subjectively rated). We provide brief description for each database and refer the reader to the cited reference for more detailed information.

The LIVE image database [115] includes 29 original 24-bits/pixel color images. Totally it consists of 982 images (779 distorted images and 203 reference images). Five types of distortions were introduced to obtain the distorted images: 1) JP2K compression, 2) JPEG compression, 3) WGN, 4) Gaussian blurring, and 5) Rayleigh-distributed bit stream errors of a JP2K compressed stream or Fastfading distortions (FF). Subjective quality scores for each image are available in the form of DMOS.

The IRCCyN/IVC subjective viewing database [116] consists of 10 original color images with a resolution of 512 × 512 pixels from which 235 distorted images have been generated, using 4 different processes: JPEG compression, JP2K compression, LAR coding, and blurring. Subjective evaluations have been performed in a normalized room with lighting conditions and display settings adjusted according to ITU recommendation BT.500-11. The viewing distance was set to six times the picture's height. A DSIS method has been used. Both distorted and original pictures were displayed sequentially.

The Toyama subjective database [117] contains 182 images of 768 × 512 pixels. Out of

all, 14 were original images (24 bit/pixel RGB) in each group. The rest of the images were JPEG and JP2K coded images (i.e. 84 compressed images for each type of distortion). Six quality scales and six compression ratios were respectively selected for the JPEG and JP2K encoders. Subjective experiments were conducted in a normalized room with low lighting conditions and display settings adjusted according to ITU-R BT.500.11. The viewing distance was set to four times the picture's height. Single stimulus absolute category rating (SSACR) method was used in these subjective experiments. The subjects were asked to provide their perception of quality on a discrete quality score that was divided into five and marked with the numerical value of adjectives: Bad (1), Poor (2), Fair (3), Good (4) and Excellent (5).

The CSIQ database [118] consists of 30 original images. The distorted images have been subjectively rated base on a linear displacement of the images across four calibrated LCD monitors placed side by side with equal viewing distance to the observer. The database contains 5000 subjective ratings from 35 different observers, and ratings are reported in the form of DMOS. Each original image in the database is distorted using six different types of distortions at four to five different levels of distortion. The distortions used in CSIQ are: JPEG compression, JP2K compression, global contrast decrements, additive pink Gaussian noise, and Gaussian blurring. This results in a total of 866 distorted versions of original images.

In the A57 database [119], 3 original images of size 512 × 512 are distorted with 6 types of distortions and 3 contrasts. These result in 54 distorted images (3 images × 6 distortion types × 3 contrasts). The distortion types used are: 1) quantization of the LH subbands of a 5-level DWT of the image using the 9/7 filters, 2) additive WGN, 3) baseline JPEG compression, 4) JP2K compression, 5) JP2K compression with the

Dynamic Contrast-Based Quantization algorithm of which applies greater quantization to the fine spatial scales relative to the coarse scales in an attempt to preserve global precedence, and 6) blurring. The subjective scores have been made available in the form of DMOS.

The Tampere Image Database (TID) database [120] involves 25 original reference color images (resolution 512 × 384) which have been processed by 17 different types of distortions: additive Gaussian noise, additive noise in color components, spatially correlated noise, masked noise, high frequency noise, impulse noise, quantization noise, Gaussian blur, image denoising, JPEG compression, JP2K compression, JPEG transmission errors, JP2K transmission errors, non eccentricity pattern noise, local block-wise distortions of different intensity, mean shift (intensity shift) and contrast change. There are 4 distortion levels and thus it consists of 1700 (25 × 17 × 4) distorted images; there are 100 images for each distortion type. Subjective quality scores are reported in the form of MOS.

The Wireless Imaging Quality (WIQ) database [121] consists of 7 undistorted reference images, 80 distorted test images, and quality scores rated by human observers that have been obtained from two subjective tests. In each test, 40 distorted images along with the 7 reference images were presented to 30 participants. The quality scoring was conducted using a DSCQS. The difference scores between reference and distorted image were then averaged over all 30 participants to obtain a DMOS for each image. The test images included in the WIQ database consist of wireless imaging artifacts, which are not considered in any of the other publicly available image quality databases.

Lastly, we used another publicly available image database [122]. It is different from all the databases discussed above, with respect to the distortion type since the distortion is

due to watermarking. It consists of 210 images watermarked in three distinct frequency ranges. The watermarking technique basically modulates a noise-like watermark onto a frequency carrier, and additively embeds the watermark in different regions of the Fourier spectrum. The subjective scores are reported as MOS.

We used video sequences from three publicly available video databases in this thesis. The first video database (we refer to it as the EPFL database) consists [70] of 6 original video sequences at CIF spatial resolution (352 × 288 pixels) encoded with H.264/AVC. For each encoded video sequence, 12 corrupted bit streams were generated by dropping packets according to a given error pattern. To simulate burst errors, the patterns have been generated at six different packet loss rates (0.1%, 0.4%, 1%, 3%, 5% and 10%) and two channel realizations were selected for each packet loss rate. The packet loss free sequences were also included in the test material, thus finally 78 video sequences were rated by 40 subjects. Subjective scores have been made available as MOSs.

The second video database we used is the LIVE video database [80]. It contains 150 distorted videos (obtained from 10 uncompressed reference videos of natural scenes) with spatial resolution being 768 × 432. The distorted videos have been obtained by using four distortion processes: (a) simulated transmission of H.264 compressed bit streams through error-prone wireless networks, (b) through error-prone IP networks, (c) H.264 compression, and (d) MPEG-2 compression. Each video was assessed by 38 human subjects and the subjective scores have been made available as DMOS.

The third video database used in this study is publicly available at [123] and we refer to it as the TUL database. It comprises of 8 reference video sequences at CIF spatial resolution encoded with H.264/AVC to result in 32 distorted video sequences. These were rated by 22 observers and the subjective quality scores have been made available as

MOS.

For reader's convenience, a brief summary of the major characteristics of the subjectively rated image and video databases used in this thesis is presented in Table A.

Table A: Major characteristics of the subjectively rated visual databases used in this thesis

	No. of reference images/videos	No. of distorted images/videos	No. of distortion types	Typical image/frame size	Subjective score format (Range)
LIVE	29	779	5	768 × 512	DMOS (0-100)
CSIQ	30	866	6	512 × 512	DMOS (0-1)
IVC	10	185	4	512 × 512	MOS (1-5)
Toyama	14	168	2	512 × 768	MOS (1-5)
A57	3	54	6	512 × 512	DMOS (0-1)
TID	25	1700	17	512 × 384	MOS (0-9)
WIQ	7	80	1	512 × 512	DMOS (0-100)
Watermarked image database	5	210	1	512 × 512	MOS (1-5)
LIVE video database	10	150	4	768 × 432	DMOS (0-100)
EPFL video database	6	78	1	352 × 288	MOS (1-5)
TUL video database	8	32	1	352 × 288	MOS (1-5)

References

[1] Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios. ITU-R Recommendation BT.601-7, 2011.

[2] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] ITU-T Rec. P. 862, “Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” Intl. Telecom. Union, 2000.

[4] ITU-T Rec. P. 863, “Perceptual Objective Listening Quality Assessment (POLQA) - An advanced objective perceptual method for end-to-end speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals,” Intl. Telecom. Union, 2010.

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

[6] Y. Huang, T. Ou, P. Su, and H. Chen, “Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric,” IEEE Trans. Circuits Syst. Video Technol., vol. 20, no.11, pp. 1614–1624, 2010.

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

[8] ITU. Specifications and alignment procedures for setting of brightness and contrast of displays. ITU-R Recommendation BT.814-1, 1994.

[9] T. Falk, H. Yuan and W. Chan, “Single-Ended Quality Measurement of Noise Suppressed

Speech Based on Kullback-Leibler Distances," Journal of Multimedia, vol. 2, no. 5, pp. 19-26, 2007.

[10] ITU. Subjective assessment of standard definition digital television (sdtv) systems. ITU-R Recommendation BT.1129-2, 1998.

[11] ITU. Worldwide unified colorimetry and related characteristics of future television and imaging systems. ITU-R Recommendation BT.1361, 1998.

[12] ITU. Subjective video quality assessment methods for multimedia applications. ITU-T Recommendation P.910, 1999.

[13] ITU. Methodology for the subjective assessment of the quality of television pictures. ITU-R Recommendation BT 500-11, 2002.

[14] B. Girod. What's wrong with mean-squared error? Pages 207–220. The MIT Press, 1993. In Digital Images and Human Vision, A.B. Watson (ed.).

[15] Z. Wang, A. C. Bovik and L. Lu, "Why is image quality assessment so difficult?" Proc. IEEE International Conference on Acoustics, Speech, & Signal Processing, pp. 3313-3316, 2002.

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

[17] W. Lin and C. Kuo, "Perceptual Visual Quality Metrics: A Survey," J. of Visual Communication and Image Representation, vol. 22, pp. 297-312, 2011.

[18] S. Winkler, "Perceptual video quality metrics—A review," in Digital Video Image Quality and Perceptual Coding, H. R.Wu and K. R. Rao, Eds. Boca Raton, FL: CRC Press, ch. 5, 2005.

[19] W. Lin and M. Narwaria, "Perceptual image quality assessment: Recent progress and trends," in Proc. SPIE, vol. 7744, p. 774403, 2010.

[20] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," IEEE Trans. Information Theory, vol. 20, no. 4, pp. 525–536, 1974.

[21] F. Lukas and Z. Budrikis, "Picture quality prediction based on a visual model," IEEE Trans. Communications, vol. 30, no. 7, pp. 679–1692, 1982.

[22] S. Daly, “The visible differences predictor: An algorithm for the assessment of image fidelity,” in Digital Images and Human Vision, A.B. Watson, Ed. Cambridge, MA: MIT Press, pp. 179–206, 1993.

[23] J. Lubin and D. Fibush, “Sarnoff JND Vision Model,” T1A1.5 Working Group Document #97-612, ANSI T1 Standards Committee, 1997.

[24] C. J. Lambrecht and O. Verscheure, “Perceptual quality measure using a spatio-temporal model of the HVS,” in Proc. SPIE Digital Video Compression: Algorithms and Technologies, vol. 2668, pp. 450–461, 1996.

[25] S. Winkler, “A perceptual distortion metric for digital color video,” in Proc. SPIE Human Vision and Electronic Imaging, San Jose, CA, January 23–29, vol. 3644, pp. 175–184, 1999.

[26] J. O. Limb, “Distortion Criteria of the Human Viewer,” IEEE Trans. On Systems, Man and Cybernetics, vol.SMC-9, no.12, pp. 778-793, 1979.

[27] W. Lin, L. Dong and P. Xue, “Visual Distortion Gauge Based on Discrimination of Noticeable Contrast Changes,” IEEE Trans. Circuits and Systems for Video Technology, vol.15, no. 7, pp. 900- 909, 2005.

[28] S. Karunasekera and N. Kingsbury, “A distortion measure for blocking artifacts in images based on human visual sensitivity,” IEEE Trans. on Image Processing, vol. 4, no. 6, pp. 713-724, 1995.

[29] P.G.J. Barten, Contrast Sensitivity of the Human Eye and Its Effects on Image Quality, SPIE, Bellingham, WA (1999).

[30] B.A. Wandell, Foundations of Vision, Sinear Associates, Sunderland, MA (1995).

[31] Z. Wang and A. Bovik. Modern Image Quality Assessment. Morgan & Claypool Publishers, 2006.

[32] D. Rouse and S. Hemami, “Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM,” in Proc. Western New York Image Processing Workshop, October 2007.

[33] Z. Wang and A. Bovik, “A universal image quality index,” IEEE Signal Processing

Letters, vol. 9, no. 3, pp. 81-84, March 2002.

[34] D. Tao, X. Li, W. Lu and X. Gao, “Reduced-Reference IQA in Contourlet Domain,” IEEE Trans. on Systems Man and Cybernetics (Part B), vol. 39, no. 6, pp. 1623-1627, 2009.

[35] H. Han, D. Kim and R. Park, “Structural Information-Based Image Quality Assessment Using LU Factorization,” IEEE Trans. Consumer Electronics, vol.55, no. 1, pp.165- 171, 2009.

[36] D. Kim and R. Park, “New Image Quality Metric Using the Harris Response,” IEEE Signal Processing Letters, vol. 16, no. 7, pp.616-619, 2009.

[37] A. Eskicioglu, A. Gusev, and A. Shnayderman, “An SVD-Based Gray-Scale Image Quality Measure for Local and Global Assessment,” IEEE Trans. on Image Processing, vol. 15, no. 2, pp. 422-429, 2006,.

[38] M. Narwaria and W. Lin, “Scalable Image Quality Assessment based on Structural Vectors,” in Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP’09), pp. 1-6, 2009.

[39] M. Sendashonga and F. Labeau, “Low complexity image quality assessment using frequency domain transforms,” Proc. Int. Conf. on Image Processing, pp. 385–388, 2006.

[40] X. Zhu and P. Milanfar, “Automatic Parameter Selection for Denoising Algorithms Using a No-Reference Measure of Image Content, ” IEEE Trans. on Image Processing, vol. 19, no. 12, pp. 3116-3132, 2010.

[41] A. Aznaveh, A. Mansouri, F. Azar and M. Eslami, “Image Quality Measurement Besides Distortion Type Classifying,” Optical Review, vol. 16, no. 1, pp. 30–34, 2009.

[42] A. Mansouri, A. Aznaveh, F. Azar and J. Jahanshahi, “Image Quality Assessment Using the Singular Value Decomposition Theorem,” Optical Review, vol. 16, no. 2, pp. 49–53, 2009.

[43] U. Engelke, M. Kusuma., H.J. Zepernick and M. Caldera, “Reduced-reference metric design for objective perceptual quality assessment in wireless imaging,” Signal Processing: Image Communication, vol. 24, pp.525-547, 2009.

[44] H. Sheikh and A. Bovik, “Image information and visual quality,” IEEE Trans. Image

Processing, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[45] D. Chandler and S. Hemami, “VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images,” IEEE Trans. on Image Processing, vol. 16, no. 9, pp. 2284-2298, 2007.

[46] C. Wee, R. Paramesran, R. Mukundan and X. Jiang, “Image Quality assessment by Discrete Orthogonal Moments,” Pattern Recognition, vol. 43, pp. 4055–4068, 2010.

[47] M. Liu and X. Yang, “Image Quality Assessment using Contourlet Transform,” Optical Engineering, vol. 48, no. 10, 107201, 2009.

[48] G. Chen, C. Yang, and S. Xie, “Gradient-based structural similarity for image quality assessment,” in Proc. Int. Conf. Image Processing, pp. 2929–2932, 2006.

[49] G. Cheng, J. Huang, C. Zhu, Z. Liu, and L. Cheng, “Perceptual image quality assessment using a geometric structural distortion model,” in Proc. Int. Conf. Image Processing, pp. 325–328, 2010.

[50] K. Okarma, “Combined Full-Reference Image Quality Metric Linearly Correlated with Subjective Assessment,” in Proc. 10th International Conference on Artificial Intelligence and Soft Computing, Part I, LNAI, vol. 6113, pp. 539-546, 2010.

[51] K. Okarma, “Color Image Quality Assessment using the Combined Full-Reference Metric,” Advances in Intelligent and Soft Computing, Computer Recognition Systems 4, vol. 95, pp. 287-296, 2011.

[52] Z. Wang and X. Shang, “Spatial Pooling Strategies for Perceptual Image Quality Assessment,” Proc. of IEEE International Conferencing on Image Processing, (ICIP), pp. 2945-2948, 2006.

[53] A. Moorthy and A. Bovik, “Visual Importance Pooling for Image Quality Assessment,” IEEE J. of Selected Topics in Signal Processing, vol. 3, no. 2, 2009.

[54] A. Ninassi, O. Lemeur, P. Callet and D. Barba, “Does where you gaze on an image affect your perception of quality? Applying Visual Attention to Image Quality Metric,” Proc. of IEEE International Conferencing on Image Processing (ICIP), pp. 169-172, 2007.

[55] E. Larson., C. Vu, and D. Chandler, “Can visual fixation patterns improve image quality assessment?” Proc. of IEEE International Conferencing on Image Processing (ICIP), pp.

2572–2575, 2008.

[56] Q. Ma and L. Zhang, “Image Quality Assessment with Visual Attention,” Proc. Int. Conf. on Pattern Recognition (ICPR), pp. 1-4, 2008.

[57] U. Engelke, V. X. Nguyen, and H. Zepernick, “Regional Attention to Structural Degradations for Perceptual Image Quality Metric Design,” Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 869–872, 2008.

[58] J. You, A. Perkis, M. Hannuksela and M. Gabbouj, “Perceptual Quality Assessment Based on Visual Attention Analysis,” Proc. of ACM International Conference on Multimedia (MM’09), pp. 561-564, 2009.

[59] L. Karam, T. Ebrahimi, S. Hemami, T. Pappas, R. Safranek, Z. Wang, and A. Watson, “Introduction to the special issue on visual media quality assessment,” IEEE Journal on Selected Topics in Signal Processing, vol. 3, no. 2, pp. 189–192, 2009.

[60] “Methodology for the subjective assessment of the quality of television pictures,” ITU-R Recommendation BT.500-11.

[61] “Subjective video quality assessment methods for multimedia applications,” Sept 1999, ITU-T Recommendation-P.910.

[62] Z. Haddad, A. Beghdadi, A. Serir and A. Mokraoui, “Image Quality Assessment Based on Wave Atoms Transform,” Proc. Int. Conf. on Image Processing, 2010, pp. 305–308.

[63] L. Zhang, L. Zhang and X. Mou, “RFSIM: A Feature Based Image Quality Assessment Metric Using Riesz Transforms,” Proc. Int. Conf. on Image Processing, , pp. 321–324, 2010.

[64] S. Winkler, Digital Video Quality: Vision Models and Metrics, John Wiley and Sons, 2005.

[65] Q. Thu and M. Ghanbari, “Modelling of spatio–temporal interaction for video quality assessment,” Signal Processing: Image Communication, vol. 25, , pp. 535-546, 2010.

[66] K. Seshadrinathan and A. Bovik, “Motion Tuned Spatiotemporal Quality Assessment of Natural Videos,” IEEE Trans. on Image Processing, vol. 9, no. 2, 2010.

[67] A. Ninassi, O. L. Meur, P. Callet, and D. Barba, “Considering temporal variations of

spatial visual distortions in video quality assessment," IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 2, pp. 253–265, 2009.

[68] M. Barkowsky, B. Bialkowski, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 2, pp. 266–279, 2009.

[69] N. Suresh and N. Jayant, "'Mean time between failures': A Subjectively Meaningful Video Quality Metric," in Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 941-944, 2006.

[70] F. Simone, M. Naccari, M. Tagliasacchi, F. C. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective Assessment of H.264/Avc Video Sequences Transmitted Over A Noisy Channel," Proc. of IEEE International Workshop on Quality of Multimedia Experience, pp. 204-209, 2009.

[71] A. Webster, C. Jones, M. Pinson, S. Voran, and S. Wolf, "An Objective Video Quality Assessment System Based on Human Perception," in SPIE Human Vision, Visual Processing, and Digital Display IV, pp. 15-26, 1993.

[72] R. Hamberg and H. Ridder, "Continuous assessment of perceptual image quality," Journal of the Optical Society of America, vol. 12, pp. 2573–2577, 1995.

[73] M. Narwaria and W. Lin, "Objective Image Quality Assessment with Singular Value Decomposition," in Proc. Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM' 10), Jan. 13-15, 2010, Arizona, U.S.A.

[74] A. Hekstra, J. Beerends, D. Ledermann, F. de Caluwe, S. Kohler, R. H. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss, "PVQM—A perceptual video quality measure," Signal Process.: Image Commun., vol. 17, pp. 781–798, 2002.

[75] A. Moorthy and A. Bovik, "Efficient Video Quality Assessment Along Temporal Trajectories," IEEE Trans. on Circuits and Syst. for Video Technol, vol. 20, no. 11, pp. 1653-1658, 2010.

[76] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," Proc. of IEEE International Conference on Acoustics, Speech and Signal Process, pp. I869-I872, 2007.

[77] Z. Wang, L. Lu, and A. Bovik, “Video quality assessment using structural distortion measurement,” Signal Processing: Image Communication, vol. 19, no. 2, 2004, pp. 121-132.

[78] Z. Wang and Q. Li, “Video quality assessment using a statistical model of human visual speed perception,” Journal of Optical Society of America, vol. 24, no. 12, pp. B61–B69, 2007.

[79] M.H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” IEEE Transactions on Broadcasting vol. 50, no. 3, pp. 312–322, 2004.

[80] LIVE Video Quality Database, 2009. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html.

[81] VQEG, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II August 2003 [Online]. Available: http://www.vqeg.org.

[82] H. R. Sheikh and A. Bovik, “A visual information fidelity approach to video quality assessment,” in First International Conference on Video Processing and Quality Metrics for Consumer Electronics, 2005.

[83] S. Wan, F. Yang, X. Zhang, and C. Jiang, “Frame-Loss Adaptive Temporal Pooling for Video Quality Assessment,” in Visual Communications and Image Processing, Proc. SPIE, vol. 7744, p. 77440W-1, 2010.

[84] J. Lubin, “The use of psychophysical data and models in the analysis of display system performance,” in Digital Images and Human Vision, A.B. Watson, Ed. Cambridge, MA: MIT Press, pp. 163–178, 1993.

[85] J. Nachmias and R. Sansbury, “Grating contrast: Discrimination may be better than detection,” Vis. Res., vol. 14, no. 10, pp. 1039–1042, 1974.

[86] N. Jayant and P. Noll, “Digital Coding of Waveforms: Principles and Applications to Speech and Video,” Englewood Clifs, New Jersey: Prentice-Hall, 1984.

[87] R. Kubichek, D. Atkinson, and A.Webster, “Advances in objective voice quality assessment, ” in Proc. IEEE Globecom, 1991, pp. 1765-1770.

[88] S. Wang, A. Sekey, and A. Gersho, “An objective measure for predicting subjective quality of speech coders, ” IEEE Journal on Selected Areas in Communications, vol. 10, no. 5, pp. 819-829, June 1992.

[89] ITU-T Rec. P.861, “Objective quality measurement of telephone-band (300-3400 hz) speech codecs, ” Intl. Telecom. Union, Aug. 1996.

[90] S. Voran, “Objective estimation of perceived speech quality - Part I: Development of the measuring normalizing block technique, ” IEEE Trans. on Speech, Audio and Lang. Processing, vol. 7, no. 4, pp. 371-382, 1999.

[91] S. Voran, “Objective estimation of perceived speech quality - Part II: Development of the measuring normalizing block technique, ” IEEE Trans. on Speech, Audio and Lang. Processing, vol. 7, no. 4, pp. 383-390, 1999.

[92] W. Zha and W.-Y. Chan, “Objective speech quality measurement using statistical data mining, ” EURASIP Journal of Applied Signal Processing, vol. 2005, no. 9, pp. 1410-1424, 2005.

[93] T. Falk, W. Chan, and P. Kabal, “Speech quality estimation using Gaussian mixture models, ” in Proc. Intl. Conf. Spoken Lang. Proc., pp.2013-2016, 2004.

[94] S. Pennock, “Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm, ” in Proc. Intl. Conf. on Measurement of Speech and Audio Quality in Networks, Jan. 2002.

[95] S. Broom, “VoIP quality assessment: taking account of the edge device, ” IEEE Trans. on Speech, Audio and Lang. Processing., vol. 14, no. 6, pp. 1977-1983, 2006.

[96] ITU-T P.862.3, “Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2, ” Intl. Telecom. Union, 2005.

[97] O. Au and K Lam., “A novel output-based objective speech quality measure for wireless communication,” in Proc. 4th Int. Conf. Signal Process, 1998, vol. 1, pp. 666–669.

[98] T. Falk and W. Chan, “Nonintrusive speech quality estimation using Gaussian mixture models,” IEEE Signal Proc. Letters, vol. 13, no. 2, pp. 108-111, Feb. 2006.

[99] T. Falk, Q. Xu, and W. Chan, “Non-intrusive GMM-based speech quality measurement”, in Proc. Intl. Conf. Acoustics, Speech, Signal Processing, vol. 1, pp. 125-128, 2005.

[100] P. Gray, M. P. Hollier, and R. E. Massara, “Non-intrusive speech-quality assessment using vocal-tract models, ” in IEE Proc. Vision, Image and Signal Processing, vol. 147, no. 6, pp. 493-501, 2000.

[101] D. Kim, “ANIQUE: An auditory model for single-ended speech quality estimation,” IEEE Trans. on Speech, Audio and Lang. Processing, vol. 13, no. 5, pp.821-831, 2005.

[102] G. Chen and V. Parsa, “Bayesian model based non-intrusive speech quality evaluation,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, pp. 385-388, 2005.

[103] V. Grancharov, Y. David, L. Jonas, W. Bastiaan “Low Complexity Non Intrusive Speech Quality Assessment,” IEEE Trans. Speech Audio Process, vol. 14, no. 6, pp. 1948–1956, 2006.

[104] Y. Hu and P.C. Loizou, “Evaluation of Objective Quality Measures for Speech Enhancement,” IEEE Trans. Speech Audio Process, vol. 16, no. 1, pp. 229–230, 2008.

[105] S. Vaseghi, Advanced Signal Processing and Digital Noise Reduction, 2nd Ed., John Wiley & Sons Ltd, 2000.

[106] ITU-T Rec. P.835, “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms,”, Intl. Telecom. Union, 2003.

[107] ETSI EG 202 396-2: “Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise; Part 2: Background Noise Transmission-Network Simulation - Subjective Test Database and Results”.

[108] ETSI EG 202 396-3: “Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise Part 3: Background noise transmission - Objective test methods”.

[109] ITU. Subjective assessment methods for image quality in high definition television. ITU-R Recommendation BT.710-4, 1998.

[110] D. Kalman, “A Singularly Valuable Decomposition: The SVD of a Matrix,” The College Mathematics Journal, vol. 27, no. 1., pp. 2-23, 1996.

[111] G. Stewart, “Stochastic Perturbation Theory,” SIAM Review, vol. 32, no.4 pp. 579-610, 1990.

[112] J. Liu, X. Liu and X. Ma, “First Order Perturbation Analysis of Singular Vectors in Singular Value Decomposition,” IEEE Trans. on Signal Processing, vol. 56, no. 7, pp. 3044-3049, 2008.

[113] A. Targhi and A. Shademan, “Clustering of singular value decomposition of image data with applications to texture classification,” in Proc. SPIE Visual Communications and Image Processing, vol. 5150, pp. 972−979, 2003.

[114] H. R. Sheikh, A. C. Bovik, and G. de Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics,” IEEE Trans. Image Process., vol. 14, no. 12, pp. 2117–2128, 2005.

[115] H. Sheikh, K. Seshadrinathan, A. Moorthy, Z. Wang, A. Bovik, and L. Cormack, “Image and video quality assessment research at LIVE.” [Online]. Available: http://live.ece.utexas.edu/research/quality/.

[116] P. Le Callet and F. Autrusseau, Subjective Quality Assessment IRCCyN/IVC Database, http://www2.irccyn.ec-nantes.fr/ivcdb/

[117] Y. Horita, Y. Kawayoke, and Z. Sazzad, “Image quality evaluation database,” http://160.26.142.130/toyama_database.zip

[118] E. Larson and D. Chandler, “Most apparent distortion: full-reference image quality assessment and the role of strategy,” Journal of Electronic Imaging Vol. 19 no.1, 2010.

[119] A57 dataset: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html

[120] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola and F. Battisti, “Color Image Database for Evaluation of Image Quality Metrics,” Proc. of Intern. Workshop on Multimedia Signal Processing, Australia, pp. 403-408, 2008.

[121] U. Engelke, H. Zepernick, and M. Kusuma, “Wireless Imaging Quality Database, ” http://www.bth.se/tek/rcg.nsf/pages/wiq-db, 2010.

[122] F. Autrusseau, “Subjective quality assessment-Fourier Subband database” [online] Available: http://www.irccyn.ec-nantes.fr/~autrusse/Databases/FourierSB/, 2009.

[123] http://amalia.img.lx.it.pt/~tgsb/H264_test/

[124] J. Bretel, T. Caelli, R. Hilz, and I. Rentschler, ‘‘Modelling perceptual distortion: Amplitude and phase transmission in the HVS,’’ Human Neurobiology, vol. 1, no. 1, pp. 61–67, 1982.

[125] Y. Tadmor and D. Tolhurst, “Both the phase and the amplitude spectrum may determine the appearance of natural images,’’ Vision Res., vol. 33, no. 1, pp. 141–145, 1993.

[126] T. Huang, J. Burnett, and A. Deczky, “The importance of phase in image processing filters,” IEEE Trans. Acoust., Speech, Signal Process., vol. 23, no. 6, pp. 529–542, 1975.

[127] A. Oppenheim and J. Lim, “The Importance of Phase in Signals,” Proc. of the IEEE, vol. 65, no. 5, pp. 529-541, 1981.

[128] X. Ni, and X. Huo, “Statistical interpretation of the importance of phase information in signal and image reconstruction,” Statistics & Probability Letters, vol. 77, no. 4, pp. 447-454, 2007.

[129] P. Kovesi, “Phase Preserving Denoising of Images,” The Australian Pattern Recognition Society Conference: DICTA'99, pp.212-172, 1999.

[130] T. Alieva and M. Calvo, “Importance of the phase and amplitude in the fractional Fourier domain,” J. of Optical Society of America, vol. 20, no. 3, pp. 533-541, 2003.

[131] M.C. Morrone and D.C. Burr, “Feature detection in human vision: A phase-dependent energy model,” in Proc. Royal Society, London Series B 235, pp. 221.245, 1988.

[132] K. Gegenfurtner, D. Braun and F. Wichmann, “The importance of phase information for recognizing natural images,” J. of Vision, vol. 3, no. 9, 2003.

[133] B.C Hansen and R.F Hess, “Structural sparseness and spatial phase alignment in natural scenes,” J. of Optical Society of America, vol. 24, no. 7, pp. 1873-1885, 2007.

[134] W. Hsiao and R. Millane, “Effects of spectral amplitude and phase errors on image reconstruction,” Proc. SPIE 5562, 27, pp. 175-180, 2004.

[135] W. Hsiao and R. Millane, “Effects of Fourier-plane amplitude and phase errors on image reconstruction. I. Small amplitude errors,” J. Opt. Soc. America. A 24, pp. 3180-3188, 2007.

[136] M. Thomson, D. Foster and R. Summers, “Human sensitivity to phase perturbations in natural images: a statistical framework,” Perception, vol. 29, pp. 1057-1069, 2000.

[137] G. Blanchet, L. Moisan and B. Rouge, “Measuring the Global Phase Coherence of the Image,” Proc. of IEEE International Conference on Image Processing, pp. 1176-1179, 2008.

[138] C. Juglin, D. Hines, “The phase correlation image alignment method,” Proc Int. Conf. Cybernetics and Soc., pp. 163-65, 1975.

[139] K. Ito, T. Aoki, E. Kosuge, R. Kawamata and I. Kashima, “Medical Image Registration Using Phase-Only Correlation for Distorted Dental Radiographs**,”** Proc. of IEEE International Conference on Pattern Recognition, pp. 1-4, 2008.

[140] V. Ojansivu and J. Heikkilä, “Image Registration Using Blur-Invariant Phase Correlation,” IEEE Signal Process. Letters, vol. 14, no. 7, pp. 449-452, 2007.

[141] K. Ito, T. Aoki, H. Nakajima, K. Kobayashi and T. Higuchi, “A Palmprint Recognition Algorithm Using Phase-Based Image Matching,” Proc. of IEEE International Conference on Image Processing, pp. 2669-2672, 2008.

[142] C. Guo, Q. Ma and L. Zhang, “Spatio-temporal Saliency Detection Using Phase Spectrum of Quaternion FT,” Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[143] A. Sao and B. Yegnanarayana, “On the use of phase of the FT for face recognition under variations in illumination,” Signal, Image and Video Processing, vol. 4, no. 3, pp. 353-358, 2009.

[144] S. Mitra, M. Savvides and A. Brockwell, “Modeling Phase Spectra Using Gaussian Mixture Models for Human Face Identification,” Pattern Recognition and Image Analysis, LNCS, vol. 3687, pp. 174–182, 2005.

[145] S. Rajagopalan and R. Robb, “Phase based Image Quality Assessment,” Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment, Proceedings of SPIE, vol. 5749, pp. 373-382, 2005.

[146] P. Skurowski, and A. Gruca, “Image Quality Assessment Using Phase Spectrum Correlation,” Lecture Notes in Computer Science, vol. 5337, Computer Vision and

Graphics, pp. 80-89, 2009.

[147] K. Castleman, Digital Image Processing. New York: Prentice-Hall, 1996.

[148] Y. Fang, W. Lin, C. Lau and B. Lee, "A Visual Attention Model Combining Top-Down And Bottom-Up Mechanisms For Salient Object Detection," in Proc. Int. Conf. Acoustics, Speech and Signal Processing, pp. 1293-1296, 2011.

[149] N. Skarbnik, C. Sagiv, and Y. Zeevi, "Edge Detection and Skeletonization using Quantized Localized phase," Proceedings of 17th European Signal Processing Conference (EUSIPCO 2009). pp. 1542-1546, 2009.

[150] R. Kakarala, "Signal Processing Approach to Fourier analysis of ranking data: the importance of phase," IEEE Trans. on Signal Processing, vol. 59, no. 4, pp. 1518-1527, 2011.

[151] S. Cakir, and A. Cetin, "Mel-cepstral feature extraction methods for image representation," Optical Engineering, vol. 49, no. 9, 097004, 2010.

[152] D. Ghiglia and M. Pritt, "Two-dimensional phase unwrapping: Theory, Algorithms and Software", John Wiley & Sons, 1998.

[153] Z. Wang, E. Simoncelli, and A. Bovik, "Multi-scale structural similarity for image quality assessment," in Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, pp. 1398–1402, 2003.

[154] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," in Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Jan. 2007.

[155] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization based image representation," IEEE Journal of Selected Topics in Signal Processing: Special issue on Visual Media Quality Assessment, vol. 3, pp. 202-211, 2009.

[156] A. Rehman, and Z. Wang, "Reduced-reference SSIM Estimation," Proc. of IEEE International Conference on Image Processing, pp. 289-292, 2010.

[157] C. Warring and X. Liu, "Face Detection using Spectral Histograms and SVMs," IEEE Trans. on Systems Man and Cybernetics (Part B), Vol. 35, No. 3, pp. 467-476, 2005.

[158] C. Gruber, T. Gruber, S. Krinninger and B. Sick, “Online Signature Verification with Support Vector Machines Based on LCSS Kernel Functions,” IEEE Trans. on Systems Man and Cybernetics (Part B), vol. 40, no. 4, pp. 1088-1100, 2010.

[159] K. Huang, D. Tao, Y. Yuan, X. Li and T. Tan, “Biologically Inspired Features for Scene Classification in Video Surveillance,” IEEE Trans. on Systems Man and Cybernetics (Part B), vol. 41, no. 1, pp. 307-313, 2009.

[160] J. Tani, R. Nishimoto and M. Ito, “Codevelopmental Learning Between Human and Humanoid Robot Using a Dynamic Neural Network Model,” IEEE Trans. on Systems Man and Cybernetics (Part B), vol. 38, no. 1, pp. 43-59, 2008.

[161] ITU. Specification of a signal for measurement of the contrast ratio of displays. ITU-R Recommendation BT.815-1, 1994.

[162] P. Gastaldo, S. Rovetta, and R. Zunino, “Objective Quality Assessment of MPEG-2 Video Streams by Using CBP Neural Networks,” IEEE Trans. on Neural Networks, vol. 13, no. 4, pp. 939-947, 2002.

[163] P. Callet, V. Christian and B. Dominique, “A Convolutional Neural Network Approach for Objective Video Quality Assessment,” IEEE Trans. on Neural Networks, vol. 17, no.5, pp. 1316-1327, 2006.

[164] A. Bouzerdoum, A. Havstad, and Beghdadi, “Image quality assessment using a neural network approach,” Proc. of the Fourth IEEE International Symposium on Signal Processing and Information Technology, pp. 330-333, 2004.

[165] P. Carrai, I. Heynderickz, P. Gastaldo, R. Zunino and P. Monza, “Image quality assessment by using neural networks,” Proc. IEEE International Symposium on Circuits and Systems, vol.5, pp. 253-256, 2002.

[166] B. Scholkopf, A. J. Smola, “Learning with kernels,” MIT Press, 2002.

[167] J.Taylor and N. Cristianini, “Kernel Methods for Pattern Analysis,” Cambridge University Press, 2004.

[168] P. Bartlett, S. Boucheron and G. Lugosi, “Model selection and error estimation,” Machine Learning, pp. 85–113, 2002.

[169] C. Mantel, T. Kunlin and P. Ladret, “The Role of Temporal Aspects for Quality Assessment,” Proc. of IEEE International Workshop on Quality of Multimedia Experience, pp. 94-99, 2009.

[170] M. Gaubatz, “Metrix MUX Visual Quality Assessment Package,” http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[171] C. Chang and C. Lin, “LIBSVM: A library for support vector machines,” 2001 [Online]. Available: http://www.csie.ntu.edu.tw/\verb~cjlin/libsvm

[172] 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, 2010.

[173] L. Rabiner and B. Juang, Fundamentals of speech recognition, Prentice-Hall, Inc., Upper Saddle River, NJ, 1993.

[174] A. Oppenheim and R. Schafer, “From Frequency to Quefrency: A History of the Cepstrum,” IEEE Signal Processing Magazine, vol. 21, no. 5, pp. 95-106, 2004.

[175] V. O’Brien, “Contour perception, illusion and reality,” Journal of the Optical Society of America, vol.48, pp. 112-119, 1958.

[176] J.H. Elder, S.W.Zucker, “Evidence for boundary-specific grouping in human vision,” Vision Research, vol. 38, no. 1, pp. 143-152, 1998.

[177] R. L.De Valois and K. K. De Valois, Spatial Vision. New York: Oxford University Press, 1990.

[178] W. K. Pratt, Digital Image Processing: PIKS Inside, 3rd ed. New York: Wiley-Interscience, 2001.

[179] M. Narwaria and W. Lin, “Objective Image Quality Assessment Based on Support Vector Regression”, IEEE Trans. on Neural Networks, vol. 21, no. 3, pp. 515-519, 2010.

[180] D. Marr and E. Hildreth, “Theory of edge detection,” Proc. Royal Soc. London B, vol. 207, no. 1167, pp. 187-217, Feb. 1980.

[181] X.Ran and N.Farvardin, “A perceptually-motivated three-component image model—Part

I: description of the model," IEEE Trans. on Image Processing, vol. 4, no. 4, 1995.

[182] D. Marr, Vision, New York: W. H. Freeman and Company, 1980.

[183] C. Li and A. Bovik, "Content-partitioned structural similarity index for image quality assessment," Signal Processing: Image Communication, vol 25, pp. 517–526, 2010.

[184] L. Liang, S. Wang, J. Chen b, S. Mac, D. Zhao and W. Gao, "No-reference perceptual image quality metric using gradient profiles for JPEG2000," Signal Processing: Image Communication, vol. 25, pp. 502–516, 2010.

[185] Z. Sazzad, Y. Kawayoke and Y. Horita, "NR image quality assessment for JPEG2000 based on spatial features," Signal Processing: Image Communication, vol. 23, pp. 257–268, 2008.

[186] D. O. Kim, H. S. Han, and R. H. Park, "Gradient information-based image quality metric," IEEE Trans. Consumer Electronics, vol. 56, no. 2, pp. 930–936, 2010.

[187] D. Rouse and S. Hemami, "Natural Image Utility Assessment Using Image Contours," Proc. Int. Conf. on Image Processing, pp. 2217–2220, 2009.

[188] D. Chandler, S. Hemami, "Suprathreshold Image Compression based on Contrast Allocation and Global Precedence," Proc. Human Vision and Electronic Imaging, 2003.

[189] S. Hemami and M. Ramos, "Wavelet coefficient quantization to produce equivalent visual distortion in complex stimuli," in Human Vision and Electronic Imaging V, Proc. SPIE, Vol. 3959, pp. 200–210, 2000.

[190] H. Terasawa, M. Slaney and J. Berger, "A Timbre Space for Speech," Proc. of INTERSPEECH, pp. 1729-1732, 2005.

[191] J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?" Proc. of Int. Symposium on Music Info. Retrieval (ISMIR), 2002.

[192] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," IEEE Trans. on Speech and Audio Processing, vol. 10, no. 5, pp. 293–302, 2002.

[193] A Watson and L. Kreslake, "Measurement of visual impairment scales for digital video," Proc. SPIE, Human Vis., Vis. Process., and Digit. Display, vol. 4299, pp. 79–89, 2001.

[194] K. Lee, J. Park, S. Lee, and A. Bovik, “Temporal Pooling of Video Quality Estimates Using Perceptual Motion Models,” Proc. of IEEE International Conferencing on Image Processing, (ICIP), pp. 2493-2496, 2010.

[195] H. Nothdurft, “Salience from feature contrast: additivity across dimensions,” Vis. Res., vol. 40, no. 10–12, pp. 1183–1201, 2000.

[196] L. Ma, F. Zhang, S. Li, and K. Ngan, “Video Quality Assessment Based on Adaptive Block-Size Transform Just-Noticeable Difference Model,” Proc. of IEEE International Conference on Image Process, pp. 2501-2504, 2010.

[197] B. Nasersharif, A. Akbari, “SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features,” Pattern Recognition Letters, vol. 28, no. 11, pp. 1320-1326, 2007.

[198] E.H. Choi, “A Generalized Framework for Compensation of Mel-filterbank Outputs in Feature Extraction for Robust ASR,” in Proc. INTERSPEECH, pp. 933-936, 2005.

[199] B. Milner, J. Darch and S. Vaseghi, “Applying Noise Compensation Methods to Robustly Predict Acoustic Speech Features from MFCC Vectors in Noise**,”** in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, pp. 3945-3948, 2008.

[200] B. Milner and J. Darch, “Robust Acoustic Speech Feature Prediction from Noisy Mel-Frequency Cepstral Coefficients**,”** To appear in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 2, pp. 338-347, 2010.

[201] K. Onoe, H. Segi, T. Kobayakawa, S. Sato, T. Imai and A. Ando, “Filter bank subtraction for robust speech recognition,” in Proc. of International Conference on Spoken Language Processing, pp. 1021-1024, 2002.

[202] I. Almajai, B. Milner, J. Darch and S. Vaseghi, “Visually-Derived Wiener Filters For Speech Enhancement,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, pp. 585-588, 2007.

[203] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and Alex Acero, “A Minimum-Mean-Square-Error Noise Reduction Algorithm on Mel frequency Cepstra for Robust Speech Recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, pp. 4041-4044, 2008.

[204] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and Alex Acero, “Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor,” IEEE Trans. on Speech and Audio Process, vol. 16, no.5, 2008.

[205] A. Stark and K. Paliwal, “Use of speech presence uncertainty with MMSE spectral energy estimation for robust automatic speech recognition,” Speech Commun., vol. 53, no. 1, pp. 51-61, 2010.

[206] H. Cho and Y. Oh, “On the Use of Channel-Attentive MFCC for Robust Recognition of Partially Corrupted Speech,” IEEE Signal Processing Letters, vol. 11, no.6, 2004.

[207] C. Lee and S. Lee, “Noise-Robust Speech Recognition Using Top-Down Selective Attention with an HMM Classifier,” IEEE Signal Processing Letters, vol. 14, no.7, 2007.

[208] Michael L. Seltzer,and Richard M. Stern, “Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 2109-2121, 2006.

[209] Yuan-Fu Liao and I-Yun Xu, “Subband Minimum Classification Error Beamforming For Speech Recognition in Reverberant Environments,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, pp. 4702-4705, 2010.

[210] W. Li, J. Dines, M. Doss and H. Bourlard, “Neural network based regression for robust overlapping speech recognition using microphone arrays,” in Proc. INTERSPEECH, pp. 2012-2015, 2008.

[211] W. Li, J. Dines, M. Doss and H. Bourlard, “Non-linear learning for multi-channel speech separation and robust overlapping speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 3921-3924, 2009.

[212] W. Li, M. Doss, J. Dines, and H. Bourlard, “MLP-based Log Spectral Energy Mapping for Robust Overlapping Speech Recognition,” in Proc .16th European Signal Processing Conference (EUSIPCO-2008), 2008.

[213] T. Fingscheidt, S. Suhadi and K. Steinert, “Towards Objective Quality Assessment of Speech Enhancement Systems in a Black Box Approach,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, pp. 273-276, 2008.

[214] S. Davis, P. Mermelstein, “Comparison of parametric representation for monosyllabic

word recognition in continuous spoken sentences," IEEE Trans. Speech Audio Process, vol. 28, no. 4, pp. 357–366, 1980.

[215] K. Audhkhasi and A. Kumar "Two Scale Auditory Feature based Nonintrusive Speech Quality Evaluation," IETE Journal of Research , vol. 56, no. 2, 2010.

[216] Y. Gong, "Speech recognition in noisy environments: A survey," Speech Commun., vol. 16, no. 3, pp. 261–291, Apr. 1995

[217] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," IEEE Signal Processing Letters, vol. 9, pp. 113–116, 2002.

[218] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," IEEE Trans. Speech Audio Proc., pp. 334–341, 2003.

[219] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 2002.

[220] R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Biermann, M. Bush, M. Clements, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan,D. Novick, M. Ostendorf, S. Oviatt, P. Price , H. Silverman, J. Splitz, A. Waibel, C. Weinstein, S. Zahorian and V. Zue, "The challenge of spoken language systems: research directions for the nineties," IEEE Trans. on Speech and Audio Process, vol. 3, no.1, pp. 1-21, 1995.

[221] G. Chen and V. Parsa, "Nonintrusive Speech Quality Evaluation Using an Adaptive Neurofuzzy Inference System," IEEE Signal Process. Letters, vol. 12, no. 5, pp. 403-406, 2005.

[222] N. Pourmand, D. Suelzle, V. Parsa, Y. Hu, and P. Loizou, "On the use of Bayesian Modeling for Predicting Noise Reduction Performance," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, pp. 3873-3876, 2009.

[223] H. Hirsch, and D. Pearce "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," ISCA ITRW ASR2000, Paris, France, September 18-20, 2000.

[224] Y. Hu and P.C. Loizou, "Subjective comparison and evaluation of speech enhancement

algorithms," Speech Commun., vol. 49, pp. 588–601, 2007.

[225] C. Yang, "Inverted pattern approach to improve image quality of information hiding by LSB substitution", Pattern Recognition, vol. 41, pp. 2674–2683, 2008.

[226] T. Gunawan, E. Ambikairajah and J. Epps, "Perceptual speech enhancement exploiting temporal masking properties of human auditory system," Speech Commun., vol. 52, pp. 381–393, 2010.

[227] 3GPP2 TS 26.094, "Adaptive multi-rate (AMR) speech codec: voice activity detector (VAD), release 6," Dec. 2004.

[228] O. Schwartz and E. Simoncelli, "Natural signal statistics and sensory gain control," Nature Neuroscience, vol. 4, pp. 819 – 825, 2001.

[229] F. Petitcolas, R. Anderson, and M Kuhn, "Information hiding—a survey," Proc. of IEEE, vol. 87, no. 7, pp. 1062–1078, 1999.

[230] Z. Wang and A. Bovik, "MSE: love it or leave it? -A new look at signal fidelity measures," IEEE Signal Processing Magazine, vol. 26, no. 1, pp. 98–117, 2009.

[231] J.-H. Chen and J. Thyssen, "The broadvoice speech coding algorithm," in Proc.Intl. Conf. Acoustics, Speech, Signal Processing, vol. 4, pp. 537-540, 2007.

[232] E. Larson and D. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," Journal of Electronic Imaging, vol. 19 no.1, pp. 1-20, 2010.

[233] J. Robinson and V. Kecman, "Combining Support Vector Machine Learning With the Discrete Cosine Transform in Image Compression", IEEE Trans. on Neural Networks, vol. 14, no. 4, pp. 950-958, 2003.

[234] J. Lee, M. Kabrisky, M. Oxley, S. Rogers, and D. Ruck, "The complex cepstrum applied to two-dimensional images," Pattern Recogn,. vol. 26, pp. 1579–1592, 1993.

[235] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," IEEE Trans. Image Process., vol. 20, no. 5, pp. 1185–1198, May 2011.

[236] L. Zhang, L. Zhang, X. Mou and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," IEEE Trans. Image Processing, vol. 20, no. 8, pp. 2378-2386,

2011.

[237] M. Lin, S. Li, and K. Ngan, "Reduced-Reference Video Quality Assessment of Compressed Video Sequences", to appear in *IEEE Transactions on Circuits and Systems for Video Technology*.

[238] S. Li, M. Lin, and K. Ngan, "Full-reference Video Quality Assessment by Decoupling Detail Losses and Additive Impairments", to appear in *IEEE Transactions on Circuits and Systems for Video Technology*.

[239] S. Li, Z. Fan, M. Lin, and K. Ngan, "Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments", *IEEE Transaction on Multimedia*. vol. 13, no. 5, pp. 935-949, 2011.

[240] M. Lin, S. Li, Z. Fan, and K. Ngan, "Reduced-Reference Image Quality Assessment Using Reorganized DCT-Based Image Representation", *IEEE Transaction on Multimedia*, vol. 13, no. 4, pp. 824-829, 2011.

[241] Z. Fan, M. Lin, S. Li, and K. Ngan, "Practical Image Quality Metric Applied to Image Coding", *IEEE Transaction on Multimedia*, vol. 13, no. 4, pp. 615-624, 2011.

[242] M. Lin, S. Li, and K. Ngan, "Visual Horizontal Effect for Image Quality Assessment", *IEEE Signal Processing Letters*, vol. 17, no. 7, pp. 627-630, 2010.

[243] http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.shtml

Publications

Journal Papers

A. M. Narwaria, W. Lin, I. McLoughlin, S. Emmanuel and C. Tien, “Fourier Transform Based Scalable Image Quality Measure”, To appear *IEEE Trans. Image Process*., 2012.
(Chapter 6 in this thesis is based upon this paper)

B. M. Narwaria and W. Lin, “Low-Complexity VQA Using Temporal Quality Variations”, *IEEE Trans. on Multimedia, Special Issue on ICME 2011*,vol. 14, no. 3, pp. 525-535, 2012.
(Chapter 7 of the thesis is based upon this paper)

C. M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel and C. L. Tien, “Nonintrusive Quality Assessment of Noise Suppressed with Mel-Filtered Energies and Support Vector Regression”, *IEEE Trans .on Audio, Speech, and Language Process*., vol. 20, no. 4, pp. 1217-1232, 2012.
(Chapter 8 is based upon this paper)

D. M. Narwaria and W. Lin, “SVD-Based Quality Metric for Image and Video Using Machine Learning”, to appear in *IEEE Trans. on Systems, Man, and Cybernetics (Part B)*, vol. 42, no. 2, pp. 347-364, 2012.
(Parts of Chapters 3 and 4 in this thesis is based upon this paper)

E. M. Narwaria, W. Lin and A. Cetin, “Scalable Image Quality Assessment with 2D mel-cepstrum and machine learning approach”, *Pattern Recognition*, vol. 45, no. 1, pp. 299-313, 2012.
(Chapter 5 of this thesis is based upon this paper)

F. M. Narwaria and W. Lin, “Objective Image Quality Assessment Based on Support Vector Regression”, *IEEE Trans. on Neural Networks*, Vol. 21, No. 3, pp. 515-519, 2010.
(Parts of Chapters 3 and 4 in this thesis is based upon this paper)

Conference Papers

G. M. Narwaria and W. Lin, “Machine Learning Based Modeling of Spatial and Temporal Factor For VQA,” in *Proc. Int. Conf. Image Processing*, 2011.

H. M. Narwaria and W. Lin, “VQA Using Temporal Quality Variations and Machine Learning”, in *Proc. Int. Conf. Multimedia and Expo*, 2011.

I. W. Lin and M. Narwaria, “Perceptual IQA: Recent progress and trends,” in *Visual Communications and Image Processing*, *Proc. SPIE*, vol. 7744, p. 774403, 2010.

J. M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel and C. L. Tien, “Non-intrusive Speech Quality Assessment with Support Vector Regression”, In *Proc. 16th Int.* Conf. *on Multimedia Modeling, Lecture Notes in Computer Science*, vol. 5916, pp. 325-335, 2010.

K. M. Narwaria and W. Lin, “ Objective IQA with Singular Value Decomposition” in *Proc. Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM’ 10)*, Jan. 13-15, 2010, Arizona, U.S.A., 2009.

L. M. Narwaria and W. Lin, " Scalable IQA based on Structural Vectors" in *Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP'09)*, pp. 1-6, 2009.