8-24-2015

Scalable Speech Coding for IP Networks

Koji Seto
Santa Clara University

Scalable Speech Coding for IP Networks

by

Koji Seto

DISSERTATION

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

in Electrical Engineering

in the School of Engineering

at Santa Clara University

Santa Clara, California

June 2014

SANTA CLARA UNIVERSITY

Department of Electrical Engineering

Date: _____

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY

Koji Seto

ENTITLED

Scalable Speech Coding for IP Networks

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING

_____
Professor Tokunbo Ogunfunmi, Committee Chair

_____
Professor Samiha Mourad

_____
Professor Nam Ling

_____
Professor Shoba Krishnan

_____
Professor Sim Narasimha

Scalable Speech Coding for IP Networks

by

Koji Seto

Acknowledgement

I would like to express my deepest gratitude to my advisor, Professor Tokunbo Ogunfunmi, for his support, valuable guidance, and patience throughout the course of my research. He also provided the excellent research environment which helped me to concentrate on my research. I would also like to thank the members of my doctoral committee, Professors Samiha Mourad, Nam Ling, Shoba Krishnan, and Madihally (Sim) Narasimha, for their valuable comments and suggestions. The discussions we had during the oral examination led to one of the ideas used in this work.

I am also thankful to my fellow students at Santa Clara University, especially to my lab colleagues for their help and insightful discussions. I greatly appreciate all participants in the subjective speech quality tests that I administered to obtain some of the results presented in this work.

Finally, I would like to express my appreciation to my family for their continuous support and encouragement. Especially, I would like to thank my wife, Kathy Peng. I could not have come this far without her understanding and support.

Abstract

Scalable Speech Coding for IP Networks

by

Koji Seto

Doctor of Philosophy in Electrical Engineering

Santa Clara University, Santa Clara

The emergence of Voice over Internet Protocol (VoIP) has posed new challenges to the development of speech codecs. The key issue of transporting real-time voice packet over IP networks is the lack of guarantee for reasonable speech quality due to packet delay or loss.

Most of the widely used narrowband codecs depend on the Code Excited Linear Prediction (CELP) coding technique. The CELP technique utilizes the long-term prediction across the frame boundaries and therefore causes error propagation in the case of packet loss and need to transmit redundant information in order to mitigate the problem. The internet Low Bit-rate Codec (iLBC) employs the frame-independent coding and therefore inherently possesses high robustness to packet loss. However, the original iLBC lacks in some of the key features of speech codecs for IP networks: Rate flexibility, Scalability, and Wideband support.

This dissertation presents novel scalable narrowband and wideband speech codecs for IP networks using the frame independent coding scheme based on the iLBC. The rate flexibility is added to the iLBC by employing the discrete cosine transform (DCT) and

the scalable algebraic vector quantization (AVQ) and by allocating different number of bits to the AVQ. The bit-rate scalability is obtained by adding the enhancement layer to the core layer of the multi-rate iLBC. The enhancement layer encodes the weighted iLBC coding error in the modified DCT (MDCT) domain. The proposed wideband codec employs the bandwidth extension technique to extend the capabilities of existing narrowband codecs to provide wideband coding functionality. The wavelet transform is also used to further enhance the performance of the proposed codec.

The performance evaluation results show that the proposed codec provides high robustness to packet loss and achieves equivalent or higher speech quality than state-of-the-art codecs under the clean channel condition.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Advances and wide acceptance of voice over Internet protocol (VoIP) [1] have been driving the evolution of telephony technologies in recent years. The transition from the legacy public switched telephone network (PSTN) to IP-based communications is already under way. Voice communication over IP networks has gained popularity and may become the dominant service for overall telephony including the wireless telephony in the near future [1]–[3]. According to [3], the Technological Advisory Council (TAC) for the Federal Communications Commission (FCC) recommended a target date of 2018 as the end of the PSTN. The Technology Transitions Policy Task Force was established on December 10, 2012 to provide recommendations to modernize the Commission's policies in response to evolving communications networks, often called the “IP transition”. AT&T filed the petition to launch a proceeding concerning the time-division multiplexing (TDM)-to-IP transition with the FCC [4] in 2012. In a recent filing with the FCC [5], AT&T proposed two trials involving the transition of two wire centers to all IP services and hopes to continue this transition in all wire centers in order to meet their goal of completing the IP transition by the end of 2020.

On the other hand, the emergence of VoIP has posed new challenges to development of speech codec. The key issue of transporting real-time voice packet over IP networks is the lack of guarantee for reasonable speech quality due to packet delay or loss. The characteristics of IP network channel is constantly changing. Especially, packet traffic on public Internet can be unpredictable and its channel is expected to produce much higher packet delay or loss rate than managed networks. Therefore, voice communication over public Internet is less reliable. Reliability of VoIP can be increased by controlling IP networks. The IP multimedia subsystem (IMS) is a network functional architecture for

multimedia service delivery and suited for controlling the multimedia traffic by utilizing quality of service (QoS). It uses a VoIP implementation based on session initiation protocol (SIP), and runs over the standard IP. By relying on managed networks using IMS, packet loss rate is reduced and bursty loss pattern of IP networks becomes less severe although the packet loss is still the main cause of performance degradation and a codec that is robust to packet loss is required.

One way to reduce packet loss is to make use of bit-rate adaptation. If the codec has the capability of multi-rate operation, the packet loss rate can be reduced by lowering the bit rate. Thus, the functionality of speech codec that allows its bit rate to adapt to the current available channel capacity is of significant importance because the efficient channel usage is maintained by adjusting the congestion of packet traffic. When the encoder adjusts the bit rate, it requires information about the current channel condition, which means that feedback is required. The RTP control protocol (RTCP) is used along with the real-time transport protocol (RTP) to provide feedback on the quality of speech transmission for VoIP applications. However, RTCP is not always enabled and feedback is slow. Therefore, the instantaneous adaptation to the current channel condition without the need of feedback is the attractive feature of VoIP application due to the requirement of short time delay for real-time communication. Bit-stream scalability is a promising technique that makes it possible to adjust the bit rate to the desired value by truncating the bit stream at any point of a communication system. Low packet delay or loss rate can be maintained by adjusting the bit rate of voice traffic instantaneously. Note that the benefits of scalability are most enjoyed by the codec used for public Internet. The jitter buffer management (JBM) can also be used to mitigate the effect of delay jitter and is achieved by buffering incoming packets at the receiver and delaying their playout so that most of the packets are received before their scheduled playout times.

1.1 Motivation

Most of the widely used narrowband codecs such as adaptive multi-rate (AMR) [6] or G.729 [7] depend on the code excited linear prediction (CELP) [8] coding technique. The

CELP technique utilizes the long-term prediction (LTP) across the frame boundaries and therefore causes error propagation in the case of packet loss and need to transmit redundant information in order to mitigate the problem. Some of the simple solutions were proposed in [9], which requires significant increase in bit rate and delay. Recent approach was introduced in [10] to reduce the error propagation after lost frames by replacing the long-term prediction with a glottal-shape codebook in the subframe containing the first glottal impulse in a given frame, and utilized in G.718 [11]. Another approach which depends on low bit-rate redundancy frames and an LTP scaling parameter can be found in the recent codec called Opus [12].

The internet low bit-rate codec (iLBC) [13], [14] employs the frame-independent coding and therefore inherently possesses high robustness to packet loss. This frame independence is achieved by applying the adaptive codebook (CB) both forward and backward in time, starting from the start state which is a short segment of samples with the highest energy within the linear prediction (LP) residual signal. The start state captures pitch information in voiced speech and accurate noise-like information in unvoiced speech, and enables the operation of the adaptive CB without depending on the history of the LP residual signal. Therefore, when packets are lost, the effect of speech quality degradation is limited without depending on transmission of redundant information. Due to its inherent robustness to packet loss, iLBC quickly became a popular choice of speech codecs for VoIP applications and was adopted by Skype and Google Talk. However, the original iLBC is a narrowband codec and operates at fixed bit rates of 13.33 kbps for the frame size of 30 ms or 15.2 kbps for the frame size of 20 ms, and therefore lacks in some of the key features of speech codecs for IP networks: **Rate flexibility**, **Scalability**, and **Wideband support** [15]. In addition, the benefit of high robustness to packet loss comes at the expense of a high bit rate.

Multi-rate operation is used in most of the recent speech codecs because it enables the speech codec to adapt its bit rate in order to achieve the best possible speech quality under the current channel condition.

Scalable speech coding techniques [16] have been the subject of intense research, and the need for scalable speech coding has been clearly recognized by the industry, resulting

in new standardization activities [15]. Indeed, bit-stream scalability facilitates the deployment of new codecs that are built as embedded extensions of widely deployed codecs. In addition, scalability gives more flexibility for VoIP applications. It allows the bit stream to be truncated at the decoder or at any point of the communication system to adapt the bit rate in response to the time-varying channel characteristics without the need of feedback. Note, however, that the bit rate scalable structure may cause the performance degradation as is the case for MPEG-4 scalable speech coding [17].

In recent years, a clear trend toward high quality voice communication has been recognized. Increasing network bandwidth and processing power have already allowed wideband speech coding to be used for IP phone and softphone over IP networks. Recent smartphones also started to adopt wideband speech codecs. Most of the recent wideband speech codecs such as ITU-T G.729.1 [18] and G.718 utilize scalable structure to encode wideband signals, which leads to higher speech quality while providing the flexibility of the bit-rate adaptation. However, they still depend on the CELP coding technique for the core layer and low bit rate operations, and therefore need to utilize the redundant information in order to limit the error propagation in the case of packet loss.

In this work, we add three functionalities to the iLBC: Rate flexibility, Scalability, and Wideband support to develop novel scalable narrowband and wideband speech codecs for IP networks using the frame-independent coding scheme based on the iLBC.

1.2 Related Work

The rate-flexible solution for iLBC was first introduced in [19] by employing the multi-pulse approach to encode the start state. The method used in the encoding procedure was based on an analysis-by-synthesis approach to search and quantize the Multi-Pulse (MP)-based start state. The effect of the adaptive CB in forward and backward directions is first captured in a non-square synthesis matrix to enable the analysis-by-synthesis approach. A variable bit rate is achieved by varying the number of pulses used to approximate the start state. The performance improved results was reported in [20]. The improvements were achieved by reallocating bits from the adaptive

CB refinement procedure, reducing the length of the start state vector, utilizing an adaptive pulse gain quantization scheme, and extending the use of entropy coding. Whereas the performance was close to AMR, the codec relied on the analysis-by-synthesis technique for start state encoding, which requires the intensive computational power. The codec does not employ scalability and supports only narrowband speech.

1.3 Contributions of Our Work

First, the rate flexibility was successfully added to the iLBC in [21]. The proposed multi-rate iLBC uses the discrete cosine transform (DCT) and entropy coding to encode the start state. Therefore, our codec has the advantage of lower computational complexity and achieved similar speech quality at most of the bit rates compared to the multi-rate iLBC previously presented by Garrido, et al. [19], [20]. Various approaches to improve performance of the multi-rate iLBC were presented in [22]. The simulation results showed that when all the improvement schemes were combined, the performance was improved at all the bit rates compared to the previous results despite the fact that the Huffman table structure was significantly simplified.

Second, the bit-rate scalable multi-rate iLBC was proposed in [23]. Its scalable structure was constructed by the addition of enhancement layer to the core layer of multi-rate iLBC in LP residual domain. The enhancement layer encodes the iLBC coding error in LP residual domain. The experimental simulation results showed that the proposed framework improved speech quality at high bit rates compared to the non-scalable version. The performance enhanced scalable narrowband multi-rate iLBC was introduced in [24], and the proposed codec was re-designed based on the subjective listening quality instead of the objective quality. In particular, perceptual weighting and the modified DCT (MDCT) with short overlap in weighted signal domain were employed along with the improved packet loss concealment (PLC) algorithm. The subjective evaluation results showed that the speech quality of the proposed codec was equivalent to that of state-of-the-art codec, G.718, under both a clean channel condition and lossy channel conditions. This result is significant considering that development of the proposed codec was still in

early stage. In order to further improve the performance, we proposed a new scalable speech codec using the discrete wavelet transform (DWT) in [25]. The DWT was employed instead of the MDCT to encode the core-layer coding error in the enhancement layer. The experimental simulation results showed that the DWT was a promising technique to use for encoding highly non-stationary signals such as the coding error.

Third, the scalable wideband speech codec based on the multi-rate iLBC was introduced in [26]. The coder adopted a split-band structure, where the input signal sampled at 16 kHz was split into two sub-bands. Both the lower- and higher-band signals were encoded by the scalable multi-rate iLBC. Based on the objective evaluation, the proposed codec provided high robustness to packet loss and achieved slightly higher voice quality than G.729.1 at the bit rate higher than about 24 kbps under clean channel condition. We proposed the performance enhanced version of the scalable wideband codec in [27]. The proposed codec utilized the time domain bandwidth extension (TDBWE) for higher-band coding to improve the low bit-rate performance and the efficient coding structure was also employed in enhancement layers. The objective test results showed that significant improvement was achieved at low bit rates and the proposed codec outperformed G.729.1 at most bit rates. In order to provide further improvements in performance, the wavelet packet transform (WPT) was employed instead of the MDCT in the enhancement layers in [28]. The proposed codec was designed based on both the objective and subjective quality measure, and the clear improvement was achieved according to the performance evaluation. The subjective test results also showed that the proposed codec outperformed G.729.1 at high bit rates under clean channel condition and had higher robustness to packet loss than G.729.1.

1.4 Organization of the Dissertation

The dissertation is organized as follows. Chapter 2 presents the background materials which are used in the development of the proposed codecs. First, the overview of speech coding is provided. Then the iLBC and the wavelet transform are introduced. The chapter

concludes with the descriptions of performance evaluation methods for speech quality assessment and packet loss robustness assessment.

In Chapter 3, we describe the details of the scalable narrowband speech codec based on the iLBC. The introduction of the multi-rate iLBC is followed by the descriptions of two types of scalable multi-rate codecs based on the iLBC: the first codec using the MDCT and the second codec using the DWT. The performance evaluation is also provided for each codec.

Chapter 4 provides the details of the scalable wideband speech codec based on the iLBC. Three types of bandwidth scalable wideband codecs are presented in the order of performance from lowest to highest. All three codecs adopt split-band structure where the input signal is decomposed into two frequency bands and the lower-band signal is encoded by the scalable narrowband coding scheme based on the iLBC. For the higher-band signal encoding, the first codec uses the scalable narrowband coding scheme based on the iLBC whereas the second and third codecs employ the time-domain bandwidth extension (TDBWE). The difference between the last two codecs is that the second codec uses the MDCT while the third codec uses the WPT in the enhancement layer coding. The performance evaluation is also included for each codec.

Finally, Chapter 5 gives the conclusions and some suggestions for future work.

Chapter 2

Background

2.1 Overview of Speech Coding

The objective of speech coding is to represent a speech signal in digital form with as few bits as possible while maintaining a certain level of speech quality required for the particular application. The performance of speech codecs can be measured by a set of properties. The fundamental codec properties are bit rate, speech quality, robustness to channel errors and packet loss, delay, and computational complexity. Whereas desired properties of a speech codec are low bit rate, high speech quality, high robustness to channel errors and packet loss, low delay, and low computational complexity, there are trade-offs among these properties. Good performance for one of the properties generally leads to lower performance for the others. For example, the iLBC has an advantage over other codecs in terms of robustness to packet loss. This advantage comes at the expense of higher bit rates. When a codec is designed, the desired values for its properties are determined depending on application needs and constraints. A common approach to develop a speech codec is to constrain all properties but one quantitatively. The design objective is then to optimize the remaining property (usually quality or rate) subject to these constraints.

Traditionally, speech codecs have been designed for narrowband speech in which the audio bandwidth is limited to about 300–3400 Hz with a sampling rate of 8 kHz. The increasing deployment of the higher-capacity end-to-end digital networks for both fixed and mobile communications has been allowed the use of wider speech bandwidth, from 50 Hz to 7 kHz, which is called “wideband”. Wideband speech provides substantial improvement in voice quality, especially in naturalness and intelligibility, compared to

narrowband speech. In order to take advantage of the clear benefits from wider speech bandwidth such as the increase in speech quality and the capability to represent speech with arbitrary background sounds including music, recent speech codecs [29], [30] have been developed for so-called super-wideband (50–14000 Hz) and fullband (20–20000 Hz) audio. This section provides the basic speech coding methods with a focus on analysis-by-synthesis coding and subband/transform coding. Subsequently, some of the key technologies to be used for VoIP codecs, especially multi-rate coding and scalable coding, are described.

2.1.1 Speech Coding Methods

Traditionally, speech coding methods have been categorized into three classes: waveform coding, parametric coding, and hybrid coding although the class distinction is not clear. The waveform codec attempts to preserve the original shape of the signal waveform. The waveform coding can be further subdivided into time-domain waveform coding and frequency-domain waveform coding. The well-known time-domain waveform coding techniques are logarithmic pulse code modulation (log PCM) and adaptive differential pulse code modulation (ADPCM) whereas the familiar representatives of the frequency-domain waveform coding methods are subband coding and transform coding. In parametric coding, the speech signal is assumed to be generated from a model and characterized in terms of a set of model parameters. The parametric codec makes no attempt to preserve the original shape of the waveform, and hence SNR is a useless quality measure. The hybrid coding combines the strength of waveform coding with that of parametric coding, and its term is usually reserved for analysis-by-synthesis coding. The analysis-by-synthesis coding technique usually utilizes the LP model and a perceptual distortion measure to reproduce perceptually important characteristics of the input speech. The coding approaches of the recent speech codecs are dominated by analysis-by-synthesis coding and subband/transform coding, which are described in detail next.

2.1.1.1 Analysis-by-Synthesis Coding

Since the early 1980s, advances in speech coding technologies have enabled speech codecs to achieve bit-rate reductions of a factor of 4 to 8 while maintaining roughly the same high speech quality. One of the most important driving forces behind this advancement is the so-called analysis-by-synthesis paradigm [31] for coding the excitation signal of predictive speech codecs. In a speech codec, the speech signal is represented by a combination of parameters. In an open-loop system, the parameters are extracted from the input signal, which are quantized and later used for synthesis. A more effective method is to use the parameters to synthesize the signal during encoding and fine-tune them so as to generate the most accurate reconstruction. Conceptually, this is a closed-loop optimization procedure, where the goal is to choose the best parameters so as to match as much as possible the synthesized speech with the original speech. Since the signal is synthesized during encoding for analysis purposes, the principle is known as analysis-by-synthesis.

The dominant and most successful analysis-by-synthesis technique is code-excited linear prediction (CELP). Figure 2.1 shows a simplified block diagram of the CELP encoder. The excitation signal is generated from a codebook with the corresponding gain and applied to the synthesis filter. The error signal is calculated by subtracting the synthesized signal from the original speech signal and passed through a perceptual weighting filter. The process is repeated for all excitation codevectors stored in a codebook. The excitation codevector and gain which produce the minimum perceptually weighted coding error are selected, and their indexes are transmitted to the decoder. A simplified block diagram of the CELP decoder is illustrated in Figure 2.2. The excitation codevector is retrieved from a codebook identical to that in the encoder and multiplied by the gain to provide the excitation signal. After the synthesized signal is generated, it is processed by the postfilter to enhance the quality of the decoded signal.

The perceptual weighting is a key to improving the subjective quality in CELP. It is used to shape the coding error spectrum so that it follows the spectrum of the input signal to some extent. Due to the noise masking effect of the human auditory system, such spectrally shaped coding error is less audible to human ears.

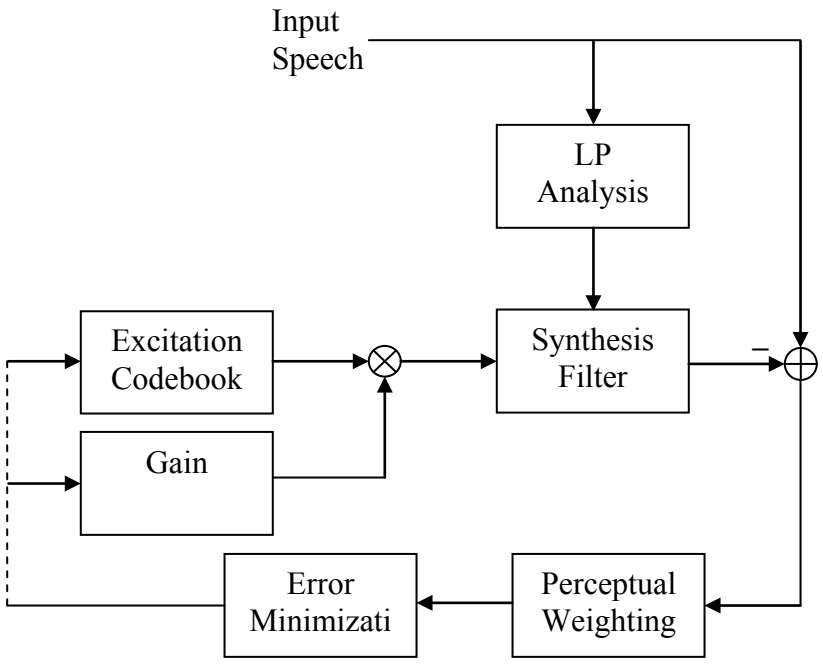


Figure 2.1: Block diagram of a CELP encoder

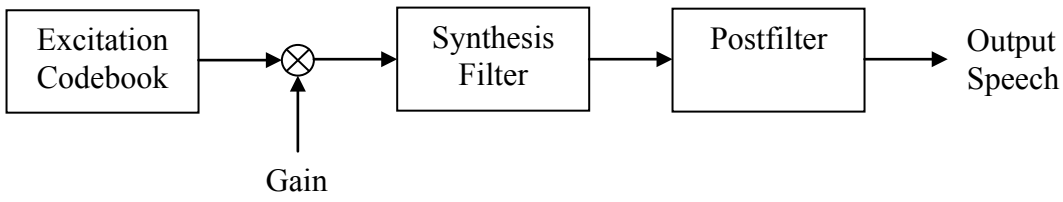


Figure 2.2: Block diagram of a CELP decoder

The CELP codec relies on the long-term and short-term linear prediction models. In the original CELP codec, the long-term synthesis filter was cascaded with the short-term synthesis filter. The long-term synthesis filter, often called the pitch synthesis filter, creates periodicity in the signal associated with the fundamental pitch frequency. The concept of the adaptive codebook was developed to replace the long-term LP analysis in order to improve performance and has become common to many CELP-based standard codecs.

2.1.1.2 Subband/Transform Coding

Whereas subband and transform coding methods have been playing a critical role in high quality audio coding since the early nineties, more recently their use for speech coding, especially for coding of wideband, superwideband, and fullband signal, has gained increasing importance. Although subband coding and the transform coding grew out of different areas using different building blocks, it became clear that they are just different views of the same underlying methodology [32]. In both coding methods, the input signal is split into frequency bands. In one point of view, the term, "subband coding", is usually used if the number of frequency bands is small; otherwise the technique is called "transform coding".

In subband coding, an analysis filter bank is used in the encoder to decompose the input signal into subbands, and each band is encoded separately. In order to preserve the data rate after filtering, each filter output is critically down-sampled. This down-sampling process may introduce aliasing distortion due to the overlapping nature of the subbands. Thus, the success of this technique depends on the design of appropriate analysis and synthesis filter banks. Quadrature mirror filter (QMF) banks allow the aliasing that occurs during filtering and down-sampling at the encoder to be cancelled at the decoder in the absence of quantization errors. The codecs used in each band can be PCM, ADPCM, or even an analysis-by-synthesis method. The advantage of subband coding is that each band can be coded to a different accuracy and that the coding error in each band can be controlled in relation to human perceptual characteristics.

In transform coding, a block of input samples is linearly transformed by a discrete transform into a set of transform coefficients. These coefficients are then quantized and transmitted to the decoder. An inverse transform yields exact reconstruction in the absence of quantization errors. Recent speech codecs employ a modified discrete cosine transform (MDCT) to encode the coding error from the core codec. The MDCT is a lapped transform, where subsequent blocks of samples are overlapped. This overlapping helps to avoid artifacts stemming from the block boundaries. The main attraction of transform coding is that it allows more bits to be allocated to the perceptually important coefficients.

2.1.2 Key Technologies for VoIP Codecs

The emergence of VoIP incurs numerous impairments including delay, jitter, packet loss and decoder clock offset, which degrade the quality of the speech. Advanced signal processing algorithms and coding technologies can combat these impairments and improve the perceived quality of a VoIP conversation.

For example, packet delay or loss can be a major source of impairments in long distance packet switched networks. A jitter buffer is an essential tool to smooth-out the inevitable delay variations caused by the network routers. It is also essential to use packet loss concealment (PLC) algorithms to alleviate the speech quality degradation.

Another approach to improve speech quality in the case of packet loss is to reduce packet loss rate. One way to reduce packet loss rate is to make use of bit-rate adaptation. If the codec has the capability of multi-rate operation, the packet loss rate can be reduced by lowering the bit rate. Thus, the functionality of speech codec that allows its bit rate to adapt to the current available channel capacity is of significant importance. When the encoder adjusts the bit rate, it requires information about the current channel condition, which means that feedback is required.

The instantaneous adaptation to the current channel condition without the need of feedback is the attractive feature of VoIP application due to the requirement of short time delay for real-time communication. Bit-stream scalability is a promising technique that makes it possible to adjust the bit rate to the desired value by truncating the bit stream at any point of a communication system. The packet delay or loss rate can be reduced by lowering the bit rate of voice traffic instantaneously without re-encoding.

2.1.2.1 Multi-Rate Coding

The term, multi-rate coding, here is also called network-controlled multimode coding, where multiple modes of speech coding are defined with each mode having a different fixed bit rate. The multi-rate codec responds to an external control signal to switch the

data rate to one of a predetermined set of possible rates. The control signal is typically generated by the network operating system in response to network conditions and the desired quality of service. For example, the network operating system can switch to using lower bit rates during network congestion to improve capacity and reduce packet delay/loss or to trade off speech bit rate for channel coding to increase channel protection. A special case of multi-rate coding called scalable coding is of particular interest and is explained in detail in the next sub-section.

The typical example of the multi-rate narrowband codecs is the adaptive multi-rate (AMR) speech codec [6] standardized by European Telecommunication Standards Institute (ETSI) in 1999 [33]. The Third Generation Partnership Project (3GPP) adopted the AMR codec as the mandatory speech codec for the third generation WCDMA system [34]. The wideband version of the AMR codec, referred to as AMR Wideband (AMR-WB) [35], was standardized in 2001 [36]. The AMR must be supported in all universal mobile telecommunications system (UMTS) and long term evolution (LTE) terminals, and the AMR-WB is also supported in all wideband-voice (HD voice) capable terminals across UMTS and LTE [37]. A new codec for enhanced voice services (EVS) [38], the successor of the AMR and AMR-WB codecs, was standardized in September 2014 [39], [40]. The EVS codec has been primarily designed for Voice over LTE (VoLTE) and provides not only enhanced voice quality and coding efficiency, but also high robustness to packet loss and delay jitter while maintaining backward compatibility to the AMR-WB codec.

In contrast to network-controlled multimode coding, source-controlled variable bit rate (SC-VBR) coding is used to achieve higher coding efficiency. The SC-VBR codecs select the encoding rate based on the characteristics of each speech frame (e.g., voiced, unvoiced, transient, background noise). The average bit rate is typically less than the bit rate of the fixed rate codec to achieve a given level of speech quality. Note that the SC-VBR coding scheme is usually combined with the network-controlled multimode coding scheme, where the average bit rate is controlled by the network. The SC-VBR codecs have been becoming increasingly common [15]. The EVS codec also provides a SC-VBR mode.

2.1.2.2 Scalable Coding

Scalable speech coding is a special case of multi-rate coding with a bitstream structured into layers which consists of a core bitstream and enhancement bitstreams. When a scalable speech codec is used, the encoder can operate at the highest bit rate, but some enhancement layers could be discarded at any point of communication systems by simply truncating the bitstream to reduce the bit rate. In contrast to a multi-rate codec, a feedback of channel conditions and re-encoding are not required for a scalable codec to reduce the bit rate. Therefore, a layered bitstream offers higher flexibility and easier adaptation to sudden change of network conditions, which can be exploited to reduce packet loss rates. In fact, enhancement layers can be used to add various types of functionality to a core layer, such as speech quality improvement (also called signal-to-noise ratio (SNR) scalability), bandwidth extension or mono to stereo extension (number of channels extension) [41]. Note, however, that a scalable speech codec generally has lower performance than a multi-rate codec with each bit-rate mode independently optimized for the highest speech quality.

There are two other advantages for employing scalable speech codecs [16]. First, scalable coding is a possible solution to cope with the heterogeneity and variability in communication systems. In fact, the telephone industry has been experiencing a transition from the PSTN to an all IP network. Currently, links having different capacities and terminals with various capabilities may coexist. A transmission path may include both wireless links and fixed links with different capacities. Using a scalable coding approach, users can receive different quality versions of the same speech according to their individually available resources and supported capabilities by simple bitstream truncation.

Secondly, the coexistence of the PSTN and IP networks with a mixture of wireless and fixed links means that transcoding at gateways is inevitable. In this situation, the bitstream scalability can be employed to ensure interoperability with different network infrastructures without requiring too much transcoding overhead. More specifically, a

scalable extension of a widely used core codec is a very attractive solution to deploy a new enhanced codec while minimizing the required transcoding overhead and maintaining interoperability and backwards compatibility with existing infrastructure and terminals.

The recent scalable wideband codec examples are ITU-T G.729.1, G.718, and G.711.1 [42]. G.729.1 and G.711.1 are scalable extensions of widely used narrowband codecs, G.729 and G.711, respectively. G.718 provides the interoperability mode with G.722.2/AMR-WB.

2.2 Internet Low Bitrate Codec

The iLBC is a speech codec developed for robust voice communication over IP networks. It is designed for narrowband speech signals sampled at 8 kHz. The codec uses a frame-independent coding algorithm and operates at fixed bit-rates of 13.33 kbps for the frame size of 30 ms and 15.2 kbps for the frame size of 20 ms.

The iLBC algorithm inherently possesses high robustness to packet loss and enables a controlled response to packet losses similar to what is known from pulse code modulation (PCM) with packet loss concealment (PLC), i.e., the ITU-T G.711 which operates at a fixed bit rate of 64 kbps. The codec overcomes the error propagation problem caused by frame dependencies of the CELP-based codecs (e.g. G.729, G.723.1, GSM-EFR and 3GPP-AMR) in the case of packet loss. Therefore, the iLBC codec is especially suitable for VoIP applications such as Skype, Yahoo Messenger, and Google Hangouts among others. Cable Television Laboratories (CableLabs) has adopted the iLBC as a mandatory PacketCable audio codec standard for VoIP over Cable applications.

The iLBC was developed by a company called Global IP Solutions (GIPS) formerly Global IP Sound (acquired by Google Inc. in 2011), and the experimental protocol is defined in request for comments (RFC): 3951 [43] published by the Internet engineering task force (IETF) in December 2004. The RTP payload format for iLBC speech is defined in RFC 3952 [44]. The iLBC was selected by Cable Television Laboratories, Inc. (CableLabs) as a codec standard suitable for packet-based communication networks, and

is supported in PacketCable as one of the recommended codecs [45]. The iLBC is available under a free software/open source license as a part of the open source WebRTC project [46].

The computational complexity of the iLBC is in the same range as the reduced complexity version of the ITU-T G.729 speech codec, i.e., G.729 Annex A. It provides not only significantly higher robustness to packet loss than G.729A, but also equivalent speech quality in a clean channel condition.

2.2.1 Codec Structure

The basic framework of iLBC is based on the linear prediction (LP) model and block-based coding of the LP residual signal using an adaptive CB as is the case with CELP-based codecs. The main difference from CELP-based codecs is that the long-term predictive coding is exploited without introducing inter-frame dependencies. Thus, the propagation of errors across frames is avoided when packets are lost, which makes the iLBC robust to packet loss. This frame independence is achieved by applying the adaptive CB both forward and backward in time, starting from the start state. The start state captures pitch information in voiced speech and accurate noise-like information in unvoiced speech, and enables the operation of the adaptive CB without depending on the history of the LP residual signal.

The benefit of using the start state comes at the expense of a large number of bits required to represent it accurately for each frame. The start state occupies 43.5 % and 56.25 % of encoded bits for 30 ms frame mode and 20 ms frame mode, respectively.

Figure 2.3 shows the block diagram of the iLBC speech encoder. The encoder operates on 20/30 ms input frames, each of which is divided into 5 ms sub-frames. The narrowband input signal is sampled at 8 kHz and pre-processed by a high-pass filter with 90 Hz cut-off frequency. For each frame, the LP analysis is performed and the LP residual signal is calculated. The two consecutive sub-frames of the LP residual signal having the highest weighted energy are identified. Within these two sub-frames, the start state is selected as either the first 57/58 samples or the last 57/58 samples of the two

consecutive sub-frames, depending on which segment has a higher energy. The start state is encoded with scalar quantization. An adaptive CB search procedure is used to first encode the 23/22 remaining samples in the two sub-frames containing the start state. Secondly, the remaining sub-frames after the start state are encoded forward in time, and lastly, the remaining sub-frames before the start state are encoded backward in time. Each adaptive CB search is repeated three times for refinement. The encoded bits are packetized into the payload to be transmitted.

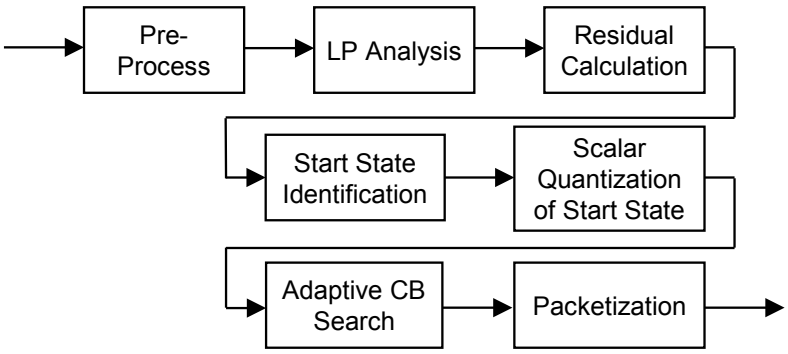


Figure 2.3: Block diagram of the iLBC encoder

The decoder is basically the inverse function of the encoder except for two added functions: enhancement of the LP residual signal and packet loss concealment (PLC). In particular, an enhancement algorithm is applied to the reconstructed LP residual signal. This enhancement augments the periodicity of voiced speech regions. The PLC operation is embedded in the decoder. The PLC operation is based on repeating LP filters and obtaining the LP residual signal by using a long-term prediction estimate from previous residual frames.

2.2.2 Codec Performance

Figure 2.4 shows the formal subjective test results performed by the Dynastat Inc., where the performance of the iLBC was compared with those of the existing coding standards G.729A and G.723.1 under lossy channel conditions. The speech quality scores

(MOS, which is explained in Section 2.4) are plotted as a function of packet loss rates. The results clearly show the iLBC's superiority when used in a real life environment, where its intrinsic packet-loss robust property results in a high quality even under adverse network conditions. We can also see that the iLBC not only performs significantly better than G.729A and G.723.1 under lossy channel conditions but also provides equivalent or higher speech quality in a clean channel condition.

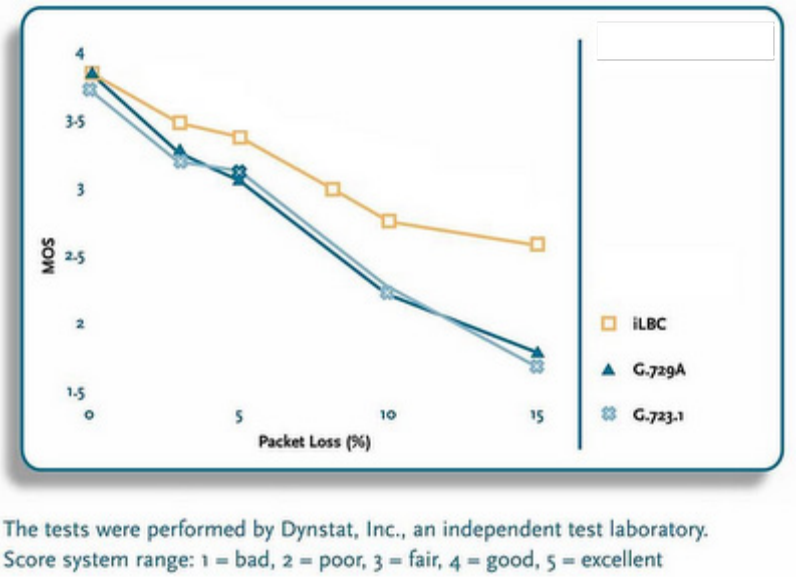


Figure 2.4: Performance comparisons of iLBC, G.729A and G.723.1 under lossy channel conditions [46]

2.3 Wavelet Transform

The Fourier-based transforms such as the Discrete Fourier Transform (DFT), DCT, and MDCT have a problem encoding highly non-stationary signals because the transform of a non-stationary signal spreads over the whole spectrum, which makes compression in the transform domain a more difficult task. In contrast to the Fourier-based transform, which uses a single analysis window, the wavelet transform [47]–[50] uses short windows at high frequencies and long windows at low frequencies. In other words, it provides good frequency selectivity for low frequencies at the cost of the temporal

resolution, and good time localization for high frequencies at the cost of the frequency resolution. Therefore, the wavelet transform can be used to better capture non-stationarities and localized waveforms in the time domain than the Fourier-based transforms.

Wavelets are localized waves. Instead of oscillating forever, they drop to zero. Wavelets are basis functions $w_{jk}(t)$ in continuous time. A basis is a set of linearly independent functions that can be used to produce all admissible functions $f(t)$:

$$f(t) = \sum_{j,k} b_{jk} w_{jk}(t) . \tag{2.1}$$

The special feature of the wavelet basis is that all functions $w_{jk}(t)$ are constructed from a single mother wavelet $w(t)$. This wavelet is a small wave (a pulse). Normally it starts at time $t = 0$ and ends at time $t = N$. A typical wavelet $w_{jk}(t)$ is compressed j times and shifted k times. Its formula is

$$w_{jk}(t) = w\left(2^j t - k\right). \tag{2.2}$$

The remarkable property that is achieved by many wavelets is orthogonality. The wavelets are orthogonal when their inner products are zero:

$$\int_{-\infty}^{\infty} w_{jk}(t) w_{JK}(t) dt = 0 . \tag{2.3}$$

In this case the wavelets form an orthogonal basis for the space of admissible functions. A perfect basis is not only orthogonal but orthonormal, which means that the functions have length 1. The rescaled wavelet that forms an orthonormal basis is

$$w_{jk}(t) = 2^{j/2} w\left(2^j t - k\right). \tag{2.4}$$

Corresponding to the low-pass filter with coefficients $h_0(k)$, there is a continuous-time scaling function $\phi(t)$. Corresponding to the high-pass filter with coefficients $h_1(k)$, there is a wavelet $w(t)$. The dilation equation that produces the scaling function $\phi(t)$ is

$$\phi(t) = 2\sum_{k=0}^{N} h_0(k) \phi(2t - k) . \tag{2.5}$$

The wavelet equation for $w(t)$ is

$$w(t) = 2\sum_{k=0}^{N} h_1(k)\phi(2t-k)\,. \tag{2.6}$$

The wavelet transform operates in continuous time on functions and in discrete time on vectors. The input is a function $f(t)$ or a vector $x(n)$. The output is the set of coefficients b_{jk}, which expresses the input in the wavelet basis. For functions and infinite signals, this basis is necessarily infinite. For finite length vectors with L components, there will be L basis vectors and L coefficients. The discrete wavelet transform, from L components of the signal to L wavelet coefficients, is expressed by an L by L matrix.

In order to calculate the coefficients b_{jk}, integrals of $f(t)$ times $w_{jk}(t)$ are used in continuous time. In discrete time we are solving a linear system. The inverse transform involves the inverse matrix.

For an orthonormal basis, the synthesis and analysis of a function $f(t)$ are

$$\text{Synthesis in continuous time:} \quad f(t) = \sum_{j,k} b_{jk} w_{jk}(t) \tag{2.7}$$

$$\text{Analysis in continuous time:} \quad b_{jk} = \int_{-\infty}^{\infty} f(t) w_{jk}(t)\,dt\,. \tag{2.8}$$

In the matrix case, the wavelets are ordinary vectors. They go into the columns of the wavelet matrix S. Each wavelet vector corresponding to the coefficient b_{jk} has a position in time given by k and a position in frequency (scale) given by j. The columns of the L by L matrix S are the discrete wavelets:

$$\text{Synthesis in discrete time:} \quad x = Sb\,. \tag{2.9}$$

The rows of the L by L matrix A contain the analyzing wavelets:

$$\text{Analysis in discrete time:} \quad b = Ax\,. \tag{2.10}$$

For all orthonormal wavelets, the columns of S are the same as the rows of A. Analysis and synthesis are related by $A = S^T$. In general they are related by $A = S^{-1}$.

Without orthogonality, the rows of $A = S^{-1}$ are biorthogonal to the columns of S:

$$(\text{row } i \text{ of } A) \cdot (\text{column } j \text{ of } S) = \delta(i-j)\,. \tag{2.11}$$

Each row of S^{-1} is orthogonal to $L - 1$ columns of S. This is biorthogonality. The columns of S are the synthesis basis, and the rows of $A = S^{-1}$ are the analysis basis.

By using the word "basis", we ensured that all these matrices are square. In the rectangular case, S would not have an inverse. There are too many columns to be independent; instead of a basis we have a frame. In this case, the pseudo-inverse S^{+} would be used instead of S^{-1}.

The wavelet basis has special properties beyond orthogonality. The scales j and $j - 1$ are closely related. By taking advantage of this relation, the multiplications by A and S can be reorganized and the fast transforms are derived, which is described next.

The recursive nature of wavelets can be seen when we construct a tree of filter banks as shown in Figure 2.5. Wavelets come from the iteration of filters (with rescaling). The link between discrete-time filters and continuous-time wavelets is in the limit of a logarithmic filter tree. The coefficients b_{jk} are obtained by processing the input signal by the high-pass filter H_1 and down-sampling by a factor of 2. These coefficients are at the end of the branch in this logarithmic tree and at the finest level. The coefficients a_{jk} are obtained by processing the input signal by the low-pass filter H_0 and down-sampling by a factor of 2. These coefficients are filtered again by H_1 and H_0. Whereas the low-frequency coefficients a_{jk} give the approximations of the signal, the high-frequency coefficients b_{jk} provide the details of the signal. This signal decomposition is called the discrete wavelet transform (DWT) and the DWT coefficients are formed by the high-frequency coefficients of each level together with the low-frequency coefficients of the last level. Because of the decimation by 2 at each level, the DWT of an L dimensional signal will produce L transform coefficients.

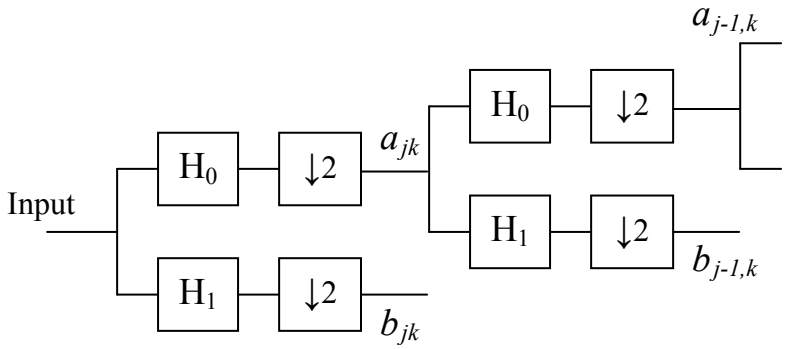


Figure 2.5: The logarithmic tree of the discrete wavelet transform

The tree of filters leads to the fast wavelet transform. The analysis matrix A can be expressed as a product of very sparse matrices which correspond to the filter bank at each level of the logarithmic tree. Since at each level, only the low-pass signal from the previous level is further decomposed, the computational complexity is significantly reduced. In fact, the complexity is linear in the number of input samples, L, independent of the depth of the tree. Therefore, the DWT is asymptotically faster than the fast Fourier transform (FFT), requiring only $O(L)$ steps instead of $O(L \cdot \log L)$.

In continuous time, the input $f(t)$ is a function instead of a vector. The output is the set of coefficients b_{jk} that multiply the wavelet basis functions $w_{jk}(t)$. These coefficients are inner products of $f(t)$ with $w_{jk}(t)$. Here we can see the beautiful connection between wavelets and filter banks. The coefficients at level $j-1$ come directly from the coefficients at level j as in discrete time by filtering and down-sampling. This is because the functions at level $j-1$ come from the functions at level j. Furthermore, a function at fine resolution j is equal to a combination of "approximation plus detail" at coarse resolution $j-1$:

$$\sum_k a_{jk}\phi_{jk}(t) = \sum_k a_{j-1,k}\phi_{j-1,k}(t) + \sum_k b_{j-1,k}w_{j-1,k}(t)\,. \tag{2.12}$$

In the DWT, only the low-pass filter is iterated. It is assumed that lower frequencies contain more important information than higher frequencies. For many signals this is necessarily not true. A wavelet packet [48]–[52] basis allows any dyadic tree structure as

shown in Figure 2.6. At each point in the tree we have the option to send the signal through the lowpass-highpass filter bank, or not.

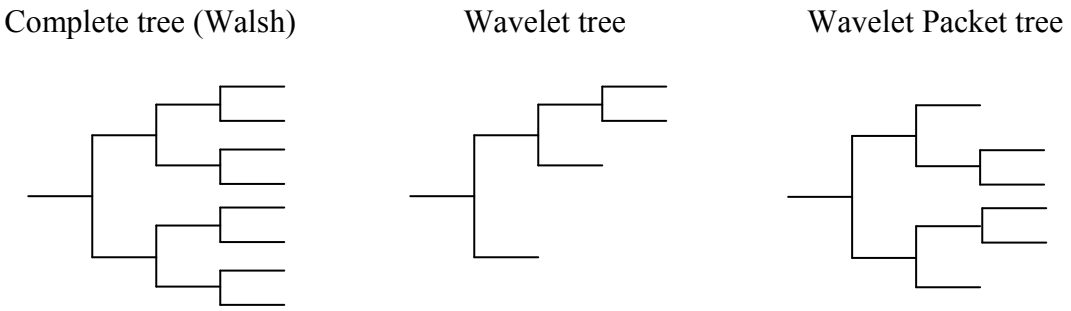


Figure 2.6: Tree structures for time-frequency decomposition

One possibility is the logarithmic tree for the DWT, with low-pass iteration only. Another possibility is the complete tree for the Walsh-Hadamard transform, which is analogous to the short time Fourier transform (STFT). Wavelet packets make up the entire family of bases. The decision to split or to merge should be aimed at achieving minimum distortion subject to constraints on the bit rate and the delay.

2.4 Performance Evaluation Methods

2.4.1 Speech Quality Assessment

Evaluation methods generally fall into two types: subjective tests and objective tests. Subjective tests involve human listeners and are difficult to organize and perform. Thus, objective evaluation methods were developed to model subjective tests for automated assessment of the speech quality.

Absolute category rating (ACR) tests are the most common type of subjective tests of speech quality. In this test, listeners rate for each speech utterance according to the scales, such as, Excellent (5), Good (4), Fair (3), Poor (2), and Bad (1). The mean opinion score (MOS) [14], [53] is the arithmetic mean of all the individual scores. Comparison category

rating (CCR) method, also known as A-B comparison tests, is also commonly used. In this test, listeners are presented with a pair of speech samples on each trial, and use the following scale to judge the quality of the second sample relative to that of the first: Much better (3), Better (2), Slightly Better (1), About the Same (0), and Slightly Worse (-1), Worse (-2), Much Worse (-3).

The perceptual evaluation of speech quality (PESQ) [14], [54] algorithm was developed to provide an objective assessment of speech quality in conversational voice communications. The PESQ is used to predict the subjective quality of speech and generate the objective score, MOS-LQO, which can be used for a linear comparison with MOS scores for both narrowband and wideband speech. Note that whereas a MOS score ranges from 1 to 5, the range of a MOS-LQO is from 1.02 to 4.55 for narrowband speech and from 1.04 to 4.64 for wideband speech. While there is no substitute for actual listening tests, the PESQ are widely used for initial codec evaluations and are highly useful. The objective tests based on perceptual objective listening quality assessment (POLQA) [55] algorithm, which is a successor to the PESQ, may present better correlation with the subjective tests; however, it is important to note that an objective test will never be a replacement for a subjective listening test.

The speech samples utilized for performance evaluation are from database in Annex B of ITU-T P.501 [56] pre-published in January 2012. The source speech was down-sampled to 8 kHz or 16 kHz depending on the bandwidth of speech signals that a codec supports and its speech level was equalized to -26 dBov using the ITU-T Software Tool Library [57]. The modified-IRS filter and any mask were not used because the target VoIP applications includes soft phones.

2.4.2 Channel Model Used for the Evaluation of Packet Loss Robustness

The Gilbert Elliot channel model [58] was employed using the ITU-T Software Tool Library [57] to simulate the bursty packet loss such as the behavior of IP networks. This channel model has two states: Good (*G*) and Bad or Burst (*B*). The probabilities

associated with the channel states are P and Q, P being the probability of transition from state G to B, and Q the probability of transition from B to G. Thus, the probability of remaining in the same state is $(1 - P)$ and $(1 - Q)$ for state G and B, respectively. When the packet error probability is P_G and P_B for state G and B, respectively, the average packet error probability, PER, generated by this channel model is given by

$$PER = \frac{P}{1-\gamma} P_B + \frac{Q}{1-\gamma} P_G \tag{2.13}$$

where

$$\gamma = 1 - (P + Q) \tag{2.14}$$

is a measure for the correlation of the packet errors, and consequently an indication of the burst or random characteristic of the channel. In this case, $\gamma \approx 0$ implies a nearly random error channel, while $\gamma \approx 1$ implies a totally bursty channel. From (2.13) and (2.14) we get

$$P = (1-\gamma) \cdot \left(1 - \frac{P_B - PER}{P_B - P_G}\right) \tag{2.15}$$

$$Q = (1-\gamma) \cdot \frac{P_B - PER}{P_B - P_G} \tag{2.16}$$

In this work, $P_G = 0$, $P_B = 0.5$ and $\gamma = 0.2$ were used. For the packet loss simulation, four different patterns of packet error sequences were generated for each PER and each speech sample, and the mean value was calculated to evaluate the speech quality for the performance comparison.

Chapter 3

Scalable Narrowband Speech Codec Based on iLBC

The scalable narrowband speech codec based on the iLBC coding scheme is developed in two steps: the addition of rate flexibility to the iLBC and the addition of scalability to the multi-rate codec based on the iLBC. The first step results in the development of the multi-rate iLBC, which is used as the core layer codec of the scalable codec developed in the second step. In this chapter, the multi-rate iLBC is introduced first. Two types of scalable multi-rate codecs based on the iLBC are subsequently described: the first codec using the MDCT and the second codec using the DWT in the enhancement layer. The performance evaluation is also provided for each codec.

3.1 Multi-Rate iLBC

The multi-rate operation of the iLBC is enabled by transforming the start state into the DCT domain and allocating different number of bits to the DCT coefficients. We refer to this narrowband multi-rate iLBC as the proposed codec N1 in this dissertation. The details of the multi-rate iLBC coding scheme are described in this section.

3.1.1 Start State Coding using the DCT

Since the start state contains the important information as explained in Section 162.2.1, the encoding process should maintain its waveform as accurately as possible. The original iLBC uses 3-bit scalar quantizer. However, a time domain waveform coding

is not flexible in terms of the bit rate reduction. A frequency domain coding technique has potential for reducing the bit rate because of the nature of the start state. The DCT is used since it has a strong energy compaction property, a fast transform algorithm is available and the start state is completely independent for each frame.

Figure 3.1 shows the block diagram of the start state encoder using the DCT, which replaces the block for scalar quantization of the start state in Figure 2.3. The N samples of the start state $x_0, \ldots, x_k$ are processed by perceptual weighting filter

$$W(z) = \frac{1}{\hat{A}\left(\frac{z}{\gamma_s}\right)} \tag{3.1}$$

where $\hat{A}(z)$ is a LP analysis filter and the filter $W(z)$ models the short-term frequency masking curve. The parameter γ_s is used to adjust the degree in which the perceptual weighting is applied. Note that the start state is in the residual domain and weighting the start state with $W(z)$ is equivalent to employing a perceptual weighting filter $\hat{A}(z)/\hat{A}(z/\gamma_s)$ in speech signal domain as used in CELP technique. The filter $W(z)$ is initialized to zero in each frame. Note also that N for the original iLBC is 57 and 58 for the 20 ms and 30 ms frame, respectively, whereas the proposed codec N1 uses the longer start state length of $N = 80$. The reason of its choice is explained in Section 3.1.2. The weighted start state samples are transformed into the DCT coefficients $X_0, \ldots, X_k$ by one-dimensional DCT according to

$$X_k = w_k \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right) \qquad k = 0,\ldots,N-1 \tag{3.2}$$

where

$$w_k = \begin{cases} 1/\sqrt{N} & k = 0 \\ \sqrt{2/N} & 1 \le k \le N-1 \end{cases} \tag{3.3}$$

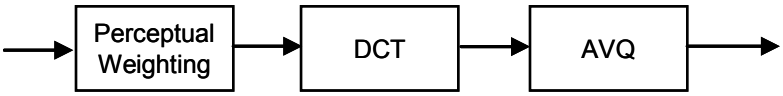


Figure 3.1: Block diagram of DCT-based start state encoder

The DCT coefficients are quantized by the scalable AVQ which is specified in G.718 [11], [59] and implemented by forming 8-dimensional vectors and using multi-rate lattice vector quantizer [60]. Codebooks of different bit rates are formed by selecting subsets of appropriate size from the RE8 lattice [61] points. Any point in a given lattice can be generated from appropriate permutation of the components of a few basic vectors called leaders. Thus, significant savings of memory requirements can be achieved. To remain within the total bit budget, DCT coefficients are divided by a global gain prior to quantization.

The multi-rate functionality is obtained by allocating different number of bits to the AVQ. Thus, when a small number of bits are available to use for the AVQ, those bits are allocated to only a limited number of sub-bands.

3.1.2 Performance Enhancement Schemes

The low bit rate operation is achieved by decreasing the number of available bits for the AVQ, which leads to rapid degradation of speech quality. Some of the schemes were already introduced in [20] to improve performance at low bit rates. One of the schemes is used in the proposed codec N1, which is to reduce the number of adaptive CB stages and reallocate bits from one or two of the adaptive CB refinement stages to start state encoding. This scheme sacrifices speech quality at high bit rates in order to achieve good speech quality at low bit rates.

Longer start state samples can capture more information and provide better frequency resolutions. Thus, high speech quality can be achieved at high bit rates. Interestingly, good speech quality can also be maintained at low bit rates by reducing the number of the adaptive CB stages and reallocating a part of bits to encode the longer start state.

Especially when the length of the start state is 80, extra 34 bits can be saved since the first target of sub-frame to be encoded using the adaptive CB in the original iLBC is completely included in the start state. Therefore, the proposed codec N1 uses the start state length of 80.

Since the start state is encoded in frequency domain, selecting only the most important information is possible and beneficial when only the limited number of bits is available. The energy of speech usually concentrates more on the lower side of frequency bands and when the start state is the part of the voiced speech, the pitch period information is included in the low frequency coefficients. The AVQ is designed to take advantage of the perceptual importance by allocating more bits to the high-energy frequency bands. However, when the available number of bits is limited, the bandwidth of the encoded speech can fluctuate for different frames, which results in annoying synthetic speech. The easiest way to resolve this issue is to select only lower frequency coefficients for the AVQ and discard higher frequency information. Note that when this scheme is used, the care has to be taken to keep good speech quality because the speech bandwidth is reduced. For a narrowband codec, the DCT coefficients corresponding to at least up to 3400 Hz should be encoded in order to avoid muffled sound.

3.1.3 Computational Complexity

The computational complexity of the DCT is $O(N \cdot \log N)$ where N is the number of DCT coefficients whereas the worst-case complexity of the AVQ specified in G.718 is $O(K \cdot \max(\log K, V))$ where K is the number of 8-dimentional DCT coefficient vectors to be quantized, and V is the Voronoi extension order which is used to extend the lattice codebook. Since $K = N/8$ and V is expected to be less than $\log N$, the computational complexity of the start state encoder is $O(N \cdot \log N)$. The start state encoder of the original iLBC has the complexity of $O(N)$ and takes only 6 % and 10 % of total computational load for 30 ms and 20 ms frame mode, respectively according to [62]. Note that when coding only low-frequency DCT coefficients, the implementation complexity of the AVQ

becomes even lower. Therefore, the start state encoder using the DCT is of reasonable complexity.

3.1.4 Packet Loss Concealment (PLC) Algorithm

The PLC algorithm used in the original iLBC is informative only. In order to improve performance under lossy channel conditions, the proposed codec N1 employed the PLC algorithm used in G.729.1. The PLC algorithm of G.729.1 was modified so that it works for the proposed codec N1 in LP residual domain. In particular, some parameters which are not available in the decoder of the proposed codec N1 are estimated. The improvement of the subjective speech quality using this new PLC algorithm in lossy channel conditions was confirmed compared to the PLC algorithm used in the original iLBC.

3.1.5 Objective Performance Evaluation

The objective tests based on the PESQ algorithm were performed. The purpose of presenting the objective test results is to show the effect of using different parameter settings and to present and discuss about the discrepancy between objective and subjective quality test results under certain conditions. It was found that the objective MOS-LQO score based on PESQ algorithm does not necessarily correlate with the subjective MOS score. Thus, the important design considerations for the proposed codec N1 which we need to take into account when utilizing objective speech quality measures are also discussed.

Figure 3.2 shows the MOS-LQO scores of the multi-rate iLBC as a function of bit rates to evaluate the effect of coding only low-frequency DCT coefficients and the effect of using different number of refinement stages for adaptive CB search process. First, the performance of three different upper frequency limits of 4.0 kHz, 3.2 kHz, and 2.4 kHz for start state encoding is compared. The trend that can be seen from the objective test results is that higher scores are obtained at lower bit rates when the coefficients are

limited to lower frequency region. However, as briefly pointed out in Section 3.1.2, the subjective quality quickly goes down when the frequency content is limited to below 3.4 kHz. The bandwidth of the start state affects the quality of the decoded signal because the LP residual signal is encoded based solely on the start state. The obvious mismatch between subjective quality and objective quality measures can be seen here. Thus, all the DCT coefficients should be used for the AVQ, which handles the efficient bit allocation by using a global gain, unless the number of available bits for the AVQ is relatively small. When only the small number of bits is available, the coefficients corresponding to at least up to 3.4 kHz should be used for the AVQ. Secondly, the significant performance difference by using different number of adaptive CB refinement stages is observed in Figure 3.2. When the bit rate is decreased while keeping the same number of adaptive CB refinement stages, speech quality goes down quickly. The relatively good performance can be maintained by reducing the number of adaptive CB refinement stages and reallocating the part of the bits from adaptive CB refinement stages to the start state encoding. The same trend can be seen from the subjective quality tests, thus this is the effective method to achieve higher performance at low bit rates.

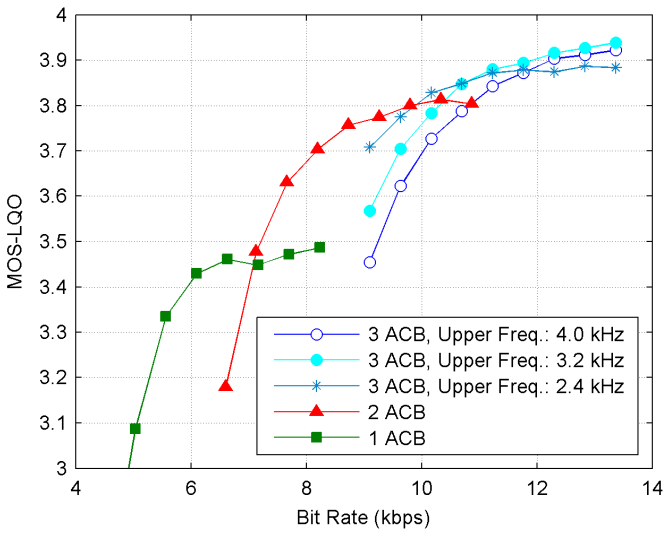


Figure 3.2: Effect of coding only low-frequency DCT coefficients and using the different number of adaptive codebook refinement stages.

Figure 3.3 shows the effect of using the different frame length. When all the other parameters are fixed to the same values for 20 ms and 30 ms mode, the performance of 20 ms frame length is higher than that of 30 ms frame length at the bit rate higher than about 11.2 kbps. The MOS-LQO curve of 30 ms mode is saturated at high bit rates while the 20 ms mode can achieve much higher performance. Since the length of the start state is 80 for the proposed codec N1, the start state occupies half of a frame for the case of 20 ms mode whereas it occupies only one third of a frame for the case of 30 ms mode. When enough bits are allocated to the start state, more accurate encoding of the original signal is possible for 20 ms case. On the other hand, if the other parameter changes are allowed such as an increase of the number of adaptive CB refinement stages, the higher performance can be achieved for 30 ms mode as you can see in Figure 3.3. A similar trend is confirmed from the subjective quality tests although it is not as obvious as Figure 3.3 indicates.

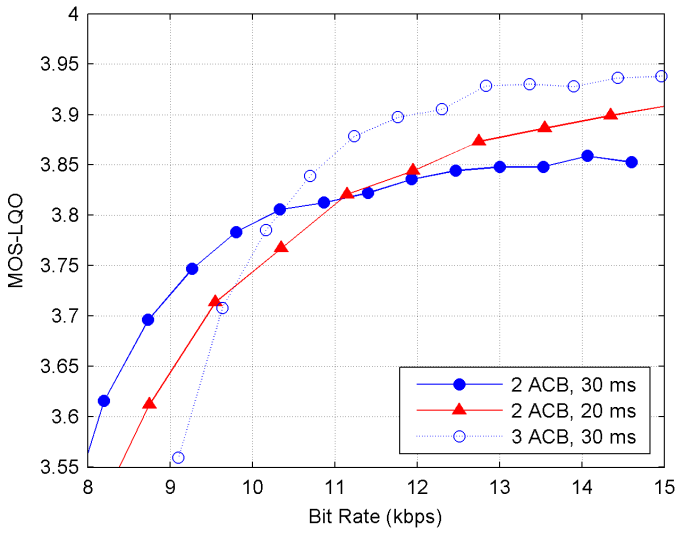


Figure 3.3: Effect of using the different frame length

3.2 Scalable Multi-Rate Codec Using the MDCT

The scalable multi-rate speech codec using the MDCT, which we call the proposed codec N2 in this dissertation, is presented in this section. The proposed codec N2 is developed by adding the scalability to the multi-rate iLBC. In particular, the multi-rate iLBC coding error is encoded by employing the MDCT and the AVQ in the enhancement layer.

3.2.1 Codec Structure

Figure 3.4 shows the block diagram of our proposed N2 encoder. The input speech signal is encoded by multi-rate iLBC encoder first. The bit-stream produced constitutes the core layer portion of the scalable bit-stream. The decoded speech signal is obtained during the iLBC encoding process. The multi-rate iLBC coding error is computed by subtracting the decoded speech signal from the original speech signal and processed by perceptual weighting filter $W_e = \hat{A}(z/\gamma_e)$ where $\hat{A}(z)$ is a LP analysis filter the parameter γ_e is used to adjust the degree in which the perceptual weighting is applied. This weighting filter is used to flatten MDCT coefficients as employed in G.729.1 and G.718. The weighted error signal is windowed and transformed into MDCT coefficients. Figure 3.5 shows the power-complementary window for 20 ms frame mode. For the overlap region, the Kaiser-Bessel derived (KBD) window is employed. To reduce the delay, the overlap is only 40 samples which correspond to 5 ms while the window size is 320 samples which is twice the frame size as shown in Figure 3.5. The effective overlap can be reduced by padding zeros on each side and the perfect reconstruction is still achieved as long as the window function satisfies the Princen-Bradley condition [63]. The overall algorithmic delay for 20 ms and 30 ms frame mode is therefore 25 ms and 35 ms, respectively. The resulting MDCT coefficients are quantized using the AVQ and the enhancement layer bit-stream is produced.

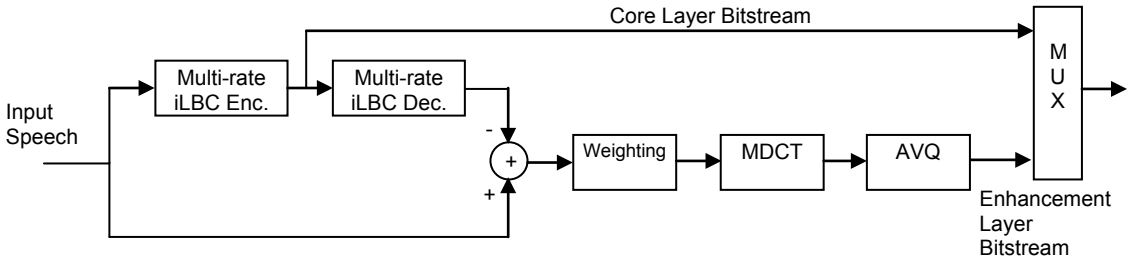


Figure 3.4: Block diagram of the proposed N2 encoder

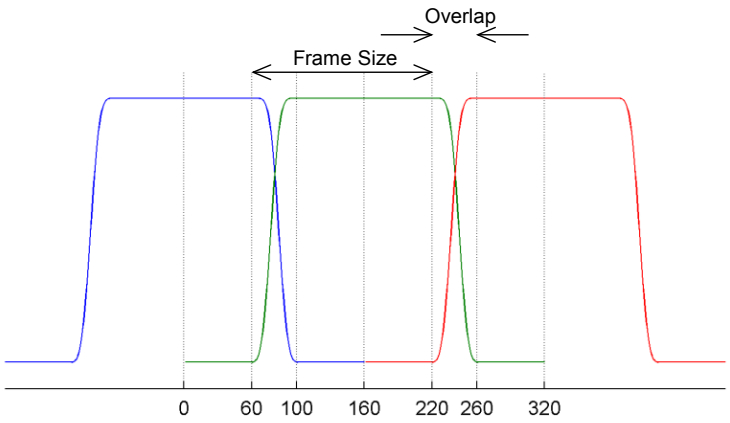


Figure 3.5: Window function with reduced overlap for 20 ms frame mode. KBD window is used for overlap region.

The block diagram of the proposed N2 decoder is shown in Figure 3.6. The AVQ parameters of enhancement layer are decoded, transformed into time domain signal using inverse MDCT (IMDCT), and the weighted overlap-and-add (WOLA) synthesis is performed to obtain the perceptually weighted error signal. The weighted error signal is inverse-weighted and processed by the pre-echo reduction module which performs the same algorithm used in [11] to obtain the decoded error signal. The decoded speech signal of the core layer is combined with the error signal decoded from the enhancement layer. The enhanced speech signal is passed through the post-filter to produce the output speech signal. The post-filter used in G.729.1 was modified to be incorporated in the proposed N2 decoder by employing open-loop pitch estimation for the integer part of the pitch delay.

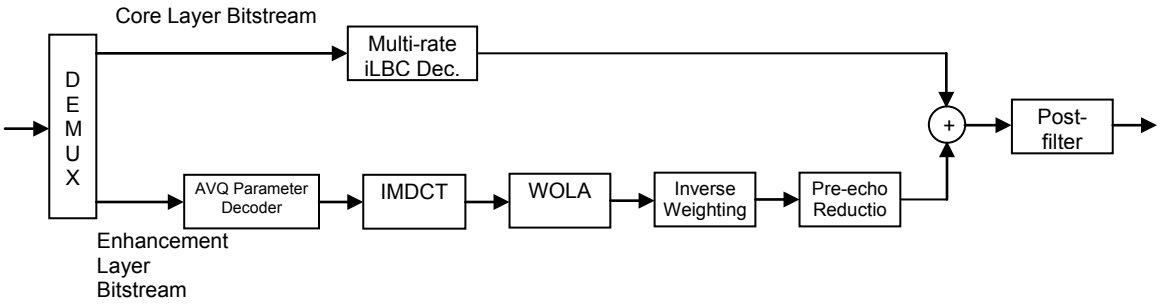


Figure 3.6: Block diagram of the proposed N2 decoder

The post-processing unit used to enhance LP residual signals in the original iLBC was modified to be employed without adding any delay, and is employed in the multi-rate iLBC decoder to achieve higher speech quality. Note that the post-processing unit needs to be included in the decoding process in the encoder as well. The post-processing unit in the original iLBC introduces 5 ms and 10 ms delay for 20 ms and 30 ms frame case, respectively in order to achieve high performance and is employed in the multi-rate iLBC when enhancement layer is not used. Therefore, the overall delay for the multi-rate iLBC without enhancement layer is 40 ms when the frame length is 30 ms, whereas the overall delay for the proposed codec N2 with enhancement layer is 35 ms. When the frame length is 20 ms, the overall delay is 25 ms for the proposed codec N2 with or without enhancement layer.

This scalable structure performs better than the structure introduced in [23] because the enhancement layer is added in speech signal domain instead of the LP residual domain. All the errors resulted from the multi-rate iLBC encoding are handled by the enhancement layer. Since the waveform matching in the adaptive CB stages for the multi-rate iLBC is performed in LP residual domain, the error in speech signal domain should still include useful information.

3.2.2 Performance Evaluation

In order to evaluate the quality of speech produced by our proposed codec N2, the two types of subjective listening tests: mean opinion score (MOS) tests and A-B

comparison tests were conducted. The objective tests based on the PESQ algorithm were also performed and their results are presented first.

Figure 3.7 shows the MOS-LQO scores of the proposed codec N2 as a function of bit rates to evaluate the effect of using the enhancement layer and the effect of coding only low-frequency MDCT coefficients in the enhancement layer. The lowest performance curve at 13.33 kbps or higher is the highest performance that the core layer codec can achieve. From this objective test results, it is seen that the enhancement layer can be used to improve the performance of the core layer even at around 12.5 kbps. However, the highest subjective quality achievable at 12.5 kbps using the current design of our proposed scalable codec N2 is slightly lower than the non-scalable version. This is another evidence of discrepancy between objective and subjective quality measures. It was found that this discrepancy is mainly caused by the quality degradation due to the modification of the post-processing unit used in LP residual domain of the core layer decoder. However, it is confirmed that the allocation of a relatively small number of bits to the enhancement layer is enough to mitigate this quality degradation. We can also observe that the method of coding only low-frequency MDCT coefficients seems to be an effective way to improve the performance since the highest performance is achieved by using the upper frequency limit of 1067 Hz. Note, however, that the performance saturation needs to be taken into consideration when coding only very low-frequency MDCT coefficients as can be seen from the upper frequency limit of 800 Hz. A similar trend was confirmed by the subjective tests especially when the available number of bits is limited to a small amount; however, the performance difference between different upper frequency limits seems to be getting smaller when a relatively large number of bits are allocated to the enhancement layer. When enough bits are available, using a higher upper frequency limit seems to lead to higher subjective speech quality.

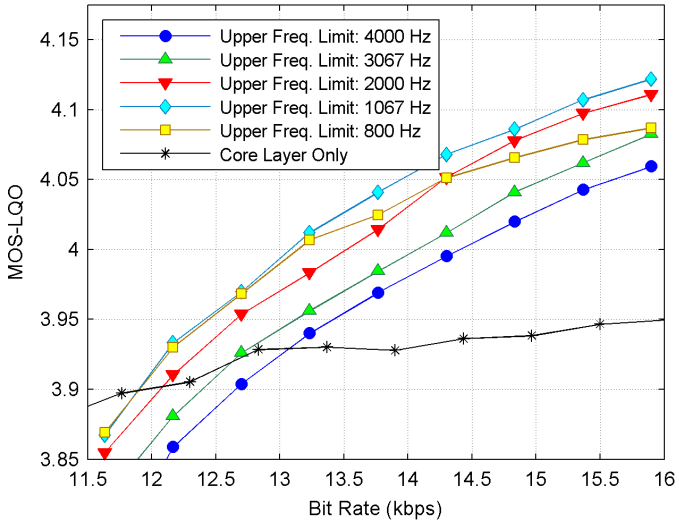


Figure 3.7: Effect of using the enhancement layer and coding only low-frequency MDCT coefficients in enhancement layer

Furthermore, the discrepancy between subjective quality and objective quality was also observed in a certain case for the performance of PLC algorithms under lossy channel conditions. The objective quality based on the POLQA algorithm, which is a successor to PESQ, may present better correlation with the subjective quality; however, from these experiments, it is important to note that an objective test based on the PESQ algorithm will never be a replacement for a subjective listening test.

Informal MOS tests (Absolute Category Rating method) and A-B tests (Comparison Category Rating method) were performed using binaural headphones. Naïve listeners were used. The test samples consisted of four sentence pairs spoken by four different speakers which include 2 male and 2 female speakers. Figure 3.8 shows the MOS scores as a function of bit rate to compare the performance of the proposed codec N2 with core layer only and with core layer plus enhancement layer, the original iLBC, and G.718. Both 30 ms and 20 ms frame length are employed for the performance comparison. The error bars represent the 95 % confidence intervals. These subjective test results are represented only to show the trend of the subjective quality of speech because the MOS scores are obtained from limited informal tests and the 95 % confidence intervals are large. Two adaptive CB refinement stages were used for all the test cases except for the core layer only codec at 11.77 kbps with 30 ms mode which used three stages. As typical

examples, bit allocations for the proposed codec N2 using 20 ms frame length with core layer only operated at 11.95 kbps and with core layer plus enhancement layer operated at 14.7 kbps are provided in Table 3.1 and Table 3.2, respectively. According to the subjective test results, the performance of the proposed codec N2 is similar to G.718 at around 12 kbps, and the subjective quality of the proposed codec N2 drops rapidly at 8 kbps. The performance of 30 ms mode is higher than that of 20 ms mode when operated at the same bit rates because the more efficient encoding is possible using longer frame length especially at low bit rates at the cost of a longer delay. The proposed codec N2 operated at about 12 kbps using the 30 ms and 20 ms mode achieves similar performance to the original iLBC with the corresponding frame length. The proposed scalable codec N2 operated at 12.7 kbps using 30 ms mode includes the core layer codec operable at 10.3 kbps. Using this parameter setting, the performance of the proposed scalable codec N2 is slightly higher than the core layer codec and equivalent to the non-scalable codec operated at 11.77 kbps. The proposed scalable codec N2 operated at 14.7 kbps with 20 ms mode also outperforms the core layer codec operable at 11.95 kbps embedded in it. Overall, the performance of the proposed codec N2 is good at the bit rate higher than 10 kbps. The speech quality differences may not be as obvious as seen in the Figure 3.8. Note that if an extra delay is allowed for the post-processing unit used in LP residual domain of the core layer decoder, the performance of the proposed codec N2 with core layer plus enhancement layer can be increased.

Table 3.1: Bit Allocation for the proposed codec N2 with core layer only when operating at 11.95 kbps using 20 ms mode

Parameter	Bits
LSF	20
Position of Start State	2
DCT global gain for Start State	7
DCT spectral parameters for Start State	160
Adaptive CB index	31
Adaptive CB Gain	18
Empty Frame Indicator	1
Total	239

Table 3.2: Bit Allocation for the proposed codec N2 with core layer plus enhancement layer when operating at 14.7 kbps using 20 ms mode

Parameter	Bits
LSF	20
Position of Start State	2
DCT global gain for Start State	7
DCT spectral parameters for Start State	160
Adaptive CB index	31
Adaptive CB Gain	18
MDCT global gain for Enh. Layer	7
MDCT spectral parameters for Enh. Layer	48
Empty Frame Indicator	1
Total	294

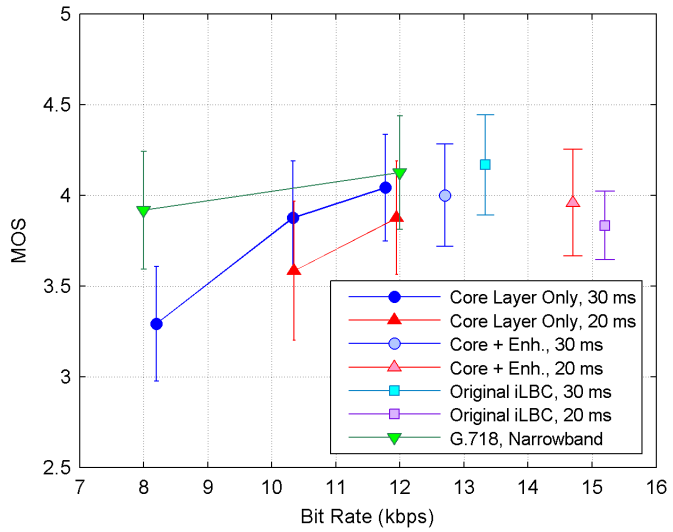


Figure 3.8: Performance comparison of the proposed codec N2 with the original iLBC and G.718

Figure 3.9 shows the results of the A-B tests where the proposed codec N2 operated at 11.95 kbps using 20 ms mode is compared with G.718 operated at 12 kbps for different packet loss rates. The performance of the proposed codec N2 using 20 ms mode is equivalent to that of G.718 under the clean channel condition and at all packet loss rates. Note, however, that the speech quality difference between the proposed codec N2 using 20 ms mode at 11.95 kbps and G.718 at 12 kbps can be observed in Figure 3.8. The performance difference between Figure 3.8 and Figure 3.9 may result from limited

informal subjective tests. From the subjective listening, the decoded speech of the proposed codec N2 at 11.95 kbps sounds more natural and thicker than G.718. It seems that reproduction of lower frequency range is more accurate with the proposed codec N2. On the other hand, G.718 produces obviously clearer sound and the objective test suggests that the performance of G.718 is better. According to the subjective speech quality test results, each listener tends to prefer the specific codec over the other under both clean and lossy channel conditions. Thus, the listener's preference may have affected the subjective test results in Figure 3.9. The PLC performance of the proposed codec N2 may be better than G.718 because the decoded speech of the proposed codec N2 under packet loss conditions seems to contain less artificial sound. Figure 3.10 shows the results of the A-B tests where the proposed codec N2 at 11.95 kbps is compared with the original iLBC both using 20 ms frame length for different packet loss rates. It is easily observed that the proposed codec N2 has higher robustness to packet loss than the original iLBC due to the better PLC algorithm. Note that the proposed codec N2 is operated at 11.95 kbps which is 3.25 kbps lower than the bit rate of the original iLBC. This shows the significant improvement of the proposed codec N2 over the original iLBC.

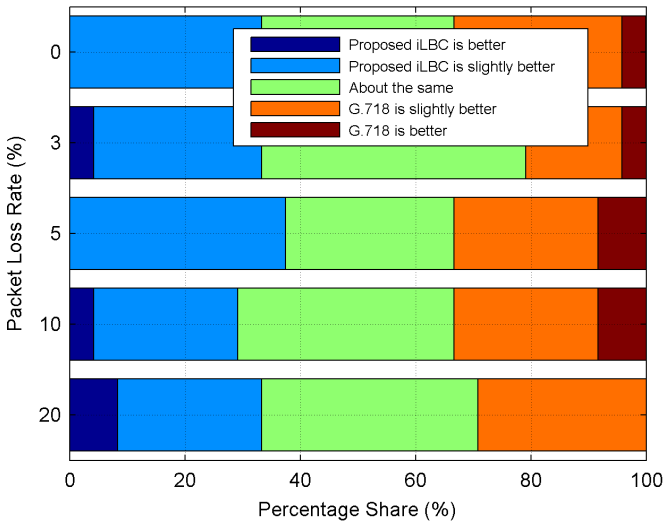


Figure 3.9: A-B comparison test results for the proposed codec N2 using 20 ms frame at 11.95 kbps vs G.718 at 12 kbps

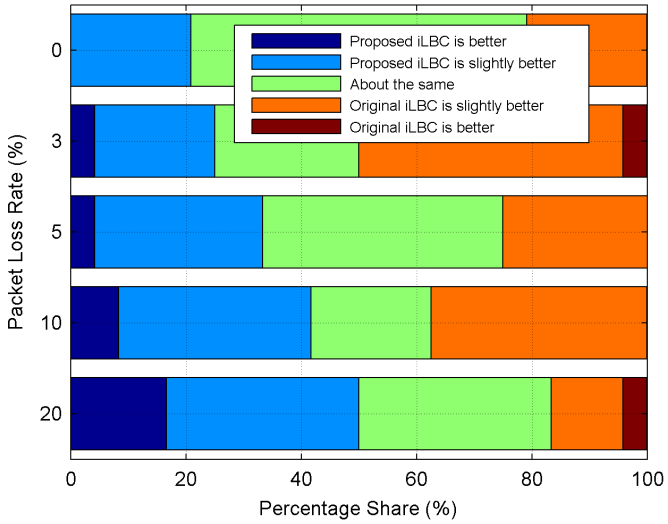


Figure 3.10: A-B comparison test results for the proposed codec N2 using 20 ms frame at 11.95 kbps vs the original iLBC at 15.2 kbps

The original iLBC structure has its own limitation in terms of performance because encoding is performed without information from the previous frame and the waveform matching during the adaptive CB search is conducted in LP residual domain. The proposed scalable structure removes the speech quality limitation by coding the iLBC coding error in speech signal domain. On the other hand, a CELP-based codec takes advantage of both long-term redundancies across frame boundaries and the waveform matching in signal domain. Therefore, the proposed scalable structure gives much more significant benefit to the core layer codec based on the iLBC. Furthermore, the improvement of packet-loss robustness may be possible because the current PLC algorithm is not fully optimized yet.

3.3 Scalable Multi-Rate Codec Using the DWT

In this section, the scalable multi-rate speech codec using the DWT, which is referred to as the proposed codec N3 in this dissertation, is introduced. The scalable multi-rate codec using the MDCT was developed in Section 3.2. Whereas the performance of the

core-layer codec was satisfactory, the speech quality improvement by adding the enhancement layer was limited. In order to improve performance from the addition of the enhancement layer, we propose the use of the DWT instead of the MDCT to encode the core-layer coding error in the enhancement layer.

The MDCT uses a single analysis window, which results in a fixed frequency resolution. When the window length is increased to achieve better frequency resolution, the time resolution becomes poor. On the other hand, the DWT uses short windows at high frequencies and long windows at low frequencies, which is better suited to encode highly non-stationary signals such as the core-layer coding error.

3.3.1 Codec Structure

The block diagram of the proposed N3 encoder is illustrated in Figure 3.11. The input speech signal is first encoded by multi-rate iLBC encoder presented in Section 3.1 and the core layer portion of the scalable bit-stream is produced. The locally decoded speech signal is subtracted from the input speech signal to calculate the multi-rate iLBC coding error. The error signal is processed by perceptual weighting filter $W_e = \hat{A}(z/\gamma_e)$ where $\hat{A}(z)$ is a LP analysis filter and the parameter γ_e is a constant which determines the degree to which the perceptual weighting is applied. This weighting filter is used to flatten DWT coefficients. The weighted error signal is decomposed in a number of sub-bands into wavelet coefficients. The resulting DWT coefficients are quantized using the scalable AVQ and the enhancement layer bit-stream is produced.

The multi-rate functionality is obtained by allocating different number of bits to the AVQ. Thus, when a small number of bits are available to allocate for the AVQ, those bits are allocated to only a limited number of sub-bands. The AVQ are designed to take advantage of the perceptual importance by allocating more bits to the high-energy frequency bands.

Figure 3.12 shows the block diagram of the proposed N3 decoder. The decoder is basically the inverse operation of the encoder except for the post-filter at the end of the decoding process. The AVQ parameters of the enhancement layer are decoded,

transformed into time domain signal using inverse DWT (IDWT), and inverse-weighted to obtain the decoded error signal. The decoded speech signal in the core layer is combined with the error signal decoded in the enhancement layer. The combined speech signal is passed through the post-filter to produce the output speech signal. Note that when the IMDCT was used in the enhancement layer in Figure 3.6, the pre-echo reduction module was used after the inverse weighting filter. A typical artifact in transform coding known as pre-echo is observed especially when the signal energy grows suddenly, like speech onsets. Pre-echo occurs when the quantization noise in the frequency domain is translated to the time domain by an inverse MDCT and is spread uniformly in the MDCT synthesis window. Pre-echo is reduced and becomes inaudible when the DWT is used because of the better time resolution at high frequencies. Thus, the pre-echo reduction module is not employed in the proposed N3 decoder in Figure 3.12.

The post-processing unit used to enhance LP residual signals in the original iLBC has a delay of 5 ms and 10 ms for the 20 ms and 30 ms mode, respectively. It was modified so that it works with no delay at the expense of slight speech quality degradation and is used in the core-layer decoder. Note that the post-processing unit needs to be included in the decoding process in the encoder as well.

The proposed codec N3 employs the PLC algorithm used in G.729.1. The PLC algorithm of G.729.1 was modified so that it works for the proposed codec N3 in LP residual domain. In particular, some parameters which are not available in the decoder of the proposed codec N3 are estimated.

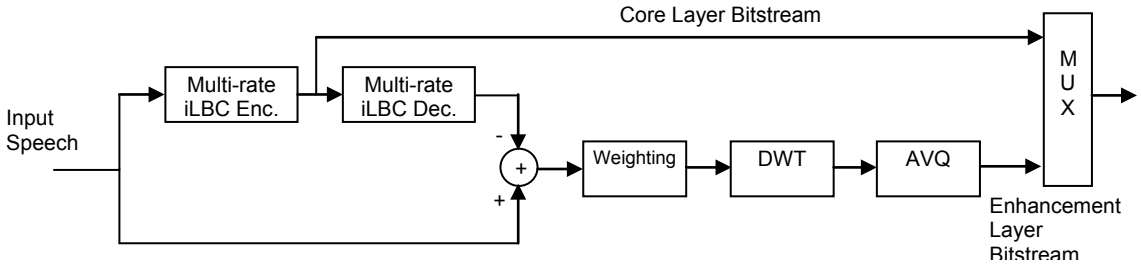


Figure 3.11: Block diagram of the proposed N3 encoder

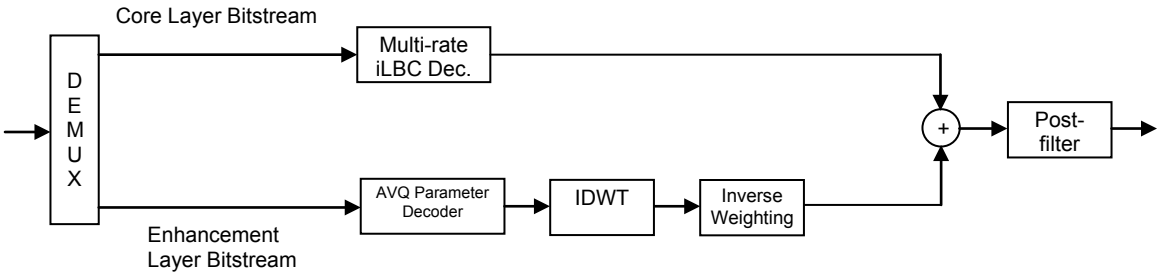


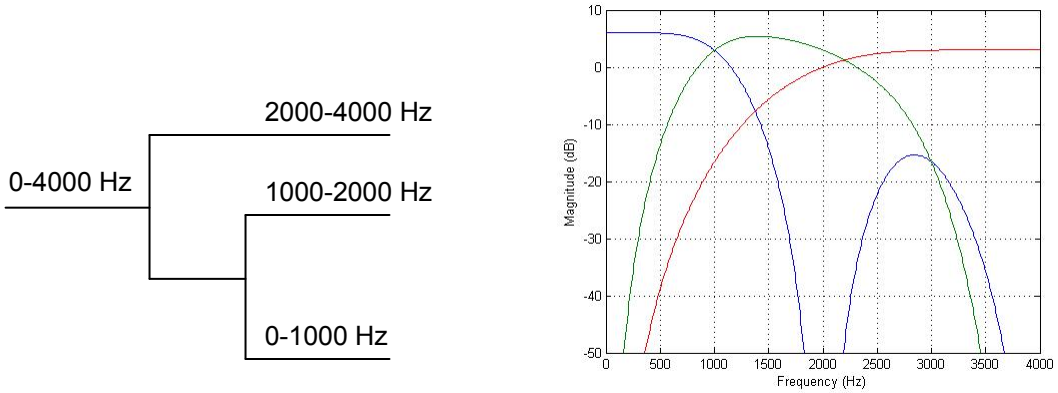
Figure 3.12: Block diagram of the proposed N3 decoder

3.3.2 Discrete Wavelet Transform

The proposed codec N3 utilizes the DWT to encode the core-layer coding error which is more likely to consist of highly non-stationary signals. Therefore, the better performance can be expected by using the DWT instead of the MDCT.

In the proposed codec N3, we used the orthogonal Daubechies wavelet [48], [64] with order 4, and 2 levels of decomposition as an initial experiment. Thus the wavelet coefficients are divided into 3 sub-bands as shown in Figure 3.13(a). The decomposition tree structure is limited to 2 levels because the delay needs to be kept small and the number of spurious sidelobes should also be kept small. The overall magnitude frequency response of the cascaded filterbank is shown in Figure 3.13(b). The scaling function, the wavelet function and four types of filter coefficients for the Daubechies wavelet with order 4 are presented in Figure 3.14.

The total delay from the DWT is 21 samples, which is about a half of the delay of 40 samples for the case of the MDCT presented in Section 3.2. The overall algorithmic delay of the proposed codec N3 for 20 ms and 30 ms frame mode is therefore 22.625 ms and 32.625 ms, respectively.



(a) (b)

Figure 3.13: DWT using the Daubechies wavelet with order 4. (a) Tree structure. (b) Magnitude frequency response.

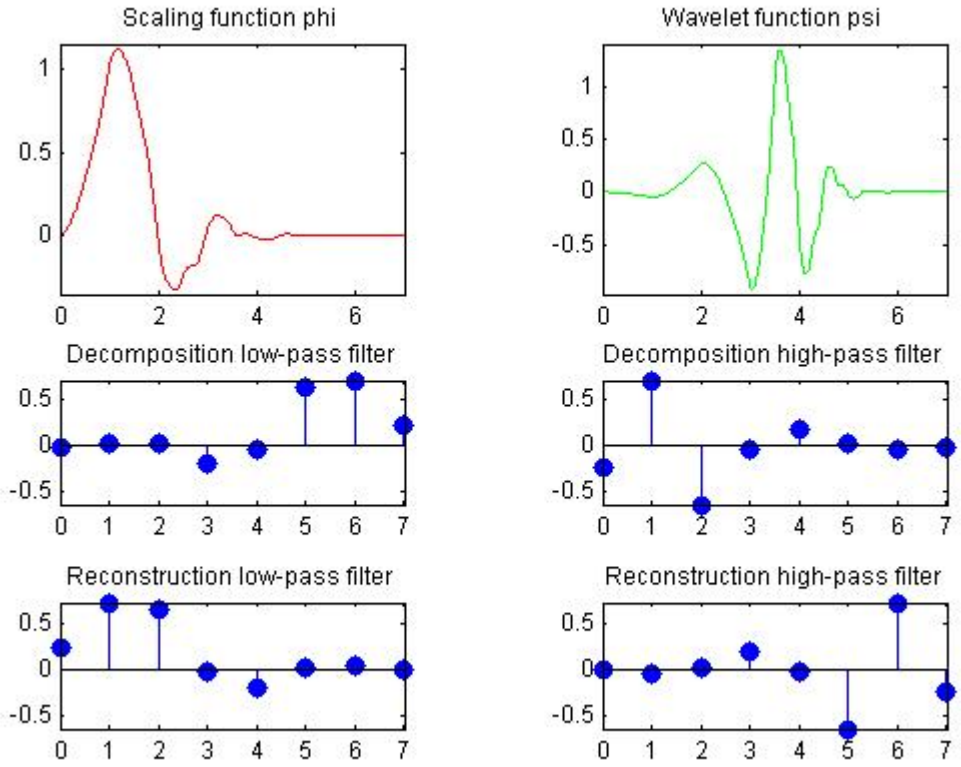


Figure 3.14: Scaling function, wavelet function and filter coefficients for Daubechies wavelet with order 4

3.3.3 Performance Evaluation

In order to evaluate the quality of speech produced by the proposed codec N3, the objective tests based on the PESQ algorithm were performed. All the results were obtained by using the frame length of 20ms, the start state length of 80, and two adaptive codebook refinement stages for the multi-rate iLBC in the core layer codec.

Figure 3.15 shows the MOS-LQO scores based on the PESQ algorithm of the proposed codec N3 as a function of bit rates to evaluate the effect of using the DWT instead of the MDCT. The same bit allocations and parameter settings were used to compare the performance of using the DWT and the MDCT. All 160 coefficients in each transform domain were used for the AVQ. It is clear that the proposed codec N3 using the DWT outperforms the proposed codec N2 using the MDCT. Note that the DWT used has only 2 levels of decomposition, yet the performance is better than that of the MDCT, which means that the beneficial effect of increased time resolution is higher than the adverse effect of decreased frequency resolution. Although the Daubechies wavelets with higher levels of decomposition and/or higher orders were also tested to evaluate the performance of the proposed codec N3, the significant improvement was not observed, which confirms that most of the performance improvement is due to good time resolution at high frequencies. Note also that the delay caused by the DWT is about a half of the delay by the MDCT.

Figure 3.16 shows the performance comparison of the proposed codec N3 with G.718 operated at 12 kbps, G.729.1 operated at 12 kbps, and AMR operated at 12.2 kbps under lossy channel conditions. Two different settings are used for the proposed codec N3: the non-scalable structure (core layer only) and the scalable structure (core layer plus enhancement layer). The proposed scalable codec was operated at 13.9 kbps, which consists of 11.15 kbps for the core layer and 1.95 kbps for the enhancement layer. The proposed non-scalable codec was operated at 11.95 kbps to compare the performance at a similar bit rate to the other codecs. All the codecs were operated using the narrowband mode.

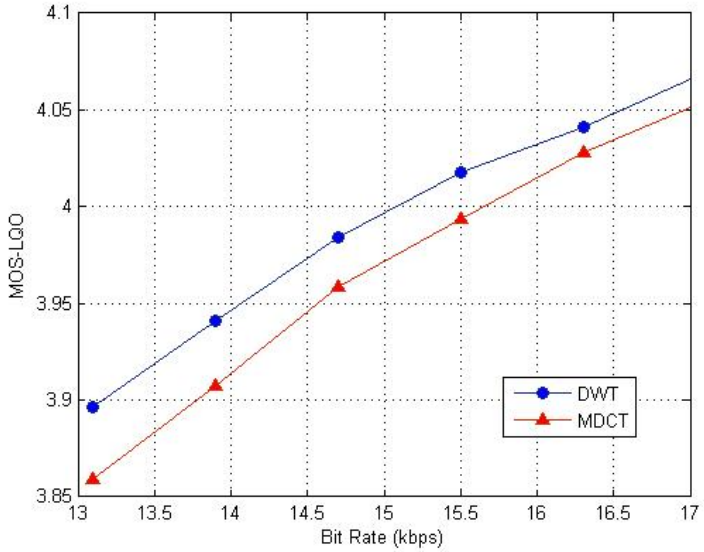


Figure 3.15: Performance comparison between the DWT and the MDCT under clean channel condition.

The performances of all the codecs in terms of packet-loss robustness are about the same except for that of AMR. If you carefully compare the performance, it is observed that the proposed non-scalable codec outperforms the proposed scalable codec at the packet loss rate higher than 5 % even though the bit rate of the non-scalable codec is lower. This is because the PLC is performed in LP residual domain and the number of bits allocated to the core layer of the proposed scalable codec is lower than that of the non-scalable codec. It is worth noting that the proposed codec N3 underperforms the other codecs under clean channel conditions, however, the performance degradation of the proposed codec N3 is more gradual than the other codecs, and the proposed codec N3 has similar performance to or even outperforms the other codecs at high packet loss rates. These results prove that the proposed codec N3 is robust to packet loss. The algorithmic delays of the codecs used for performance comparison in Figure 3.16 are given in Table 3.3. The proposed codec N3 with core layer only has a delay of 25 ms because the post-processing unit in the original iLBC, which has a delay of 5 ms, was employed to maintain high speech quality. The performance of G.718 against packet loss is relatively high, but it has the longest delay of 33.875 ms. In contrast, AMR provides low performance in terms of packet loss robustness, but has the shortest delay of 20 ms. Note

that the delay of AMR for the other bit rate modes is 25 ms, and this delay is usually maintained for the bit rate of 12.2 kbps to allow seamless frame-wise mode switching with the rest of rates. The proposed scalable codec N3 with core and enhancement layers has a relatively short delay of 22.625 ms although the bit rate used for comparison is higher than the other codecs. The proposed scalable codec N3 shows high performance under lossy channel conditions; however, the operating bit rate may be relatively high for the narrowband codec in order to achieve high speech quality as can be seen in Figure 3.15. More efficient quantization and bit allocation may be possible such as allocating a different number of bits to a different frequency band, and it may be more effective especially when the scalable structure is extended to higher frequency for the development of the wideband codec.

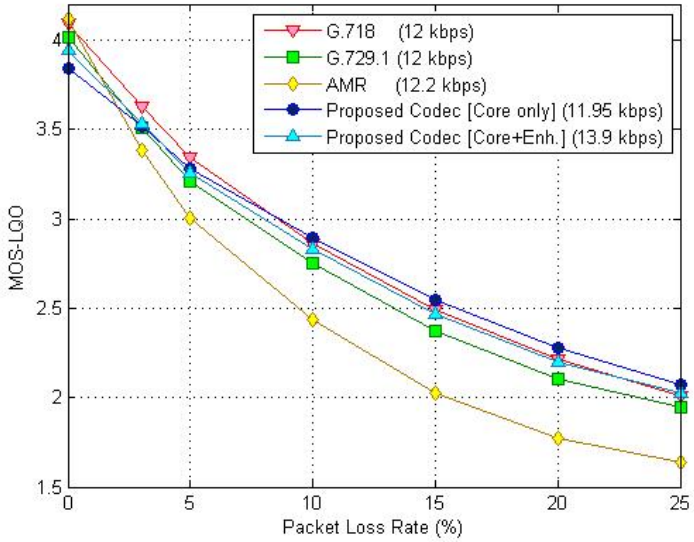


Figure 3.16: Performance comparison of the proposed codec N3 with G.718, G.729.1, and AMR under lossy channel conditions.

Table 3.3: Algorithmic delay of various codecs in Figure 3.16

Codec	**Delay (ms)**
G.718 (12 kbps)	33.875
G.729.1 (12 kbps)	25
AMR (12.2 kbps)	20
Proposed Codec N3 [Core only] (11.95 kbps)	25
Proposed Codec N3 [Core + Enh.] (13.9 kbps)	22.625

Chapter 4

Scalable Wideband Speech Codec Based on iLBC

The scalable wideband speech codec based on the iLBC is developed by employing bandwidth scalability to extend the capability of the core layer codec for wideband support. In this chapter, three types of bandwidth scalable wideband codecs are presented in the order of performance from lowest to highest. All three codecs adopt split-band structure where the input signal is decomposed into two frequency bands. The first codec uses the scalable narrowband coding scheme based on the iLBC to encode both lower-band and higher-band signals. The second and third codecs employ the same lower-band coding scheme as the first codec, however, they use the time-domain bandwidth extension (TDBWE) to encode the higher-band signal in order to improve performance at low bit rates. Whereas the first and second codecs use the MDCT for the enhancement-layer coding, the third codec employs the WPT to further enhance performance. The performance evaluation is also included for each codec.

4.1 Bandwidth Scalable Codec Based on iLBC

In this section, the bandwidth scalable wideband codec based on the iLBC is presented. The codec adopts a split-band structure, where the input signal sampled at 16 kHz is split into two sub-bands. Both the lower-band and higher-band signals are encoded by the scalable narrowband coding scheme based on the iLBC. We refer to this wideband codec as the proposed codec W1 in this dissertation.

4.1.1 Codec Structure

Figure 4.1 shows the block diagram of the proposed W1 encoder. The encoder takes the input signal sampled at 16 kHz, which is split into two sub-bands using a quadrature mirror filter (QMF) analysis filter bank: lower band and higher band. Both bands are encoded by the bit-rate scalable multi-rate iLBC described in Section 3.2 except that the shorter start state is used, which is described below. Note that the lower-band and higher-band enhancement encoders consist of the MDCT-based enhancement layer coding blocks in Figure 3.4.

The lower-band signal is first processed by a high-pass filter with 50 Hz cut-off frequency and encoded by the multi-rate iLBC using 56 start state samples. The multi-rate iLBC coding error is encoded by the lower-band enhancement encoder. The higher-band signal is first spectrally folded and processed by low-pass filter with 3 kHz cut-off frequency. The low-pass filtered signal is encoded by the multi-rate iLBC using 40 start state samples and the coding error is encoded by the higher-layer enhancement encoder. The use of the shorter start state for the higher-band multi-rate iLBC encoder is due to weak pitch characteristics of the higher-band signal, which allows reducing the number of allocated bits. In order to allow flexible bit allocation and to achieve higher speech quality, the bit-streams of both the lower-band and higher-band enhancement encoders are further divided into two separate layers. The bit-stream separation is performed by dividing MDCT coefficients in half and employing AVQ for each sub-band separately. The layering structure is summarized in Table 4.1.

The block diagram of the proposed W1 decoder is illustrated in Figure 4.2. Each layer of bit-streams is decoded by the respective decoders and the decoded signals are added to generate the lower-band signal and the higher-band signal. After post-filtering the decoded lower-band signal and spectrally folding the decoded higher-band signal, both resulting signals are combined using a QMF synthesis filter bank.

When packet losses occur, the PLC algorithm is applied to both lower-band and higher-band signals only in the multi-rate iLBC decoder. In order to evaluate the inherent packet-loss robustness of the iLBC coding scheme, the simple PLC algorithm specified in original iLBC specification [43] is employed, which is similar in performance to the PLC

algorithm in G.711 Appendix I [65]. Therefore, the robustness to packet loss of the proposed codec W1 can be easily improved by using the more advanced PLC algorithm.

Table 4.2 compares the algorithmic delay of the proposed codec W1 with G.729.1 and G.718 for wideband input and wideband output. The algorithmic delay of the proposed codec W1 is 28.9375 ms and 38.9375 ms for the frame size of 20 ms and 30 ms, respectively. It consists of 25 ms and 35 ms delay for 20 ms frame and 30 ms frame, respectively from the scalable multi-rate iLBC as explained in Section 3.2 and 3.9375 ms delay for the QMF filter bank. Owing to the short overlap for the MDCT, the codec delay is much shorter than the delay of 48.9375 ms and 42.875 ms for G.729.1 and G.718, respectively when wideband inputs and wideband outputs are used.

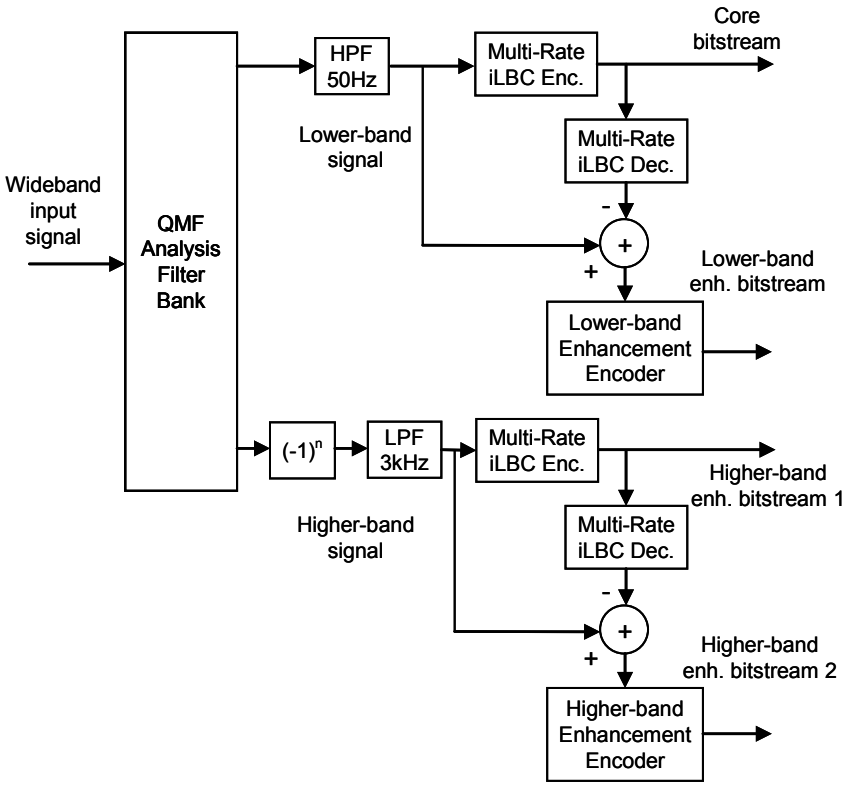


Figure 4.1: Block diagram of the proposed W1 encoder.

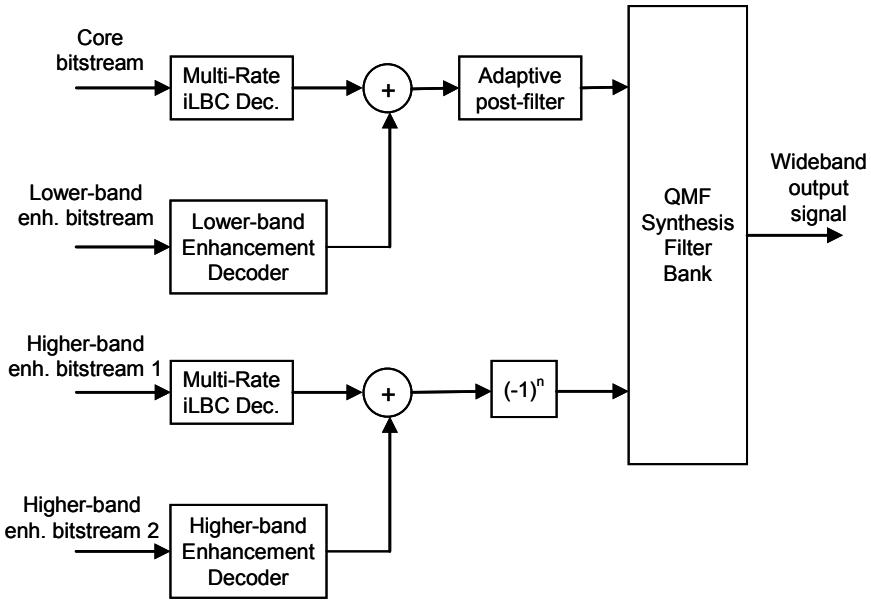


Figure 4.2: Block diagram of the proposed W1 decoder.

Table 4.1: Layer Structure of the proposed codec W1 bitstream

Layer	Description (Technique)
Layer 1	Core layer (Multi-rate iLBC)
Layer 2	Higher-band enhancement layer 1 (Multi-rate iLBC)
Layer 3	Lower-band enhancement layer 1st half sub-band (MDCT)
Layer 4	Lower-band enhancement layer 2nd half sub-band (MDCT)
Layer 5	Higher-band enhancement layer 2 1st half sub-band (MDCT)
Layer 6	Higher-band enhancement layer 2 2nd half sub-band (MDCT)

Table 4.2: Algorithmic delay of the proposed codec W1 compared with G.729.1 and G.718 for wideband input and wideband output

Codec	Delay (ms)
Proposed Codec W1 (20 ms frame)	28.9375
Proposed Codec W1 (30 ms frame)	38.9375
G.729.1	48.9375
G.718	42.875

4.1.2 Performance Evaluation

In order to evaluate the performance of the proposed codec W1 for wideband input and wideband output, the objective tests based on the PESQ algorithm were performed. Five different configurations are used for evaluation and those are described in Table 4.3. For example, case 1 configuration uses the frame size of 20 ms, 3 layers of bit-streams, and 3 adaptive CB stages for the multi-rate iLBC in the core layer. The number of adaptive CB stages for the multi-rate iLBC in the higher-band enhancement layer 1 was fixed to one. The number of start state samples was fixed to 56 and 40 for the multi-rate iLBC in the core layer and the higher-band enhancement layer 1, respectively as described in the previous section.

Figure 4.3 shows the MOS-LQO scores of 5 different configurations of the proposed codec W1 and G.729.1 as a function of bit rates to compare their performances. For each case of configurations, the MOS-LQO score at the lowest bit rate corresponds to the performance of the codec in which only layer 1 and layer 2 are employed. Adding another layer gives the score at next higher bit rate for each case. The score at the lowest bit rate is lower than that of G.729.1 at a similar bit rate because the proposed codec W1 uses the same type of iLBC-based coding scheme as the core layer for the higher-band enhancement layer 1, which leads to high robustness to packet loss in exchange for the requirement of extra bits. Whereas the performance of the proposed codec W1 is relatively low at low bit rates, higher MOS-LQO scores than G.729.1 can be achieved at high bit rates.

Table 4.3: Summary of configurations used for evaluation of the proposed codec W1

Configuration	Frame size	Number of layers	# of adaptive CB stages for core codec
Case 1	20 ms	3	3 stages
Case 2	20 ms	4	2 stages
Case 3	30 ms	4	3 stages
Case 4	30 ms	4	2 stages
Case 5	30 ms	5	2 stages

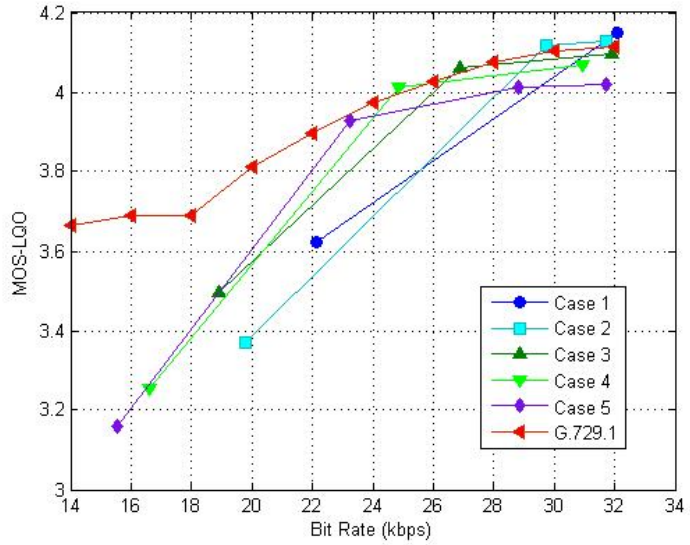


Figure 4.3: Performance comparison of 5 different configurations of the proposed codec W1 with G.729.1 under clean channel condition.

Another interesting observation is that case 1 can achieve the highest score at about 32 kbps. Note that case 1 uses only 3 layers. This indicates that the accurate representation for lower frequency contents in speech is more important than higher frequency information. A similar conclusion can be drawn from the observation that the MOS-LQO scores for each case suddenly start to saturate after adding another layer on top of first 3 layers. This sudden saturation also indicates that the bit allocation is not

optimum while the high score of using only 3 layers shows high potential of the proposed codec W1.

Figure 4.4 shows the MOS-LQO scores of the case 1 configuration of the proposed codec W1 at 32.1 kbps and G.729.1 at 32 kbps as a function of packet loss rates to compare their performances of packet-loss robustness. The performance of the proposed codec W1 is almost the same as that of G.729.1 at all packet loss rates. Note, however, that the PLC algorithm of G.729.1 is optimized specifically for its codec whereas the PLC algorithm employed in the proposed codec W1 is a simple algorithm that can be used for any frame-independent codec as pointed out in the previous section. **Therefore, the high performance of the proposed codec W1 in terms of robustness to packet loss comes from the inherent nature of the iLBC coding scheme.**

The performance comparison of the proposed codec W1 with G.729.1 at about 24 kbps under lossy channel condition is shown in Figure 4.5. The case 4 configuration at 24.8 kbps is used as the performance of the proposed codec W1. The performance curve of the case 1 configuration at 22.2 kbps is also included in Figure 4.5 to compare the performance in terms of packet-loss robustness. About the same performance is achieved for the case 4 configuration of the proposed codec W1 and G.729.1 at around 24 kbps as in the results at around 32 kbps. What is more remarkable is that the performance of the proposed codec W1 with the case 1 configuration at 22.2 kbps is also the same as that of G.729.1 at 24 kbps at the packet loss rate of 25 % despite the fact that the performance under clean channel condition is lower than that of G.729.1 by about 0.4 point in MOS-LQO score. Therefore, it is clear that the proposed codec W1 inherently has higher robustness to packet loss.

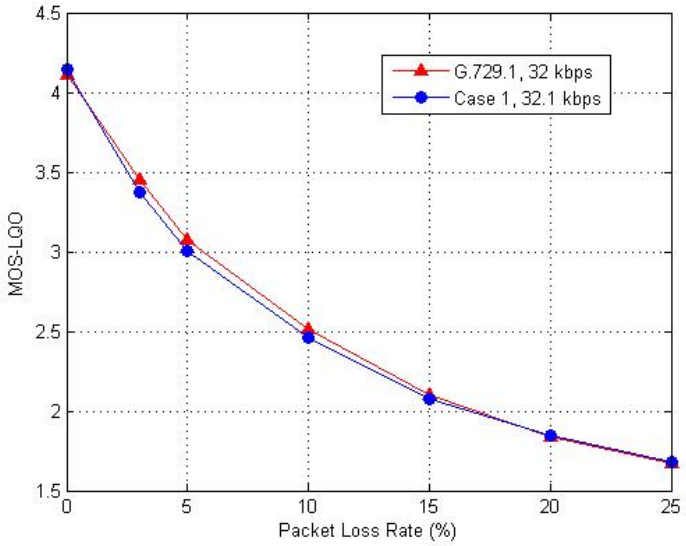


Figure 4.4: Performance comparison of the case 1 configuration of the proposed codec W1 with G.729.1 at around 32 kbps under lossy channel condition.

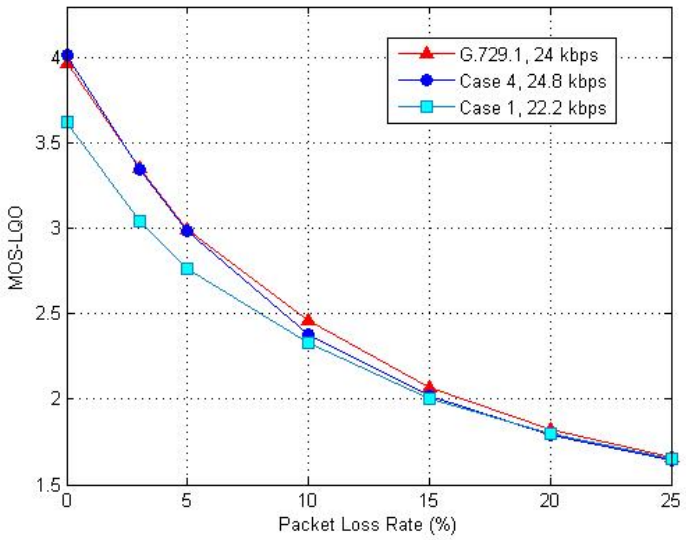


Figure 4.5: Performance comparison of the case 4 configuration of the proposed codec W1 with G.729.1 at around 24 kbps under lossy channel condition. The performance curve of case 1 configuration at 22.2 kbps is also included for comparison.

4.2 Performance-Enhanced Wideband Codec Using the MDCT

The scalable wideband speech codec based on the iLBC (the proposed codec W1) presented in the previous section achieved speech quality equivalent to ITU-T G.729.1 at high bit rates. However, the performance was limited at low bit rates. In this section, various approaches are applied to the previously developed codec in order to improve performance especially at low bit rates and the performance-enhanced scalable wideband codec using the MDCT is developed and is referred to as the proposed codec W2 in this dissertation. In particular, the time-domain bandwidth extension (TDBWE) is used to encode higher-band signal, and the efficient coding structure is employed in enhancement layers.

4.2.1 Codec Structure

The proposed codec W2 is a scalable wideband extension of the multi-rate codec based on the iLBC (the proposed codec N1) described in Section 3.1. Figure 4.6 shows the block diagram of the proposed W2 encoder. The encoder operates on 20 ms input frames. The wideband input signal is sampled at 16 kHz and split into two sub-bands using a quadrature mirror filter (QMF) analysis filter bank.

The lower-band signal is first processed by a high-pass filter with 50 Hz cut-off frequency and encoded by the multi-rate iLBC using 80 start state samples and three adaptive CB refinement stages, which generates the core layer (Layer 1) bit-stream. The multi-rate iLBC coding error is computed by subtracting the decoded speech signal from the high-pass filtered lower-band signal and processed by perceptual weighting filter. The weighted error signal is transformed into frequency domain by the MDCT with the reduced overlap of 5 ms as employed in Section 3.2.

The higher-band signal is first spectrally folded and processed by a low-pass filter with 3 kHz cut-off frequency. The low-pass filtered signal is encoded by the TDBWE

encoder and Layer 2 bit-stream is generated. The MDCT is applied to the coding error from the TDBWE encoder and the MDCT coefficients are obtained.

The resulting two sets of the MDCT coefficients are concatenated to cover whole frequency range of wideband signals. Those MDCT coefficients are divided into two parts at either 1 kHz or 2 kHz and each part is separately quantized using the scalable AVQ and Layer 3 and Layer 4 bit-streams are produced. In order to further improve performance, the quantization errors from Layers 3 and 4 are encoded by the scalable AVQ, which generates Layer 5 bit-stream.

The bit-stream generated by the encoder is scalable. The enhancement layers can be truncated during transmission and speech signal is still decoded with decreased quality.

Note that the TDBWE algorithm used is the same as the one employed in G.729.1 except that a predefined fixed sequence is used for the TDBWE excitation signal in the decoder instead of an artificially generated signal based on received parameters so that the TDBWE coding error can be used to improve performance. The fixed sequence was generated from random variables uniformly distributed between -1 and 1, center clipped at 0.6 and ternary level quantized to {-1, 0, 1}. 62.5 % of the samples are zeros.

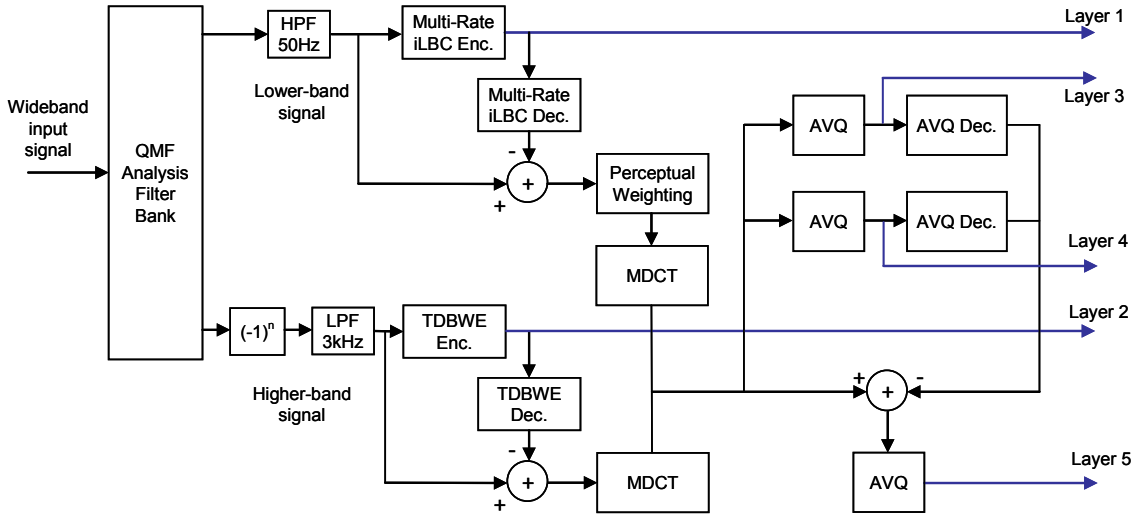


Figure 4.6: Block diagram of the proposed W2 encoder

The block diagram of the proposed W2 decoder is illustrated in Figure 4.7. Each layer of bit-streams is decoded by the respective decoders and the decoded signals are added to generate the lower-band signal and the higher-band signal. After post-filtering the decoded lower-band signal and spectrally folding the decoded higher-band signal, both resulting signals are delay-adjusted and combined using a QMF synthesis filter bank.

The enhancement unit in LP residual domain used in the original iLBC decoder is employed in the multi-rate iLBC, which causes 5 ms delay. Therefore, the overall algorithmic delay is 38.9375 ms, which consists of 20 ms for input frame, 10 ms for the enhancement unit in the encoder and the decoder, 5 ms for the overlap-add operation after the IMDCT, and 3.9375 ms for the QMF analysis-synthesis filterbank.

In order to improve performance under lossy channel conditions, the proposed codec W2 employs the PLC algorithm used in G.729.1. In the lower band, some parameters which are not available in the decoder of the proposed codec W2 are estimated. In the higher band, instead of shaping an artificially generated excitation signal according to the previously received time and frequency envelopes, only the TDBWE mean-time envelope and the frequency envelopes of the previous frame are used to shape a predefined fixed signal when the frame is not received. The energy of the concealed signal is gradually decreased for the consecutive lost frames.

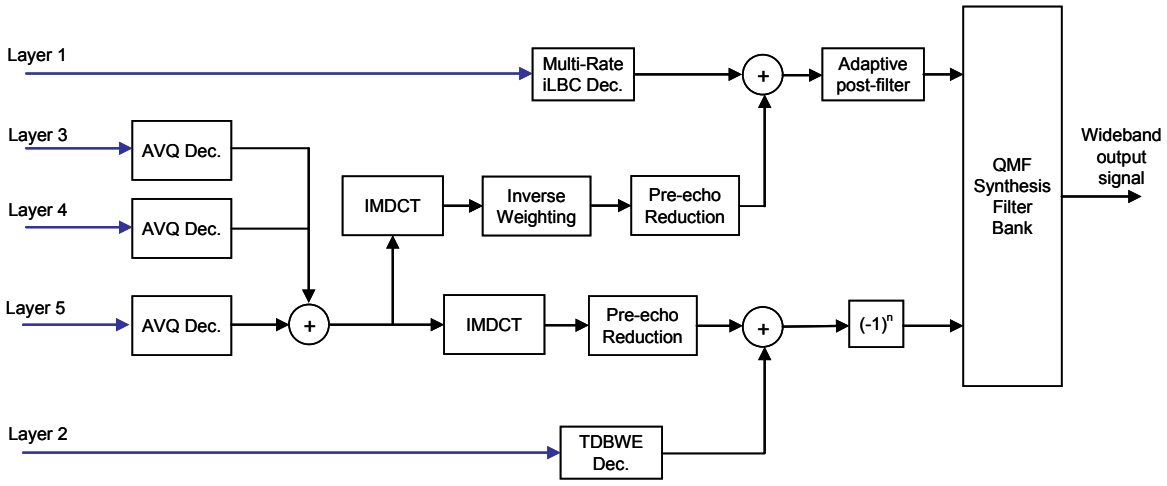


Figure 4.7: Block diagram of the proposed W2 decoder

4.2.2 Performance Evaluation

In order to evaluate the quality of speech produced by the proposed codec W2, the objective tests based on the PESQ algorithm were performed. All the results were obtained for wideband input and wideband output.

Four different modes are used for performance evaluation and the bit allocation of each mode is presented in Table 4.4. For example, in Mode 1, 276, 33, 231, and 87 bits are allocated to Layer 1, 2, 3 and 4, respectively, and Layer 5 is not used. Note that the term "Mode" is used here instead of "Case" used in Table 4.3 in order to differentiate a choice in bit allocations for "Mode" from a choice in parameter settings for "Case". The frequency boundary between Layer 3 and Layer 4 is 2 kHz in Mode 1, and 1 kHz in Mode 2, 3, 4. The frequency range of the MDCT coefficients in Layer 5 is limited to 0–4 kHz in all modes. Note that at least both Layer 1 and Layer 2 are required to encode wideband signals.

Figure 4.8 shows the MOS-LQO scores computed by PESQ algorithm as a function of bit rates to compare the performance of the proposed codec W2 using Mode 1 to 4, G.729.1 and the codec previously presented in Section 4.1. Only the Case 1 and 2 in Figure 4.3 are included for fair comparisons using the same frame size of 20 ms. Mode 1 achieves the best performance among four modes at high bit rates whereas Mode 4 achieves the best performance at the low bit rate of about 18 kbps. The proposed codec W2 performs significantly better than the previous results at low bit rates. It is obvious that the proposed codec W2 benefits from the TDBWE for higher performance at low bit rates. The performance of the proposed codec W2 is even higher than that of G.729.1 at the bit rate of 18 kbps or higher although the performance gap gets smaller as the bit rate increases. The sudden drop of the codec performance below 18 kbps is mainly because the iLBC-based codec generally underperforms the CELP-based codec when operated at the same low bit rate as a core-layer codec. However, it is possible for the proposed codec W2 to achieve similar performance to G.729.1 at low bit rates if the performance of the core-layer codec and the performance at high bit rates are allowed to be sacrificed.

Table 4.4: Bit allocation of experimental modes for the proposed codec W2

	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
Mode 1	276	33	231	87	0
Mode 2	276	33	135	71	103
Mode 3	260	33	103	71	151
Mode 4	244	33	87	71	199

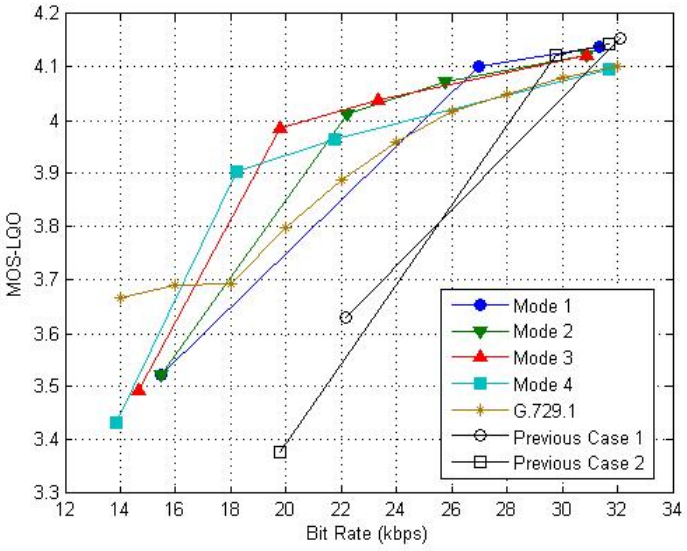


Figure 4.8: Performance comparisons of the proposed codec W2, G.729.1, and Case 1 and 2 (results in Figure 4.3) of the previously presented codec (the proposed codec W1) in Section 4.1 under clean channel condition.

Figure 4.9 shows the performance comparison of the proposed codec W2 operated at 31.35 kbps using Mode 1 and G.729.1 operated at 32 kbps under lossy channel conditions where the MOS-LQO scores are plotted as a function of packet loss rates. The proposed codec W2 slightly outperforms G.729.1 at all packet loss rates. Both codecs employ basically the same PLC algorithm; however, the proposed codec W2 needs to estimate some parameters which are not available at the decoder, including all the frame erasure

concealment (FEC) parameters. In other words, the PLC algorithm is not optimized for the proposed codec W2. Therefore, we can see that the proposed codec W2 is inherently more robust to packet loss than G.729.1 and higher performance can be expected for the proposed codec W2 using the optimized PLC algorithm.

In Figure 4.10, the proposed codec W2 operated at 13.85 kbps using Mode 4 and G.729.1 operated at 14 kbps are compared in terms of packet-loss robustness. With no packet loss, G.729.1 outperforms the proposed codec W2 by about 0.23 point in terms of MOS-LQO score; however, the proposed codec W2 performs better at packet loss rates higher than 5 % and the performance gap becomes bigger as the packet loss rate increases. Note that G.729.1 transmits the FEC parameters in Layer 2, 3, and 4, and the bit-stream at 14 kbps comprises Layer 1, 2, and 3. Thus, when G.729.1 is operated at 14 kbps, one of the FEC parameters in Layer 4 is not available at the decoder, which leads to the performance degradation in packet loss situations. Note also that at low bit rates, the performance difference of the core-layer codecs clearly comes to the surface. It is clear that the proposed codec W2 has higher robustness to packet loss than G.729.1, and it is worth considering using the frame-independent coding such as the iLBC-based coding in the core layer codec for VoIP applications.

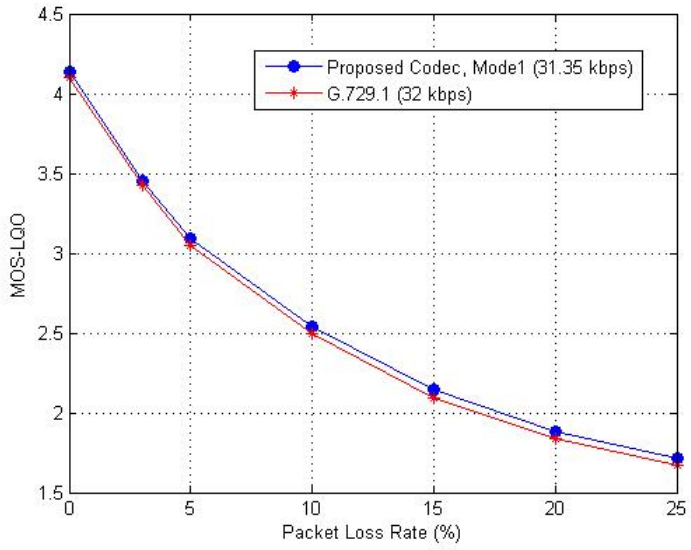


Figure 4.9: Performance comparison of the proposed codec W2 using Mode 1 at 31.35 kbps and G.729.1 at 32 kbps under lossy channel conditions.

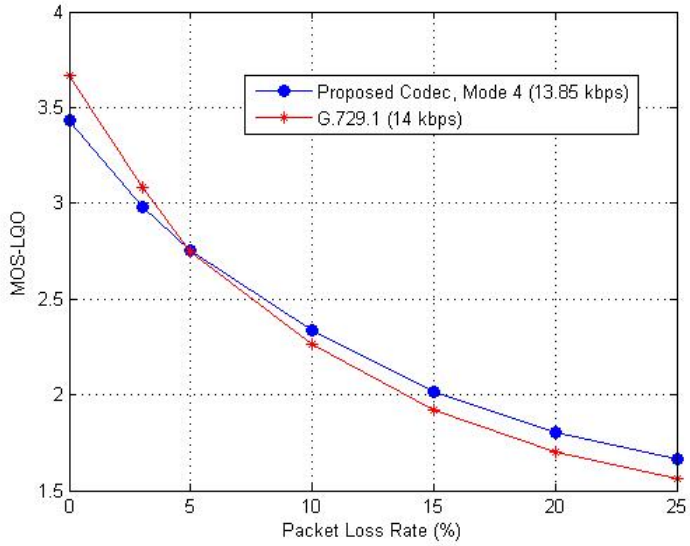


Figure 4.10: Performance comparison of the proposed codec W2 using Mode 4 at 13.85 kbps and G.729.1 at 14 kbps under lossy channel conditions.

4.3 Performance-Enhanced Wideband Codec Using the WPT

The performance-enhanced scalable wideband codec using the MDCT (the proposed codec W2) was presented in the previous section, and the objective quality evaluation showed that the codec provided higher robustness to packet loss than G.729.1 and even outperformed G.729.1 at most bit rates except for low bit rates under clean channel condition. In this section, we propose a novel scalable wideband speech codec which provides further improvements in performance to the codec presented in the previous section. In particular, the wavelet packet transform (WPT) is employed instead of the MDCT in the enhancement layers to improve performance. This codec is referred to as the proposed codec W3 in this dissertation. The proposed codec W3 is designed based on both the objective and subjective quality measure.

4.3.1 Codec Structure

The proposed codec W3 is a scalable wideband extension of the multi-rate iLBC (the proposed codec N1) described in Section 3.1 and has basically the same structure as the proposed codec W2 introduced in Section 4.2. Figure 4.11 shows the block diagram of the proposed W3 encoder. The encoder operates on 20 ms input frames. The wideband input signal is sampled at 16 kHz and split into two sub-bands using a quadrature mirror filter (QMF) analysis filter bank.

The lower-band signal is first processed by a high-pass filter with 50 Hz cut-off frequency and encoded by the multi-rate iLBC using 80 start state samples and three adaptive CB refinement stages, which generates the core layer (Layer 1) bitstream. The multi-rate iLBC coding error is computed by subtracting the decoded speech signal from the original speech signal and processed by perceptual weighting filter. The weighted error signal is decomposed into wavelet coefficients by the WPT.

The higher-band signal is first spectrally folded and processed by low-pass filter with 3 kHz cut-off frequency. The low-pass filtered signal is encoded by the TDBWE and Layer 2 bitstream is generated. The WPT is applied to the coding error from the TDBWE and the wavelet coefficients are obtained.

The resulting two sets of wavelet coefficients are concatenated to cover whole frequency range of wideband input signal. Those wavelet coefficients are divided into two parts at either 1 kHz or 2 kHz and each part is separately quantized using the scalable AVQ and Layer 3 and Layer 4 bitstreams are produced. In order to further improve performance, the quantization errors from Layers 3 and 4 are encoded by the scalable AVQ, which generates Layer 5 bitstream.

The bitstream produced by the encoder is scalable. The enhancement layers can be truncated during transmission and speech signal is still decoded with decreased quality.

Note that the TDBWE algorithm used is the same as the one employed in G.729.1 except that a predefined fixed sequence is used for the TDBWE excitation signal in the decoder instead of an artificially generated signal based on received parameters for the

narrowband decoder so that the TDBWE coding error can be used to improve performance. This TDBWE codec provides higher performance than the original one when the enhancement layer is used at the cost of slightly lower speech quality in the case of using only the TDBWE decoder for higher-band signal decoding. The fixed sequence was generated from random variables uniformly distributed between -1 and 1, center clipped at 0.6 and ternary level quantized to {-1, 0, 1}. 62.5 % of the samples are zeros.

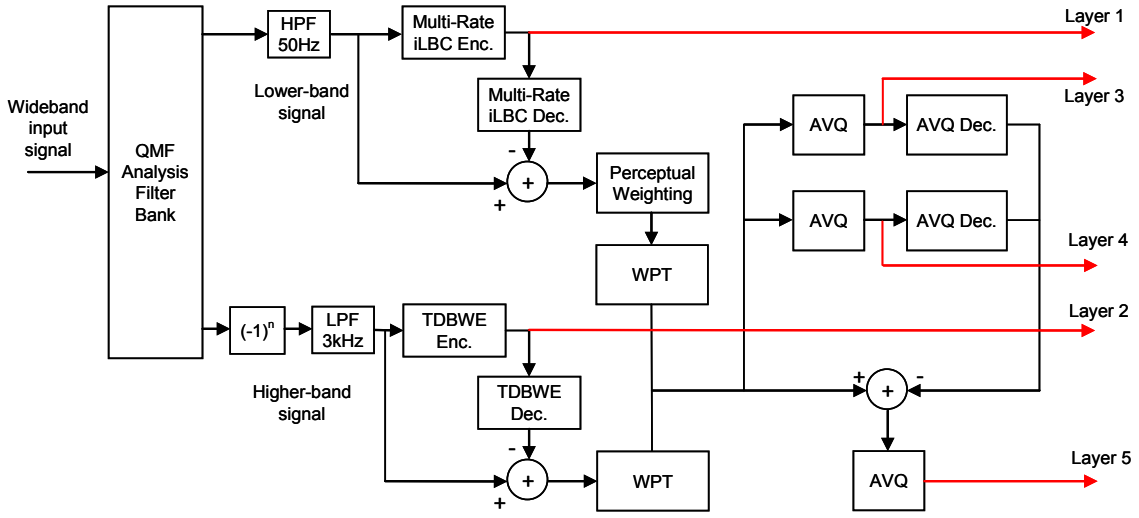


Figure 4.11: Block diagram of the proposed W3 encoder

The block diagram of the proposed W3 decoder is illustrated in Figure 4.12. Each layer of bitstreams is decoded by the respective decoders and the decoded signals are added to generate the lower-band signal and the higher-band signal. After post-filtering the decoded lower-band signal and spectrally folding the decoded higher-band signal, both resulting signals are delay-adjusted and combined using a QMF synthesis filter bank. The pre-echo reduction blocks used in Figure 4.7 was removed because the WPT has the good time localization property.

The enhancement unit in LP residual domain used in the original iLBC decoder is employed in the multi-rate iLBC, which causes 5 ms delay. Therefore, the overall algorithmic delay is 40.3125 ms, which consists of 20 ms for input frame, 10 ms for the enhancement unit in the encoder and the decoder, 6.375 ms for the wavelet filters, and

3.9375 ms for the QMF analysis-synthesis filterbank. The delay caused by the wavelet filters is explained in Section 4.3.2.

In order to improve performance under lossy channel conditions, the proposed codec W3 employs the PLC algorithm used in G.729.1 with the necessary modification. In particular, in the lower band, some parameters which are not available in the decoder of the proposed codec W3 are estimated. In the higher band, in addition to using a predefined fixed excitation signal, the attenuation factor applied to the concealed frames in the case of consecutive frame losses was fine-tuned.

The computational complexity of the original iLBC is in a range of G.729A according to [46]. The Multi-rate iLBC is only slightly more complex than the original iLBC [24]. The implementation of the WPT takes at most $O(N \cdot \log N)$ where N is the length of input signals, as compared to $O(N \cdot \log N)$ for the MDCT. Furthermore, the structure of the proposed codec W3 is similar to G.729.1. Therefore, the computational complexity of the proposed codec W3 should be comparable to that of G.729.1.

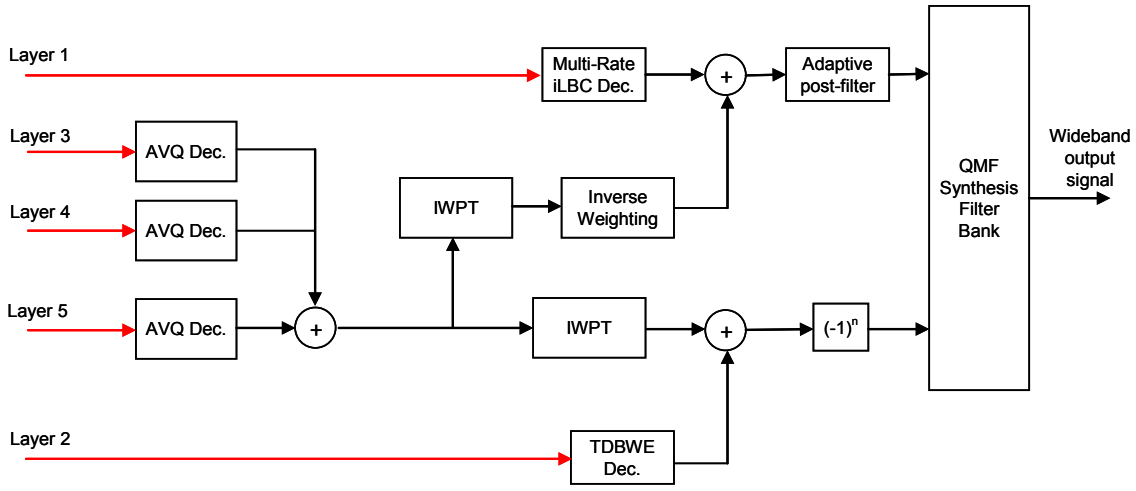


Figure 4.12: Block diagram of the proposed W3 decoder

4.3.2 Wavelet Packet Transform

The WPT can be used to better capture localized waveforms in time domain than the Fourier-based transforms such as the MDCT. The proposed codec W3 utilizes the WPT

to encode the lower-band and higher-band coding error which is more likely to consist of highly non-stationary signals. Therefore, better performance can be expected by replacing the MDCT with the WPT.

In the proposed codec W3, we used the reverse biorthogonal spline wavelet [48] with order 6 and 8 for decomposition and reconstruction, respectively, for lower-band signals and the biorthogonal spline wavelet with order 6 and 8 for reconstruction and decomposition, respectively, for higher-band signals. The scaling function, the wavelet function and four types of filter coefficients for the former wavelet are presented in Figure 4.13. The advantage of biorthogonal over orthogonal wavelet is that wavelet filter coefficients can be symmetric. Hence, the wavelet filters used have linear phase. The tree structure for the WPT is shown in Figure 4.14. This decomposition structure is designed to roughly resemble the critical band divisions except for low frequencies. The only two levels of decomposition for lower-band signal were chosen because the delay of wavelet filters needs to be kept small and any further decomposition results in a greater number of significant sidelobes. Although frequency selectivity can be enhanced by using filters with narrower transition bands, more filter taps cause a larger delay, which is often not acceptable for interactive speech applications. Since capturing time-localized waveforms is the main purpose of using the WPT instead of the MDCT, frequency selectivity can be sacrificed without significant performance degradation.

The delay from the wavelet filters is 6.375 ms, which is only 1.375 ms longer than the delay of 5 ms caused by the MDCT with reduced-overlap window used in the codec presented in Section 4.2.

In the next sub-section, the performance evaluation of the proposed codec W3 using various wavelets and tree structures is provided, which resulted in the selection of the reverse biorthogonal spline wavelet with order 6 and 8 for decomposition and reconstruction, respectively, for lower-band signals.

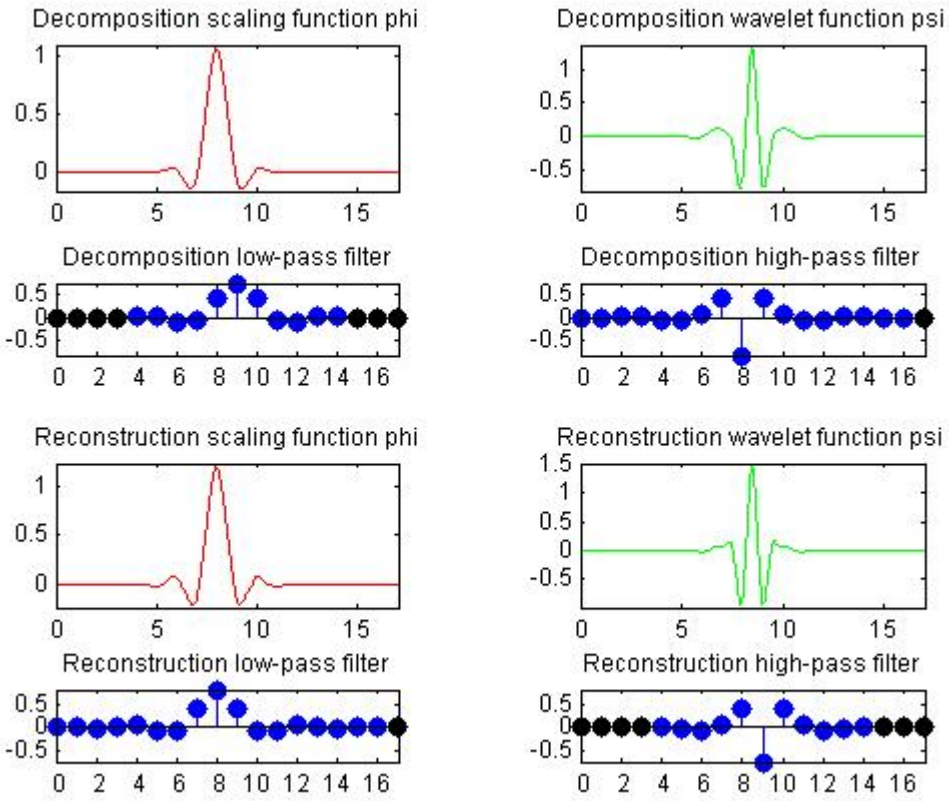


Figure 4.13: Scaling function, wavelet function and filter coefficients for the reverse biorthogonal spline wavelet with order 6 and 8 for decomposition and reconstruction, respectively.

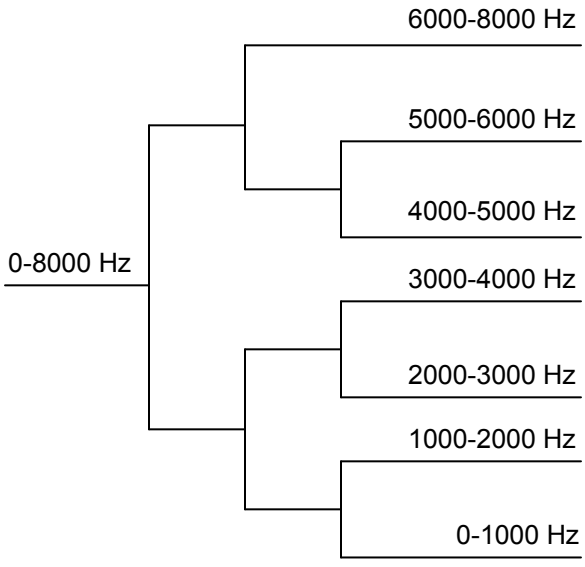


Figure 4.14: Tree structure for the WPT

Selection of Wavelet and Tree Structure

In order to evaluate the performance of the proposed codec W3 using various wavelets and tree structures, the objective tests based on the PESQ algorithm were performed. The block diagram of the proposed W3 encoder in Figure 4.11 is slightly modified so that the wavelet coefficients for the lower-band and higher-band signals are separately quantized by the AVQ to properly evaluate the effect of different wavelets and tree structures for the lower-band signal and the higher-band signal independently. The wavelets used for performance evaluation are as follows: Haar, Daubechies (db), Symlets (sym), Coiflets (coif), Biorthogonal spline (bior), and Reverse biorthogonal spline (rbio) wavelets.

First, the wavelets and tree structures for the lower-band signal are explored. The tree structures for the WPT of the lower-band signal shown in Figure 4.15 are used to compare the performance. The number of decomposition levels directly affects the amount of delay. When the number of decomposition levels is larger than three, the delay is often too large for the interactive speech applications although the performance is barely improved or degraded because of the larger number of spurious sidelobes. When it is one, the performance is almost always lower than the case for larger decomposition levels. Thus, we used either two or three levels of decomposition for performance comparisons. All the parameters are fixed except for two choices of the upper frequency limit for the lower-band wavelet coefficients: 2000 Hz and 4000 Hz. If we chose the upper frequency limit of 2000 Hz, only the wavelet coefficients from 0 to 2000 Hz are used for the AVQ, whereas the upper frequency limit of 4000 Hz means that all the lower-band wavelet coefficients from 0 to 4000 Hz are used. The number of bits allocated to the AVQ for the lower-band wavelet coefficients are varied for performance evaluation to plot the MOS-LQO scores as a function of bit rate.

Figure 4.16 shows the performance comparisons of the proposed codec W3 using various tree structures in Figure 4.15 when the Daubechies wavelet with order 4 is employed. The performances of some cases are exactly the same because their tree structures are the same for the frequency range of interest. Note that the performance of

using the upper frequency limit of 2000 Hz is higher than that of using whole frequency range of 0 to 4000 Hz. This is because the bits are allocated to only a limited number of coefficients when the available number of bits is small. Thus, the performance difference gets smaller and eventually two performance curves cross as the bit rate increases. The performance of using the tree structure (b) is better than that of using (a). The use of three levels of decomposition barely outperforms the use of two levels of decomposition although the delay more than doubles. The performances of using (d) and (e) are about the same, and they are better than the performance of using (c) when the upper frequency limit is 2000 Hz as expected. Note that the tree structure of (e) roughly, but most accurately among 5 choices in Figure 4.15, approximates the critical band divisions of the human auditory system except for low frequencies, however, the use of (d) slightly outperforms the use of (e), probably because the number of spurious sidelobes is larger and/or the benefit of good time localization is higher than that of good frequency resolution. Therefore, the tree structure (b) or (c) is a good choice and (b) was selected eventually using the different choice of wavelet. However, since the delay from the WPT using the Daubechies wavelet with order 4 and three levels of decomposition is 49 samples, which is only 9 samples longer than the delay for the proposed codec W2 using the MDCT in Section 4.2, the tree structure (c), which provides the highest performance, is used for performance comparisons when the Daubechies wavelet with order 4 is employed.

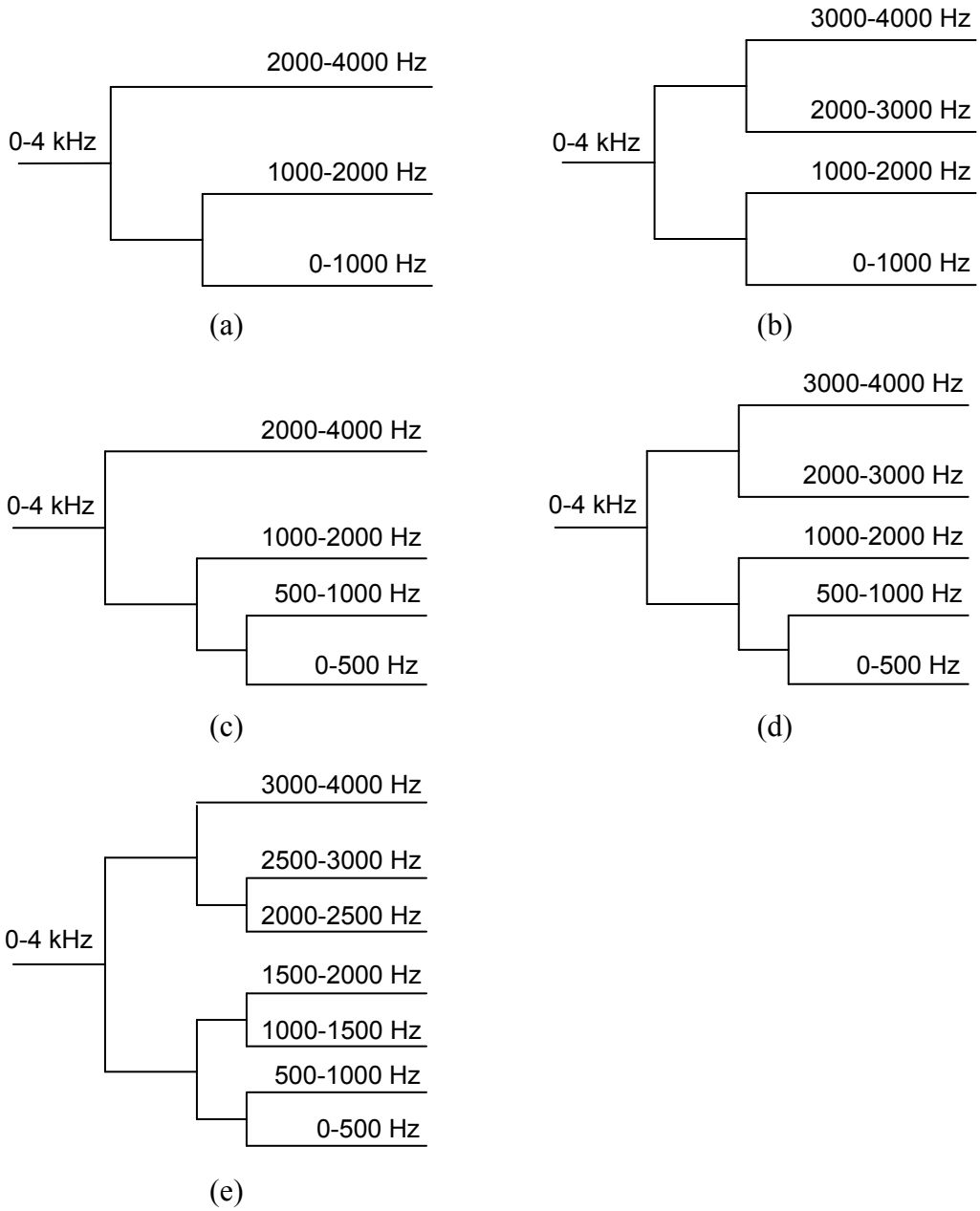


Figure 4.15: Tree structures for the WPT of the lower-band signal that are used for performance comparisons

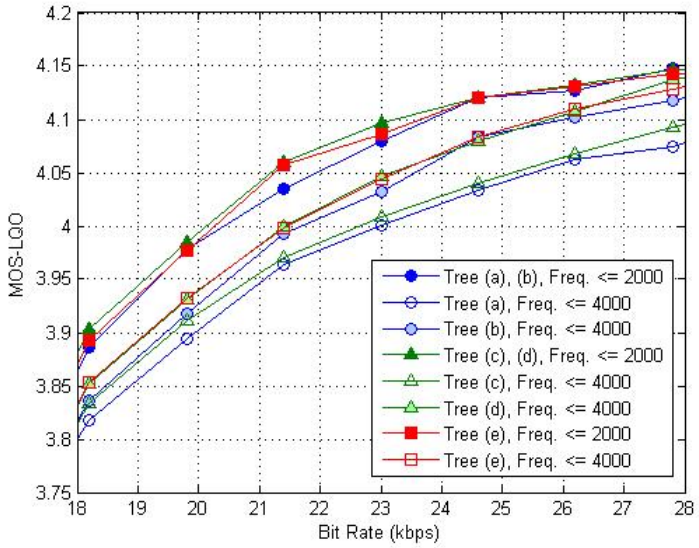


Figure 4.16: Effect of using different tree structures in Figure 4.15 with Daubechies wavelet with order 4.

Figure 4.17 shows the performance comparisons of the proposed codec W3 using the Daubechies wavelet with different order: 1, 2, 4, and 6 (db1, db2, db4, and db6), when the tree structure (d) is employed. The performance becomes higher as the order increases from 1 to 4. The performance peaks when the order is 4, and it becomes smaller when the order is larger than 4. A similar trend was confirmed when the tree structure (b) was used. A similar trend was also confirmed when the symlet wavelets are used. The symlet wavelets are a modified version of Daubechies wavelets with least asymmetry. Thus, the associated scaling filters are near linear-phase filters and this family of wavelets is probably better suited for speech applications.

Figure 4.18 illustrates the performance comparison of the proposed codecs W3 using various wavelets with order and the tree structure selected to achieve the highest performance. The specific wavelets used are the Daubechies wavelet with order 4 (db4), the symlet wavelet with order 4 (sym4), the coiflet wavelet with order 1 (coif1), the biorthogonal spline wavelet with order 6 and 8 for reconstruction and decomposition, respectively (bior6.8), and the reverse biorthogonal spline wavelet with order 6 and 8 for decomposition and reconstruction, respectively (rbio6.8). The performances of these wavelets are similar; however, the biorthogonal wavelet filters have the benefit of linear

phase. Furthermore, since rbio6.8 has slightly better performance overall than bior6.8, rbio6.8 was selected to be used for lower-band signals. The delay from the WPT with rbio6.8 and the tree structure of (b) is 51 samples, which is reasonable.

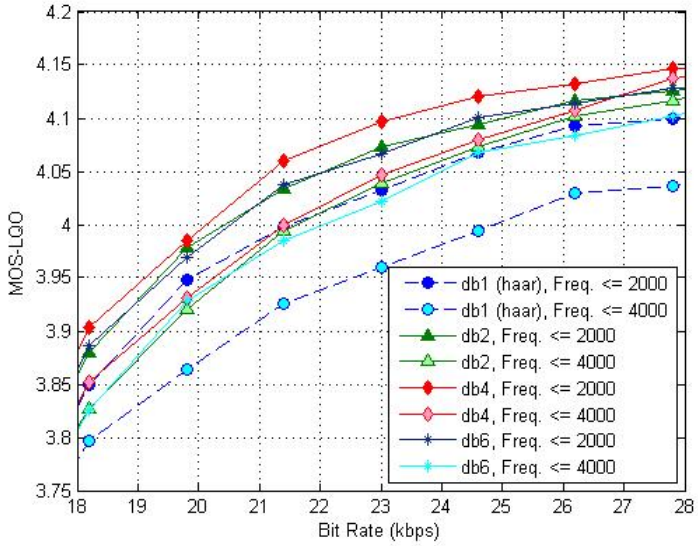


Figure 4.17: Effect of using the Daubechies wavelet with different orders when the tree structure (d) is employed.

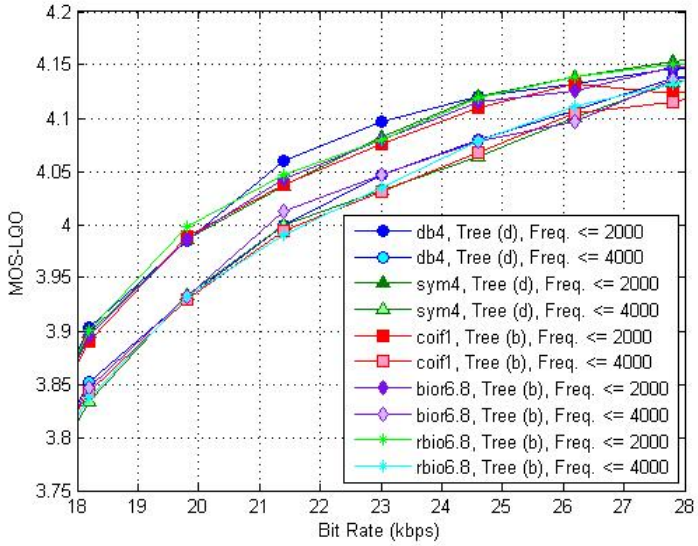


Figure 4.18: Performance comparison of the proposed codec W3 using various wavelets with order and the tree structure selected to achieve the highest performance for the lower-band signal.

The wavelets and tree structures for the higher-band signal are similarly explored next. The performance comparison of the proposed codec W3 using various tree structures for the higher-band signal revealed that the tree structures for the DWT provided higher performance than the other tree structures. Thus, the only parameter for the tree structure is the number of decomposition levels. Figure 4.19 shows that the performance comparison of the proposed codec W3 using various wavelets with different order and different number of decomposition levels for the higher-band signal. The order and number of decomposition levels are selected so that the high performance is achieved without causing too large delay. The delay caused by the wavelet filters is also provided for each wavelet in terms of the number of samples in Figure 4.19. When 3 levels of decomposition is used, the performance increases more rapidly as the bit rate starts to increase, however, the performance of using 2 levels of decomposition catches up quickly and exceeds that of using 3 levels of decomposition as the bit rate increases. The db9 and bio6.8 provide the highest performance at high bit rates, and should be reasonable choices although the sym5 may give higher performance at the cost of the larger delay when the codec is operated at relatively low bit rates. Since the biorthogonal wavelet filters have the benefit of linear phase, the biorthogonal spline wavelet with order 6 and 8 for reconstruction and decomposition, respectively (bio6.8), was selected to be used for the higher-band signal.

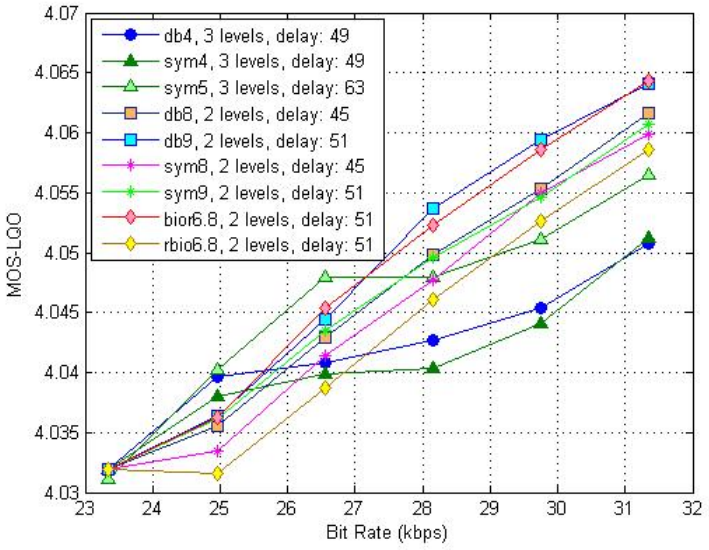


Figure 4.19: Performance comparison of the proposed codec W3 using various wavelets with different order and different number of decomposition levels for the higher-band signal.

4.3.3 Performance Evaluation

In order to evaluate the quality of speech produced by our proposed codec W3, the objective tests based on the PESQ algorithm and the informal subjective listening tests were performed. All the results were obtained for wideband input and wideband output.

Four different modes are used for performance evaluation and the bit allocation of each mode is presented in Table 4.5. For example, in Mode 1, 260, 33, 247, and 103 bits are allocated to Layer 1, 2, 3 and 4, respectively, and Layer 5 is not used. The frequency boundary between Layer 3 and Layer 4 is 2 kHz in Mode 1, and 1 kHz in Mode 2, 3, 4. Note that at least both Layer 1 and Layer 2 are required to encode wideband signals. It is also important to note that having many operating modes provides the ability to adjust the bit rate based on the requirements. In order to evaluate the benefit of using the WPT instead of the MDCT, the proposed codec W2 using the MDCT presented in Section 4.2 are also used for comparisons. Note that the bit allocations for the WPT and the MDCT are slightly different for each mode in order to optimize performance.

Table 4.5: Bit allocation of experimental modes for the proposed codec W3

	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
Mode 1	260	33	247	103	0
Mode 2	260	33	151	103	87
Mode 3	260	33	103	71	167
Mode 4	228	33	103	71	199

4.3.3.1 Objective Evaluation

Figure 4.20 shows the MOS-LQO scores obtained by the PESQ algorithm as a function of bit rates to compare the performances of the proposed codecs W3 and W2 using the WPT and the MDCT respectively with G.729.1 under clean channel condition. The performance of the WPT clearly exceeds that of the MDCT at all bit rates. The proposed codec W3 also outperforms G.729.1 at the bit rate of 18 kbps or higher. As explained in Section 4.2.2, the sudden drop of the codec performance below 18 kbps is mainly because the iLBC-based codec generally underperforms the CELP-based codec when operated at the same low bit rate as a narrowband codec. The slightly lower performance of the TDBWE decoder with a fixed excitation sequence also affected the MOS-LQO score at the low bit rates when only the Layer 1 and 2 were used. However, it is possible for the proposed codec W3 to achieve slightly better performance than G.729.1 at 16 kbps or lower using Layer 3 if the performance of the core-layer codec and the performance at high bit rates are sacrificed. Note that at 14.65 kbps the MOS-LQO score of the WPT is slightly better than that of the MDCT even though any transform coding is not used. This is because the difference in delays caused by the WPT and the MDCT affects the performance of the post-filter, however, we confirmed that the performance gain from the post-filter diminished or even became negative as the bit rate increases.

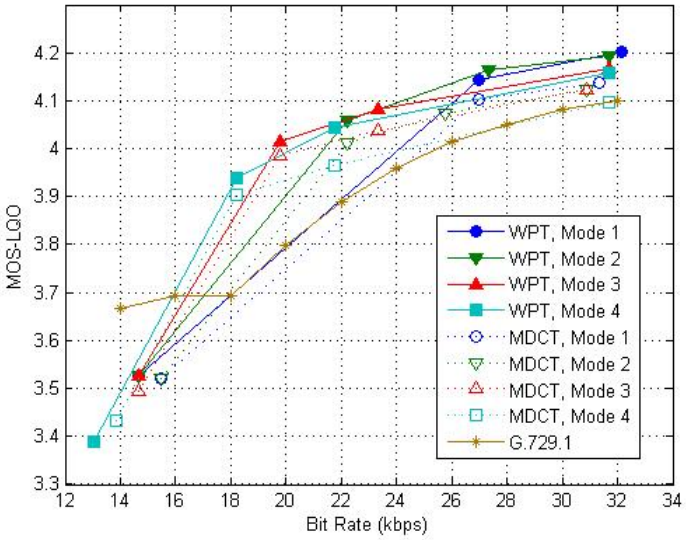


Figure 4.20: Performance comparisons of the proposed codecs W3 and W2 using the WPT and the MDCT respectively with G.729.1 under clean channel condition

Figure 4.21 shows the performance comparison of the proposed codec W3 operated at 31.7 kbps using Mode 2 and G.729.1 operated at 32 kbps under lossy channel conditions where the MOS-LQO scores are plotted as a function of packet loss rates. The proposed codec W3 outperforms G.729.1 at all packet loss rates. As explained in Section 4.3.1, both codecs employ basically the same PLC algorithm; however, the necessary modification causes the PLC performance of the proposed codec W3 to degrade compared to that of G.729.1. In addition, this PLC algorithm was designed to perform well with the frame erasure concealment (FEC) parameters that only G.729.1 can utilize. Therefore, we can see that the robustness to packet loss of the proposed codec W3 results from the frame-independent iLBC-based coding and higher performance can be expected for the proposed codec W3 by using the optimized PLC algorithm.

In Figure 4.22, the proposed codec W3 operated at 14.65 kbps using Mode 2 and G.729.1 operated at 16 kbps are compared under lossy channel conditions. Without packet loss, G.729.1 outperforms the proposed codec W3 by about 0.2 point in terms of MOS-LQO score; however, the proposed codec W3 performs better when the packet loss rates are higher than 5 %. Note that G.729.1 transmits the FEC parameters in Layer 2, 3, and 4, and the bit-stream at 16 kbps comprises Layer 1 to 4. Thus, when G.729.1 is

operated at 16 kbps, all the FEC parameters are transmitted. Since the proposed codec W3 and G.729.1 have the almost identical PLC algorithm for higher-band signals, it is obvious that the frame-independent iLBC-based coding can offer higher robustness to packet loss than the CELP-based coding with FEC parameters at the cost of higher bit rates. Therefore, it is worth considering the iLBC-based codec as a core-layer codec for VoIP applications.

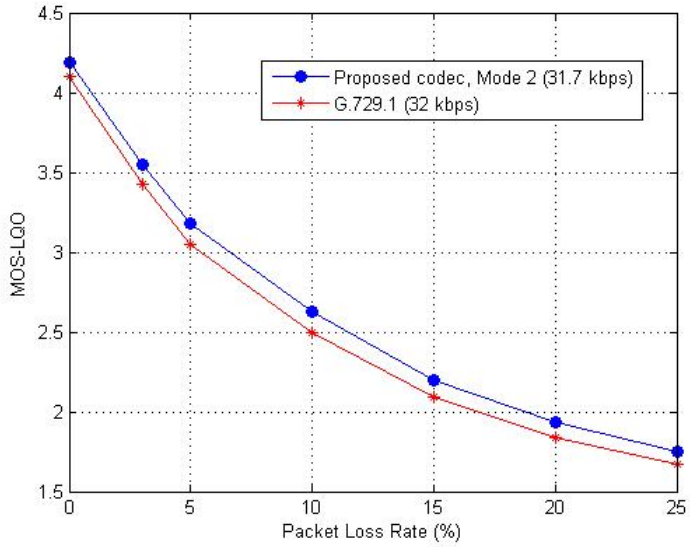


Figure 4.21: Performance comparison of the proposed codec W3 using Mode 2 at 31.7 kbps and G.729.1 at 32 kbps under lossy channel conditions

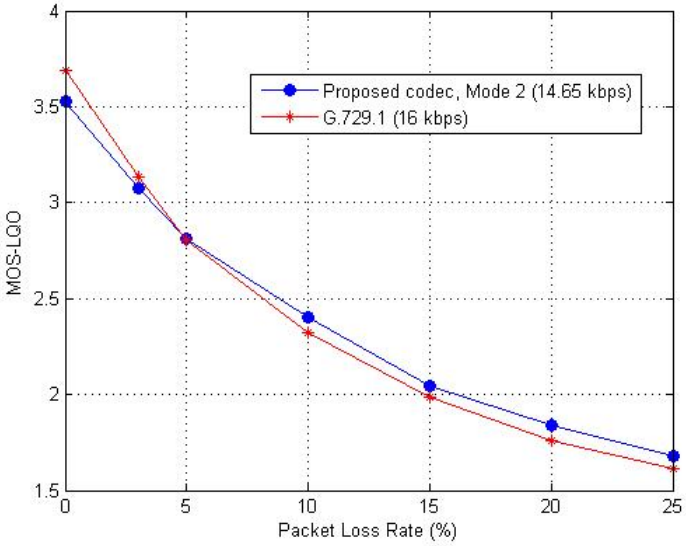


Figure 4.22: Performance comparison of the proposed codec W3 using Mode 2 at 14.65 kbps and G.729.1 at 16 kbps under lossy channel conditions.

4.3.3.2 Subjective Evaluation

In order to evaluate the subjective quality performance of the proposed codec W3, the two types of subjective listening tests: Informal mean opinion score (MOS) tests based on absolute category rating (ACR) method and A-B comparison tests based on comparison category rating (CCR) method [53] were performed. Each test was conducted in American English with untrained listeners using binaural headphones. The test samples consisted of four sentence pairs spoken by 2 male and 2 female speakers.

Figure 4.23 shows the MOS scores as a function of bit rate to compare the performance of the proposed codecs W3 and W2 using the WPT (Mode 2) and the MDCT (Mode 2) respectively, and G.729.1. The error bars represent the 95 % confidence intervals. These subjective test results are presented only to show the trend of the subjective quality of speech because the MOS scores are obtained from limited informal tests and the 95 % confidence intervals are large. Although all three codecs have similar performance, we can see the trend that G.729.1 performs slightly better than the proposed codecs W3 and W2 at around 16 kbps and the proposed codecs W3 and W2 outperform G.729.1 at around 32 kbps. This trend matches with the objective test results. There are

also mismatches between the objective quality and the subjective quality. For example, the MOS score of the WPT is lower than that of the MDCT at around 32 kbps, and the MOS score of G.729.1 is equivalent or higher than those of the proposed codecs W3 and W2 at around 22 kbps although the objective test results indicate the contrary. These mismatches probably result from the limited number of informal subjective tests.

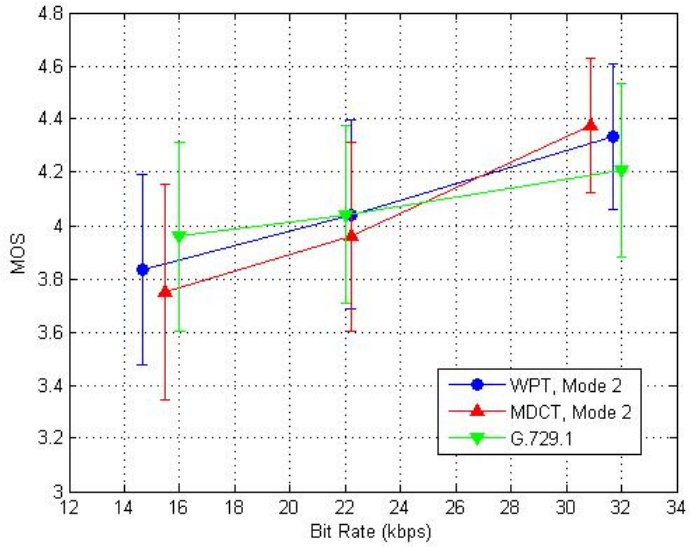


Figure 4.23: Subjective test results for the proposed codecs W3 and W2 using the WPT and the MDCT respectively, and G.729.1

In order to reliably compare the subjective quality of these codes, A-B comparison tests were performed. Figure 4.24 shows the results of the A-B comparison tests when the proposed codec W3 with the WPT using Mode 2 (A) is compared with the proposed codec W2 with the MDCT using Mode 2 (B) when operated at around 22 kbps and 32 kbps. The following five categories were used to compare the subjective quality of "A" and "B":

(1) A is better than B

(2) A is slightly better than B

(3) A is about the same as B

(4) B is slightly better than A

(5) B is better than A

Therefore, from Figure 4.24, we can tell what percentage each category occupies. It is clear that the WPT performs slightly better than the MDCT at around 22 kbps and 32 kbps. These results match with the objective test results. Figure 4.25 illustrates the A-B comparison test results when the proposed codec W3 using Mode 2 and G.729.1 are compared at the bit rate of around 16 kbps, 22 kbps, and 32 kbps. The performance of G.729.1 is slightly better than that of the proposed codec W3 at around 16 kbps, whereas the proposed codec W3 slightly outperforms G.729.1 at around 22 kbps and 32 kbps. These results correspond to the performances of the objective quality.

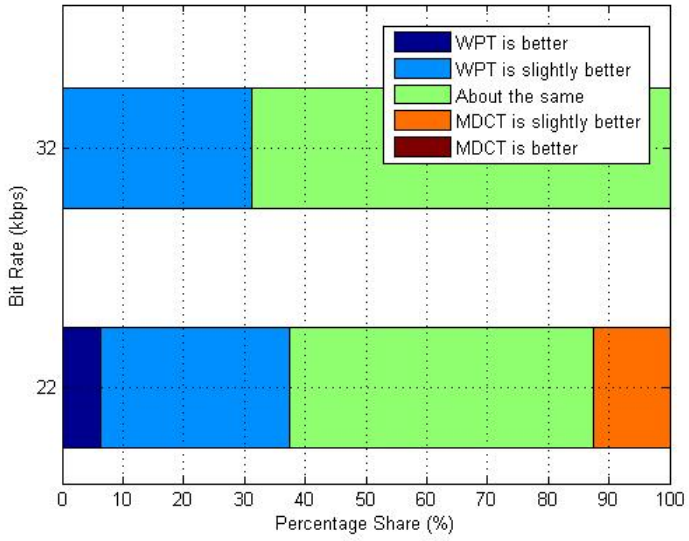


Figure 4.24: A-B comparison test results for the proposed codec W3 using the WPT vs. the proposed codec W2 using the MDCT

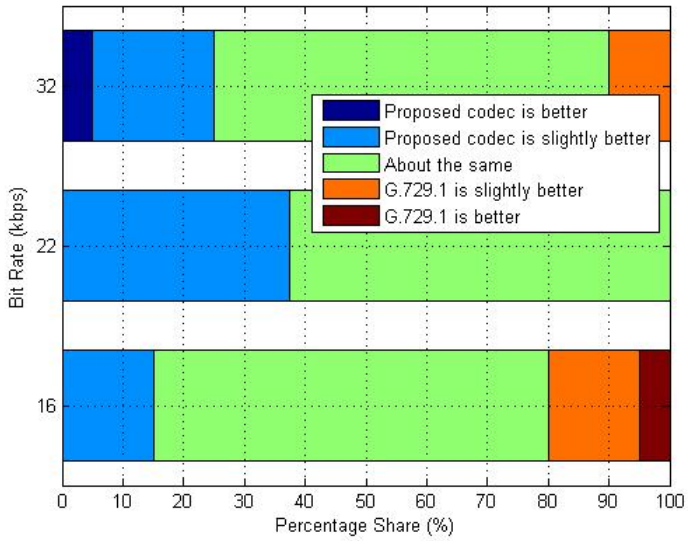


Figure 4.25: A-B comparison test results for the proposed codec W3 vs. G.729.1

Figure 4.26 shows the results of the A-B comparison tests when the proposed codec W3 operated at 31.7 kbps using Mode 2 is compared with G.729.1 operated at 32 kbps for different packet loss rates. The proposed codec W3 performs slightly better without packet loss and at all packet loss rates, which matches with the objective test results. Figure 4.27 depicts the A-B comparison test results when the proposed codec W3 operated at 14.65 kbps using Mode 2 and G.729.1 operated at 16 kbps are compared for different packet loss rates. Without packet loss, G.729.1 performs slightly better the proposed codec W3. The proposed codec W3 slightly outperforms G.729.1 at all packet loss rates except that the performances are about the same at the packet loss rate of 5 %. This trend almost matches the performance of the objective quality except for the case of 3 % packet loss.

It is found from the objective and subjective evaluation results that the MOS-LQO scores appear to provide a good estimate of the actual MOS scores in these tests, contrary to some cases in Section 3.2.2.

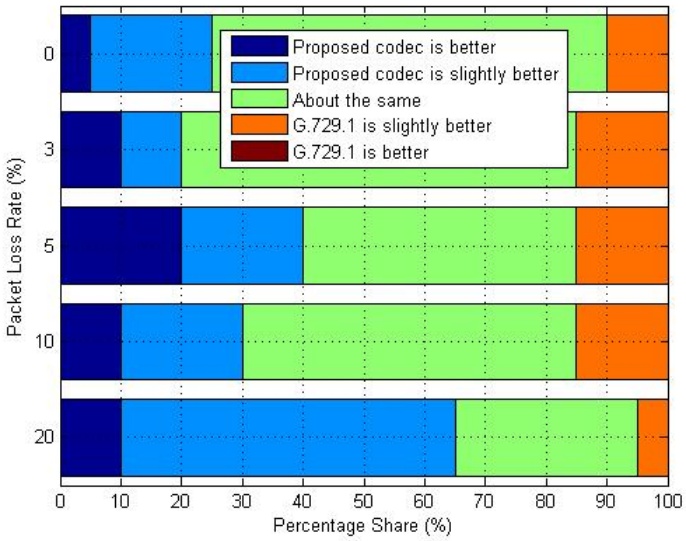


Figure 4.26: A-B comparison test results for the proposed codec W3 at 31.7 kbps vs. G.729.1 at 32 kbps under lossy channel conditions

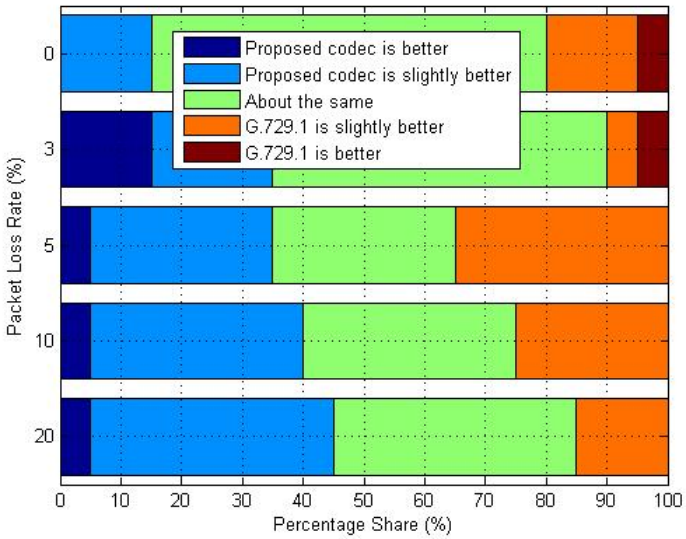


Figure 4.27: A-B comparison test results for the proposed codec W3 at 14.65 kbps vs. G.729.1 at 16 kbps under lossy channel conditions

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this dissertation, we developed novel scalable narrowband and wideband speech codecs for IP networks using the frame independent coding scheme based on the iLBC. The performance evaluation results showed that the developed codecs provided high robustness to packet loss and achieved equivalent or higher performance than the state-of-the-art codecs.

The iLBC features the frame-independent coding and therefore inherently possesses high robustness to packet loss. However, the original iLBC is a narrowband fixed bit-rate codec, and thus lacks in some of the key features of speech codecs for IP networks: Rate flexibility, Scalability, and Wideband support.

These missing functionalities were added to the original iLBC. In particular, the rate flexibility was added to the iLBC by employing the discrete cosine transform (DCT) and the scalable algebraic vector quantization (AVQ) and by allocating different number of bits to the AVQ. The bit-rate scalability was obtained by adding the enhancement layer to the core layer of the multi-rate iLBC. The enhancement layer encodes the weighted iLBC coding error in the modified DCT (MDCT) domain. The proposed wideband codec employed the bandwidth extension technique to extend the capabilities of existing narrowband codecs to provide wideband coding functionality. The wavelet transform was also used to further enhance the performance of the proposed codec.

The developed codecs achieved the high performance under both clean channel and lossy channel conditions. These are remarkable results considering that the original iLBC is designed to achieve high robustness to packet loss at the expense of the high bit rates.

Therefore, it is worth considering the iLBC-based codec as a core-layer codec for VoIP applications.

5.2 Future Work

We employed bandwidth scalability to support wideband speech signals when a scalable wideband speech codec based on the iLBC was developed. Therefore, the scalable extension of bandwidth to support super-wideband (50–14000 Hz) and fullband (20–20000 Hz) audio is the natural next step for further research. On the other hand, a speech codec generally achieves higher performance when it is designed and optimized specifically for encoding wideband speech signals. Thus, higher performance can be expected by re-designing the iLBC specifically for wideband speech signals, which is worth researching.

Another research direction could be to improve and optimize the PLC algorithm for the iLBC-based codec. The proposed codecs in this work used the PLC algorithm specified in G.729.1. All the missing parameters for the PLC algorithm are estimated in the decoder. Thus theoretically we should be able to improve performance under lossy channel conditions by optimizing the PLC algorithm specifically for the iLBC-based codec.

List of Publications Related to Thesis

Journals

K. Seto and T. Ogunfunmi, "Scalable Speech Coding for IP Networks: Beyond iLBC," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 11, pp. 2337-2345, 2013.

K. Seto and T. Ogunfunmi, "A Scalable Wideband Speech Codec Based on the iLBC," submitted to IEEE Transactions on Audio, Speech, and Language Processing, 2015.

Book Chapter

T. Ogunfunmi and K. Seto, "Scalable and Multi-Rate Speech Coding for Voice-over-Internet Protocol (VoIP) Networks," Chapter 3 in T. Ogunfunmi, R. Togneri and M.J. Narasimha, editors, Advances in Speech and Audio Processing for Coding, Enhancement and Recognition, Springer, 2015.

Conference Proceedings

K. Seto and T. Ogunfunmi, "Multi-rate iLBC using the DCT," Proceedings of the IEEE workshop on SiPS, pp. 478–482, 2010.

K. Seto and T. Ogunfunmi, "Performance Enhanced Multi-Rate iLBC," Proceedings of the 45th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, pp. 590–594, 2011.

K. Seto and T. Ogunfunmi, "Scalable Multi-Rate iLBC," Proceedings of IEEE International Symposium on Circuits and Systems, pp. 1034–1037, 2012.

K. Seto and T. Ogunfunmi, "Scalable Wideband Speech Coding for IP Networks," Proceedings of the 46th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, pp. 77–81, Nov. 2012.

K. Seto and T. Ogunfunmi, “Packet-Loss Robust Scalable Speech Coding Using the Discrete Wavelet Transform,” Proceedings of IEEE International Symposium on Circuits and Systems, pp. 129–132, 2014.

K. Seto and T. Ogunfunmi, “Performance Enhanced Scalable Wideband Speech Coding for IP Networks,” Proceedings of the 48th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, pp. 833–837, Nov. 2014.

Bibliography

[1] T. Ogunfunmi and M.J.Narasimha, “Speech over VoIP Networks: Advanced Signal Processing and System Implementation,” IEEE Circuits and Systems Magazin, Vol. 12, no. 2, pp. 35 – 55, 2012.

[2] FCC, Meeting presentation of the Technological Advisory Council, https://transition.fcc.gov/oet/tac/TACMarch2011mtgfullpresentation.pdf, March 30, 2011.

[3] FCC, Meeting presentation of the Technological Advisory Council, https://transition.fcc.gov/oet/tac/TACJune2011mtgfullpresentation.pdf, June 29, 2011.

[4] AT&T, “Petition to launch a proceeding concerning the TDM-to-IP transition,” FCC filing, http://www.att.com/Common/about_us/files/pdf/fcc_filing.pdf, Nov. 7, 2012.

[5] AT&T, “AT&T Proposal for Wire Center Trials,” FCC filing, http://apps.fcc.gov/ecfs/document/view?id=7521090526, Feb. 27, 2014.

[6] Ekudden, E., R. Hagen, I. Johansson, and J. Svedberg, ”The adaptive multi-rate speech coder” Proceedings of IEEE Speech Coding Workshop, pp. 117-119, 1999.

[7] R. Salami, et al., ”Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder” IEEE Transactions on Speech and Audio Processing, March 1998.

[8] M. R. Schroeder and B. S. Atal, “Code-excited linear prediction (CELP): High-quality speech at very low bit rates,” Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 937–940, 1984.

[9] R. Lefebvre, P. Gournay, and R. Salami, “A study of design compromises for speech coders in packet networks,” Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing, vol. I, pp. 265–268, 2004.

[10] V. Eksler and M. Jelinek, “Glottal-shape codebook to improve robust-ness of CELP codecs,” IEEE Transactions on Audio, Speech and Language Processing., vol. 18, no. 6, pp. 1208–1217, Aug. 2010.

[11] T. Vaillancourt *et al*., “ ITU-T G.EV-VBR: A Robust 8–32 kb/s Scalable Coder for Error Prone Telecommunications Channels,” Proceedings of the Eusipco, Lausanne, Switzerland, Aug. 2008.

[12] J.-M. Valin, K. Vos, and T. Terriberry, Internet Engineering Task Force RFC6716, Sep. 2012.

[13] S.V. Andersen, W.B. Kleijn, R. Hagen, J. Linden, M.N. Murthi, and J. Skoglund, ”iLBC-A Linear Predictive Coder with Robustness to Packet Losses,” In 2002 IEEE Speech Coding Workshop Proceedings, pp.23-25.

[14] T. Ogunfunmi and M.J.Narasimha, Principles of Speech Coding, CRC Publishers, 2010.

[15] T. Ogunfunmi, R. Togneri and M.J. Narasimha, Advances in Speech and Audio Processing for Coding, Enhancement and Recognition, Springer, 2015.

[16] B. Geiser et al. “Embedded Speech Coding: From G.711 to G.729.1,” Chapter 8 in Advances in Digital Speech Transmission, Wiley, Jan. 2008.

[17] K. Brandenburg, O. Kunz, and A. Sugiyama, “MPEG-4 natural audio coding,” Signal Processing: Image Communication, Vol. 15, pp. 423–444, 2000.

[18] S. Ragot, B. Kovesi, R. Trilling, D. Virette, N. Duc, D. Massaloux, S. Proust, B. Geiser, M. Gartner, S. Schandl, H. Taddei, Y. Gao, E. Shlomot, H. Ehara, K. Yoshida, T. Vaillancourt, R. Salami, M. S. Lee, and D. Y. Kim, “ITU-T G.729.1: An 8–32 kb/s scalable coder interoperable with G.729 for wideband telephony and voice over IP,” Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 529–532, 2007.

[19] C. M. Garrido, M. N. Murthi, and S. V. Andersen, “Towards iLBC Speech Coding at Lower Rates through A New Formulation of The Start State Search,” Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing, Vol. 1, pp. 769 - 772, March 2005.

[20] C. M. Garrido, M.N. Murthi, and S.V. Andersen, “On variable rate frame independent predictive speech coding: Re-engineering iLBC,” Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing, Vol. 1, pp. 717–720, 2006.

[21] K. Seto and T. Ogunfunmi, “Multi-rate iLBC using the DCT,” Proceedings of the IEEE workshop on SiPS, pp. 478–482, 2010.

[22] K. Seto and T. Ogunfunmi, “Performance Enhanced Multi-Rate iLBC,” Proceedings of the 45th Asilomar Conference, pp. 590–594, 2011.

[23] K. Seto and T. Ogunfunmi, “Scalable Multi-Rate iLBC,” Proceedings of IEEE International Symposium on Circuits and Systems, pp. 1034–1037, 2012.

[24] K. Seto and T. Ogunfunmi, “Scalable Speech Coding for IP Networks: Beyond iLBC,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 11, pp. 2337-2345, 2013.

[25] K. Seto and T. Ogunfunmi, “Packet-Loss Robust Scalable Speech Coding Using the Discrete Wavelet Transform,” Proceedings of IEEE International Symposium on Circuits and Systems, 2014.

[26] K. Seto and T. Ogunfunmi, “Scalable Wideband Speech Coding for IP Networks,” Proceedings of the 46th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, pp. 77–81, Nov. 2012.

[27] K. Seto and T. Ogunfunmi, “Performance Enhanced Scalable Wideband Speech Coding for IP Networks,” Proceedings of the 48th Annual Asilomar Conference on Signals, Systems, and Computers, pp. 833–837, Pacific Grove, Nov. 2014.

[28] K. Seto and T. Ogunfunmi, “A Scalable Wideband Speech Codec Based on the iLBC,” submitted to IEEE Transactions on Audio, Speech, and Language Processing.

[29] L. Laaksonen, M. Tammi, V. Malenovsky, T. Vaillancourt, M. S. Lee, T. Yamanashi et al., “Superwideband Extension of G.718 and G.729.1 Speech Codecs,” Proceedings of Interspeech, Tokyo, Japan, 2010.

[30] M. Xie, P. Chu, A. Taleb and M. Briand, “ITU-T G.719: A New Low-Complexity Full-Band (20 KHz) Audio Coding Standard for High-Quality Conversational Applications,” 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, October 2009.

[31] W. C. Chu, Speech Coding Algorithms: Foundation and Evolution of Standardized Coders, Hoboken, NJ: Wiley-Interscience, 2003.

[32] M. Bosi and R. Goldberg, Introduction to Digital Audio Coding and Standards, Norwell, MA: Kluwer, 2002.

[33] ETSI, Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) Speech Transcoding, GSM 06.90, version 7.1.1, Release 1998, Dec. 1999.

[34] 3GPP TS 26.090 V 3.0.0, Release 1999, Mandatory Speech Codec Speech Processing Functions; AMR Speech Codec; Transcoding Fucntions, June 1999.

[35] B. Bessette et al., “The Adaptive Multirate Wideband Speech Codec (AMR-WB),” IEEE Trans. on Speech and Audio Processing, vol. 10, no. 8, pp. 620–636, Nov. 2002.

[36] 3GPP TS 26.190 V 5.0.0, Release 5, Speech Codec Speech Processing Functions; AMR Wideband speech codec; Transcoding functions, Mar. 2001.

[37] K. Järvinen et al., "Media Coding for the Next Generation Mobile System LTE," Elsevier Computer Communications, vol. 33, no. 16, pp. 1916-1927, 2010.

[38] M. Dietz et al., "Overview of the EVS Codec Architecture," Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 5698–5702, Apr. 2015.

[39] S. Bruhn et al., "Standardization of the New 3GPP EVS Codec," Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 5703–5707, Apr. 2015.

[40] 3GPP TS 26.445 V12.0.0, Release 12, Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description, Sep. 2014.

[41] Y. Hiwasaki, T. Mori, H. Ohmuro, J. Ikedo, D. Tokumoto, and A. Kataoka, "Scalable Speech Coding Technology for High-Quality Ubiquitous Communications," NTT Technical Review, vol. 2, no. 3, pp. 53–58, Mar. 2004.

[42] Y. Hiwasaki et al., "G.711.1: A Wideband Extension to ITU-T G.711," Proceedings of the EUSIPCO, Lausanne, Switzerland, Aug. 2008.

[43] S. Andersen et al., "Internet Low Bit Rate Codec (iLBC)," RFC3951, IETF, Dec. 2004. Available Online: https://tools.ietf.org/html/rfc3951

[44] A. Duric et al., "Real-time Transport Protocol (RTP) Payload Format for internet Low Bit Rate Codec (iLBC) Speech," RFC3952, IETF, Dec. 2004. Available Online: https://tools.ietf.org/html/rfc3952

[45] CableLabs, PacketCable™ 1.5 Specifications, Audio/Video Codecs, Apr. 2012.

[46] http://www.webrtc.org/license-rights/ilbc-freeware

[47] O. Rioul and M. Vetterli, "Wavelets and signal processing," IEEE Signal Processing Mag., pp. 14–38, Oct. 1991.

[48] I. Daubechies, Ten Lectures on Wavelets, Philadelphia, PA: SIAM, 1992.

[49] G. Strang and T. Nguyen, Wavelet and Filter Banks, Wellesley-Cambridge Press, 1996.

[50] M. Vetterli and J. Kovačević, Wavelets and Subband Coding, Englewood Cliffs, NJ: Prentice-Hall, 1995.

[51] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, “Signal processing and compression with wavelet packets”, Technical report, Dept. of Math., Yale University, 1991.

[52] R. R. Coifman, Y. Meyer, and M. V. Wickerhauser, “Wavelet analysis and signal processing,” in M. B. Ruskai et al, editor, Wavelets and their Applications, pages 153–178, Jones and Barlett, Boston, 1992.

[53] ITU-T P.800 ” Methods for subjective determination of transmission quality.”

[54] ITU-T P.862 ”Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.”

[55] ITU-T P.863 ”Perceptual objective listening quality assessment.”

[56] ITU-T P.501 ”Test signals for use in telephonometry.”

[57] ITU-T G.191 (2010) “Software tools for speech and audio coding standardization.”

[58] E.N Gilbert. Capacity of a burst-noise channel. Bell Syst.Tech.J., pages 1253–1265, 1960.

[59] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, “AMR-WB+: a new audio coding standard for 3rd generation mobile audio services,” Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing, vol. 2, pp. 1109–1112, March 2005.

[60] S. Ragot, B. Bessette and R. Lefebvre, “Low-complexity multi-rate lattice vector quantization with application to wideband speech coding at 32 kbit/s,” Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing, Vol. 1, pp. 501-504, May 2004.

[61] G.D. Forney, “Coset codes. I. Introduction and geometrical classification,” IEEE Trans. on Information Theory, Sep 1988 Vol. 34 , no. 5, pp. 1123 – 1151.

[62] F. Chen and K. Kuo, “Complexity Scalability Design in the Internet Low Bit Rate Codec (iLBC) for Speech Coding,” IEICE Transactions on Information and Systems Vol.E93-D No.5 pp.1238-1243, May 2010.

[63] J. Princen and A. Bradley, “Analysis/synthesis filter bank design based on time domain aliasing cancellation,” IEEE Trans. Acoust., Speech, Signal Process., vol. 34, no. 5, pp. 1153–1161, Oct. 1986.

[64] I. Daubechies, “Orthonormal bases of compactly supported wavelets,” Commun. Pure Appl. Math., vol. 41, pp. 909-996, 1988.

[65] ITU-T Rec. G.711 Appendix I, “A high quality low-complexity algorithm for packet loss concealment with G.711,” Sep. 1999.