ON ENHANCEMENT AND QUALITY ASSESSMENT OF AUDIO AND VIDEO IN COMMUNICATION SYSTEMS

Andreas Rossholm

Blekinge Institute of Technology
Doctoral Dissertation Series No. 2014:16

Department of Applied Signal Processing

On Enhancement and Quality Assessment of Audio and Video in Communication Systems

Andreas Rossholm

Blekinge Institute of Technology Doctoral Dissertation Series
No 2014:16

On Enhancement and Quality Assessment of Audio and Video in Communication Systems

Andreas Rossholm

Doctoral Dissertation in
Applied Signal Processing

Department of Applied Signal Processing
Blekinge Institute of Technology
SWEDEN

Department of Applied Signal Processing
Publisher: Blekinge Institute of Technology
SE-371 79 Karlskrona, Sweden
Printed by Lenanders Grafiska, Kalmar, 2014
ISBN: 978-91-7295-295-9
ISSN: 1653-2090
urn:nbn:se:bth-00604

"I dont feel that it is necessary to know exactly what I am. The main interest in life and work is to become someone else that you were not in the beginning. If you knew when you began a book what you would say at the end, do you think that you would have the courage to write it? What is true for writing and for love relationships is true also for life. The game is worthwhile insofar as we dont know where it will end."

Michel Foucault

Abstract

The use of audio and video communication has increased exponentially over the last decade and has gone from speech over GSM to HD resolution video conference between continents on mobile devices. As the use becomes more widespread the interest in delivering high quality media increases even on devices with limited resources. This includes both development and enhancement of the communication chain but also the topic of objective measurements of the perceived quality. The focus of this thesis work has been to perform enhancement within speech encoding and video decoding, to measure influence factors of audio and video performance, and to build methods to predict the perceived video quality.

The audio enhancement part of this thesis addresses the well known problem in the GSM system with an interfering signal generated by the switching nature of TDMA cellular telephony. Two different solutions are given to suppress such interference internally in the mobile handset. The first method involves the use of subtractive noise cancellation employing correlators, the second uses a structure of IIR notch filters. Both solutions use control algorithms based on the state of the communication between the mobile handset and the base station.

The video enhancement part presents two post-filters. These two filters are designed to improve visual quality of highly compressed video streams from standard, block-based video codecs by combating both blocking and ringing artifacts. The second post-filter additionally performs sharpening.

The third part addresses the problem of measuring audio and video delay as well as skewness between these, also known as synchronization. This method is a black box technique which enables it to be applied on any audiovisual application, proprietary as well as open standards, and can be run on any platform and over any network connectivity.

The last part addresses no-reference (NR) bitstream video quality prediction using features extracted from the coded video stream. Several methods have been used and evaluated: Multiple Linear Regression (MLR), Artificial Neural Network (ANN), and Least Square Support Vector Machines (LS-SVM), showing high correlation with both MOS and objective video assessment methods as PSNR and PEVQ. The impact from temporal, spatial and quantization variations on perceptual video quality has also been included, together with the trade off between these, and for this purpose a set of locally conducted subjective experiments were performed.

Preface

This doctoral thesis summarizes my work in the field of audio and video signal processing. The work has been carried out at the Department of Signal Processing at Blekinge Institute of Technology and some parts also with Ericsson Mobile Platforms AB. This thesis comprises four parts where the first two parts are in the field of audio and video enhancement, and the last two parts are in the field of video quality assessment using objective and subjective methods;

Parts

A On Audio Enhancement in Mobile Devices.

B On Video Enhancement in Mobile Devices.

C On Audio and Video Delay and Sync Measurement.

D On Audio and Video Quality Assessment.

Furthermore, parts of the work has been reported in a Licentiate thesis entitled "On the Enhancement of Audio and Video in Mobile Equipment", also published at BTH.

Acknowledgments

I wish to express my special thanks to Professor Ingvar Claesson, for his support and inspiration and for letting me start as a PhD candidate. Also, sincere gratitude to my friend and co-supervisor Dr. Benny Lövström for his guidance, support, and for all the interesting and constructive discussions. I'm both proud and happy that we managed to reach the goal despite changes in setup and jobs. I really hope we will continue our collaboration in the future.

Also, I want to thank my colleague Muhammad Shahid at BTH who became a close partner in the final parts of the thesis. Together, we have had tons of long discussion over Skype or Mobiles solving issues 24/7, it has sometimes been hard but mostly really fun.

I am thankful to Ericsson Mobile Platforms AB for making me an *industrial* PhD-student; Björn Ekelund for sanctioning it, Jim Rasmusson for his commitment, and John Philipsson for his interest and for allowing me to spend time on my research.

I also wish to thank my dear friend Per Rosengren who collaborated with me on my Master Thesis, which came to be the starting point for this research. Thanks also to Dr. Kenneth Andersson at Ericsson AB in Stockholm for his extensive support.

At Skype I would like to thank Dr. Sylvain Tourancheau and Dr. Mattias Nilsson for well considered inputs and helpful discussions. Also, I am very grateful for valuable inputs from Chris Owen

I thank all my old colleagues at both Ericsson and BTH for always giving me support and assistance and new colleagues at Skype for interesting discussions and thoughtful inputs during our *fika* pauses.

I also want to send my gratitude to Matt Berninger, Jim James, Thom Yorke, Damon Albarn, and Bernard Sumner and their colleagues for always keeping me company, and giving me strength and passion during my reading at buses and subways, writing or programming at cafes, or when I have sneaked away at home to the bedroom or kitchen just to get time to work on the thesis. Thanks for always being there for me.

Finally, I would like to thank my family for their support, and especially my loving wife Elisa for always encouraging me to believe in myself and supporting me through the whole process *"All I know is I'm loving you for all the right reasons"*.

Andreas Rossholm
Karlskrona, November 16, 2014

Contents

Publication List

Part A is published as:

I. Claesson and A. Nilsson (Rossholm), *GSM TDMA Frame Rate Internal Active Noise Cancellation.*, in International Journal of Acoustics and Vibration (IJAV), September 2003.

I. Claesson and A. Nilsson (Rossholm), *Notch Filtering of humming GSM mobile telephone noise.*, at International Conferences on Information, Communications and Signal Processing (ICICS), December 2005.

Parts of Part A has been published as:

I. Claesson and A. Nilsson (Rossholm), *Cancellation of Humming GSM Mobile Telephone Noise.*, at International Conferences on Information, Communications and Signal Processing (ICICS), December 2003.

Part B is published as:

A. Rossholm and K. Andersson, *Adaptive De-blocking De-Ringing Post Filter.*, at International Conference on Image Processing (ICIP), September 2005.

A. Rossholm, K. Andersson, and B. Lövström, *Low-Complex Adaptive Post Filter for Enhancement of Coded Video.*, at International Symposium on Signal Processing and its Applications (ISSPA), February 2007.

Publication related to Part B, but not included in the thesis:

U. Engelke, A. Rossholm, H.-J. Zepernick, and B. Lövström, *Quality Assessment of an Adaptive Filter for Artifact Reduction in Mobile Video Sequences.*, at IEEE International Symposium on Wireless Pervasive Computing, February 2007.

Part C is accepted as:

A. Rossholm, and B. Lövström, *A Robust Method for Estimating Synchronization and Delay of Audio and Video for Communication Services.*, accepted for

publication in Multimedia Tools and Applications (DOI: 10.1007/s11042-014-2306-6), October 2014.

Parts in D is published as:

A. Rossholm, and B. Lövström, *A New Video Quality Predictor Based on Decoder Parameter Extraction.*, at International Conference on Signal Processing and Multimedia Applications and Media (SIGMAP), July 2008.

A. Rossholm, M. Shahid, and B. Lövström, *Analysis of Impact from Temporal and Spatial Artifacts on Perceptual Video Quality.*, at IEEE Network Operations and Management Symposium (NOMS), 2014, May 2014.

Parts in D is submitted as:

A. Rossholm, M. Shahid, B. Lövström, *Comparison of Machine Learning Methods for Quality Estimation of Videos with Diversity in Temporal, Spatial, and Quantization Domains.*, submitted to EURASIP Journal on Image and Video Processing, Nov 2014.

Publications related to Part D, but not included in the thesis:

A. Rossholm, and B. Lövström, *A New Low Complex Reference Free Video Quality Predictor.*, at International Workshop on Multimedia Signal Processing (MMSP), October 2008.

M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, *No-reference image and video quality assessment: a classification and review of recent approaches.*, EURASIP Journal on Image and Video Processing, vol. 2014, no. 40, August 2014.

M. Shahid, A. Rossholm, and B. Lövström, *A no-reference machine learning based video quality predictor.*, in Fifth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 176 – 181, 2013.

M. Shahid, A. Rossholm, and B. Lövström, *A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor.*, at 4th International Congress on Image and Signal Processing, October 2011.

Additional publications:

M. Shahid, A. Rossholm, and B. Lövström, *A High Quality Adjustable Complexity Motion Estimation Algorithm For Video Encoders.*, at International Congress on Image and Signal Processing (CISP), October 2011.

M. Shahid, A. Rossholm, and B. Lövström, *A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor.*, at 4th International Congress on Image and Signal Processing, October 2011.

T. Minhas, M. Shahid, A. Rossholm, B. Lövström, H.-J. Zepernick, and M. Fiedler, *QoE Rating Performance Evaluation of ITU-T Recommended Video Quality Metrics in the Context of Video Freezes.*, accepted in the Australian Journal of Electrical & Electronics Engineering, 2014.

T. Minhas, M. Shahid, A. Rossholm, B. Lövström, H.-J. Zepernick, and M. Fiedler, *Assessment of the Rating Performance of ITU-T Recommended Video Quality Metrics in the Context of Video Freezes.*, at Australasian Telecommunication Networks and Applications Conference, October 2013.

M. Shahid, A. K. Singam, A. Rossholm, and B. Lövström, *Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos.*, at 5th International Congress on Image and Signal Processing (CISP), October 2012.

L. Spaanenburg, D. Zhang, M. Chen, and A. Rossholm, *Commanding the Cloud by Moving a Camera Phone.*, at International Journal of Handheld Computing Research (IJHCR), Vol. 1, Issue 3, 2010.

Patent applications have been filed in collaboration with Ericsson related to the different parts as follows.

Related to Part A:

A. Rossholm (Nilsson), I. Claesson, P. Rosengren, P. Ljungberg, J. Uden, P. Lakatos, *System and Method for Noise Suppression in a Communication Signal*, priority filed 3 Nov 1999, US patent US 6,865,276, granted 8 March 2005, EP patent EP 1228572, granted 31 Aug. 2005.

Related to Part B:

A. Rossholm, K. Andersson, *Adaptive De-Blocking De-Ringing Post Filter*, filed 22 Dec 2004, US Patent US 7,136,536, granted 14 Nov. 2006.

Related to Part D:

A. Rossholm, M. Pettersson, *Methods of and Arrangements for Processing an Encoded Bit Stream*, priority filed Mar 13, 2009 JP patent JP 5513532, granted 4 April 2014.
(Corresponding patent applications in other countries are still pending, published patent application no's are for ex. US2011/0299593 or WO2010/104432.)

A. Rossholm, M. Pettersson, *Technique for Video Quality Estimation*, filed Feb. 1, 2010, (P30339).
(All filed patent applications in various countries are still pending, published patent application no's are for ex. US2012/0281142 or WO2011/082719.)

Introduction

The development of audio and video communication during the last decades has been significant and the growth shows no tendency to decrease. In this context many different research fields interact and it has been in this environment the thesis has been created. In this introduction an overview of the audio and video communication field is given and the fundamental concepts are described. This includes an overview of the audio and video communication processing chain and its different parts including speech coding and transmission in GSM networks and fundamentals of video coding. Further, the definition and concept of Quality of Experience (QoE) will be discussed and an overview of definitions and categories will be given together with how different factors in the audio and video communication processing chain influence the QoE. Additionally, the area of QoE assessment will be discussed both addressing subjective and objective assessment methods. Finally an overview of the thesis is given including motivation, structure, summary and contribution, and a conclusion.

1 Audio and Video Communication

Audio and video communication usage is growing rapidly and its development is continuously ongoing. Regardless of whether the communication chain is looked upon from a straight speech call perspective or from an audio and video communication perspective, the processing chain can be represented by acquisition, compression, transmission over network, and reconstruction stages, where each stage includes fundamental parts enabling the communication, see illustration in Fig. 1. This processing chain can be applied to several audio and video communication systems. These can be audio or speech only, e.g. GSM and VoIP (voice over IP), or audiovisual, e.g. streaming or two-way real-time communication. When divided into the four processing parts many similarities are seen between them, the delay aspect, however, represents a major difference where the real-time aspects impacts the communication chain by the requirement of low latency or processing delay compared to e.g. streaming.

In the acquisition step of a typical application the audio signal is captured by the microphone and the video signal by the camera sensor. Both the audio and video can be pre-processed or enhanced by different algorithms. The audio part starts with digitalizing, analog-to-digital (A/D) conversion,

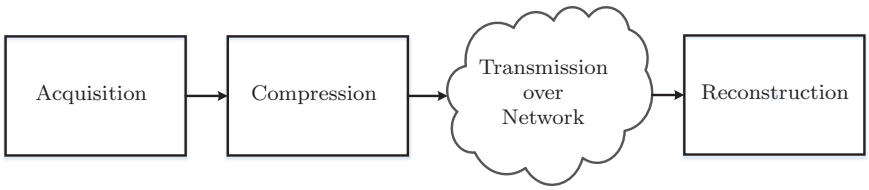


Figure 1: *Processing chain for audiovisual communication.*

and then includes different algorithms such as automatic gain control (AGC), active noise cancellation (ANC), and acoustic echo cancellation (AEC). For video the corresponding algorithms can be stabilization and noise reduction for simplifying the encoding or other adjustments to enhance the perceived quality. In this step the adaption to correct audio bandwidth and video spatial and temporal resolution, i.e. frame size and frame rate, is made.

In the next step compression is performed, where different codecs are used for different purposes and requirements. For audio one significant factor is the audio bandwidth, and a differentiation is seen between narrowband (NB), $0-3.5$kHz, wideband (WB), $0-7$kHz, and superwideband (SWB), $0-11$kHz. An overview of some of the most common present audio and video codecs is given in Table 1. Further in the compression stage consideration must be taken to both available network bandwidth and computational resources on the device. This is handled by rate controller, bandwidth estimator, and some kind of resource manager.

In the transmission stage the network connectivity and packetization is included, but also encryption to disable eavesdropping. Protection against packet loss, enabled with a resending mechanism or with forward error correction (FEC), should also be considered here. A major difference is seen between circuit switched (CS) and packet switched (PS) networks, where CS

Media	Codec Name	Standardized by	Note
Audio	GSM-HR, -FR, -EFR	ETSI(GSM)	NB
Audio	G.729A	ITU-T	NB
Audio	G.711	ITU-T	NB
Audio	AMR-NB	ETSI(GSM)	NB
Audio	AMR-WB(G722.2)	3GPP(ITU-T)	WB
Audio	AAC-family	MPEG	NB, WB, SWB
Audio	G.719	ITU-T	SWB
Audio	G.722	ITU-T	WB
Audio	Silk	IETF	NB, WB, SWB
Video	H.263	ITU-T	
Video	MPEG-4	ITU-T/MPEG	
Video	H.264(AVC)	ITU-T/MPEG	a.k.a MPEG-4 Part 10
Video	H.265(HEVC)	ITU-T/MPEG	a.k.a MPEG-H Part 2
Video	DivX	MPEG	following MPEG-4 Part 2
Video	Xvid	MPEG	following MPEG-4 Part 2, Advanced Simple Profile
Video	VP8	IETF	
Video	VP9	IETF	Draft Feb. 18, 2013

Table 1: Audio and Video Codecs

establish a dedicated communication channel before the communication starts while PS divides the data into packets before these are sent independently of each other over the network, meaning that the network links are shared instead of dedicated to only one session as is the case for CS. For network connectivity there are both cellular network standards like the ones from the 3rd Generation Partnership Project (3GPP) family e.g. GSM (Global System for Mobile Communications, originally Groupe Spécial Mobile), 3G and 4G (LTE) and wireless local area networks (WLANs) e.g. the IEEE 802.11 standards (WiFi), or wire-based communication technology e.g. the fixed telephone networks, cable television, and fiber-optics. The packetization is then

depending on whether the application is running CS or PS. For CS the transmission is controlled by specifications from e.g. GSM, 3G, and 4G while for PS the internet protocol suite is usually used, including SIP, UDP, TCP, and RTP, to enable the end-to-end communication over the network. For PS a differentiation of how the communication is organized can also take place where peer-to-peer, client-server or a hybrid of these are used, where the main difference is if the communication is controlled and transmitted via a server or not.

In the last part the reconstruction buffering and reordering (for PS) is enabled together with retransmitted packets or FEC before decompression is performed. After decompression error concealment algorithms are applied if data is lost and can not be reconstructed by FEC or resent. Also other post-processing algorithms can be applied to increase the perceived quality e.g. deblocking filter for video or noise reduction of audio. Then audio and video is rendered via loudspeaker and display.

1.1 Speech Coding and Transmission in GSM Networks

In the digital wire-line telecommunication system the analog speech signal is encoded by sampling and quantization which divides the signal into discrete time and discrete levels. This is simple and sufficiently effective for the wire-line system. In most digital speech encoders the speech signal is sampled at 8kHz resulting in a bandwidth of approximately 3400 Hz. However, mobile phones require a more effective encoder, since the transmission bandwidth is limited to 9.6 Kbps for GSM. In the 2nd generation cellular phone system GSM, the first introduced speech codec was a Regular Pulse Excitation with Long-Term Prediction (RPE-LTP). This speech codec, called *GSM full rate* [1], uses a speech producing model, consisting of spectral shape coding, excitation signal coding, and residual error coding. The speech producing model is created as a model of the human speech mechanism from the lungs, through the the vocal tract, including glottis and tongue, and the radiation of the lips. Since the speech organs usually change slowly, it is approximated that the filter parameters representing the speech organs are constant for 20 ms. Therefore, the speech codec processes frames of 20 ms at the time. Speech is acquired by the microphone and digitalized, with a sampling rate of 8kHz and 13 bits quantization. This means that 160 samples buffered to represent 20 ms. These samples are sent to the speech encoder, which compresses every frame of 160 samples to 260 bits. This will then be transmitted over the radio interface.

In the GSM system the transmission is performed on chunks of data, *bursts*. The sharing of the radio spectrum between several users, multiple access, is a mixed Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA) system. The modulation used is Guassian Minimum Shift Keying (GMSK). In the mixed TDMA and FDMA system the bursts (148 bits for a normal burst) are sent at a specific instant of time, *time slot*, where one time slot has a duration of 3/5200 seconds ($\approx$577μs) with a specific frequency. Time slots are organized in a cyclic fashion, in which the cycle can differ with different usage of the radio channel (data transport or signaling), as illustrated in Fig. 2. Eight time slots form a TDMA frame

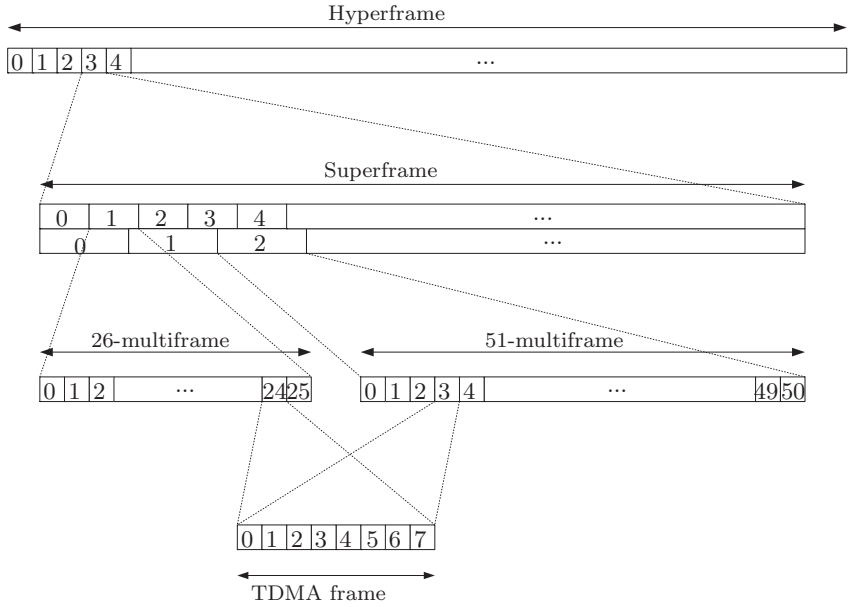


Figure 2: *The speech transmission model.*

(120/26 or 4.615 ms). The time slots within a TDMA frame are numbered from 0 to 7, and a particular time slot is referred to by its Time slot Number (TN). The TDMA frames are then numbered by a Frame Number (FN). There are two types of multiframes: a 26-multiframe (120 ms), consisting of 26 TDMA frames, used to support traffic and associated control channels,

and a 51-multiframe (3060/13 ms), consisting of 51 TDMA frames, used to support broadcast, common control and stand alone dedicated control (and their associated control) channels. A superframe is formed by 26×51 TDMA frames, and $26 \times 51 \times 2048 - 1 = 2715647$ TDMS frames forms a hyperframe (which is the longest cycle).

Since the standardization of the *GSM full rate*, several speech codecs have been adapted to the GSM specification; *GSM Half rate* [2] which doubles the capacity of the GSM system, *GSM Enhanced Full Rate* [3] with improved speech quality, *Adaptive Multi Rate (AMR)* [4] where an adaptation of the speech coding is performed based on the radio channel quality (also an improved speech quality).

1.2 Video Coding Basics

There have been many areas in which digital video technology has been applied over the last few years, from video recordings and playback to integration in applications as video conferencing and video streaming. Depending on user scenario and type of device the requirements differs when it comes to computational power, memory, and acceptable latency, for real time usage, as well as the limited bandwidth when radio transmission is requested.

A digital video sequence is generated when a series of images or frames of a real scene are sampled both temporally and spatially. This results in a large amount of data if no compression is made. Three fundamental steps are usually performed to increase the compression for a video sequence. The first step, performed before a frame is processed, is a color conversion from RGB to YCbCr, where Y is the luminance component and Cb and Cr represent color, or chrominance, difference for blue and red. Also, due to the fact that the human visual system is more sensitive to luminance than to color, the colors are represented with lower resolution. The second step is to exploit the high redundancy, correlation, between successive frames. The most common way to accomplish this is to use a similar principle to DPCM (Differential Pulse Code Modulation) where each sample or pixel is predicted from previous transmitted samples. This is achieved by calculating the difference between the actual pixel and the corresponding pixels in a previous frame and transmit this difference to the receiver. According to the typical temporal correlation this difference or prediction error will be small and have less energy to code. However, since the video scene most often includes motion, the DPCM is improved to compensate for this motion by translating or warping the samples of the previous frame to minimize the prediction error.

The third step to increase compression involves exploting the spatial redundancy, or high correlation, between pixels in the difference frame. The aim of the transform is to reduce this correlation by transforming the samples into visually significant transform coefficients and a large number of insignificant transform coefficients which can be discarded without decrease of the visual quality. In video context the second part where the temporal correlation is exploited the resulting frames are called *Inter frames*, denoted P or B, where P is a forward predicted frame and B is a bi-directional predicted frame. In the third part where spatial correlation is exploited, the resulting frames are called *Intra frames*, denoted I. In the intra frame, I, no prediction from surrounding frames, DPCM, is performed. All the video codecs in Table 1 are based on this concept which is often referred to as a hybrid Intra/Inter coding method.

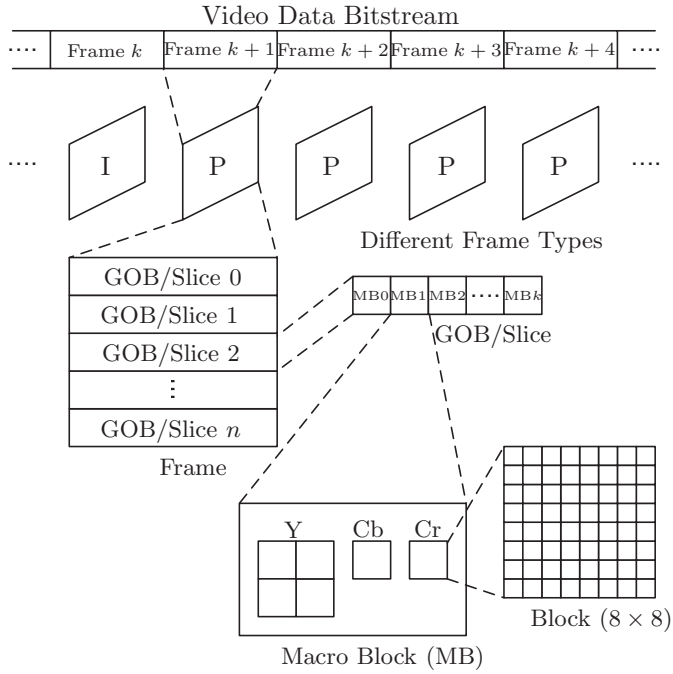


Figure 3: *A scheme over the basic layers in the video data stream.*

For both the temporal and the spatial compression the frame is subdivided

into smaller units before processing. Fig. 3 shows a scheme of the basic layers into which the frame is divided.

Depending on codec the smallest unit can be of different size. For example, in H.263 [5] and MPEG-4 Visual Simple Profile [6] the smallest units are blocks defined as a set of 8×8 pixels, while for some codec these can be further divided into even smaller blocks e.g H.264 [7] supporting 8×4, 4×8, and 4×4. As stated before, the chrominance has lower resolution and thereby each chrominance block corresponds to four luminance blocks and forms a Macro Block (MB). An integer number of MBs forms a Group Of Blocks (GOB) if the size and layout is fixed by a standard, or a slice (which does not have a fixed layout). GOBs are not used in H.263 or H.264. A number of GOBs or slices forms an I, P, or B frame.

A block diagram of a block-based hybrid Intra/Inter video codec is shown in Fig. 4. All standards mentioned follow this scheme whereby the different

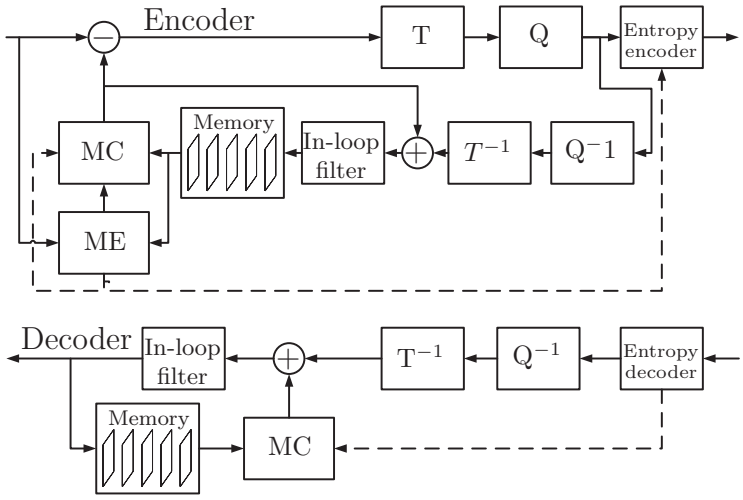


Figure 4: *A block diagram of a standard block based video coder.*

blocks are:

- **ME (Motion Estimation):** The block based ME compares a block in the current frame with blocks from the previous reconstructed frame to find the best match, i.e. minimize the residual. The residual is

calculated by subtraction with the original block after reconstruction by the MC.

- **MC (Motion Compensation):** The results from motion estimation are used to reconstruct the current block from a block from the previous frame.

- **T (Transform):** The most popular block-based transform is the Discrete Cosine Transform (DCT), which has low memory and computational requirements. Also, since it is block-based, it is well suited for block-based motion estimation. The transform is performed on the residual or an original block.

- **Q (Quantization):** The quantization is a lossy compression that reduces the amount of transform coefficients and lowers the precision. Thus, it decides the amount of compression obtained.

- **In-loop filter (Deblocking filter):** In-loop deblocking filter is mandatory for some standards e.g. H.264/AVC. This means that a filter is always applied on macro-block level to reduce blocking artifacts both on encoder and decoder side.

- **Memory:** The memory stores previously reconstructed frames for motion estimation/compensation.

- **Entropy coding:** The entropy coding algorithm is a lossless compression applied to meta data, as transform coefficients and motion vectors, to reduce the bitstream.

In scenarios where there is requirement to limit the size of the video sequences, due to e.g. limited radio bandwidth or limited computational power, these requirement can result in high compression. Thus, a high quantization is needed. When this is performed with a block-based hybrid Intra/Inter codec the video codec introduces artifacts. Two of the main artifacts are blocking and ringing. The blocking artifact is seen as an unnatural discontinuity between pixel values of neighboring blocking. The ringing artifact is seen as high frequency irregularities around the image edges. There are two main procedures to minimize these effects; for the standards enabling in-loop deblocking filter this will be used, and in absence of in-loop filter to detect and compensate for it, a post-filter can be used after the decoder, and thereby reducing the amount of high frequency variations in a controlled way.

2 Quality of Experience

With the increasing usage of communication technology systems and services delivering media to humans, the need for evaluation has become essential. This includes both media-to-human systems, e.g. streaming, and human-to-human systems, e.g. two-way audiovisual applications. The research around the concepts really took off in the early twenty-first century including researchers from different disciplines e.g. signal processing, telecommunications, psychophysics, and psychology. One of the outcomes from this work was construction of the term "Quality of Experience" used for evaluation of media transmission systems, services or applications and where the primary aim was to consider the perceived quality from the engineering's point of view. The most frequently used standardized definition is ITU-T's [11]:

Definition 2.1 (QoE (ITU-T))**.** *The overall acceptability of an application or service, as perceived subjectively by the end-user.*
Note 1: Includes the complete end-to-end system effects.
Note 2: May be influenced by user expectations and context

The definition of QoE differs from the concept of "Quality of Service" where an explicit view from the perspective of a system's or service's operator is considered, see definition of ITU-T [10]:

Definition 2.2 (QoS (ITU-T))**.** *Totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.*

However, there is a dependence between QoE and QoS, where QoS can be seen as a contributor to the potential QoE since it is reflecting the network performance. It should also be mentioned that in a comparison between QoE and User Experience (UX), which may occur is some contexts since there are similarities, one of the general differences is the relationship between QoE and QoS, which is mainly technology driven, compared to UX which is human-centered and have its origins from the field of Human-Computer Interaction [13].

In ITU-T's definition of QoE some possible impediments were found in the formulation. To address this and to be more specific around the context of users of applications and services a new definition of QoE was performed by the European Network on Quality of Experience in Multimedia Systems and Services, Qualinet. This started in 2011 and resulted in Qualinet White Paper on Definitions of Quality of Experience [12]:

Definition 2.3 (QoE (Qualinet))**.** *The degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the users personality and current state.*
Application: A software and/or hardware that enables usage and interaction by a user for a given purpose. Such purpose may include entertainment or information retrieval, or other.
Service: An episode in which an entity takes the responsibility that something desirable happens on the behalf of another entity. (Dagstuhl Seminar 09192, May 2009, cited after Möller, 2010)

When QoE or quality is looked upon, one fundamental area is the human perception. This includes the brain activity where some incidence of stimuli reaching one or multiple of human sensory organs, and how the stimuli converts into neural signals and transforms into more abstract or symbolic representations. This is the area of neuroscience and not covered further in this thesis, for detailed overview see [13].

Another area that has been studied is related to the factors that influence the QoE in the context of users of applications and services. In the work of Qualinet this has also been defined as [12]:

Definition 2.4 (Influence Factor (Qualinet))**.** *Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user.*

The influence factors (IF) can be grouped in three categories; Human IF, System IF, and Context IF, where the definitions are [12]:
Human IF (HIF): Any variant or invariant property or characteristic of a human user. The characteristic can describe the demographic and socio-economic background, the physical and mental constitution, or the users emotional state.
System IF (SIF): Refer to properties and characteristics that determine the technically produced quality of an application or service. They are related to media capture, coding, transmission, storage, rendering, and reproduction/display, as well as to the communication of information itself from content production to user.
Context IF (CIF): Factors that embrace any situational property to describe the users environment in terms of physical, temporal, social, economic, task, and technical characteristics.
These IFs must not be regarded as isolated as they may interrelate. Also,

it should be noted that the impact a specific IF has on QoE is not deterministic. Some examples of influence factors from the three categories could be for HIF e.g. gender, age, and expertise level, for CIF e.g. time of day, duration, costs of service, brand of the service/system, level of focus, alone or with other people, or technical interconnectivity. The SIF category is related to the technical part of producing QoE and has been divided further into four sub-categories; Content-related, referring to the content type and content reliability, Media-related, referring to media configuration factors, Network-related, referring to data transmission over a network, and Device-related, referring to the end systems or devices involved along the end-to-end communication path, including system specifications, equipment specifications, device capabilities and provider specification and capabilities. For a more detailed discussion see [13].

2.1 Influence Factors of Audio and Video Communication Chain on QoE

In section 1 an audio and video communication service or application was presented as a processing chain in a generic way and described divided into four different blocks, acquisition, compression, transmission over network, and reconstruction, see Fig. 1. A service or application can be influenced by factors having impact on the QoE, and can thereby be categorised according to the above presented groups. Further, in this context the decscribed processing chain can be categorised under SIF and its sub-categories, which are described below.

The acquisition block includes *content-related IFs* e.g. audio signal bandwidth and dynamic range, and video spatial and temporal information, *device-related IFs* e.g. the quality and performance of camera and microphone, introduced delay, and *media-related IFs* e.g. in the pre-processing steps of audio and video such as noise and echo cancelation, other factors are sampling rate, frame rate, and resolution resulting in aliasing, bandwidth limitation, and jerkiness, but also overall introduced delay in the process.

The compression block includes *media-related IFs* e.g. compression related influences such as blocking and ringing artifacts, but also factors from the rate controller performance, bandwidth estimator, and resource manager resulting in temporal and spatial scaling, aliasing artifacts and bandwidth limitation, and overall introduced delay in the process.

The transmission over network block includes *network-related IFs*. These could be generated in different ways depending on the type of network and

connection but some generic IFs would be e.g. bandwidth fluctuation and congestions, packet loss, jitter, and delay.

The reconstruction block includes *device-related IFs* e.g. the quality and performance of the display and its scaling, the loudspeaker, computational power to perform decoding, introduced delay, and *media-related IFs* e.g. the post-processing steps for enhancement and concealment like de-blocking and de-ringing, and overall introduced delay in the process.

Finally, the end-to-end perspective must also be taken into account. Here media-related IFs of delay and audio and video synchronization are significant factors, especially in a two-way audio and video communication scenario.

When the processing chain is divided into IFs as described it can be seen that some factors may originate from several places and in some cases it can be problematic to distinguish between these origins, e.g. for delay and for aliasing. Also, this highlights the importance to take the whole processing chain into account when QoE optimizations are taking place.

3 Video Quality Assessment

Video quality assessment (VQA) can be divided into subjective and objective VQA. This chapter starts with a brief presentation of subjective VQA, followed by a more detailed description on objective VQA.

3.1 Subjective Quality Assessment

The legitimate judges of visual quality are humans as end users, the opinions of which can be obtained by subjective experiments. Subjective experiments involve a panel of participants which are usually non-experts, also referred to as test subjects, to assess the perceptual quality of given test material such as a sequence of videos. Subjective experiments are typically conducted in a controlled laboratory environment. Careful planning and several factors including assessment method, selection of test material, viewing conditions, grading scale, and timing of presentation have to be considered prior to a subjective experiment. For example, Recommendation (ITU-R) BT.500 [8] provides detailed guidelines for conducting subjective experiments for the assessment of quality of television pictures. The outcomes of a subjective experiment are the individual scores given by the test subjects. These scores are used to compute mean opinion score (MOS) and other statistics. The obtained MOS, in particular, represents a ground truth for the development

of objective quality metrics. In ITU-R BT.500 and related recommendations, various types of subjective methods have been described. These types include either single stimulus or double stimulus based methods. In single stimulus methods, the subjects are shown variants of the test videos and no reference for comparison is provided. In some situations, a hidden reference can be included but the assessment is based only on a no-reference scoring of the subjects. In double stimulus methods a pair of videos comprising the reference video and a degraded video are presented twice and the subject rate the quality or change in quality between the two video streams. Similar procedure for multimedia applications are described in ITU-T P.910 [9] including absolute category rating (ACR) for single stimulus and ACR-H including hidden reference and degradation category rating (DCR) for double stimulus.

3.2 Objective Quality Assessment

Due to the time-consuming nature of executing subjective experiments, large efforts have been made to develop objective quality metrics, alternatively called objective quality methods. The purpose of such objective quality methods is to automatically predict MOS with high accuracy. Objective quality methods may be classified into psychophysical and engineering approaches [15]. Psychophysical metrics aim at modeling the human visual system (HVS) using aspects such as contrast and orientation sensitivity, frequency selectivity, spatial and temporal pattern, masking, and color perception. These metrics can be used for a wide variety of video degradations but the computation is generally demanding. The engineering approach usually uses simplified metrics based on the extraction and analysis of certain features or artifacts in a video but do not necessarily disregard the attributes of the HVS as they often consider psychophysical effects as well. However, the conceptual basis for their design is to do analysis of video content and distortion rather than fundamental vision modeling.

A set of features or quality-related parameters of a video are pooled together to establish an objective quality method which can be mapped to predict MOS. Depending on the degree of information that is available from the original video as a reference in the quality assessment, the objective methods are further divided into full reference (FR), reduced reference (RR), and no-reference (NR) methods as follows:

- FR methods: With this approach, the entire original video is available as a reference. Accordingly, FR methods are based on comparing a

distorted video with the original video.

- RR methods: In this case, it is not required to give access to the original video but only to provide representative features of the characteristics of the original video. The comparison of the reduced information from the original video with the corresponding information from the distorted video provides the input for RR methods.

- NR methods: This class of objective quality methods does not require access to the original video but searches for artifacts with respect to the pixel domain of an video, utilizes information embedded in the bitstream of the related video format, or performs quality assessment as a hybrid of pixel-based and bitstream-based approaches.

The remaining part of this chapter gives a further presentation of NR methods.

3.2.1 No-reference quality assessment

In recent years, there has been increasing interest in the development of NR quality assessment methods due to the widespread use of multimedia services in the context of wireless communications and telecommunication systems. Applications of NR methods include the following areas:

- Network operators and content providers have a strong interest to objectively quantify the level of service quality delivered to the end user and residing inside the network nodes. NR methods will provide the data needed to adopt network settings such that customer satisfaction is secured and hence churn can be avoided.

- The involvement of multiple parties between content providers and the end users gives rise to establishment of service-level agreements (SLA) under which an agreed level of quality has to be guaranteed. In this respect, NR methods are a suitable choice for in-service quality monitoring in live systems.

- Real-time communication and streaming applications can benefit from using NR methods for collecting information regarding delivered quality together with other statistics.

The current literature in the area of methods of NR video quality assessment is quite diverse. Hence, it is a challenging task to classify these methods

into a well-structured and meaningful categorization. A good categorization of such methods is given by Reibman et al. [16] who classify NR methods as either stemming from statistics derived from pixel-based features and call them NR pixel (NR-P) type or computed directly from the coded bitstream and call them NR bitstream (NR-B) type. This is a useful classification which can serve as an effective basis for constructing a broader classification, this is further elaborated in [14].

In the case of NR-P-based methods, one relevant method to classify available approaches is to investigate these in terms of the employment of certain artifacts that are related to a specific kind of degradation of the visual quality. Quantification of such artifacts has been used as a measure for the quality assessment. However, a NR-P-based method will require more computational power than the NR-B-based methods

The NR-B-based methods are relatively simpler to compute than NR-P-based methods, and the quality values can often be computed in the absence of a full decoder. However, such methods can have limited scope of application as they are usually designed for a particular coding technique and bitstream format, e.g., H.264/AVC standard. Such methods are based on either the encoding information derived from the bitstream or the packet header information, or a combination of both. These methods are quite suitable for network video applications such as realtime video communication and streaming.

The performance of NR-B-based methods for quality assessment can be improved by adopting an approach of adding some input from NR-P-based quality assessment. Such composites of NR-P- and NR-B-based methods are called hybrid methods. These methods inherit the computational simplicity of NR-B-based methods and depend on NR-P-related data to gain further robustness.

NR-B-based methods can be divided further into three categories based on the level of information used for processing, in accordance with the standardized models recommended by telecommunication standardization sector of International Telecommunication Union (ITU-T), as discussed in [17, 18]. This includes parametric models (parametric planning model and parametric packet-layer model) and bitstream layer model. In the former type, extrinsic features of a video that are of parametric nature such as bitrate, frame rate, and packet loss rate are used. Bitstream layer models have detailed access to the payload and intrinsic features related to a video such as coding modes, quantization parameter, and DCT coefficients. The standardization of these models includes the methods designed for estimation of audio quality as well, but is here in the following subchapters limited to video quality only. An

overview of the methods mapped tos audio and video communication chain is shown in Fig. 5.

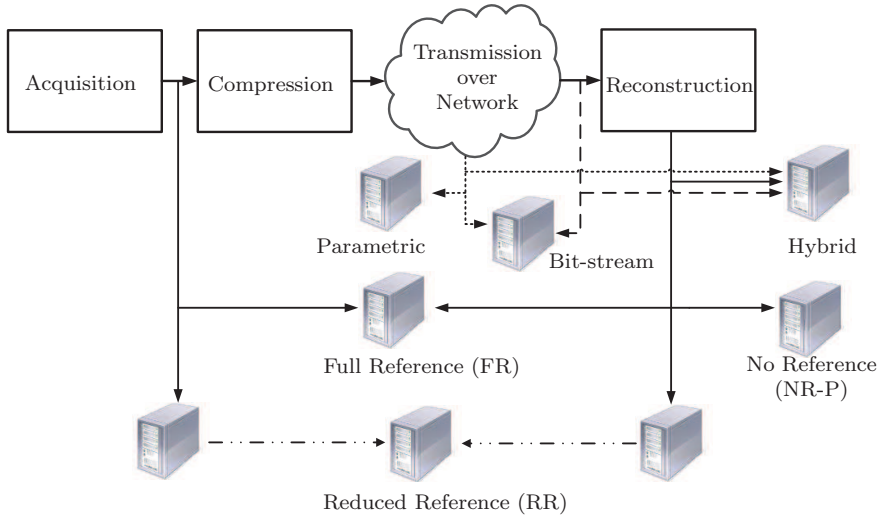


Figure 5: *Processing chain for audiovisual communication, visualizing classes of objective quality assessment methods.*

3.2.2 Parametric models

The parametric models can divided into planning and packet layer models. The parametric planning models have rather low complexity as they do not access the bitstream but utilize bitrate, codec type, and packet loss rate for making a crude estimation of video quality. The work item related to this category in ITU-T is known as Opinion model for video-telephony applications, G.1070 [19]. ITU-T Recommendation G.1070 proposes a method for the assessment of videophone quality, based on speech and video parameters, that can be used by the network performance planners to ensure the given level of end-to-end quality of the service.

The packet layer models have access to the packet header of the bitstream and can extract a limited set of parameters including bitrate on sequence or frame level, frame rate and type, and packet loss rate. Parametric packet-layer models are also known as QoS-based methods. The work item related to this category in ITU-T is known as non-intrusive parametric model for the assessment of performance of multimedia streaming (P.NAMS) [20].

3.2.3 Bitstream layer model

In the bitstream-based methods, bitstream layer models have access to most of the data that can be used for the video quality estimation. The work item parametric non-intrusive bitstream assessment of video media streaming quality (P.NBAMS) [21] in its mode 1 (Parsing mode) is related to the bitstream layer models. In this mode, it is allowed to do any kind of analysis of the bitstream except the usage of the pixel data. The input information includes parameters extracted from the packet header and payload. Besides the parameters included in the parametric models, this model uses quantization parameter (QP), DCT coefficients of the coded video, and pixel information. This makes the model comparatively more complex but it generally offers better performance.

Bitstream-based methods of video quality assessment have recently received a significant attention for their computational simplicity and applications in the online quality monitoring. Potentially, the main advantage of these methods is the variety in choice of the features which can be used for quality estimation, that in turn means the privilege of adapting to the desired level of complexity. Compared to pixel-based processing, the bitstream-based methods have the special advantage of having access to readily available information such as bitrate, frame rate, QP, motion vectors, and various types of information regarding the impacts of network impairments. However, these methods are coding scheme specific that makes them less generally applicable. In the case of parametric planning models, the performance of quality estimation remains limited due to the constraints of the information that can be obtained from the allowed level of access to the bitstream. Packet layer models have better performance with popular application in intermediate nodes of a network as they do not need complex processing and decryption of the data. Bitstream layer models are superior in the performance and the complexity can be flexible depending upon the desired level of accuracy.

3.2.4 Hybrid of NR-P and NR-B methods

There are no-reference video quality estimation methods which combine features from the coded bitstream and some statistics from the decoded media. This type of methods inherits the simplicity of computation from the bitstream-based approaches, and further accuracy in quality estimation is achieved by adding input from the pixel-based approaches. Therefore, such methods can avoid some of the difficulties involved in the pixel and bitstream-based methods [22]. One such example is the fusion of artifacts like blocking or blurring with parameters derived from motion vectors to build up a quality estimation method. The work item P.NBAMS [21] in its mode 2 (full decoding mode) is related to the hybrid models where the information from the coded bitstream as well as reconstructed video can be used.

4 Thesis Overview

In this overview a motivation to the thesis work is given together with a description of the structure of the remaining parts of the thesis. Also, a summary of each part is presented including their contribution to the thesis. Finally conclusions of the outcome are given.

4.1 Motivation

During the last decades the growth of audio and video communication has been significant on both computers and mobile devices. The amount of applications and user scenarios has emerged as a result of advances in new technologies and devices. During the last 20 years communication has gone from speech over GSM to HD resolution video conference between continents on mobile devices. In this development it has been motivated from an applied signal processing point of view to find ways to develop and enhance the communication chain. Also, with this huge increase in exposure of audio and video in different scenarios, the interest in delivering high quality media has increased naturally and thereby the topic of quantifying the quality has become a growing field. This can be seen from a perceived quality of experience (QoE) point of view where the overall quality is looked upon, or from different influence factors that has impact on the QoE.

In this dissertation the aim has been: 1. Remove the impact from audio interfering signal generated by the switching nature of TDMA cellular telephony, 2. Reduce the impact of blocking and ringing artifacts generated by

highly compressed video streams in mobile devices, 3. Develop a black-box method to measure audio and video delay and synchronization for audiovisual communication applications, 4. Find a no-reference method to predict perceptual video quality based on extracted video features from the encoded video stream.

In this work several methodologies have been used. The emphasis is on objective measures such as signal-to-noise ratio (SNR) and peak signal-to-noise ratio (PSNR), but also filter design, system design, subjective experience, subjective data analysis and objective modeling have been addressed and used.

4.2 Structure

In this thesis different areas related to audio and video communication has been addressed and these are structured as follows. Part A includes speech enhancement is GSM, and consists of two different methods for cancelation of interference to enhance audio quality. In Part B enhancement of a decoded video signal is presented where one specific de-blocking and de-ringing filter and one more generic postprocessing filter also including sharpening is presented. Part C addresses measurement of audio and video delay and synchronization and a black-box method is presented. In Part D of this thesis the focus is on NR-B-based video quality assessment, but a subjective quality assessment is also presented, which was performed and enabled the development of the last method.

4.3 Summaries and Contributions

In this sub chapter a summary is presented for each of the four parts of the thesis, and their contributions are given.

PART A - On Audio Enhancement in Mobile Devices

PART A.1 - GSM TDMA Frame Rate Internal Active Noise Cancellation

This part describes two different solutions designed to suppress the interfering signal generated by the switching nature of TDMA cellular telephony, where the radio circuits are switched on and off. The interfering signal is transmitted with the speech signal to the receiver. Due to the humming sound of the interfering signal, it is commonly denoted the *Bumblebee*.

Methods used include Notch Filtering, which is multiplicative in frequency domain, and subtractive Noise Cancellation, which is an alternative method employing correlators. The fundamental switching rate is approximately 217 Hz. Since the frequency components of the disturbing periodic humming noise are crystal generated and accurately known, it is possible to estimate the cosine- and the sine- parts of these with correlators. This is done by correlating the microphone signal with sinusoids with the same crystal generated frequencies as the disturbing frequencies. By generating the cosine- and sine-signals with correct signed amplitudes and then subtracting these from the microphone signal, the humming "Bumblebee" is almost perfectly suppressed in the microphone signal.

PART A.2 - Notch Filtering of Humming GSM Mobile Telephone Noise

Part A.2 proposes an alterative solution to the problem of an interfering signal generated by the switching nature of TDMA cellular telephony (addressed in Part A.1). This part proposes a dual cascaded notch filter solution, using internal knowledge of the GSM transmission pattern and transmitter state to suppress the interfering signal, *the Bumblebee*. Compared to A.1 this is a more memory efficient solution.

The basic idea is to use two notch filters, whereby one of the filters has a slightly larger notch-bandwidth. The first filter is only used to insert the distortion during the idle slot, which is the problem with a single notch filter since it consists of poles (autoregressive), which give feedback of the output signal continuously. These samples are then used to replace the samples in the original signal during the idle slot. The idle slot is located by using internal knowledge of the GSM transmission pattern and transmitter state. This results in the presence of the bumblebee signal during the idle slot. It therefore follows that the signal is periodic with the TDMA frame rate. The second filter is then used to notch the new signal with the periodic bumblebee. The reason for the difference in bandwidth is to make sure that we do not add any distortion that is not suppressed.

PART B - On Video Enhancement in Mobile Devices

PART B.1 - Adaptive De-blocking De-Ringing Post Filter

In Part B.1 an adaptive de-blocking and de-ringing post filter is proposed. This post filter is designed to improve visual quality of highly compressed

video streams from standard, block-based video codecs by combating both blocking and ringing artifacts. The proposed solution is designed with consideration of mobile equipments with limited computational power and memory. Also, the solution is computationally scalable if there are different limitations of CPU resources in different user cases. The filter consists of a reference filter that has coefficients that determine the filtering function, and these coefficients are selectively modified by the weight generator considering the amount of quantization in the frame, pixel location, if it is a block border or not, and where the filter strength in different scenarios can be changed. Part B.1 has been verified by implementation in large volumes of several models of mobile devices.

PART B.2 - Low-Complex Adaptive Post Filter for Enhancement of Coded Video

In Part B.2 an adaptive filter is presented that removes blocking and ringing artifacts and also enhances the sharpness of decoded video. Loss of sharpness may occur when zeroing high-frequency DCT coefficients in the encoder. This is a further development of Part B.1 using the same filter structure but with updated modification of the reference filter. Thus, the resulting filter characteristics can not only vary from weak to strong low-pass filtering depending on the reference filter output. In this design the resulting filter characteristics can also vary from weak to strong high-pass filtering, based on the output from the reference filter's and additional information e.g QP value, position of a pixel in the frame etc. Weak low-pass or all-pass filtering is implemented as in Part B.1. As a consequence, the proposed filter can achieve low-pass filtering as well as sharpening, depending on amount of compression location in the frame etc.

PART C - On Video Delay and Sync Measurement

PART C.1 - A Robust Method for Estimating Synchronization and Delay of Audio and Video for Communication Services

In part C.1 a method is proposed to estimate the influence factors for audio and video delay as well as synchronization. The method as such is an out-of-service or black box technique which enables it to be applied on any audiovisual application, proprietary as well as open standards, and can be run on any platform and over any network connectivity.

The method is using an audio and video reference stream, where audio and video frames are marked with frame numbers which are decoded on the receiver side to enable calculation of synchronization and delay. The audio part is using fundamentals from Part A.1 by using stamps that consists of a sum of sinusoidal base functions while the video stamp is a binary pattern embedded into the frame. The method has been verified with a real two-way communication application running in an open network.

PART D - On Video Quality Assessment

PART D.1 - A New Low Complex Reference Free Video Quality Predictor

In Part D.1 a low complex no-reference multi-linear regression video quality prediction method is presented, that has been applied to several perceptual quality metrics of coded video sequences. These metrics are PSNR, SSIM, VSSIM, VSSIM modified, NTIA, and PEVQ. The video features used for prediction is extracted from the bitsteam, readily available at the receiver side of a communications channel. Since these parameters are extracted from the coded video bit stream the model can be used in user scenarios where it is normally difficult to estimate the quality due to the reference not being available, as in streaming video and mobile TV applications. The predictor turns out to give good results for both the PSNR and the PEVQ metrics.

PART D.2 - Analysis of Impact from Temporal, Spatial and Quantization Variations on Perceptual Video Quality

In Part D.2 an analysis of impact from temporal, spatial and quantization variations on perceptual video quality is performed. A subjective quality assessment experiment was conducted with five original sequences, having 38 different combinations of bitrates, frame rates, and resolutions. The experiment involved 32 subjects. Both direct and statistical evaluation was made of the MOS scores, where MOS scores versus the bitrates are studied, and ANOVA was used for statistical analysis. The experiment outcomes reveals that preserving the spatial resolution throughout the process has the highest significance even in the scenarios with high temporal information.

PART D.3 - Comparison of machine learning methods for quality estimation of videos with diversity in temporal, spatial, and quantization domains.

In Part D.3 an evaluation of three different machine learning methods for no-reference video quality estimation is performed. The methods used were multi-linear regression, artificial neural networks, and least square support vector machine. Based on the state-of-the art two-way video communication applications and streaming services, the models were applied to test data from Part D.2 where the impact of differentiation between both temporal and spatial resolution and quantization level were considered. In the feature selection process both forward greedy and statistical approaches were used and for evaluation Pearson linear correlation coefficient (PCC), Spearman rank order correlation coefficient (SROC) and outlier ratio (OR) were applied.

It could be seen that the two non-linear methods overall performed better and achieved better generalization, especially for the statistical approach, than multi-linear regression. The difference in performance is larger when the amount of features is reduced.

4.4 Conclusion

From the Summaries and Contributions section the following conclusion can be drawn. A solution to remove the impact from audio interfering signal generated by the switching nature of TDMA cellular telephony has been presented. An post-filter reducing blocking and ringing artifacts generated by highly compressed video streams in mobile devices together with an integrated sharpening filter was designed. The de-blocking and de-ringing filter was verified by implementation in large volumes of several models. A robust black-box methods was developed to measure audio and video delay and synchronization for audiovisual communications applications, and the method has been proven robust in evaluation of real application running over public networks. And finally several no-reference methods to predict perceptual video quality based on extracted video features from the encoded video stream was presented. In the context of developing no-reference methods a subjective experiment was also conducted to evaluate the trade off between spatial and temporal resolution and compression.

References

[1] "Digital cellular telecommunication system (phase 2+); half rate speech transcoding, GSM 06.10 version 8.1.0," European Standard, 1999, ETSI.

[2] "Digital cellular telecommunication system (phase 2+); full rate speech; transcoding, GSM 06.20 version 8.0.1," European Standard, 1999, ETSI.

[3] "Digital cellular telecommunication system (phase 2+); enhanced full rate (EFR) speech transcoding, GSM 06.60 version 8.0.0," European Standard, 1999, ETSI.

[4] "Digital cellular telecommunication system (phase 2+); adaptive multi rate (AMR) speech transcoding, GSM 06.90 version 7.2.1," European Standard, 1998, ETSI.

[5] "ITU-T Recommendation H.263, Video coding for low bit rate communication," 2005, ITU.

[6] "ISO/IEC 14496-2:2004, Information technology - Coding of audio-visual objects - Part 2: Visual," 2009, ISO.

[7] "ITU-T Recommendation H.264, Advanced video coding for generic audiovisual services," 2014, ITU.

[8] ITU, ITU-R Recommendation BT.500-13 Methodology for the subjective assessment of the quality of the television pictures. (2012)

[9] "Subjective video quality assessment methods for multimedia applications," September 1999, ITU-T, Recommendation ITU-R P910.

[10] ITU, ITU-T Recommendation P.10 (Amendment 2). (2008)

[11] ITU, ITU-T Recommendation P.10 (Amendment 3). (2011)

[12] P. L. Callet, S. Möller, A. Perkis (eds), *Qualinet White Paper on Definitions of Quality of Experience, Output from the fifth Qualinet meeting*, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003). (2013)

[13] S. Möller, A. Raake, (eds) *Quality of Experience: Advanced Concept, Applications and Methods*, T-Labs Series in Telecommunication Services,. (Springer, 2014)

[14] M. Shahid, A. Rossholm, B. Lövström, H-J. Zepernick, *No-reference image and video quality assessment: a classification and review of recent approaches.*, EURASIP Journal on Image and Video Processing, (Aug., 2014)

[15] H. R. Wu, K. R. Rao, *Digital Video Image Quality and Perceptual Coding.*, Signal Processing and Communications. (CRC, Boca Raton, 2005)

[16] A. R. Reibman, S. Sen, J. V. der Merwe, Analyzing the spatial quality of internet streaming video, in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, (Scottsdale, Arizona, USA January, 2005)

[17] A. Takahashi, D. Hands, V. Barriac, Standardization activities in the ITU for a QoE assessment of IPTV. IEEE Commun. Mag. **46**(2), 78–84 (2008)

[18] F. Yang, S. Wan, Bitstream-based quality assessment for networked video: a review. IEEE Commun. Mag. **50**(11), 203–209 (2012)

[19] ITU-T, ITU-T Recommendation G.1070: opinion model for videotelephony applications. http://www.itu.int/rec/T-REC-G.1070. Accessed 11 April 2014 (2012)

[20] ITU, ITU-T Recommendation P.1201: Parametric non-intrusive assessment of audiovisual media streaming quality. http://handle.itu.int/11.1002/1000/11727. Accessed 11 April 2014 (2012)

[21] ITU, ITU-T Recommendation P.1202: parametric non-intrusive bitstream assessment of video media streaming quality. http://handle.itu.int/11.1002/1000/11730. Accessed 11 April 2014 (2012)

[22] S. Winkler, P. Mohandas, The evolution of video quality measurement: from PSNR to hybrid metrics. IEEE Trans. Broadcasting **54**(3), 660–668 (2008)

Part A

On Audio Enhancement in Mobile Devices

Part A consists of:

Part A.1: GSM TDMA Frame Rate Internal Active Noise Cancellation
Part A.2: Notch Filtering of Humming GSM Mobile Telephone Noise

Part A.1

GSM TDMA Frame Rate Internal Active Noise Cancellation

Part A.1 is published as:

I. Claesson and A. Nilsson (Rossholm), *GSM TDMA Frame Rate Internal Active Noise Cancellation.*, in International Journal of Acoustics and Vibration (IJAV), September 2003.

Parts of Part A has been published as:

I. Claesson and A. Nilsson (Rossholm), *Cancellation of Humming GSM Mobile Telephone Noise.*, at International Conferences on Information, Communications and Signal Processing (ICICS), December 2003.

GSM TDMA Frame Rate Internal Active Noise Cancellation

Andreas Rossholm and Ingvar Claesson

Abstract

A common problem in the world's most widespread cellular telephone system, the GSM system, is the interfering signal generated by the switching nature of TDMA cellular telephony in handheld and other terminals. Signals are sent in chunks of data, speech frames, equivalent to 160 samples of data corresponding to 20 ms at 8 kHz sampling rate.

This paper describes a study of two different software solutions designed to suppress such interference internally in the mobile handset. The methods are Notch Filtering, which is multiplicative in frequency, and subtractive Noise Cancellation, which is an alternative method employing correlators. The latter solution is a straigt-forward, although somewhat unorthodox, application of "in-wire" active noise control.

Since subtraction is performed directly in the time domain, and we have access to the state of the mobile, it is also possible to consider a recurring pause in the interference caused by the idle frame in the transmission, when the mobile listens to other base stations communicating. More complex control algorithms, based on the state of the communication between the handset and the base station, can be utilized.

1 Introduction

In GSM mobile telephony it is a common problem that an interfering signal is introduced into the microphone signal when the mobile is transmitting. This interfering signal is transmitted along with the speech signal to the receiver. Due to the humming sound of the interfering signal it is commonly denoted the *Bumblebee*.

Since interleaving of data is utilized and since control data transmission is also necessary, the connection between transmitter/receiver frames and speech

frames is somewhat complicated. Data from a speech frame of 20 ms is sent in several bursts, each occupying 1/8 of a transmitting frame. The radio circuits are switched on and off with the radio access rate frequency. An electromagnetic field pulsating with this frequency and its harmonics disturbs its own microphone signal, as well as electronic equipment in the vicinity, producing in some cases annoying periodic humming noise in the uplink speech from the handset to the base station.

The Bumblebee is generated by the switching nature of TDMA cellular telephony, where the radio circuits are switched on and off. During the time the radio is switched on, denoted a time slot, the mobile transmits its information by sending electromagnetic impulses. These impulses are induced in the microphone path and generate interference, which consists of the fundamental frequency and its harmonics. The fundamental switching rate is approximately 217 Hz, more specifically, $5200/(3 \cdot 8)$Hz, according to the GSM standard [1].

Since the frequency components of the disturbing periodic humming noise are crystal generated and accurately known, it is possible to estimate the cosine- and the sine- parts of these with correlators. This is easily done by correlating the microphone signal with sinusoids having the same crystal generated frequencies as the disturbing frequencies. By generating the cosine- and sine- signals with correct signed amplitudes and then subtracting these from the microphone signal, the humming "Bumblebee" is almost perfectly suppressed in the microphone signal. This is a classical example where in-wire subtractive active noise control is beneficial [2, 3]

Depending on the power level the mobile telephone is transmitting, how it is held and if one uses portable hands-free equipment or not, the amplitudes and phases of the fundamental and its harmonics will vary. When the mobile changes time slot, i.e. during a hand-over between base stations, the amplitudes and phases will also change abruptly.

Earlier solutions of this problem have utilized different hardware constructions, i.e., better placement of the components, usage of special electronics and microphones, reconstruction of analog parts, etc. However, this is expensive, time absorbing and becomes increasingly harder when the mobiles constantly shrinks in size, thus causing the microphone to be situated closer to the transmitting antenna.

The solution to the problem presented in this paper makes use of the fact that the disturbance, after a Fourier series expansion, can be accurately described by a sum of sinusoids with well-defined frequencies. Two time domain software solutions to attenuate these frequency components of the digitized

microphone signal directly in the base band, syncronized correlators and notch filtering, are evaluated.

The best results were achieved by estimating the amount of the different sinusoids with correlators, and then subtract these sinusoidal estimates from the microphone signal, as opposed to conventional notch filtering. This is an illustrative example of an application where subtraction of disturbances, typical for Active Noise Control [2, 4], is suitable.

2 Problem background and Signal Model

The humming "Bumblebee" disturbance is a result of the transmitting technique used in GSM, Time Division Multiple Access (TDMA). The handheld mobile, formally denoted the Mobile Equipment (ME), sends information during the time slot that it is assigned. Eight time slots make one TDMA-frame, in which the time slots are numbered $0-7$. A mobile uses the same time slot in every TDMA-frame until the network orders it to another time slot, i.e. when the traffic is rerouted via another base station, a handover. The duration of a time slot is $3/5200$ seconds, and the period time of the TDMA-frames is $8 \cdot 3/5200$ seconds. During the assigned time slot the mobile transmits its information by sending electromagnetic bursts. These are induced in the analog microphone path and produce an annoying periodic interference in the uplink speech. The fundamental frequency is $1/(8 \cdot (3/5200)) \approx 217$ Hz in Full Rate (FR).

There is another case that is not so common but still worth mentioning, Half Rate transmission (HR), where the radio access pattern differs considerably from FR. This communication scheme offers cheaper traffic with slightly decreased speech quality, but approximately twice as many connections in the ideal case. The period of the interference in this case is $1/(8{\cdot}2{\cdot}(3/5200)) \approx 108$ Hz, which is half the frequency of the FR, since the mobile is only transmitting during every other time slot.

Some mobile networks supports a feature denoted Discontinuous Transmission (DTX), which is a mechanism allowing the radio transmitter to be switched off most of the time during speech pauses. During these pauses the background noise is averaged and only Silent Descriptor (SID) frames are transmitted to the receiver. A SID frames contains hereby no disturbing frequencies, and consequently, the algorithm is not allowed to run during DTX.

2.1 Analysis of the Bumblebee

A typical recorded disturbed signal from a silent room can be seen in Fig. 1. The interfering signal is periodic but somewhat complicated since, in the

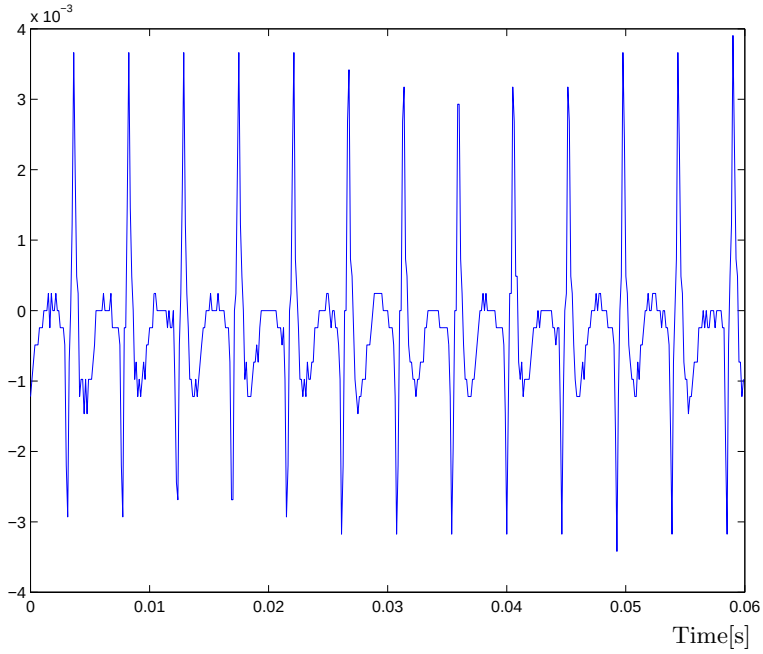


Figure 1: *Interfering signal at the microphone A/D converter recorded in a silent room with no speech.*

case of FR, there is no transmission when the mobile is listening to other base stations. Such silent frames occur once every 26 TDMA-frames and are denoted *idle* frames. Idle frames are illustrated in Fig. 2-3.

In the HR case the disturbance pattern is even more complex, but we refrain from detailed analysis here. We observe that since the state of the communication between the mobile and base station is known, sufficient information to ascertain whether estimation and/or cancellation should take place or not is always at hand.

The simple radio access pattern for Full Rate (FR) as well as the more

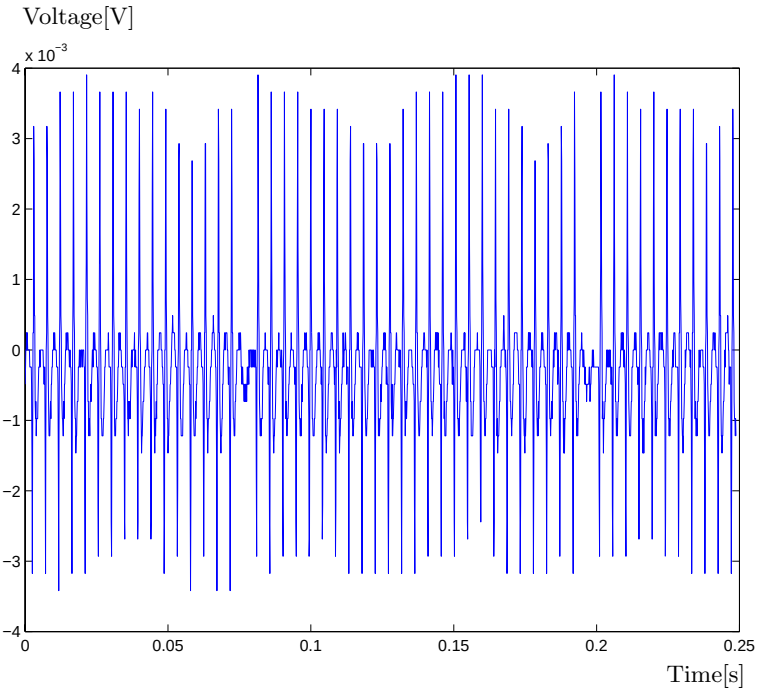


Figure 2: *Pattern for interfering signal recorded in a silent room,* **Full Rate.**

complex pattern for an even Half Rate (HR) channel can also be seen in Fig. 2 and 3 respectively.

Obviously, the idle frame should be considered when eliminating interference. Since the disturbance is periodic, it can be viewed as a Fourier series expansion

$$x_p(n) = \sum_{k=1}^{K} C_k \sin\left(2\pi k(f_0/f_s)n + \theta_k\right) \tag{1}$$

where K denotes the number of tones (fundamental plus harmonics), f_s is the sample frequency, and f_0 represents the frequency of the fundamental tone.

The number of tonal components K that are needed to represent the disturbance are limited by the sampling rate of the signal, which is 8 kHz. Consequently, the interfering signal after sampling will only consist of frequencies below 4 kHz since aliasing is carefully avoided in the mobile. Further filters

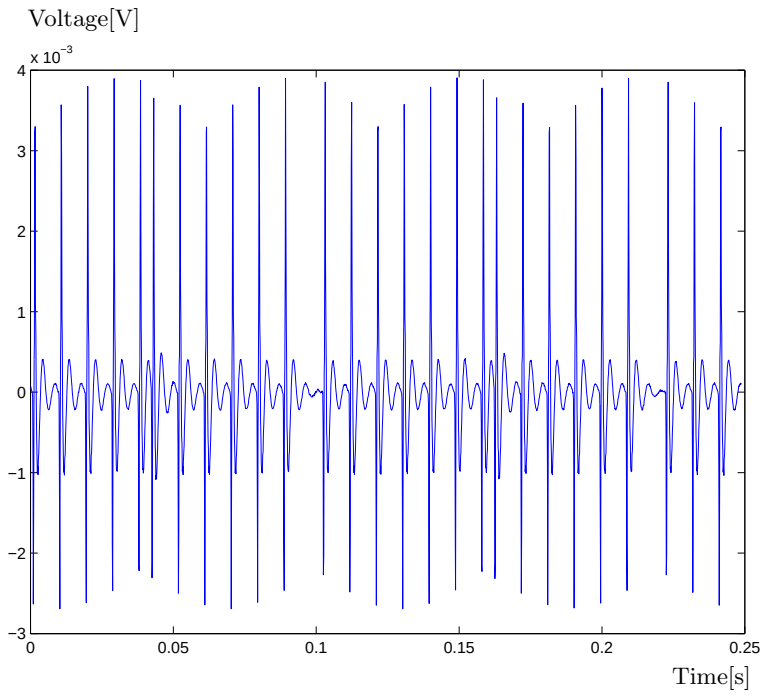


Figure 3: *Pattern for interfering signal recorded in a silent room,* **Half Rate.**

connected to the A/D conversion and the speech coder also band limit all signals, including Bumblebee disturbance, to approximately $300 - 3400$ Hz. Hence, the fundamental tone and the 15:Th harmonic will be slightly attenuated, see Fig. 4. A similar Fourier series expansion can of course be carried out for the HR case but the details are omitted in this paper. However, we observe that in this case the fundamental frequency, f_0, equals half the fundamental frequency in the full rate case. Hence, almost the double amount of harmonics is needed within the telephone frequency range to represent the disturbance. A comprehensive description illustrating the transmission patterns for both full rate and half rate transmission are given in Fig. 5.

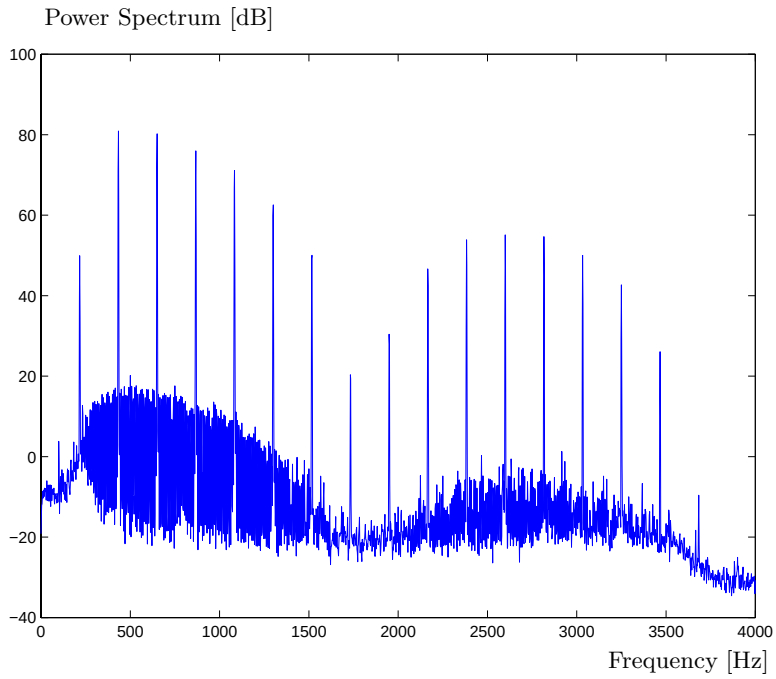


Figure 4: *Spectrum of periodic "Bumblebee" disturbance in random noise background.*

3 Solution proposals

Two different methods to eliminate the Bumblebee disturbance are proposed, both working in the time domain.

These methods are Linear Time-Invariant Notch filters, which work on a sample-by-sample basis, and Noise Canceling Correlators, which work frame-wise on 160 samples in each time slot of 20 ms duration, i.e. the standardized slot duration in GSM at 8 kHz sampling rate.

3.1 Notch filters

A notch filter consists of a number of deep notches, or ideally nulls, in its frequency response, see Fig. 6. Such a filter is useful when specific frequency

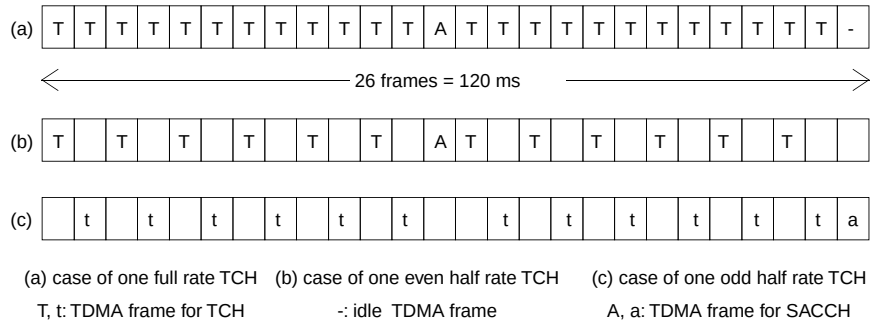


Figure 5: *Transmission patterns in GSM full-rate and half-rate. TCH denotes Traffic Channel, and SACCH Slow Associated Control Channel.*

components of known frequencies must be eliminated [5,6]. To eliminate the frequencies at $\omega_n, n = [1, \ldots, N]$, pairs of complex-conjugated zeros are placed on the unit circle at the angles ω_n

$$z_{n_{1,2}} = r_b e^{\pm j\omega_n}, \qquad r_b = 1 \tag{2}$$

This results in a crude FIR notch filter with the system function

$$H(z) = B(z) = b_o \prod_{n=1}^{N} (1 - r_b e^{j\omega_n} z^{-1})(1 - r_b e^{-j\omega_n} z^{-1}). \tag{3}$$

The b_0 constant is chosen as

$$b_0 = \frac{1}{\sum_{n=1}^{N} b_n} \tag{4}$$

to normalize the gain. To control the bandwidth of the FIR notches, poles are placed at the same angle as the zeros but with a slightly smaller magnitude. The positions of the poles are thus

$$p_{n_{1,2}} = r_a e^{\pm j\omega_n}, \qquad 0 \ll r_a < r_b. \tag{5}$$

Consequently, the system function of the resulting notch filter is

$$H(z) = \frac{B(z)}{A(z)} = b_o \prod_{n=1}^{N} \frac{(1 - r_b e^{j\omega_n} z^{-1})(1 - r_b e^{-j\omega_n} z^{-1})}{(1 - r_a e^{j\omega_n} z^{-1})(1 - r_a e^{-j\omega_n} z^{-1})} \tag{6}$$

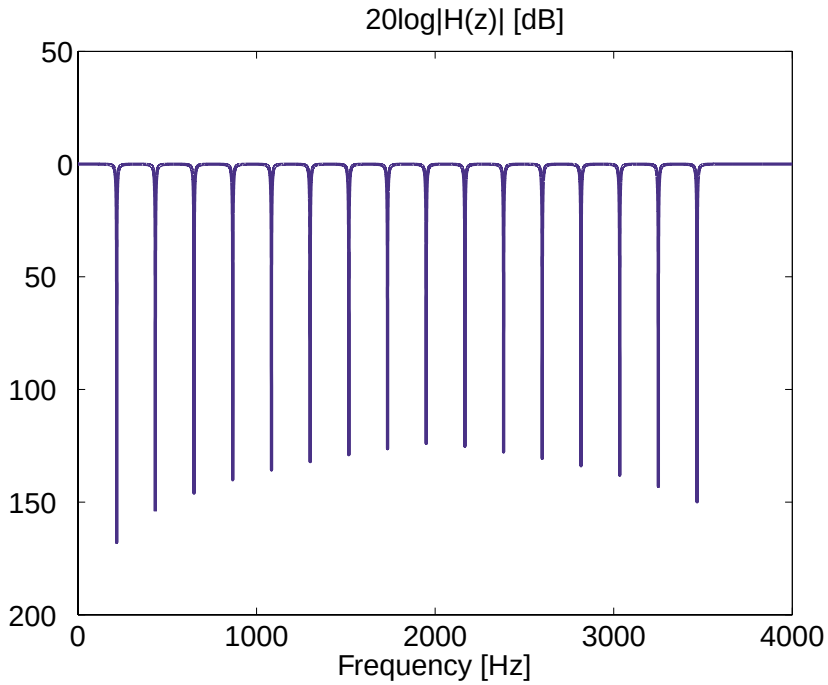


Figure 6: *Frequency response of an FIR notch filter with* $r_b = 1,\ N = 16$ *and* $\omega_n = n \cdot 2\pi \cdot (5200/(8 \cdot 3))$

where

$$b_0 = \frac{\sum_{n=1}^{N} a_n}{\sum_{n=1}^{N} b_n} \tag{7}$$

The frequency response of the filter in Equation (6) is plotted in Fig. 6. However, even sharp IIR notch filters have a non-negligible bandwidth, which leads to signal attenuation at frequencies also in the vicinity of the notches.

3.2 Orthogonal Correlators or Length-480 FFT Coefficients

Any band-limited periodic signal can be represented by a finite sum of sinusoids. Since we have periodic disturbance $x_p(n)$ superimposed on aperiodic speech $w(n)$, the model assumption for the input signal is given by

$$x(n) = x_p(n) + w(n) = \sum_{k=1}^{K} C_k \sin(nk\omega_0 + \theta_k) + w(n) \tag{8}$$

or alternatively

$$x(n) = \sum_{k=1}^{K} R_k \cos(2\pi f_k n) + I_k \sin(2\pi f_k n) + w(n) \tag{9}$$

where $f_k = k \cdot f_0$ and f_0 is the fundamental frequency of the disturbance. Since the disturbance frequencies are known, only the coefficients of the cosine- and sine- parts, R_k and I_k, need to be estimated.

The Maximum Likelihood (ML) estimate of known sinusoids in white noise background is given by correlation or matched filtering. This is equivalent, in our situation, to finding the Fourier Expansion coefficients, or in the discrete-time case, the FFT coefficients at the exact frequencies where the periodic disturbances are. Even if speech cannot be regarded as a white disturbance, it is still an attractive Least Squares (LS) solution to correlate out the sinusoidals [7]- [9].

In order to inherently achieve unbiased LS estimates, correlation can be made over a whole number of periods for each sinusoidal. This corresponds to that each disturbing frequency is situated exactly at an FFT bin. This is achieved if correlation (FFT bin calculation) is made over 480 samples (3 frames) in the full-rate situation, and 960 samples (6 frames) in the half-rate case. Performing a pruned FFT with lengths of other lengths than factors of 2:s (2^M), in this case N=480 or N=960 is certainly not straightforward. Neither is it desirable in the present context, since we are only interested in the FFT bins where the periodic disturbance is present, typically only in 16 of the bins. Hence, an FFT is not the most efficient way to calculate the correlations in this case.

A sinusoidal correlator estimator consists mainly of a bank of dual product-adders, one for each frequency, one for each cosine- and sine part, in total $2 * K$ ($K = 16$) correlators of length N=480 in the full-rate case. This makes

it easy to estimate and compensate the Bumblebee disturbance in "real time", frame by frame, by adding the correlation contribution of the most recent 160 samples, the present frame, and subtracting the correlation contribution of the 160 samples (3 frames back) in the frame leaving the estimation interval, i.e. the most recent 480 samples. To do this, the cosine- and sine- parts of the different frequencies are estimated by correlation in accordance with Fig 7, yielding the estimates $\hat{R}_k$ and $\hat{I}_k$, respectively in the two branches. These

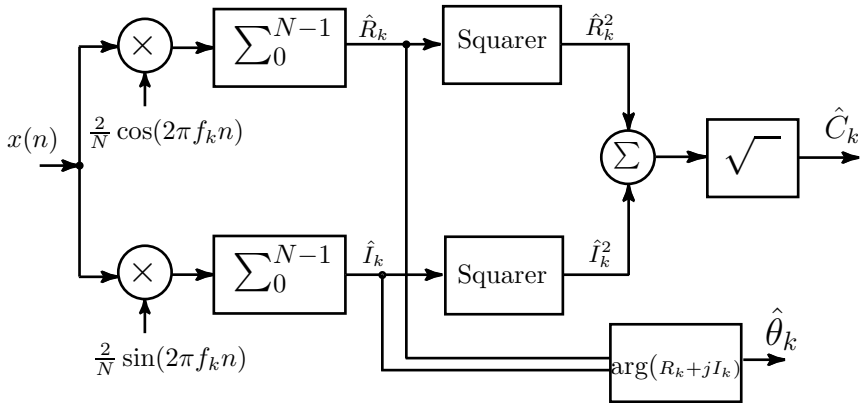


Figure 7: *Sinusoidal estimation with correlators*

signals are then subtracted from the input signal yielding

$$y(k) = x(k) - \sum_{k=1}^{K} \hat{R}_k \cos(2\pi f_k n) + \hat{I}_k \sin(2\pi f_k n) \tag{10}$$

If the amplitude and phase are required instead, we proceed by

$$\sqrt{\hat{R}_k^2 + \hat{I}_k^2} = \hat{C}_k \tag{11}$$

and the corresponding phase estimate of θ_k by calculating the four-quadrant angle

$$\hat{\theta}_k = \arg\left(\hat{R}_k + j\hat{I}_k\right). \tag{12}$$

3.3 Implementation Aspects

This estimation is carried out block-wise using correlators. The amount of data in each block that is used for the estimation should preferably be done over an integer number of fundamental periods in order to avoid bias from incomplete periods. For the fundamental tone, which has the lowest frequency and thus requires most samples, we need 480 samples to fulfill the requirement in the FR case. This is easily derived, since the frequency of the fundamental tone is $1/(8 \cdot (3/5200) \cdot 8000)$, and the sample rate is 8 kHz. This gives to $f_0/f_s = 13/480$ implying that 480 samples are needed to represent an integer number (13) of the fundamental periods with an integer number (3) of slots of 160 samples. In other words, to fulfill the biasfree requirement of whole periods, 13 fundamental periods are required which gives the block size 480, which is also the equivalent of 3 GSM frames, each with 160 samples. Since the length is given by 480 samples and only 16 tonal components are to be calculated, the is no need to use FFT algorithms. Instead, a more straightforward route is taken.

In discrete time, we simply correlate the received signal with the 16*2 basis functions of the correlators (cosines and sines) in order to obtain the coefficients for the cosines and sines. These estimates are subsequently used as coefficients for the amount each sinusoid should be subtracted from the received signal.

If estimation is performed during speech, the estimate of the Bumblebee disturbance will be incorrect, since the speech contains high energy at the same frequencies as the disturbance. This problem is solved by only making estimates during speech pauses, a Voice Activity Detector (VAD) is thus required. Fortunately, the mobile is already equipped with a VAD, which therefore can be easily utilized, see Fig. 8.

The VAD information is further elaborated on for several GSM frames, since a VAD algorithm works on 160-sample frames. A flag is set to one if speech is present. To consider the present frame as non-speech ,the three most recent frames (480 samples) and the following frame must all have VAD=0. The reason for this is that even if VAD=0 for the past three frames, it is wise to check the following frame (n), since there may be the beginning of speech at the end of the present, most recent tentative estimation frame (n-1) which otherwise would destroy the estimation. As a result, the correlation estimate will be one frame older (delayed), but this is still a better solution. If the VAD conditions are not fulfilled, it is often much better to keep an old estimate than an erroneous one partially disturbed by speech, since the coefficients of

the cosines and sines normally only varies slowly during operation.

More important to observe is that the speech will not be delayed. To avoid any signal delay of the speech we only estimate/correlate sinusoids on the three previous GSM frames, with delayed samples, though the subtraction is performed on the present GSM frame, see Fig. 8. The idle and silent states

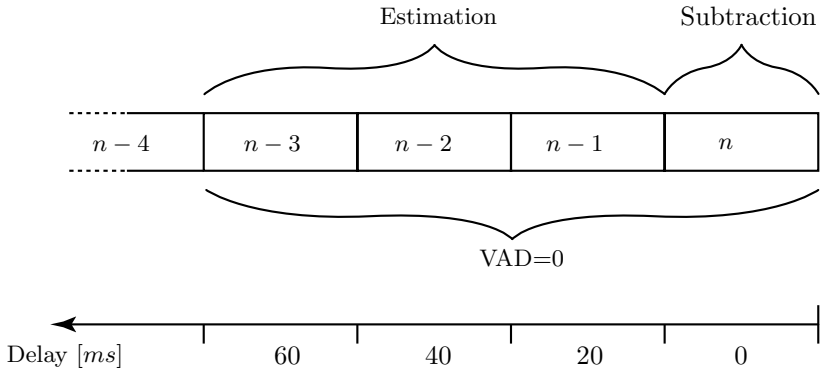


Figure 8: *Estimation and subtraction when the VAD algorithm is used*

should also be considered. This is done by inhibiting disturbance subtraction during idle mode and preventing from estimation/correlation during silent frames. Since the transmission state is locally known in the mobile as well as the structure of the frames, Fig. 5, these states are easily handled in a software implementation.

4 Cancellation results on recorded signals

The problem with the Bumblebee distubance is not just to eliminate it, but to do so without impairing speech quality. The following analysis is based on data recorded from the Digital Audio Interface (DAI) in an Ericsson mobile. The DAI is the interface after the A/D-converter where the signal is Pulse Code Modulated. This is the signal that enters the DSP, which is processed by the algorithm.

The frequencies that will be attenuated in the tests are: $k \cdot \omega_0, \quad k = [1, \ldots, 16]$ and ω_0 is the fundamental tone of the Bumblebee disturbance, $5200/(8 \cdot 3)$ Hz. With $K = 16$, the fundamental tone and 15 of its harmonics

will be eliminated. This will span a range up to 3467 Hz which covers the frequency range of the telephone frequency range.

4.1 Notch filter

Since the frequencies which constitute the Bumblebee are well defined, we first apply a notch filter directly in the signal path to reduce the interference.

4.1.1 Implementation

The notches are made as deep as possible, so that ideally the frequencies in question are totally eliminated. This results in the following system function:

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{k=1}^{16} a_k}{\sum_{k=1}^{16} b_k} \prod_{k=1}^{16} \frac{(1 - r_b e^{jk\omega_0} z^{-1})(1 - r_b e^{-jk\omega_0} z^{-1})}{(1 - r_a e^{jk\omega_0} z^{-1})(1 - r_a e^{-jk\omega_0} z^{-1})} \qquad (13)$$

The calculations are made recursively on the whole data set. This will result in a convergence period at the start up and also when a handover between base stations occurs. Unfortunately, the notch filter is active also under idle frames, a drawback resulting from the fact that it works sample-by-sample and recursively, leading to unwanted artifacts during idle frames, when trying to subtract a disturbance that is not present, i.e. a negative disturbance is added, see Fig. 9.

It can be seen that the Bumblebee disturbance is considerably attenuated. However, this solution does not give a satisfactory result, since a portion of the speech is also attenuated, resulting in a "canned" or metallic sound. This can be seen in Fig. 9-10. Another problem with this solution is that the periodic idle frame cannot be handled resulting in a new periodic interference, 26 times lower in frequency, see Fig. 11. The reason for this is that the notch filter consists of poles (autoregressive), which give feedback of the output signal ($y(t)$) continuously. Consequently, the Bumblebee is added during the idle frame, according to the tails of impulse responses of IIR filters.

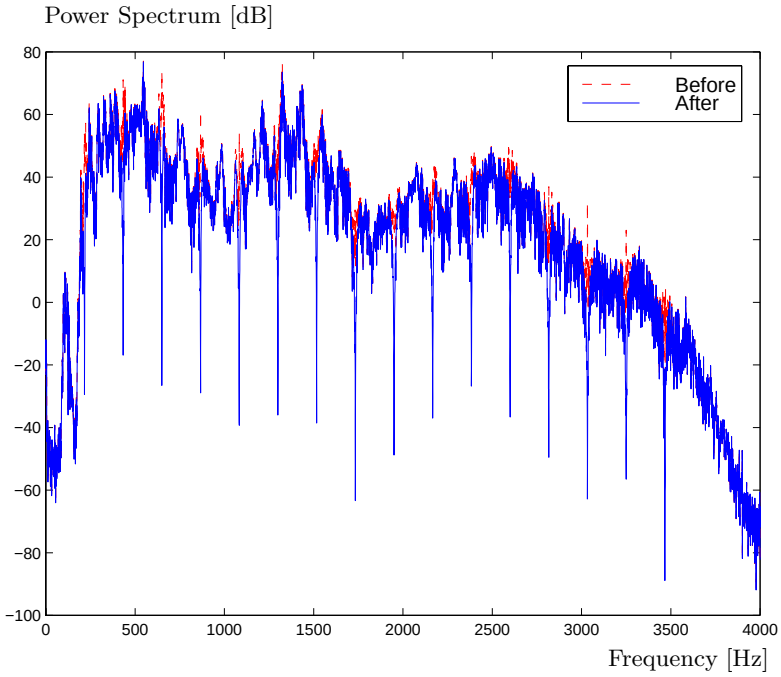


Figure 9: *Cancellation of the Bumblebee with* **notch filter** *in speech.* **Full Rate, with speech**.

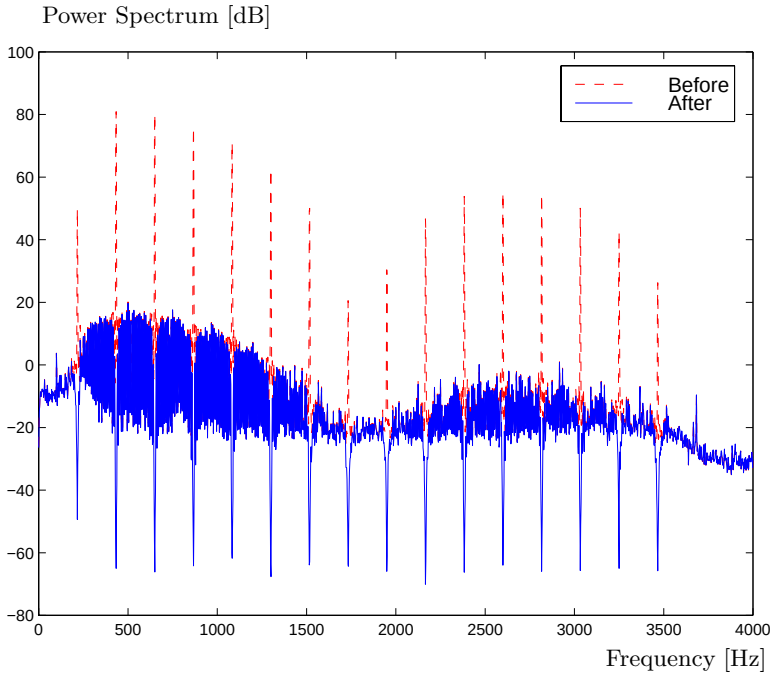


Figure 10: *Cancellation of the Bumblebee with* **notch filter**. *The Bumblebee was recorded in a silent room.* **Full Rate, no speech**.

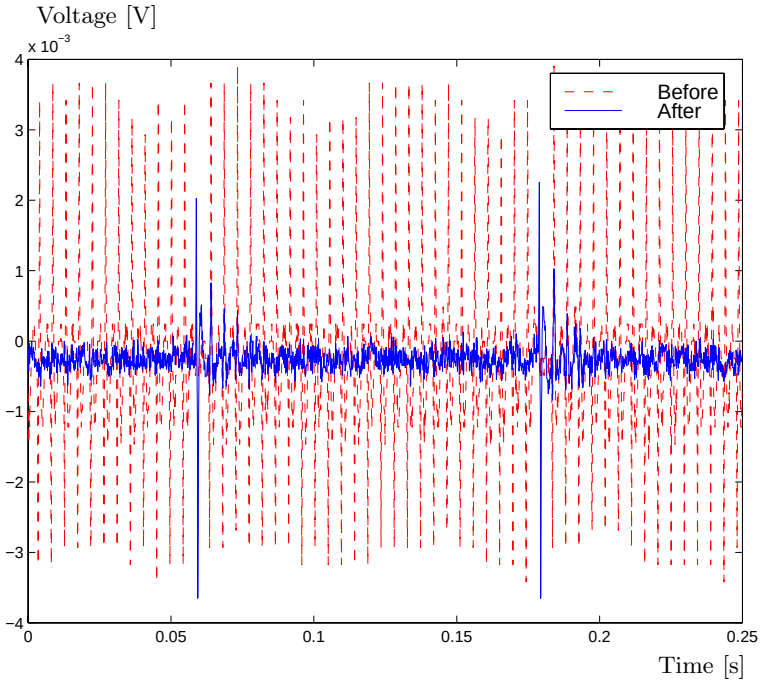


Figure 11: *Time signal of the notched Bumblebee.*

4.2 Correlators

The data set that has been used is identical to that used when evaluating the notch filter. That is, the first test is done on data recorded both with speech and in a silent room, see Fig. 12-13. The metallic sound and the

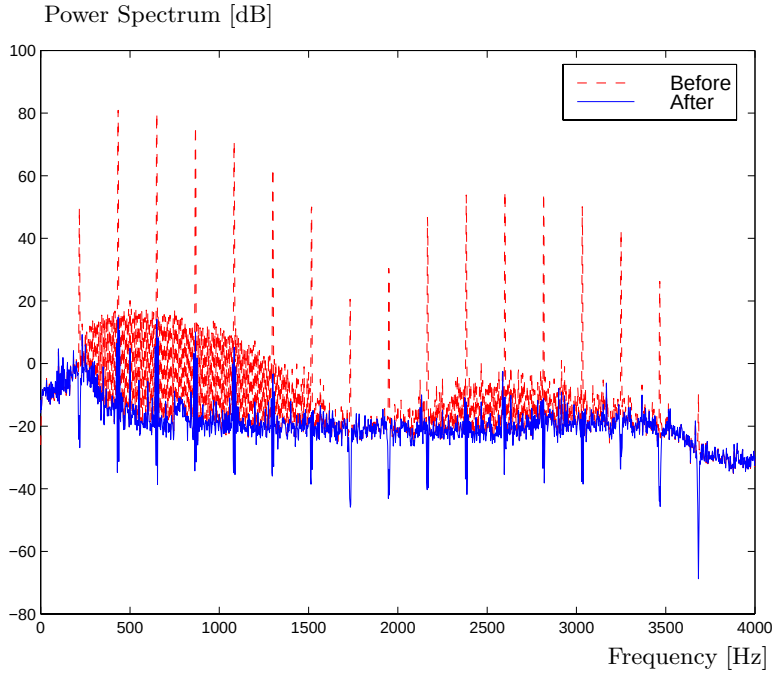


Figure 12: *Cancellation of the Bumblebee with* **correlators** *where idle mode has been taken into consideration. The Bumblebee was recorded in a silent room.* **Full Rate, no speech**.

periodic interference that appeared in the notch tests from the idle frame are also avoided, thanks to time-limited subtractive nature of block correlation canceling, thus avoiding long-tailed (recursive) impulse responses. This gives a highly satisfactory result. Observe in Fig. 13 that only the Bumblebee disturbance is attenuated. A corresponding and even more impressive result is also presented for the HR case, Fig. 14. Finally, an alternative type of comparison is introduced in Fig. 15-16 illustrating P_{out}/P_{in}, which gives the

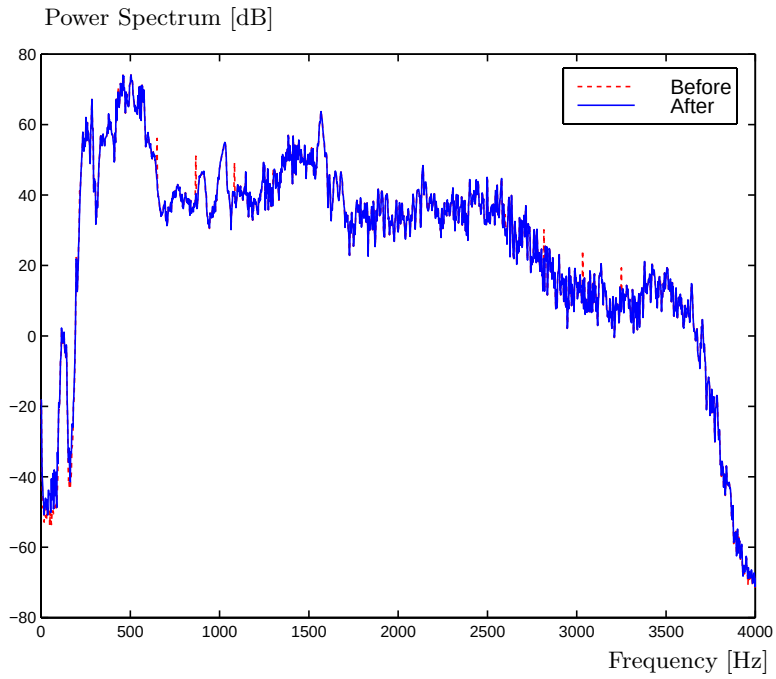


Figure 13: *Cancellation of the Bumblebee in speech with* **correlators** *where VAD and idle mode have been taken into consideration.* **Full Rate, with speech**.

over-all system attenuation both for the notch filter and the correlator.

It can be observed that the notch filter gives both a deeper and wider attenuation, which explains the metallic sound and inferior quality as compared with when correlators are used.

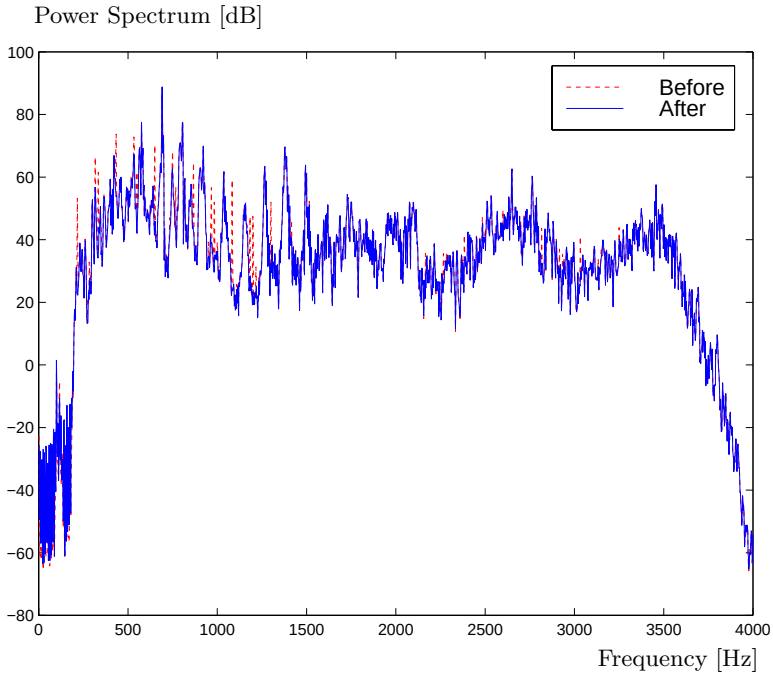


Figure 14: *Cancellation of the Bumblebee in speech with* **correlators** *in the* **Half Rate** *case where the VAD and idle frame have been taken into consideration.* **With speech**.

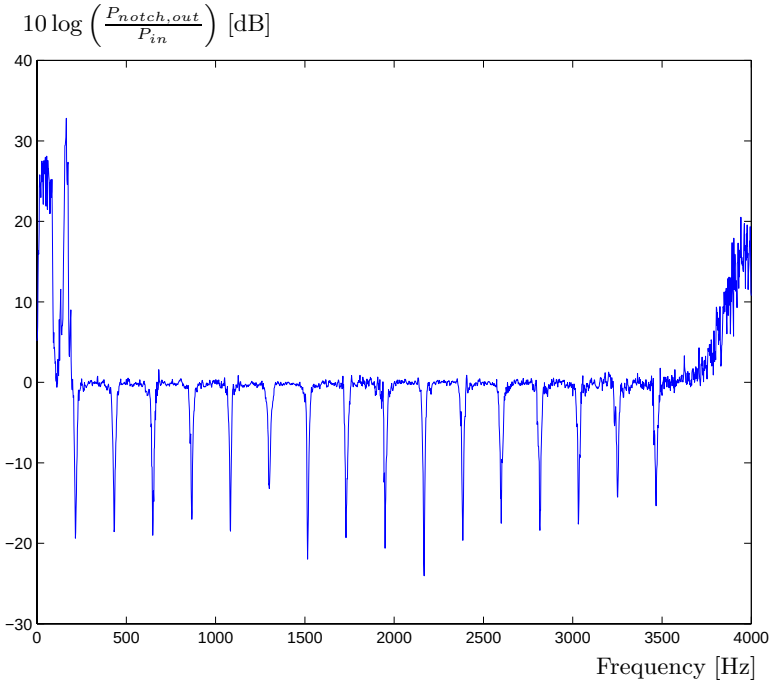


Figure 15: *Divided power estimates with notch filter, no speech.*

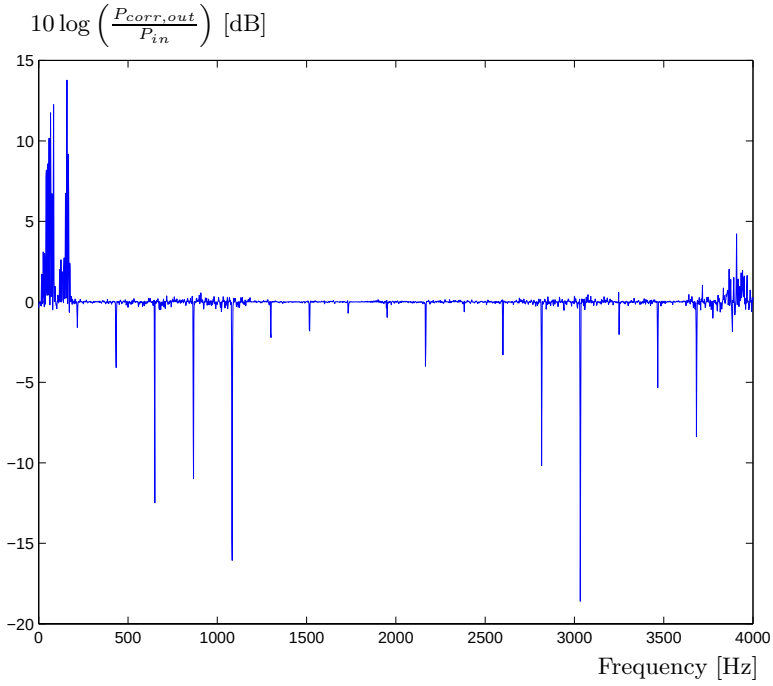


Figure 16: *Divided power estimates with correlators, no speech.*

5 Complexity and Implementation aspects

Complexity estimates have only been made for the correlators since this solution was preferred. The most commonly used unit when performing complexity estimates is MIPS (Millions of Instructions Per Second). However, this can be a misleading measure because of the varying amounts of work done by an instruction. That is, an instruction on one processor may accomplish far more work than an instruction on another. This is especially important for DSP processors, which often have highly specialized instruction sets.

Similarly, MOPS (Millions of Operations Per Second) suffer from related problems: what counts as an operation and the number of operations needed to accomplish useful work varies greatly from processor to processor.

A third performance unit that can be used is MACS (Multiply ACcumulates per Second). Most DSP processors can complete one MAC per instruction cycle, making this unit equivalent to MIPS for DSPs. Furthermore, MAC estimates disregard the important data movement and processing required before and after.

After considering the various drawbacks, we selected the MIPS measure to be used, which is given in Table 2.

The complexity calculations are based on the attenuation of 16 sinusoids. The estimation is performed on 480 samples, and the subtraction of the estimated signal on 160 samples. This is the way it should be done in the mobile to avoid a delay. The sinusoids and the cosinusoids are stored in a Read Only Memory (ROM) as a table. Another solution could be to use a digital sinusoidal oscillator. A such solution does not require as much ROM memory as the table approach, but is much more complex and does not generate the sinusoids and the cosinusoids perfectly.

To build up the 480 samples long sinusoids, the table should contain an integer number of periods for each frequency. That is, $480/k$ samples with the exception of the frequencies stated in Table 1. If $K = 16$, a ROM of 6452 words, is required and the complexity is approximately 1.3 MIPS, see Table 2. Control code and data transfers will also be needed. A very conservative estimation of the total complexity is 2 MIPS.

As mentioned before, the fundamental tone (and the first harmonic for HR) are already severely attenuated because of the filter, A/D converter and the speech coder. This makes it possible to also ignore these tones without degrading the result. Symmetries in sinusoidal base functions and recursive estimation where the estimates are updated with the recent frame data of 160 samples can reduce the computational load by more than 50%. With this in

k	Samples needed
7	480
9	160
11	480
13	480
14	240

Table 1: Samples needed for the frequencies $k \cdot f_0$

Task	Instructions / 20ms	MIPS
Correlation	$16 \times 2 \times 480$	0.768
Building $\hat{b}$	$16 \times 2 \times 2 \times 160$	0.512
Subtracting	160	0.008
Total	**25760**	**1.288**

Table 2: Complexity of the Table approach

mind we conclude that correlation canceling is a cheap and convenient way of coping with the problem of humming Bumblebee noise in GSM cellular telephony.

6 Summary, Conclusions and Future Work

In this paper we have compared two methods for eliminating an annoying self-disturbance in mobile telephone microphone signals originating from the telephones's own antenna. Such disturbance is caused by TDMA switching in GSM cellular telephones. The Active Noise Control approach which subtracts disturbances, instead of filtering them out has shown great potential. The aim is now to implement the algorithm in fixed-point precision.

References

[1] GSM Standard (GSM 05.01 version 7.0.0 Release 1998) *Digital cellular telecommunications system (Phase 2+), Physical layer on the radio path (General description).*

[2] M. Kuo, D. R. Morgan, *Active Noise Control Systems*, John Wiley & Sons, Inc., 1996.

[3] Chaplin; George B. B.,Smith; Roderick A. Method and apparatus for cancelling vibrations, United States Patent no 4,490,841 Chaplin , Dec 25, 1984

[4] B.Widrow, S. D. Stearns *Adaptive Signal Processing* Prentice Hall, 1985

[5] Proakis, J.G. and Manolakis, D.G. *Digital signal processing*, pp. 343-345, 1996, Prentice-Hall Inc.

[6] Simon Haykin *Digital communications*, 1988, John Wiley & Sons Inc.

[7] Peyton Z. Peebles, Jr. *Probability, random variables, and random signal principles*, 1993, McGraw-Hill Inc.

[8] Steven M. Kay *Fundamentals of statistical signal processing: Estimation Theory*, pp. 183-198, 1993, Prentice-Hall Inc.

[9] Per Eriksson *On estimation of the amplitude and the phase function (Technical Report TR-148)*, 1981, University of Lund / SWEDEN.

Part A.2

Notch Filtering of Humming GSM Mobile Telephone Noise.

Part A.2 is published as:

I. Claesson and A. Nilsson (Rossholm), *Notch Filtering of Humming GSM Mobile Telephone Noise.*, at International Conferences on Information, Communications and Signal Processing (ICICS), December 2005.

Notch Filtering of Humming GSM Mobile Telephone Noise.

Andreas Rossholm and Ingvar Claesson

Abstract

A common problem in the world's most widespread cellular telephone system, the GSM system, is the interfering signal generated in TDMA cellular telephony. The infamous "bumblebee" is generated by the switching nature of TDMA cellular telephony, the radio circuits are switched on and off at a rate of approximately 217 Hz (GSM).

This paper describes a study of two solutions for eliminating the humming noise with IIR notch filters. The simpler one is suitable for any exterior equipment. This method still suffers from a small residual of the noise, resulting from the IDLE slots of the sending mobile. The more advanced IIR structure for use within the mobile also eliminates this residual.

1 Introduction

In GSM mobile telephony it is a common problem that an interfering signal is introduced into the microphone signal when the mobile is transmitting. This interfering signal is transmitted along with the speech signal to the receiver. Due to the humming sound of the interfering signal it is commonly denoted the *Bumblebee*.

Since interleaving of data is utilized and since control data transmission is also necessary, the connection between transmitter/receiver frames and speech frames is somewhat complicated. The interference consists of the fundamental frequency and its harmonics, where the fundamental switching rate is approximately 217 Hz, more specifically, $5200/(3 \cdot 8)$Hz, according to the GSM standard [1]. Signals are sent in chunks of data, speech frames, equivalent to 160 samples of data corresponding to 20 ms at 8 kHz sampling rate. Data

from a speech frame of 20 ms is sent in several bursts, each occupying 1/8 of a transmitting frame. The radio circuits are switched on and off with the radio access rate frequency. An electromagnetic field pulsating with this frequency and its harmonics disturbs its own microphone signal, as well as electronic equipment in the vicinity (within 1-2 meters) of the sending handset antenna, such as radios and active loudspeakers as well as hearing aids, producing in some cases annoying periodic humming noise in the uplink speech from the handset to the base station.

It has been proposed that for internal cancelation in the mobile, the periodic distorsion can be removed by subtraction of an estimate of the distorsion employing correlators and subtraction, similar to Active Noise Control [2–4]. This estimate can be done, since it is known at what frequencies the disturbance will occur, by correlating the block of data with a number of base functions. These base functions are blocks of data corresponding to the fundamental tone and its harmonics. The results of the correlations are used to estimate the amplitude and phase of the bumblebee.

However for equipment with no access to the internal data sending structure of the GSM mobile, notch filters is still the most straight-forward solution.

2 Background and Analysis of the Bumblebee

A typical recorded disturbed signal from a silent room can be seen in Fig. 1. The interfering signal is periodic but somewhat complicated since, in the case of Full Rate transmission, FR, there is no transmission when the mobile is listening to other base stations. Such silent frames occur once every 26 TDMA-frames and are denoted *idle* frames, see Fig. 2.

In densely populated areas, such as Hong Kong, an alternative is sometimes used, Half Rate Transmission (HR), offering cheaper traffic with slightly decreased speech quality. In this case, the period of the interference is $1/(8 \cdot 2 \cdot (3/5200)) \approx 108$ Hz, which is half the frequency of the FR, since the mobile is only transmitting during every other time slot, thus enabling almost twice the number of calls as compared to Full Rate Transmission.

In the HR case the disturbance pattern is thus even more complex, see Fig. 3, but observe that since the state of the communication between the mobile and base station is known, sufficient information to perform internal cancellation is always at hand.

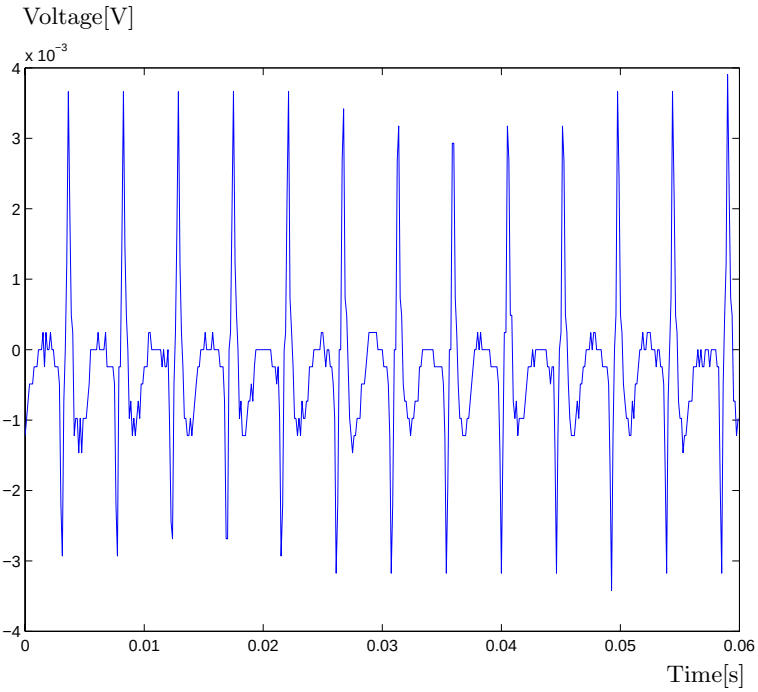


Figure 1: *Interfering signal at the microphone A/D converter recorded in a silent room with no speech.*

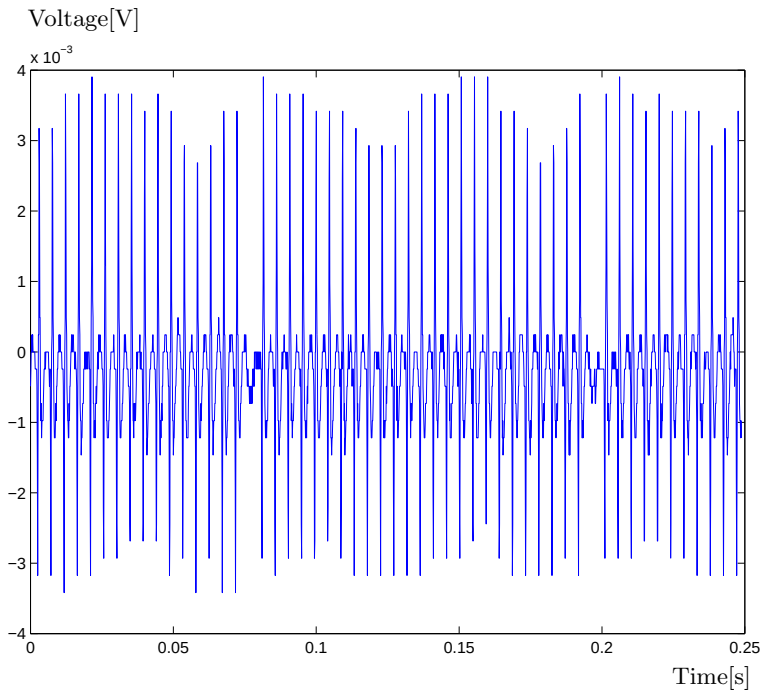


Figure 2: *Pattern for interfering signal recorded in a silent room,* **Full Rate.**

Suppressing the "bumblebee" noise by analog means is a costly, time-consuming and difficult work. It may also require non-optimal system settings in, e.g., the microphone gain, as well as more expensive components. If a digital method is employed, it must be able to continuously track variations in the amplitude and phase of the disturbing periodic signal. The reason for this is that the conditions may change during a call, e.g., the amplitudes are a function of the output power level, and the phases a function of the timing towards the air IF (time slot). Since these parameters change during a call, we must be able to cope with this. Making a Fourier series expansion of the disturbing periodic signal, it is seen that the frequency components decay as 1/f2, which is very slow. In other words there are approximately 15 frequency components that must be suppressed in the band below 3.4 kHz.

By using base function correlation [4] we must save blocks of data in memory, which is negligible with notch filters. Also, this estimation can only

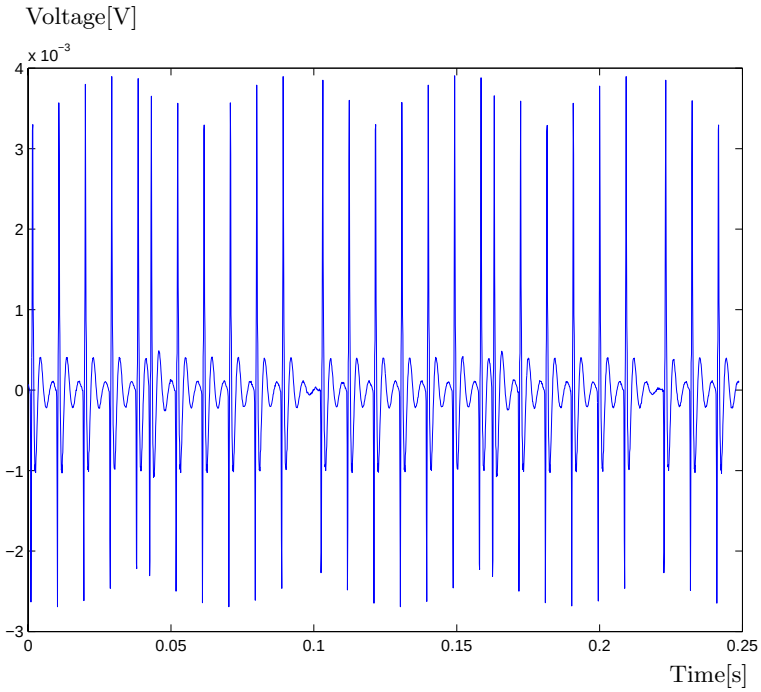


Figure 3: *Pattern for interfering signal recorded in a silent room,* **Half Rate.**

be done during speech pauses, which makes it dependent on side information like Voice Activity Detection(VAD).

A notch filter contains deep notches, in its frequency response. Such a filter is useful when specific frequency components of known frequencies must be eliminated [5, 6]. To eliminate the frequencies at $\omega_n, n = [1, \ldots, N]$, pairs of complex-conjugated nulls and zeros are placed on and just inside the unit circle at the angles ω_n

$$z_{n_{1,2}} = r_b e^{\pm j\omega_n}, \qquad r_b = 1 \tag{1}$$

Consequently, the system function of the resulting notch filter is

$$H(z) = \frac{B(z)}{A(z)} = b_o \prod_{n=1}^{N} \frac{(1 - r_b e^{j\omega_n} z^{-1})(1 - r_b e^{-j\omega_n} z^{-1})}{(1 - r_a e^{j\omega_n} z^{-1})(1 - r_a e^{-j\omega_n} z^{-1})} \tag{2}$$

where

$$b_0 = \frac{\sum_{n=1}^{N} a_n}{\sum_{n=1}^{N} b_n} \tag{3}$$

Using a single, simple straight-forward notch filter will reduce the disturbance significantly, but not totally. This problem is related to the radio access pattern in GSM. In GSM, the mobile makes one radio access every 4.615 ms. Unfortunately, the mobile does not transmit during every time slot. In one 120 ms multiframe, there are 26 TDMA frames of 4.615 ms each, i.e., there are 26 possible occasions for the mobile to transmit. However, only 24 of them are required for transmission of speech coded data (frames 0-11 and 13-24), and one for transmission of the SAACH control data (frame 12). The problem is TDMA frame 25, the idle frame, in which there is no radio transmission. During the idle frame (or idle time slot), the mobile measures neighboring cells. Since the radio of the mobile is not transmitting during the idle frame, the disturbance is zero during this period, and the IIR filters are trying to cancel a noise that is not there.

A simple Notch filter is an IIR (infinite-duration impulse response) filter, attenuating the "bumblebee", but introduces a new residual disturbance. The frequency of the introduced disturbance is approximately 8 Hz (= 1 / 120ms). Although the disturbing power is much attenuated compared to the original "bumblebee" signal, the fluttering characteristic of the introduced noise is still perceived. Because of the absence of radio transmission in the idle frame the "bumblebee" noise is not exactly periodic with the TDMA frame rate, even though it appears so when listening to it.

3 Notch Solutions

3.1 Simple Notch filter

We first apply a notch filter directly in the signal path to reduce the interference. The notches are made as deep as possible, so that ideally the frequencies in question are totally eliminated. This results in the following system function:

$$\frac{B(z)}{A(z)} = \frac{\sum_{k=1}^{16} a_k}{\sum_{k=1}^{16} b_k} \prod_{k=1}^{16} \frac{(1 - r_b e^{jk\omega_0} z^{-1})(1 - r_b e^{-jk\omega_0} z^{-1})}{(1 - r_a e^{jk\omega_0} z^{-1})(1 - r_a e^{-jk\omega_0} z^{-1})} \tag{4}$$

The calculations are made recursively on the whole data set. This will result in a convergence period at the start up and also when a handover between base stations occurs. Unfortunately, the notch filter is active also under idle frames, a drawback resulting from the fact that it works sample-by-sample and recursively, leading to small, residual artifacts during idle frames, when trying to subtract a disturbance that is not present, i.e. a negative disturbance is added, see Fig. 4-6.

It can however be seen that the Bumblebee disturbance is considerably attenuated. However, this solution can be further improved to even more satisfactory results, by handling the residual periodic interference, 26 times lower in frequency, see Fig. 5. The reason for this is that the notch filter consists of poles (autoregressive), which give feedback of the output signal ($y(t)$) continuously. Consequently, the Bumblebee is added during the idle frame, according to the tails of impulse responses of IIR filters.

3.2 Advanced Notch filter

We propose the following solution to the problem for internal cancelation the mobile. We make use of our a priori knowledge that the disturbing signal consists of a sum of sinusoids of very well known frequencies, i.e., the disturbing signal can be expressed as

$$e(k) = \sum_{1}^{N} A_n \sin(2\pi k n f_0 / f s + \varphi_n) \tag{5}$$

where $f_0 = 216.66...Hz (= 3 \cdot 8/5200$ ms), is the fundamental frequency, and $f_s = 8$ kHz, the sampling frequency in GSM), $1 \leqslant n \leqslant 15$, and finally A_n and φ_n are the amplitude and phase of frequency component n, respectively.

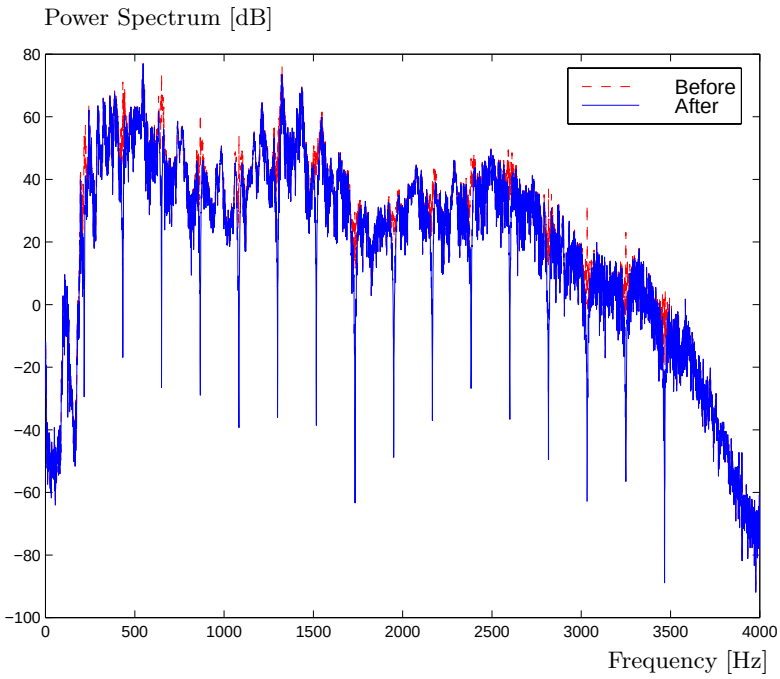


Figure 4: *Cancelation of the Bumblebee with notch filter in speech* **Full Rate**.

We again make use of our knowledge about the location of the idle frame in the PCM sample stream. This can be done since communication to the DSP during a call is performed with code and decode commands from the host ASIC. A code command requires a reply from the DSP containing speech coded data from the 160 latest received PCM samples. The code commands arrive to the DSP on the average every 20 ms. However, their exact time arrivals follow the pattern (18.465 ms, 18.465 ms, 23.070 ms), i.e., over a period of three code commands the average distance is 20 ms.

In order for the DSP to be able to synchronize its PCM buffers properly, the code commands contain information on the time to the next code command, the "syncInfo". This information can take six different values, and is carried in the code commands in a three bit field. When synchronizing the PCM buffers, we only make use of the two least significant bits in the field, since they contain sufficient information for that task.

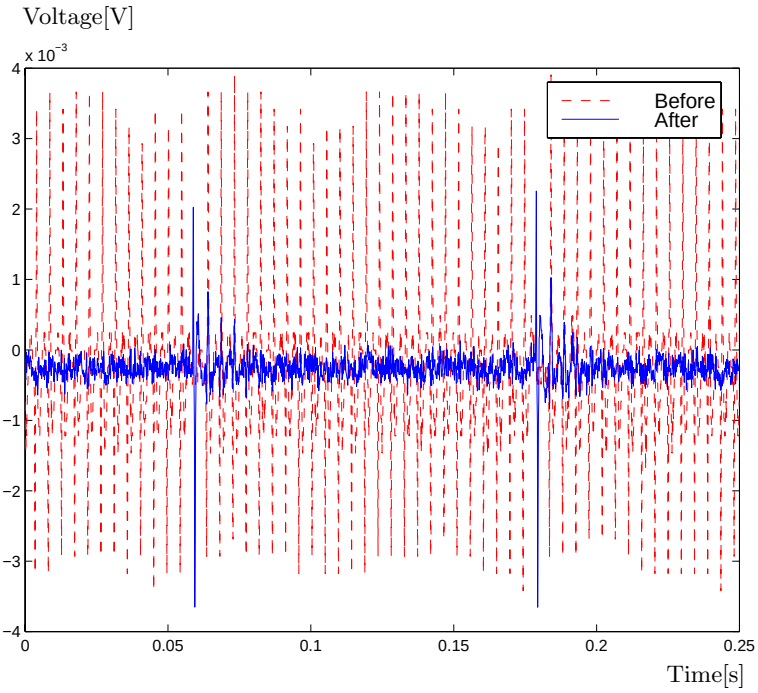


Figure 5: *Time signal of the simple notched Bumblebee. Observe residual in blue in idle slots, which is eliminated by advanced solution.*

The interesting fact about the "syncInfo" information is that each of the six possible numbers corresponds to a certain position in the 120 ms multi-frame structure. (Each multi-frame of 120 ms corresponds to six code commands to the DSP.) Thus, given the "syncInfo" information in the code commands, it is possible to calculate the position of the idle burst!

The basic idea is to now use two notch filters, that notch the bumblebee, and the "syncInfo", see Fig. 7. The difference between the notch filters is that one of the two filters has slightly larger notch-bandwidth. The first filter is only used to insert the distorsion during the idle slot, which was the problem with a single notch filter. These samples are then used to replace the samples in the original signal during the idle slot. The idle slot is located by using "syncInfo". This results in that the bumblebee signal gets present during the idle slot and from that it follows that the signal is periodic with the TDMA

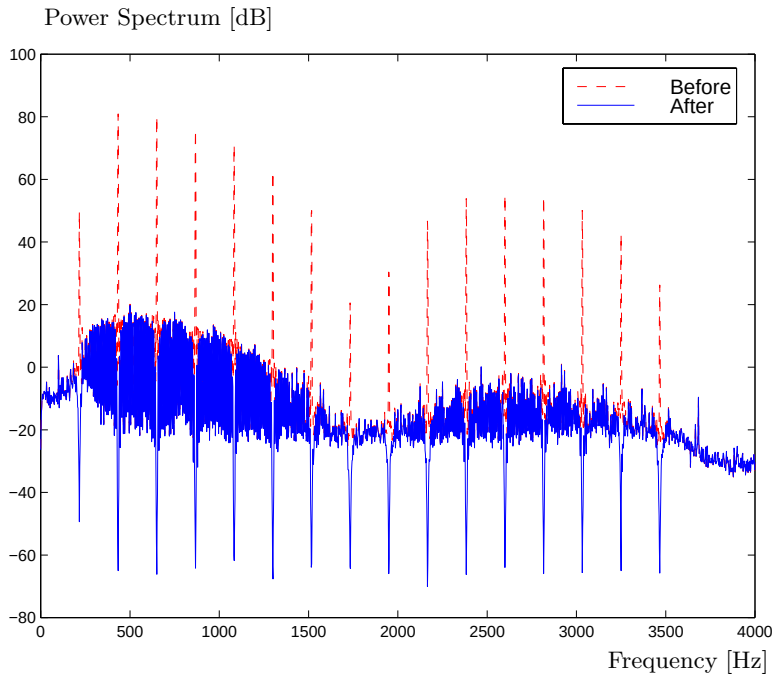


Figure 6: *Cancelation of the Bumblebee with notch filter. The Bumblebee was recorded in a silent room.* **Full Rate**, *no speech.*

frame rate. The second filter is then used to notch the new signal with the periodic bumblebee. The reason for the difference in bandwidth is to make sure that we do not adding any distorsion that is not suppressed.

The first filter in Fig. 7, denoted "Notch 1", has the smallest bandwidth. It is used to notch the input signal x(n). During the idle slot, the samples in x(n) are replaced by samples from the notched signal x'(n), containing the residual bumblebee ringing out from the states in the IIR filters. By changing these samples, the new signal (x"(n)) includes a complete bumblebee, even during the idle slot. The second filter, "Notch 2", then notches this signal, which suppresses the bumblebee without any residual disturbance and negligible distorsion.

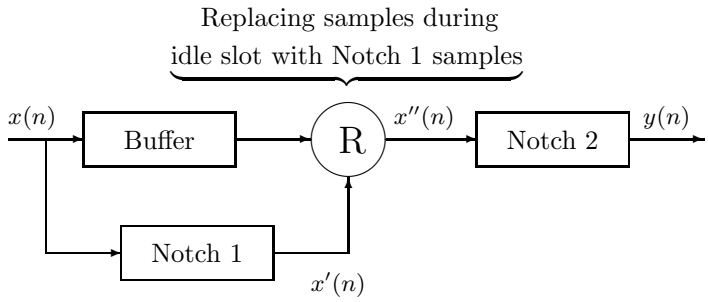


Figure 7: *Two-stage notch filtering.*

4 Summary and Conclusions

This paper presents two notch filter based solutions to reduce the humming disturbance in GSM mobile telephony. The first is a straight-forward solution with notch filters, reducing the disturbance considerably, but not totally. The second solution is a dual cascaded notch filter solution with internal knowledge of the GSM transmission pattern and transmitter state. With this method a full elimination of the "Bumblebee" can be achieved.
While the simple method is appropriate for exterior electronic equipment, the second more advanced cancelation is suited for internal cancelation in the mobile telephone.

References

[1] GSM Standard (GSM 05.01 version 7.0.0 Release 1998) *Digital cellular telecommunications system (Phase 2+), Physical layer on the radio path (General description).*

[2] B.Widrow, S. D. Stearns *Adaptive Signal Processing* Prentice Hall, 1985

[3] M. Kuo, D. R. Morgan, *Active Noise Control Systems*, John Wiley & Sons, Inc., 1996.

[4] I. Claesson, A. Nilsson GSM TDMA Frame Rate Internal Active Noise Cancellation. International Journal of Acoustics and Vibration (IJAV), vol. 8, no. 3, 2003.

[5] Proakis, J.G. and Manolakis, D.G. *Digital signal processing*, pp. 343-345, 1996, Prentice-Hall Inc.

[6] Simon Haykin *Digital communications*, 1988, John Wiley & Sons Inc.

[7] Peyton Z. Peebles, Jr. *Probability, random variables, and random signal principles*, 1993, McGraw-Hill Inc.

[8] Steven M. Kay *Fundamentals of statistical signal processing: Estimation Theory*, pp. 183-198, 1993, Prentice-Hall Inc.

[9] Per Eriksson *On estimation of the amplitude and the phase function (Technical Report TR-148)*, 1981, University of Lund / SWEDEN.

Part B

On Video Enhancement in Mobile Devices

Part B consists of:

Part B.1: Adaptive De-blocking De-Ringing Post Filter
Part B.2: Low-Complex Adaptive Post Filter for Enhancement of Coded Video

Part B.1

Adaptive De-Blocking De-Ringing Post Filter.

Part B.1 is published as:

A. Rossholm and K. Andersson, *Adaptive De-Blocking De-Ringing Post Filter.*, at International Conference on Image Processing (ICIP), September 2005.

Adaptive De-Blocking De-Ringing Post Filter.

Andreas Rossholm and Kenneth Andersson

Abstract

In this paper an adaptive filter for reducing blocking and ringing artifacts is presented. The solution is designed with consideration of Mobile Equipment with limited computational power and memory. Also, the solution is computationally scalable if there is limited CPU resources in different user cases.

1 Introduction

In the Mobile Equipment (ME) today the use of video becomes more and more common. To make it possible to view a video clip or streaming video, or to make a video telephony call, it is important to compress the data as much as possible. Most codecs, video enCOders and DECoders, used today are designed as block-based motion-compensated hybrid transform coders, like MPEG-4 and H.263, where the transformation is done by a Discrete Cosine Transforms (DCT) on blocks of 8x8 pixels. The reason to segment the image into 8x8-sized blocks is to exploit local characteristics of the images and to simplify the implementation.
One way for these kinds of codecs to reduce the bit rate is to change the strength of the quantization, on the encoder side. The quantization means that the DCT coefficients are divided with a fixed quantization parameter (QP). The quotient is then rounded to the nearest integer level to form a quantized coefficient. In the inverse quantization step, on the decoder side, the quantized coefficients are then multiplied by the quantization value to reproduce the real coefficients. However, the reproduced transform coefficients will differ from the original due to the quantization operation. This difference or error is referred to as the "quantization error".

Two of the main artifacts from the quantization of the DCT are blocking and ringing. Blocking artifacts are also due to motion compensation. The blocking artifact is seen as an unnatural discontinuity between pixel values of neighboring blocks. The ringing artifact is seen as high frequency irregularities around the image edges. In brief; the blocking artifacts are generated due to the blocks being processed independently, and the ringing artifacts due to the coarse quantization of the high frequency components [1], see Fig. 1. To reduce blocking artifacts, two-dimensional (2D) low-pass filtering of pixels

Figure 1: *Example on blocking and ringing artifacts on a frame from the Foreman sequence at QCIF (*176×144*)*

on block boundaries of the decoded image(s) was suggested in [2]. The 2D space-invariant static filtering described in that paper reduces blocking artifacts but can also introduce blurring artifacts when true edges in the image are low-pass filtered.

To avoid blurring of true edges in the image and also to be computationally efficient, the amount of low-pass filtering may be controlled by table-lookup as described in [3]. Large differences between initial pixel values and filtered pixel values are seen as natural image structure, and thus filtering is weak so that the image is not blurred. Small pixel differences are seen as coding artifacts, and thus stronger filtering is allowed to remove the artifacts. Based on data from other equipment, the amount of filtering can be controlled by using additional filter tables. The algorithm modifies the output of a low-pass-filtered signal with the output of a table-lookup using the difference between a delayed input signal and the filtered signal as an index into the table, and different degrees of filtering are achieved by only providing additional tables. A combined de-blocking and de-ringing filter was proposed in [4]. The pro-

posed filter used filter strengths on block boundaries that were different from filter strengths inside blocks, allowing for stronger filtering at block boundaries than inside blocks. This was achieved by using a metric that used different constants when computing the output values of block boundary pixels versus the output values of pixels inside the block boundary. The metric also included the QP value.
These and most other current algorithms handle de-blocking and de-ringing artifacts sequentially. This requires filtering in two steps to handle both artifacts, e.g., first process a decoded image with a de-blocking filter to remove artifacts on block boundaries, and then apply a de-ringing filter to remove ringing artifacts. Such double filtering can have a negative impact on computational complexity and memory consumption, which are parameters of particular importance in many devices, such as mobile communication devices.
Moreover, removal of blocking and ringing artifacts can add visually annoying blurring artifacts as described above. It is thus important to be careful with strong image features that likely are natural image features and not coding artifacts.

2 The Adaptive Filter

The proposed filter is developed with two main considerations; limiting the computational complexity, and limiting the amount of working memory. The idea is to filter rows of pixels of an image in a vertical direction, store the results in row vectors, and then filters the row vectors in the horizontal direction, and display the results. In the following part the adaptive filter is described in one of the above directions. Coefficients of a reference filter are modified based on the output from the reference filter passed through a table-lookup process that accesses a table of modifying weight coefficients. The output of the modified filter is added to a delayed version of the input to provide the adaptive filter output. A block diagram of the adaptive filter is shown in Fig. 2.

2.1 Overview of the filter

In Fig. 2 an input stream of pixel data is provided to a switch that directs the input pixels to either the output of the filter or to a delay element and a reference filter. The operation of the switch is responsive to additional data,

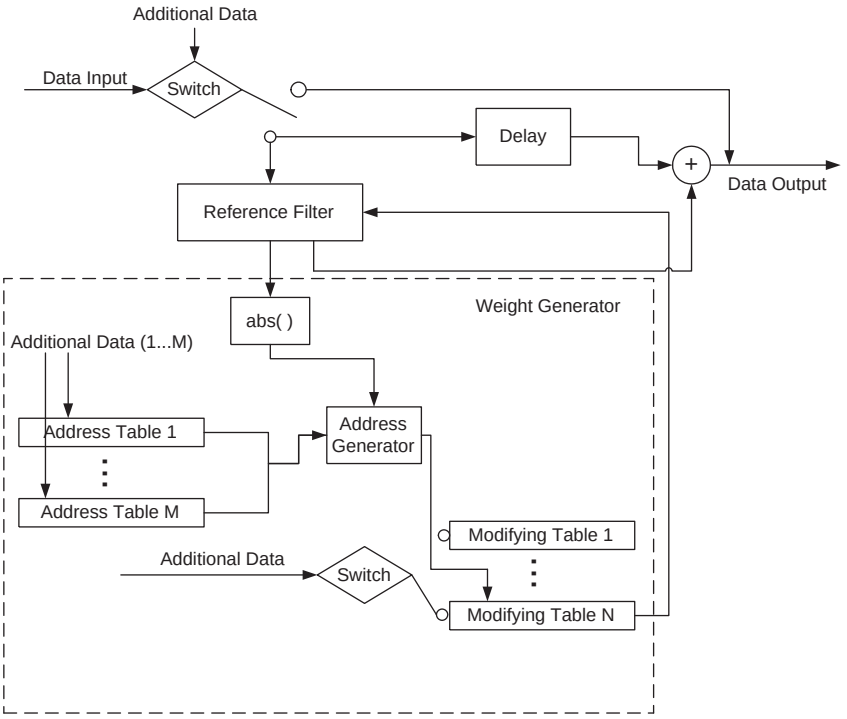


Figure 2: *A block diagram of the adaptive filter.*

in particular, whether the input pixels belong to an error-concealed block or if the amount of filtering is limited based on location in the frame, as described in more detail below. The reference filter has coefficients that determine the filtering function, and these coefficients are selectively modified. The output of the reference filter is provided to an adder that combines the output with the delayed input produced by the delay element, thereby generating the output of the adaptive filter.

The modification of the output of the reference filter is performed by a weight generator that produces weights that selectively modify the coefficients of the filter based on the filter output to the weight generator. A signal corresponding to the absolute value of the reference filter output is produced, and this

signal is provided to an address generator. The absolute value together with additional data provided by M suitable address tables, generates addresses into N tables of modifying weight coefficients, as described in more detail below. As a set of modifying weight coefficients is retrieved from the selected table, it is provided by the weight generator to the filter, and the transfer function of the reference filter is modified accordingly. Through this modification, the filter adapts to the input stream of pixels.

2.2 Reference Filter

In the adaptive filter a 5-tap reference filter is used, [1 1 − 4 1 1]. The number of filter taps chosen is the result of a trade-off between the amount of low-pass filtering that can be performed, locality in filtering, and computational complexity. The filter coefficients are chosen to detect variations in pixel value in the filter neighbourhood with as low complexity as possible. The same filter is used for filtering luminance, denoted Y, and chrominance blocks, denoted Cb and Cr, although luminance blocks are more important to filter than chrominance blocks. The modification of the reference filter is performed with a set of modifying weights and the resulting adaptive filter response is illustrated in Fig. 3a. It can be seen that the same modification is made to each coefficient. In the figure, the sign and magnitude of a filter coefficient or a weight are indicated by the length of the respective vertical line segment and its position above or below the horizontal reference line. The "+" sign indicates the operation of the adder. In Fig. 3b the modifying weight is shown as 0.5 and the other coefficients are fixed. Comparing Fig. 3a and Fig. 3b, it will be seen that a "weaker" adaptive filter is achieved when the reference filter coefficients are scaled by a factor of 0.5, i.e., neighboring pixels have less influence on the modified-filter output for a pixel.
If the modifying weights are such that all filter coefficients are modified in the same way (see, e.g., FIG. 3b), also used in this implementation, the output of the modified reference filter is simply a scaling of the output of the unmodified reference filter. Otherwise, the output of the modified reference filter is calculated using the input pixels and the modified reference filter transfer function.

2.3 Weight Generator

The weight generator handles the adaptive part of the filter. It is divided into three main parts:

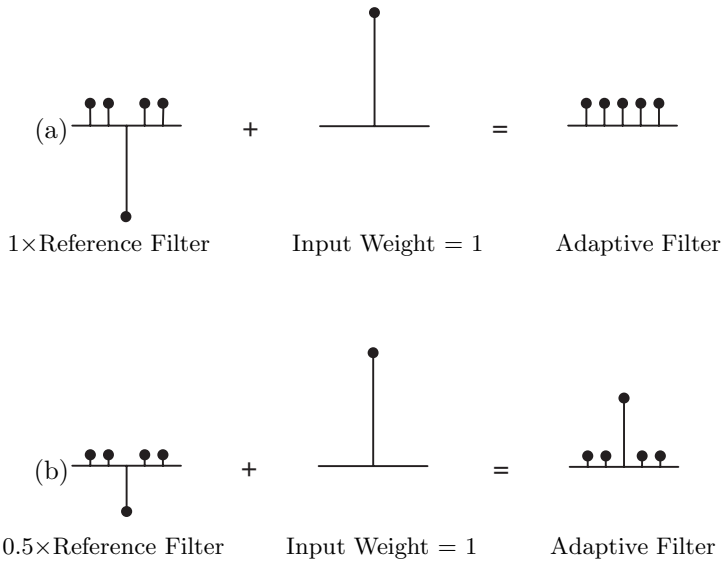


Figure 3: *Depiction of reference filter modification and adaptive filter response.*

1 The address tables with additional data.

2 The address generator.

3 Modifying tables with switch and additional data.

The first part, address table, uses the QP for the block as additional data and the address table length correspond to the range of the QP data. The output from the address table are positive for low QP values and negative for high QP values, resulting in potentially weaker and stronger filtering, respectively, depending on the magnitude of the reference filter output. Several address tables can be used if different sessions needs different strength of filtering.
The second part, address generator, produces a signal corresponding to the absolute value of the filter output, together with the output from the address tables, to generate addresses into one of the modifying tables with weight coefficients. The address generator also check the validity of the addresses generated, confirming that an address is inside the range of the modifying tables.

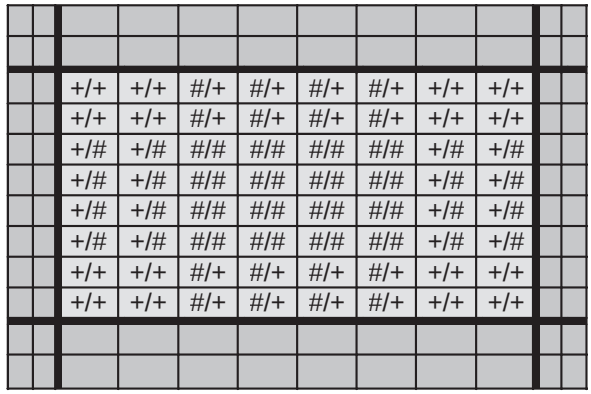


Figure 4: *Depiction of a block of pixels.*

The third part, modifying tables, provides sets of weight coefficients to modify the transfer function of the reference filter, resulting in a modified, or adapted, transfer function for the adaptive filter as described in subsection 2.2. The length (i.e., the address range) of a modifying table corresponds to the range of the reference filter output. In this implementation small address values give weights close to 1/5 and large address values give weights close to 1/260. The result is thus variation from flat low-pass filtering to very weak low-pass filtering over the filter output range.

The additional data that is input to the switch, selecting modification table, is based on the position of a pixel in its block. As indicated by Fig. 4, which depicts a block of pixels, outer boundary pixels (indicated by + in the figure) select a table that corresponds to stronger filtering than the table selected for inner block pixels. Furthermore, the weights of the selected table for inner pixels (indicated by # in Fig. 4) decreases more quickly with increasing index than the weights in the boundary pixels table. This results in reduction of blocking and ringing artifacts without blurring the image too much. x/y in Fig. 4 describes filtering in horizontal/vertical direction.

2.4 Further considerations

The first switch in Fig.2 makes it possible to limit the amount of filtering for different combinations of applications for a given device. The priority of filtering is given from low to high priority as, all luminance and chrominance blocks may be filtered, only luminance blocks may be filtered, outer boundary pixels may be filtered, and only block border pixels may be filtered.

3 Results

The performance of the adaptive filter is evaluated against using no post filtering and filtering as recommended in H.263 App. III [4]. The algorithms are processed on decoded H.263 profile 0 bit streams for two different sequences each presented at four different bit rates at 15 frames per second (fps) and of size 176×144 (QCIF). The size, bitrates and framerate are chosen to correspond with the use in todays 2G and 3G networks. The peak signal-to-noise ratio (PSNR) is calculated for the post processed images and an average for the complete sequence. The PSNR of an 8-bit $M \times N$ image is given by

$$PSNR = 10 \log \frac{MN \times 255^2}{\sum_{m,n} \|f(m,n) - f_{org(m,n)}\|^2}$$

The sequences used are "Foreman" and "Mother and Daughter", presented in Table 1 and Table 2.

In the tables it is shown that the adaptive filter always keeps or increases the PSNR compared to the original decoded sequences. The adaptive filter gives significantly better visual quality as can be seen in Fig. 5 and Fig. 6. As shown in the tables H.263 App. III gives slightly higher PSNR than the adaptive filter but also gives somewhat blurred results compared to the adaptive filter, see Fig. 5 and Fig. 6. It shall also be noted that the complexity of the adaptive filter is about 18 cycles per filtered pixel including 2 multiplications, 10 additions, 4 shifts and 2 abs, which is significantly lower than for the H.263 App. III filter. H263 App. III will require at least 34 cycles per filtered pixel including 4 divisions, 14 multiplications and 16 additions.

Foreman			
Bitrate [kbit/s]	**Filter**	**Average PSNR [dB] for YCbCr**	**Average PSNR [dB] for Y**
32	No Post Filter	30.0038	28.6846
32	H.263 App. III	30.0960	28.7631
32	Adaptive Filter	30.0482	28.7219
48	No Post Filter	31.0114	29.7355
48	H.263 App. III	31.1552	29.8716
48	Adaptive Filter	31.0611	29.7799
64	No Post Filter	31.8652	30.6329
64	H.263 App. III	32.0412	30.8038
64	Adaptive Filter	31.9159	30.6797
128	No Post Filter	34.4384	33.3140
128	H.263 App. III	34.6722	33.5539
128	Adaptive Filter	34.4716	33.3475

Table 1: *Results from de-blocking and de-ringing on Foreman. All sequences have a QCIF resolution and 15 fps.*

Mother and Daughter			
Bitrate [kbit/s]	**Filter**	**Average PSNR [dB] for YCbCr**	**Average PSNR [dB] for Y**
32	No Post Filter	34.3828	33.2452
32	H.263 App. III	34.5634	33.4313
32	Adaptive Filter	34.4442	33.2975
48	No Post Filter	35.6795	34.5960
48	H.263 App. III	35.8562	34.7795
48	Adaptive Filter	35.7138	34.6301
64	No Post Filter	36.6913	35.6550
64	H.263 App. III	36.8635	35.8284
64	Adaptive Filter	36.6943	35.6641
128	No Post Filter	39.4050	38.5265
128	H.263 App. III	39.5368	38.6457
128	Adaptive Filter	39.4314	38.5492

Table 2: *Results from de-blocking and de-ringing on Mother and Daughter. All sequences have a QCIF resolution and 15 fps.*

Figure 5: *Luminance output from Foreman in QCIF format, coded at* 32 *kbps and* 15 *fps. From left, No Post Filter PSNR* 28.30 *dB, H.263 App. III PSNR* 28.51 *dB, Adaptive Filter PSNR* 28.40 *dB.*

Figure 6: *Luminance output from Mother and Daughter in QCIF format, coded at* 32 *kbps and* 15 *fps. From left, No Post Filter PSNR* 34.20 *dB, H.263 App. III PSNR* 34.35 *dB, Adaptive Filter PSNR* 34.22 *dB.*

4 Conclusion

This paper has described an adaptive filter that can improve visual quality by combating both de-blocking and de-ringing artifacts as generated by standard block based coders. The filter has further low complexity and can be used in MEs with limited computational power and memory.

References

[1] Michael Yuen, H.R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, pp. 247-278, July 1998.

[2] H. C. Reeve III, Jae S. Lim, "Reduction of Blocking Effect in Image Coding," *Proc. ICASSP*, pp. 1212- 1215, Boston, Mass. 1983.

[3] US patent No. 5,488,420 to G. Bjontegaard for "Cosmetic filter for smoothing regenereted pictures, , e.g. after Signal Compression for Transmission in a Narrowband Network".

[4] ITU-T Recommendation H.263 Appendix III: "Examples for H.263 Encoder/Decoder Implementations," June 2000.

Part B.2

Low-Complex Adaptive Post Filter for Enhancement of Coded Video

Parts of Part B.2 has been published as:

A. Rossholm, K. Andersson, and B. Lövström, *Low-Complex Adaptive Post Filter for Enhancement of Coded Video.*, at International Symposium on Signal Processing and its Applications (ISSPA), February 2007.

Low-Complex Adaptive Post Filter for Enhancement of Coded Video

Andreas Rossholm, Kenneth Andersson, and Benny Lövström

Abstract

In this paper an adaptive filter that removes de-blocking and de-ringing artifacts and also enhances the sharpness of decoded video is presented. The solution is designed with consideration of Mobile Equipment with limited computational power and memory. Also, the solution is computationally scalable to be able to handle limited computational resources in different user cases. In the paper it is shown that the adaptive filter always keeps or increases the image quality, compared to the original decoded sequences, and that the amount of sharpening decreases with an decrease of bit-rate to limit amplification of coding artifacts or noise.

1 Introduction

In the Mobile Equipment (ME) today the use of video becomes more and more common. To make it possible to view a video clip or streaming video, or to make a video telephony call, it is important to compress the data as much as possible. Most video codecs, video enCOders and DECoders, used today are designed as block-based motion-compensated hybrid transform coders, like MPEG-4 and H.263, where the transformation is done by a Discrete Cosine Transforms (DCT) on blocks of 8x8 pixels. The DCT coefficients are quantized with a quantization parameter (QP). Two of the main artifacts from the quantization of the DCT are blocking and ringing. Blocking artifacts are also due to motion compensation. The blocking artifact is seen as an unnatural discontinuity between pixel values of neighboring blocks. The ringing artifact is seen as high frequency irregularities around the edges in the image. In brief; the blocking artifacts are generated due to the blocks being processed

independently, and the ringing artifacts due to the coarse quantization of the high frequency components [2].
To reduce blocking artifacts, two-dimensional (2D) low-pass filtering of pixels on block boundaries of the decoded image(s) was suggested in [3]. The 2D space-invariant static filtering described in that paper reduces blocking artifacts but can also introduce blurring artifacts when true edges in the image are low-pass filtered. To avoid blurring of true edges in the image and also to be computationally efficient, the amount of low-pass filtering may be controlled by table-lookup as described in [4]. Large differences between initial pixel values and filtered pixel values are seen as natural image structure, and thus filtering is weak so that the image is not blurred. Small pixel differences are seen as coding artifacts, and thus stronger filtering is allowed to remove the artifacts. Based on data from other equipment, the amount of filtering can be controlled by using additional filter tables. The algorithm modifies the output of a low-pass-filtered signal with the output of a table-lookup using the difference between a delayed input signal and the filtered signal as an index into the table, and different degrees of filtering are achieved only by providing additional tables. A combined de-blocking and de-ringing filter was proposed in [5]. The proposed filter used filter strengths on block boundaries that were different from filter strengths inside blocks, allowing for stronger filtering at block boundaries than inside blocks. This was achieved by using a metric that used different constants when computing the output values of block boundary pixels versus the output values of pixels inside the block boundary. The metric also included the QP value. These and most other current algorithms handle de-blocking and de-ringing artifacts sequentially. Such double filtering can have a negative impact on computational complexity and memory consumption, which are parameters of particular importance in many devices, such as mobile communication devices. Moreover, removal of blocking and ringing artifacts can add visually annoying blurring artifacts as described above. It is thus important to be careful with strong image features that likely are natural image features and not coding artifacts. In [6] an adaptive non-linear filter is proposed. The proposed filter handles both the coding artifacts and performs sharpening on true details. However, this filter uses a rational function for the control of the filter function based on measures of variance. This gives good results but is a too complex solution for implementation in a ME.
In this paper, we propose a filter that performs enhancement on the coded video stream including both de-blocking, de-ringing and sharpening based on the output from a reference filter, which requires much less computational power than the state of art approach. This filter is a further development of

our adaptive de-blocking and de-ringing filter published in [1].

2 The Adaptive Filter

The proposed filter is developed with two main considerations; limiting the computational complexity, and limiting the amount of working memory. The idea is to filter rows of pixels of an image in a vertical direction, store the results in row vectors, and then filter the row vectors in the horizontal direction, and display the results. In the following part the adaptive filter is described in one of the above directions. Coefficients of a reference filter are modified based on the output from the reference filter passed through a table-lookup process that accesses a table of modifying weight coefficients. The output of the modified filter is added to a delayed version of the input to provide the adaptive filter output. A block diagram of the adaptive filter is shown in Fig. 1.

2.1 Overview of the filter

In Fig. 1 an input data stream of pixel data is provided to a switch that directs the input pixels to either the output of the filter or to a delay element and a reference filter. The operation of the switch is responsive to additional data, in particular, whether the input pixels belong to an error-concealed block or if the amount of filtering is limited based on location in the frame, as described in more detail below. The reference filter has coefficients that determine the filtering function, and these coefficients are selectively modified. The output of the reference filter is provided to an adder that combines the output with the delayed input produced by the delay element, thereby generating the output of the adaptive filter.

The modification of the output of the reference filter is performed by a weight generator that produces weights that selectively modify the coefficients of the filter based on the filter output to the weight generator. A signal corresponding to the absolute value of the reference filter output is produced, and this signal is provided to an address generator. The absolute value together with additional data provided by M suitable address tables, generates addresses into N tables of modifying weight coefficients, as described in more detail below. As a set of modifying weight coefficients is retrieved from the selected table, it is provided by the weight generator to the filter, and the transfer function of the reference filter is modified accordingly. Through this modification,

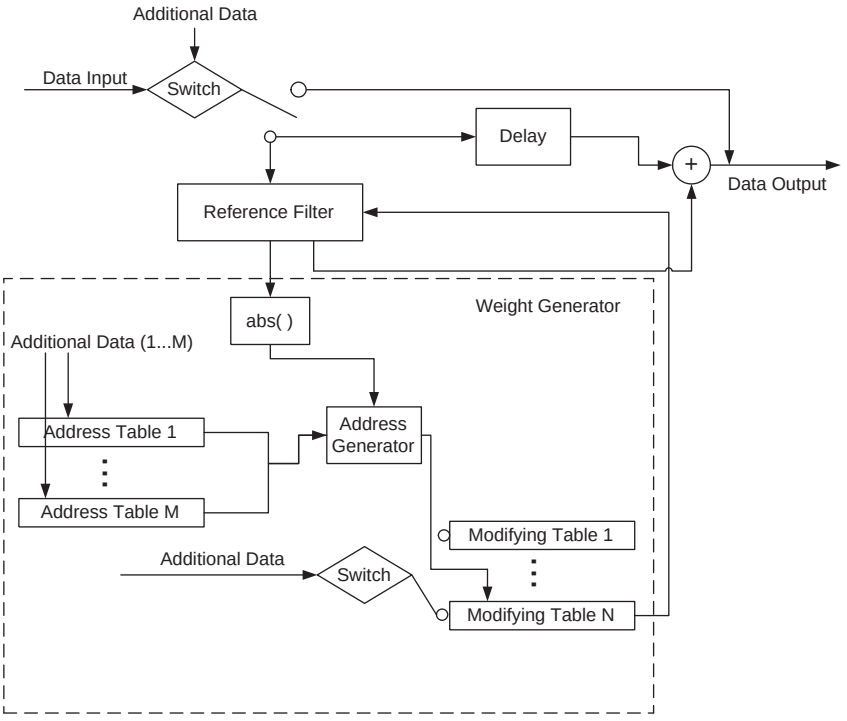


Figure 1: *A block diagram of the adaptive filter.*

the filter adapts to the input stream of pixels.

2.2 Reference Filter

In the adaptive filter a 5-tap reference filter is used, $[1 \quad 1 \quad -4 \quad 1 \quad 1]$. The number of filter taps chosen is the result of a trade-off between the amount of low-pass filtering that can be performed, locality in filtering, and computational complexity. The filter coefficients are chosen to detect variations in pixel value in the filter neighborhood with as low complexity as possible. The same filter is used for filtering luminance, denoted Y, and chrominance blocks, denoted Cb and Cr, although luminance blocks are more important to filter than chrominance blocks. The modification of the reference filter is per-

formed with a set of modifying weights which results in de-blocking/de-ringing or sharpening. If the modifying weights are such that all filter coefficients are modified in the same way, the output of the modified reference filter is simply a scaling of the output of the unmodified reference filter. Otherwise, the output of the modified reference filter is calculated using the input pixels and the modified reference filter transfer function.

2.2.1 De-blocking and De-ringing

For de-blocking and de-ringing the resulting adaptive filter response is illustrated in Fig. 2a. It can be seen that the same modification is made to each

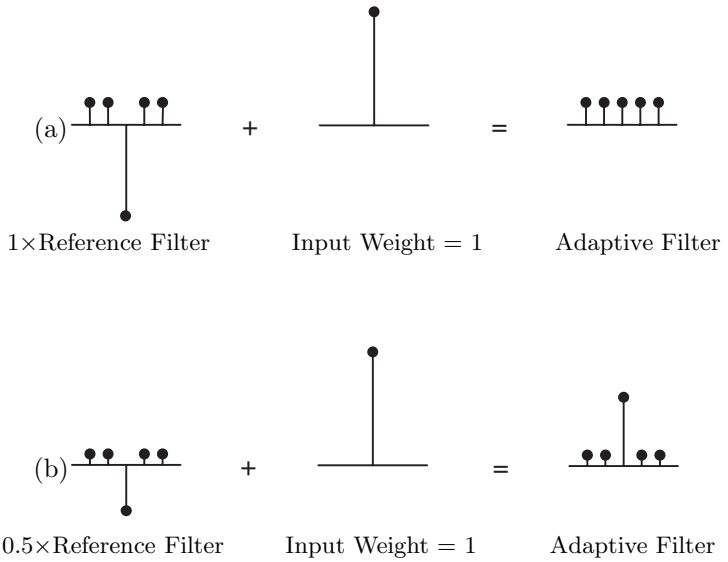


Figure 2: *Depiction of reference filter modification and adaptive filter response.*

coefficient. In the figure, the sign and magnitude of a filter coefficient or a weight are indicated by the length of the respective vertical line segment and its position above or below the horizontal reference line. The "+" sign indicates the operation of the adder. In Fig. 2b the modifying weight is shown as 0.5 and the other coefficients are fixed. Comparing Fig. 2a and Fig. 2b, it will be seen that a "weaker" adaptive filter is achieved when the reference

filter coefficients are scaled by a factor of 0.5, i.e., neighboring pixels have less influence on the modified-filter output for a pixel.

2.2.2 Sharpening

For sharpening the same concept as described above can be used by changing the sign of the reference filter, which generates a high-pass filter compared to the above-described low-pass filter. This is illustrated in Fig. 3. Comparing

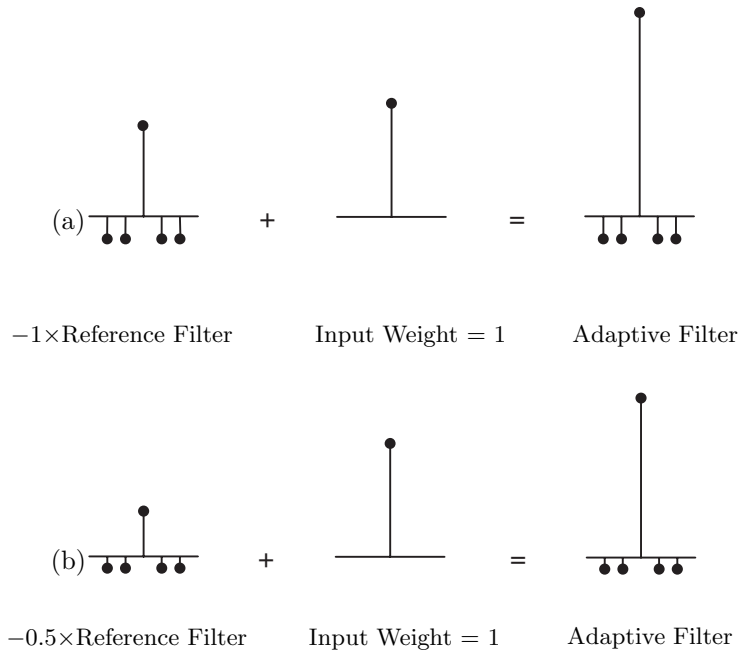


Figure 3: *Depiction of reference filter modification and adaptive filter response.*

this figure with Fig. 2 the difference between an adaptive sharpening, or high-pass, filter and an adaptive low-pass filter will be recognized. As in Fig. 2 the Fig. 3 depicts how modification of the coefficients of the reference filter with a set of modifying weights modifies the adaptive filter response. In the case illustrated by the figure, the same modification is made to each coefficient. In Fig. 3a, the modifying weight is shown as -1 and the other coefficients

are fixed. In Fig. 3b, the modifying weight is shown as -0.5 and the other coefficients are fixed. Comparing Fig. 3a and Fig. 3b, it will be seen that a "weaker" adaptive filter is achieved when the reference filter coefficients are scaled by a smaller negative factor, i.e., neighboring pixels have less influence on the modified-filter output for a pixel.

2.3 Weight Generator

The weight generator handles the adaptive part of the filter. It is divided into three main parts:

1 The address tables with additional data.

2 The address generator.

3 Modifying tables with switch and additional data.

The first part, address table, uses the QP for the block as additional data and the address table length correspond to the range of the QP data. The output from the address table are positive for low QP values and negative for high QP values, resulting in potentially weaker and stronger filtering, respectively, depending on the magnitude of the reference filter output. Several address tables can be used if different sessions needs different strength of filtering.
The second part, address generator, produces a signal corresponding to the absolute value of the filter output, together with the output from the address tables, to generate addresses into one of the modifying tables with weight coefficients. This input also determines wether the filter is low-pass or high-pass, de-blocking/de-ringing or sharpening.
The third part, modifying tables, provides sets of weight coefficients to modify the transfer function of the reference filter, resulting in a modified, or adapted, transfer function for the adaptive filter as described in subsection 2.2. The length (i.e., the address range) of a modifying tables corresponds to the range of the reference filter output.
The additional data that is input to the switch, selecting modification table, is based on the position of a pixel in its block. A block of pixels is illustrated in FIG. 4 where the outer boundary pixels are indicated by + and the inner pixels are indicated by #. Also, the x/y in FIG. 4 describes filtering in horizontal/vertical direction.
In the de-blocking and de-ringing case stronger filtering is performed on border pixels than the table selected for inner block pixels. Furthermore, the

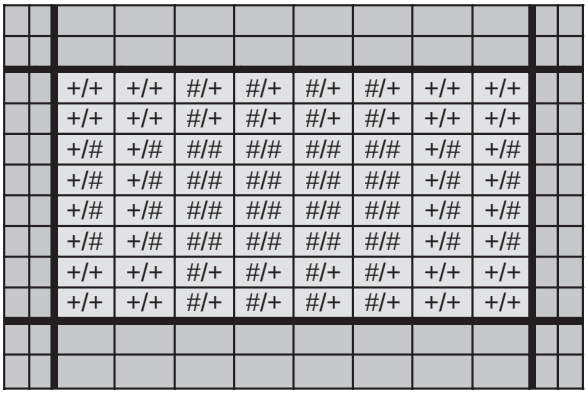


Figure 4: *Depiction of a block of pixels.*

weights of the selected table for de-blocking and de-ringing decreases more quickly with increasing index than the weights in the boundary pixels table. This results in reduction of blocking and ringing artifacts without blurring the image too much. For sharpening more weights are given to the inner pixels and thereby sharpening central parts of the block with normally less prominent coding artifacts. The block boundary pixels are not sharpened at all or only weakly sharpened to avoid amplifying block artifacts.
The resulting filter characteristics can hereby vary with the absolute value of the reference filter output by using QP as additional data in tables and position of a pixel in its block gradually changing from strong low-pass filtering (large positive weight), to weak low-pass filtering (small positive weight), to weak high-pass filtering (small negative weight), and to strong high-pass filtering (large negative weight). Weak or all-pass filtering (small negative/positive or zero weight) is implemented when the reference filter output magnitude is relatively large.

2.4 Further considerations

The first switch in Fig.1 makes it possible to limit the amount of filtering for different combinations of applications for a given device. The priority of filtering is given from low to high priority as, all luminance and chrominance

blocks may be filtered, only luminance blocks may be filtered, outer boundary pixels may be filtered, and only block border pixels may be filtered.

3 Results

In [2] the performance of the de-blocking and de-ringing part of the adaptive filter was evaluated against using no post filtering and filtering as recommended in H.263 App. III [5]. It was shown that the adaptive filter improves visual quality by combating both de-blocking and de-ringing artifacts and also that the peak signal-to-noise ratio (PSNR) was comparable with the results from using the H.263 App. III filter.
Here the adaptive filter, including the sharpening part, is evaluated by examining the PSNR value and the perceptual quality both against; No filtering, only de-blocking and de-ringing. The algorithms are processed on decoded H.263 profile 0 bit streams for two different sequences, each presented at four different bit rates at 15 frames per second (fps) and of size 176×144 (QCIF). The size, bit-rates and frame rate are chosen to correspond to the use in todays 2G and 3G networks. The PSNR is calculated for the post processed images as an average for the complete sequence. The PSNR of an 8-bit $M \times N$ image is given by

$$PSNR = 10 \log \frac{MN \times 255^2}{\sum_{m,n} \|f(m,n) - f_{org(m,n)}\|^2}$$

The sequence used is "Foreman" and the results are shown in Table 1. In the table it is shown that the adaptive filter always keeps or increases the PSNR for low bit-rates compared to the original decoded sequences and that the PSNR slightly decreases when the amount of sharpening increases which is notified for the higher bit-rates. However, the perceptual quality increases for these bit-rates, visualized in Fig. 5-7. In Fig. 5 there is very little sharpening performed and therefore almost no visible effects can be seen. In Fig. 6 and Fig. 7 the sharpening effects are more obvious and there is an increase of perceptual quality even though the decrease of PSNR.

Foreman		
Bitrate [kbit/s]	**Filter**	**Average PSNR [dB] for YCbCr**
48	No Post Filter	33.078
48	Adaptive Post Filter [2]	33.129
48	Proposed Adaptive Filter	33.077
64	No Post Filter	33.481
64	Adaptive Post Filter [2]	33.548
64	Proposed Adaptive Filter	33.478
128	No Post Filter	35.795
128	Adaptive Post Filter [2]	35.875
128	Proposed Adaptive Filter	35.138
196	No Post Filter	37.508
196	Adaptive Post Filter [2]	37.562
196	Proposed Adaptive Filter	36.442

Table 1: *Results from de-blocking and de-ringing on Foreman. All sequences have a QCIF resolution and 15 fps.*

Figure 5: *Luminance output from Foreman in QCIF format, coded at* 64 *kbps and* 15 *fps. From left, No Post Filter, Adaptive Post Filter [2], and Proposed Adaptive Filter.*

Figure 6: *Luminance output from Foreman in QCIF format,coded at* 128 *kbps and* 15 *fps. From left, No Post Filter, Adaptive Post Filter [2],and Proposed Adaptive Filter.*

Figure 7: *Luminance output from Foreman in QCIF format, coded at* 196 *kbps and* 15 *fps. From left, No Post Filter, Adaptive Post Filter [2], and Proposed Adaptive Filter*

4 Conclusion

This paper has described an adaptive filter that can remove de-blocking and de-ringing artifacts and also enhance the sharpness of decoded video. The adaptive filters described here uses the reference filter output to control the filter function, which gives a low computational power and memory consumption. Experiments show an increase in perceptual quality, and especially for high bit-rate video the sharpening effect is obvious.

References

[1] A. Rossholm and K. Andersson, "Adaptive De-blocking De-ringing Filter," *IEEE International Conference on Image Processing 2005*, pp. 1042-5., Genoa, Italy, 2005.

[2] M. Yuen, H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, pp. 247-278, July 1998.

[3] H. C. Reeve III, Jae S. Lim, "Reduction of Blocking Effect in Image Coding," *Proc. ICASSP*, pp. 1212- 1215, Boston, Mass. 1983.

[4] US patent No. 5,488,420 to G. Bjontegaard for "Cosmetic filter for smoothing regenereted pictures, , e.g. after Signal Compression for Transmission in a Narrowband Network".

[5] ITU-T Recommendation H.263 Appendix III: "Examples for H.263 Encoder/Decoder Implementations," June 2000.

[6] G. Scognamiglioa, G Ramponia, A. Rizzi, "Enhancement of coded video sequences via and adaptive non-linear post- processing," *Image Communications*, vol. 18, pp. 127-139, 2003

Part C

On Audio and Video Delay and Synchronization Measurement

Part C consists of:

Part C.1: A Robust Method for Estimating Synchronization and Delay of Audio and Video for Communication Services

Part C.1

A Robust Method for Estimating Synchronization and Delay of Audio and Video for Communication Services

Part C.1 has been accepted as:

A. Rossholm, and B. Lövström, *A Robust Method for Estimating Synchronization and Delay of Audio and Video for Communication Services.*, accepted for publication in Multimedia Tools and Applications (DOI: 10.1007/s11042-014-2306-6), October 2014.

A Robust Method for Estimating Synchronization and Delay of Audio and Video for Communication Services

Andreas Rossholm, and Benny Lövström

Abstract

One of the main contributions to the quality of experience in streaming services or in two-way communication of audio and video applications is synchronization. This has been shown in several studies and experiments but methods to measure synchronization are less frequent, especially for situations without internal access to the application and independent of platform and device. In this paper we present a method for measuring synchronization skewness as well as delay for audio and video. The solution incorporates audio and video reference streams, where audio and video frames are marked with frame numbers which are decoded on the receiver side to enable calculation of synchronization and delay. The method has been verified in a two-way communication application in a transparent network with and without inserting known delays, as well as in a network with 5% and 10% packet loss levels. The method can be used for both streaming and two-way communication services, both with and without access to the internal structures, and enables measurements of applications running on e.g. smartphones, tablets, and laptops under various conditions.

1 Introduction

The widespread use of video communication and video streaming in a growing number of application areas gives emphasis to the topics of Quality of Service (QoS) and Quality of Experience (QoE). Several factors influence the QoE, including source quality, encoding degradation, network behaviour, and decoder and rendering performance. The methods to measure the quality, especially in

a quantitative way, is a large research topic having many aspects. One of the most important factors for the audiovisual quality is synchronization of audio and video. This is of high importance when it comes to streaming services as well as real time applications for one-way or two-way communication of audio and video. Several studies have been published focused on both the effects of audio and video skewness and on reference models to handle synchronization in different ways, at IP level as well as application level [7], [1], [23], [25]. It is shown that viewers perceive audio and video to be synchronized with an audio-to-video skew up to about 80 ms but also that there is a higher tolerance for video ahead of audio than vice versa. Further, the type of content as well as the quality of the video, e.g. video resolution, quantization level, and frame rate are also impacting the perceived skewness [22], [6]. In the case of real time two-way communication another important contribution to the quality of experience is delay. The delay can impact synchronization especially in the case of separate audio and video streams, but also the delay itself has impact on the QoE of users sharing information instantly and continuously [24]. It has been shown that a delay less than 100-150 ms is preferred and above 400 unacceptable, which also is stated in several specifications [5]. However, recent studies on speech have shown that interactivity has a big impact on the perceived quality and that people can adapt to the current situation [18] compared to quality scale used in [12].

To evaluate synchronization and delay for video communication applications like Microsoft's Skype, Google's Hangouts, Apple's FaceTime it is required that several issues are taken into account to get a complete evaluation and support different kind of scenarios, e.g. different network conditions, different platforms and devices running under different constraints with different operating systems. These requirements results in the need for a method that is robust to e.g. different packet losses and jitter as well as different compressions levels and CPU constraints. In this paper a novel, robust, out-of-service measurement method is presented, and since it is a standalone application it is supporting both different platforms and different devices. The method can be used for measuring both synchronization between audio and video and delay. The method uses a pre-generated test signal fed into the sender's audio and video input, and on the receiver side the audio and video output are captured and processed. However, it would also be possible to apply the prestored frame codes to the incoming audio and video signals on the sender side to construct an in-service measurement which would enable synchronization measurement at the receiver side.

The paper is organized as follows. A technical background is given in

Section 2, and in Section 3 published work related to this paper is discussed. In Section 4 the proposed method is presented including a more detailed description of the audio and video stamps and how the detection of the reference signal is performed. In Section 5 a description of a proof of concept is given and in Section 6 the results are presented. Finally, in Section 7 summary and conclusions are given.

2 Technical background

When a video sequence with related audio content is streamed over a network there are two main systems that can cause audio and video to be out of sync, or skewed. One is the transport over the network, and the other is the sending and receiving equipment which usually processes audio and video separately. The transport employed in most streams today is packet-based, using Internet Protocol (IP) with for e.g. User Datagram Protocol (UDP) and Real-time Transport Protocol (RTP) or Transmission Control Protocol (TCP), where audio and video can be handled in different paths as well as multiplexed. In transmission of audio and video data over a network a number of trade-off decisions are made, such as between getting acceptable delays or have a low packet loss rate and jitter, or considerations regarding bitrates, frame-rates and resolution. This all affects the quality the user finally experiences, and among the parameters affecting the experience are delay and synchronization. In addition, even if there is a big impact from the transport on delay and synchronization, also acquisition, compression, transmission, and reconstruction must be included when evaluating the impact since all of these stages will to some extent have impact on the accumulated end-to-end delay and audio and video skewness. This means that even if the data stream would be unencrypted it would not be possible to only use significant information from used packets, e.g. RTP [8] or TCP, for reliable measurements.

There are several recommendations for the accuracy of synchronization and delay, varying for different user scenarios and also between the recommending bodies. In the television context ITU-T recommended synchronization thresholds in J.100 [10], which are 20 ms for audio lead and 40ms for audio lag. This recommendation provides a fixed figure for all content types and is intended to ensure that synchronization errors remain imperceptible. For real-time two-way low bitrate video communication, ITU recommends the asynchrony to be less than 100 ms. Further ITU sets the preferred one-way end-to-end delay to be 100 ms and the upper limit to be 400 ms [13]. For the

Rec. body	User scenario	Rec. nr.	Skew [ms]	Delay [ms]	Ref
ITU-T	Television	J.100	<20 (audio lead) <40 (audio lag)	-	[10]
ITU-T	Two-way video com.	H-series Supp.1	<100	<100 (pref.) <400 (limit)	[13]
ETSI	End2End video-telephony	ETR 297	<20 (audio lead) <40 (audio lag)	<800 (round trip)	[4]
ITU-R	Broad-casting	BT.1359-1	<90 (audio lead) <185 (audio lag)	-	[9]
ITU-T	Video-phone	G.1010	< 80	< 100 (pref.) < 400 (limit)	[11]

Table 1: Table showing recommendations for skew and one-way delay in different user scenarios and by different recommendation bodies

same application ETSI [4] recommends less than 40 ms skew. For one-way broadcasting ITU recommends in [9] the skew to be less than 185 ms when video arrives first, and less than 90 ms when audio arrives first. This recommendation does not specify any preferred delay for the broadcasting scenario. For videophone ITU recommends a skew up to about 80 ms and end-to-end delay to be 150 ms with the upper limit 400 ms [11]. A summary of the presented recommendations is given in Table 1.

3 Related work

The impact of synchronization and delay on both QoE and QoS is well documented. Nevertheless, means or methods to measure this under real conditions, with a generic setup without limitations regarding e.g. platforms, operating systems (OS), devices and applications, is a topic that has not been

published to the same extent. The topic can be seen from a broader network perspective or from the perspective of the communication application. In [21] peer-to-peer media streaming is discussed and the impact of delay and synchronization on QoS is addressed. Also, in the growing field of supporting different multimedia applications over heterogeneous networks the topic of delay and synchronization is addressed, where also the impact from the devices are taken into account [27], [26]. However, even if the frameworks addressing these topics and enabling several parts from a network and QoS perspective, there are still several impediments remaining when an evaluation or analysis should be performed. This comes mainly from the proprietary nature of the majority of applications resulting in lack of access to, and understanding of, protocol and system parameters which make insight and analysis problematic [21]. Further, as a result of today's tremendous dissemination of applications and platforms these different capabilities have a significant impact on the end result.

Methods for measuring synchronization and delay from the application perspective can be divided into two categories, out-of-service and in-service measurements, where out-of-service is also denoted as black-box tests. The difference is whether the method requires access to the application to enable the test method, which is the case for in-service methods, or if the method can be applied from the outside without any interaction or impacts of the application. Different out-of-service methods have been presented. A user centric measurement is presented in [14] enabling video delay measurements using QR codes, which was also further developed in [15] where audio measurements were added and thereby enabling synchronization measurements. A capture-to-display latency and frame rate estimation was presented in [2] where barcodes were used in a JAVA application running on the sender device. These methods enable measurements but in a limited range of scenarios. Also, the methods may affect the application resulting in an altered behaviour. Using QR codes or barcodes can, besides giving a complex decoding process, be limited in supporting e.g. high compression, small displays, and decreased frame resolution. All these limitations will reduce the scope of possible test scenarios.

In the area of in-service methods the use of watermarking or feature extraction has been suggested in e.g. [19], [16]. These in-service methods are usually limited by lack of robustness and they require sending of meta-data including relative temporal alignment information. A scenario with packet loss or fluctuating network bandwidth, resulting in temporal discontinuities or spatial fidelity change in the decoded sequences, can result in complications

when it comes to extracting the features and to be able to preserve the correlation in a window of several seconds. Also, since in-service solutions require processing of the data inside the application this makes them unusable for evaluation of a not accessible application, e.g. a proprietary solution, since it requires the possibility to manipulate the application.

4 Proposed method

The main idea of the proposed method is to measure the audio and video delay and synchronization by marking the audio and video stream, before encoding, on the sender side with a signature or stamp in form of a frame number. This signature or stamp is designed not to be corrupted in the encoding/decoding process or the transport. After decoding of the received audio and video stream these stamps are detected and decoded and used together with time information to calculate audio/video delay and synchronization. The frame number is generated by a binary code of M bits resulting in 2^M original numbers, which is coded with Gray codes [20] to reduce the impact on the video processing as well as being robust to different resolutions and compression levels. Further, with a frame rate of F_r fps this gives a maximum delay or skewness that can be detected of $2^M/F_r$ seconds.

The audio frames are marked with a sum of windowed sinusoids of different frequencies from a set of M sinusoidal base functions $b_1(n), ..., b_M(n)$ with center frequencies $f_1, ..., f_M$. The length of a base function is N samples. The set is chosen for each frame to form a code number for the frame, in line with the Gray coded frame numbering used for the video sequence. Additionally, to detect the beginning of a new audio frame an additional base function b_s or audio frame synchronization signal is added in the beginning of each frame. The audio signal is at the receiver side filtered in a bank of matched filters, in a similar way as in [3], having impulse responses $b_m(N-n), m = 1, ..., M$, after which it can be decided which combination of base functions is present in a specific audio frame. In a corresponding manner the video frames are tagged by adding a pattern of black and white squares spatially distributed to each frame, where the pattern codes the frame number. In the proposed method a frame rate, F_r, of 15 fps is chosen to update frame number. This comes from a trade off between, on one side, as high refresh rate as possible of new measurements and, on the other side, sufficient support for different frame rates and camera capturing times for the video and reasonable length of the base functions for the audio.

This setup results in a robust measurement method both from a packet loss perspective and with respect to fluctuating bandwidth resulting in different compression levels and varying temporal or spatial resolution. It gives a robust solution with a possibility of new measurement every 1/15 second, and limited effect from fidelity change on the detection of the audio and video signatures and from limitations in the capturing device. The proposed method thereby enables the possibility to measure delay and synchronization in different kind of network and configurations. Also, any application, proprietary or not, can be tested.

4.1 The audio frame stamps

The base functions inserted in each audio frame will be sent through speech coders and decoders, and may be subject also to additional distortions. For the method to be robust the selection of base functions is essential. Due to the usual speech coder frequency range we have decided to keep the base functions in the frequency range of 700 to $2100Hz$. To get low correlation between the base functions we distribute them equidistant in this range, and choose long base functions. For a frame rate of F_r the number of samples in a frame will be F_s/F_r where F_s is the audio sampling rate. To have room for audio frame sync data in each frame and to have margins for consecutive frames, the length of the base functions were chosen as $N = F_s/(2 * F_r)$, i.e. half the frame length.

For the implementation in this paper the following values of the parameters F_s, F_r, N, and M were set. The audio channel sample rate is $F_s = 18000Hz$, and the frame rate is $F_r = 15$ fps. This gives the base function length $N = 600$ samples. To be able to uniquely number up to 256 frames the number of base functions is set to $M = 8$. The separation of the base functions in frequency domain can be seen in Figure 1 where the sinusoids have been windowed with rectangular window and Hanning window [17], respectively. In order to get a robust detection system the interference between the base functions needs to be low. The figure shows that as expected the spectrum for each base function is widened by the use of Hanning window, but the better suppression of the side lobes makes Hanning window a better choice in order to reduce interference. The timing within the audio frames is seen in Figure 2.

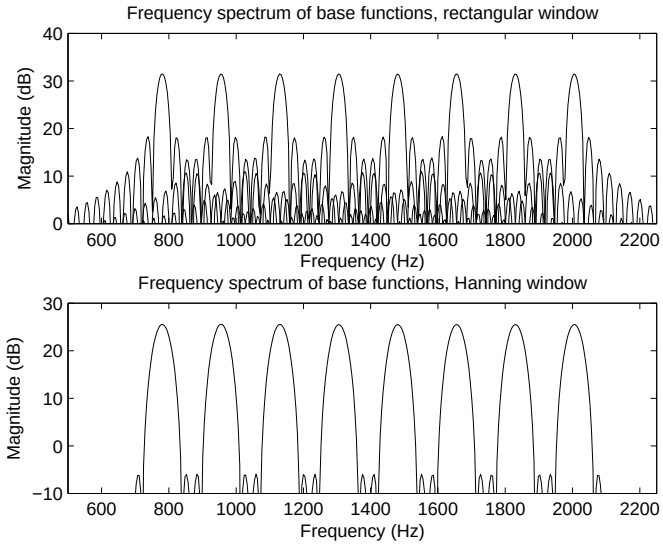


Figure 1: Frequency functions of the base functions b_m, windowed by rectangular window (top) and Hanning window (bottom).

4.2 The video frame stamps

The black and white squares representing the frame number is positioned in a video frame as shown in Figure 3. To minimize the impact on the video encoder with discontinuities between two consecutive frames the frame number is coded with Gray codes resulting in that only one square is changed between each frame. The video sequence is updated with the same frame rate, $F_r = 15$ fps, and with the same frame number as the audio sequence. This gives a robust detection system since the squares are easy to detect and the impact from quantization and resolution change is small.

4.3 The synchronization and delay detector

At the receiver side the audio and video is captured. For calculating synchronization the received audio signal $\hat{A}$ and the video signal $\hat{V}$ are used, and to enable delay measurements also the audio and video reference signals A and V are needed. A block diagram of the system at the receiver is shown in Figure

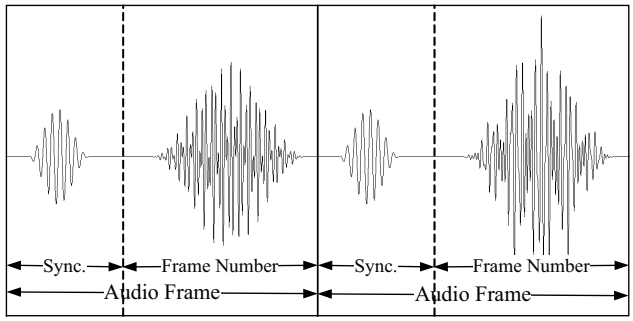


Figure 2: Illustration of the position of the sum of base functions in the audio frames.

4. There are two pre-processing blocks, Split, which is used to separate the audio and video signal, and Re-Sampling, which is applied to both audio and video signals to decimate the data in order to reduce the processing needed in subsequent steps. The block for detecting audio frame number in Figure 4 is shown in detail in Figure 5, and the block for detecting video frame number is shown in detail in Figure 6.

Here it can be seen that the audio sequence is filtered with matched filters for both the synchronization signal and the base functions. The results from the synchronization filter are used in the frame splitting for further base function detection. These data are then digitized and converted from a Gray code to the actual frame number together with the corresponding time index captured from the synchronization match. For the video detection in Figure 6 the video signal first is sent to the Region of interest (ROI) detection that is used to find where the binary squares are located and map the right square to the corresponding digit, resulting in separate vectors for each digit. These are individually normalized and digitized and then converted from a gray-code to the actual frame number together with the time index corresponding to the frame capture time. This set of data with frame number and time index for A, V, $\hat{A}$, and $\hat{V}$ is used to calculate the audio delay (A and $\hat{A}$), the video delay (V and $\hat{V}$) and the synchronization skewness ($\hat{A}$ and $\hat{V}$) with corresponding time index, using simple matching of frame number and difference calculation of corresponding time indices.

Figure 3: Code word entered into a video frame.

4.4 Reference signal

The reference signal is generated in MATLAB where the video stamps are added as an overlay of any video sequence, while the audio reference is generated as sequence of audio frames. The audio and video frames shown in Figures 2-3 are then merged together in perfect alignment regarding frame number and time index, which is also verified. Any video sequence can be used for the video reference, in this case Ice. The choice of video sequence could impact the complexity of the video process and shall be based on the type of application that is tested, i.e. emulating a realistic scenario. Also, the video stamp squares should cover only a small amount of the video frame resolution to minimize impact on complexity of the video process.

4.5 Implementation of the method

In the implementation of the method both MATLAB and Python environment has been used and the setup is presented in Figure 7, where the application to be tested is represented by the Sender and Receiver blocks. In the A/V Reference the audio and video signals are generated separately and merged in MATLAB, see Subsection 4.4, and since the same frame period and frame numbering is used for both signals this results in a fully synchronized reference stream. The A/V reference signal is played out to the Sender from a separate computer where the audio is electrically feed to the microphone input and the video is played out on a 60 fps display and captured by the Sender's camera.

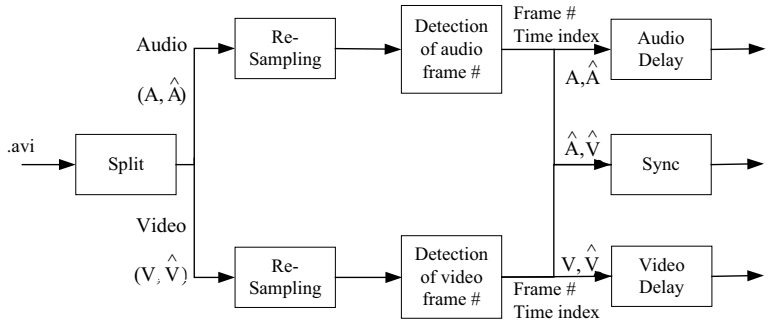


Figure 4: Block diagram of the system measuring the audio/video delay and skew.

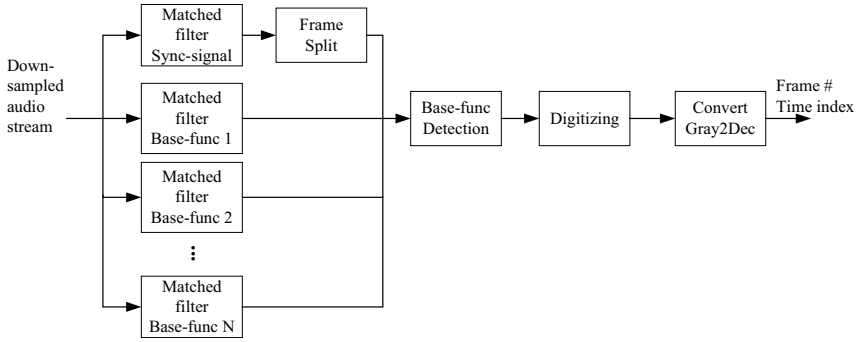


Figure 5: Block diagram of the detection of the audio frame number.

The data is sent over a network and rendered and played out by the Receiver. The A/V Capture Device is a 60 fps HD-DV camera with a microphone input. For video, the Capturing Device is set up so that both the received, $\hat{V}$, and the reference signal, V, are included side by side in the same captured frame. For capturing audio, the audio reference A is recorded on the left microphone channel while the received signal, $\hat{A}$, was recorded on the right channel. This recorded data was then fed into the Delay Sync Measure system. The Delay Sync Measure block in Figure 7 was implemented in Python. Since the squares representing the positions of the marking of V and $\hat{V}$ are spatially fixed, in this implementation the position is marked manually before the processing starts. The reference signal is without skew, which was verified by using the

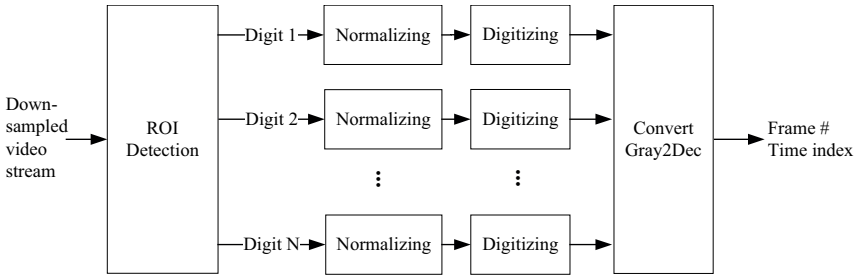


Figure 6: Block diagram of the detection of the video frame number.

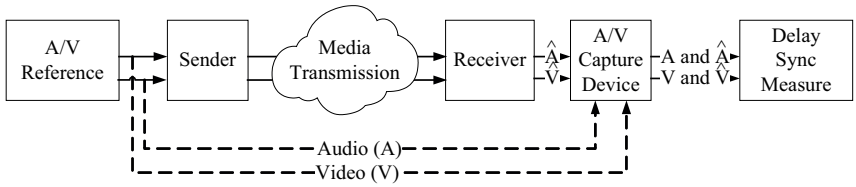


Figure 7: The setup of the delay and synchronization test.

reference both as input signal and degraded signal, so the decoded audio and video frame number together with referring time stamps can be used directly for calculating delay and skew. It should also be noted that skewness is only calculated for frames where both an audio and video number is detected, e.g. during a detected video freeze no skew result is generated. The limitation in the setup that should be taken into account is mainly video related: the refresh rate on the reference display, the shutter time of the camera at the sender side, the refresh rate of the receiver display, and finally the video shutter time of the A/V capture device. The display refresh rate used in this setup is 60 fps. These capturing limitations results in a maximum accuracy of 16.6 ms minus mis-alignment in the test setup chain, which make the impact of the audio setup limitations negligible considering audio sampling rate (F_s). Having $M = 8$ results in a maximum delay of 17 s which is then also the theoretical maximum skew detectable.

5 Test setup for proof of concept

In order to evaluate the proposed method the test setup in Figure 7 was used, and a VoIP based audio and video two-way communication application with encrypted UDP data was chosen for the test. The factors impacting the delay and synchronization are many, e.g. packet losses, network jitter, compressions levels, CPU constraints, different internal delays in different devices and the OS. The most important conditions from the methods point of view is delay and audio/video skew together with packet loss and compression from which all impact factors can be derived.

To enable the test and verify it during realistic conditions, the application is running over a real network. This is carried out both with a transparent network and with a network with emulated packet loss. In addition to the artifacts generated by the packet loss as such, packet loss will also cause the application to decrease the bandwidth usage and thereby increase the compression. This results in the method being verified during realistic conditions, which means that the frame code information is going through all the signal processing steps, is sent over a network and finally played out and captured with real devices.

5.1 Test scenarios

To evaluate the proposed method five different test scenarios were constructed using both a real transparent network with no packet loss and a real network with emulated packet loss.

1. No added delay.
2. An artificial delay of 16 frames, which is approximately 1.07 s, is introduced after 20 s for both audio and video.
3. An artificial delay. After 20 s, a delay of 16 frames was added to video, and after 30 s audio is delayed with 31 frames, approximately 2.07 s.
4. A uniformly distributed 5% packet loss is introduced, with no added delay. Resulting in increased compression.
5. A uniformly distributed 10% packet loss is introduced, with no added delay. Resulting in increased compression.

The first scenario will address the delay of the application with low impact from the network. It should be noted that this metric not only gives the QoS timing coming from the network impact usually taken into account, but also the processing time for capturing and encoding at the sender time and for the decoding and rendering at receiver side. The second scenario will test that the audio and video measurements are aligned and that the method is not introducing any synchronization skewness. The third scenario, with different audio and video delay applied at different moments in time, will test the ability to measure the skewness. The fourth and fifth scenarios, with 5% and 10% packet loss, respectively, will evaluate the robustness from the perspectives of packet loss as well as compression, since the increased packet loss rate will also result in an increased compressing level.

6 Results

All result comes from one application, but the method has been applied to and verified for several different applications. The method was applied several times for each of the different test scenarios. One representative session from each scenario is presented in Figures 8-12, where every calculated value is presented with a star in the graphs. It can be seen in the figures that the method is behaving in the way described and expected in Section 5.1. In Figure 8 the scenario with no added delay, transparent network, is presented and thereby showing the application's audio and video delay together with its synchronization skewness with low impact from the network. These results will represent an estimate of the impact from accumulated processing time on sender and receiver side. In Table 2 the corresponding average values are presented.

Measured	Avg. [ms]	Std [ms]
Audio Delay	119.7	0.65
Video Delay	176.7	11.1
Synchronization	57.0	11.1

Table 2: Audio and video delay and synchronization for transmission over transparent network, see Figure 8

It can be seen in Table 2 that there is an overall longer delay of the video than the audio, resulting in a positive synchronization skewness of $\approx$ 57ms

which according to the recommendations is an acceptable skewness. In Figure 9 it is shown that the audio and video measurements are aligned when the delay is applied, visualized with a stable synchronization skewness. In Figure 10 when the delay is applied at different points and with different amounts it can be seen that the method can also detect this, resulting in skewness. Further, it can be seen that the delay in Figure 9-10 is the accumulation of the delay from Figure 8 and the artificial delay as expected. The maximum accuracy of 16.6 ms can also be seen in the video delay figures by the quantized delay values. In Figures 11-12 packet losses are applied which is also observed by the decreased amount of correct detected frames. For video, packet loss results in lost frames which leads to freezes, this is indicated in the figures by a linearly increasing video delay. This is noticeable for the 10% packet loss case. For audio, the packet loss leads to a lower number of detected frames. It can also be seen that the audio delay variation continues to be small even in the 10% packet loss case which indicates that the variation is handled by the audio jitter buffer. From the figures it can also be seen that as long as an audio and video frame is successfully decoded with the same frame number a synchronization value can be calculated. This makes the method robust since only one frame, in this case representing 1/15s, needs to pass the whole system in order to be able to compile a new measurement, which is beneficial in demanding situations.

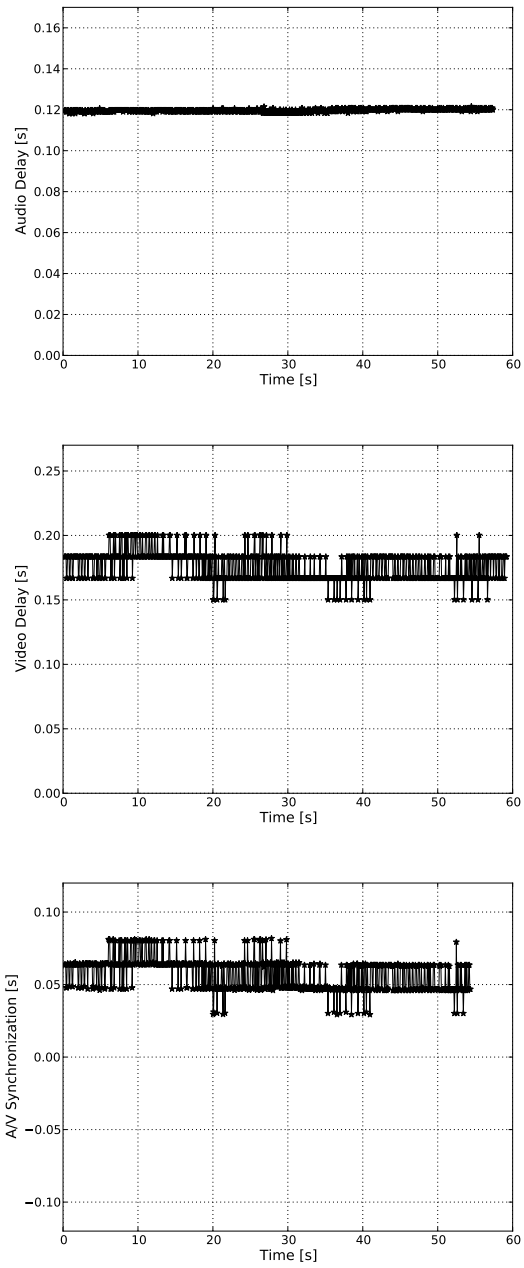


Figure 8: Transmission over transparent network. Top: audio delay. Center: video delay. Bottom: Synchronization skewness

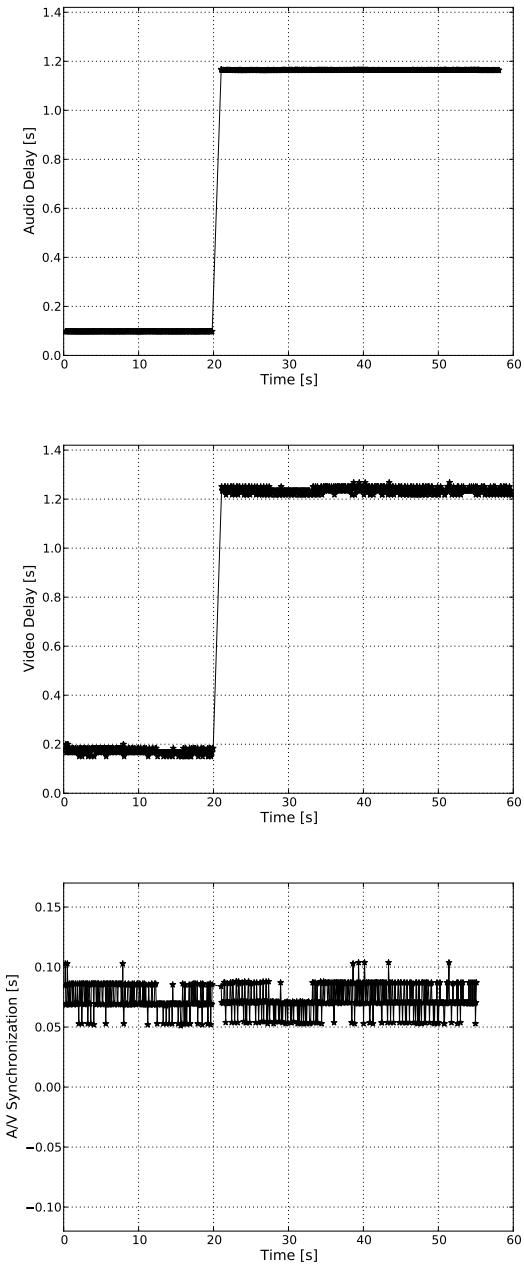


Figure 9: Transmission over transparent network with artificial delay. The audio and video delay is added simultaneously. Top: audio delay. Center: video delay. Bottom: Synchronization skewness.

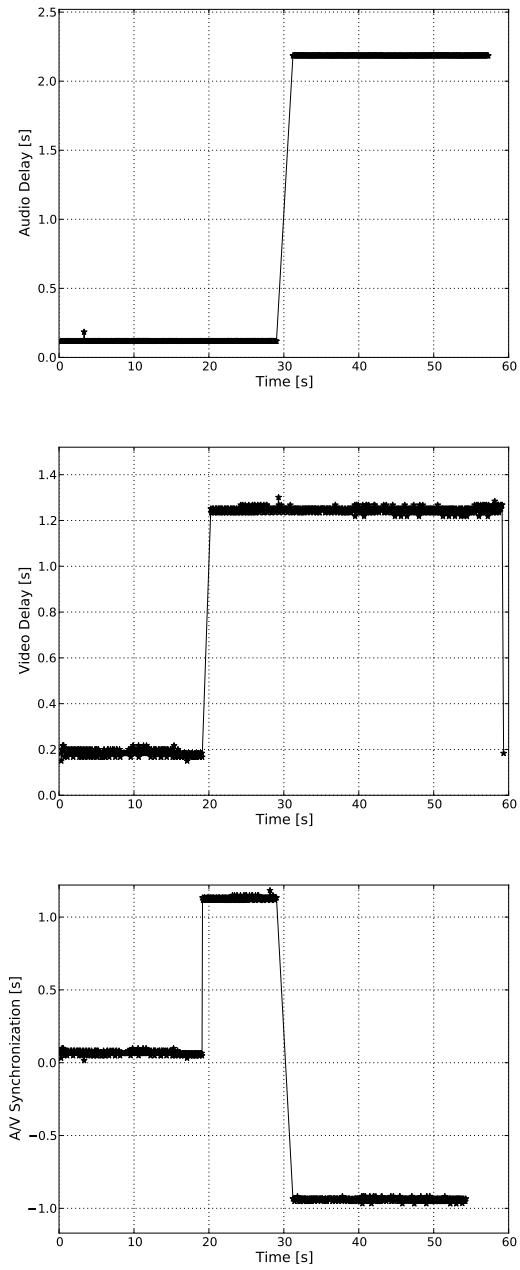


Figure 10: Transmission over transparent network with artificial delay where the delays are added separately, first audio and then video. Top: audio delay. Center: video delay. Bottom: Synchronization skewness.

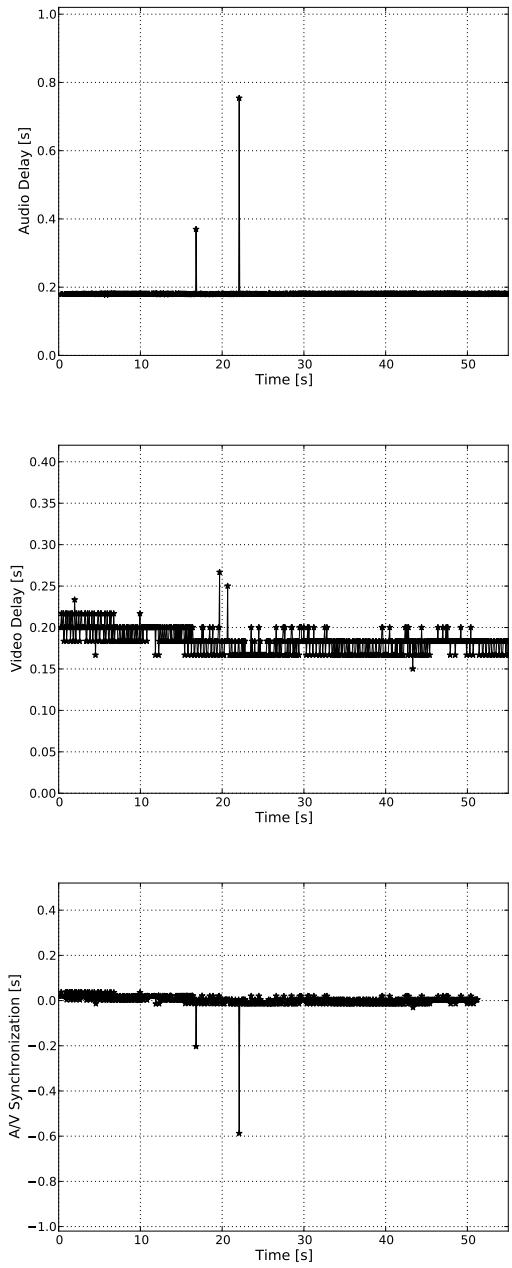


Figure 11: Transmission over network with 5% packet loss. Top: audio delay. Center: video delay. Bottom: Synchronization skewness

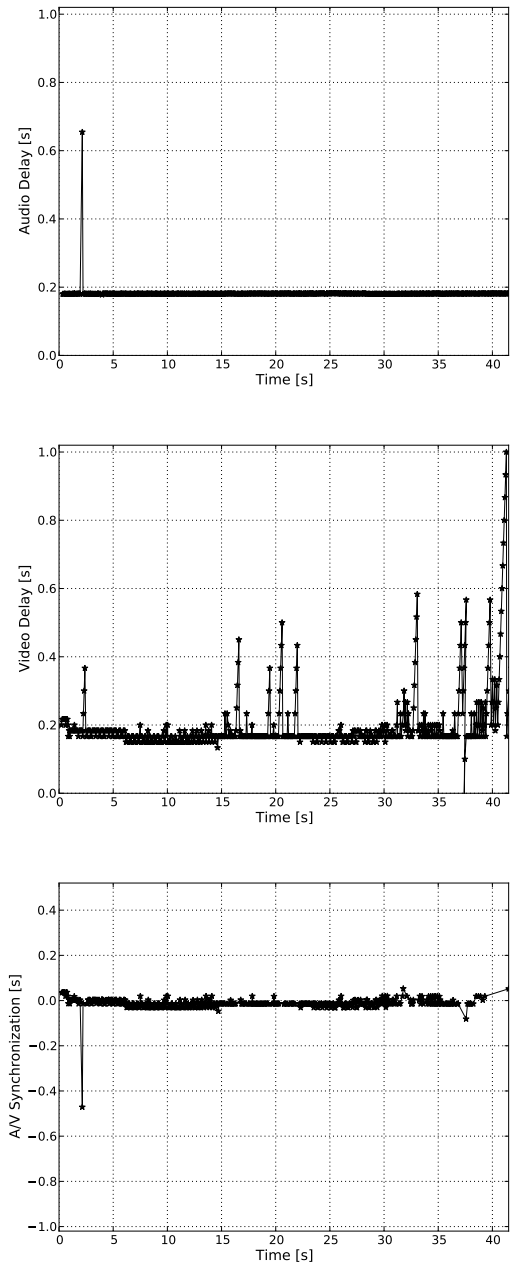


Figure 12: Transmission over network with 10% packet loss. Top: audio delay. Center: video delay. Bottom: Synchronization skewness.

7 Conclusions

In the area of quality of experience (QoE) for audio and video streaming and two-way audio and video communication the skewness in audio and video synchronization together with the audio and video delay have a major importance. A novel method for measuring both synchronization and delay, without need of access to application internal data, has been presented. The solution consists of an audio and video reference stream, where audio and video frames are marked with frame numbers, which are decoded on the receiver side to enable calculation of synchronization and delay. The audio stamps consists of a sum of sinusoidal base functions while the video stamp is a binary pattern embedded into the frame. The method has been verified with a real two-way communication application in a transparent network where the results show that both delays introduced by the application and the network as well as synchronization skewness can be detected. Additional artificial delays were added to verify that the method was able to detect and behave correctly in these controlled test scenarios. To verify robustness against packet losses both 5% and 10% packet loss levels were applied and it was shown that both audio and video delays were detected successfully. Further, the test system can be adapted to be used in several ways depending on whether the requirements come from a streaming service or a two-way communication application. When only the skew is of interest, no reference signal is needed for the calculation at the receiver side which makes it useful for streaming applications. The method presented opens up opportunities to build and design tests for many test scenarios, enabling evaluation of many kind of applications running on any platform, OS and device, e.g. for all applications developed for tablets and smartphones, under any conditions.

8 Future Work

As stated in the conclusions the method can be used in many kinds of scenarios, but it can also be adapted for in-services usage. The method can be further developed by increasing the precision depending on the user scenario being requested, e.g. increasing the frame rates of capturing and reference rendering. The method will also be used for evaluation of new and upcoming WebRTC solutions regarding their robustness against packet loss as well as fluctuating network conditions.

References

[1] Blakowski, G., Steinmetz, R.: A media synchronization survey: reference model, specification, and case studies. Selected Areas in Communications, IEEE **14**(1), 5–35 (1996)

[2] Boyaci, O., Forte, A., Baset, S., Schulzrinne, H.: vDelay: A tool to measure capture-to-display latency and frame rate. In: Multimedia, 2009. ISM '09. 11th IEEE International Symposium on, pp. 194–200 (2009)

[3] Claesson, I., Rossholm (formerly Nilsson), A.: GSM TDMA frame rate internal active noise cancellation. International Journal of Acoustics and Vibration **8**, 159–166 (2003)

[4] ETSI ETR 297: Human Factors (HF); Human Factors in Videotelephony (1996)

[5] ETSI TR 102 643: Human Factors (HF); Quality of Experience (QoE) requirements for real-time communication services (2010)

[6] Hollier, M.P., Rimell, A.N., Hands, D.S., Voelcker, R.M.: Multi-modal perception. BT Technology Journal **17**, 35–46 (1999)

[7] Huang, Z., Nahrstedt, K., Shu, L., Steinmetz, R.: Evolution of temporal multimedia synchronization principles: A historical viewpoint. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) **9**(34), 40–47 (2013)

[8] IETF RFC 3550: RTP: A Transport Protocol for Real-Time Applications (2003)

[9] ITU-R BT.1359-1: Relative Timing of Sound and Vision for Broadcasting (1998)

[10] ITU-T: Recommendation J.100: - Tolerances for transmission time differences between vision and sound components of a television signal (1990)

[11] ITU-T Series G: Recommendation G.1010: End-user multimedia QoS categories (2001)

[12] ITU-T Series G: Recommendation G.114: One-way transmission time (2003)

[13] ITU-T Series H: Audiovisual and multimedia systems, Supplement 1 (05/99) Application profile - Sign language and lip-reading real-time conversation using low bit-rate video communication (1999)

[14] Jansen, J., Bulterman, D.C.A.: User-centric video delay measurements. In: Proceeding of the 23rd ACM Workshop on Network and Operating Systems Support for Digital Audio and Video, pp. 37–42. ACM (2013)

[15] Kryczka, A., Arefin, A., Nahrstedt, K.: AvCloak: A tool for black box latency measurements in video conferencing applications. In: Multimedia (ISM), 2013 IEEE International Symposium on, pp. 271–278 (2013)

[16] Liu, Y., Sato, Y.: Recovering audio-to-video synchronization by audio-visual correlation analysis. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pp. 1–4 (2008)

[17] Proakis, J.G., Manolakis, D.G.: Digital signal processing : [principles, algorithms and applications], 4. ed. edn. Pearson Prentice Hall, Upper Saddle River, N.J. (2007)

[18] Raake, A., Schoenenberg, K., Skowronek, J., Egger, S.: Predicting speech quality based on interactivity and delay. In: In: Proc. INTERSPEECH 2013, Lyon, France, pp. 1549–1552 (2013)

[19] Radhakrishnan, R., Terry, K., Bauer, C.: Audio and video signatures for synchronization. In: Multimedia and Expo, 2008 IEEE International Conference on, pp. 1549–1552 (2008)

[20] Savage, C.: A survey of combinatorial gray codes. SIAM Review **39**(4), pp. 605–629 (1997)

[21] Shen, Z., Luo, J., Zimmermann, R., Vasilakos, A.: Peer-to-peer media streaming: Insights and new developments. Proceedings of the IEEE **99**(12), 2089–2109 (2011)

[22] Steinmetz, R.: Human perception of jitter and media synchronization. Selected Areas in Communications, IEEE **14**(1), 61–72 (1996)

[23] Winkler, S., Mohandas, P.: The evolution of video quality measurement: From psnr to hybrid metrics. Broadcasting, IEEE Transactions on **54**(3), 660 –668 (2008)

[24] Yamagishi, K., Hayashi, T.: Analysis of psychological factors for quality assessment of interactive multimodal service. In: Electronic Imaging, vol. 5666, pp. 130–138 (2005)

[25] You, J., Reiter, U., Hannuksela, M.M., Gabbouj, M., Perkis, A.: Perceptual-based quality assessment for audiovisual services: A survey. Signal Processing: Image Communication **25**(7), 482 – 501 (2010)

[26] Zhou, L., Chao, H.C., Vasilakos, A.: Joint forensics-scheduling strategy for delay-sensitive multimedia applications over heterogeneous networks. Selected Areas in Communications, IEEE Journal on **29**(7), 1358–1367 (2011)

[27] Zhou, L., Xiong, N., Shu, L., Vasilakos, A.V., Yeo, S.S.: Context-aware middleware for multimedia services in heterogeneous networks. Intelligent Systems, IEEE **25**(2), 40–47 (2010)

Part D

On Audio and Video Quality Assessment

Part D consists of:

Part D.1: A New Low Complex Reference Free Video Quality Predictor
Part D.2: Analysis of Impact from Temporal and Spatial Artifacts on Perceptual Video Quality
Part D.3: Comparison of Machine Learning Methods for Quality Estimation of Videos with Diversity in Temporal, Spatial, and Quantization Domains

Part D.1

A New Video Quality Predictor Based on Decoder Parameter extraction

Part D.1 is published as:

Andreas Rossholm, and Benny Lövström *A New Video Quality Predictor Based on Decoder Parameter extraction.*, at SIGMAP, February 2008.

A New Video Quality Predictor Based on Decoder Parameter extraction

Andreas Rossholm, and Benny Lövström

Abstract

In the mobile communication area there is a demand for reference free perceptual quality measurements in video applications. In addition low complexity measurements are required. This paper proposes a method for prediction of a number of well known quality metrics, where the inputs to the predictors are readily available parameters at the decoder side of the communication channel. After an investigation of the dependencies between these parameters and between each parameter and the quality metrics, a set of parameters is chosen for the predictor. This predictor shows good results, especially for the PSNR and the PEVQ metrics.

1 Introduction

There is a growing demand for objective quality measurement techniques estimating perceived video quality in mobile devices. This is of interest to, among others, the mobile phone industry, mobile network operators and software developers. The quality of a video encoder and decoder can be measured with different metrics. Frequently metrics which require a reference together with the processed image or video in order to evaluate the perceived quality are used [1]. Two of the most commonly used metrics are the objective metrics peak signal-to-noise ratio (PSNR) and the mean-squared-error (MSE) which can be calculated for each decoded frame and then averaged for the complete sequence.

Since both MSE and PSNR are based on a pixel-by-pixel comparison the metrics have some issues regarding the relation to the perceptual quality. This has resulted in the development of several new metrics as SSIM [2], a video

adapted version of SSIM denoted VSSIM [3], NTIAs VQM [4], and Opticoms PEVQ [5]. All these metrics use the original frame as reference, or some kind of reduced reference information, to calculate a relation between this and the decoded frame.

In many situations where perceptual quality is of interest, e.g. streaming video, video telephony, MBMS, and DVB-H, the original frames are not available. Thus there is a need for reference free quality metrics, which can be implemented entirely at the decoder side of a transmission line. There are a number of such metrics that have been developed but they often focus on one parameter such as blur [7], blockiness [8], or motion [9], and they require some processing of the received frame and are thereby often less useable in real application.

In this paper a solution is proposed to estimate the quality without having access to the original frames or reduced reference and without the requirement of processing of the received frame. The estimation is based on predicting the above mentioned full or reduced reference quality metrics.

2 The proposed idea

When a video sequence is encoded to fulfil the required properties such as bit rate, frame rate and resolution, the encoder sets and adjusts a number of parameters. Some of these are set for the whole sequence while some are adjusted for each frame or within frames. The coding results in a bit stream consisting of motion vector parameters, coded residual coefficients and header information, e.g. frame rate and quantization parameters (QP) value. From the bit stream it is also possible to calculate the number of intra blocks, number of inter blocks, number of skipped blocks, etc. The idea proposed in this paper is to predict the video quality using these parameters. The predictor is built by setting up a model and adapt its coefficients using a number of training sequences. The parameters used are available at the decoder and therefore the quality predictor is reference free.

Throughout this paper the video sequences are coded using the H.264 standard, since this is one of the most used video encoder for mobile equipment. The parameters chosen for evaluation of contribution to the predictor are

1. Average QP value (Avg QP)
2. Bitrate /Frame rate (Bits/Frame)

3. Number of intra blocks (Intra [%])
4. Number of inter blocks (Inter [%])
5. Number of skipped blocks (Skip [%])
6. Frame rate
7. Number of inter blocks of size 16x16 (P16x16[%])
8. Number of inter blocks of size 8x8, 16x8, and 8x16 (P8x8 [%])
9. Number of inter blocks of size 4x4, 8x4, and 4x8 (P4x4 [%])
10. Average motion vector length (Avg MV [%])

Also, other parameters could be extracted and evaluated but these where chosen based on their expected potential contribution to the perceptual quality.

3 The Metrics Predicted

The proposed model will in this paper be evaluated in predicting the following quality metrics; PSNR, SSIM, VSSIM, NTIA VQM, and PEVQ.
PSNR, the peak signal-to-noise ratio, is defined as

$$PSNR(n) = 10 \cdot \log \frac{MAX_I^2}{MSE(n)} \tag{1}$$

where MAX_I is the maximum value a pixel can take (e.g. 255 for 8-bit images) and the MSE is the average of the squared differences between the luminance values of corresponding pixels in two frames. MSE is defined as

$$MSE = \frac{1}{UV} \sum_{u=1}^{U} \sum_{v=1}^{V} [I_R(u,v) - I_D(u,v)]^2 \tag{2}$$

where $I_R(u,v)$ denotes the intensity value at pixel location (u,v) in the reference video frame, $I_D(u,v)$ denotes the intensity value at pixel location (u,v) in the distorted video frame, U is the number of rows in a video frame, and V is the number of columns in a video frame. To get a measure for a video

sequence a simple averaging over a video sequence of length N frames is made as.

$$PSNR = \frac{1}{N}\sum_{n=1}^{N} PSNR(n) \tag{3}$$

SSIM, the Structural SIMilarity index, considers image degradations as perceived changes in the variation of structural information by combining measures of the distortion in luminance, contrast and structure between two frames, [2], as

$$SSIM(n) = \frac{[2\mu_{I_R}(n)\mu_{I_D}(n)+C_1][2\sigma_{I_R I_D}(n)+C_2]}{[\mu^2_{I_R}(n)+\mu^2_{I_D}(n)+C_1][\sigma^2_{I_R}(n)+\sigma^2_{I_D}(n)+C_2]} \tag{4}$$

where $\mu_{I_R}(n), \mu_{I_D}(n)$ and $\sigma_{I_R}(n), \sigma_{I_D}(n)$ denote the mean intensity and contrast of the n-*th* reference video frame I_R and distorted video frame I_D, respectively. The constants C_1 and C_2 are used to avoid instabilities in the structural similarity comparison that may occur for certain mean intensity and contrast combinations.

Similar as with PSNR, the SSIM value for an entire video sequence of length N may be calculated as

$$SSIM = \frac{1}{N}\sum_{n=1}^{N} SSIM(n) \tag{5}$$

VSSIM, the Video Structural SIMilarity index, is an adaption of the SSIM metric to quality evaluation for video. VSSIM was developed using the VQEG (Video Quality Experts Group) Phase I test data set for FR-TV video quality assessment [6] and calculated as

$$Q_i = \frac{\sum_{j=1}^{R_S} w_{ij} SSIM_{ij}}{\sum_{j=1}^{R_S} w_{ij}} \tag{6}$$

where Q_i denotes the quality index measure of the i-*th* frame in the video sequence. The weighting value w_{ij} is given to the j-*th* sampling window in the i-*th* frame based on the observation that dark regions usually do not attract fixations and should therefore be assigned smaller weighting values. R_S is

the number of sampling windows per video frame that has been used. The VSSIM value for the entire video sequence of length N is then calculated as

$$VSSIM = \frac{\sum_{i=1}^{N} W_i Q_i}{\sum_{i=1}^{N} W_i} \tag{7}$$

where W_i is the weighting value assigned to the i-*th* frame based on global motion and w_{ij}. Since the metric was developed using the VQEG Phase I test data it consists of larger frame sizes (SD-resolutions, 525-line and 625-line) than the QCIF used in this paper, therefore a modified VSSIM has also been used in the proposed solution to adapt it to smaller resolution. This is accomplished by scaling the weighting coefficient K_M, used to calculate W_i, and its connected thresholds with a factor of 8, from 16 to 2 [3].

NTIA VQM, the National Telecommunications and Information Administrations general purpose Video Quality Model general model, is a reduced reference method containing linear combination of seven objective parameters for measuring the perceptual effects of a wide range of impairments such as blurring, block distortion, jerky/unnatural motion, noise (in both the luminance and chrominance channels), and error blocks [4]. The perceptual impairment is calculated using comparison functions that have been developed to model visual masking of spatial and temporal impairments. Some features use a comparison function that performs a simple Euclidean distance between two original and two processed feature streams but most features use either the ratio comparison function or the log comparison function. The VQM general model was included in the Video Quality Experts Group (VQEG) Phase II Full Reference Television (FR-TV) tests [6].

PEVQ, the Perceptual Evaluation of Video Quality from Opticom, calculates measures from the differences in the luminance and chrominance domains between corresponding frames. Also motion information is used in forming the final measure [5]. PEVQ has been developed for low bit rates and resolutions as CIF (352×288) and QCIF (176×144). PEVQ is a proposed candidate for standardization of a FR video model within VQEG which is in the process of starting verification tests for future standardization.

4 The mathematical model

The problem can be presented as an observation matrix, $X = [\mathbf{x_1 x_2} \cdots \mathbf{x_N}]$, where $\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_N}$ are a number of feature vectors that has been generated

with different video content and codec setups. Each feature vector $\mathbf{x_n}$ consists of extracted codec parameters denoted $x_1, x_2, \ldots x_K$. The corresponding quality measures for the different video content, PSNR, PEVQ, SSIM, VSSIM, and NTIM then correspond to the desired $Y = [y_1 y_2 \cdots y_N]$. X and Y can be viewed as training data for a classification, mapping or regression problem. It is desired to find a function $Z = f(\mathbf{x})$ that maps the given values in $\mathbf{x}$ to a specific value Z, e.g. an estimation of PSNR.
There are several different models solving the problem, that are more or less computational complex. Because a low complex solution is required in order to have the possibility for an implementation in a mobile device, multi-linear regression is selected.
The multi-linear model is formulated as:

$$Y = \beta \mathbf{x} + \epsilon \tag{8}$$

where ϵ represents the unpredicted variation. The multi-linear regression estimates the values for β denoted $\widehat{\beta}$ that can be used to predict Z as

$$\widehat{Z} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_K x_K \tag{9}$$

4.1 Predicted Metric Evaluation

To be able to evaluate the accuracy of the predicted metric *Pearson linear correlation coefficient* is used. It is defined as follows:

$$r_P = \frac{\sum(\widehat{Z}_i - \widehat{Z}_{mean})(Z_i - Z_{mean})}{\sqrt{\sum(\widehat{Z}_i - \widehat{Z}_{mean})^2}\sqrt{\sum(Z_i - Z_{mean})^2}} \tag{10}$$

where $\widehat{Z}_{mean}$ and Z_{mean} are the mean value of estimated and true data set respectively, and $\widehat{Z}_i$ and Z_i are the estimated and true data values for each sequence. This assumes a linear relation between the data sets.

5 Video Source Sequences

To generate training and verification data different sequences with different characteristic (amount of motion, color, heads, animations) were used. The source sequences had QCIF (176×144) resolution and were generated with different frame rates, 30, 15, 10, and 7.5 frames per second (fps), and bitrates,

approximately: 30, 40, 50, 100, 150, and 200 kilobits per second (kbps). The video sequences were approximately 3 seconds long (90, 45, 30, and 23 frames) and they were encoded with the H.264/MPEG-4 AVC reference software, version 12.2 generated by JVT [10] using the baseline profile.

The sequences for training were: Foreman, Cart, Mobile, Shine, Fish, Soccer goal, and Car Phone resulting in 168 sequences for training. For verification five different parts from a cropped version of the *3G*-sequence was used, where the five parts have different characteristics. The cropping was made to QCIF without the original letter box aspect ratio. Varying the bitrate and the frame rate in the same way as for the training data results in 120 verification sequences.

6 Results

In the first step the linear regression described above is applied to the 168 training sequences for each of the parameters separately, giving a measure of the correlation between this parameter and the quality metric. The outcome is shown in Fig. 1, where it can be seen that the parameters have considerably varying correlation, but also that it differs between the metrics.

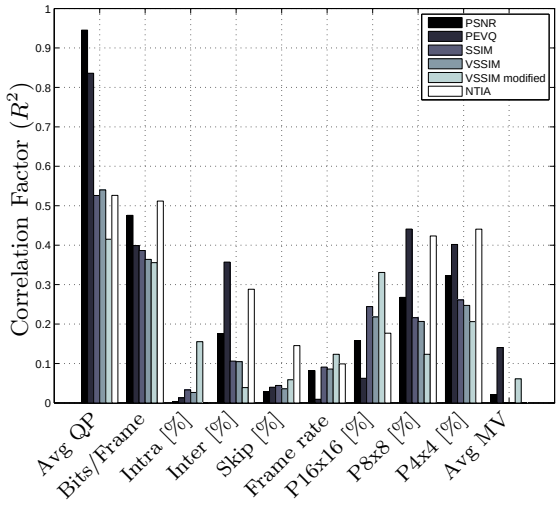


Figure 1: The correlation factor R^2 between each of the parameters and the metrics used

	Avg MV	P4x4 [%]	Frame rate	P16x16 [%]	Avg QP	Intra [%]	Skip [%]	Bits/ Frame	P8x8 [%]
Avg MV	1.000	-0.041	0.247	0.451	0.036	0.098	0.505	0.193	0.454
P4x4 [%]	-0.041	1.000	-0.180	0.545	-0.218	0.556	0.234	-0.577	-0.475
Frame rate	0.247	-0.180	1.000	0.032	0.124	0.143	0.079	0.206	0.374
P16x16 [%]	0.451	0.545	0.032	1.000	0.140	0.710	0.784	0.116	0.189
Avg QP	0.036	-0.218	0.124	0.140	1.000	0.160	0.429	0.652	0.539
Intra [%]	0.098	0.556	0.143	0.710	0.160	1.000	0.697	-0.162	0.294
Skip [%]	0.505	0.234	0.079	0.784	0.429	0.697	1.000	0.323	0.597
Bits/Frame	0.193	-0.577	0.206	0.116	0.652	-0.162	0.323	1.000	0.473
P8x8 [%]	0.454	-0.475	0.374	0.189	0.539	0.294	0.597	0.473	1.000

Table 1: The correlation matrix of the evaluated parameters.

Metrics	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	Scale
PSNR	66.89	-0.92	-0.07	-0.03	-0.01	-0.09	-0.07	0.01	$1.0 \exp -0$
SSIM	109.67	-0.46	0.16	-0.03	-0.38	0.26	0.49	-0.03	$1.0 \exp -2$
VSSIM	112.76	-0.52	0.10	-0.01	-0.33	0.19	0.41	-0.01	$1.0 \exp -2$
VSSIM modified	111.74	-0.48	0.00	0.00	-0.33	0.17	0.32	0.37	$1.0 \exp -2$
NTIA	66.13	-0.66	1.23	0.39	-0.82	1.34	-1.61	0.41	$1.0 \exp -2$
PEVQ	55.94	-0.92	0.14	-0.21	-0.07	0.23	0.26	-0.26	$1.0 \exp -1$

Table 2: The values of β_i in Eq. (9) for the different metrics, resulting from the regression

In the second step an evaluation of the different parameters are performed. In this the correlation between the parameters is calculated. Before the correlation is calculated "Inter [%]" is removed since this parameter is a summation of "P16x16 [%]", "P8x8 [%]", and "P4x4 [%]" and therefore redundant. The result from the correlation are shown in Tab. 1. It can be seen that "Intra [%]", "Skip [%]", "Frame rate", and "Avg MV" have the lowest correlations in Fig. 1. If these are analyzed it can also be seen that both "Intra [%]" and "Skip [%]" have the highest correlation with "P16x16 [%]" while neither "Frame rate" nor "Avg MV" correlations are that high. This makes it possible to reduce the parameter set further by excluding "Intra [%]" and "Skip [%]".
Performing the regression with the reduced parameter set using the training sequences gives a prediction function $\hat{Z}$ for each metric. The resulting coefficients in this function $\hat{Z}$ (see Eq. (9)) are shown in Tab. 2. The mapping of the $\widehat{\beta}_k$ in Tab. 2 to the actual parameters are shown in Tab. 3. These

$\widehat{\beta}_k$	Parameter
$\widehat{\beta}_0$	Constant
$\widehat{\beta}_1$	Avg QP
$\widehat{\beta}_2$	Bits/Frame
$\widehat{\beta}_3$	Frame rate
$\widehat{\beta}_4$	P16x16 [%]
$\widehat{\beta}_5$	P8x8 [%]
$\widehat{\beta}_6$	P4x4 [%]
$\widehat{\beta}_7$	Avg MV

Table 3: Mapping of the $\widehat{\beta}_k$ in Tab. **??** to the parameters used in the regression.

prediction functions, $\hat{Z}$, are applied to the verification sequences to predict the quality metric for these. Further, the quality metrics are calculated according to their definitions, and the *Pearson correlation coefficient*, r_P, from Eq. (10) is calculated and shown i Tab. 4. In the Fig. 2 – 5 the true metrics are plotted versus the predicted metrics. Note that the scale differs between the figures since the metrics have different range.

It can be seen from the table and the figures that the best prediction is obtained for the PSNR metric. This is expected since the JM encoder

Metric	r_P
PSNR	0.99
SSIM	0.62
VSSIM	0.61
VSSIM modified	0.71
NTIA	0.74
PEVQ	0.95

Table 4: The Pearson correlation coefficient, r_P, for the prediction of the different quality metrics

uses rate distortion optimization where the distortion measure is correlating with the PSNR. Also the PEVQ metric is well predicted, with a correlation coefficient of 0.95. This gives the possibility to implement the proposed no-reference metric in environments where full or reduced reference metrics are not possible to implement. The metric is also of low complexity, since the prediction is a simple calculation of the function in Eq. (9).

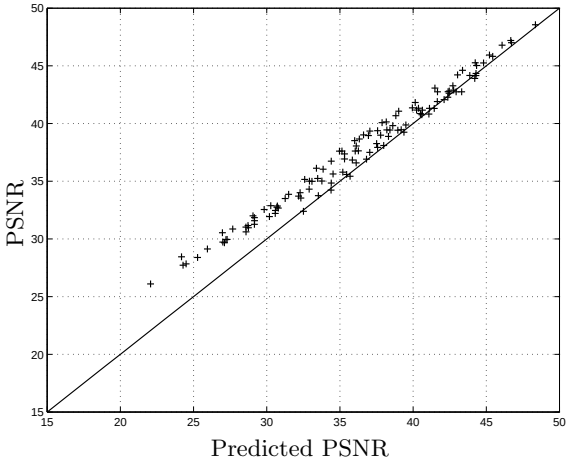


Figure 2: PSNR in dB vs. predicted PSNR for each verification sequence, $r_P = 0.99$

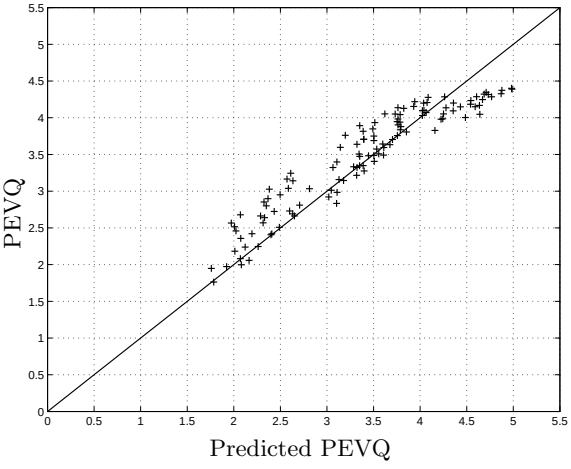


Figure 3: PEVQ vs. predicted PEVQ for each verification sequence, $r_P = 0.95$

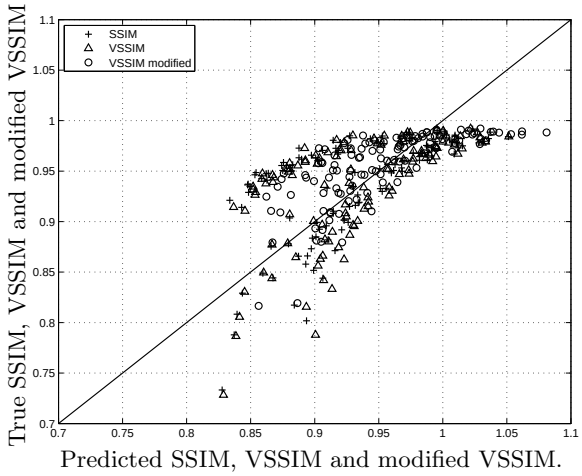


Figure 4: SSIM, VSSIM and modified VSSIM vs. the predicted values for each verification sequence, $r_P = 0.62,\ 0.61$ and 0.71

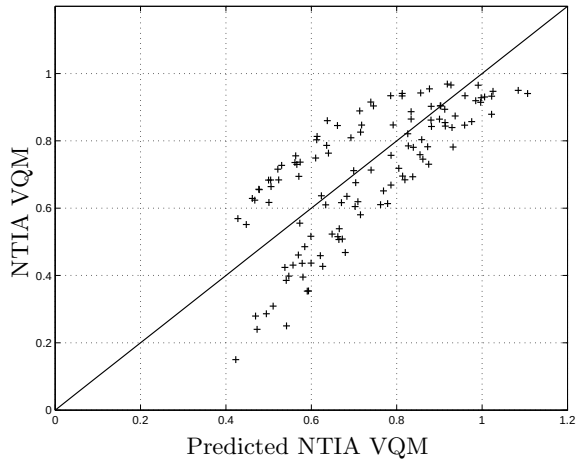


Figure 5: NTIA VQM vs. predicted NTIA VQM for each verification sequence, $r_P = 0.74$

7 Conclusion

A low complex, reference free method to predict perceptual quality metrics of coded video sequences has been suggested. For the PSNR and PEVQ metrics a very good precision is achieved, while for the other metrics the correlation is weaker. The result for PSNR is expected since rate distortion optimization is used in the encoder, while the result for PEVQ was not obvious beforehand and shows great potential. The precision of the prediction for PSNR may be considered to be of limited practical use since the correlation of PSNR to subjective perceptual quality is known to be low in many cases. On the other hand, the good precision for PEVQ prediction is promising since PEVQ is developed to measure the perceptual quality for low resolution and low bitrates, and is also proposed for standardization. The main result of this paper is the ability of the proposed method to predict quality metrics, and the final value of using this prediction depends on the value of the chosen quality metric.

In constructing the predictor an investigation has been performed to choose the most promising parameters to base the prediction on. This has been performed by evaluating the correlation both between each extracted parameter and the actual quality metric and between each parameter. The outcome from this has made it possible to restrict the number of parameters to seven and still achieve promising result. The parameters finally chosen are: Avg QP, Bits/Frame, Frame rate, P16x16 [%], P8x8 [%], P4x4 [%], and Avg MV.

To get a more general predictor where also other encoders are included the proposed model can be used. It will be obtained by increasing the training and verification set with sequences encoded using additional codecs. Then a new evaluation of which parameters to choose would be needed, resulting in a new set of $\widehat{\beta}$ values.

References

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

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity" IEEE Transactios on Image Processing, vol. 13, no. 4, pp.600-612, Apr. 2004.

[3] Z. Wang, L. Lu and A.C. Bovik, "Video quality assessment based on structural distortion measurement", Signal Processing: Image Communication, Special issue on Objective video quality metrics, vol. 19, no. 2, February 2004.

[4] M. H. Pinson, S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality". IEEE Transactions on Broadcasting, vol. 50, no. 3, Sept. 2004.

[5] "PEVQ Advanced Perceptual Evaluation of Video Quality (PEVQ Whitepaper)", Opticom, http://www.pevq.org, (verified 2008-01-18).

[6] "FR-TV Phase II Final Report", 2003-08-25, VQEG: The Video Quality Experts Group, http://www.vqeg.org/projects/frtv_phaseII/, (verified 2008-01-18).

[7] P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahimi, "A No-Reference Perceptual Blur Metric", IEEE International Conference on Image Processing 2002, volume 3, pp. III-57 - III-60, Rochester, USA, Sept. 2002.

[8] W. Zhou, A. C. Bovik, B. L. Evans. "Blind measurement of blocking artifacts in images", IEEE International Conference on Image Processing 2000, volume 3, pp. 981 - 984, Vancouver, Canada, Sept. 2000.

[9] M. Ries, O. Nemethova, M. Rupp, " Motion Based Reference-Free Quality Estimation for H.264/AVC Video Streaming", 2nd International Symposium on Wireless Pervasive Computing (ISWPC '07), pp. 355 - 359, San Juan, Puerto Rico, USA, February 2007.

[10] "H.264/MPEG-4 AVC REFERENCE SOFTWARE", Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6), 24th Meeting: Geneva, CH, 29 June - 5 July, 2007.

Part D.2

Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

Part D.2 is published as:

A. Rossholm, M. Shahid, and B. Lövström, *Analysis of Impact from Temporal and Spatial Artifacts on Perceptual Video Quality.*, at IEEE Network Operations and Management Symposium (NOMS), 2014, May 2014.

Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

Andreas Rossholm, Muhammad Shahid, and Benny Lövström

Abstract

The growing consumer interest in video communication has increased the users' awareness in the visual quality of the delivered media. This in turn increases, at the service provider end, the need for intelligent methodologies of optimal techniques for adapting to varying network conditions. Recent studies show that constraints on the bandwidth of transmission media should not always be translated to an increase in compression ratio to lower the bitrate of the video. Instead, a suitable option for adaptive streaming is to scale down the video temporally or spatially before encoding to maintain a desirable level of perceptual quality, while the viewing resolution is constant. Most of the existing studies to examine these scenarios are either limited to low resolution videos or lack in provisioning of subjective assessment of quality. We present here the results of our campaign of subjective quality assessment experiments done on a range of spatial and temporal resolutions, up to VGA and 30 frames per second respectively, under a number of bitrate conditions. The analysis shows, among other things, that keeping the spatial resolution is perceptually preferred among the three parameters that have impact on the video quality, even in the case with high temporal activity.

1 Introduction

As video communications constantly continues to grow both regarding its share of all data traffic and the amount of data in absolute terms, the consumers demand on perceived quality also increases. Also, new cellular wireless

technology evolves and an increasing share of all data communication will be wireless. This results in many new scenarios with different services and requirements where the provider want to optimize the perceived quality or quality of experience (QoE). One new challenge with new mobile networks like 3G and 4G is that even if high peak link rates are possible the cellular wireless networks experience rapid link rate variation and occasional long delays in one or both direction. This requires either long receiver buffers, resulting in long end-to-end delay, or fast adaptation, resulting in need for the possibility to change used band width [1]. In this context the need of optimizing the delivered quality of experience by a service provider is raised. To this end, one significant issue to be resolved is finding the best trade-off among spatial resolution, temporal resolution, and quantization level, giving the optimal value of QoE in a given scenario. In practice, this includes applications such as adaptive streaming [2], [3], as well as different real time video communication services where maintaining the desired level of perceived quality is required in fluctuating network conditions. Also, there is a growing demand for objective quality measurement or monitoring techniques estimating perceived video quality in these scenarios, especially to be able to compare different spatial and temporal resolutions. An overview of various types of contemporary objective Video Quality Assessment (VQA) is presented in [4].

The quest of discovering the optimal trade-off has been the subject of video scalability for assuring stipulated level of visual quality. For service providers, it is useful to ascertain the best QoE of a video at a given bandwidth capacity. In order to optimally address any fluctuations in the transmission network, it becomes pertinent to determine the parameter that can be scaled up or down with minimal deviation in the level of delivered visual quality. To serve this matter, a number of studies have been made that focus on examining the impacts of changes in the aforementioned three parameters of a video. Subjective quality assessment of low resolution, QCIF ($176x144$) and CIF ($352x288$), videos encoded using H.264/AVC has been reported in [5] for 150 test scenarios. Under low bitrate conditions, it was concluded that small frame size is mostly preferred. For CIF resolution or high temporal (30 fps) resolution at low bitrates, it was found that it was most efficient to reduce quantization except for video sequences containing very low spatial activity. It was also pointed out that a minimum threshold value of 0.1 Bits Per Pixel (BPP) is required to achieve good or excellent perceptual quality. Subjective experiments conducted using low resolution videos, CIF, in [6] show that frame rate can be compromised to maintain the perceptual quality by keeping the compression ratio at low value. Similar results can be observed in the

study reported in reference [7]. Impact of encoding strategy on the quality of MPEG-2 encoded videos, QCIF and CIF, while transmitted over lossy network has been investigated in [8]. It has there been observed that videos with high spatial activity are perceptually preferred with higher spatial resolution, and videos with higher temporal activity are preferred in full frame-rates. The validity of these results needs to verified in the case of videos encoded by H.264/AVC.

Considering the case of high resolution video conferencing applications, video scalability has been tested for high definition videos ($1920x1080$) in [9]. It was observed that the quality level can be maintained by decreasing frame rate and frame resolution to cater the constraints of the transmission bandwidth. Hence, high compression rates can be avoided. Moreover, as the bandwidth begins to grow, it is perceptually preferred to increase the frame rate up to a certain higher level first and the frame resolution can be increased afterwards. Unfortunately, these conclusions have been drawn only from the results of objective metrics, with no subjective assessment to support the results. A detailed discussion and a review of the studies performed on the video scalability for quality can be found in [10] and the references therein.

By examining the existing drives to investigate the impacts of three basic parameters of video encoding, the requirement of a comprehensive study on a wider range of videos, in a highly interesting bandwidth range, supported by subjective assessment of quality becomes evident. Therefore, we present here the details of an extensive campaign of subjective quality assessment experiments of videos encoded using combinations of multiple levels of the bitrate, frame rate and resolution. This enables examinations of e.g. the perceptual trade off between spatial and temporal resolution at a certain bitrate. The rest of this paper is organized as follows. In Section 2 the video sequences used in the test are described, encoding configuration, as well as the subjective assessment setup. Also the pre-processing of the sequences before the assessment is described. In Section 3 the findings from the subjective tests are given, and finally in Section 4 conclusions are drawn.

2 Test Stimuli and Subjective Video Quality Assessment

To perform a comprehensive subjective quality assessment that can be used to infer useful conclusions, it is imperative to select the SouRCe sequences

(SRCs) carefully. Such SRCs should possess a variety of spatio-temporal characteristics to be representative of most commonly used videos. To this end, we followed the ITU recommendation P.910 [11] for the selection of SRCs based on spatial perceptual information (SI) and temporal perceptual information (TI). The SI and TI values are calculated in the luminance plane of a video. The five SRCs used in this study are Children, City, Elisa, Ice, and Soccer, all of 10 s duration. The starting frame of each of the sequences is shown in Fig. 1, and Table 2 gives a short description of the content and lists the original frame rate of the sequences as well as their SI and TI characteristics.


Children


City


Elisa


Ice


Soccer

Figure 1: The first frame of each source sequence used for generation of test sequences

Sequence	Frame rate [fps]	Description
Children	30	Two children sitting on the floor, slowly moving, low SI and low TI
City	25	Panning view over a city from an airplane, high SI and medium TI
Elisa	30	Head and shoulder of a talking woman, medium SI and low TI
Ice	25	Several persons skating on white ice, low SI and high TI
Soccer	25	Close up view of soccer game, panning, low SI and high TI

Table 1: The original frame rate and a brief description of the SRCs used in the experiment

2.1 Encoding configuration

The SRCs have been encoded following the standard H.264/AVC using the JM reference software to produce Processed Video Sequences (PVSs). For the

encoding of the PVSs a number of combinations of resolutions (Res), frame rates (FR) and bitrates (BR) have been used, based on several considerations. For the bitrates, the band width fluctuation and the built in limitations running realtime communication over cellular wireless network was taken into count. Based on this and the requirements of a realistic BPP value, and also de facto configurations from industry, the resolution and frame rate was limited, as shown below.

- BR: 50, 150, 300, 600, and 900 kbps
- Res: VGA = 640×480, HVGA (Half VGA) = 480×320, QVGA (Quarter VGA) = 320 × 240, and MVGA (mobile VGA) = 192 × 144
- FR: A: 30, 15, 10 fps, and B: 25, 12.5, 8.33 fps

All used combinations of bitrate, resolution, and frame rate are shown in Table 2, where columns A and B shows the different combinations for the videos with original frame rates 30 fps and 25 fps, respectively. It can be seen in Table 2 that it results in 38 combination for every SRC. To conduct a suitable subjective test the combination of resolution, frame rate and bitrate is based on realistic combinations used in practice, which means that combinations with too low or very high BPP are excluded.

2.2 Pre-processing the test sequences

Before executing the subjective assessment all the processed video sequences (PVSs) are pre-processed. The reason for this is to enable a more realistic test scenario as in streaming or realtime video applications, where the viewing resolution is usually fixed even if the source data is changed, e.g. down sampled, in the context of adapting to fluctuating bandwidth. Therefore all PVSs with spatial resolution MVGA, QVGA, and HVGA were up-scaled to VGA ($640x480$), performed with bicubic filtering as it produces sufficient quality and does not require too much of processing power. Also, all files with subsampled temporal resolution from the original 25fps or 30fps were up-sampled to the original frame rate by frame repetition. This was performed to limit difference in play out during the subjective assessment between the PVSs.

2.3 Subjective Video Quality Assessment Setup

The subjective quality assessment has been performed on 32 test subjects with video sequences described in the previous subsection. Since not all combinations of bitrates and frame rates are used, this results in a total of 190

sequences being used in the test. The setup of the subjective quality assessment follows ITU recommendations as given by ITU-R BT 500-12 [12] for the lab setup of our experiments. Particularly, the method followed was the single stimulus quality evaluation where a test video sequence is shown once without the presence of any explicit reference, corresponding to the reality where users see only the processed version of the video. Overall, the adopted methodology and lab setup has been summarized in [7]. The subjects who participated in the tests were of both genders, mainly students at the university and some staff members, and all of them were considered to be non-expert in the area of video quality assessment. In order to obtain reliable results out of the raw subjective scores on the quality scale of 1 to 100, a screening of the observers scores was employed to discard observers that are considered as outliers. The algorithmic details of these steps are reported in Annex 2 of [12]. After screening of our data no subject had to be rejected. Finally, the mean opinion score (MOS) was calculated and used in this work.

A			B		
Resolution	FR[fps]	BR[kbps]	Resolution	FR[fps]	BR[kbps]
MVGA	10	50	MVGA	8.33	50
MVGA	10	150	MVGA	8.33	150
MVGA	10	300	MVGA	8.33	300
MVGA	15	50	MVGA	12.5	50
MVGA	15	150	MVGA	12.5	150
MVGA	15	300	MVGA	12.5	300
MVGA	30	150	MVGA	25	150
MVGA	30	300	MVGA	25	300
QVGA	10	50	QVGA	8.33	50
QVGA	10	150	QVGA	8.33	150
QVGA	10	300	QVGA	8.33	300
QVGA	10	600	QVGA	8.33	600
QVGA	15	50	QVGA	12.5	50
QVGA	15	150	QVGA	12.5	150
QVGA	15	300	QVGA	12.5	300
QVGA	15	600	QVGA	12.5	600
QVGA	30	150	QVGA	25	150
QVGA	30	300	QVGA	25	300
QVGA	30	600	QVGA	25	600
HVGA	10	150	HVGA	8.33	150
HVGA	10	300	HVGA	8.33	300
HVGA	10	600	HVGA	8.33	600
HVGA	10	900	HVGA	8.33	900
HVGA	15	150	HVGA	12.5	150
HVGA	15	300	HVGA	12.5	300
HVGA	15	600	HVGA	12.5	600
HVGA	15	900	HVGA	12.5	900
HVGA	30	300	HVGA	25	300
HVGA	30	600	HVGA	25	600
HVGA	30	900	HVGA	25	900
VGA	10	300	VGA	8.33	300
VGA	10	600	VGA	8.33	600
VGA	10	900	VGA	8.33	900
VGA	15	300	VGA	12.5	300
VGA	15	600	VGA	12.5	600
VGA	15	900	VGA	12.5	900
VGA	30	600	VGA	25	600
VGA	30	900	VGA	25	900

Table 2: The PVS combinations

3 Results

In the context of adaptive streaming an estimation of the variable bandwidth over a channel is used as a restriction for available bitrates to use for the video codec. With this in mind, the MOS results for the five test sequences with their 38 combinations is presented in Fig. 2-6 where the MOS scores are plotted versus the bitrate. To be able to identify different resolutions and frame rates in the figures the resolution is color coded and the frame rate is marked by different symbols.

3.1 Observations from the MOS results

Some observations can be made directly by studying the MOS results. It can be seen that the sequences with lowest temporal information (TI), Elisa, City, and Children, have clear differentiation between the different resolutions, indicating that the resolution has high significance. There is though a difference regarding City versus Elisa and Children, where the later have the highest spatial information, that for City it clearly differentiates between resolutions even for VGA and HVGA which is not the case for Elisa and Children even if highest resolution is always preferred. It can also be seen for these sequences that higher resolution over increased frame rate is always preferred. For the two sequences with highest temporal information (TI), Soccer and Ice, the tendency is the same but not to the same extent. It can be seen that for increased spatial resolution at lower frame rate is preferred over increase of frame rate but keeping the spatial resolution.

3.2 Analysis of Variance Based Comparison

To further evaluate the MOS scores ANalysis Of VAriance (ANOVA) [13] was used. ANOVA analysis is used to determine whether or not different factors or variables are statistically significant. We considered Res, FR, and BPP for this analysis to see their statistical significance on the MOS scores, where BPP can be seen as an indicator of level of compression. To resolve the relative importance of these variables the multiway ANOVA technique was used and the variables are stated significant if the p-value was below 0.05. We used the Matlab function `anovan` for this purpose. The result from the ANOVA comparison is shown in Table 3. It can be seen in Table 3 that for all the cases resolution (Res) has the highest significance which was also confirmed in the evaluations of the MOS scores illustrated in Fig. 2-6. Further the

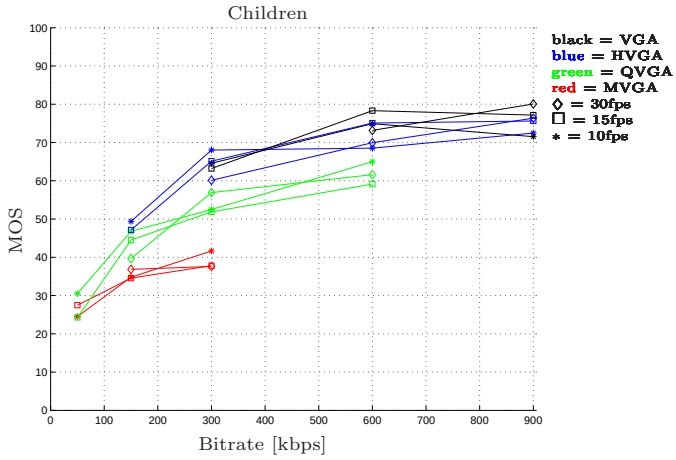


Figure 2: MOS vs. bitrate for different frame rate and resolutions where Children is characterised to have low SI and low TI.

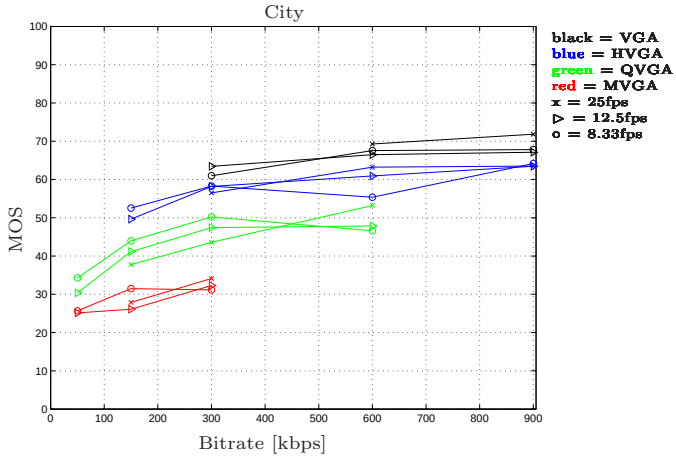


Figure 3: MOS vs. bitrate for different frame rate and resolutions where City is characterised to have high SI and medium TI.

result indicates that BPP is the second most important variable, i.e. the compression level, except for Soccer and Ice which are the two sequences with

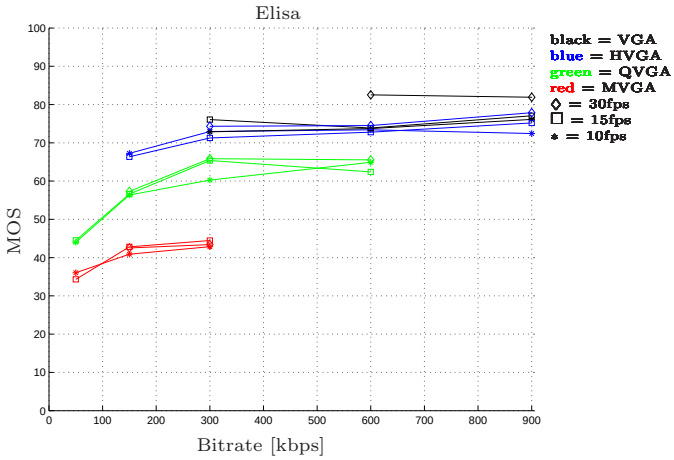


Figure 4: MOS vs. bitrate for different frame rate and resolutions where Elisa is characterised to have medium SI and low TI.

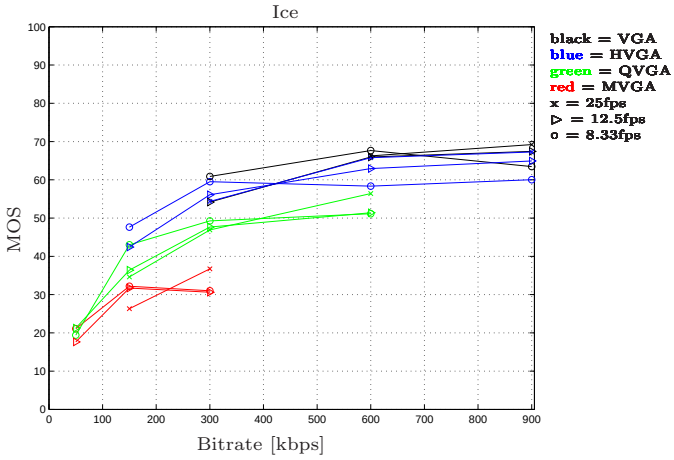


Figure 5: MOS vs. bitrate for different frame rate and resolutions where Ice is characterised to have low SI and high TI.

highest temporal information where the frame rate (FR) has the same or higher significance.

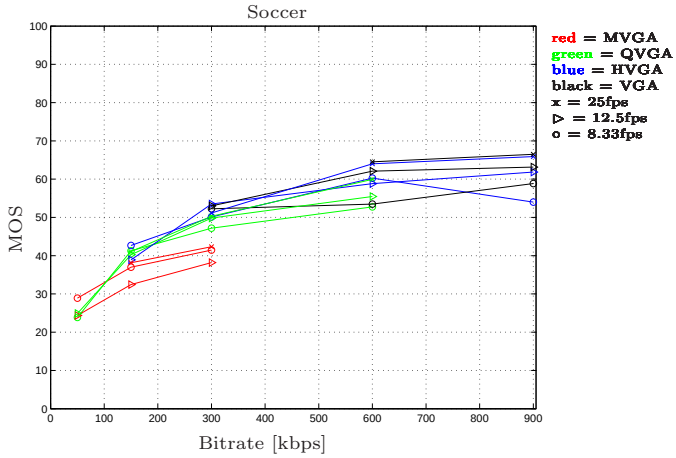


Figure 6: MOS vs. bitrate for different frame rate and resolutions where Soccer is characterised to have low SI and high TI.

Sequences	Variable	Prob>F
All	FR	7.55e-11
	Res	5.81e-33*
	BPP	1.062e-19
Children	FR	4.22e-06
	Res	1.51e-09*
	BPP	9.75e-07
City	FR	0.0026
	Res	0*
	BPP	0.0035
Elisa	FR	0.0013
	Res	0*
	BPP	0.0308
Ice	FR	0.0001
	Res	0*
	BPP	0.0001
Soccer	FR	0.0002
	Res	0*
	BPP	0.0008

Table 3: ANOVA applied to the sequences. The "*" marks the most significant variable.

4 Conclusion

In this paper we have addressed the increasing interest of video communication and its attempt to maximize the perceptual quality during fluctuating bandwidth conditions. In many scenarios of streaming, realtime video communication, or other video applications, adaptive streaming is used to handle fluctuating network bandwidths. A suitable option for adaptive streaming is to scale down the video temporally or spatially before encoding to maintain a desirable level of perceptual quality while viewing resolution is constant. In a subjective assessment with five original sequences, 38 different combination of bitrate, frame rate, and resolution, 32 subjects were used. Both direct and statistical evaluation was made of the MOS scores, where MOS scores were plotted versus the bitrate, and ANOVA was used for statistical analysis. The result shows that preserving the spatial resolution throughout the process has the highest significance even in the scenarios with high temporal information. In comparison, in most studies when increasing the bitrate for a sequence with high SI this results in a preference for increased resolution or decreased quantization, while for sequences with high TI it results in a preference for increased frame rate. One of the reasons to this could be that all sequences were assessed at the same or limited number of different spatial resolutions. In our study, however, four different spatial resolutions were used, and all sequences were assessed at a fixed spatial viewing resolution. Future work planned includes using the presented results to develop a bit-stream based no-reference quality metric, as well as conducting a subjective study using higher resolutions to investigate the same parameters of video coding in other user scenarios.

References

[1] K. Winstein, A. Sivaraman, and H. Balakrishnan, "Stochastic forecasts achieve high throughput and low delay over cellular networks," *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2013)*, April 2-5 2013.

[2] D. Robinson, Y. Jutras, and V. Craciun, "Subjective video quality assessment of HTTP adaptive streaming technologies," *Bell Labs Technical Journal*, vol. 16, no. 4, pp. 5–23, 2012.

[3] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 20–27, 2012.

[4] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, June 2011.

[5] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1316 –1324, nov. 2008.

[6] J. Korhonen, U. Reiter, and J. You, "Subjective comparison of temporal and quality scalability," in *Third International Workshop on Quality of Multimedia Experience (QoMEX)*, 2011, pp. 161–166.

[7] M. Shahid, A. K. Singam, A. Rossholm, and B. Lovstrom, "Subjective quality assessment of H.264/AVC encoded low resolution videos," in *5th International Congress on Image and Signal Processing*, Oct. 2012, pp. 63 –67.

[8] R. Shmueli, O. Hadar, R. Huber, M. Maltz, and M. Huber, "Effects of an encoding scheme on perceived video quality transmitted over lossy internet protocol networks," in *IEEE Transactions on Broadcasting*, vol. 54, Sept 2008, pp. 628–640.

[9] A. Ciancio, J. F. L. De Oliveira, C. D. Estrada, and E. A. B. da Silva, "Impact of encoding configurations on the perceived quality of high definition videoconference sequences," in *IEEE International Symposium Circuits and Systems (ISCAS)*, May 2012, pp. 1716–1719.

[10] J.-S. Lee, F. De Simone, T. Ebrahimi, N. Ramzan, and E. Izquierdo, "Quality assessment of multidimensional video scalability," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 38–46, 2012.

[11] "Subjective video quality assessment methods for multimedia applications," September 1999, iTU-T, Recommendation ITU-R P910.

[12] "ITU-R Radio communication Sector of ITU, Recommendation ITU-R BT.500-12," 2009.

[13] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 8th ed. Ames, IA: Iowa State Univ. Press, 1989.

Part D.3

Comparison of Machine Learning Methods for Quality Estimation of Videos with Diversity in Temporal, Spatial, and Quantization Domains

Part D.3 is submitted as:

A. Rossholm, M. Shahid, B. Lövström *Comparison of Machine Learning Methods for Quality Estimation of Videos with Diversity in Temporal, Spatial, and Quantization Domains.*, submitted to EURASIP Journal on Image and Video Processing, Nov 2014.

Comparison of Machine Learning Methods for Quality Estimation of Videos with Diversity in Temporal, Spatial, and Quantization Domains

Andreas Rossholm, Muhammad Shahid, and Benny Lövström

Abstract

During the last years there has been a growing need for no-reference (NR) methods of video quality assessment due to the widespread use of multimedia services and applications. These demands come from different areas such as network operators and content providers having a strong interest to objectively quantify the level of service quality, or in the context of establishing service-level agreements (SLA) between different parties under which an agreed level of quality has to be guaranteed, or in two-way or streaming audiovisual applications requesting possibility to collect information regarding delivered quality. In this article, a NR objective video quality metric is presented that uses extracted features from the video bitstream as input to different machine learning methods for prediction of perceived quality. A common scenario for video communication services and applications today is that the video stream varies between different temporal and spatial resolutions and quantization levels, and to address these different scenarios locally conducted subjective experiments were performed where video streams up to VGA resolution at 30 fps and 900 kbps were considered. Three different machine learning methods have been tested; Multi-Linear Regression (MLR), Artificial Neural Networks (ANN), and Least Squares Support Vector Machines (LS-SVM), together with different strategies for feature selection. These method are also compared with full-reference perceptual quality estimation methods and the outcome shows the most promising results are obtained using LS-SVM.

1 Introduction

Multimedia services that have gained wide interest include digital television broadcasts, video streaming applications, and realtime audio and video services over the Internet. The global mobile data traffic grew by 81% in 2013, and during 2014, the number of mobile-connected devices is expected to exceed the number of people on earth, according to predictions made by Cisco. The video portion of the mobile data traffic was 53% in 2013, and will exceed 67% by 2018 [1]. With this huge increase in exposure of image and video to the human eye, the interest in delivering quality of experience (QoE) will increase naturally. Videos captured in raw format, called original video, can possess the best possible visual quality but are usually huge in size and can contain variably large amounts of redundant data. To make an efficient use of the available storage capacity and transmission bandwidth it is required to process the raw video to decrease its size, which is accompanied by a drop in the visual quality. This process comes mainly from compression introducing different types of distortions like blockiness, ringing, and blurriness [2]. However, to reduce the amount of data being compressed the raw video can be pre-processed by subsampling it both spatially and temporally. This will introduce distortion like blurriness and ringing for the spatial part and jerkiness from the temporal part [3]. Such a drop in quality is valuable to estimate for many reasons, ranging from the applications that involve making the compression process efficient to applications that involve the delivery of a certain threshold of quality to the end user. The optimum approach to make such a quality estimate is to get the processed videos inspected by human subjects. This inspection is performed by displaying the videos to a panel of people under a recommended setup of the testing environment. However, because of the time and resources required to conduct such procedures for the completion of subjective experiments, objective methods that can make the quality estimate automatically are necessary. For this purpose, based on the extent of the information used from the original video, three different types of methods can be adopted to estimate the visual quality of the processed video. These three approaches are categorized as full-reference (FR), reduced-reference (RR) and no-reference (NR) methods [4]. FR methods require full access to the original video in order to make a quality estimate of the processed video. RR methods require access to a certain set of features that can represent the original video. NR methods, however, perform quality estimation using only the processed video. Two of the most commonly used FR methods are peak signal-to-noise ratio (PSNR) and the mean-squared-error (MSE) which can be calculated for

each frame of the processed video. Since both MSE and PSNR are based on an aligned pixel-by-pixel comparison without any regards to perceptual impacts, such as the impact from spatial and temporal changes, these methods tend to produce results that have poor correlation with human perceptual quality. This has resulted in the development of several new methods as SSIM [5], a video adapted version of SSIM denoted VSSIM [6], VQM [7], and PEVQ [8]. RR methods offer an advantage over FR methods since they need only access to a set of suitable features of the original video besides full access to the processed video. Such a set of features have to be transmitted using an ancillary channel between the server and client end, and performance of RR methods can become limited in the event of any failure in delivery of the required features. NR methods are more versatile in their applications as in many scenarios the access to the original video is not possible, and quality estimation can be made only using the processed video [9].

For many applications involving online and realtime quality monitoring, NR methods emerge as a natural choice for video quality estimation. NR methods can be further classified into three types depending on the level of information about the processed video being used for making an estimate of the quality. These are pixel based, NR-P, bitstream based, NR-B, and a hybrid of both NR-P and NR-B methods [10]. The current paper is related to NR-B based video quality assessment and we highlight its contributions in the following subsection.

1.1 Motivation and Contributions

In recent years, there has been increasing interest in development of NR methods due to the widespread use of multimedia services in the context of wireless communications and telecommunication systems. Applications of NR methods include a wide range of areas such as: network operators and content providers that have a strong interest to objectively quantify the level of service quality delivered to the end-user and inside the network nodes. Also, the involvement of multiple parties between content providers and the end users gives rise to establish service level agreements (SLA) under which an agreed level of quality has to be guaranteed. Another area could be to perform real-time objective quality assessment where resources are limited such as frequency spectrum in wireless communications. To support the wide range of NR applications the main contributions of this paper are the following:

- Our NR models are based on locally conducted subjective experiments

where video streams have variations in temporal and spatial resolutions and quantization levels. This trade-off is of substantial importance in many user scenarios but has previously not been evaluated to a large extent.

- All processed video sequences (PVSs) were scaled to same viewing resolution (VGA) to reflect realistic user scenarios.
- Only extracted features from the H.264/AVC encoded bitstream were used.
- Two Machine Learning (ML) techniques were chosen for evaluation against Multi-Linear Regression (MLR), based on previous studies showing an advantage of non-linear models for predicting video quality, namely Artificial Neural Networks (ANN), and Least Squares Support Vector Machines (LS-SVM). For comparison of the obtained results three FR-methods has been used, PSNR, SSIM, and PEVQ, and the results indicat the superiority of the LS-SVM approach.
- For feature selection both statistical and forward greedy approach showed beneficial results for LS-SVM. This resulted in a reasonably long yet compact feature list.

This differentiation of the test content between temporal and spatial resolutions and quantization levels is usually missing in data bases available and not often addressed in publications. However, in the context of e.g. two-way realtime communication application for mobile devices or tablets, this is the most common scenario with constrained bandwidth conditions.

1.2 Organization of the Paper

The remaining part of this paper is organized as follows. We present an account of related work in Section 2. Relevant details on how the required test-stimuli annotated with subjective assessment was prepared are provided in Section 3. In order to build the proposed models, video feature extraction was performed, as described in Section 4. In Section 5, descriptions of our proposed models are given. The model training including the feature selection procedure is described in Section 6, and finally the Results, and Summary and Conclusion are given.

2 Related Work

Bitstream based methods use readily available data from the encoded bitstream of a processed video and may not require the full decoding of the video for making a quality estimate. The fundamental advantage of the NR-B methods is the lower computational requirement for their mode of operation. With reference to the standardized models recommended by telecommunication standardization sector of International Telecommunication Union (ITU-T), as discussed in [11, 12], a concise review of the state-of-the art techniques for NR-B methods is presented in the following. This includes parametric models (parametric planning model and parametric packet-layer model) and bitstream layer model. A review of state-of-the art ML-based video quality predictors is also given.

2.1 Parametric Planning Model

A rough estimate of the video quality can be made by using a minimal amount of information such as bitrate, codec type, and packet loss rate of a processed video. Such approaches have been classified as parametric planning models. The work item related to this category in ITU-T is known as Opinion model for video-telephony applications, G.1070 [13].

2.2 Parametric Packet-layer Model

The parametric planning based methods can be enhanced in performance by combining additional input from the packet headers of the bitstream about the processed video, and the resulting type of methods is known as packet-layer models. The packet layer models can extract a limited set of parameters including bitrate on sequence or frame level, frame rate and type, and packet loss rate. Parametric packet-layer models are also known as quality of service (QoS) based methods. The work item related to this category in ITU-T is known as Non-intrusive parametric model for the assessment of performance of multimedia streaming (P.NAMS) [14].

2.3 Bitstream Layer Model

The packet-layer models are good only up to a certain extent and further improvement in the performance can be made by adding more information available in the bitstream of a processed video. This way, bitstream layer

methods can use virtually all of the useful information available at bitstream level. The work item Parametric non-intrusive bitstream assessment of video media streaming quality (P.NBAMS) [15] in its Mode 1 (Parsing mode) is related to the bitstream layer models. In these models it is allowed to do any kind of analysis of the bitstream except decoding the bitstream and using the pixel data. The input information includes parameters extracted from the packet header and payload. Besides the parameters included in the parametric models, these models utilize information from the encoded bitstream to analyze the characteristics of the source, e.g. quantization parameter (QP), DCT coefficient, motion vectors (MV), and other pixel information. This makes the model comparatively more complex but it generally offers better performance.

2.4 Machine Learning-Based Video Quality Prediction

Recently machine learning techniques, such as Multi-Linear Regression (MLR), Artificial Neural Networks (ANN), and Support Vector Machines (SVM) have been more commonly applied to construct prediction models. In [16] an overview of the advantages and limitation of using ML for visual quality prediction is given. A low complexity solution of video quality prediction based on bitstream extracted parameters is found in [17]. The features used are mainly related to the encoding parameters and are extracted on sequence level. Low complexity has been achieved by using a simple multi-linear regression system for developing the relationship between the parameters and quality values. A set of 48 bitstream parameters related to slice coding type, coding modes, various statistics of motion vectors and QP value was used in [18] to predict the quality of HDTV video encoded by H.264/AVC. Partial Least Square Regression (PLSR) was here used as a tool for learning by regression between the feature set and subjective assessment. In [19] an improvement approach of [17] is presented where the required number of parameters have been reduced for computational efficiency and the prediction accuracy has been improved by the virtue of the usage of an ANN to predict FR assessment methods. ANN is also used in [20] to estimate perceived quality of H.264/AVC encoded QCIF and CIF resolution sequences using pixel information and using pixel information and extracted features from both original and coded videos. A further improvement of [19] is found in [21] where a larger features set is used and LS-SVM is applied for prediction of subjective MOS for resolution QCIF and CIF. In [22] a novel No-Reference bitstream-based objective video quality metric is presented that is constructed by Genetic Programming-based Sym-

bolic Regression, based on a set of tree-based regression models at different resolution and data bases. In [23] SVM is used for modeling the interaction effect between spatial and temporal quality factors affecting perceived video and image quality using Singular Value Decomposition (SVD) for feature detection in pixel domain. In [24] SVM has also been used to predict video quality, different NR and RR parameters are extracted from the bitstream to predict FR objective video quality metrics.

The performance and test-stimuli in the case of most of the aforementioned contributions in this section are shown in Table 1. It is observed that they use various schemes to generate test data but none of them actually cover wide variety to compare alternatives of quantization levels, and temporal and spatial resolutions prior to compression in order to find perceptually preferred scenarios. To this end, in this paper, we have used a targeted test-stimuli generated in order to find such preferences.

3 Subjective Video Quality Assessment

In the procedure of creating a no-reference method that captures differentiation between both temporal, spatial, and quantization variations an appropriate reference is required for training and validating of the proposed method. Also, the variations should as much as possible reflect realistic scenarios from user situations. This was achieved by conducting a subjective quality assessment where these variations were controlled as well as the video content, the encoding parameters, and the pre-processing of the sequences before viewing and assessment.

3.1 Source sequences

The video content of the sequences should possess a variety of spatio-temporal characteristics to be representative of most commonly used videos. To this end, we followed the ITU recommendation P.910 [25] for the selection of sequences based on spatial perceptual information (SI) and temporal perceptual information (TI). The SI and TI values are calculated in the luminance plane of a video.

For SI, the formula is:

$$SI = \max_{time}[std_{space}[sobel(f_n)]] \tag{1}$$

Metric	PCC	SROCC	RMSE	Ground-truth	Reg.	Test-stimuli
[17]	0.94	–	–	PEVQ	Linear	QCIF, 6 bitrates, 4 frame-rates, 288 samples
[18]	0.93	0.85	0.08	MOS	PLSR	HD, 4 bitrates, 32 samples
[19]	0.97	–	0.06	PEVQ	ANN	QCIF, 6 bitrates, 4 frame-rates, 288 samples
[21]	0.98	0.96	0.97	MOS	LSSVM	QCIF and CIF, 5 bitrates and 2 frame rates each, 120 samples
[20]	0.89	–	–	MOS	ANN	CIF, 7 bitrates, 5 frame rates
[22]	0.88	0.88	–	MOS	GP-symb.	HD, 48 impairment scenarios, 8 SRCs , 384 samples
[24]	0.85	–	0.36	MOS	SVM	CIF, 15 samples

Table 1: Machine learning based related work

where:

- $sobel(f_n)$ is frame at time n, filtered with Sobel filter [26]
- $std_{space}[.]$ is the standard deviation over all pixels in a frame
- $max_{time}[.]$ is the maximum value in the time series

For TI the corresponding formula is

$$TI = \max_{time}[std_{space}[M_n(i,j)]] \tag{2}$$

where:

- $M_n(i,j)$ is the difference between pixel values in space for sequential frames, i.e. difference between the pixel values at the same location (i,j) in space but at successive frames f_n and f_{n-1}.
- $std_{space}[.]$ is the standard deviation over the motion values
- $max_{time}[.]$ is the maximum value in the time series

The five source sequences used in this study are Children, City, Elisa, Ice, and Soccer, all having 10 s duration. The starting frame of each of the sequences is shown in Fig. 2, and Table 2 gives a short description of the content and lists the original frame rate of the sequences as well as their SI and TI characteristics. In Fig. 1, the calculated SI and TI value is presented and it can be seen that the characteristics of the sequences are well distributed

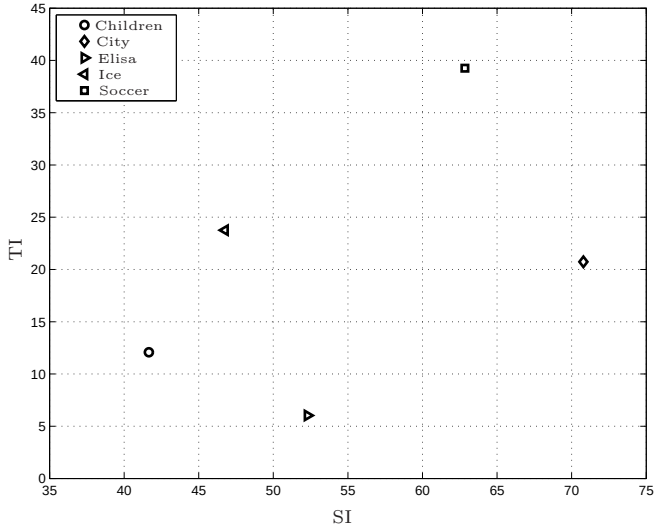
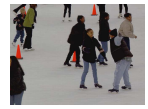


Figure 1: SI and TI plot for the luminance component of the test videos

3.2 Encoding configuration

The source sequences have been encoded following the standard H.264/AVC using the JM reference software to produce Processed Video Sequences (PVSs) [27]. For the encoding of the PVS a number of combinations of resolutions




Children City Elisa Ice Soccer

Figure 2: The first frame of each source sequence used for generation of test sequences

Sequence	Frame rate	Description
Children	30 fps	Two children sitting on the floor, slowly moving, low SI and low TI
City	25 fps	Panning view over a city from an airplane, high SI and medium TI
Elisa	30 fps	Head and shoulder of a talking woman, medium SI and low TI
Ice	25 fps	Several persons skating on white ice, low SI and high TI
Soccer	25 fps	Close up view of soccer game, panning, low SI and high TI

Table 2: The original frame rate and a brief description of the source sequences

(Res), frame rates (FR) and bitrates (BR) have been used, based on several considerations. For the bitrates, the bandwidth fluctuation and the built in limitations running realtime communication over cellular wireless network was taken into account. Based on this, and the requirements of a realistic bits per pixel (BPP) value and the *de facto* configurations from industry, the selection of resolution and frame rate were limited, as shown below.

- BR (kbps): 50, 150, 300, 600, and 900
- Res (pixels):
  - VGA = 640×480
  - HVGA (Half VGA) = 480×320
  - QVGA (Quarter VGA) = 320×240
  - MVGA (Mobile VGA) = 192×144

- FR (fps):
  - A: 30, 15, 10
  - B: 25, 12.5, 8.33

All used combinations of bitrate, resolution, and frame rate are shown in Table 3, where columns A and B show the different combinations for the videos with original frame rates 30 fps and 25 fps, respectively. It can be seen in Table 3 that it results in 38 combination for every source sequence. To conduct a suitable subjective experiment the combination of resolution, frame rate and bitrate is based on realistic combinations used in practice, which means that combinations with too low or very high BPP are excluded.

3.3 Adapting the test sequences

Before executing the subjective assessment all the PVS are adapted for viewing. This is performed to enable a more realistic test scenario, as in streaming or realtime video applications the viewing resolution is usually fixed even if the source data is changed, e.g. downsampled, in the context of adapting to fluctuating bandwidth. Therefore all PVS with spatial resolution MVGA, QVGA, and HVGA were up-scaled to VGA, performed with bicubic filtering as it produces sufficient quality and does not require too much of processing power. Also, all files with sub-sampled temporal resolution from the original 25fps or 30fps were up-sampled to the original frame rate by frame repetition. This was performed to limit difference in playout during the subjective assessment among the PVSs.

3.4 Subjective Video Quality Assessment Setup

The subjective quality assessment has been performed with 32 test subjects for the video sequences described in the previous subsection. Since not all combinations of bitrates and frame rates are used, this results in a total of 190 sequences being used in the test. The setup of the subjective quality assessment follows ITU recommendations as given by ITU-R BT 500-12 [28] for the setup of our experiments. Particularly, the method followed was the single stimulus quality evaluation where a test video sequence is shown once without the presence of any explicit reference, corresponding to a real situation where users see only the processed version of the video. Overall, the adopted methodology and lab setup has been summarized in [29]. The subjects who

A			B		
Resolution	FR[fps]	BR[kbps]	Resolution	FR[fps]	BR[kbps]
MVGA	10	50	MVGA	8.33	50
MVGA	10	150	MVGA	8.33	150
MVGA	10	300	MVGA	8.33	300
MVGA	15	50	MVGA	12.5	50
MVGA	15	150	MVGA	12.5	150
MVGA	15	300	MVGA	12.5	300
MVGA	30	150	MVGA	25	150
MVGA	30	300	MVGA	25	300
QVGA	10	50	QVGA	8.33	50
QVGA	10	150	QVGA	8.33	150
QVGA	10	300	QVGA	8.33	300
QVGA	10	600	QVGA	8.33	600
QVGA	15	50	QVGA	12.5	50
QVGA	15	150	QVGA	12.5	150
QVGA	15	300	QVGA	12.5	300
QVGA	15	600	QVGA	12.5	600
QVGA	30	150	QVGA	25	150
QVGA	30	300	QVGA	25	300
QVGA	30	600	QVGA	25	600
HVGA	10	150	HVGA	8.33	150
HVGA	10	300	HVGA	8.33	300
HVGA	10	600	HVGA	8.33	600
HVGA	10	900	HVGA	8.33	900
HVGA	15	150	HVGA	12.5	150
HVGA	15	300	HVGA	12.5	300
HVGA	15	600	HVGA	12.5	600
HVGA	15	900	HVGA	12.5	900
HVGA	30	300	HVGA	25	300
HVGA	30	600	HVGA	25	600
HVGA	30	900	HVGA	25	900
VGA	10	300	VGA	8.33	300
VGA	10	600	VGA	8.33	600
VGA	10	900	VGA	8.33	900
VGA	15	300	VGA	12.5	300
VGA	15	600	VGA	12.5	600
VGA	15	900	VGA	12.5	900
VGA	30	600	VGA	25	600
VGA	30	900	VGA	25	900

Table 3: *The PVS combinations*

participated in the tests were of both genders, mainly university students but also staff members, and all of them were considered to be non-expert in the area of video quality assessment. In order to obtain reliable results out

of the raw subjective scores on the quality scale of 1 to 100, a screening of the observers scores was employed to discard observers that are considered as outliers. The algorithmic details of these steps are reported in Annex 2 of [28]. After screening of our data no subject had to be rejected. Finally, the mean opinion score (MOS) was calculated and used as one of the target quality measures to predict.

3.5 Distribution of Subjective and Objective Scores

The distribution of the MOS scores from the subjective experiment is seen in Fig. 3. The moments of distribution are calculated and presented in Table 4.

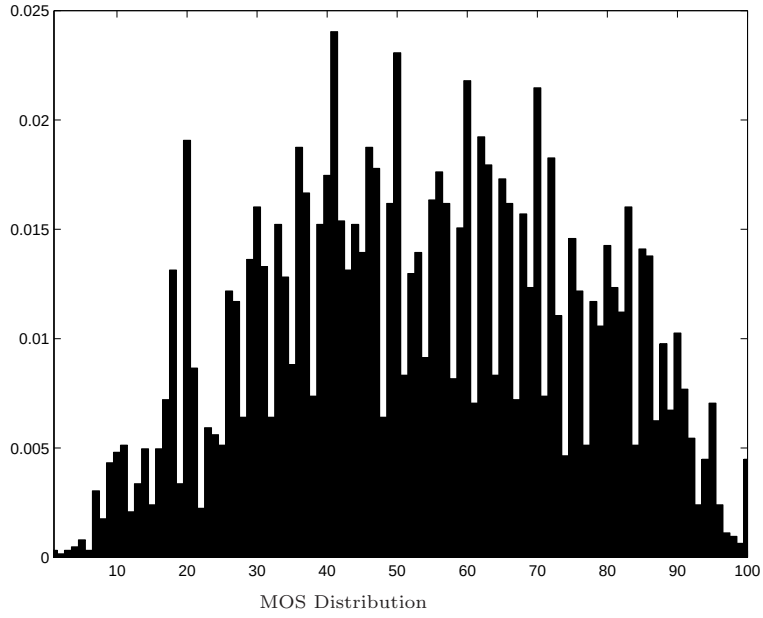


Figure 3: Distribution of MOS from subjective experiments.

It can be seen both from the distribution graph and the moments that there is not a uniform distribution, which is expected since the most extreme combinations of resolution, frame rate, and quantization were excluded from the subjective experiment. The reason for this was to cover as many realistic

Moments	Value
1st Moment (Mean)	53.7
2nd Moment (Variance)	488
(Std.dev.)	22.1
3rd Moment (Skew)	−0.02

Table 4: The 1st, 2nd and 3rd moment of the distribution of MOS scores

scenarios in the context of a subjective experiment as possible. Also a small negative skew is seen.

3.6 Result from Subjective Experiments

The MOS values from the subjective experiments are all shown in Fig. 4. From this figure it can be observed that for the sequences with low TI the resolution is of high significance for the MOS values, whereas for the sequences with high TI the resolution have less impact. It can also be seen that increased resolution is preferred over increased frame rate. A deeper analysis of the implications can be found in [30], where also an analysis of the MOS scores using Analysis of Variance (ANOVA) [31] is reported. ANOVA has been used to discover the statistical significance of different variables to determine the value of the target variable. In the final outcome a three-way ANOVA test was used where Resolution (Res), Bitrate (BR), and Bits per pixel (BPP) were used in the ANOVA test, seen Table 5. It can be seen in Table 5 that for all the cases resolution (Res) has the highest significance which was also confirmed in the evaluations of the MOS scores illustrated in Fig. 4. Further the result indicates that BPP is the second most important variable, i.e. the compression level, except for Soccer and Ice which are the two sequences with highest temporal information where the frame rate (FR) has the same or higher significance.

Sequences	Variable	Prob>F
All	FR	7.55e-11
	Res	5.81e-33*
	BPP	1.062e-19
Children	FR	4.22e-06
	Res	1.51e-09*
	BPP	9.75e-07
City	FR	0.0026
	Res	0*
	BPP	0.0035
Elisa	FR	0.0013
	Res	0*
	BPP	0.0308
Ice	FR	0.0001
	Res	0*
	BPP	0.0001
Soccer	FR	0.0002
	Res	0*
	BPP	0.0008

Table 5: ANOVA applied to the sequences. The "*" marks the most significant variable.

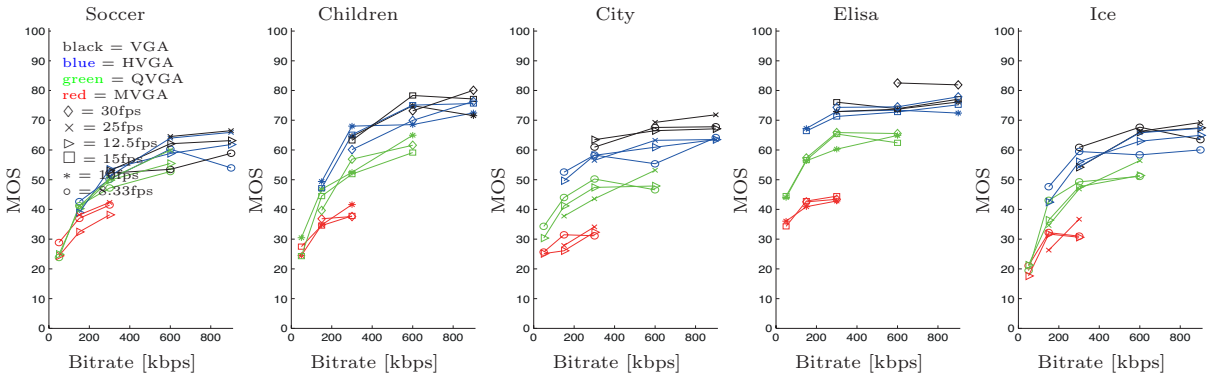


Figure 4: MOS values for the 38 combinations of frame rate and resolution for each of the five sequences, plotted as a function of bitrate. For expanded versions of these figures, see Fig. 2-6 in Part D.2.

3.7 Comparison with state-of-the-art assessment methods

For comparison correlation between MOS result and state-of-the-art FR-assessment method has been analysed. The following FR video quality assessment metrics were considered; PSNR, SSIM, and PEVQ.

3.7.1 PSNR

PSNR, the peak signal-to-noise ratio, is defined as

$$PSNR(n) = 10 \cdot \log \frac{MAX_I^2}{MSE(n)} \tag{3}$$

where MAX_I is the maximum value a pixel can take (e.g. 255 for 8-bit images) and the MSE is the average of the squared differences between the luminance values of corresponding pixels in two frames. MSE is defined as

$$MSE = \frac{1}{UV} \sum_{u=1}^{U} \sum_{v=1}^{V} [I_R(u,v) - I_D(u,v)]^2 \tag{4}$$

where $I_R(u,v)$ denotes the intensity value at pixel location (u,v) in the reference video frame, $I_D(u,v)$ denotes the intensity value at pixel location (u,v) in the distorted video frame, U is the number of rows in a video frame, and V is the number of columns in a video frame. To get a measure for a video sequence a simple averaging over the N frames of a video sequence is calculated as

$$PSNR = \frac{1}{N} \sum_{n=1}^{N} PSNR(n) \tag{5}$$

3.7.2 SSIM

SSIM, the Structural SIMilarity index, considers image degradations as perceived changes in the variation of structural information by combining measures of the distortion in luminance, contrast and structure between two

frames, [5], as

$$SSIM(n) = \frac{[2\mu_{I_R}(n)\mu_{I_D}(n)+C_1][2\sigma_{I_R I_D}(n)+C_2]}{[\mu^2_{I_R}(n)+\mu^2_{I_D}(n)+C_1][\sigma^2_{I_R}(n)+\sigma^2_{I_D}(n)+C_2]} \quad (6)$$

where $\mu_{I_R}(n), \mu_{I_D}(n)$ and $\sigma_{I_R}(n), \sigma_{I_D}(n)$ denote the mean intensity and contrast of the n-*th* reference video frame I_R and distorted video frame I_D, respectively. The constants C_1 and C_2 are used to avoid instabilities in the structural similarity comparison that may occur for certain mean intensity and contrast combinations. Similar as with PSNR, the SSIM value for an entire video sequence of length N may be calculated as

$$SSIM = \frac{1}{N}\sum_{n=1}^{N} SSIM(n) \quad (7)$$

3.7.3 PEVQ

PEVQ, the Perceptual Evaluation of Video Quality from Opticom, is a FR method based on approximations of human visual responses. PEVQ is based on five indicators working in the temporal, spatial, luminance and chrominance domain and where perceptual masking is performed in several stages. Also motion information is used in forming the final measure. PEVQ has been developed for bitrates up to 4 Mbit/s and for resolutions up to VGA. PEVQ is part of the ITU-T Recommendation J.247 [8].

The result from correlation between MOS results and the assessment metrics is shown in Fig 5. Here, it can be seen that objective metrics do not correlate so well with the subjective assessment. One of the main reasons behind this poor performance of the objective metrics is the fact that none of these metrics were originally designed to account for the changes that occur after temporal or spatial upscaling. As it appears, PEVQ does not perform significantly better than PSNR and SSIM in the case of the considered adaptation scenarios.

4 Extracted features

The video features extracted from the bitstream are used to model the impact of various impairments on the video quality. These impairments can

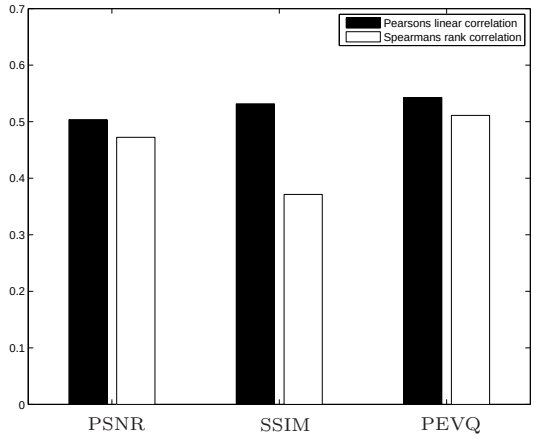


Figure 5: Correlation between MOS and PSNR, SSIM, and PEVQ.

occur due to various sources of distortion, in this context source coding and scaling issues. Driven by this fact, a variety of video features are collected from the signal, bitstream, that may have an impact on perceptual video quality. Table 6 summarizes the used features, their description, and the attributes through which they are related towards video quality. As the features that were selected for this study are related to both compression and resolution, a suitable level of the granularity at which various features can be computed is different for different features. Some features, related to motion vectors, are more suitably computed at macro-block (MB) level. Hence, a bottom-up approach for computing most of these features at MB level has been performed and subsequently averaged at slice level. However, the computational cost of processing features at MB level becomes unnecessary high. As a result of MB level computing, a large amount data will be collected and it can be very complex to build models by the processing of the obtained results. Thus, in order to avoid such complexity issues, we computed the average values of the features to obtain their values at slice level, which is equal to frame level in this test. Moreover, the frame-level feature values were averaged further to obtain their values at video sequence level. In H.264/AVC based coding, several coding modes are typically dependent on the content of the video that is being encoded. Mainly, the coding starts with the prediction of one part of a

Feature	Description	Attribute
1. Avg QP	Average value of quantization parameter	Compression rate
2. std QP	Standard deviation of QP	Variation of Compression
3. Intra [%]	The percentage of I coded MBs in a frame	Content Structure
4. I4 × 4inIfrm [%]	The percentage of MBs of size 4 × 4 in I frame	Content Structure
5. I16 × 16inIfrm [%]	The percentage of MBs of size 16 × 16 in I frame	Content Structure
6. IinPfrm [%]	The percentage of I coded MBs in a P frame	Content Structure
7. P [%]	The percentage of P coded MBs in a frame	Content Structure
8. PSkip [%]	The percentage of MBs coded as *P_Skip* in a frame	Content Structure
9. P16 × 16 [%]	The percentage of MBs coded with no sub-partition of MBs in a frame	Content Structure
10. P8 × 16 [%]	The percentage of MBs coded with 8 × 16 and 16 × 8 partition of MBs in a frame	Content Structure
11. P8 × 8 [%]	The percentage of MBs coded with 8 × 8 partition of MBs in a frame	Content Structure
12. P8 × 8Sub [%]	The percentage of MBs coded with 8 × 8 in a sub-partition of MBs in a frame	Content Structure
13. P4 × 8 [%]	The percentage of MBs coded with 4×8 and 8×4 sub-partition of MBs in a frame	Content Structure
14. P4 × 4 [%]	The percentage of MBs coded with 4 × 4 sub-partition of MBs in a frame	Content Structure

15-16. ΔMV_x, ΔMV_y	The average measures of MV difference values for x and y direction in a frame	Content Motion
17-18. $avg(MV_x)$, $avg(MV_y)$	The average measures of MV values for x and y direction in a frame	Content Motion
19. MV_O [%]	The percentage of MV values equal to zero for x and y direction in a frame	Content Motion
20. ΔMV_O [%]	The percentage of MV difference values equal to zero in a frame	Content Motion
21. Motion Intensity 1	$\sum_{i=1}^{N} \sqrt{avg(MV_x)_i^2 + avg(MV_y)_i^2}$	Content Motion
22. Motion Intensity 2	$\sqrt{avg(MV_x)^2 + avg(MV_y)^2}$	Content Motion
23-24. $\lvert avg(MV_x)\rvert$, $\lvert avg(MV_y)\rvert$	The average measures of abs. MV values for x and y direction in a frame	Content Motion
25. Motion Intensity 3	$\sum_{i=1}^{N} \sqrt{\lvert MV_x\rvert_i^2 + \lvert MV_y\rvert_i^2}$	Content Motion
26. Motion Intensity 4	$\sqrt{\lvert MV_x\rvert^2 + \lvert MV_y\rvert^2}$	Content Motion
27-28. $std(\lvert MV_x\rvert)$, $std(\lvert MV_y\rvert)$	Standard deviation of $\lvert MV_x\rvert$, $\lvert MV_y\rvert$	Content Motion
29. Frame-rate	Average frames per second	Temporal fluidity
30. Resolution	Spatial Resolution	Spatial clarity
31. Bitrate	Coding bitrate	Compression level
32. BPP	Bits per pixel (based on original frame size)	Intra-frame fidelity

Table 6: A Brief Description of the Compression Related Features

video frame usually using adjacent frames of the same sequence to reduce any temporal redundancies. The first frame is intra (I) coded, followed by a predetermined sequence of forward predictive (P) frames. These predictions can

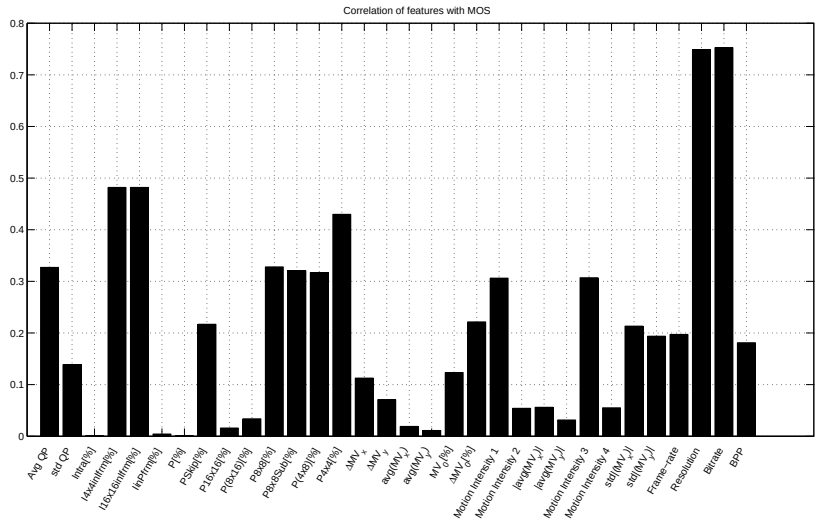


Figure 6: Correlation between features and MOS. The features in the figure appear in the same order as in Table 6.

be applied on a MB, i.e., a 16×16 region of pixels, or on its sub-sized blocks. The available information regarding these coding modes provides a source for estimation of the structural content of a video. The features that we computed from the bitstream and can be grouped in this category are listed from 3 to 14 in Table 6. The features 1-2 represent the sequence mean compression level together with variation in the sequence. In addition, inter frame prediction, which takes advantage from the temporal redundancy between neighboring frames, involves the determination of motion vector information. Such type of information describes the relative movements of blocks of video frames. Besides using the absolute values of the motion vectors of the video, some basic statistics were computed from the bitstream so as to represent the motion content of a video. The selected features of this category are listed from 15 to 28 in Table 6. Feature 29-30 represent temporal and spatial resolution, and 31 is average bitrate which together with 29-30 enable bits per pixel shown as feature 32.

Further, to get a more comprehensive understanding of the features ability to be the input to a NR model that predicts the impact of various impair-

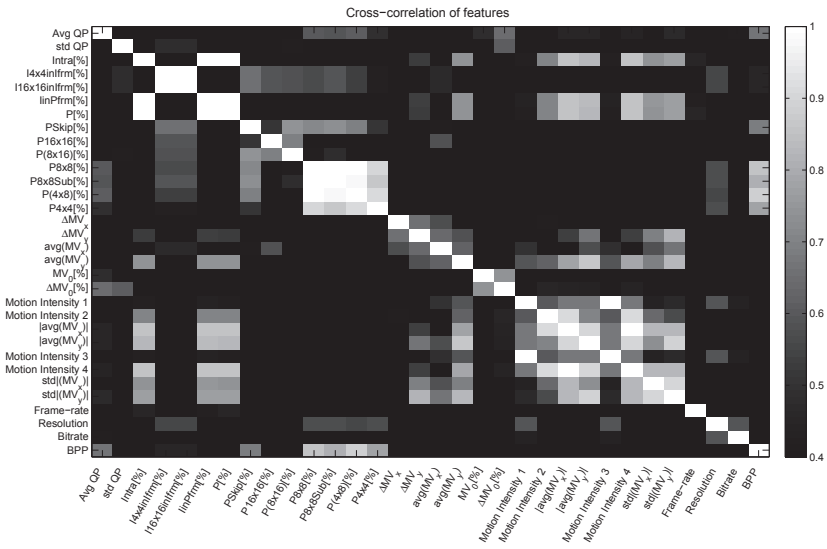


Figure 7: Cross correlation of the selected features. The features in the figure appear in the same order as in Table 6.

ments and create an objective score of the perceived video quality, correlation between the features and MOS values are shown in Fig. 6. Also, the cross-correlation between the features are calculated and shown in Fig. 7.

5 Prediction Models

A NR video quality assessment method based on bitstream-layer features is constructed by building a model that takes a given set of feature values as an input vector $\mathbf{x}$ and produces appropriate values of the quality as an output vector $\mathbf{y}$. In particular, a set of suitable features are extracted from the coded bitstream of the given videos. These features are fed to a system that is capable of interpreting these feature values into a quality score which is representative of the perceptual quality of the videos. A number of FR quality metrics or subjective experiment MOS values are used to compute the ground truth, output vector, for the perceptual quality values. This setup with an input vector and with corresponding output vector or target vector can be seen

as a supervised learning problem and when a continuous variable is desired, as in this case, it becomes a regression problem. In practice, such a mathematical system is required to be trained before it can be used for prediction based on a suitable set of features. Mathematically, the problem can be presented as an observation matrix, $X = [\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \cdots \mathbf{x_N}]$, where $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots \mathbf{x_N}$ are column vectors of a set of M features. Each feature vector $\mathbf{x_n}$ consists of values of the particular feature, in some cases pre-processed, see Section 4, for all the videos, denoted by $x_1, x_2, x_3, \ldots x_M$. The corresponding quality values, target value, for each video is represented by $Y = [y_1, y_2, y_3, \cdots y_N]^T$. In practice, the given set of N videos are divided into two sets namely training set and test or validation set. Once the system has been trained using the training set, its performance is validated using the data from the test set. The ability to perform regression on new input data different from the training set is known as generalization. In the literature this is a well known problem with several solutions both with linear and non-linear methods. The choice of a particular solution depends upon the trade-off preferences made between the complexity and performance of a method. In this paper, we have demonstrated the performance of our proposed model of NR quality estimation using three different methods, as described in the following.

5.1 Multiple Linear Regression (MLR)

With reference to the previous discussion on the prediction problem at hand, the general form of multi-linear regression model is formulated as a function of $\beta = [\beta_0, \beta_1, \beta_2, \beta_3, \ldots, \beta_M]^T$ and $x = [1, x_1, x_2, x_3, \cdots x_M]^T$ where

$$y = x^T \beta \tag{8}$$

The solution is found by minimize the sum-of-square error function and known as the normal equation.

$$\widehat{\beta} = (XX^T)^{-1} XY \tag{9}$$

The normal equation can be derived from the maximization of the likelihood function under gaussian noise, which is equal to minimizing the sum of squares, or to minimize the euclidian distance of the sum-of-square error in a geometrical interpretation of the model. This solution is then used to compute an estimate of the quality vector as follows.

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \ldots + \widehat{\beta}_M x_M \tag{10}$$

This linear function can also be extended by considering linear combinations of fixed nonlinear functions of the input variables as

$$y = \beta_0 + \sum_{j=1}^{J-1} \beta_j \phi_j(x) \tag{11}$$

where $\phi_j(x)$ denotes the nonlinear functions or base functions resulting in J number of parameters. This enables the regression model to be a nonlinear function of the input vector x and it will have the same normal equation as solution. However, since the number of basis functions and their characteristics are fixed and not adaptable during training this limits the applicability for large scaled problems and non-linear models.

5.2 Artificial Neural Network (ANN)

Linear regression is simple but unable to efficiently deal with the possible non-linearities of complex relationships between various feature values and the corresponding quality value. An artificial neural network (ANN) uses several layers of linear combinations of a fixed number of nonlinear basis function which can be adjusted during training. ANNs have been used in different types of applications, besides image processing, to cater for non-linear systems when required. In this paper, in contrast to the contemporary networks such as [32] used for video quality prediction, a reasonably simpler architecture for the ANN model is proposed, using a two-layer feed-forward network with hidden sigmoid neurons and one linear output neuron, as depicted in Fig. 8.

The two-layer feed-forward network is given by

$$y(\mathbf{x}, \omega) = \mathbf{h}\left(\sum_{\mathbf{j=1}}^{\mathbf{D}} \omega_{\mathbf{j}}^{(\mathbf{2})} \sigma\left(\sum_{\mathbf{i=1}}^{\mathbf{M}} \omega_{\mathbf{ji}}^{(\mathbf{1})} \mathbf{x_i} + \mathbf{b}^{(\mathbf{1})}\right) + \mathbf{b}^{(\mathbf{2})}\right) \tag{12}$$

where D is the number of hidden neurons and $\sigma(a)$ the sigmoid function defined as

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \tag{13}$$

and h the linear function. The superscript (1) and (2) in Eq. (12) refers to the first and second layer of the network, respectively. The network is trained

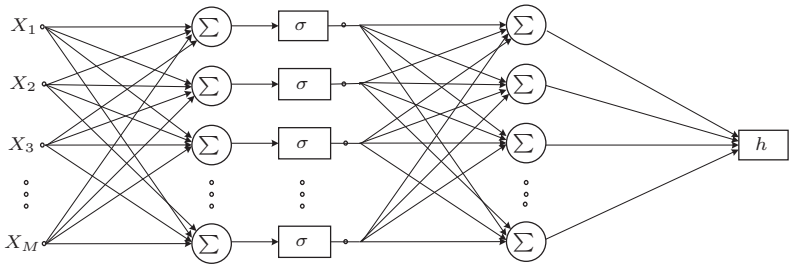


Figure 8: The ANN architecture.

with Levenberg-Marquardt backpropagation algorithm which is considered to be a fast and robust method. This training can be written as

$$\omega_{k+1} = \omega_k - \left(J^T J + \mu I\right)^{-1} J^T e \tag{14}$$

where J is the Jacobian matrix, e is a vector of network errors, I is the identity matrix, and μ is a damping factor.

A properly trained network gives reasonable predictions when it is tested on new inputs not being part of the training. Typically, a new input leads to an expected output if the new input data has some similar characteristics with regards to the data used for training. To this end, the network should be trained to achieve proper level of generalization. In practice, generalization is achieved by regularization where the most widely used technics are Bayesian regularization or early stopping. As a result of the limited amount of subjective experiment data Bayesian regularization was used, adding a regularization term to the error function. The regularization term, which controls the complexity in order to avoid over-fitting, consist of a controlled amount of the sum of squares of the weights, also known as the weight decay [33]. To find the number of hidden neurons, D, that results in best prediction performance and generalization there are no strict design rules. Therefore, for different combinations of features, M, as input to the network, it was trained with various number of hidden neurons to find an efficient system with high performance and no over-fitting.

5.3 Least Square Support Vector Machines (LS-SVM)

The use of neural network for image quality prediction has given results better than the use of linear regression technique [19]. However, neural network techniques often face the problem of over-fitting and the computational complexity of neural network based methods can grow with increase in dimension of input space. Also, neural network based methods are limited by the fixed amount of basis functions. In order to circumvent such issues, support vector machine (SVM) technique is used [34]. SVM belongs to a class of supervised kernel based learning algorithm techniques that solve any problem by mapping the input data set into high dimensional feature space via linear or nonlinear mapping, referred to as the kernel trick. In kernel based learning algorithms the training data is not discarded, instead this is used, or a subset of it, when applied on new data. The usage of support vector machine can be ascertained by considering an intuitive example of the case where non-linear dependencies can be converted to linear dependencies by converting the input data into a higher dimension such as squared values. Similarly, when a support vector machine is used for this type of prediction, the input data is projected to higher dimensional space for building a relationship between the feature values and the measures of quality values. During recent years, a few kernel based models were proposed such as support vector machines, kernel fisher discriminant and kernel principal component analysis which are used for regression, classification, dimensionality reduction. SVM is commonly used for classification and regression analysis [35]. Regression computations in SVM are done based on structural risk minimization (SRM) principle which employs complexity control, the amount of function needed to achieve good generalization, i.e. to avoid the over-fitting issue. In order to simplify the implementation of SVM, in which a solution of inequality constraints is sought, Suykens et al. [36] developed a variant of SVM called least square support vector machines (LS-SVM). It reformulates the standard SVM of solving the Karush-Kuhn-Tucker equation systems to a Lagrangian equation and it helps decrease computational complexity. In essence, this is achieved by modifying SVM through changing the inequality constraints to equalities and instead of slack variables introduce an error variable for which a squared loss function is taken. This transforms the quadratic programming into a set of linear equations which are easier to solve. In our work, we adopted LS-SVM algorithm for regression analysis due to its known ability of handling non-linear data and the simplicity of computation. To motivate the use of support vector machine for solving our prediction problem, we present here the mathematical

procedure behind it. Given a matrix X of N videos of M features with the corresponding N values of the quality measure presented in vector $\mathbf{y}$, where n and m are real positive numbers, as expressed by the following:

$$X = \{\mathbf{x_1}, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N\} \quad X \epsilon R^{MxN} \tag{15}$$

$$\mathbf{y} = \{y_1, y_2, y_3, ..., y_N\}T \quad \mathbf{y} \epsilon R^{Nx1} \tag{16}$$

This data can be taken as a set of features for training data points given by $[\mathbf{x_i}, y_i]_{i=1}^{N}$, where $\mathbf{x}_i \epsilon R^M$ is an input vector that consists of M features of a video and $y_i \epsilon R$ is the corresponding target quality value. LS-SVM estimation is based on a primal-dual formulation in its standard framework. A linear estimation is obtained in a kernel-induced feature space by the following [37]:

$$y = \omega^T \varphi(\mathbf{x}) + b \tag{17}$$

where $\varphi(.)$ represents the mapping to a high dimensional kernel induced feature space, b denotes the bias term and the weight vector ω is in primal weight space. Similar to standard SVM, optimization problem in LS-SVM is formulated as given below.

$$\min_{w,b,e} J(\omega, e) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{i=1}^{N} e_i^2$$

subject to:

$$y_i = \omega^T \varphi(\mathbf{x}_i) + b + e_i \tag{18}$$

where $e_i \epsilon R$ is the error variable. This expression is essentially a ridge regression cost function formulated in the feature space and it can not be solved if ω becomes infinite dimensional. Therefore, a dual problem for this expression is formulated after constructing Lagrangian function for it. The typical convex optimization problem at hand can be solved with the help of Lagrangian multipliers method [38], defined by:

$$L(\omega, b, e; \alpha) = J(\omega, e) - \sum_{i=1}^{N} \alpha_i [\omega^T \varphi(\mathbf{x}_i) + b + e_i - y_i] \tag{19}$$

with Lagrangian multiplier $\alpha_i \epsilon R$. The conditions for optimality are given by

$$\begin{cases} \frac{\partial L}{\partial \omega} = 0 \rightarrow \omega = \sum_{i=1}^{N} \alpha_i \varphi(\mathbf{x}_i) \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0 \\ \frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma_i e_i \\ \frac{\partial L}{\partial \alpha_i} = 0 \rightarrow y_i = \omega^T \varphi(\mathbf{x}_i) + b + e_i. \end{cases} \tag{20}$$

Elimination of ω and e gives the following:

$$\begin{bmatrix} 0 & \mathbf{l}^T \\ \mathbf{1} & \mathbf{K}+\boldsymbol{\Delta} \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix} \tag{21}$$

where $\mathbf{y} = (y_1, y_2, y_3.....y_N)^T$,
$\mathbf{K} = \{k_{ij}\}_{N\times N} = \varphi(\mathbf{x_i})^{\mathbf{T}}\varphi(\mathbf{x_j})$,
$\boldsymbol{\Delta}$=diag$(\frac{1}{\gamma}, ..., \frac{1}{\gamma})$,
$\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)^T$, $\mathbf{1} = (1, ..., 1)^T$

Finally, the prediction function of LS-SVM model in dual space is given by the following.

$$y(\mathbf{x}) = b + \sum_{i=1}^{N} \alpha_i K(\mathbf{x}, \mathbf{x}_i) \tag{22}$$

where K(.,.) is known as kernel function, α_i and b are the solution to the linear system in Eq. (21). We selected radial basis functions (RBF) kernel function which is known for its good performance, see [39], [40], and is given by

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2} \tag{23}$$

where σ is width or insensitive zone of RBF kernel. After training the LS-SVM model using the training data, the values of α (support vector) and bias b are obtained, along with other kernel parameters to be explained in the next paragraph. These values are used afterwards while the model is put under test for performance evaluation.

RBF was selected as kernel function for realization of implicit mapping of given input data into higher dimensional feature kernel space, which results in obtaining better training and testing errors. An optimization algorithm was employed for tuning the hyper parameters in order to achieve better generalization performance and prediction accuracy. There are several optimization algorithms which can be employed on the kernel and since this requires tuning of the hyper parameters for building LS-SVM model, simplex optimization algorithm is used. The mechanism of the tuning process is operated by coupled simulated annealing (CSA) which is better than multi start gradient descent optimization [41]. We applied leave one out cross validation (LOOCV) as a cost function for estimating the performance of LS-SVM model. This algorithm is controlled by the performance metric mean square error (MSE) and offers good generalization ability [42].

The performance and accuracy of LS-SVM model depends on setting of (σ^2, γ), where σ^2 is width of kernel and γ is regularization parameter. For each pair of hyper parameters (σ^2, γ), LOOCV method is performed on training set to estimate the prediction error. Therefore, a robust model is obtained by selecting those optimal pair of hyper parameters which give the lowest MSE.

6 Model Training

In the process of training the models both cross validation and different approaches for the feature selection have been used. A description of these approaches is given in the following.

6.1 Cross Validation Strategy

In order to properly train a prediction model with sufficient capacity of generalization, the training data should be large enough. Moreover, besides the need of high number of data samples, ample variety in the content of SRC videos is required. Although the aforementioned test stimuli can be considered fairly large in sample size, it is problematic to account for the variety in the content itself as well as all the possible variations in the quality levels based on variations in the video encoding configuration. For example, it is usually appropriate to exclude the rather impractical scenarios of quality level such as encoding a VGA video at 50 kbps and 30 fps, which the encoder can not perform as a result of a maximum QP value. Still, when considering performing of the subjective experiment, for practical reasons constraints of the number of variations must be taken. Thus, cross validation has been applied in order to develop our proposed models. We considered leave one out cross validation (LOOCV) for this purpose [43]. In general n-fold and bootstrap are commonly used cross validation (CV) techniques and differences between these techniques lie in selection of data for the training and testing process. The n-fold CV, where n represents total number of instances in the data set, is also referred as the jackknife approach by [44] and as LOOCV by [42]. It is achieved by leaving one instance for testing and the rest of all instances for training and this process is reiterated until all the instances are left out once. This has been applied through the test sequences resulting in a 5-fold CV where the instances was chosen to fulfill the available variety. The following list shows each set of the performed validation, by indicating the SRC of the test sequences used for training and testing.

- CV1: Training [Children, City, Elisa, Ice], Test[Soccer]
- CV2: Training [Soccer, City, Elisa, Ice], Test[Children]
- CV3: Training [Soccer, Children, Elisa, Ice], Test[City]
- CV4: Training [Soccer, Children, City, Ice], Test[Elisa]
- CV5: Training [Soccer, Children, City, Elisa], Test[Ice]

These CV structures were kept both during training and verification to extract as much information as possible from the data set.

6.2 Feature selection

Two procedures have been used to select features, namely the forward greedy selection algorithm and analysis of correlation between features and MOS scores together with cross correlation between the features. In the forward greedy selection procedure the feature that is added at each iteration is the one that generates the largest contribution to the desired criteria, which in this case is Pearson correlation coefficient (PCC). This have been applied for all test cases from 5-fold CV.

To select optimum number of features, early stop is applied in order to reduce the total number of features used when combining the results from CV. In this context this means that PCC is analysed for every added feature, and when PCC is starting to decrease no more features is added. To avoid strange initialization behaviour this is applied from the third iteration and forwards, e.g. see CV 4 ANN in Fig 10. For ANN in the stage of adding feature, different number of neurons in the hidden layer, D, is tested.

In Fig. 9-11 the correlation is shown for each ML method and each added feature.

It can be seen that that PCC is increasing to some point and then starts to decrease. In these figures the blue star marks this early stop point. The red circle marks the maximum correlation and in some of the cases this point occurs later then the blue star. This can be seen as a result of a wrongly chosen feature in an earlier iteration, but where its impact decreases when more features are added later.

The correlation values from max correlation and early stop decision can be seen in Table 7. To select a final feature set the feature selection resulting from these five CVs are combined into one set for each method to form a union ($\bigcup$) of all features occurring in the five test cases. These sets are then used to

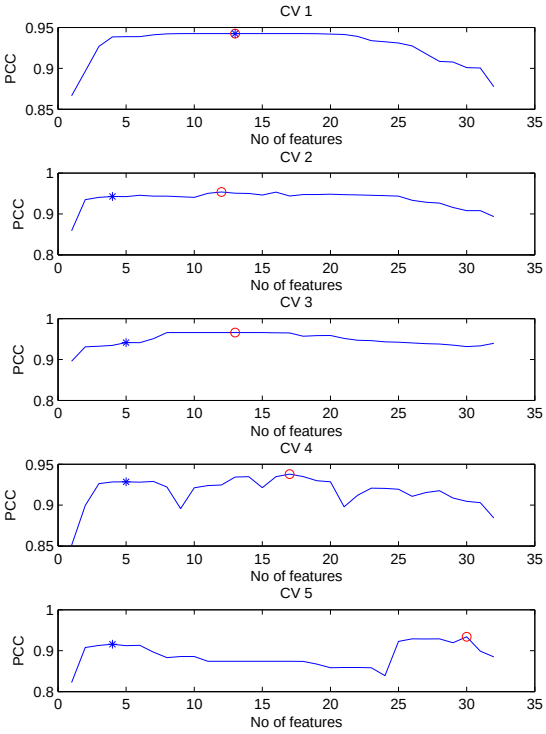


Figure 9: Correlation after adding features for MLR.

obtain final performance evaluation for the different ML methods. It can be seen that the amount of features decreases from 32 to 27 for ANN and to 22 for LS-SVM for max correlation but for MLR still all features are included. For early stop the feature sets are further reduced to 21 for ANN and 15 for LS-SVM, also MLR is decreased to 21 features. It can be seen that there is high similarity between the two non-linear methods.

The second approach utilizes a statistical relationship between the features and the MOS, and also among the features. To this end, correlation between individual feature and MOS is considered. Moreover, the cross-correlation among the features has been considered as well. This approach first adds

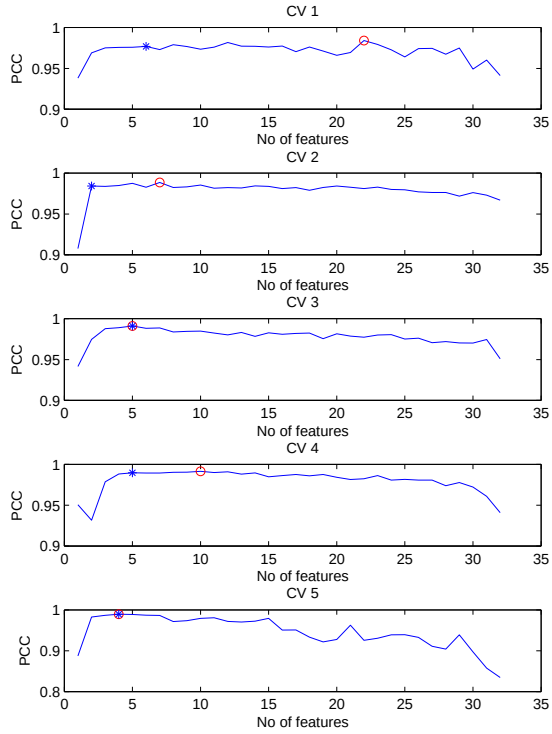


Figure 10: Correlation after adding features for ANN.

feature with the highest correlation with MOS whereupon only one is retained from any two features with high cross-correlation, see Fig 6 and 7. In the process a correlation threshold at 0.3 was set and the features making the cut was then grouped based on its character, e.g. intra or inter related, motion intensity, quantization, resolution, and bitrates. The cross correlation inside these groups, if more than two, was then analysed and high cross-correlation with same MOS correlation was removed. This resulted in 7 features being kept. The resulting features are shown in Table 8.

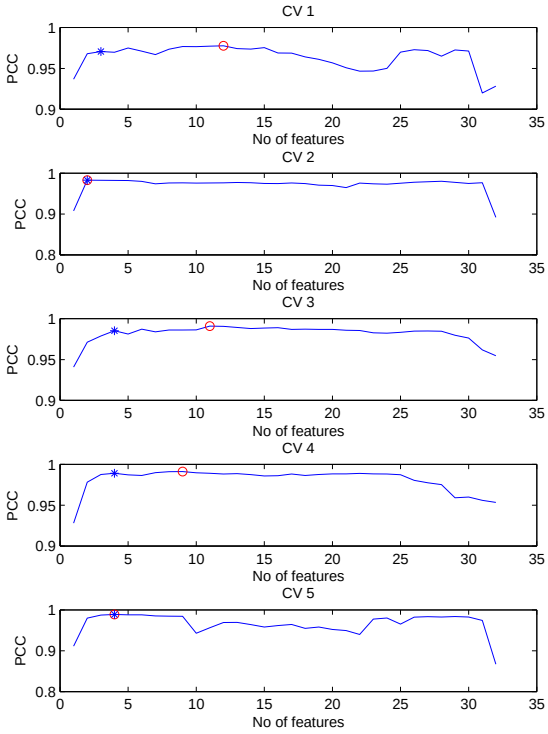


Figure 11: Correlation after adding features for LS-SVM.

7 Results

The feature selection procedures presented in Table 7 show that ML models in the individual 5-fold cross validation achieve very high average correlation, 98-99% for early stop and max correlation for both ANN and LS-SVM, with a small advantage for ANN, and 93-94% for MLR. This shows that non-linear models have an advantage in prediction of the perceived video quality. However, in this stage all cross validation cases used different feature sets. This was addressed in the next phase of the features selection procedure where a union of the selected features was constructed both from the forward greedy approach and from a statistical point of view, and presented in Table 8. In

Procedure	Decision	Method	CV1	CV2	CV3	CV4	CV5	Mean
Forward Greedy	Max PCC	MLR	0.94	0.95	0.97	0.94	0.93	0.95
Forward Greedy	Max PCC	ANN	0.98	0.99	0.99	0.99	0.99	0.99
Forward Greedy	Max PCC	LS-SVM	0.98	0.98	0.99	0.99	0.99	0.99
Forward Greedy	Early Stop	MLR	0.94	0.94	0.94	0.93	0.92	0.93
Forward Greedy	Early Stop	ANN	0.98	0.98	0.99	0.99	0.99	0.99
Forward Greedy	Early Stop	LS-SVM	0.97	0.98	0.99	0.99	0.99	0.98

Table 7: Pearson Correlation from different methods and procedures

this procedure a common set of features was compiled and the 5-fold cross validation process was applied. To evaluate the methods with these feature sets, three parameters recommended by VQEG [45] were used: for description of prediction accuracy Pearson linear correlation coefficient (PCC), for monotonicity Spearman rank order correlation coefficient (SROC), and for consistency the outlier ratio (OR) was used. The final sets of features were used to evaluate the performance of the proposed prediction models. Both mean and standard deviation was calculated and presented in Table 9.

In the table it can be seen that ML methods drop to 0.90 (PCC) for the union of features when max correlation decision is made, and when the early stop is applied this results in a further drop for ANN, similar for MLR, while LS-SVM has some increase. Similar tendencies are seen for SROC. For OR all method has high values but they decrease for all methods when using early stop, especially LS-SVM shows large improvement. This indicates that for LS-SVM the generalization increases when the feature sets decreases, which also could be a result of a limited amount of test data while this effect is not achieved for ANN. For the statistical approach, only using 7 features, ANN and LS-SVM have the highest correlation for both PCC and SROC. Also the OR is decreased for ANN even if it is not reaching LS-SVM level. MLR performs considerably worse for the statistical approach with limited amount of features, as a result if its inability to account for any non-linear

Feature selection	Decision	Method	Selected feature number with reference to Table 6
Union	Max PCC	MLR	All
Union	Max PCC	ANN	1, 2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 32
Union	Max PCC	LS-SVM	1, 2, 3, 6, 7, 8, 9, 10, 11, 15, 18, 20, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
Union	Early Stop	MLR	2, 3, 4, 5, 9, 11, 12, 13, 17, 18, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32
Union	Early Stop	ANN	1, 2, 8, 12, 13, 15, 19, 20, 21, 24, 25, 27, 30, 31, 32
Union	Early Stop	LS-SVM	1, 3, 6, 8, 11, 15, 24, 30, 31, 32
Statistical	Corr. based	All	1, 4, 11, 14, 25, 30, 31

Table 8: Selected Features

relationship between the feature space and the corresponding MOS values. For the objective methods the two image based methods, PSNR and SSIM, perform as expected poorly. PEVQ on the other hand performs good expect for CV5, Ice sequence, where it fails to predict the quality.

8 Summary and Conclusion

We have presented an evaluation of three different machine learning methods used for no-reference video quality estimation. The methods that have been evaluated are multi-linear regression (MLR), artificial neural networks (ANN), and least square support vector machines (LS-SVM). Based on the state-of-the art of present two-way video communication applications and streaming services, the models were applied on test data consisting of subjective MOS for sequences with differences in temporal and spatial resolution as well as in quantization level. These differences are usually not fully covered in the available databases. The models use selected quality-relevant features from the encoded bitstream based on the rationale of encoding fidelity, structural

information of contents, motion information, coding distortion and spatio-temporal complexity. In the feature selection process both forward greedy and statistical approach were used. For the evaluation Pearson linear correlation coefficient (PCC), Spearman rank order correlation coefficient (SROC) and outlier ratio (OR) were used.

It could be seen that the two non-linear methods overall performed better and achieved better generalization, especially for the statistical approach, than MLR. When comparing ANN and LS-SVM, LS-SVM has the highest correlation for both PCC and SROC and also reaches the lowest OR. The FR image based methods, PSNR and SSIM, performs overall poorly as expected since they do not consider temporal aspects in the sequences. However, the FR VQA method PEVQ shows good and stable result except for CV5 which is the only outlier. In CV5, Ice sequence is used as test sequence and causes problem for several of the methods. For example, ANN encounter difficulties with the reduced amount of features in early stop approach, and MLR for the statistical approach. For feature selection Forward Greedy, with and without early stop, and a statistical approach is used where the best result is achieved by the statistical approach which also uses least features.

Future work in this area is to create a merged method based on the test data and apply it on other video sequences. Also, the generic machine learning based methods in this paper, and modifications of these, should be applied to video sequences compressed using Scalable Video Coding. The results in the paper indicates that the methods have potential to successfully predict perceptual quality also when large variations in resolutions and quantization occur.

Proc.	FC	SC	Method	EM	CV1	CV2	CV3	CV4	CV5	Mean(std)	HN
FG	Union	PCC_{max}	MLR		0.88	0.89	0.94	0.88	0.88	0.90(0.03)	-
FG	Union	PCC_{max}	ANN	PCC	0.79	0.93	0.95	0.91	0.73	0.86(0.10)	28
FG	Union	PCC_{max}	LSSVM		0.94	0.97	0.96	0.88	0.74	0.90(0.10)	-
FG	Union	PCC_{max}	MLR		0.92	0.91	0.97	0.92	0.92	0.93(0.02)	-
FG	Union	PCC_{max}	ANN	SROC	0.71	0.92	0.93	0.92	0.77	0.85(0.10)	28
FG	Union	PCC_{max}	LSSVM		0.95	0.97	0.95	0.86	0.81	0.91(0.07)	-
FG	Union	PCC_{max}	MLR		0.58	0.34	0.37	0.89	0.39	0.52(0.23)	-
FG	Union	PCC_{max}	ANN	OR	0.97	0.34	0.63	1.00	0.92	0.77(0.28)	28
FG	Union	PCC_{max}	LSSVM		0.66	0.24	0.21	0.42	0.89	0.48(0.29)	-
FG	Union	ES	MLR		0.88	0.91	0.90	0.89	0.87	0.89(0.01)	-
FG	Union	ES	ANN	PCC	0.91	0.92	0.92	0.92	0.49	0.83(0.19)	18
FG	Union	ES	LSSVM		0.85	0.96	0.91	0.90	0.94	0.91(0.04)	-
FG	Union	ES	MLR		0.91	0.94	0.94	0.94	0.91	0.93(0.01)	-
FG	Union	ES	ANN	SROC	0.90	0.89	0.91	0.90	0.43	0.81(0.21)	18
FG	Union	ES	LSSVM		0.82	0.95	0.91	0.89	0.96	0.90(0.05)	-
FG	Union	ES	MLR		0.63	0.29	0.34	0.68	0.47	0.48(0.17)	-
FG	Union	ES	ANN	OR	0.82	0.32	0.42	0.79	0.87	0.64(0.25)	18
FG	Union	ES	LSSVM		0.34	0.13	0.29	0.29	0.26	0.26(0.08)	-
Stat.	-	-	MLR		0.84	0.89	0.93	0.83	0.29	0.76(0.26)	-
Stat.	-	-	ANN	PCC	0.91	0.96	0.94	0.95	0.92	0.93(0.02)	8
Stat.	-	-	LSSVM		0.93	0.95	0.95	0.92	0.94	0.94(0.01)	-
Stat.	-	-	MLR		0.91	0.93	0.96	0.91	0.25	0.79(0.31)	-
Stat.	-	-	ANN	SROC	0.92	0.93	0.93	0.94	0.90	0.92(0.02)	8
Stat.	-	-	LSSVM		0.91	0.94	0.90	0.90	0.87	0.90(0.02)	-
Stat.	-	-	MLR		0.45	0.55	0.76	0.76	0.79	0.66(0.15)	-
Stat.	-	-	ANN	OR	0.71	0.24	0.21	0.53	0.66	0.47(0.23)	8
Stat.	-	-	LSSVM		0.47	0.21	0.13	0.58	0.32	0.34(0.18)	-
-	-	-	PSNR		0.35	0.43	0.25	0.91	0.25	0.44(0.28)	-
-	-	-	PEVQ	PCC	0.80	0.81	0.87	0.95	0.04	0.68(0.41)	-
-	-	-	SSIM		0.43	0.71	0.22	0.95	0.04	0.45(0.39)	-
-	-	-	PSNR		0.34	0.41	0.10	0.97	0.20	0.40(0.34)	-
-	-	-	PEVQ	SROC	0.84	0.83	0.87	0.96	0.06	0.71(0.37)	-
-	-	-	SSIM		0.36	0.66	0.18	0.91	0.06	0.43(0.35)	-

Table 9: Final result of combined features, including procedure (Proc.) to combine features, feature combination (FC) method in CV, stop criteria (SC) for forward greedy (FG) procedure, used regression method or video quality assessment method, evaluation method (EM), cross validation sets (CV), and finally mean and standard deviation (std). For ANN the number of neuron in hidden layers is also presented (HN).

References

[1] Cisco Visual Networking Index, "Global mobile data traffic forecast update, 2013-2018," *Cisco white paper*, 2014.

[2] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. Wiley, 2005.

[3] S. Möller and A. Raake, Eds., *Quality of Experience: Advanced Concepts, Applications and Methods*. Cham: Springer, 2014.

[4] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactios on Image Processing*, vol. 13, pp. 600–612, April. 2004.

[6] Z. Wang, L. Lu, and A. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication, Special issue on Objective video quality metrics*, vol. 19, pp. 121–132, Feberuary. 2004.

[7] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, pp. 312–322, September. 2004.

[8] I.-T. R. J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference," 2008.

[9] Z. Wang and A. Bovik, "Reduced- and no-reference image quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29–40, Nov 2011.

[10] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 40, Aug. 2014.

[11] A. Takahashi, D. Hands, and V. Barriac, "Standardization activities in the ITU for a QoE assessment of IPTV," *IEEE Communications Magazine*, vol. 46, no. 2, pp. 78 –84, february 2008.

[12] F. Yang and S. Wan, "Bitstream-based quality assessment for networked video: a review," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 203–209, 2012.

[13] "ITU-T Recommendation G.1070: Opinion model for videotelephony applications," http://www.itu.int/rec/T-REC-G.1070, 2012, [Online; Accessed 11-April-2014].

[14] "ITU-T Recommendation P.1201: Parametric non-intrusive assessment of audiovisual media streaming quality," http://handle.itu.int/11.1002/1000/11727, 2012, [Online; Accessed 11-April-2014].

[15] "ITU-T Recommendation P.1202: Parametric non-intrusive bitstream assessment of video media streaming quality," http://handle.itu.int/11.1002/1000/11730, 2012, [Online;Accessed 11-April-2014].

[16] P. Gastaldo, R. Zunino, and J. Redi, "Supporting visual quality assessment with machine learning," *EURASIP Journal on Image and Video Processing*, September 2013.

[17] A. Rossholm and B. Lövström, "A new low complex reference free video quality predictor," in *IEEE 10th Workshop on Multimedia Signal Processing,*, oct. 2008, pp. 765 –768.

[18] C. Keimel, M. Klimpke, J. Habigt, and K. Diepold, "No-reference video quality metric for HDTV based on H.264/AVC bitstream features," in *Image Processing (ICIP), 18th IEEE International Conference on*, sept. 2011, pp. 3325 –3328.

[19] M. Shahid, A. Rossholm, and B. Lövström, "A reduced complexity no-reference artificial neural network based video quality predictor," in *International Congress on Image and Signal Processing*, vol. 1, oct. 2011, pp. 517 –521.

[20] H. E. Khattabi, A. Tamtaoui, and D. Aboutajdine, "Video quality assessment measure with a neural network," *International Journal of Computer and Information Engineering*, vol. 4, no. 3, pp. 167–171, 2010.

[21] M. Shahid, A. Rossholm, and B. Lövström, "A no-reference machine learning based video quality predictor," in *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 176–181.

[22] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, "Constructing a no-reference h.264/avc bitstream-based video quality metric using genetic programming-based symbolic regression," *Circuits and Systems for Video Technology, IEEE Transactions*, vol. 23, pp. 1322–1333, April 2013.

[23] M. Narwaria and W. Lin, "Svd-based quality metric for image and video using machine learning," *IEEE Transactions on Systems, Man,and Cybernetics - Part B: Cybernetics*, vol. 42, p. 347 364, April 2012.

[24] B. Wang, D. Zou, and R. Ding, "Support vector regression based video quality prediction," in *IEEE International Symposium on Multimedia (ISM)*, December.

[25] "Subjective video quality assessment methods for multimedia applications," September 1999, iTU-T, Recommendation ITU-R P910.

[26] Gonzalez and P. Wintz, *Digital Image Processing, 3rd edition.* Prentice Hall, 2008.

[27] "H.264/AVC JM Reference Software, ver. 18.2," http://iphome.hhi.de/suehring/tml, [Online; accessed: 04-April-2014].

[28] "ITU-R Radio communication Sector of ITU, Recommendation ITU-R BT.500-12," 2009.

[29] M. Shahid, A. K. Singam, A. Rossholm, and B. Lovstrom, "Subjective quality assessment of H.264/AVC encoded low resolution videos," in *5th International Congress on Image and Signal Processing*, Oct. 2012, pp. 63 –67.

[30] A. Rossholm, M. Shahid, and B. Lovstrom, "Analysis of the impact of temporal, spatial, and quantization variations on perceptual video quality," in *Network Operations and Management Symposium (NOMS), 2014 IEEE*, May 2014, pp. 1–5.

[31] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 8th ed. Ames, IA: Iowa State Univ. Press, 1989.

[32] D. Culibrk, D. K. D, P. Vasiljevic, M. Pokric, and V. Zlokolica, "Feature selection for neural-network based no-reference video quality assessment," *Lecture Notes in Computer Science*, vol. 5769/2009, pp. 633–642, 2009.

[33] M. Hagan, H. Demuth, and M. Beale, *Neural Network Design*. Boston: MA: PWS Publishing, 1996.

[34] J. Suykens, J. D. Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: robustness and sparse approximation," *NEUROCOMPUTING*, vol. 48, pp. 85–105, 2002.

[35] Vapnik.V, *Statistical learning theory*. Wiley, 1998.

[36] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," pp. 293–300, Jun. 1999.

[37] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, and B. D. M. J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore, 2002.

[38] H. Xing and T. Jin, "Detection of weak signal in chaotic clutter using advanced ls-svm regression," in *2nd International Congress on Image and Signal Processing*, oct. 2009, pp. 1–5.

[39] W. Huang, F. Huang, and J. Song, "An SVM model for water quality monitoring using remote sensing image," in *Proceedings of The Second International Symposium on Networking and Network Security*, april 2010.

[40] I. Steinwart, D. Hush, and C. Scovel, "An explicit description of the reproducing kernel hilbert spaces of gaussian RBF kernels," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4635 –4643, oct. 2006.

[41] S. Xavier-de Souza, J. Suykens, J. Vandewalle, and D. Bolle, "Coupled simulated annealing," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 2, pp. 320 –335, Apr. 2010.

[42] I. H.Witten and E.Frank, *Data mining: Practical machine learning tools and techniquess*, 3rd ed. Morgan Kaufmann, 2011.

[43] G. Rubio, H. Pomares, I. Rojas, L. J. Herrera, and A. Guillén, "Efficient optimization of the parameters of LS-SVM for regression versus cross-validation error," in *Proceedings of the 19th International Conference on Artificial Neural Networks: Part II*, 2009, pp. 406–415.

[44] R. O.Duda, P. E.Hart, and D. G.Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, 2000.

[45] "Full reference television phase II subjective test plans," 2002, objective Quality Model Evaluation Criteria.

ABSTRACT

The use of audio and video communication has increased exponentially over the last decade and has gone from speech over GSM to HD resolution video conference between continents on mobile devices. As the use becomes more widespread the interest in delivering high quality media increases even on devices with limited resources. This includes both development and enhancement of the communication chain but also the topic of objective measurements of the perceived quality. The focus of this thesis work has been to perform enhancement within speech encoding and video decoding, to measure influence factors of audio and video performance, and to build methods to predict the perceived video quality.

The audio enhancement part of this thesis addresses the well known problem in the GSM system with an interfering signal generated by the switching nature of TDMA cellular telephony. Two different solutions are given to suppress such interference internally in the mobile handset. The first method involves the use of subtractive noise cancellation employing correlators, the second uses a structure of IIR notch filters. Both solutions use control algorithms based on the state of the communication between the mobile handset and the base station.

The video enhancement part presents two post-filters. These two filters are designed to improve visual quality of highly compressed video streams from standard, block-based video codecs by combating both blocking and ringing artifacts. The second post-filter additionally performs sharpening.

The third part addresses the problem of measuring audio and video delay as well as skewness between these, also known as synchronization. This method is a black box technique which enables it to be applied on any audiovisual application, proprietary as well as open standards, and can be run on any platform and over any network connectivity.

The last part addresses no-reference (NR) bitstream video quality prediction using features extracted from the coded video stream. Several methods have been used and evaluated: Multiple Linear Regression (MLR), Artificial Neural Network (ANN), and Least Square Support Vector Machines (LS-SVM), showing high correlation with both MOS and objective video assessment methods as PSNR and PEVQ. The impact from temporal, spatial and quantization variations on perceptual video quality has also been included, together with the trade off between these, and for this purpose a set of locally conducted subjective experiments were performed.

ISSN 1653-2090
ISBN: 978-91-7295-295-9

2014:16