uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Libin Cai

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Master of Computer Science

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Speech quality evaluation using digital watermarking

TITRE DE LA THÈSE / TITLE OF THESIS

Jiying Zhao

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Abdulmotaleb El Saddik

Chung-Horng Lung

Gary W. Slater

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES /
DEAN OF THE FACULTY OF GRADUATE AND POSTDOCORAL STUDIES

Speech quality evaluation using digital watermarking

by

Libin Cai

A thesis submitted to
the Faculty of Graduate and Postgraduate Studies
in partial fulfillment of
the requirements for the degree of

Master of Computer Science

Ottawa-Carleton Institute for Computer Science
School of Information Technology and Engineering
University of Ottawa

Ottawa, Ontario, Canada

Your file Votre référence
ISBN: 0-494-14888-8
Our file Notre référence
ISBN: 0-494-14888-8

Abstract

Speech quality evaluation is a very important research topic. The Mean Opinion Score (MOS) is reliable but the listening test is very expensive, time consuming, and even impractical for some applications. Objective quality evaluation methods require either the original speech or a complicated computation model, which makes some applications of quality evaluation impossible.

Different from the perceptual model used by the Perceptual Evaluation of Speech Quality (PESQ), in this thesis, we propose to use digital audio watermarking to evaluate the quality of speech. Based on quantization, watermark bits are embedded and extracted in the Discrete Wavelet Transform (DWT) domain. By comparing the original and the extracted watermark, we predict the quality of speech that has undergone MP3 compression, Gaussian noise addition, low-pass filtering, or packet loss. Our quality evaluation method does not need the original signal or a computation model.

For the quality evaluation, we use the PESQ MOS as a reference. We predict the speech quality from the PCEW (Percentage of Correctly Extracted Watermark bits) based on the mapping between ITU-T P.862 PESQ MOS and the PCEW. To evaluate the performance of our objective quality evaluation method, we introduce the correlation coefficient and residual error to evaluate the correlation between the predicted MOS and the PESQ MOS. The experiments show that the method yields very promising evaluation results which are very close to the results of the PESQ.

Contents

List of Tables

List of Figures

List of Acronyms

DFT	Discrete Fourier Transform
IDFT	Inverse Discrete Fourier Transform
HAS	Human Auditory System
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
IDWT	Inverse Discrete Wavelet Transform
LSB	Least Significant Bit
PN	Pseudo Noise
PSNR	Peak Signal to Noise Ratio
SNR	Signal to Noise Ratio
MOS	Mean Opinion Score
PCEW	Percentage of Correctly Extracted Watermark
PSQM	Perceptual Speech Quality Measure
PAQM	Perceptual Audio Quality Measure
PESQ	Perceptual Evaluation of Speech Quality
GMM	Gaussian Mixture Models
QoS	Quality of Service

Acknowledgement

I would like to deeply thank my supervisor, Professor Jiying Zhao, for bringing me to the interesting research topic: digital watermarking. I really appreciate his valuable guidance, strong support and prompt feedback during my work.

I would also like to thank the members of the Multimedia Communications Research Laboratory for their friendship and suggestions.

Finally, I would like to thank my family for their continuous support in these years.

Chapter 1

Introduction

1.1 Background

The evaluation of audio and speech quality is of critical importance in today's computer network control, e-commerce and telephony networks. The main reason is that the quality is a key determinant of customer satisfaction and key indication of computer network condition. Traditionally, the only way to measure the perception of quality of a speech signal was through the use of subjective testing [4], in which the average of these scores is the subjective MOS (Mean Opinion Score). This is the most reliable method of speech quality assessment but it is highly unsuitable for online monitoring applications and is also fairly expensive, time-consuming, and labor-intensive. Due to these reasons, objective methods have been developed in recent years, classified into two categories: signal based methods and parameters based methods [5]. Signal based methods use the reference and degraded signals as the input to the measurement, such as the state-of-the-art objective measurement algorithm, PESQ (perceptual evaluation of speech quality) [1]. Meanwhile, parameters based methods predict the speech quality through

a computational model instead of using real measurement. For example, Falk and Chan [4] proposed an approach to objective speech quality measurements using Gaussian mixture models (GMMs). This category of methods need a large training database to construct good estimators of subjective listening quality, and different training database may result in different model.

The Quality of Service (QoS) of IP telephony networks has been discussed in terms of both subjective quality and objective quality. The subjective quality corresponds to users perceptions of transmitted speech, while the objective quality assessment has been developed by measuring the physical characteristics of transmitted speech. So far, the objective methods must employ either the original speech or complex training to determine the strength of distortion. However, it is often impossible to obtain original speech for use in objective quality assessment for in-service testing. Therefore, it is desirable to develop a method that uses only degraded speech to estimate speech quality [6].

Digital watermarking technology has been around for more than ten years, which has been used in copyright protection, content authentication, copy control, broadcast monitoring, etc. However, for our best knowledge, so far, there is no application of digital watermarking oriented for audio or speech quality evaluation.

1.2 Contributions

In this thesis, we propose a new application of digital watermarking, speech quality evaluation. The basis of the method is that the carefully embedded watermark in a speech will suffer the same distortions as the speech does. The proposed method needs neither original speech, nor training database. Using PESQ as a reference, the

experimental results show that the proposed method gives very accurate quality evaluation that the correlation between $DWMOS$ (Predicted MOS score in our method) and MOS are 0.9759, 0.9727, 0.8493 and 0.9744 for the effects of Gaussian noise, MP3 compression, Low-pass filtering and Packet loss respectively. Furthermore, without the complicated signal processing on both original and degraded speeches, such as time alignment, equalization and FFT filtering, the implementation of this quality assessment is very fast. For example, for a one-minute speech (16KHz, 16-bit, Mono), the PESQ needs an average of 29 seconds to calculate its MOS score, while our method only takes 9 seconds. Hence, it well satisfies the strict time requirement of real time quality assessment systems.

Publications generated from the research:

Refereed journals

1. Libin Cai and Jiying Zhao, Evaluation of speech quality using digital watermarking, IEICE Electronics Express (ELEX), Vol.1, No.13, pp. 380-385, October 2004.

Refereed proceedings

1. Libin Cai and Jiying Zhao, Speech quality evaluation: a new application of digital watermarking, IEEE Instrumentation and Measurement Technology Conference (IMTC) 2005, Ottawa, Ontario, Canada, pp. 726-731, May 17-19, 2005

2. Libin Cai and Jiying Zhao, Speech quality assessment using digital watermarking, Proceedings of HAVE2004 - IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications, Ottawa, Ontario, Canada, October 2–3, pp. 177-182, 2004.

3. Libin Cai and Jiying Zhao, Audio quality measurement by using digital watermarking, Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) 2004, Niagara Falls, Ontario, Canada, pp. 1159-1162, May 2–5, 2004.

1.3 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2, we introduce the basic knowledge and background of speech and digital watermarking. In Chapter 3, we briefly review the audio watermarking technology and speech quality assessment. In Chapter 4, we describe the proposed algorithm and our implementation strategies. In Chapter 5, we show our experimental results and evaluation. Lastly in Chapter 6, we make conclusions and give suggestions for future works.

Chapter 2

Overview of Speech Acoustics and Digital Watermarking

2.1 Speech Acoustics

Speech is a form of measures that human beings use to communicate intention and emotion. It is convenient and natural [7]. According to the perspective of acoustics, speech is the sound perceived by the auditory system. Sound is variations in air pressure detectable by human ears. The motive power for the human speech system is provided by the lungs which force air out through the vocal tract. This consists of the larynx, pharynx, nasal and oral cavities, and the lips. Within the larynx are the vocal folds. The rate at which the vocal folds vibrate controls the pitch of the sound, which may vary from 70 Hz for an average low male voice to 1000 Hz for a high pitched female voice [8].

The pharynx, oral and nasal cavities act as resonant chambers, concentrating the sound energy around particular frequencies. These concentrations of the sound energy

are generally known as formants. When speech is produced, the articulators move and change the formant frequencies and the amount of energy which is associated with each formant. This is perceived as a change in the speech sound, In addition to the voiced sounds, a number of unvoiced sounds are also produced. Other important features are stops and plosives which are produced by suddenly closing off the air flow in some part of the vocal tract, or suddenly releasing a constricted airflow respectively. This is called Acoustic-Phonetics.

Based on Acoustic-Phonetics theory, phoneticians and linguists decompose a spoken language into elements of linguistically distinctive sounds - the phonemes. Phonemes are determined and taxonomically classified according to their corresponding articulatory configurations [9]. The phoneme categories include diphthongs, semivowels, nasals, stops, fricatives, affricates, and whisper. As in many classical studies, the taxonomy was established for a systematic investigation of the properties of the "element" of speech sounds. Such properties of sounds are often referred to as acoustic-phonetic features. An alternative way to classify the phonemes is to use the broad phonetic class according to key acoustic-phonetic feature dimensions.

2.1.1 Normal Distortions to Speech

Taking account of transmission, the distortion factors include environmental noise, sidetone, talker echo, frequency loss, circuit noise, transmission errors (random bit errors and erased frames), non-linear distortion, propagation time delay, harmful effects of voice-operated devices, distortions of the time scale arising from packet switching, time-varying degradations of the communication channel, etc. Combinations of two or more of such factors also have to be considered.

Sidetone means the sound of the speaker's own voice (and background noise) as heard in the speaker's telephone receiver. Its volume is usually suppressed relative to the transmitted volume. Talker echo occurs when a talker's speech energy, transmitted down the primary signal path, is coupled into the receive path from the far end. The talker then hears his/her own voice, delayed by the total echo path delay time. If the 'echoed' signal has sufficient amplitude and delay, the result can be annoying to the talker and the listener, and interfere with the normal speech process. Talker echo is a direct result of the 2-wire to 4-wire conversion that takes place through 'hybrid' transformers. The frequency loss usually occurs on the high frequency. The low-pass filter is a classic frequency loss distortion on the speech. Circuit noise is a fundamental phenomenon in electronic circuits caused by the small fluctuations in currents and voltages that occur within the devices in the circuit. The fluctuations are due mainly to the discontinuous nature of electric charge. Determining the effects of noise is very important, as noise often represents the fundamental limit of circuit or system performance. Nonlinear distortion is caused by a deviation from a linear relationship between specified input and output parameters of a system or component. The propagation time delay is the time required for a signal to travel from one point to another.

2.1.2 Speech Quality Measurement

There are two main classes of speech quality metrics: subjective and objective [10]. At the very beginning, the traditional engineering tools, such as signal to noise ratio that is often abbreviated to SNR or S/N, were used to assess the speech quality. Nevertheless, they are clearly insufficient in predicting subjective quality. As a result, subjective listening tests have traditionally been necessary to quantify results.

Subjective measures involve humans listening to a live or recorded conversation and

assigning a rating to it. The MOS is such a useful metric. Although MOS is not the only subjective measure, it is one of the most widely used and recognized methods. The MOS defines a method to derive a mean opinion score of voice quality after collecting scores between 1 (bad) and 5 (excellent) from human listeners. It consists of n subjects listening to specific signals in order to rate their quality. Those subjects, named "Gold Ear", are trained to build a mapping between a 5-level-quality-scale and a set of processed speech signals. In the test laboratory, they are seated in separate sound-proof cabinets near the point from which the experiment is controlled. The test room and cabinets are favorably decorated to recreate a natural environment. The ambient noise level is kept as low as possible. Environmental noise is fed in with the required spectrum at the required level (e.g. 50 DBA). DBA is the abbreviation of Adjusted Decibel. The measured noise levels in dB will not reflect the actual human perception of the loudness of the noise. A specific circuit is added to the sound level meter to correct its reading in regard to this concept. This reading is the noise level in dBA. The letter A is added to indicate the correction that was made in the measurement. It is essential to ensure that the loudspeakers and amplifiers are capable of faithfully reproducing the required noise. Furthermore, the room noise and internal vehicle noise are also simulated at a proper spectral density.

For the quality measurement, the five judgement scales are most frequently used for ITU-T applications. The subjects allocate the following values to the scores: Excellent = 5, Good = 4, Fair = 3, Poor = 2, and Bad = 1. The arithmetic mean of any collection of these opinion scores is called the mean conversation-opinion score, and is represented by the symbol MOS. In each subjective experiment, the MOS scores may differ, even under the same condition. Clearly, a metric such as MOS that uses human subjects can be a good measure of perceived speech quality. However, since human listeners are

involved, in particular, such subjective testing is very expensive and time-consuming. Some researchers or organizations may not have the resources to conduct the tests. Certainly, such metrics cannot be used in any sort of real-time or online application. These shortcomings, as well as other reasons, have led to the development of objective metrics.

It became evident that objective measuring techniques that can predict the outcome of these listening tests would be useful. Over approximately the last 20 years, several researchers have addressed this issue [1][11][12][13][14]. These techniques typically use auditory models to either predict a threshold of audible distortion, or establish a metric to measure the difference between a reference signal and a test signal [1][12][14]. The goal of this thesis is to find a speech quality metric that accurately predicts human perception under conditions, such as Gaussian noise and packet loss, etc. In Section 3.2, we will discuss the objective speech quality measurement more in detail.

2.2 Digital Watermarking

2.2.1 Introduction

Digital watermarking describes methods and technologies that allow to hide information, for example a number or text, in digital media, such as images, video, audio, and other cover works. The embedding takes place by manipulating the content of the digital data.

Although mostly, the hiding process has to make the modifications of the media imperceptible, the watermark can be classified into two sub-types: visible watermark and invisible watermark. Visible watermarks change the signal altogether such that the watermarked signal is totally different from the actual signal, e.g., adding an image as

a watermark to another image. Stock photography agencies often add a watermark in the shape of a copyright symbol to previews of their images, so that the previews do not substitute for high-quality copies of the product included with a license. Meanwhile, invisible watermarks do not change the signal to a perceptually great extent, i.e., there are only minor variations in the output signal. For images this means that the modifications of the pixel values have to be invisible. An example of an invisible watermark is that some bits added to an image only modify its least significant bits.

Furthermore, the watermark has to be robust or fragile, depending on the applications. For robustness, we refer to the capability of the watermark to resist manipulations of the media, such as lossy compression, scaling, and cropping, just to enumerate some. Fragility means that the watermark should not resist tampering, or only up to a certain extent.

There are various spatial and frequency domain techniques used for adding watermarks to and removing them from signals. Purely spatial techniques are not robust to some attacks to the signal like cropping and zooming, whereas most frequency domain techniques and mixed-domain techniques are quite robust to such attacks. Watermarking system includes watermark embedder and detector. Regardless of whether it is informed detector or blind detector, basically, the embedding process consists of two steps. Firstly, the watermark is mapped into an added pattern which is the same size as cover work. Then, the pattern is added to the cover work to produce the watermarked work. For the watermark detection, if it is informed, the original cover work will be subtracted from the received work to obtain the noisy watermark pattern. It is then decoded by watermark decoder with a watermark key. On the other hand, for the blind detector, the received watermark work is viewed as a corrupted version of the added pattern. With the correlation between the added pattern and the received work, we

can determine whether the watermark exists or not.

2.2.2 Properties

Watermarking systems have a number of properties [15][16]. The relative importance of each property depends on the application. For example, the watermark used for the copyright protection needs to be robust, while the authentication applications require the fragile watermark to identify the modification. Based on the processes of watermarking, the properties are categorized and associated with watermark embedding, detection and security feature of the watermark scheme [17]. In this section, we highlight some of them.

2.2.2.1 Fidelity

The fidelity of a watermarking system refers to the perceptual similarity between the original and watermarked versions of the cover work. In other words, the changes to the original cover work should be imperceptible to human. Different applications may have very different requirements to the definition of the fidelity. In some applications, we may accept the middle perceptible watermarks for the exchange of higher robustness or lower cost. Conversely, take the HDTV and DVD as example, for the signals in a very high quality, it requires much higher fidelity watermarks.

2.2.2.2 Data Payload

Data payload refers to the number of watermark bits a watermark encodes within a unit of time or within a work. Similar as fidelity, different applications may require very different data payloads. Copy control applications may require just 4 to 8 bits of information to be received over a period of, every 10 seconds for music. Television

broadcast monitoring requires much more bits of information than the copy control applications to identify all commercials. In the watermarking literature, some systems just determine whether the watermark is present or not. In this case, only one bit of payload is enough. On the other hand, data payload is conflict with the fidelity. More data embedded in the cover work, more effects to the fidelity of the original signal.

2.2.2.3 Blind or Informed Detection

The blind or informed detection is categorized by whether the original cover work is used for watermark extraction. The informed detection uses the original cover work for registration, to obtain the watermark pattern and to counteract any temporary or geometric distortions that might have been applied to the watermarked copy. The blind detection is performed without the need for the original work.

2.2.2.4 Robustness

Robustness refers to the ability to detect the watermark after common signal processing operations, such as lossy compression, geometric distortions, low-pass filtering, additive noise distortions, etc. Not all watermarking applications require robustness to all possible signal processing operations. Clearly, this is application dependent. In some scenario, the robustness is highly required, while some others may be completely irrelevant with robustness. Fragile watermark is one designed for the non-robustness which usually used for authentication and quality distortion measurement.

2.2.2.5 False Postive/Negative Rate

False positive means that the system detects a watermark which is not actually present. When we talk about the false positive rate, we refer to the number of false positives

we expect to occur in a given number of runs of the detections [18]. On the contrary, false negative rate is the occurrence rate that the watermark detector fails to extract an embedded watermark. Watermark detection should be accurate. Hence, the rate of false positives, the detection of a non-marked work, as well as the rate of false negatives, the non-detection of a marked work, should be low. Low false positive and false negative rates are usually in conflict with low embedding distortion because reducing false positive and false negative rates usually means increasing the amount of watermark, which inevitably will inflict higher distortion on the quality of the watermarked media.

2.2.2.6 Security

Watermarks should survive deliberate attempts to remove them [16]. The security of watermark refers to its ability to resist hostile attacks – unauthorized removal, embedding and detection. For unauthorized removal, the attacker obtains a number of copies of given work, each with a different watermark, and compare them to get a copy with no watermark. The unauthorized embedding means that the attacker embeds illegitimate watermarks into works that should not contain them. And the unauthorized detection refers to the adversary detection of an embedded message.

To provide security, watermarking algorithm can be designed to use secret keys. Takes the image watermark as an example. We use the secret key as the seed to generate the pseudo-random (PN) noise pattern. The watermark embedding and detection process use the same PN pattern. Ideally, it should not be possible to detect the presence of a watermark in a Work without knowledge of the key, even if the watermarking algorithm is known. Further, by restricting knowledge of the key to only a trusted group, it should become extremely difficult for an adversary to remove a watermark without causing significant degradation in the fidelity of the cover Work.

2.2.2.7 Cost

The economics of deploying watermark embedders and detectors can be extremely complicated and depends on the business models involved. From a technological point of view, the two principal issues of concern are the speed with which embedding and detection must be performed and the number of embedders and detectors that must be deployed. For example, in broadcast monitoring, both embedder and detector must work in real time. They may need a few embedders but must have several hundred detectors. Conversely, for some other applications, such as transaction-tracking implemented by DiVX (a digital video compression format based on the MPEG-4 technology), there would be millions of embedders and only a handful of detectors.

2.2.3 Applications

Digital watermarking draws more and more attention as one of the key technology elements for content management, copyright protection and copy control of digital contents [19]. In many other areas, digital watermarking is also bloomed and has a broad range of uses across many industries [17][20]. The following are some applications of digital watermarking.

2.2.3.1 Copyright Protection

Copyright protection also means owner identification [21]. Digital watermarking technology was first used in this area in 1954. Emil Hembrooke of the Muzac Corporation filed a patent entitled "Identification of sound and like signals" [20]. He described a method for imperceptibly embedding an identification code into music for the purpose of ownership proving. Digital watermark is superior to textual copyright notice be-

cause it can be made imperceptible and inseparable from the cover work. Once the copyright information is embedded in the work, the users supplied with watermark detector can very easily identify the owner even the work has been modified by common signal processing.

2.2.3.2 Transaction Tracking

In transaction tracking, or fingerprinting, the watermark records one or more transactions on the cover Work. It can be used to identify the legal use of the work and track the source of misused content. For example, watermarks for transaction tracking implemented by DiVX DVD player allows fewer than 5 copies to be effective [17]. Furthermore, each player generates a unique watermark. On one hand, the user can have no more than 5 copies for their own use. On the other hand, if the user sells his/her copies in the black market, the DiVX corporation could identify the adversary by decoding the watermark.

2.2.3.3 Copy Control

Simply, copy control prohibits the illegal copy of the work. Because watermarks are embedded in the content itself, they are present in every representation of the content and therefore might provide a better method of implementing copy control. If every recording device were fitted with a watermark detector, the devices could be made to prohibit recording whenever never-copy watermark is detected at its input [17]. The recording device contained a watermark detector could use the watermark to prevent copying of copyrighted material. Such a system has been used on video DVDs by the Copy Protection Technical Working Group.

2.2.3.4 Content Authentication

Through some software, it becomes easier to tamper the digital works, such as image, audio and video. A fragile digital watermark can be embedded in the cover work to identify the modification. Once the cover work is modified, the watermark is affected too. Therefore, according to the analysis of the watermark, we can examine how and where the cover work was tampered. Furthermore, as a special subset of fragile watermark, reversible watermark enables the recovery of the original, un-watermarked content after the watermarked content has been detected to be authentic. Such reversibility to get back un-watermarked content is highly desired in sensitive imagery, such as military data and medical data [22].

Chapter 3

Literature Review

Along with the blooming of digital watermarking technology, the audio watermarking algorithms have also been proposed in several categories, such as quantization based and spread-spectrum based algorithms. On the other hand, the speech quality assessment methods have been developed from the traditional subjective testing, the most reliable method, to the objective methods with lots of advantages, such as low-cost, fast, and convenient. In this chapter, we will give a brief review to the development of these two technologies.

3.1 Audio Watermarking

Along with the blooming of digital watermarking, some audio watermarking schemes have been also proposed in the last decade. For example, there have been quantization based schemes, spread-spectrum based schemes, two-set based schemes, replica based schemes, self-marking schemes and so on.

The quantization based watermarking schemes quantize the sample values to embed

and detect watermark. The spread-spectrum based schemes embed a pseudo-random number and detect watermark by computing the correlation between the embedded pseudo-random noise and received watermarked audio. The two-set based schemes make two sets of audio blocks with different energies. Based on the difference of the means between these two sets, we can conclude whether the watermark is present or not. The replica based schemes embed part of original signal in frequency domain as a watermark. The detector can also generate the replica from the received watermarked audio. From the correlation between the original replica and watermarked replica, we can determine the presence of the watermark. The self-marking schemes embed a special signal into audio, or change signal shapes in time or frequency domain.

Based on these technologies, quite a number of applications have been developed in the area of audio watermarking.

Chen *et al.* [23] proposed dither modulation for the watermarking system in 1999. In dither modulation, such as quantization index modulation (QIM), the embedded information modulates the dither signal of a dithered quantizer. They developed a framework within which one can analyze performance trade-offs among robustness, distortion, and embedding rate. The results show that QIM and dither modulation systems have considerable performance advantages over previously proposed spread-spectrum and low-bit(s) modulation systems in terms of the achievable performance trade-offs among distortion, rate, and robustness of the embedding [23].

To place more difficulties on watermark removal for attackers, Cox *et al.* [24][25] inserted the watermark into the spectral components, hiding a narrow band signal in a wideband signal. The watermark is spread over very many frequency bins, hence the energy in any bin is very small and undetectable. Meanwhile, spreading the watermark throughout the spectrum of a signal ensures security. Therefore, the watermark is

difficult for an attacker to remove, even when several individuals conspire together with independently watermarked copies of the data. It is also robust to common signal and geometric distortions such as digital-to-analog and analog-to-digital conversion, resampling, quantization, dithering, compression, rotation, translation, cropping and scaling. The same digital watermarking algorithm can be applied to all three media under consideration with only minor modifications, making it especially appropriate for multimedia products [24]. Furthermore, the well placed watermark in the frequency domain of a sound track will be practically impossible to hear.

Kirovski [26] described several novel mechanisms to encode and detect direct-sequence spread-spectrum watermarks in audio signals. They proposed to improve robustness, imperceptiveness and prevent de-synchronization and removal attacks of the watermark. The watermark is embedded in the frequency domain and its energy is distributed throughout the entire synthesis block, but it may be audible in the quite periods. To solve this problem, a procedure is used to decide whether to use a particular block in the watermark embedding/detection process. They also modified the traditional spread-spectrum watermark detector to force the adversary to add an amount of noise proportional in amplitude to the recorded signal if they want to successfully remove the watermark.

Boney [27] proposed to embed the watermark in the frequency domain with the masking technology. The watermark is produced by filtering a PN-sequence with a filter that approximates the frequency masking characteristics of the human auditory system. It is then weighted in the time domain to account for temporal masking. The results show that the watermarking scheme is robust in the presence of additive noise, lossy coding/decoding, VQ distortion, multiple watermarks, resampling and time scaling.

Echo hiding [28] is one of the most well known audio data hiding techniques. It is based on the fact that human auditory system cannot distinguish an echo from the original audio when the delay and amplitude of the echo are appropriately controlled. Oh *et al.* [29] proposed to embed large-energy echoes while the host audio quality is not deteriorated, so that it is robust to common signal processing modifications and resistant to tampering. Subjective and objective evaluations confirmed that the proposed method could improve the robustness without perceptible distortion.

In 2003, Yeo *et al.* [30] presented the modified patchwork algorithm (MPA), a statistical technique for audio watermarking in the transform domain, especially in discrete cosine transform (DCT) domain. This proposal is robust to attacks defined by secure digital music initiative (SDMI), a union formed by more than 160 companies from the music industry. Furthermore, the embedded watermarks are made almost inaudible by adjusting the region of embedding and the strength of the noise. In order to test the quality of the proposed algorithm, 10 bits of copyright information are embedded into every 10 seconds. Experimental results show that the proposed watermarking algorithm is sustainable more than 98 percent against compression algorithms such as MP3 and AAC of 64 kbps, as well as common signal processing manipulations specified by SDMI.

In general, many watermarking algorithms have been proposed for various applications. However, to our best knowledge, there is no audio watermarking technology used for speech quality evaluation. In this thesis, we propose a new application for digital watermarking: speech quality evaluation.

3.2 Objective Audio/Speech Quality Evaluation

Generally, audio quality evaluation is carried out by either subjective or objective methods. The subjective methods [31] measure speech intelligibility, or the overall perceived quality using listening-only methods of subjective testing which we introduced in Section 2.1.2. Intelligibility tests include modified rhyme test (MRT) and diagnostic rhyme test (DRT). Due to the fact that the subjective methods are high cost and time-consuming, objective speech quality evaluation methods have been developed in the recent 20 years, classified into two categories: signal based methods and parameters based methods [5]. The signal based methods use the reference and degraded signals as the input to the measurement, while the parameters based methods predict the speech quality through a computational model instead of using real measurement. The PESQ, the state-of-the-art objective measurement algorithm, and E-model [32], a computational model for use in transmission planning, are defined by ITU-T in the class of signal based method and parameter based method respectively. In this section, we will firstly review the existing objective audio/speech quality assessment methods in both classes, and then introduce in more detail the PESQ and E-model.

3.2.1 Signal Based Methods

For signal based methods, in order to achieve an estimate of the perceived quality, a measurement should employ as much understanding of human perception and human judgement as possible. The common idea behind perceptual quality measures is to mimic the situation of a subjective test, where human beings would have to score the quality of sound samples in a listening laboratory environment.

Perceptual measurement is still in its infancy, but it has been the core focus of

OPTICOM since 1995. In 1993, the intrusive algorithm to calculate the perceptual speech quality measure (PSQM) was devised by Beerends [33]. In 1995, the PSQM algorithm was tested by Study Group 12 of ITU-T for the purpose of international verification. Consequently, PSQM [34] was recommended by the ITU-T in 1996 for the objective quality measurement of telephone band speech codecs. Since then, PSQM has been used intensively for voice quality testing applications. To calculate PSQM, first of all, input signals (sampling rates of 8 KHz or 16 KHz) are transformed from time domain to frequency domain with FFT. After a Hanning window is applied, the (linear) frequency scale is transformed to a pitch scale ("frequency warping"). And then, both the reference and test signals are filtered with the transfer characteristics of the receiving device, while a "Hoth noise" signal is added to simulate the background noise present in a typical office environment. The subsequent process of "intensity warping" leads to a representation of a compressed loudness as a function of pitch and time. By subtracting the two signal representations, an estimate of the audible error is derived. Furthermore, the "asymmetry processing" is taken into account that distortions, which are introduced by the device under test, are more easily perceived than signal components that are left out by the codec.

When PSQM was standardized as P.861, its scope is just for assessing speech codecs while VoIP was not considered. To meet the new requirements arising from the next generation networks such as VoIP, in 1998, PSQM+, the advanced perceptual speech quality measure (intrusive) according to ITU-T COM 12-20 has been recommended.

However, the PSQM+ still had significant problems with the compensation of the varying delay although it works well on the burst errors. Consequently, to overcome such problems, an advanced delay tracking feature has been added by OPTICOM in OPERA, called "PSQM/IP". It achieves PSQM results for the speech quality of VoIP

networks.

In 2001, with the new ITU standard P.862 (PESQ) [35], this problem is finally eliminated. We will discuss in more detail in Section 3.3.

In a joint cooperation under the leadership of International Telecommunication Union (ITU), a group of leading sound quality experts have developed a new measure for sound quality: perceptual evaluation of audio quality (PEAQ). In 1999, PEAQ has been defined as the new ITU-R recommendation BS.1387 [36], thus providing an advanced quality metrics to the world wide audio industry.

3.2.2 Parameters Based Methods

Besides perceptual measurement, some other parameters based methods, such as Gaussian mixture models, artificial neural networks and E-models, have also been developed for the audio/speech quality assessment.

Falk *et al.* [4] proposed a novel approach to objective speech quality measurements using Gaussian mixture models (GMMs). In this approach, firstly, from the distortion surface between the original speech signal and the degraded speech signal, a large pool of feature measurements is extracted and created. Secondly, good features are then chosen using a statistical data mining method, multivariate adaptive regression splines (MARS). Thirdly, the joint density of these selected features are modelled with the subjective MOS as a Gaussian mixture. Finally, using this model, the least squares estimate of the subjective MOS value is derived. This approach outperforms PESQ in RMSE but the improvement in correlation between the subjective MOS and predicted MOS is small.

Mohamed *et al.* [37] employed the artificial neural networks (ANNs) to assess the audio quality in packet networks with the concern of several distortion parameters on

the transmitted audio, such as packet loss rate, arrival jitter, end-to-end delay, sample rate and the number of bits per sample of codec algorithm, echo, crosstalk effect, etc. To build such a tool for the quality assessment, at first, the most effective quality-affecting parameters of the packet switched networks should be chosen. For each parameter, give the typical values and ranges. Then, a simulation environment is implemented to send audio samples from a source to the destination. By inviting a group of N people to listen to the very distorted sample and grade the subjective MOS, establish the relation between samples and parameter values. After that, a suitable neural network (three-layered feedforward network) is defined. Once a stable neural network configuration is obtained, the trained ANN will take the given parameter values and correspondingly compute the subjective MOS quality score.

Ding *et al.* [5][38] has extended the E-model in speech quality prediction in VoIP scenarios. The E-model, which will be introduced in Section 3.4, is a computation model used for the parameter-based methods to predict the speech quality. It includes a set of parameters characterizing the end-to-end voice transmission as its input, and the output transmission rating factor R can be transformed into a MOS scale. In this approach, the packet size was set as 10, 20, 30, 40 and 50 ms and three error concealment methods, repetition, silence and built-in were considered. Good accuracy is achieved by this extended E-model formula, especially for the separated impairments; the prediction errors lie in between ± 0.10 MOS for most cases. For the combined impairments, the formula still gives good prediction when the packet loss rate is below 10%, the errors range between -0.20 and $+0.10$ MOS.

3.3 Perceptual Evaluation of Speech Quality (PESQ): A Typical Signal Based Method

With the introduction to new technology used for telephony services, especially for VoIP (Voice over Internet Protocol), determining the subjective speech quality of a transmission system has always been an expensive, unrepeatable, unstable and laborious process. In order to provide a rapid and repeatable method for speech quality measurement, since 1990, ITU has proposed several objective algorithms for standardization, such as PSQM, which was accepted as Recommendation P.861 in 1996.

PSQM, the intrusive algorithm is an adapted version of the more general perceptual audio quality measure (PAQM), optimized for telephony speech signals. Its scope was to assess speech codecs, used primarily for mobile transmission, like GSM. With the new demands arising from next generation networks, especially VoIP, much higher distortions than with GSM codecs (e.g. unconstant delay between the reference and the test signal) needs to be handled. In this case, ITU developed PSQM+ by revising the P.861 standard. Later, to deal with significant problems with the compensation of the varying delay, an advanced delay tracking feature called "PSQM/IP" has been added by OPTICOM in OPERA. This algorithm extends PSQM results for the speech quality of VoIP networks.

Subsequently, with PESQ which was selected in May 2000 as draft of ITU-T recommendation P.862 and replaced P.861 early in 2001 [39][40], the most eminent problem, unconstant delay is finally eliminated. PESQ combines the excellent psycho-acoustic and cognitive model of PSQM+ with a time alignment algorithm adopted from PAMS, which handles varying delays perfectly. However, PESQ cannot fully replace PSQM+ because it is not designed for streaming applications.

PESQ directly uses MOS to express the speech quality. Different from the subjective method, PESQ MOS (P.862) defined by ITU ranges from -0.5 (worst) to 4.5 (best) comparing with the subjective MOS scope, from 1.0 to 5.0. This is caused by their distinct definition. MOS stands for mean opinion score but PESQ simulates the listener test and its MOS is optimized to reproduce the average result of all listeners which the best average was proved to be 4.5 instead of 5.0. However, PESQ provides significantly higher correlation (0.935) with subjective opinion than the models of P.861, PSQM and MNB. It gives accurate predications of subjective quality in a very wide range of conditions, including background noise, analogue filtering, and/or variable delay as well as coding distortions and errors [2][40].

3.3.1 Overview of PESQ Model

As illustrated in Figure 3.1, PESQ predicts the speech quality through the process of level aligning, input filtering, time aligning, auditory transform, disturbance processing, identifying bad intervals, and cognitive modelling. Except for time aligning and bad intervals identification, the other steps are either inherited from or similar to PSQM+. The basis of the whole signal processing is the simulation of the listening test. In this model, the original and degraded signals are mapped onto an internal representation using a perceptual model. The difference of those representation determines the audible difference and is used to predict the perceptual speech quality (PESQ MOS) of the degraded signal by the cognitive model. Generally, PESQ is implemented in two consecutive partitions, level/time alignment pre-processing and perceptual model. The details about these consecutive steps are described in the following sections.

3.3.2 Input Signal Pre-processing

It is important that the input signals in PESQ are representative of the real signals carried by communication networks. Hence the pre-processing is often necessary to take account of filtering in the sending path of a handset, and to ensure that power levels are set to be appropriate.

3.3.2.1 Level Alignment

PESQ assumes that the subjective listening level is constant, about 79 dB SPL (Sound Pressure Level) at the ear reference point. Both the original and degraded signals need to be scaled to the same constant power level. The level alignment is carried out based on the power of bandpass-filtered versions (300-3000Hz) of both the reference and degraded signals. Besides the level alignment in the time domain, on the other hand, it is also necessary to do level alignment in the frequency domain after the time-frequency analysis. This is implemented by the windowed FFT with a frame length of 32-ms. The sine wave used is generated with the frequency of 1000 Hz and amplitude of 40-dB SPL. Therefore, the same 40-dB SPL reference tone is then used to calibrate the psychoacoustic loudness scale. The resulting output signals $X_s(t)$ and $Y_s(t)$ are the scaled version of the input reference and degraded signals $X(t)$ and $Y(t)$.

3.3.2.2 IRS Filtering

It is assumed that the listening tests were carried out using an Intermediate Reference System (IRS) receiving or a modified IRS receiving characteristic in the handset. This filtering action simulates the frequency response of a typical telephone handset earpiece. It has a bandpass characteristic, with -3 dB points near 400 Hz and 3200 Hz, and a fairly flat passband. A perceptual model of the human evaluation of speech quality must

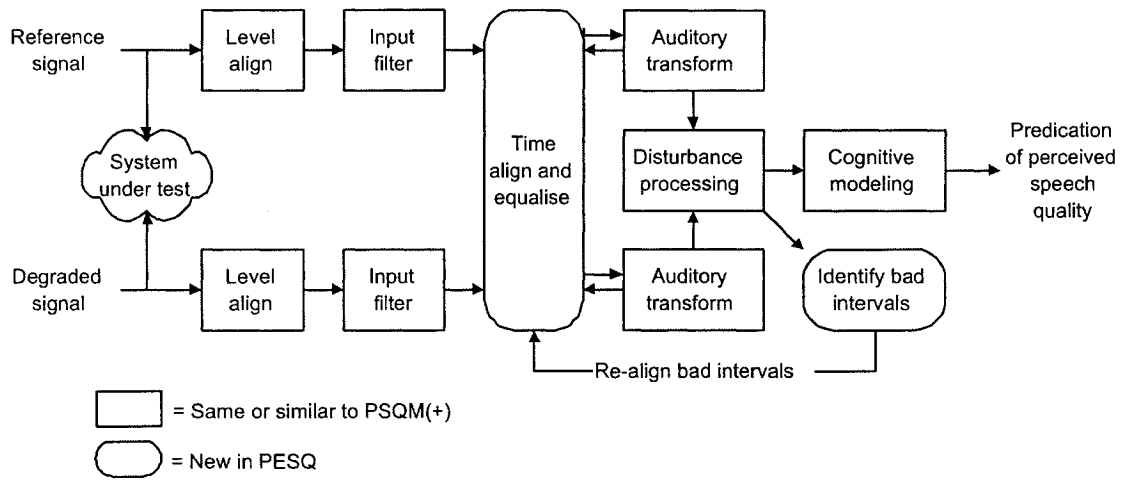


Figure 3.1: Structure of the perceptual evaluation of speech quality (PESQ) [1].

take account of this to model the signals that the subjects actually heard. In PESQ this is implemented by a FFT over the length of the file, filtering in the frequency domain with a piecewise linear response similar to the (unmodified) IRS receive characteristic (P.830), followed by an inverse FFT over the length of the speech file. This results in the filtered versions $X_{IRSS}(t)$ and $Y_{IRSS}(t)$ of the scaled input and output signals $X_S(t)$ and $Y_S(t)$. A single IRS-like receive filter is used within PESQ irrespective of whether the real subjective experiment used IRS or modified IRS filtering. The reason for this approach was that in most cases the exact filtering is unknown, and that even when it is known the coupling of the handset to the ear is not known. It is therefore a requirement that the objective method should be relatively insensitive to the filtering of the handset. The IRS filtered signals are used both in the time alignment procedure and the perceptual model.

3.3.2.3 Time Alignment

The time alignment calculates the time delay values which will be used for the perceptual model to allow the corresponding signal parts of the original and degraded speeches to be compared. The following are the main steps of time alignment:

- Narrowband filtering to both signals to emphasize perceptually important parts which will only be used for time alignment;
- Envelope-based delay estimation calculated from the scaled original and degraded signals $X_s(t)$ and $Y_s(t)$;
- Division of original signal into a number of subsections known as utterances;
- Enveloped-based delay estimation on each utterances;
- Fine correlation/histogram-based identification of delay to nearest sample on each utterance; The delay changes during silent periods are also considered. The result of delay value together with the start and end points of each utterance allows the delay of each frame to be identified in the perceptual model;
- Splitting utterances and re-aligning the time intervals to search for delay changes during speech.

The delay of utterances will be further used for computing the delay of frames in the auditory transform.

3.3.3 Perceptual Model

The perceptual model is used to compute the speech quality score, the distance between the original and degraded signal. The PESQ score is mapped to a MOS-like scale, a

single number in the range of -0.5 to 4.5, normally between 1.0 and 4.5, through a monotonic function.

3.3.3.1 Computation of Active Speech Time Interval

The large silent intervals in the start or end of the signal could influence the calculation of certain average distortion values over the file length, no matter the original or degraded speech. In this case, the estimation of the silent parts has to be made to find the start and end points of the active interval. The PESQ locates these points by searching the positions at which the sum of five successive absolute sample values exceed 500 at the beginning and end of the original speech file. The interval between this start and end is defined as the active speech time interval. In order to save computation cycles and/or storage size, some computations can be restricted to the active interval.

3.3.3.2 Auditory Transform

This is a psychoacoustic model to map the signals into a representation of perceived loudness in time and frequency domain. As shown in Figure 3.2, the following are the main steps.

Bark spectrum (Short-term FFT and Time axis modification). The human ear performs a time-frequency transformation which is modelled by PESQ using a short-term FFT with a Hanning window over 32-ms frames. The successive frames are 50% overlapped. The start points of the frames in the degraded signal are shifted over the delay. The time axis of the original speech is left as is. If the delay increases, parts of the degraded signal are omitted from the processing, whereas for the decreases in the delay parts are repeated.

Calculation of the pitch power densities. The bark scale reflects that at low fre-

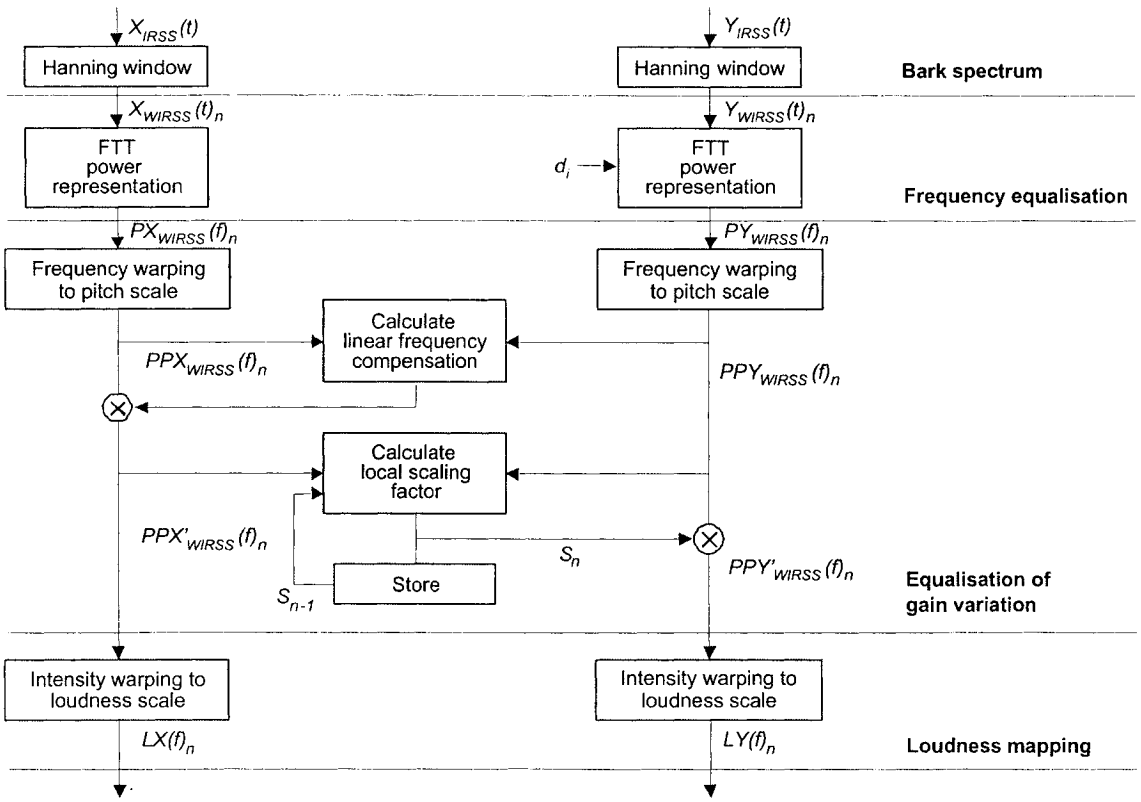


Figure 3.2: Structure of auditory transform [2].

quencies, the human hearing system has a finer frequency resolution than it has at high frequencies. This is implemented by binning FFT bands and summing the corresponding powers of the FFT bands with a normalization of the summed parts. The resulting signals are known as the pitch power densities $PPX_{WIRSS}(f)_n$ and $PPY_{WIRSS}(f)_n$.

Compensation of the original pitch power density. The compensation of the original pitch power density is also called frequency equalization. The power spectrum of the original and degraded pitch power densities are averaged over time for active speech frames which are time-frequency cells whose power is more than 1000 times the absolute hearing threshold. The ratio of the degraded spectrum to the original spectrum gives the partial compensation factor which is never more than 20 dB. The

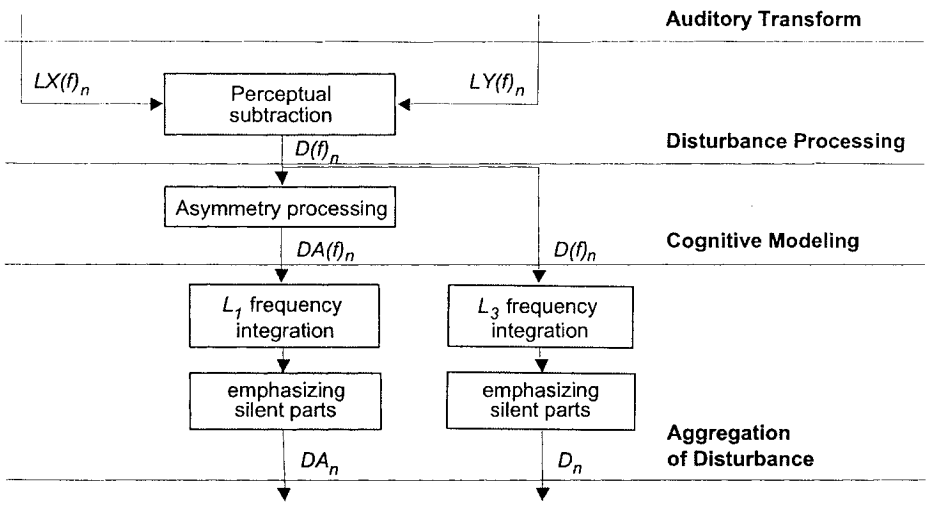


Figure 3.3: Structure of disturbance processing and cognitive modeling [2].

original pitch power density $PPX_{WIRSS}(f)_n$ of each frame n is then multiplied with this partial compensation factor to equalize the original to the degraded signal with the results in an inversely filtered original pitch power density $PPX'_{WIRSS}(f)_n$.

Compensation of the degraded pitch power density. This is used for time-varying gain. The ratio of the power in the original and the degraded files is calculated and used to identify gain variations. A first order low pass filter (along the time axis) is applied to this ratio. The distorted pitch power density in each frame, n, is then multiplied by this ratio, resulting in the partially gain compensated distorted pitch power density $PPY'_{WIRSS}(f)_n$.

Loudness mapping. The original and degraded pitch power densities are transformed to a sone loudness, including a frequency-dependent threshold and exponent. This gives the perceived loudness in each time-frequency cell.

3.3.3.3 Disturbance Processing

The absolute difference D_n between the degraded and the original loudness density for each time-frequency cell (frame number n) is computed to give a measure of audible error. If it is positive, components such as noise have been added, while the negative one indicates components have been omitted from the original signal. The minimum of the original and degraded loudness density is computed for each time-frequency cell. These minima are multiplied by 0.25 as the masking threshold τ_n under which is inaudible. The disturbance density $D(f)_n$ is computed based on the following rules.

$$D(f)_n = \begin{cases} D_n - \tau_n & \text{if } D_n > \tau_n \\ 0 & \text{if } |D_n| <= \tau_n \\ D_n + \tau_n & \text{if } D_n < -\tau_n \end{cases} \tag{3.1}$$

3.3.3.4 Cognitive Modelling

Unlike PSQM, PESQ computes two different error averages with and without an asymmetry factor. The asymmetry effect is caused by the codec distortion. When the codec leaves out a time frequency component the resulting output signal cannot be decomposed in the same way and the distortion is less objectionable. This effect is modelled by calculating an asymmetrical disturbance density $D_A(f)_n$ per frame by multiplication of the disturbance density $D(f)_n$ with an asymmetry factor. This asymmetry factor equals the ratio of the distorted and original pitch power densities raised to the power of 1.2. If the asymmetry factor is less than 3 it is set to zero. If it exceeds 12 it is clipped at that value.

3.3.3.5 Aggregation of the Disturbance Densities over Frequency and Silent Interval

The disturbance density $D(f)_n$ and asymmetrical disturbance density $D_A(f)_n$ are integrated (summed) along the frequency axis using two different Lp norms and a weight on soft frames (having low loudness). The frame disturbances, D_n and DA_n are aggregated as following:

$$D_n = M_n \sqrt[3]{\sum_{f=1,\cdots,number\ of\ Barkbands} (|D(f)_n|W_f)^3}$$

$$DA_n = M_n \sum_{f=1,\cdots,number\ of\ Barkbands} (|DA(f)_n|W_f)$$

where M_n is a multiplication factor, $[(power\ of\ original\ frame + 10^5)/10^7]^{-0.04}$, resulting in an emphasis of the disturbances that occur during silences in the original speech fragment, and W_f is a series of constants proportional to the width of the modified Bark bins. After this multiplication the frame disturbance values are bounded with the upper limit of 45.

3.3.3.6 Realignment of Bad Intervals

The structure of re-alignment is illustrated in Figure 3.4. Consecutive frames with a frame disturbance above a threshold are called bad intervals. In a minority of cases the objective measure predicts large distortions over a minimum number of bad frames due to incorrect time delays observed by the preprocessing. For those so-called bad intervals a new delay value is estimated by maximizing the cross correlation between the absolute original signal and absolute degraded signal adjusted according to the delays observed by the preprocessing. When the maximum cross correlation is below a threshold, it

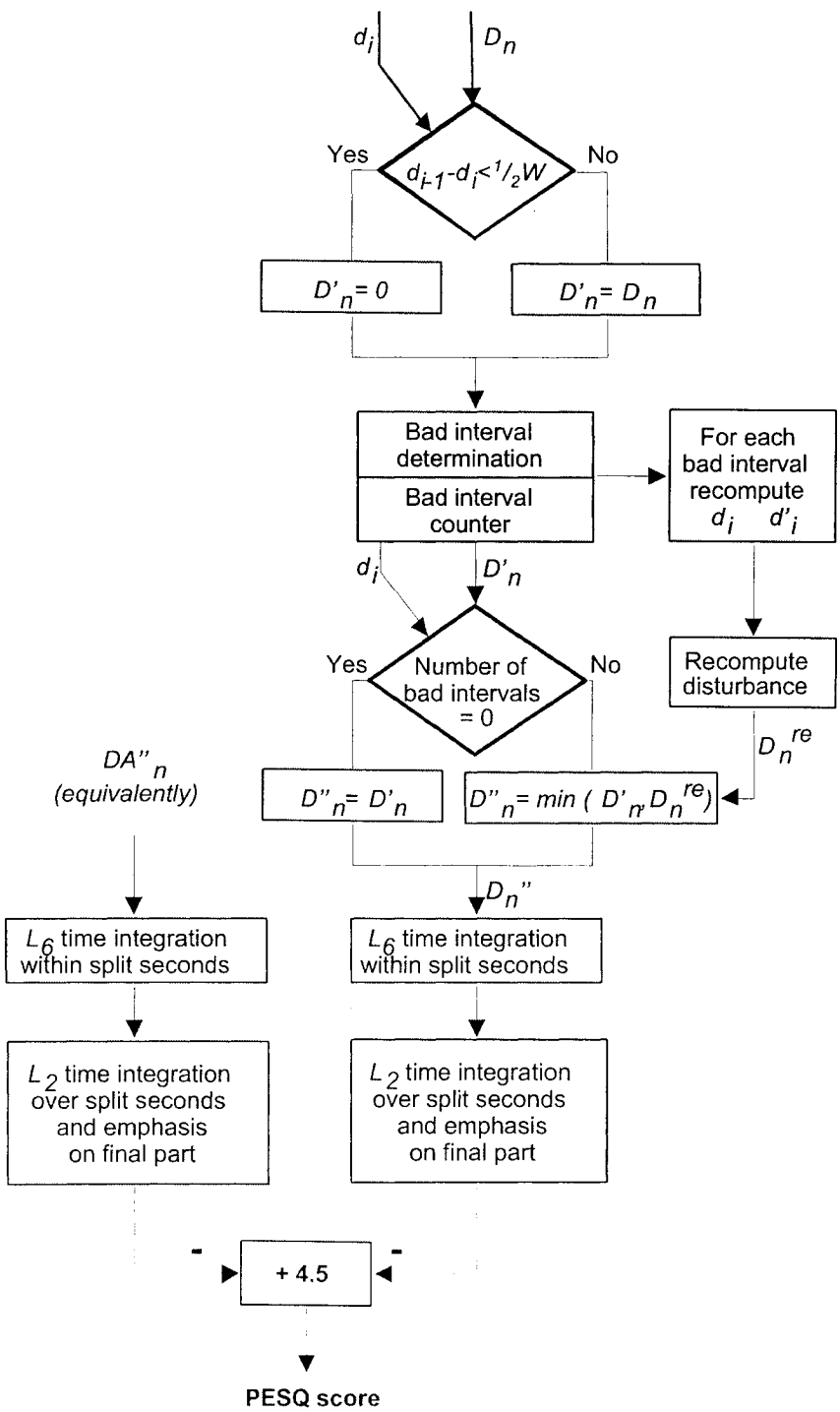


Figure 3.4: Structure of re-alignment and computation of PESQ score [2].

is concluded that the interval is matching noise against noise and the interval is no longer called bad, and the processing for that interval is halted. Otherwise, the frame disturbance for the frames during the bad intervals is recomputed and, if it is smaller replaces the original frame disturbance. The result is the final frame disturbances D''_n and DA''_n that are used to calculate the perceived quality.

3.3.3.7 Disturbances Aggregation and MOS Predication

First, the frame disturbance values and the asymmetrical frame disturbance values are aggregated over split-second intervals of 20 frames (accounting for the overlap of frames, approximately 320 ms). These split-second intervals also overlap 50 percent and no window function is used. Next, the split-second disturbances are aggregated over the complete active time interval of the speech files. Finally, PESQ score, the MOS predication is a linear combination of two parameters: average symmetric disturbance value (d_{SYM}) and average asymmetric disturbance value (d_{ASYM}). Final training is performed on a database of 30 subjective tests, giving the following output mapping used in PESQ [40]:

$$PESQMOS = 4.5 - 0.1d_{SYM} - 0.0309d_{ASYM}$$

The range of PESQ score is -0.5 to 4.5, but for the most cases the output range between 1.0 (bad) and 4.5 (no distortion).

3.4 E-model: A Typical Parameters Based Method

The E-model (ITU-T G.107) is a tool for predicting how an"average user" would rate the voice quality of a phone call with known characterizing transmission parameters. It estimates the user satisfaction at a narrow-band, handset conversation, as perceived by the listener. The E-model calculates the transmission rating factor R with Equ. (3.2), using the network impairment factors, which were obtained after an extensive set of subjective experiments. Typical network impairment factors used in the VoIP cable telephony are codecs, delay, and packet loss. A higher R factor corresponds to a better telephone quality, zero being the worst value, 70 toll quality, and 100 excellent quality. One novel feature of the E-Model is the assumption that sources of impairment which are not correlated to each other can be added on a psychological scale. This allows to trade off different sources of impairment (e.g. loss versus delay) against each other.

$$R = R_0 - I_s - I_d - I_e + A \tag{3.2}$$

where, R_0 is the basic signal-to-noise ratio based on sender and receiver loudness ratings, the circuit, and room noise; I_s is the sum of real-time or simultaneous speech transmission impairments, e.g. loudness levels, sidetone and PCM quantizing distortion; I_d is the sum of delay impairments relative to the speech signal, e.g., talker echo, listener echo and absolute delay; I_e is the equipment impairment factor for special equipment; A is the advantage factor that adds to the total and improves the R-value for new services.

Assuming that echo is properly controlled by an echo cancellation module, one can review the impairments of the E-model in terms of delay, codec impairments, and packet loss.

After computing the R-value based on the impairment factors, the R-value is converted into a MOS score based on the Equ. (3.3).

$$MOS = \begin{cases} 1 & R < 0 \\ 1 + 0.035R + R(R-60)(100-R) \cdot 7 \cdot 10^{-6} & 0 < R < 100 \\ 4.5 & R > 100 \end{cases} \tag{3.3}$$

Since the E-model is based on the measurements of impairments, it is appropriate for root-cause analysis in terms of impairment factors as well as network segments, and can be easily incorporated with the Network Management System (NMS). The E-model is also scalable because it does not require the speech samples to estimate the voice quality.

The E-model consists of several models that relate specific impairment parameters and their interactions to end-to-end performance. The total end-to-end performance, taking into account all factors, is estimated using the Impairment Factor method.

Chapter 4

Proposed Algorithm and Implementation Strategies

As reviewed in Chapter 3, the main drawbacks of ITU's quality models are as the following. The PESQ algorithm is not able to predict the speech quality at run-time nor does it take into account end-to-end delays. Furthermore, it is often impossible to obtain original speech for use in objective quality assessment for in-service testing. As well as being computationally complex it is also patented. On the other hand the E-Model considers operational parameters which are not known or not relevant to the application. It does not consider the impairment due to dynamic adaptations. Furthermore it assumes tandem coding (transcoding) conditions (ITU G.108, 1999) and as a result it leads to an imprecise correlation between loss rate and speech quality. Therefore, this category of methods need a large training database to construct good estimators of subjective listening quality, and different training database may result in different model.

In this thesis, we propose a new method of speech quality evaluation using digital

watermarking. The basis of the method is that the carefully embedded watermark in a speech will suffer the same distortions as the speech does. The proposed method needs neither original speech, operational parameters, nor training database.

In Section 4.1, we will introduce the employed techniques; in Section 4.2, we describe our watermarking scheme in detail; Section 4.3 is quality evaluation; in Section 4.4, we introduce two performance indices; finally, in Section 4.5, we introduce our implementation strategies.

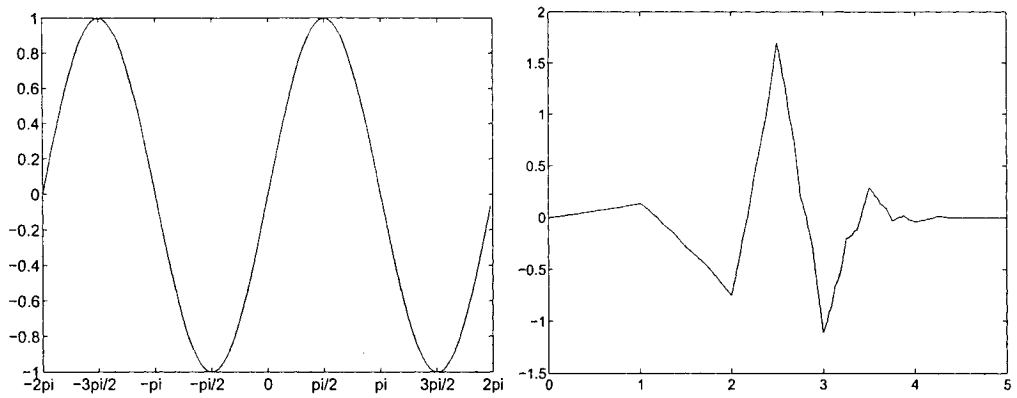
4.1 Techniques Employed

In this section, we introduce the three main techniques, discrete wavelet transform (DWT), quantization and adaptive Control, employed in this thesis.

4.1.1 Discrete Wavelet Transform

Wavelet transform provides a time-frequency representation of the signal. It uses multi-resolution technique by which different frequencies are analyzed with different resolutions, while STFT (Short-Time Fourier Transform) gives a constant resolution at all frequencies. The Fast Fourier Transform (FFT) uses the wave which is an oscillating function of time or space and is periodic. In contrast, the wavelet transform analysis is based on wavelet functions which are localized waves. Wavelets have their energy concentrated in time or space and are suited to analysis of transient signals. Figure 4.1 demonstrates an example of a wave and a wavelet.

The Discrete Wavelet Transform (DWT) is a fast computation of wavelet transform. Based on sub-band coding, it is easy to implement and reduces the computation time and resources required. In the case of DWT, a time-scale representation of the digital



(a) A Wave (b) A Wavelet

Figure 4.1: Example of a wave and a wavelet.

signal is obtained using digital filtering techniques. In other words, filters of different cutoff frequencies are used to analyze the signal at different scales.

As shown in Figure 4.2, the DWT is computed successively by low-pass and high-pass filtering of the discrete time-domain signal, connecting continuous-time mutiresolution to discrete-time filters. In this figure, sequence $x[n]$ denotes the original signal; n is an integer; G_0 is the low pass filter while H_0 denotes the high pass filter.

Figure 4.2 demonstrates a three-level DWT decomposition tree. At each level, the high pass filter produces detail information, $d[n]$, while the low pass filter associated with scaling function produces coarse approximations, $a[n]$. For many signals, the low-frequency content is the most important part because it contains main features of the signal. Meanwhile, the high frequency imparts flavor of nuance. In wavelet analysis, we often speak of approximations and details, and here, it refers to $a[n]$ and $d[n]$.

At each decomposition level, the half band high/low pass filters produce signals spanning only half the frequency band. This doubles the frequency resolution as the uncertainty in frequency is reduced by half. If we reconstruct the signal with this frequency resolution, we end up with twice as much data as the original. Therefore,

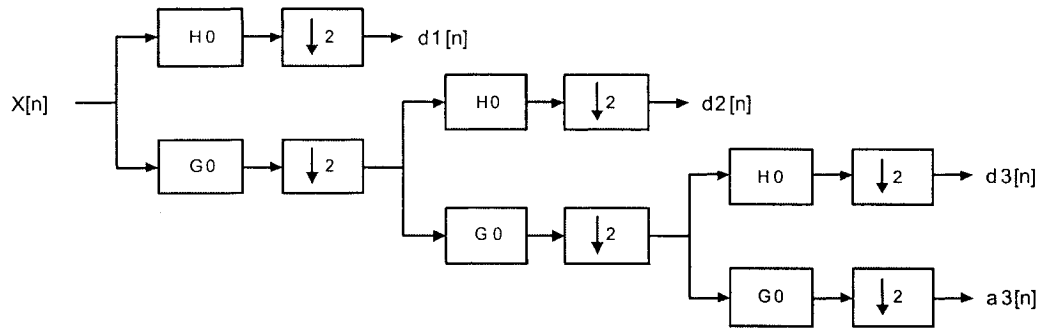


Figure 4.2: Three-level wavelet decomposition tree.

we discard half the samples with no loss of information. This decimation by 2 halves the time resolution as the entire signal is now represented by only half the number of samples. It is called downsampling.

The filtering and downsampling process is continued until the desired level is reached. The maximum number of levels depends on the length of the signal. The DWT of the original signal is then obtained by concatenating all the coefficients, $a[n]$ and $d[n]$, starting from the last level of decomposition.

Figure 4.3 shows the reconstruction of the original signal from the wavelet coefficients. Basically, the reconstruction is the reverse process of decomposition. The approximation and detail coefficients at every level are upsampled by two, passed through the low pass and high pass synthesis filters and then added together. This process is continued through the same number of levels as in the decomposition process to obtain the original signal.

4.1.2 Quantization

Quantization refers to the process of approximating the continuous set of values in the signal data with a finite (preferably small) set of values. The input to the quantizer,

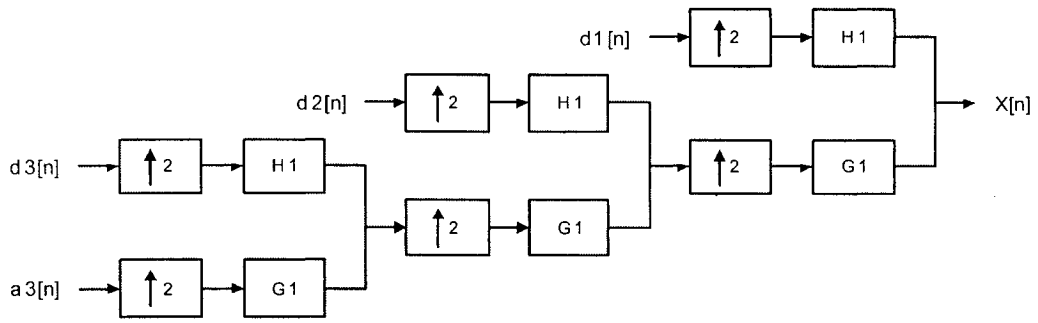


Figure 4.3: Three-level wavelet reconstruction tree.

is the original data, and the output is always one among a finite number of levels. Different from sampling, quantization loses information. A good quantizer is one that represents the original signal with minimum loss or distortion.

There are two types of quantization: scalar quantization and vector quantization. In scalar quantization, each input symbol is treated separately in producing the output, while in vector quantization the input symbols are clubbed together in groups as vectors, and processed to give the output.

For the scalar quantization, any real number x can be rounded off to the nearest integer, say $q(x) = round(x)$, mapping the real line $\Re$ (a continuous space) into a discrete space. More generally, regular quantizer S_i has disjoint intervals given by Equ. (4.1), where a_i (called thresholds) forms an increasing sequence. The width of a cell S_i is its length, $a_i - a_{i-1}$.

$$S_i = (a_{i-1}, a_i] \tag{4.1}$$

If the input range is divided into levels of equal spacing, then the quantizer is termed as a uniform quantizer, otherwise, it is termed as a non-uniform quantizer. For example, in the rounding off quantizer $S_i = (i - 1/2, i + 1/2]$ and $y_i = i$ for all integers

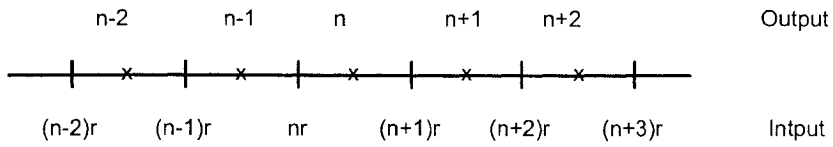


Figure 4.4: A uniform quantizer.

i, a quantizer is uniform if the output levels $\{y_i; i \in \Gamma\}$ are spaced an equal distance Δ apart, and the thresholds a_i are midway between adjacent levels.

A uniform quantizer can be easily specified by its lower bound and the step size. Also, implementing a uniform quantizer is easier than a non-uniform quantizer. In this thesis, we also use a uniform quantizer. Take a look at the uniform quantizer shown in Figure 4.4. If the input falls between $n * r$ and $(n + 1) * r$, the quantizer outputs the symbol n.

Just the same way as a quantizer partitions its input and outputs discrete levels, a dequantizer is one which receives the output levels of a quantizer and converts them into normal data, by translating each level into a 'reproduction point' in the actual range of data.

4.1.3 Adaptive Control

In everyday language, "to adapt" means to change a behavior to conform to new circumstances. Intuitively, an adaptive controller is thus a controller that can modify its behavior in response to changes in the dynamics of the process and the character of the disturbances.

In practice this implies that an adaptive controller is a controller with adjustable parameters, which is tuned on-line according to some mechanism in order to cope with time-variations in process dynamics and changes in the environment.

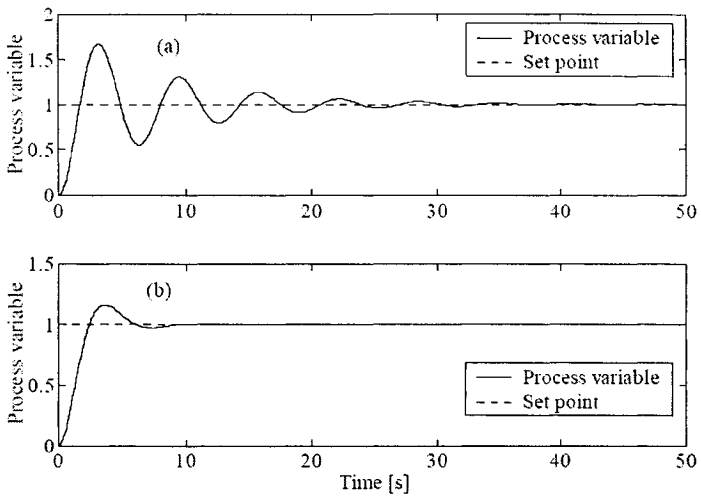


Figure 4.5: Objective of linear adaptive control. An oscillatory control response around the set point (a) is changed to a specified control response (b). The specified control response settles sooner on the set point [3].

Control design requires a dynamic process model. Optimal control design is possible only if the process model is accurate [3]. The designed controller frequently requires on-line refinements to the controller parameters and set points. An adaptive linear controller maintains a specified control response (i.e., corrective action) around a set point during process changes. For non-linear processes, a set of Proportional-Integral-Derivative (PID) controller (PID is a standard feedback loop component in industrial control applications. It measures an "output" of a process and controls an "input", with a goal of maintaining the output at a target value, which is called the "setpoint".) parameters can only maintain the specified control response for a limited range of process conditions. Process changes in non-linear processes may cause the control response to be oscillatory around the set point, as illustrated in Figure 4.5 (a). Adaptive linear control tunes the PID controller parameters, which corrects the oscillatory response in Figure 4.5 (a) to the specified response in Figure 4.5 (b).

Adaptive control does not change the set points that largely determine the economic

return. Set points are selected during the design based on an optimization of dynamic model equations. The optimization considers both economic return and controllability. However, process changes during operation may make the current set points economically sub-optimal. An evolutionary operation (EVOP) was proposed to challenge the use of constant set points in a continuously changing process. The EVOP monitors the process and improves operation by changing the set points towards the economic optimum. The EVOP makes a number of small set point changes that do not disrupt production and use an experimental design to determine the number of set point change experiments. Adaptive control and EVOP may be combined in a two-step methodology to track a changing economic optimum.

4.2 Watermarking Scheme

In this thesis, we present an objective speech quality evaluation method using digital audio watermarking, which can evaluate the quality of speech that is distorted by Gaussian noise, MP3 compression, low-pass filtering, and packet loss. It is applicable to both female and male speakers, as well as different languages.

The proposed method is based on DWT and quantization. There is a difference between the schemes for evaluating the speech quality after packet-loss and the other three distortions. We only embed watermark on sample values instead of DWT coefficients for the packet-loss distortion which will be discussed in Section 4.2.3.

In the following scheme as illustrated in Figure 4.6, we embed watermark in the DWT coefficients of the speech. The watermark will undergo the same distortions as the speech does. Therefore, we can evaluate the quality of the speech that has undergone distortions by evaluating the percentage of correctly extracted watermark bits (PCEW).

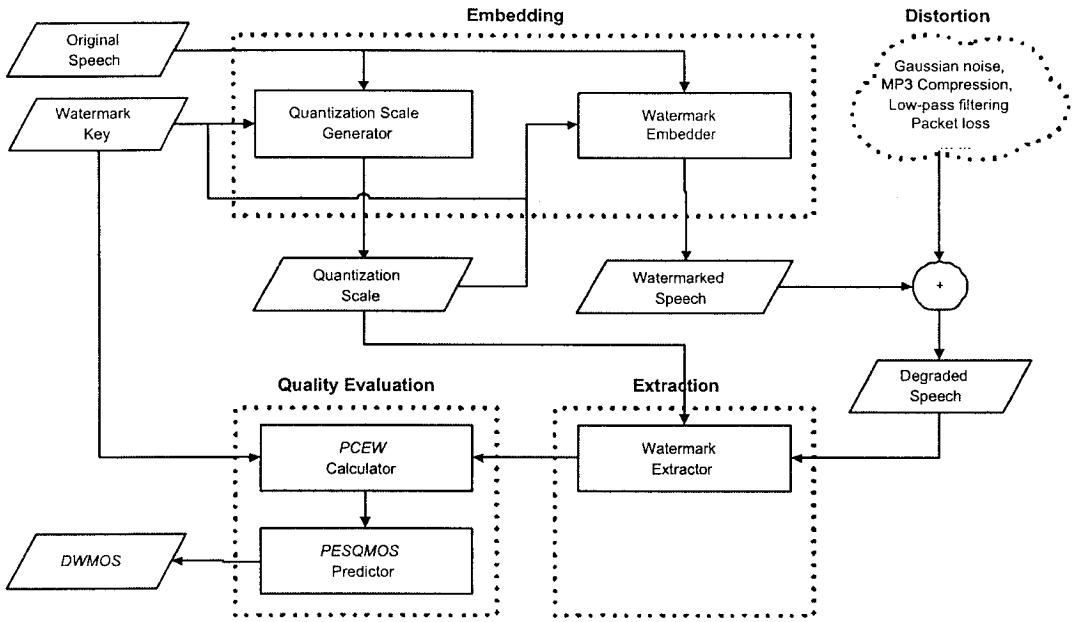


Figure 4.6: Evaluation of speech quality using digital watermarking.

As shown in Figure 4.6, at the embedding side, before embedding, we use the adaptive control algorithm to obtain the optimal quantization scale (QS) (refer to Section 4.5.3), which satisfies both good fidelity of the signal and approximation of 100% watermark detection rate before the distortion applied, for each input speech signal. And then, the watermark, a PN-sequence which generated by a seed (key), is embedded into wavelet domain of the speech to generate a watermarked signal. Different DWT sub-bands of the speech contains different number of watermark bits. Subsequently, the signal distortion was simulated. Finally, after the watermark extraction, we can use PCEW to predict the speech quality.

The following introduces the proposed algorithms of watermark embedding, watermark extraction, speech quality assessment, and performance evaluation.

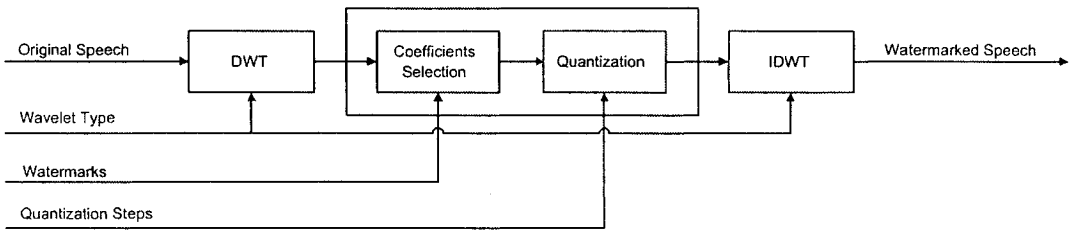


Figure 4.7: Watermark embedding process.

4.2.1 Watermark Embedding Process

As demonstrated in Figure 4.7, the watermark embedding scheme consists the following process. Note: that in Figures 4.7 and 4.9, Quantization Step Δ is calculated according to Equ. (4.2).

$$\Delta = \frac{\max_V - \min_V}{QS} \tag{4.2}$$

where $\max_V$ and $\min_V$ are the maximum and minimum value of the coefficients respectively in a specific decomposition level, and QS is quantization scale.

1. **Watermark Generation:**

   Generated the watermark with a pseudo-random noise (PN) code generator. The seed of the generator is used as a secret key

2. **DWT Decomposition:**

   Apply DWT by choosing a mother wavelet function to get the Lth-level discrete wavelet decomposition of the original audio.

3. **Block Division:**

   On each DWT level, divide the decomposition coefficients into blocks. Each block

will be embedded one watermark bit, so that the number of blocks is equal to the number of watermark bits embedded for each level.

4. **Coefficient Selection:**

   The larger the coefficient, the more the energy to embed a watermark. According to the experiment of [41], we choose 50 largest coefficients to embed the same watermark bit.

5. **Coefficient Quantization:**

   The definition of quantization is the division of a quantity into a discrete number of small parts, often assumed to be integral multiples of a common quantity. The input to a quantizer is the original data, and the output is always one among a finite number of levels. The quantizer is a function whose set of output values are discrete, and usually finite.

   For any discrete wavelet transform, the coefficients are real numbers. The quantization procedure could be performed on these coefficients [42]. In quantization step, every real number is assigned a binary number 0 or 1. We quantize an arbitrary coefficient with the following equation:

$$Q(e) = \begin{cases} 0 & \text{if } k \times \Delta \leq e < (k+1) \times \Delta \\ & (k = 0, \pm 2, \pm 4, ...) \\ 1 & \text{if } k \times \Delta \leq e < (k+1) \times \Delta \\ & (k = 1, \pm 3, \pm 5, ...) \end{cases} \tag{4.3}$$

   where e is the value of the coefficient, while Δ is a positive real number called quantization parameter. During the watermark embedding, after quantizing the selected coefficients, if $Q(e)$ equal to the watermark bit (0 or 1), no change will

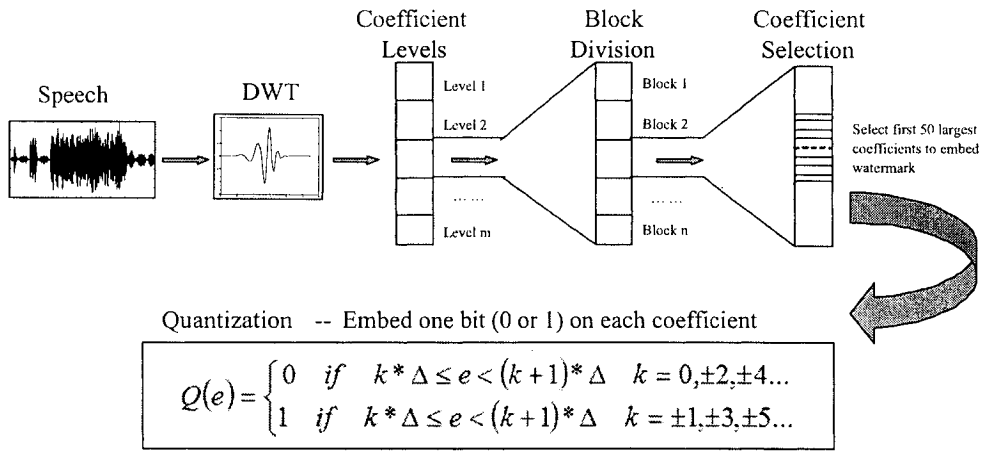


Figure 4.8: Embedding process for one watermark bit.

be made to the coefficient, otherwise, it will be added a Δ to make $Q(e)$ and watermark bit match.

6. **Watermark Embedding:**

   Then, the watermark is embedded by using the following rules. In these rules, $w(i)$ is the watermark bit to be embedded, $Q(e)$ is the quantized value of the selected coefficients.

   - If $Q(e) = w(i)$, no change will be made to the coefficient.
   - If $Q(e) \neq w(i)$, the coefficient e will be forcibly changed so that $Q(e) = w(i)$, using the function $e = e + \Delta$, where Δ is the same quantization step as in the above assignment.

   So far, one watermark bit is embedded into one selected coefficient as shown in Figure 4.8.

7. Iteration:

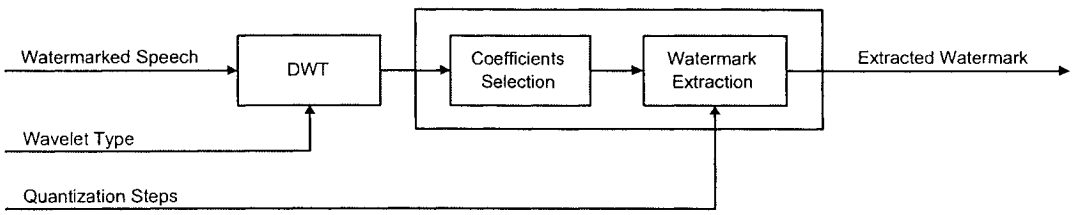


Figure 4.9: Watermark extraction process.

There are three iterations which are included one by one.

- In each block, repeat the steps from step 5 (Coefficient Quantization) to step 6 (Watermark Embedding) to embed the same watermark bit on all the selected coefficients.
- On each level, repeat the steps from step 3 (Block Division) to step 6 (Watermark Embedding) to embed watermark bits into all blocks.
- Repeat the above two iterations until all watermark bits are embedded in all levels.

8. **IDWT Reconstruction:**

   Apply the inverse discrete wavelet transform to reconstruct the watermarked speech with the watermarked wavelet coefficients.

4.2.2 Watermark Extraction Process

The watermark extraction process is similar to the embedding as shown in Figure 4.9. It is just an inverse of embedding process except some differences which are discussed in detail as following.

1. **DWT Decomposition:**

   On watermarked speech, apply discrete wavelet transform by using the same

mother wavelet function as it is used in embedding process to get the Lth-level discrete wavelet decomposition.

2. **Block Division:**

   On each DWT level, divide discrete wavelet decomposition into blocks.

3. **Coefficient Selection:**

   In each block, choose 50 largest coefficients for watermark extraction.

4. **Coefficient Quantization and Watermark Extraction:**

   Firstly, apply the quantizer $Q(e)$ in Equ. (4.1) to selected coefficients. Then, following the rule in [43] as Equ. (4.4), extract the watermark bit.

$$W'(i) = \begin{cases} 1 & \text{if } N_1 > N_0 + 8 \\ 0 & \text{if } N_0 > N_1 + 8 \\ 2 & o.w. \end{cases} \tag{4.4}$$

   where N_1 is the number of binary ones in 50 quantization values of one block; N_0 is the number of binary zeros. The state 2 represents a grey area where a decision cannot be made.

5. **Iteration:**

   There are three iterations which are included one by one.

   - In each block, repeat the steps of coefficient quantization and watermark extraction to extract watermark bits from the other selected coefficients.
   - On each level, repeat the steps from step 2 (Block Division) to step 4 (Watermark Extraction) to extract watermark bits from all blocks.

- Repeat the above two iterations until all of the watermark bits are extracted in all levels.

4.2.3 Watermarking Algorithm for Evaluating Effects of Packet-loss

On the packet-switch network, the main degradations that affect the QoS of IP telephony are coding distortion, packet loss, non-optimal loudness, delay, and talker echo [6]. The PESQ has been verified for impairment due to coding loss and normally distributed packet losses [44]. However, the effects of packet-loss depend on the parameters of packet schemes, such as bandwidth, average bit-rate, frame size and frame rate. and device implementation, so it is difficult to predict the subjective quality from the observed packet-loss rate. Furthermore, The strength of codec distortion depends on the codec models. G.711 [45] is traditionally telephony codec and provides very good speech quality. The μ-law form of G.711 is common in North America and A-law form is used in other parts of the world. This codec is a lossless compression algorithm and does not affect the speech quality. ITU G.729 [46] uses a Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CSACELP) algorithm to compress speech at a rate of 8 kbit/s. It is a high compression codec and degrades the speech quality around 0.6 MOS according to the test result using PESQ.

To suppress codec and surrounding influences that are not the scope of study of this thesis, codec distortion is not considered and background noise is not added to the speech samples. Hence, based on the digital watermarking technology, we proposed to embed watermarks on the time domain to assess the objective quality. After the watermarks are embedded, when the packet losses, the watermark losses too. There-

fore, through the assess of correct extracted watermark, the subjective quality can be predicted.

For the watermarking scheme on packet-loss, we still employ the technology of quantization which used for the above three attacks. Because packet-loss affects the speech quality in the time domain, over the speech file, the spectrum information (high, middle and low frequency) has the equal opportunity to be lost. The advantage of using the frequency information to consider the watermark embedding is not obvious here. Therefore, we adopt the quantization based watermarking scheme that quantizes the sample values into valid sample values (to which the quantization step need to be added) and invalid ones (to which the quantization step need not to be added) [23]. The watermark embedding and extraction are done in a way very similar to the one in Section 4.2.1 and 4.2.2, except for that the watermark bits are embedded in the temporal domain.

This quantization scheme quantizes a sample value V(s) and assigns new value to the sample V(s) based on the quantized sample value. The watermarked sample value V(w) is calculated from the Equ. (4.5)

$$V(w) = \begin{cases} V(s) & \text{if } Q(s) = w(i) \\ V(s) + \Delta & \text{if } Q(s) \neq w(i) \end{cases} \tag{4.5}$$

where $w(i)$ is the watermark bit (1 or 0) to be embedded, $Q(s)$ is the quantization value (1 or 0) of the selected sample point and Δ is the quantization step.

For the only considering of packet-loss distortion, the watermark is very robust if there is no packet lost. So we can use the same quantization scale for all speeches. Based on the experiment, we choose 1000, on which the watermark disturbances on host speech signal are beyond the sensing of human ears and we can still extract 100% watermark bits before the watermarked speech passes into the network.

4.3 Quality Evaluation

After watermark extraction, the $PCEW$ is calculated by comparing the extracted watermark with the original one using Equ. (4.6):

$$PCEW = \frac{\sum_{j=1}^{N} W(j) \oplus W^*(j)}{N} \tag{4.6}$$

where W is the original watermark, W^* is the extracted watermark, N is the length of the watermark, and $\oplus$ is exclusive-OR operator. The $PCEW$ value lies between 0 and 1.

In the process of quality assessment, we predict the speech quality from $PCEW$ based on the mapping between ITU-T P.862 PESQ MOS (hereinafter referred to collectively as "MOS") and $PCEW$, which we obtained from the test sample speeches. Hence, to give correct predictions, a mapping between them must be calibrated. In our method, we divide the mapping into 10 segments with a $PCEW$ interval of 0.1 and calculate the corresponding MOS at the points where the $PCEW$ values are equal to 0.1, 0.2, $\cdots$, and 1. After the $PCEW$ is calculated from the watermark extraction process, its location in the mapping segments is found based on Equ. (4.7):

$$\begin{cases} P_S = p & \text{if } p \leq PCEW < p + 0.1 \\ & (p = 0, 0.1, ..., 0.9) \\ P_E = P_S + 0.1 & \end{cases} \tag{4.7}$$

where P_S and P_E are the percentages at the start and end point of the mapping segment respectively.

And then, the PESQ score is predicted by the following equation:

$$DWMOS = MOS_S + \frac{MOS_E - MOS_S}{P_E - P_S} \times PCEW \tag{4.8}$$

where $DWMOS$ is the predicated MOS score; MOS_S and MOS_E are MOS values at the start and end point of the mapping segment; and P_S and P_E are defined in Equ. (4.7).

4.4 Two Performance Indices

To evaluate the performance of our objective quality assessment method, we introduce the correlation coefficient and residual error to evaluate the correlation between $DWMOS$ and MOS.

4.4.1 Correlation Coefficient

The fit between $DWMOS$ and $PESQMOS$ can be measured by calculating the correlation coefficient which indicates the performance of the algorithm. The correlation coefficient is calculated with Pearson's formula [35]:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{4.9}$$

In this formula, x_i is the condition $PESQMOS$ for condition i, and $\bar{x}$ is the average over the values of x_i; y_i is the averaged $DWMOS$ for condition i, and $\bar{y}$ is the average over the values of y_i. In our experiments under the effects of Gaussian noise, for 100 known benchmark experiments (10 speech samples, taking account of 10 different disturbance parameters) the average correlation was 0.9773. For the unknown set of 100 experiments used in the final validation - experiments that were unknown during

the development of DWESQ - the average correlation was 0.9759 which is pretty close to that of known set.

4.4.2 Residual Errors

The residual error is used to assess the measurement accuracy. The regression mapping removes any systematic offset between the $DWMOS$ and the $PESQMOS$, minimizing the mean square of the residual errors:

$$e_i = x_i - y_i \tag{4.10}$$

Various measures may be applied to the residual errors to give an alternative view of the closeness of $DWMOS$ to $PESQMOS$. For example, the histogram of the absolute residual errors $|e_i|$ provides a quick view of how frequently errors of different magnitudes occur. Take Gaussian noise distortion as an example. We use a set of 100 samples in the final validation. The average residual error distribution showed that the absolute errors of 23% examples are less than $0.05MOS$ and 98% of samples are less than $0.50MOS$. More details are shown in Table 5.6.

4.5 Implementation Strategies

In this Section, we discuss the strategies used to implement our digital audio watermarking. The watermark is embedded by quantizing the selected coefficients with an optimized quantization parameter QS in the discrete wavelet transform domain. We make use of the discrete wavelet transform domain to embed the watermark because it provides both the frequency and the temporal information at the same time. Based

on the frequency information, we can locate the watermark bits in different frequency levels, as discussed in Section 4.2.1. On the other hand, making use of the temporal information, at each frequency level, the wavelet coefficients are divided into blocks in each of which one watermark bit is embedded. In this way, the watermark is embedded in all the possible frequency and spatial regions of an speech. The localization of the watermark can identify the regions of watermarked speech that have undergone tampering. The globally spread watermark bits are sensitive to all kinds of possible attacks.

4.5.1 Location of Embedding Watermark

In ordinary DWT watermarking scheme, the middle frequency coefficients and blocks are usually selected for embedding watermark data, so it can achieve the balance between robustness and imperceptibility [41][43]. Since we want to develop a speech quality measurement method which should assess any potential attack over the entire speech signal, we have to embed as many watermark bits as possible in the DWT coefficient levels. Nevertheless, no matter what level is used to embed watermark, there are disadvantages. If we embed in high frequency coefficients, the watermark can be easily removed by distortions such as low-pass filtering. If embed in low frequency levels, the watermark strongly affects the perceptual quality. If embed in middle frequency levels, the watermark becomes very robust and it is difficult to reflect the degradation of the speech quality. To balance, for measuring the effects of MP3 compression, Gaussian noise addition, and low-pass filtering, we embed a different number of watermark bits in different coefficient levels, that is, more in middle frequency levels and less in high- and low-frequency levels [47]. The watermarks embedded in different locations have different robustness. By doing so we can measure the quality of the watermarked

speech.

However, this strategy is not applicable to the evaluation of speech quality under the effects of packet loss because the lost packet contains all spectrum of frequencies. According to the assignment of watermark location mentioned above, most watermark bits are embedded in the middle frequency levels. At the same packet-loss rate, more the middle frequency level clusters contended in the lost packet(s), more watermark bits will be removed. Subsequently, the *PCEW* value will vary a lot. This results in an incorrect MOS predication with the dispersing *PCEW*.

Considering that the packet-loss effect is more related to the temporal information, to solve this problem, we embed the watermark in time domain instead of frequency domain. In this case, for packet-loss distortion, we proposed a special watermarking method presented in Section 4.2.3.

4.5.2 Balance between Fidelity of Watermarked Speech and Accuracy of Prediction

There are two important parameters, the *MOS* of undistorted watermarked speech and the data payload (total number of watermark bits), which significantly affect the accuracy of the MOS prediction.

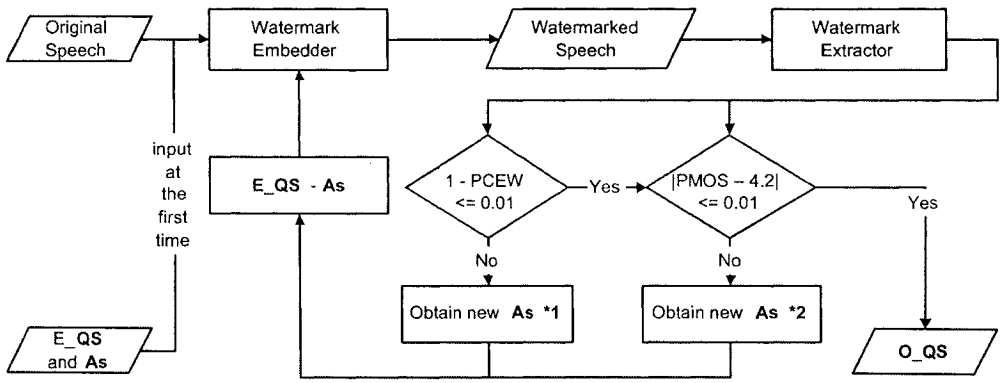
An ideal watermarking algorithm generates watermarked signal with almost the same perceptual quality as the original signal. However, this may result in the extreme fragility of the watermark. Hence, the relationship between the fidelity and robustness must be balanced. For our method, this requirement is more important. If the *MOS* of undistorted watermarked speech is very high, e.g. 4.4 *MOS*, the watermark can be easily removed even when the distortion is not so strong. So we cannot accurately

predict the MOS when the quality is not good. On the other hand, low MOS of undistorted watermarked speech, e.g. 4.0, results in a relatively strong robustness of the watermark. We cannot expect the system to give a correct predication of the speech quality if the $PCEW$ is always 100% when the MOS is over 3.8. In this case, when we detect 100% watermarks, we can only predict the MOS in the range of [3.8, 4.5], which is not accurate. Through the experiments, we obtained the optimum MOS value, 4.2, for undistorted watermarked speeches.

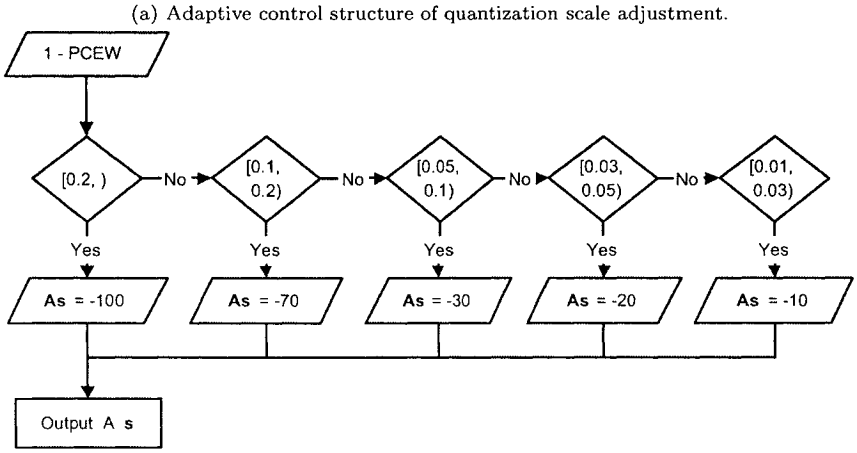
On the other hand, for the payload of the watermarking algorithm, not only should we consider the fidelity of the watermarked signal, but also the accuracy of $PCEW$ and the number of the available samples. The more watermark bits are embedded, the more accurate the $PCEW$ is, but the worse the quality of the watermarked speech. Meanwhile, the volume of the sample data suggests the maximum number of watermark bits. For example, in a 1-minute speech (16KHz), there are $16000 \times 60 = 960000$ samples. If we embed watermark with a 10-level DWT, the lowest level (cD10) has only $960000 \div 2^{10} = 938$ samples, which can only carry no more than 4 watermark bits in our method. Based on our strategy in Section 4.5.1, we embed more watermark bits in middle frequency levels and less watermark bits in high- and low-frequency levels, we can embed maximum 450 watermark bits. Actually, we embedded 400 bits in our experiments.

4.5.3 Optimization of Quantization Parameter

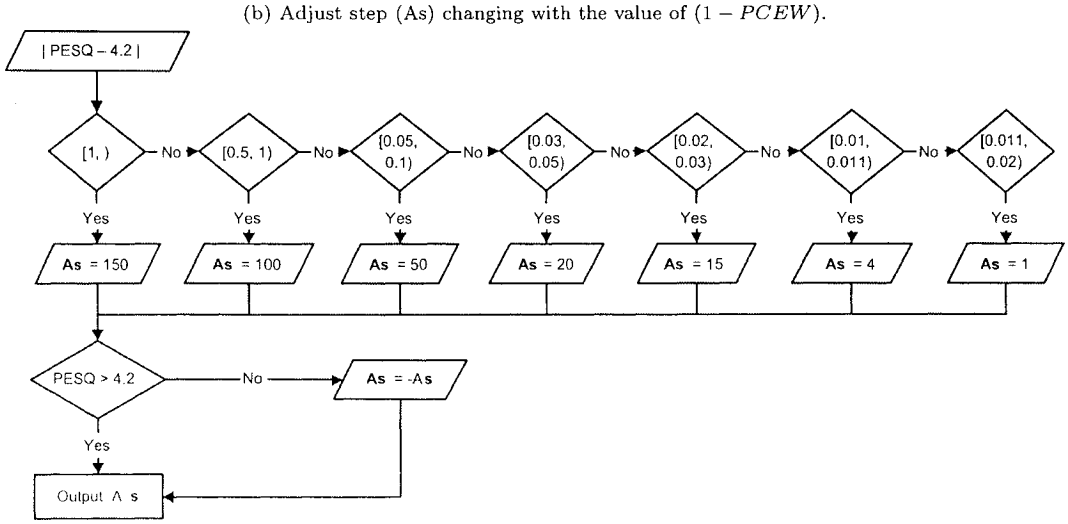
Different speech signals comprise different frequencies and amplitudes, therefore they have different robustness to the same distortion. In addition, for any discrete wavelet transform, the DWT coefficients are real numbers. Their ranges vary and depend on both decomposition level and speech itself. If we use quantization technique, we have



(a) Adaptive control structure of quantization scale adjustment.



(b) Adjust step (As) changing with the value of $(1 - PCEW)$.



(c) Adjust step (As) changing with the value of $|PESQMOS - 4.2|$.

Figure 4.10: Quantization scale adjustment.

to calculate the quantization step for each level of each speech. This will result in too many parameters sending to the watermark extractor. In fact, for quantization, it is not the size of quantization step but the scale between the quantization step and the difference of maximum and minimum coefficient values affects the fidelity and the result. Hence, in our method, we introduce the term of quantization scale (QS) to obtain the quantization step Δ by using the following equation (repetition of Equ. (4.2)):

$$\Delta = \frac{\max_V - \min_V}{QS} \tag{4.11}$$

where Δ is quantization step, $\max_V$ and $\min_V$ are the maximum and minimum value of the coefficients respectively in a specific decomposition level, and QS is quantization scale. We use the same QS for all levels and for sure, each quantization step has the same scale to the difference between the highest and lowest coefficients. In other words, the quantization step Δ will normally be different because each level has different max_V and min_V.

To obtain the optimal QS for each speech signal, we employed the adaptive control method, as shown in Figure 4.10, to conduct the recursive test on both $PCEW$ and PESQ score which ranges from -1 (bad) to 4.5 (excellent). To keep the excellent quality for the watermarked signal, we set the required PESQ score as 4.2 before distortion which we have discussed in Section 4.5.2. Furthermore, in order to obtain the stable and accurate correlation between $PCEW$ and MOS, all undistorted watermarked signal should have the same fidelity. Therefore, we estimate the QS first, and then employ the adaptive control method to carry out the recursive watermark embedding and extraction to guarantee that the $PCEW$ is in the range of (0.99, 1) and the PESQ MOS is in the range of (4.19, 4.21), before any distortion is applied. As a result, when

the two requirements are reached, the algorithm outputs the optimal QS. In Figure 4.10, E_QS and O_QS are the estimated and optimized QS respectively, $PMOS$ represents PESQ MOS, and A_S is the adjust step.

There are two strategies to accelerate the speed of quantization scale adjusting algorithm.

First, we get the estimated QS from the waveform property of the speech. According to our experiments, we found that the QS is related to the sample value of the speech signal. Mostly, the larger the average peak sample value of the speech, the bigger the QS. For instance, as shown in Figure 4.11, the average peak sample value of speech 4 and speech 8 are approximately 0.35 and 0.16. Correspondingly, our algorithm obtains the optimized quantization scales being 373 and 181 respectively. Therefore, at the first time of input in the quantization scale adjust scheme, we can set the estimated quantization scale as 350 and 160 for speech 4 and speech 8. In this case, we can decrease the adjusting times.

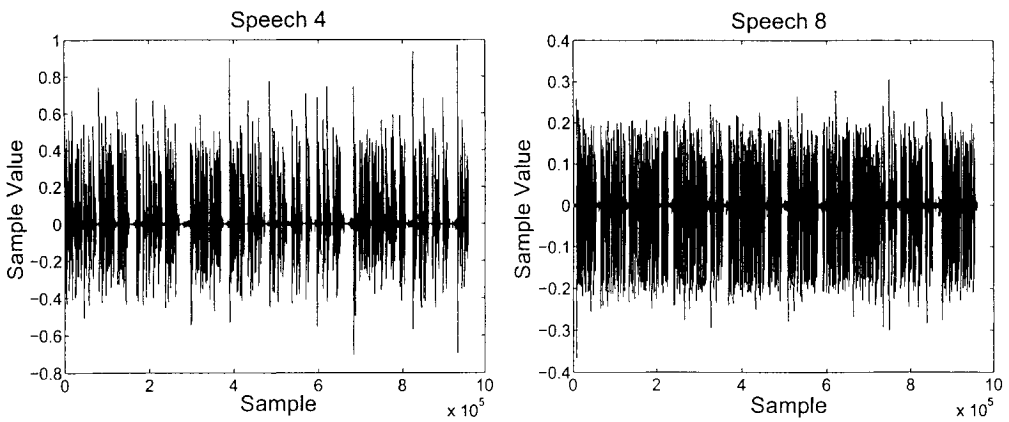


Figure 4.11: Sample data of speech 4 and 8.

Second, we vary the adjusting step. The adjusting step is the difference between the two Quantization Scales at two consecutive adjusting points of time during the process of quantization scale adjusting. A fixed adjusting step would be awfully time-consuming

and is difficult to force the difference between the real value and target value fall into a predefined range (in adaptive control, the algorithm cannot reach the exact target value and normally there is a little difference). For example, we define the allowed error, 0.01, to the target MOS value, 4.2. During the adjusting process, if the MOS is in the range of [4.19, 4.21], we can say that the target MOS is reached. To reduce the recursive adjusting time, this algorithm dynamically change the Adjusting step (A_S) according to the value ($1-PCEW$) and $|PESQMOS-4.2|$. The larger the difference, the wider the As. For example, as shown in Figure 4.10 (b), when the value ranges of ($1-PCEW$) are $[0.2,)$, $[0.05, 0.1)$ and $[0.01, 0.03)$, the changing steps are -100, -30 and -10 respectively. $|PESQMOS - 4.2|$ also affects the value of As and we can calculate QS based on Equ. (4.12).

$$QS(t) = \begin{cases} QS(t-1) - As & \text{if } PESQMOS > 4.2 \\ QS(t-1) + As & \text{if } PESQMOS < 4.2 \end{cases} \tag{4.12}$$

where $QS(t)$ is the QS at the adjusting time t. If $PESQMOS$ is larger than 4.2, the watermark energy is weak. So we have to increase the quantization step Δ to decrease the QS. Hence, in iteration t, As is subtracted from $QS(t-1)$ to get $QS(t)$. For similar reason, As is added to $QS(t-1)$ when the $PESQMOS$ is smaller than 4.2. As demonstrated in Figure 4.13 (c), we can see that the closer to the target PESQ MOS score, 4.2, the smaller the Adjusting step. With the changing As, $PESQMOS$ is approaching the target value exponentially and very fast. In this case, the adjusting time is different for different speech. As shown in Figure 4.12, we can see that the maximum adjusting time is 12 iterations (Speech 2), and the minimum adjusting time is 1 iteration (Speech 1).

From Figure 4.13 (a) and (b), we can see that the PESQ score is exponentially ap-

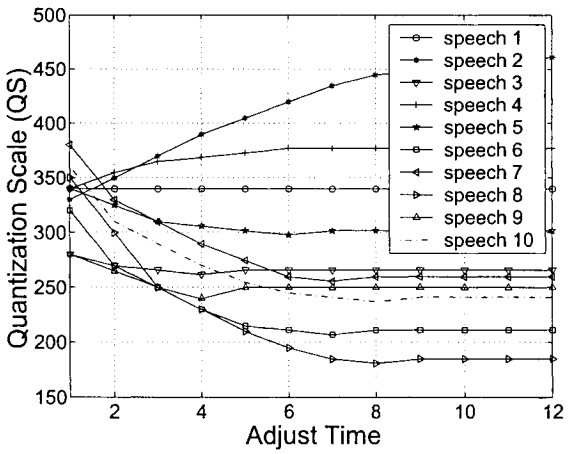
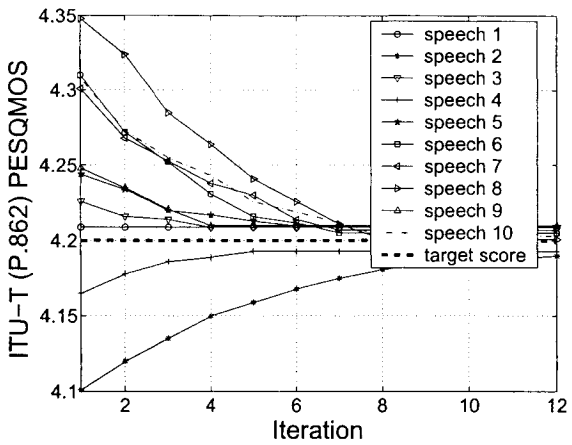
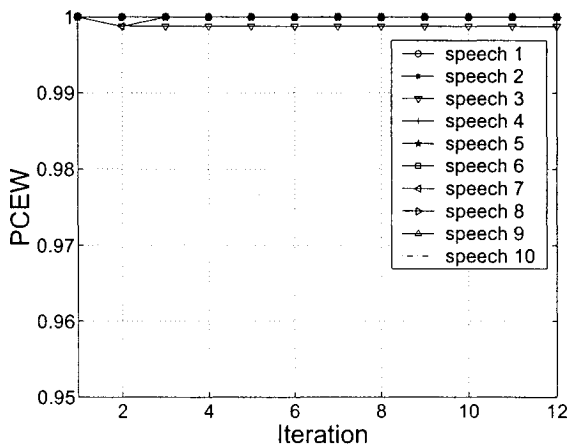


Figure 4.12: Quantization scale optimization.

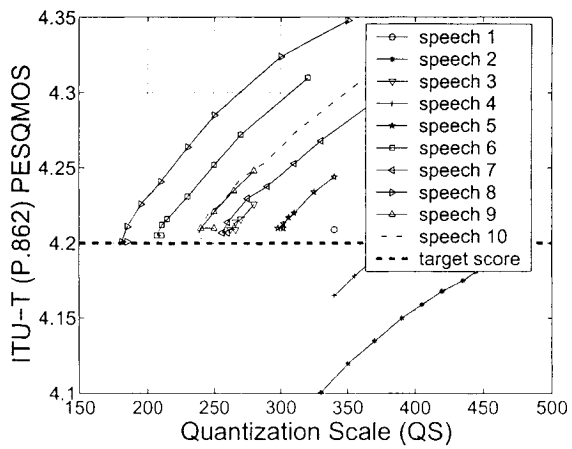
proaching the target line, $PESQMOS = 4.2$. Meanwhile, the $PCEW$ is approximately equal to 1.0 with the permitted error range of 0.01.



(a) *PESQMOS* approaching the target value, 4.2.



(b) *PCEW* approaching the target value, 1.



(c) Changing of Adjust step (As).

Figure 4.13: Process of adjusting the quantization scale (QS).

Chapter 5

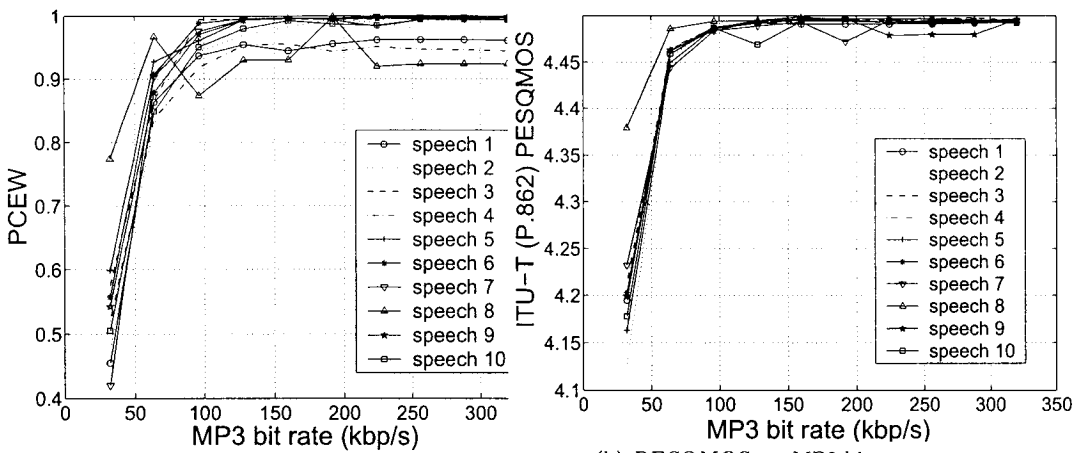
Experimental Results and Evaluation

In this chapter, we first discuss how to select the source material. Then we evaluate the performance of the proposed scheme against noise addition, filtering, MP3 compression and packet-loss. Those are the major distortions on the internet and network transmissions. Finally, we will illustrate the performance results.

5.1 Source Material

Following the criteria given by ITU-T P.830 [48] and P.800 [31], test signals are selected to include speech bursts (typically 1-3 seconds duration) separated by silent periods. These speeches are active between 40 percent and 80 percent of the time.

We selected two sets of samples that include both female and male speeches for different purposes. Set 1 was used for the linear mapping calibration between $PCEW$ and $PESQMOS$. Set 2 was used for the validation test. Both sets contain 10 speeches,



(a) *PCEW* vs. MP3 bit rate.

(b) *PESQMOS* vs. MP3 bit rate.

Figure 5.1: *PCEW* and PESQ MOS under MP3 compression.

which are stored in 16-bit, 16 KHz linear PCM format. For the distortions, we set SNR from 5 to 50 with an interval of 5 for Gaussian noise, bit rate from 32 to 320 Kbps with an interval of 32 Kbps for MP3 compression, threshold frequency from 1 to 29 KHz with an interval of 4 KHz for low-pass filtering, and packet-loss rate from 5% to 50% with an interval of 5% for packet loss distortion. Therefore, for the validation test, there are 100 speech samples for Gaussian noise, MP3 compression and packet loss respectively, and 80 for low-pass filtering.

Table 5.1: Mapping between *PCEW* and *PESQMOS* under MP3 compression.

PCEW	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PESQMOS	3.9272	3.9921	4.0571	4.1221	4.1871	4.2520	4.3324	4.4127	4.4712	4.4937

5.2 MP3 Compression

For the MP3 compression, as illustrated in Figure 5.1, both our *PCEW* and the ITU-T's PESQ score curves show that over 128K bit rate, the quality is approximately at

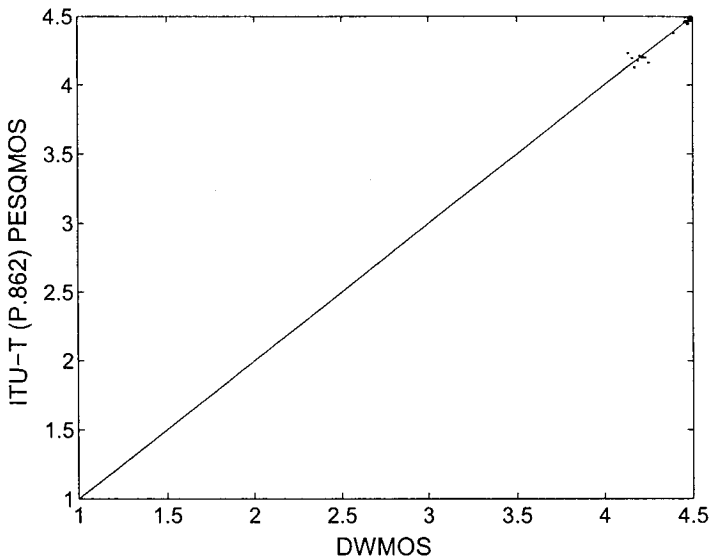
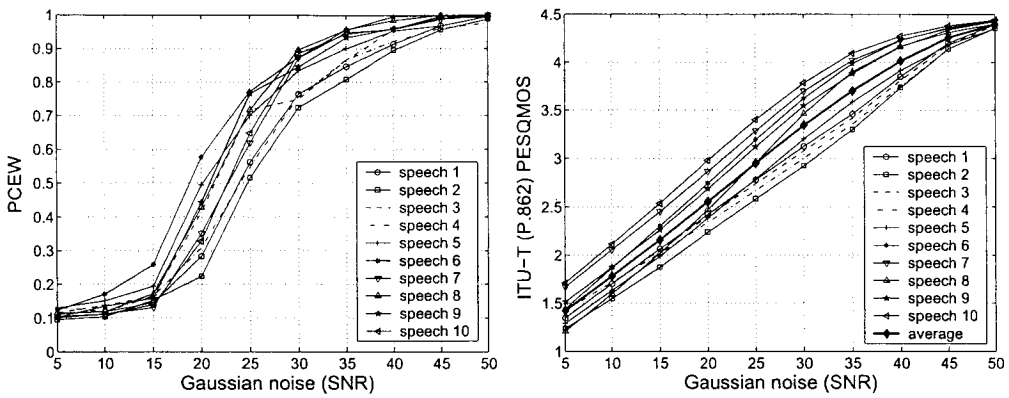


Figure 5.2: Predicated *MOS* vs PESQ MOS for MP3 compression. (100 samples)

the same "excellent" level. This means, the *PCEW* value is approximately 100% and *MOS* is almost 4.5. When the bit rate is between 128Kbps and 64Kbps, the quality decrease linearly. When the bit rate is under 64 Kbps, the quality decreases very fast. The PCEW ranges from 0.42 to 1, while the MOS varies between 4.13 and 4.5.

Based on these experimental results, we get the mapping between *PCEW* and *MOS* as shown in Table 5.1. Hence, in practice, once the *PCEW* value is computed out from the watermark extraction system, we can predict the MOS score with Equ. (4.8) in Section 4.3 by referring to the mapping table. The experimental results suggest that *DWMOS* and *PESQMOS* have very close correlation (refer to Figure 5.2), and that our quality evaluation has a pretty good accuracy on MP3 compression (refer to Section 5.6). Figures 5.2, 5.5, 5.6 and 5.10 show the correlation between the *PESQMOS* and *DWMOS* for different kind of distortions. If the *DWMOS* and *PESQMOS* have an absolute match, all the sample points should be on the solid line from point (1,



(a) *PCEW* vs. SNR. (b) *PESQ* vs. SNR.

Figure 5.3: PCEW and PESQ MOS under Gaussian noise distortion.

1) to point (4.5, 4.5). The closer the sample points to the solid line, the better the performance of the *DWMOS* is.

In addition, we implemented the further test on non-speech samples, such as music and pop songs. We found that the *PCEW* curves of those sample audio signals are also very close. Furthermore, the *PCEW* values have the same distribution and descent trend as that of speech signals. This indicates that this method can also be used for audio quality measurement.

5.3 Gaussian Noise

Under the effects of Gaussian noise, the *PCEW* and *MOS* have different descending speed. The MOS curves are more linear than the PCEW curves, as shown in Figure 5.3 (a) and 5.3 (b). In the experiment, we choose 10 sample speeches. For each speech, we simulate 10 different Gaussian Noise effects. The strength of those effects is from $SNR = 5$ to $SNR = 50$ with an interval of 5. As shown in Figure 5.3 (a), the curves of *PCEW* of the sample speeches have three distinct decreasing speeds, which distributed

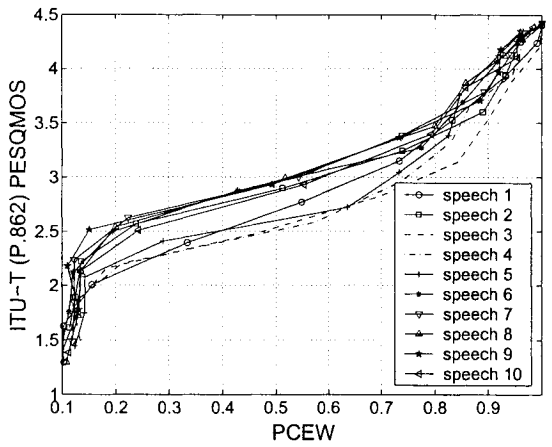


Figure 5.4: Mapping between PCEW and PESQ MOS for Gaussian noise distortion.

in the ranges of [5,15], [15,25], [25,50]. On the other hand, the curves of MOS are linear and almost 45 degrees, refer to Figure 5.3 (b). However, the differences do not affect the accuracy of DWMOS because both the $PCEW$ and $PESQMOS$ decrease with increasing noise strength, and all the curves are very close. We can obtain a perfect mapping between PCEW and MOS, as shown in Figure 5.4. We can use the $PCEW$ values to accurately predict the PESQ scores according to the mapping table as shown in Table 5.2.

Table 5.2: Mapping between $PCEW$ and $PESQMOS$ under Gaussian noise.

$PCEW$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$PESQMOS$	1.5576	1.7911	2.0158	2.1464	2.2691	2.4061	2.5962	2.8675	3.3273	4.3909

The test results indicate that MOS can be accurately predicted. As being illustrated by Figure 5.5 and Table 5.6, the $DWMOS$ has very close correlation to the $PESQMOS$, since all the sample points are distributed close to the solid straight line.

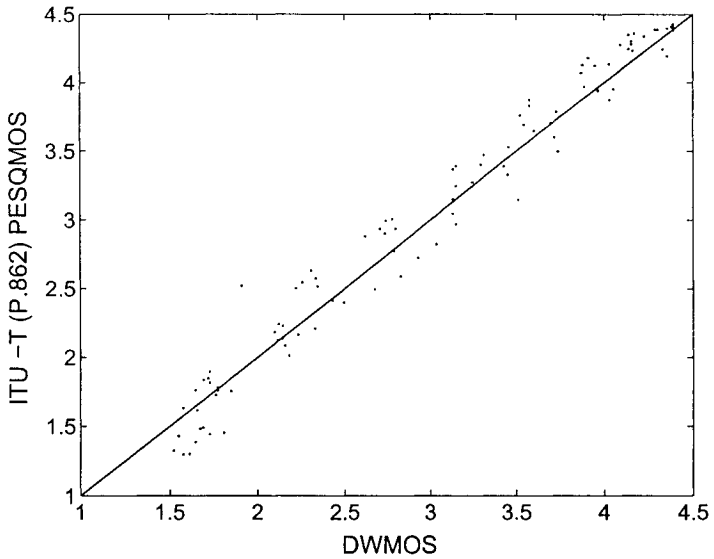


Figure 5.5: Predicted MOS vs PESQ MOS for Gaussian noise. (100 samples)

5.4 Low-pass Filtering

Under the low-pass filtering distortions, the PCEW curves are close and almost like straight lines with the values ranging from 0.2 to 0.99. Meanwhile, the $PESQMOS$ is not affected much by the low-pass filtering, with a lowest value around 4.07. When the threshold frequency is over 9 KHz, the PESQ MOS values are approximately the same and near 4.5. When the threshold frequency is below 5 KHz, the effect is bit more obvious. However, because the mapping curves between $PCEW$ and $PESQMOS$ are very close, we can predicate the MOS with extremely small errors, as shown in Table 5.6. From Figure 5.6, we can see that the $DWMOS$ and $PESQMOS$ have a very good correlation.

Table 5.3: Mapping between *PCEW* and *PESQMOS* under low-pass filtering.

PCEW	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PESQMOS	4.1042	4.2472	4.3903	4.4772	4.4887	4.4915	4.4916	4.4919	4.4927	4.4930

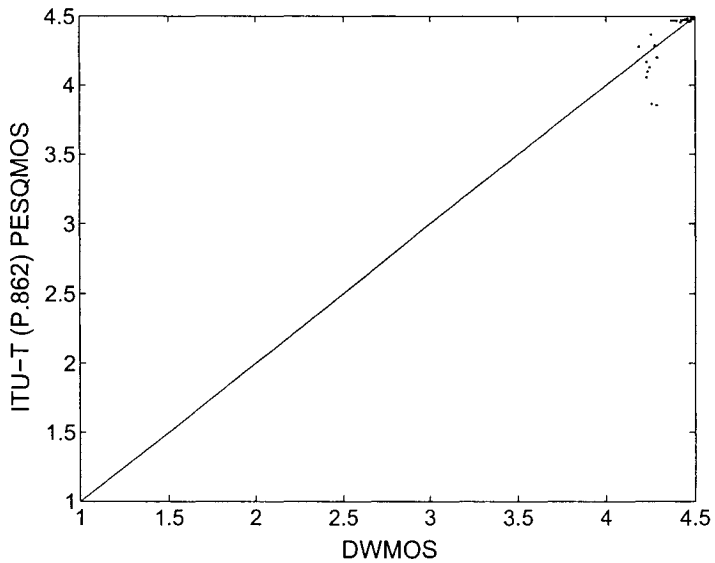


Figure 5.6: Predicted *MOS* vs PESQ MOS for low-pass filtering. (80 samples)

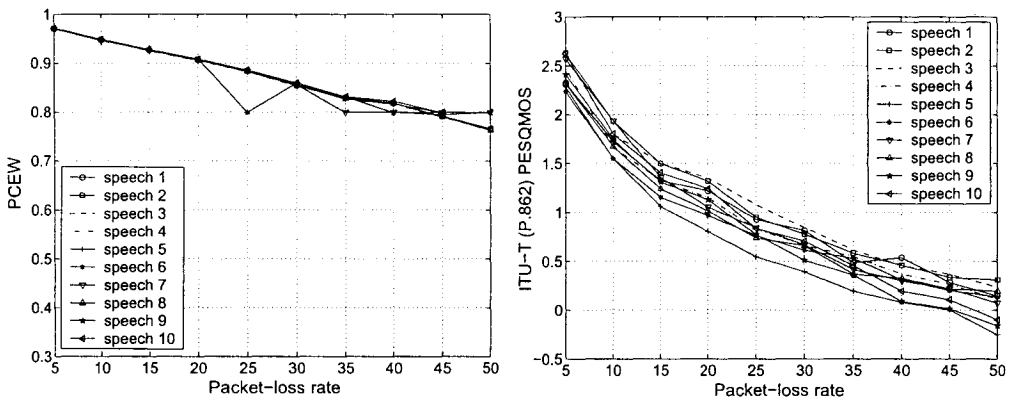
5.5 Packet-Loss

The distribution of lost packet is simulated randomly and the loss rate ranges from 5% to 50% with an interval of 5%. Besides the watermarking algorithm, the test signal is also different from the other experiments of attacks. We only need 8 seconds for each sample speech because watermark is embedded in the sample value directly and it is enough to provide a very good assessment of distortions with the certain amount of watermark.

As shown in Figure 5.7, the effects of packet loss on speech quality are very obvious for the *PESQMOS*, it can make *PESQMOS* vary from -0.5 to 2.7. But it can

Table 5.4: Mapping between *PCEW* and *PESQMOS* under packet loss distortion.

PCEW	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PESQMOS	0.2266	0.3252	0.3910	0.5950	0.7812	1.0475	1.2997	1.8011	2.5753	4.5000



(a) PCEW (b) PESQ

Figure 5.7: *PCEW* and PESQ MOS under the packet-loss attack. (100 samples)

only make the *PCEW* vary from 0.75 to 0.97. However, between *PESQMOS* and *PCEW*, there is still a pretty good mapping relationship which can be used to predict the speech quality, as shown in Table 5.4. According to the experiment, we have to take into account that this mapping is very sensitive to the silence interval in the speech file. The *PCEW* values are distributed in the range from 0.75 to 0.97. However, if the speech has more silence intervals, the *PCEW* maybe abnormal and the accuracy of quality prediction would be degraded. By analyzing the *PCEW* results, it is easy to find that the abnormal *PCEW* occurs in the area when loss rate is over 15%. In this area, the range of normal *PCEW* is from 0.75 to 0.90.

To solve this problem, two compensation methods are introduced. The first method is to arbitrarily set the *PCEW* to be one of the random number between 0.75 and 0.90 when the *PCEW* is less than 0.75, as shown in Figure 5.8 (a). The other method,

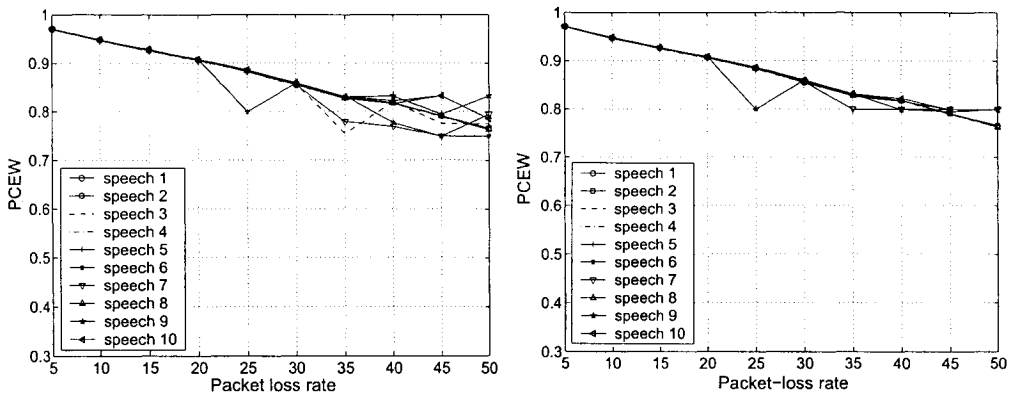
as illustrated in Figure 5.8 (b), is to arbitrarily force the abnormal $PCEW$ to be 0.8 because it is closer to the normal value. The experimental results show that the second method is better and it provides a very high correlation between $DWMOS$ and $PESQMOS$. From Table 5.5, the second method provides a higher correlation, 0.9744, between $DWMOS$ and $PESQMOS$. In the table, "Known" was used for linear mapping calibration and "Unkown" was for validation test. "ARE" shorts for Absolute Residual Error; and "$ARE \leq C$" means that the percentage of samples for which the ARE between the $DWMOS$ and PESQ MOS is less than or equal to C. More interesting is that if the loss rate is less than 25%, the correlation can reach 0.9928. The direct comparison between $DWMOS$ and $PESQMOS$ is illustrated in Figure 5.9. The samples are 10 different speeches under 10 different rates of packet-loss distortions, so there are 100 test speeches. In the figure, every ten samples (from 1 to 10, 11 to 20, $\cdots$) are one speech under 10 different packet-loss rate distortions. The curve marked with "o" is the DWMOS curve, and the other one is the PESQMOS curve. When the MOS is over 0.5, the two curves are almost identical. As illustrated in Figure 5.10, we found that $DWMOS$ has a perfect correlation with $PESQMOS$.

Table 5.5: $DWMOS$ accuracy comparison between random and average compensations for packet loss distortion.

	Corr. coeff.		Mean	*ARE*	*ARE*	*ARE*
Compensation	Known	Unknown	*ARE*	≤ 0.05	≤ 0.25	≤ 0.5
Random	0.9303	0.9531	0.1640	21%	81%	96%
Average	0.9720	0.9744	0.1832	26%	81%	99%

5.6 Performance Results

We use correlation coefficient and residual error between $DWMOS$ and PESQ MOS to quantify the performance of our digital watermarking based speech quality evaluation



(a) Randomly set the abnormal $PCEW$ in the range of 0.75 to 0.90. (100 samples)

(b) Arbitrarily set the abnormal $PCEW$ to 8.0. (100 samples)

Figure 5.8: Compensation of the abnormal $PCEW$.

Table 5.6: Overall indicators of $DWMOS$ accuracy.

Effects	*Corr. coeff.*		Mean	*ARE*	*ARE*	*ARE*
	Known	Unknown	*ARE*	≤ 0.05	≤ 0.25	≤ 0.5
MP3 Compression	0.9839	0.9727	0.0101	98%	100%	100%
Gaussian Noise	0.9773	0.9759	0.1711	24%	85%	98%
Low-pass filtering	0.8501	0.8493	0.0361	80%	98 %	100%
Packet loss	0.9720	0.9744	0.1832	26%	81%	99%

method. Table 5.6 shows the results for MP3 compression, Gaussian noise addition, low-pass filtering and packet loss. From Table 5.6, we can see that our $DWMOS$ is very well correlated to PESQ MOS.

5.7 Summary

In this chapter, we first discussed the selection of source material and the distortion strength added to the sample speeches. Then we tested our watermarking scheme with different kind of attacks. In the last, through the error analysis, the performance results show that our watermarking scheme provides accurate predictions of subjective quality for speech signals.

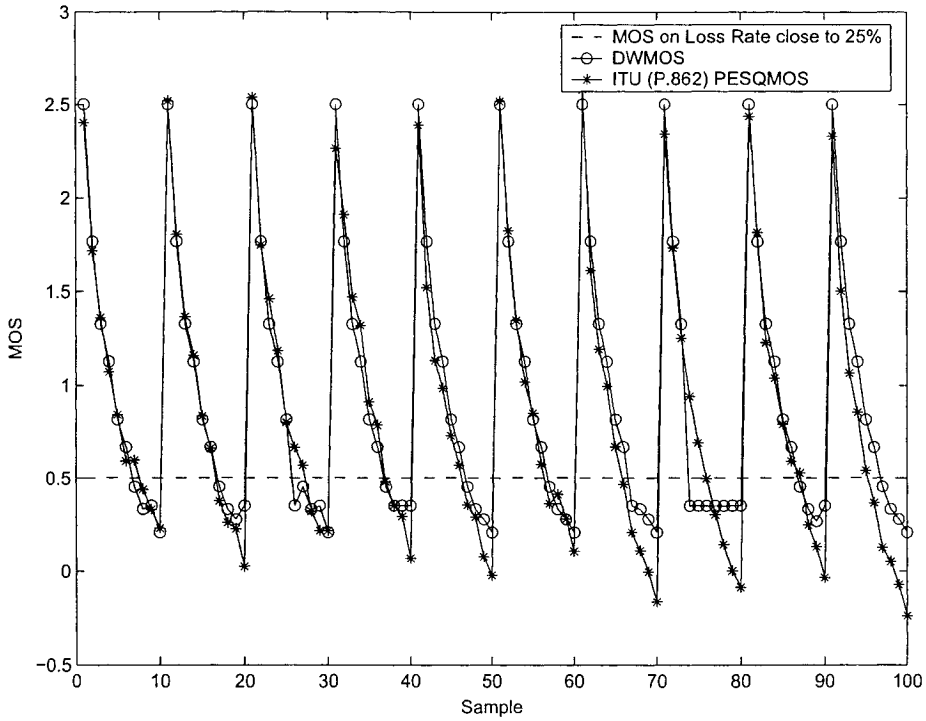


Figure 5.9: *DWMOS* vs. *PESQMOS* on sample points (packet-loss attack).

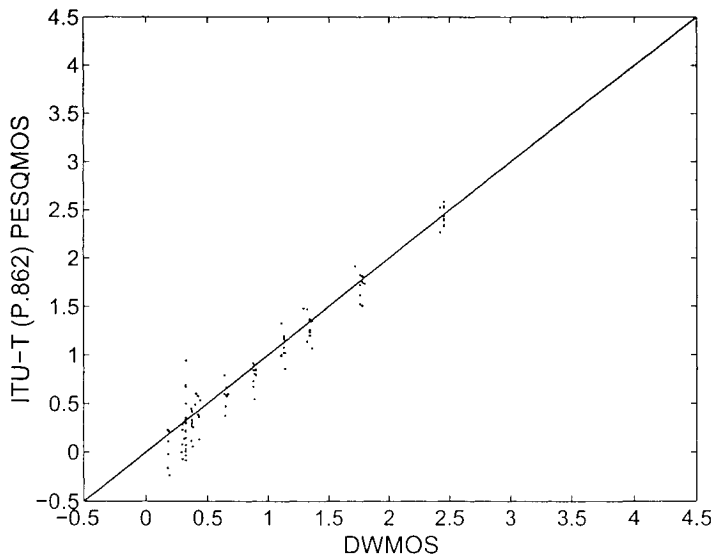


Figure 5.10: Predicted *MOS* vs PESQ MOS for packet loss. (100 samples)

Chapter 6

Conclusions and Future Works

In this thesis, we proposed an objective speech quality evaluation method using digital audio watermarking. This method is based on the DWT and quantization. It is applicable to both male and female speakers and different languages. Furthermore, the original speech signal is not needed, neither the training database.

The quantization plays an important role in this scheme. We use it to embed and extract the watermark. The selection of quantization parameter affects the efficiency and accuracy of this method. We introduced an adaptive control algorithm to obtain the optimal quantization parameter - Quantization Scale, which satisfies both the good fidelity of the signal and approximately 100% watermark detection rate before any distortion is applied, for each input speech signal.

For the quality evaluation, we built a linear mapping between $PCEW$ and $PESQMOS$. Therefore, after the $PCEW$ is obtained from the watermark extraction system, we can predict the MOS score, $DWMOS$ from the mapping. Comparing the $DWMOS$ and ITU-T's PESQ (P.862) scores, we validated the accuracy of this method through the performance analysis. The experimental results show that this method

gives accurate predictions of subjective quality for speech signals. Furthermore, based on our test on non-speech samples, it can also be used for audio quality measurement.

For the speech quality evaluation on packet-switch network, it is important that the digital watermark should survive the distortion of most codec algorithms, such as ITU-T G.721, G.729, and so on. As for our future work, we will improve our algorithm to make it applicable in this area.

Bibliography

[1] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq) - a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, USA*, pp. 749–752, May 2001.

[2] J. G. Beerends, A. W. Rix, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment, part ii psychoacoustic model," in *Journal of the Audio Engineering Society*, vol. 50, pp. 765–778, October 2002.

[3] A. v.E. Conradie, R. Miikkulainen, and C. Aldrich, "Adaptive control utilising neural swarming," in *Proceedings of the 2002 Genetic and Evolutionary Computation Conference (GECCO-2002), New York, USA.*, pp. 60–67, 2002.

[4] T. H. Falk and W.-Y. Chan, "Objective speech quality assessment using gaussian mixture models," in *Proceedings of the 22nd Biennial Symposium on Communications, Kingston, Ontario, Canada*, pp. 169–171, June 2004.

[5] L. Ding and R. Goubran, "Assessment of effects of packet loss on speech quality in voip," in *Proceedings of the IEEE International Workshop on Haptic, Audio*

and Visual Environments and their Applications (HAVE 2003), Ottawa, Ontario, Canada, pp. 49–54, September 2003.

[6] A. Takahashi, H. Aoki, and H. Yoshino, "Standardization of speech quality assessment of ip telephony," *Global Standardization Activities*, vol. 2, no. 3, pp. 82–84, Mar 2004.

[7] S. ANN, "Current status and prospects of speech processing technology in korea," in *1994 Intemational Symposium on Speech, Image Processing and Neural Networks*, vol. 1, pp. 133–136, April 1994.

[8] H. P.M. and H. M.C., "Speech synthesis and the spectral estimation problem," in *IEEE Colloquium on Spectral Estimation Techniques for Speech Processing*, vol. 2, pp. 1–7, Feburary 1989.

[9] B.-H. Juang and S. FURUI, "Automatic recognition and understanding of spoken language - a first step toward natural human-machine communication," in *Proceedings of the IEEE*, vol. 88, pp. 1142–1165, August 2000.

[10] T. A. Hall, "Objective speech quality measures for internet telephony," in *Voice over IP (VoIP) Technology, Proceedings of the International Society for Optical Engineering (SPIE)*, vol. 4522, pp. 128–136, August 2001.

[11] M. Schroeder, B. Atal, and J. Hall, "Optimizing digital speech coders by exploiting the masking properties of the human ear," in *Journal of the Acoustical Society of America (J.A.S.A.)*, vol. 66, pp. 1647–1652, Dec. 1979.

[12] M. Karjalainen, "Sound quality measurements of audio systems based on models

of auditory perception," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 211–215, 1984.

[13] K. Brandenburg and T. Sporer, "Nmr and masking flag: Evaluation of quality using perceptual criteria," in *Proceedings of the Audio Engineering Society (AES) 11th International Conference*, no. 1192 in Special Issue on Authentication,Copyright Protection and Information Hiding, 1192, pp. 169–176.

[14] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," in *Journal of Audio Engineering Societ*, vol. 40, pp. 477–480, Dec. 1992.

[15] I. Cox, J. Kilian, F. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, 1997.

[16] L. R. Matheson, S. G. Mitchell, T. Shamoon, R. E. Tarjan, and F. Zane, "Robustness and security of digital watermarks," in *Financial Cryptography*, pp. 227–240, 1998.

[17] I. Cox, M. L. Miller, and J. A. Bloom, *Digital watermarking*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002.

[18] I. J. Cox, M. L. Miller, and J. A. Bloom, "Watermarking applications and their properties," in *Proceedings of the IEEE International Conference on Information Technology: Coding and Computing*, pp. 6–10, 2000.

[19] N. Morimoto, "Digital watermarking technology with practical applications," *In-*

forming Science, Special Issue on Multimedia Informing Technologies, vol. 2, no. 4, pp. 107–111, 1999.

[20] I. J. Cox and M. L. Miller, "Electronic watermarking: The first 50 years," in *Proceedings of the IEEE 2001 International Workshop on MultiMedia Signal Processing*, pp. 126–132, 2001.

[21] C.-H. Lee and Y.-K. Lee, "An adaptive digital image watermarking technique for copyright protection," in *IEEE Transactions on Consumer Electronics*, vol. 45, pp. 1005–1015, November 1999.

[22] J. Tian, "Wavelet-based reversible watermarking for authentication," in *Proceedings of the International Society for Optical Engineering (SPIE), Security and Watermarking of Multimedia Contents IV*, vol. 4675, pp. 679–690, April, 2002.

[23] B. Chen and G. Wornell, "Dither modulation: A new approach to digital watermarking and information embedding," *Proceedings of the International Society for Optical Engineering (SPIE): Security and Watermarking of Multimedia Contents,*, vol. 3657, pp. 342–353, 1999.

[24] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio and video," in *Proceedings of the IEEE International Conference on Image Processing ICIP-96*, pp. 243–246, 1996.

[25] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," in *IEEE Transactions on Image Processing*, vol. 6, pp. 1673–1687, 1996.

[26] D. Kirovski and H. S. Malvar, "Spread-spectrum watermarking of audio signals," in *IEEE Transactions on Siginal Processing*, vol. 51, pp. 1020–1033, April, 2003.

[27] L. Boney, A. H. Tewfik, and K. N. Hamdy, "Digital watermarks for audio signals," *IEEE International Conference on Multimedia Computing and Systems, Hiroshima, Japan*, pp. 473–480, June 1996.

[28] D. Gruhl, A. Lu, and W. Bender, "Echo hiding," in *Information Hiding 96*, pp. 295–315, 1996.

[29] H. Oh, J. Seok, J. Hong, and D. Youn, "New echo embedding technique for robust and imperceptible audio watermarking," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1341–1344, 2001.

[30] I.-K. Yeo and H. Kim, "Modofied patchwork algorithm: A novel audio watermarking scheme," in *IEEE Transactions on Speech and Audio Processing*, vol. 11, 2003.

[31] "Methods for subjective determination of transmission quality," *ITU-T Recommendation P.800*, 1996.

[32] "The e-model, a computational model for use in transmission planning," *ITU-T G.107*, 2000.

[33] J. Beerends and J. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," in *Journal of the Audio Engineering Society*, vol. 42, 1994.

[34] "Objective quality measurement of telephone-band (300-3400hz) speech codecs," *ITU-T Recommendation P.861*, August, 1996.

[35] "Perceptual evaluation of speech quality (intrusive)," *ITU-T Recommendation P.862*, 2001.

[36] "Method for objective measurements of perceived audio quality," *ITU-R recommendation BS.1387*, 1999.

[37] F. C.-P. Samir Mohamed and H. Afifi, "Audio quality assessment in packet networks: an 'inter-subjective' neural network model," in *The Proceedings of the 15th International Conference on Information Networking (ICOIN'01)*, pp. 579–586, 2001.

[38] L. Ding and R. Goubran, "Speech quality prediction in voip using the extended e-model," in *Proceedings of the IEEE Globecom 2003 Conference, San Francisco*, vol. 7, pp. 3974–3978, 2003.

[39] "Performance of the integrated kpn/bt objective speech quality assessment model.," *ITU-T Study Group 12*, May 2000.

[40] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq) - a new method for speech quality assessment of telephone networks and codecs.," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 749–752, May 2001.

[41] R. Tu and J. Zhao, "A novel semi-fragile audio watermarking scheme," in *IEEE International Workshop on Haptic, Audio and Visual Environments and their Applications*, pp. 89–94, 2003.

[42] L. Cai and J. Zhao, "Paudio quality measurement by using digital watermarking," in *Proceedings of the IEEE Canadian Conference on Electrical and Computer En-*

gineering (CCECE 2004), Niagara Falls, Ontario, Canada, pp. 1159–1162, May 2004.

[43] R. Tu and J. Zhao, "A semi-fragile audio watermarking scheme based on wavelet transform and quantization," *The Proceedings of CSEE*, vol. 25, no. 12, pp. 78–85, 2005.

[44] C. Hoene, B. Rathke, and A. Wolisz, "On the importance of a voip packet," in *Proceedings of International Speech Communication Association (ISCA) Tutorial and Research Workshop on th Auditory Quality of Systems*, Herne, Germany, April 2003.

[45] "Pulse code modulation (pcm) of voice frequencies," *ITU-T Recommendation G.711*, November 1988.

[46] "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (cs-acelp)," *ITU-T Recommendation G.729*, March 1996.

[47] D. Zheng and J. Zhao, "Image quality measurement by using digital watermarking," in *IEEE International Workshop on Haptic, Audio and Visual Environments and their Applications*, pp. 65–70, 2003.

[48] "Subjective performance assesment of telephone-based and wideband digital codecs," *ITU-T Recommendation P.830*, 1996.