Packet Loss Concealment for Speech Transmissions in Real-Time Wireless Applications

B.XU

Technische Universiteit Delft

TU Delft Delft University of Technology

Packet Loss Concealment for Speech Transmissions in Real-Time Wireless Applications

By

Boliang Xu

in partial fulfilment of the requirements for the degree of

Master of Science

in Electrical Engineering

at the Delft University of Technology,

to be defended publicly on Friday July 21, 2017 at 15:00

Student number: 4476026

Supervisors: Dr.ir. R. Heusdens, TU Delft
Dr.ir. R. C. Hendriks TU Delft

Thesis committee:

Dr.ir. R. Heusdens,	TU Delft
Dr.ir. R. C. Hendriks,	TU Delft
Dr.ir. J.C.A. van der Lubbe,	TU Delft
Ir. H. van der Schaar,	Bosch Security System

Preface

This thesis is the final work I have carried out as a master student at the Circuits and Systems Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology and an intern at Bosch Security System, Netherlands. I started this thesis from November, 2016 and finished it in July, 2017. All the information about my master thesis are presented in the following report.

Boliang Xu
July, 2017

ACKNOWLEDGEMENT

I would first like to thank my thesis supervisors Dr.ir. R.Heusdens and Dr.ir. R.C.Hendriks of the Circuits and Systems Group, Faculty of Electrical Engineering, Mathematics and Computer Science at Delft University of Technology. We have regular meetings every Tuesday, and they gave me so many valuable remarks and suggestions about my research or writing. They consistently allowed this paper to be my own work, but steered me in the right direction whenever they thought I needed that.

I would also like to thank experts H.van der Schaar and Patrick Timmermans in the Bosch Security System. They always helped and supervised me whenever I ran into a trouble spot or had a question about my works. Without their passionate helps, my thesis could not have been smoothly finished.

My classmates as well as colleagues also supported me during this period and we shared a lot of happy moments together.

Finally, I must express my very profound gratitude to my parents and to my girlfriend for providing me with unfailing support and continuous encouragement throughout these years of study and through the process of finishing this thesis.

Boliang Xu
July, 2017

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

Over the past three decades, two fundamental communication technical evolutions laid foundations for my thesis topic. The first is that the analog speech signal can be represented in digital form [22], which is able to be processed, stored and transmitted easily. The second is the development of packet-switched networks, including the interconnection we know currently as the Internet [29]. Figure 1.1 demonstrates a typical packet-switched network. These two techniques firstly merged in 1992 to achieve reliable voice transmission over the Internet [5]. Recently, we have witnessed the development of an Internet conferencing architecture [14], such as protocol for interactive application [39], multicast communication [40]. However, when designing applications such as interactive conference system, we have to face its trade-off: statistical multiplexing versus transmission latency and packet loss which significantly degrade the Quality-of-Service (QoS) issues. In this chapter, the basic principle of packet-switching network, reasons of packet loss during the transmission, some appropriate approaches used to conceal the packet loss, and objectives and an outline of the whole thesis, are presented. The objectives of my thesis provides comprehensive understandings on packet loss and packet loss concealment.

1.1 Motivation and Scope

Voice over Internet Protocol (VOIP) has become one of rapid-growing technologies which commonly uses the real-time transport protocol (RTP) to deliver voice packets over the

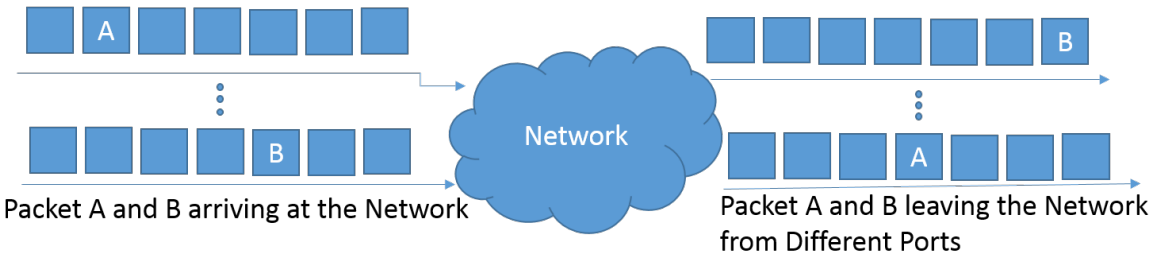


FIGURE 1.1. A packet-switched network receiving packets to route out

packet-switched network. RTP typically run on top of the user datagram protocol (UDP), which is not an absolute reliable transmission protocol so that packet loss may occur during the transmission. However, real-time applications require timely delivery of data or information and always tolerate packet loss to achieve this aim.

Packet communication which can handle speech in telecommunication system, such as conferencing system, has become more and more popular. In a typical wireless telecommunication system, analog speech signals are collected by microphones, then they are processed and encoded before arriving at the packets buffer. After transmitting to the receiver, signals are processed again and decoded to analog signals to be played-out by speakers. A diagram of typical wireless telecommunication applications is shown in Figure 1.2. The packet communication network's merit is the flexible applicability in multicast transmission with low cost [4].

1.2 Research Challenge and Approach

One inevitable problem with packet communication systems is that the messages or speech segments may be lost or do not reach the receiver during wired or wireless transmission. Figure 1.3 is used to describe the packet loss phenomenon in the transmission, the speech signal are divided into packets by black dashed lines, some packets were lost and imposed gap in the sentence during the transmission. The packet loss may occur in wireless networks for various reasons: First, congestion of routers and gateways. Second, the load in workstation is too heavy. Third, channel noise and interferences during the transmission. Referring several tests conducted to evaluate the effects of packet loss on

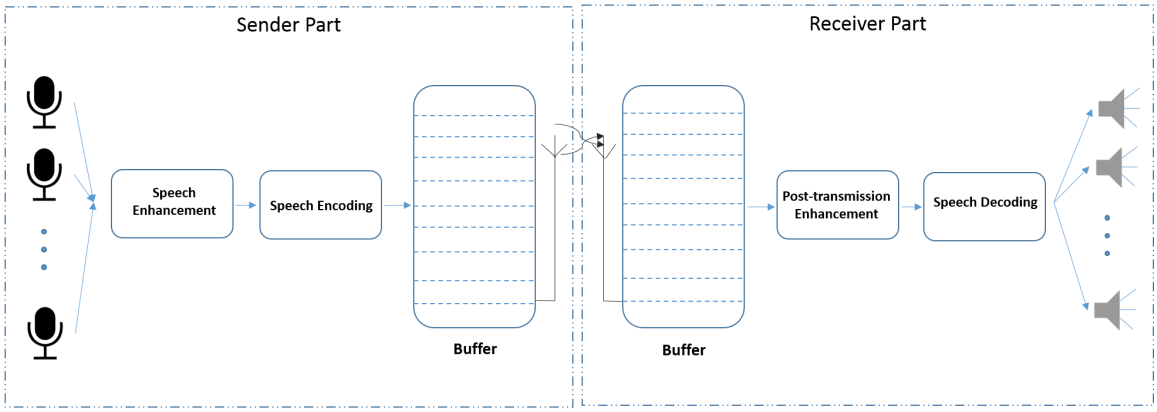


FIGURE 1.2. Diagram of a typical wireless telecommunication application

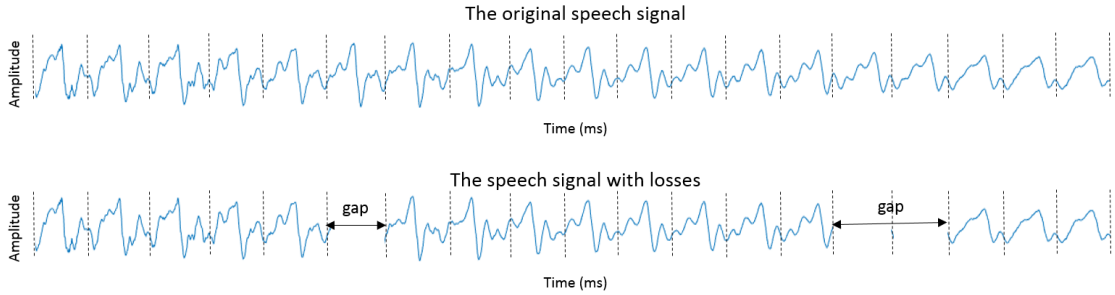


FIGURE 1.3. Illustration of packet loss in a speech transmission over an IP network

speech quality [12][13], silence replaced missing parts, and it was determined the packet loss rates up to 1 percent were acceptable. However, in the wireless transmission process, the loss rate can be really high (up to 10 percent or even more). In that case the quality of the transmitted speech may be severely impaired.

Although the packet loss exists ubiquitously, several packet loss recovery techniques are able to repair the loss in packet switched networks. Such techniques can be roughly divided into two categories: sender-driven repair techniques and receiver-based repair techniques.

Most sender-based PLC techniques work by re-transmitting lost packets or adding redundancies and additional information to packets in order to recover the loss from these extra parts. While sender-based approaches always consume undesirable resources such as network bandwidth and CPU capacity, etc. the consumption of bandwidth causes more load on the network and may potentially impose the loss of more packets. Apart from the resources consumption, sender-driven approaches typically introduce more delay into the transmission. A large delay is unacceptable for most real-time interactive applications.

As a result, we focus on the receiver-based approaches to conceal the loss. Depending on earlier researches, there are a number of approaches used to deal with the loss, such as silence and noise substitution [39], simply repeating previous packets or, if the codec information is known, synthesizing speech from received segments [44]. In recent years, interpolation-based schemes are getting more and more popular because of the simplicity and efficiency, such as pitch waveform replication [43] and waveform similarity overlap-and-add (WSOLA) technique [38].

In this thesis, the packet loss recovery technique should operate on conventional PCM packets without any codec information in order to fulfil the requirement of the system and could fully work in local devices without introducing any unacceptable delay as well as resources consumption. Of course, the high speech qualities after recovery must be guaranteed when facing to different packet loss conditions. For instance, the PLC algorithm is designed for Bosch Dicentis real-time wireless conference system, which uses simple A/D and D/A converter without any codec information. The latency budget in a complete transmission, including up-link and down-link in this system, should be as low as possible. Here, the challenges consist of three main components: 1) The reconstructed waveform should be similar to the original. 2) The discontinuity at frame boundaries should be smoothed after reconstructions. 3) The entire PLC algorithm should not add too many latencies to the system. Additionally, the algorithm should not be very complex to implement because of the practical use. Figure 1.4 gives a simple diagram to describe a typical wireless telecommunication application with receiver-based packet loss concealment unit and packet merging unit.

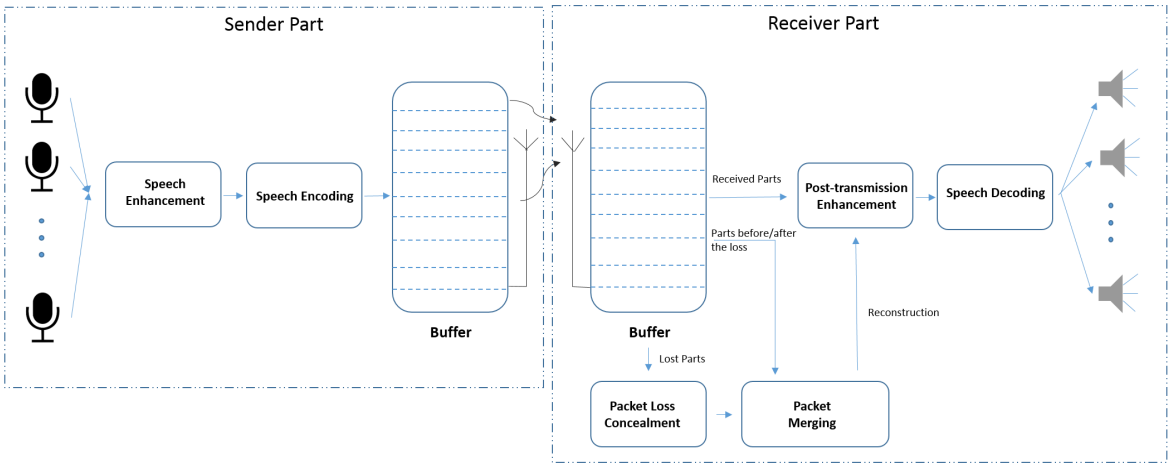


FIGURE 1.4. Diagram of a typical wireless telecommunication application with PLC

1.3 Thesis Objectives

As stated packet loss concealment as a research challenge in packet-switched network, this master thesis covers the following objectives: First, packet loss characteristics in realistic wireless network and existing PLC algorithms should be investigated in order to design a better PLC algorithm. Second, a PLC scheme with high quality, low latency and low resources consumption should be proposed. Third, the performance of proposed PLC should be evaluated and compared with performances of other techniques. Fourth, summaries and conclusions about this thesis should be given. Fifth, future works which may follow this thesis should be proposed.

1.4 Thesis Outline

This thesis is composed of five chapters and three appendixes.

The structure of my thesis is as follows. Chapter 2 introduces characteristics of packet loss in VOIP applications, along with a review to some existing receiver-based PLC algorithms. Chapter 3 presents a complete description of the proposed PLC scheme. Chapter 4 evaluates the performance of different PLC algorithms in different scenarios through a subjective listening test. After presenting results in each scenario, a brief discussion is

performed. Finally, chapter 5 gives conclusions about this thesis and proposes possible future works which can follow my present work.

At the end of this report, there are three appendixes, which contain detailed explanations about some topics referred in this report. Appendix A describes a four-state Markov model, which was used to generate packet loss patterns used to corrupt the audio file. Appendix B describes the details in State-of-Art PLC (ANSI T1.521a-2000). Appendix C gives a review to existing objective measures for evaluating speech quality.

CHAPTER 2

BACKGROUND

2.1 Packet-Loss Characteristics

Packet transmission channels usually suffer a combination of single packet losses and burst losses (several consecutive packet losses). Clear understandings of characteristics of packet loss are necessary to design a corresponding packet loss concealment algorithm. Here, we roughly categorized them into two main classes: single loss and burst loss.

- **Single loss**: With single loss we lose packets randomly. We experienced single packet loss when losses are independent from one instant to the next.
- **Burst loss**: With burst loss we suppose losses will last for a time period so that we lose one or more consecutive packets. Although the occurrences of burst loss are relatively rare with the occurrences of single loss, Loss bursts of several consecutive packets is one of main reasons of speech quality degradation caused by packet loss. As the network condition becomes bad, the probability of burst loss may increase.

Looking into the realistic wireless network, we usually experience a combination of the single packet loss and the burst packet loss. When we use standard Wi-Fi-based applications to transmit the data packets, the data packets will enter a certain channel of Wi-Fi channels. Then, the chance to transmit these packets successfully is depending on the real-time condition, such as the load, of this certain channel. The Wi-Fi environment changes from time to time, we never know the state of certain channel is "good" or "bad" at certain instant. As a result, we randomly suffer the single loss and burst loss in one

transmission.

We consider using a logic analyser to measure the packet loss situation in real-network environment. In our measurement, Agilent logic analyser is used to capture the change of voltage at the end of up-link (up-link is the wireless transmission link from devices to the access point). Here, only transitions between high voltage (packets are received) and low voltage (packets are lost) will be recorded and put into the buffer, however, the sampling keeps going on with the rate defined by users (used to get the duration between two adjacent voltage changes). The sampling will stop until users stop the measure manually or the buffer is full. The voltage level is taking on values in the set X={0, 1}, where 0 represents the loss and 1 represents the reception. The measurement is taken from 9am to 3pm (working hours) in Bosch Security System building. The devices are Dicentis wireless unit and Dicentis access point. All channels in Wi-Fi 2.4GHz are enabled.

In order to investigate the packet loss characteristics in realistic networks and represent the data clearly, we plotted the probability mass function (PMF) of packet loss (Bad States) with different run-length and run-length of received packets (Good States) between losses in log scale. As shown in the Figure 2.1, both run-length of Good States and Bad States followed somewhat experiential decays. As for packet loss, the major loss was the single loss, which approached 85% in this particular measurement. Additionally, burst loss, which was up to hundreds milliseconds, happened with really low probability. As for received packets between losses, probabilities of short run-length of received packets were higher.

When we monitored the Wi-Fi environment by Wi-spy (a tool to monitor the Wi-Fi environment) in real-time, we found that the channel used to transmit our data packets was congested or even fully occupied by other devices in a short time period which is called bursty period, so that the data packets could not get through the Wi-Fi link and reach the destination successfully. In bursty period, the packets were lost continuously (burst loss) or sort of alternatively (loss at a high rate in a short time period) until conditions in this channel became good again to transmit packets. However, the Wi-Fi environment was fine or normal to transmit the data packets most of the time.

In conclusion, we can derive two main characteristics: First, occurrences of single loss are way more than occurrences of burst loss. Most losses consist of singe packet loss,

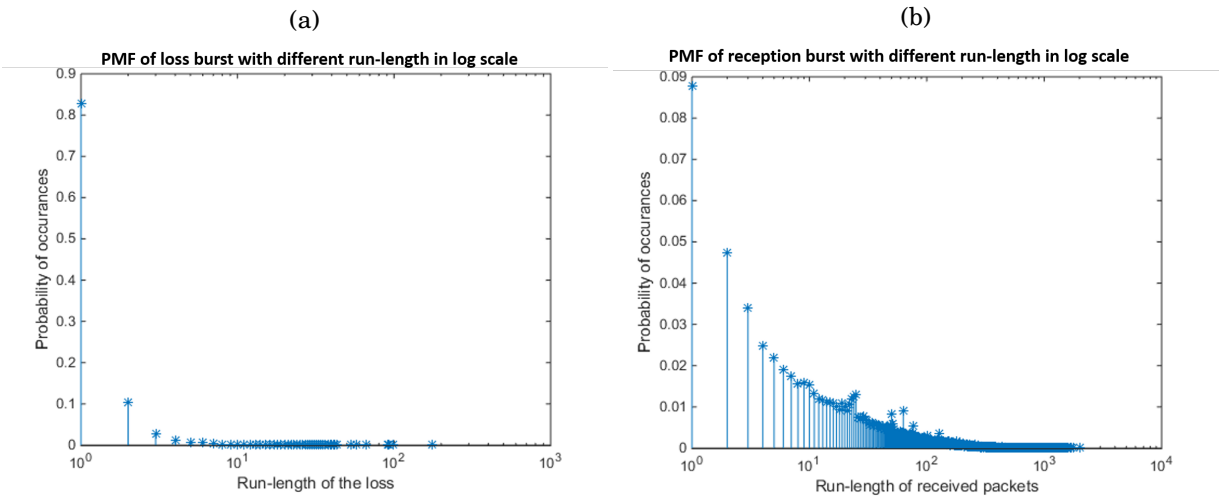


FIGURE 2.1. (a) PMF of loss burst with different run-length in log scale (b) PMF of received packets between losses with different run-length in log scale

even in bursty periods, run-lengths of individual loss are short and fall close together. Second, burst losses, including short and long bursts of loss, cannot be ignored in real-network and have to be considered in designing the packet loss model and the packet loss concealment algorithm.

In this thesis, a four-state Markov model is used to model the packet loss in wireless networks and generate the loss pattern which are used to corrupt the audio file. Details are presented in appendix A.

2.2 Existing Packet Loss Concealment Algorithms Review

The main constraint introduced by the interactive applications is the latency, which is supposed to be minimized as small as possible. A part of sender-based packet loss recovery techniques, including retransmission and interleaving, cannot be used any more, since these techniques impose too much latency to make the transmitter be ready for the transmission. Due to the fast and efficient transmission is required, IP-based interactive applications always use User Datagram Protocol (UDP) as their transmission protocol instead of Transmission Control Protocol (TCP). In UDP, the samples are transmitted in packets, which will be completely lost when the packet loss occurs, so it's useless to

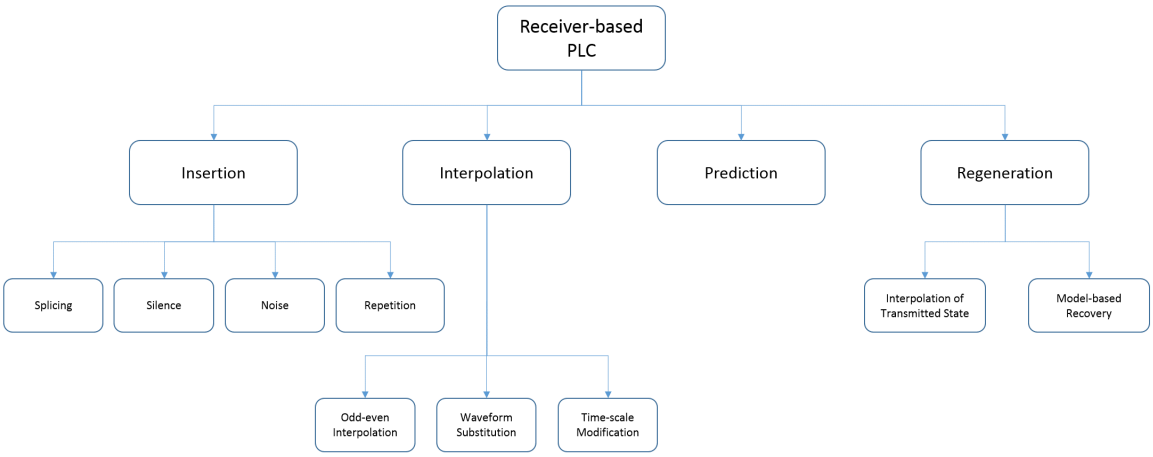


FIGURE 2.2. Receiver-based repair techniques taxonomy

piggyback any extra information to itself for repairing the loss. Meanwhile, piggybacking information to any other packet will introduce undesirable latency in the end since the reconstruction process has to start after receiving the packet with that piggybacked information in stead of reconstructing the loss immediately. As a result, the technique like Forward Error Correction (FEC) become failing. Additionally, another main reason for avoiding sender-based loss recovery techniques is that some existing approaches require additional network bandwidth. Since the data bandwidth increases, the network becomes more congested and more packets may be dropped. Above constraints force us to focus on the receiver-based PLC techniques.

The Figure 2.2 gives a taxonomy of receiver-based packet loss concealment (PLC). Receiver-based packet loss concealment techniques do not require any assistance from the sender part. Such techniques are useful when the sender cannot recovery all the loss or when the sender is unable to participate in the recovery process.

The basic goal of PLC is to create replacements for lost packets which are similar to the original. In the conference system, these techniques are feasible since speech signals are comprised of lots of short-term self-similarities.

As studied in earlier researches, the most of PLC techniques can be split into four categories

- Insertion-based schemes recover losses by inserting fill-in packets. The noise or silence are commonly used as fill-in packets. Such techniques are easy to implement at expense of the speech quality after repairs.
- Interpolation-based schemes recover losses by finding matching patterns or interpolation to derive replacement packets which are similar to the original. These techniques' computations are complicated but performances are better than insertion-based schemes.
- Prediction-based schemes recover losses by predicting them from previous packets according to training data or entropy theory in the language [28].
- Regeneration-based schemes take advantage of the knowledge of the audio compression algorithm to collect codec parameters and recover the loss by using those parameters of packets surrounding the loss to produce replacement packets. Such recoveries are expensive to implement but give satisfying results.

Although Prediction-based schemes and regeneration-based schemes perform better than insertion-based and interpolation-based schemes in packet loss concealment, both of them are too complicated and expensive to implement. Insertion-based schemes are easy to implement but with the expense of the speech quality after repairs. After balancing the performance and feasibility, I mainly focus on the interpolation-based techniques, which consist of following three categories.

- ***Odd-even interpolation***: Using Odd-even interpolation to recover the packet loss is first presented by Nuggehally [21].The speech samples are divided into adjacent odd-sample and even-sample packets, respectively. Samples in the lost packet can be interpolated by odd-samples or even-samples in this packet. Odd-even interpolation works pretty good for recovering single loss, however, both the odd-sample and even-sample packets in a pair can be lost. When this happens, odd-even interpolation will fail to recovery the loss.
- ***Waveform Substitution***: Waveform substitution techniques use the speech signal before, or after, the loss to find an alternative to recover the lost segment. Waveform substitution is effective when the speech signal doesn't change a lot during a missing packet [11]. When a reliable pitch detection is available, a new kind of

waveform substitution called Pitch waveform replication (PWR) technique is proposed [43]. If the pitch period P ms is available for a voiced speech, then the missing segment can be filled by repeatedly copying the last P ms of received speech before the loss, after the loss or both. For unvoiced speech segments, simple repetition of previous packets is employed instead . Naofumi Aoki presented a modified two-side PWR technique [1], which takes the coherency of the pitch waveforms between the backward and forward frames into account. In that paper, the study assumes that the variation between backward and forward frames can be modeled as a linear function. In other words, the pitch period is changed with respect to the time when we repeat the waveform from backward or forward frames. This technique, to some degree, alleviates the phase mismatches inside recovered segments and produces better results than the others.

However, there are still some problems existing in PWR technique degrade the performance of recovery. In the following part, we mainly take two problems into account.

- Metal/Tinny sound
  The loss gap may consist of several consecutive packets instead of single packet. In another word, the pitch waveform replication may be applied several times in order to cover the gap. PWR is seemed as a kind of repetition but with pitch information, however, the substitution waveform is completely the same with the previous waveform. As a result, concealment generates highly periodic signal, which degrades the perceptual quality significantly. Users can distinguish that sounds are heard not that natural after the concealment and such sounds are called "metal" or "tinny". This problem is inevitable in any repetition-based technique.
- Wrong pitch period detection
  Pitch detection is an indispensable procedure of PWR technique. A reliable pitch detection guarantees the high recovery quality. However, the task of estimating pitch period is still very difficult now because a) the human vocal tract is very flexible and varies from people to people. b) The emotional state of speaker also influences the pitch period. c) Pronunciation(accent) can also change the pitch period. d) pitch period varies from 1.25ms to 40ms [25]. Therefore, no one algorithm so far perfectly derives the pitch period, especially

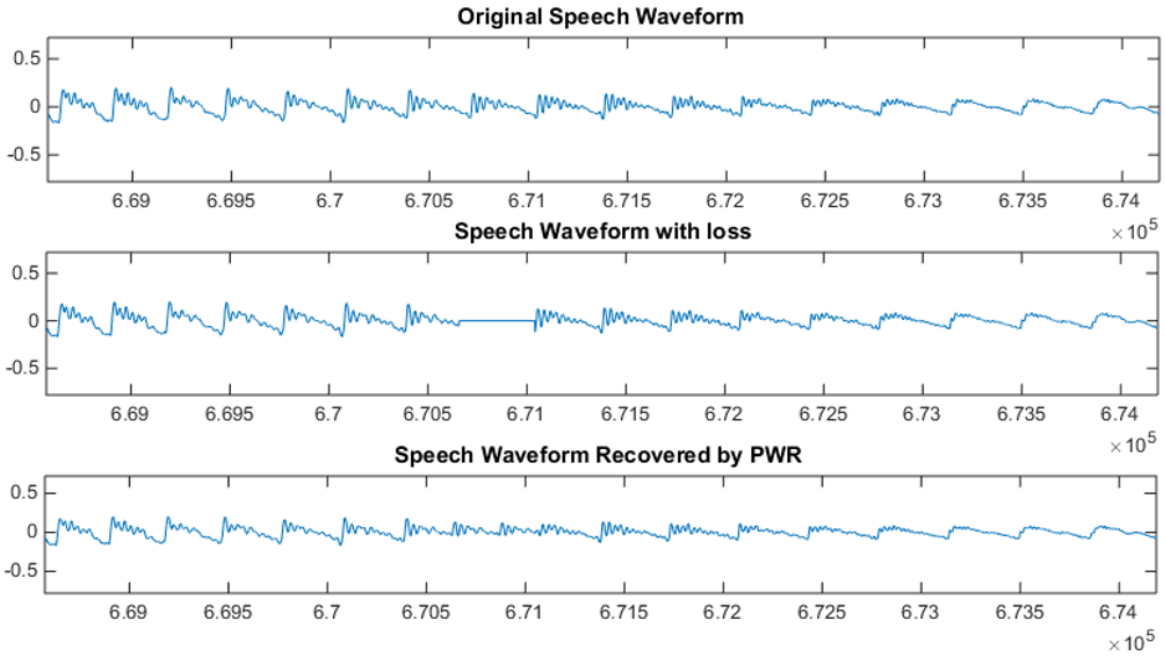


FIGURE 2.3. Wrong pitch detection in PWR

in a noisy environment. In another word, wrong pitch period detection is an inevitable problem. From the Figure 2.3, the pitch period is determined as a wrong number so that the substitution waveform doesn't look similar to the original.

- ***Time Scale modification(TSM)***: Time scale modification allows segments on either side of the loss to be "stretched" to fill the gap. One of the most advanced version of TSM is called waveform similarity overlap-and-add (WSOLA).

  Depending on the previous researches, WSOLA proposed by Werner Verhelst et al. [42] is a member of TSM techniques and seems like a ideal method for covering losses [38]. WSOLA finds overlapping waveform on either side of the loss and averages where they overlap to cover the loss. WSOLA can preserve the naturality of sound better than traditional time-scale modification method, because it doesn't simply stretch sound segments and change the pith period.

The operation of the WSOLA technique is illustrated in figure 2.4 and explained

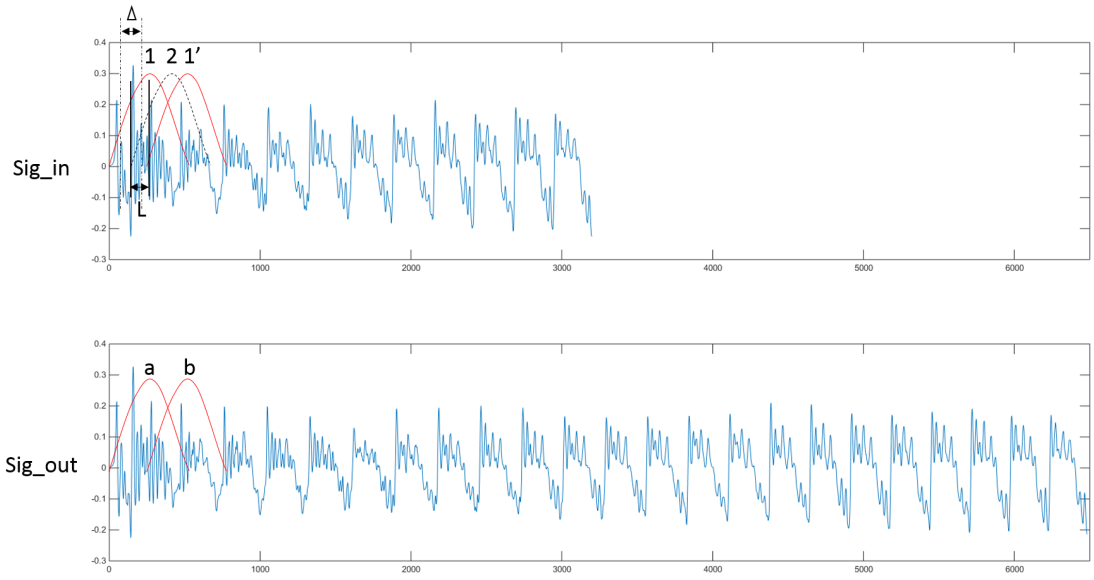


FIGURE 2.4. Illustration of a WSOLA algorithm for stretching signal

below.

Proceeding in a left-to right direction, suppose frame **1** is the last frame that is received from the input and directly added to the output (frame **a** = frame **1**). **1'** is the adjacent frame coming after frame **1** and overlap-add with frame **1** in a synchronized way. WSOLA then needs a frame **b** which can overlap-add with **a** in a synchronized way and is able to be extracted from the input. Since frame **a** = frame **1**, WSOLA just needs to go back **L** samples (depending on the ratio of stretching signal) and locate a frame that resembles **1'** as closely as possible within the tolerance interval Δ) in the input. The position of frame **2** is located by maximizing a similarity measure between the sample underlying frame **1'** and the input segment. Assume the most similar frame with **1'** is frame **2**, then add it to the output as frame **b** and overlap-add with **a** in a natural way. After overlap-adding **b** with **a** in the output, WSOLA continues to process the "neighbourhood" of **2** and stretch it. From the figure above, it's easy to see the length of the original is stretched by WSOLA. Suppose the loss happens from Time 3200 to Time 6500, then this gap is fully covered by WSOLA. Since frames are overlap-added in a synchronized way and the original pitch period is maintained, the sound after stretching will hear naturally. This is the basic starting point to use WSOLA as a

packet loss concealment technique.

In practice, the parameters in WSOLA have to be chosen carefully to ensure the high recovery quality. There are five parameters, including window size, overlap ratio, scaling factor, tolerance interval and length of signal to be stretched, existing in WSOLA techniques:

Window size: Window size is chosen to be 20ms. First, 50Hz is the lower bound on frequency of human speech and its corresponding time period in time domain is 20ms. In another word, 20ms speech segment must contain at least one entire pitch period, which ensure the pitch period will not be changed after the overlap-adding. Second, Speech signal may be stationary when it is viewed in blocks of 10-30 ms. Window size should be chosen from this interval so that the similar waveform can be extracted and overlap-added in a natural way.
Overlap ratio: Overlap ratio is chosen to be 50% to make the sum of adjacent window functions maintains 1. In another word, the amplitude of original waveform won't change after overlap-addings.
Scaling factor: Scaling factor is used to determine the ratio of stretching the original waveform. When the size of gap is known, it is used to divide by the length of signal to be stretched to get the scaling factor.
Tolerance interval: Tolerance interval is the Δ in figure 2.4, which gives a interval to let WSOLA search for the most similar waveform with the "neighbourhood" and use it in the overlap-add procedure later.
Length of signal to be stretched: In the buffer, it's called history buffer. History buffer is used to be stretched in order to cover the gap.

In practice, the most challenging part is to balance the scaling factor, tolerance interval and length of signal to be stretched so that the stretched signal can cover the gap but doesn't exceed a lot, since the excess part will destroy the coming signal. Because of the tolerance interval, user never knows the exact length of the stretched signal. Larger tolerance interval means more opportunities in finding the best match to overlap-add the signal with expense of more uncertainties. When tolerance interval is increased, the scaling factor or length of history buffer should also be increased in order to ensure the stretched signal covers the gap. In this scenario, the overlap-add quality is maintained but the uncertainty is enlarged. As

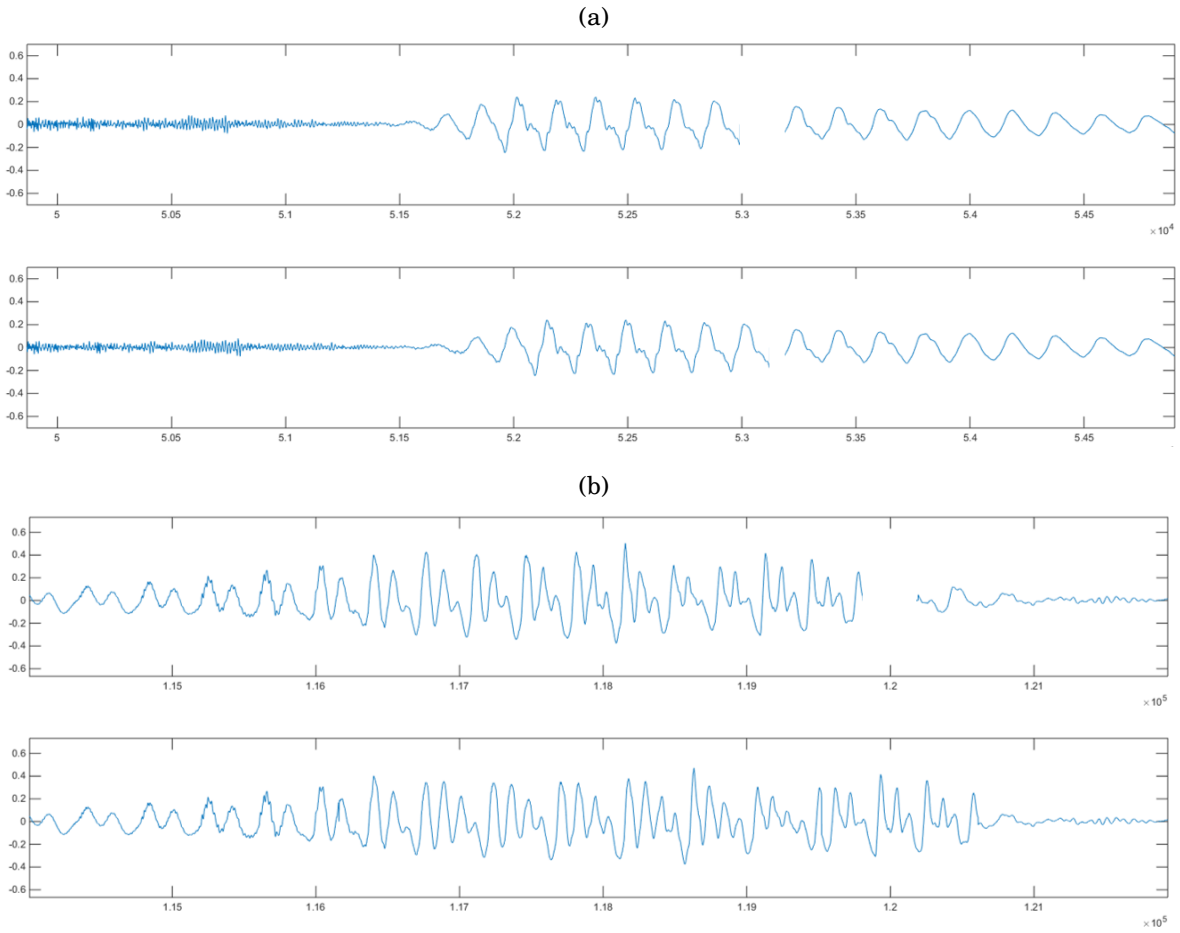


FIGURE 2.5. (a) Stretched signal is not long enough to cover the gap (b) Stretched signal is too long to destroy the coming signal

shown in the figure 2.5 (a) and (b), the stretched signal is too short and too long to fill-in the gap, respectively. Our goal is to cover the gap so only (b) is allowed in the implementation. Scaling factor is large and the coming part may suffer distortions. While if the tolerance is minimized, the uncertainty is also minimized, however, the overlap-add quality is degraded as shown in the figure 2.6. The stretched signal part distorts a lot and sounds weird after the recovery. Even the parameters are tuned into good situation and make one of the recoveries work perfectly, however, they may not be efficient in the others.

Apart from the uncertainty introduced by tolerance interval, latency is another significant drawback of WSOLA technique. Reasons of the length of the signal to be stretched should be long enough are twofold: First, the longer length means

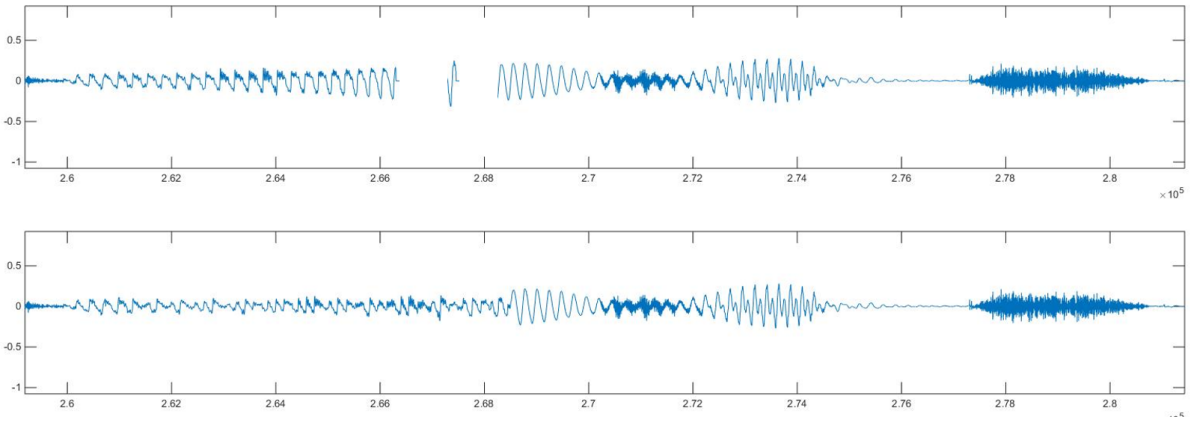


FIGURE 2.6. Small tolerance interval degrades overlap-add quality to introduce distortion

more times to stretch the signal in order to fully cover the gap. Second, the length should be long enough in case a long loss happens. The signal to be stretched can handle the repair without excess stretch, and excess stretch makes recovered signals sound quite different with the original. In the buffer, this part of the signal need to wait to be played-out after the recovery so that the whole speech segment sounds smoothly. Additionally, the merge part between stretched signal and coming signal imposes latency as well, and this part of latency is unknown because of the uncertainty. Also, the size of gap need to be known to get the scaling factor and this size can only be determined once the system successfully receives the packet. These latencies are unacceptable for some interactive applications which really need real-time communications.

In conclusion, WSOLA is not suitable when a number of losses happen and these losses are too short or too long compared with the signal to be stretched because of the uncertainties introduced by the tolerance interval. In addition, WSOLA technique adds too many latencies to play-out the received speech segments and makes real-time communication process belated.

CHAPTER 3

PROPOSED PACKET LOSS CONCEALMENT SCHEME

In this chapter, a packet loss concealment scheme is proposed to conceal the packet loss in the speech transmission under constraints given by real-time applications. Firstly, motivations to implement proposed PLC scheme are presented. Secondly, the structure of the proposed scheme is described. Thirdly, a complete and detailed description of PLC algorithm is performed. Eventually, the signal processing steps and latency consumptions of this scheme are analysed.

3.1 Motivations

The motivations of proposing a new adaptive PLC algorithm for real-time applications are twofold:

First, existing PLC algorithms, such as two-side pitch waveform replication, WSOLA, etc., they reconstruct packet loss based on the known size of gap. In order to derive the size of gap, above algorithms have to take actions until a packet is successfully received after the loss. Such PLC schemes impose large latencies, which are unacceptable for real-time applications.

Second, as shown in the investigations of packet loss characteristics and existing PLC algorithms in the chapter 2, most losses consist of single packet. These single losses can be easily repaired by odd-even interpolation because the original waveform could be

recovered by simply interpolating even or odd samples of itself. Odd-even interpolation fails when both odd and even samples of one packet are lost. However, waveform-based interpolations can be applied to erase effects caused by such short burst loss. As described in the packet loss characteristics section, transmission may also suffer long loss runs. If the burst loss is beyond certain length, the reconstructed waveform will diverge from the original and produce artificial effects which degrade the speech quality significantly. Then, an attenuation function should be employed to alleviate these artificial effects. Another packet loss characteristic mentioned in chapter 2 is that short loss runs get close together in bursty periods. Waveform-based interpolations reconstruct the packet loss based on the waveform information in previous received parts. Assume there are many losses including single loss and burst loss in one period, early losses will be reconstructed by waveform-based interpolations and used to recover later losses. Waveform-based interpolations introduce distortions (difference between the substitution and the original) in the reconstruction and, to some degree, destroy the waveform information used to recover the next loss so that the distortions as well as artificial effects are accumulated. With increasing of the number of distortions, the artificial effects will become distinguishable and impair the speech quality. However, odd-even interpolation can repair a part of losses in bursty periods to make the recovered waveform resemble the original as closely as possible.

As a result, we proposed a scheme named continuous update, which introduced extremely small latency to the system. By using this scheme, an adaptive PLC algorithm combining odd-even interpolation, waveform similarity matching and silence substitution, was proposed to deal with the packet loss in real-time applications.

In section 3.2, the use of continuous update scheme is described. Section 3.3 shows all details in the adaptive packet loss concealment algorithm. Section 3.4 analyses the complexity, latency and resources consumptions introduced by this algorithm.

3.2 Continuous Update

In common streaming speech transmission, the speech samples will be first stored in a buffer and then played-out until the buffer is full. Due to the packet loss in the transmission, some loss gaps occur in the buffer. PLC unit (see Figure 1.4) then needs to repair the loss and forward the recovered waveform to the next step. However, for real-time

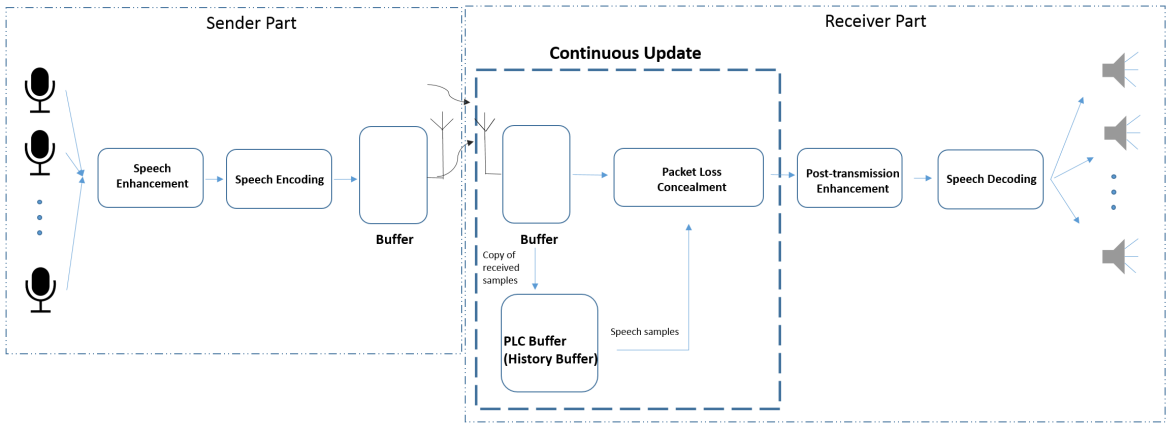


FIGURE 3.1. Illustration of the continuous update scheme

applications, above procedures impose the unacceptable latency.

In continuous update, as shown in the figure 3.1, buffer is used in different way from streaming speech transmissions. Samples in the buffer will be overwritten when new speech samples come in, but PLC buffer (History Buffer) will keep a copy of received samples so that the packet loss concealment unit can use them to reconstruct lost samples. Received speech samples will forward to the next step instead of staying in the buffer. When a missing part is indicated, it will be recovered immediately by using samples in the PLC buffer. Once the missing part is reconstructed, it will be updated to the PLC buffer in order to recover the coming potential loss. If the next part is lost as well, the same reconstruction process occurs until the proper packet arrives.

Figure 3.2 is used to describe the way of reconstructing waveform in continuous update scheme. When part 6 was indicated as a missing part, PLC algorithm could reconstruct it immediately and update a copy of it to the PLC buffer. Part 7 was lost as well, then samples in updated buffer were able to be used to reconstruct part 7. Whole recovery stopped once part 8 was successfully received. Since we did no changes to received samples and did not have to know the size of the missing part, continuous update procedure introduced negligible latency to the whole system. The specific latency consumption in this scheme is presented in the section 3.4.

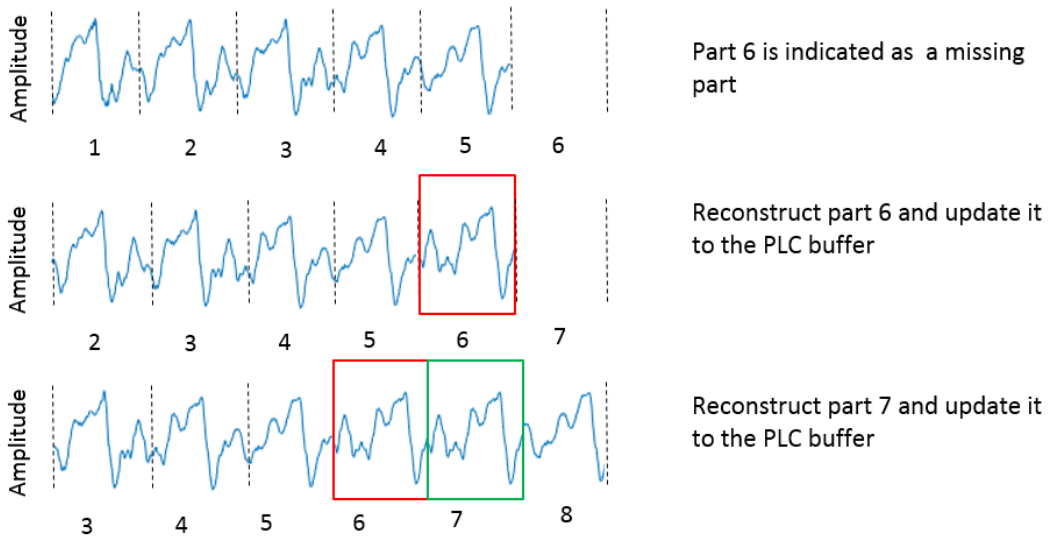


FIGURE 3.2. The way of reconstructing waveform in continuous update scheme

3.3 Adaptive Packet Loss Concealment

The reason of giving adaptive packet loss concealment as a name to the proposed algorithm is that the proposed PLC algorithm will make the decision to use which algorithm for concealing the packet loss based on the number of consecutively lost packets. Figure 3.3 gives an illustration of the adaptive packet loss concealment algorithm, which is integrated into the block named packet loss concealment in the Figure 3.1. As shown in the Figure 3.3, a pair of odd-sample and even-sample packets are hopefully received at the receiver side from the sender side. If both of them are received, PLC will simply assemble them back to the normal order and forward them to the next step. If one of the packets is received, PLC then upsamples the received one with an integer factor 2. If both of them are lost, PLC then chooses to use WSM or repetition instead. If the previous segment is detected as unvoiced/silence, simple repetition will be employed, if not, then WSM will be used to find ideal substitutions. Reconstructed waveform will be merged with previous speech samples in order to erase discontinuities at boundaries. With increasing of the number of reconstructions, the reconstructed waveform will attenuate to zero once the gap exceeds a certain length. When a incoming packet arrives, it will be merged with the last recovery immediately and forwarded to the next step. Details in

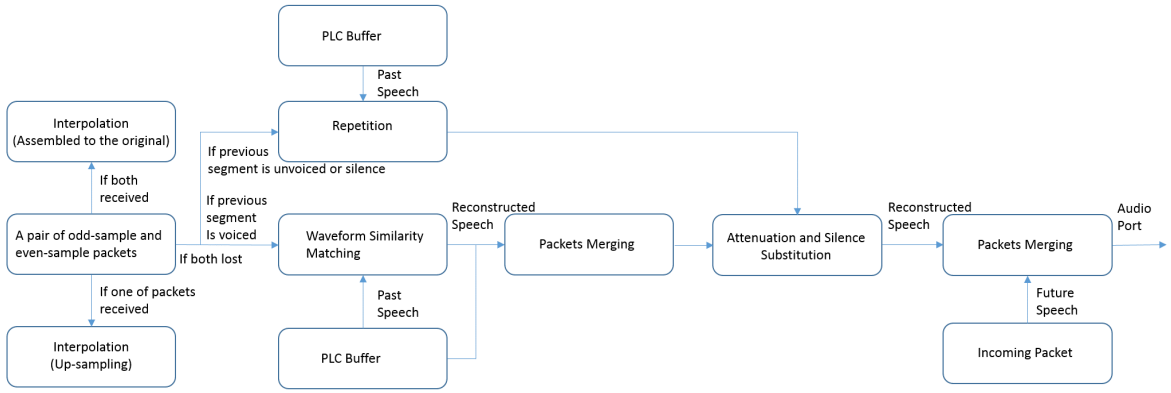


FIGURE 3.3. Illustration of the adaptive packet loss concealment algorithm

the adaptive PLC are as follows.

3.3.1 Odd-even Interpolation

The speech samples are partitioned into adjacent odd-sample and even-sample packets, respectively. When transmitting both odd-sample and even-sample packets, there are four scenarios at the receiver part as shown in the Figure 3.4. In scenario 1, both packets are successfully received, samples just need to be put back into the original order to recover the primary speech segment. In scenario 2 and 3, only odd-sample or even-sample packet is received. Then, interpolation with low-pass filter is applied to interpolate the packet to the original length. An expander is shown in the Figure 3.5 to explain steps in the interpolation, the sequence separated by zeros enter a low-pass filter (LPF) with gain factor L=2 in our case. Then, the sequence is interpolated with the gain factor L=2. The waveforms constructed by upsampling the odd-sample or even-sample packet are shown in the Figure 3.6, the red line is the original waveform, while black one and blue one are reconstructed by interpolating the even samples and odd samples, respectively. However, if both packets are lost as shown in the scenario 4, then a loss with two-packet length occurs and has to be reconstructed by other techniques.

The performance of odd-even interpolation will degrade when the signal frequency is higher than the (original sampling rate)/4. If we want to up-sample the sequence (odd-sample or even-sample packet whose sampling frequency is (original sampling rate)/2) by a factor 2, then the gain in the low-pass filter (LPF) is 2 and corresponding cut-off

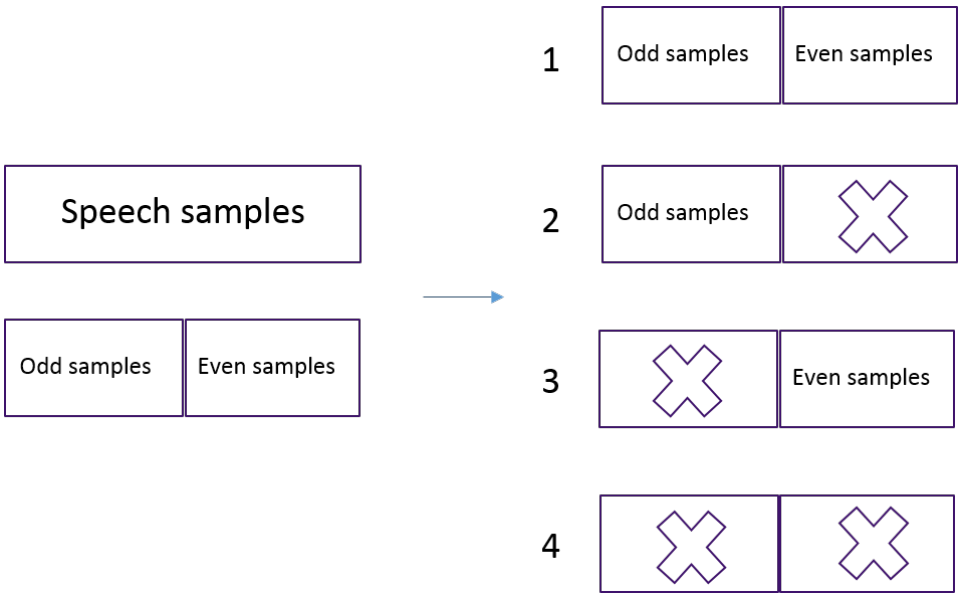


FIGURE 3.4. Scenarios in receiving odd-sample and even-sample packets at the receiver

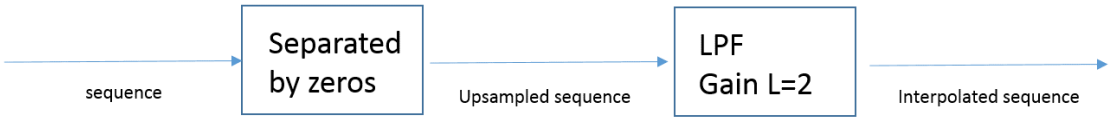


FIGURE 3.5. An expander to explain steps in the interpolation

frequency is $\pi/2$. In another word, if the original sampling rate is 48kHz, the frequency which is higher than 12kHz will be removed by the LPF. This issue introduces errors to the implementation of odd-even interpolation when signals contain very high-frequency parts. However, in telephony, the usable voice frequency ranges from 300 Hz to 3400 Hz, so the above problem will not happen to speech signals.

3.3.2 Waveform Similarity Matching

In this section, first, the motivation to choose waveform similarity matching in stead of other techniques in this adaptive PLC algorithm is presented. Second, details in WSM are described.

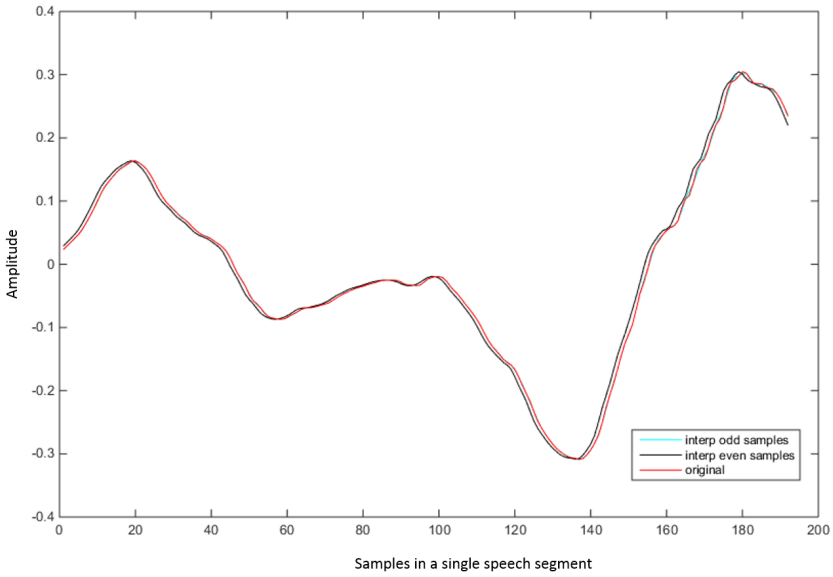


FIGURE 3.6. Waveform constructed by odd-even interpolation

Suppose both odd-sample and even-sample packets are lost in a transmission, odd-even interpolation then has no information about the current speech segment. With this happens, WSM including segment repetition will stuff the missing part based on the speech information in the history buffer (PLC buffer). Considering the continuous update scheme, simple repetition (simply repeat the last received segment to fill-in the gap) can be implemented, modified one-side pitch waveform replication (back to one pitch period just before the loss, and take a part of/couple of pitch pattern/patterns with two-packet length to fill-in the gap) can be applied as well.

As for simple repetition, the last received packet cannot be used to predict the coming packet because it can't make sure that this packet can connect itself in a synchronized way, especially for the voiced speech segment. As shown in the Figure 3.7, repeating the speech segment in the part A after itself will produce a "jump" called DC offset in the red dashed block. This DC offset produces distinguishable artificial effects, even it is smoothed by some merging techniques. However, repetition can be applied to unvoiced speech segment without producing perceivable artificial effects [11], since unvoiced speech is noise-like without any periodic nature. By looking the waveform, the amplitude level of unvoiced speech segment is so low that the offset introduced by repetition can, to

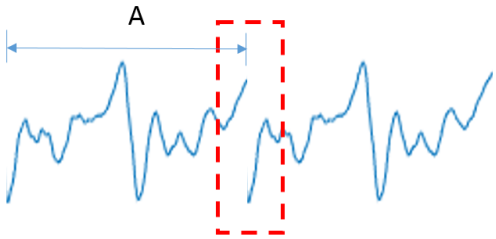


FIGURE 3.7. DC offset introduced by repetition

some extent, be ignored.

As for pitch waveform replication (PWR), some inevitable problems listed in chapter 2 also occur in continuous update scheme: First, the performance of PWR relies on the performance of the pitch detection procedure, which gives accurate results for highly periodic voiced speech signal. However, it may derive wrong pitch period with lowly periodic signal or noisy signal. Also, the pitch detection algorithm adds complexity to the PWR. Second, the simple replication of the same pitch period cause artificial effects (i.e. tinny/metal sounds), since highly periodic signal is generated and the same tone is sustained.

Considering all issues mentioned above, We proposed a PLC algorithm named waveform similarity matching, which performs in continuous update scheme and doesn't only rely on last pitch period to fill-in the gap. The basic assumption of waveform similarity matching is that the signal is quasi-periodic in the time domain and its spectral characteristics are relatively invariant for a short duration. In another word, we can pick the one which can connect the previous packet in the most natural way instead of only using the segment in last pitch period. As shown in Figure 3.8, waveform similarity matching uses the last part before the loss as the template to find the most similar part with itself (template') in the history buffer, and copy the following part (candidate) with two packets length in our case to substitute the loss.

WSM algorithm is composed of several components , such as voice activity detection,

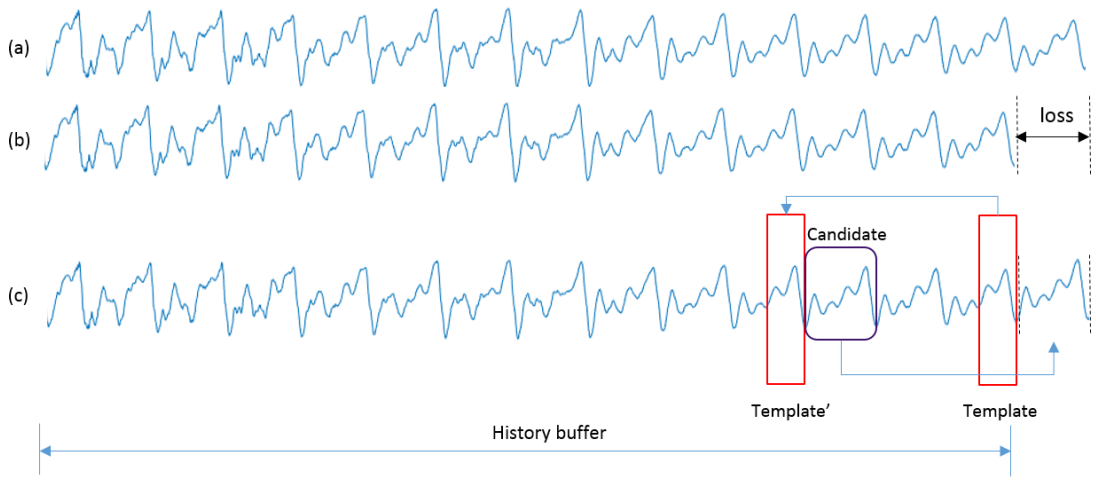


FIGURE 3.8. Illustration of Waveform Similarity Matching (WSM): (a) the original signal; (b) the corrupted signal; (c) the signal recovered by WSM

similarity matching algorithm, parameters selection and packet merging technique.

- **Voice activity detection** The first step in WSM is voice activity detection. If the lost segment is estimated to be the voiced speech, then a missing packet will be reconstructed by WSM, if the lost segment is unvoiced speech, a simple repetition of the last speech segment will be applied instead.

  Here, a combination of zero-crossing rate calculation and energy calculation is used to determine the voice activity.

  The zero-crossing rate (ZCR) is a measure of number of times in a given frame that the amplitude of speech signals crosses through a value of zero. Zero-crossing analysis is widely used in the signal processing field as a typical parameter. Voiced speech is produced because of excitation of vocal tract by the periodic flow of air at the glottis and usually shows a low zero-crossing count, whereas the unvoiced speech is produced by the constriction of vocal tract narrow enough to cause turbulent airflow which results in noise and shows a high zero-crossing count [2][20].

ZCR is defined formally as

$$ZCR = \sum_{l=1}^{L-1} \left| \frac{sgn[x(l)] - sgn[x(l-1)]}{2} \right| \tag{3.1}$$
$$where, \;\; sgn[x(l)] = \begin{cases} 1, x(l) \geq 0 \\ -1, x(l) < 0 \end{cases}$$

where **x** is a signal of length **L**.

However, zero-crossing of a waveform are sensitive to the formants, noise, and any DC level in the waveform [9]. Another parameter, the short time energy (STE), is also widely used for speech activity detection. Short time energy is a simple but effective classifying parameters for separating voiced and unvoiced speech segments based on the energy of speech signals in the frame [7]. Voiced speech signals always have higher energy in a frame than unvoiced speech signals do [30].

Energy of signal in a single frame is defined formally as

$$STE = \sum_{l=1}^{L} x(l)^2 \tag{3.2}$$

where **x** is a signal of length **L**.

As explained before, If one speech segment has low energy and large or zero zero-crossing rate, this segment is considered as the unvoiced/silence speech segment, otherwise the segment is voiced speech segment. The Block diagram of the voice activity detection scheme is shown in Fig.3.9.

The thresholds of energy and zero-crossings rate for making the decision should be set in advance. In our simulation, we analysed both of them in 4ms speech segment.

The threshold for energy was chosen to be 0.5 based on the energy of speech signals used in our simulations. The energy level of one speech signal is plotted in the Figure 3.10. From the figure, the energy of most of unvoiced/silence segments are lower than 0.5, while most of voiced segments have higher energy level.

For zero-crossing, the threshold was chosen to be 10. With a 20ms unvoiced speech signal, the zero-crossing rate is supposed to be higher than 50 [3], so in our case,

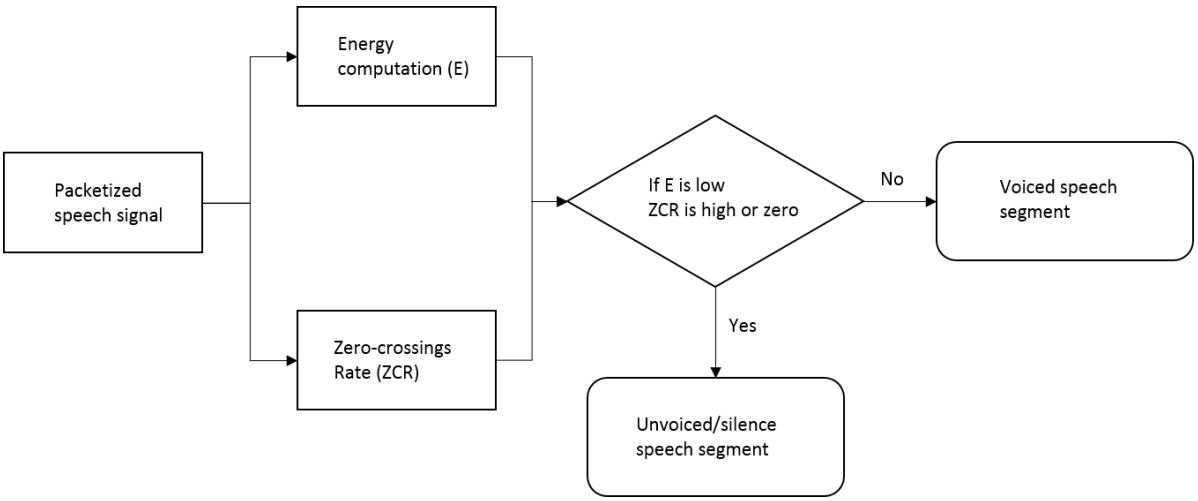


FIGURE 3.9. Voice activity detection

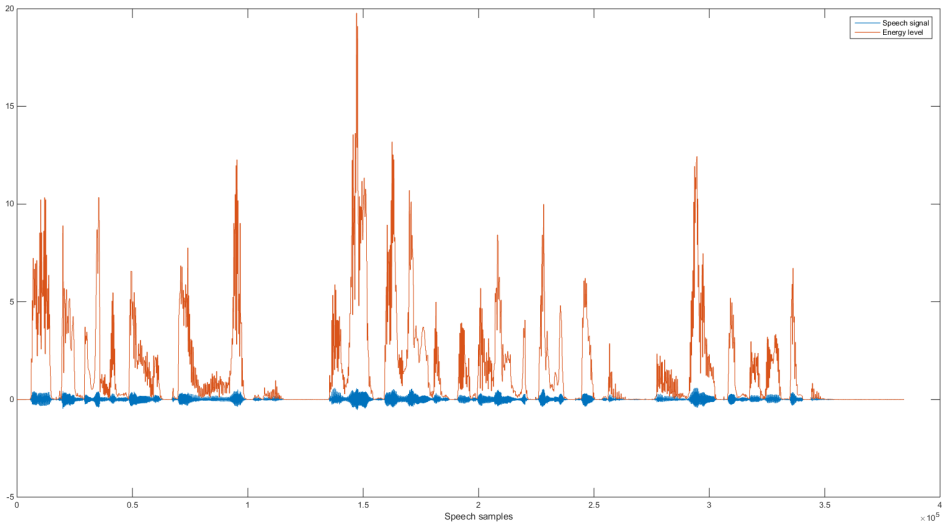


FIGURE 3.10. Energy level of the speech signal used in our simulations

the threshold of zero-crossing rate in a 4ms segment is set to be 10. The criteria for determining the voice activity was that if the energy was smaller than the energy threshold and the zero-crossing rate was larger than the ZCR threshold or the ZCR is equal to zero (silence), then the segment was assumed to be unvoiced/silence speech segment, otherwise, it was voiced speech segment. As shown in the Figure

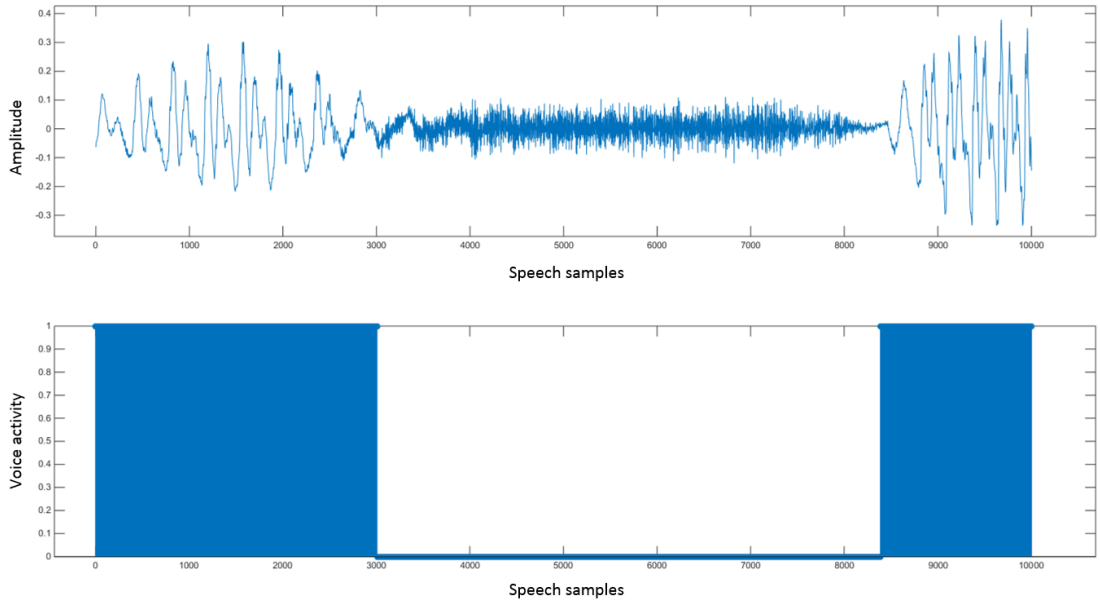


FIGURE 3.11. Simulation of voice activity detection

3.11, a speech segment containing 10000 samples was analysed by the proposed voice activity detection. In the detection algorithm, the unvoiced/silence speech signal was set as 0, while voiced speech signal was set as 1. From the figure, the speech segment was perfectly classified into two classes.

Additionally, there are two points need to be pointed out: First, in normal VAD algorithm, the speech is divided into frames by window functions and frames are normally longer than 4ms. In each frame, the speech signal is considered to be stationary and has identical voice activity. In my case, packets arrive at the receiver side one by one and are indicated as the loss or not immediately. If a loss occurs, the voice activity of this segment needs to be predicted based on the previous signal in order to take any action for repairing it. The experiments show that the voice activity of a speech segment can be precisely detected in the 4ms segment with appropriate thresholds and can be used to predict the voice activity of the next part. Second, the criteria for the unvoiced/silence segment detection should be more critical than the one for voiced segment, since WSM can be applied to unvoiced/silence segment with no harm to the reconstructed

signal (even there is no similarity in unvoiced speech signals, WSM can still find a part which substitutes the lost unvoiced signal), however, if the voiced segment is detected as unvoiced/silence, repetition may degrade the speech quality as mentioned before.

- **Similarity matching**: One essential factor influencing the performance of waveform similarity matching algorithm is the accuracy of finding the template' (see Figure 3.8), which best matches the template. There are a number of methods used to calculate the waveform similarity. Template' slides from the beginning of history buffer to the end sample by sample.

  One simple method is the cross-correlation, which measures the similarity between template and shifted copies of template'. If the number of samples in template is T, starting position of template' is n, history buffer size is N, x and y indicates the template and template', respectively. then the cross-correlation formula is:

$$S_{xcorr}(n) = \sum_{t=1}^{T} x(t)y(n+t), \quad n = 1,2,...,N-T \tag{3.3}$$

  The result of the match is the value of n corresponding to the maximum $S_{xcorr}(n)$. However, the value of S(n) can be really affected by the signal amplitude and lead a mismatch, which significantly degrades the performance of this algorithm. In order to get results which are sensitive to the waveform shapes rather than signal amplitudes, normalized cross-correlation (Equation 3.4) has to be applied. In my simulation, normalized cross-correlation worked way better than the non-normalized one in this algorithm.

$$S_{xcorr_{norm}}(n) = \frac{\sum_{t=1}^{T} x(t)y(n+t)}{\sqrt{\sum_{t=1}^{T} x^2(t) \sum_{t=1}^{T} y^2(n+t)}},, \quad n = 1,2,...,N-T \tag{3.4}$$

  Another commonly used method is the average magnitude difference function (AMDF), which is defined as

$$S_{amdf}(n) = \sum_{t=1}^{T} |x(t) - y(n+t)|, \quad n = 1,2,...,N-T \tag{3.5}$$

The result of the match is the value of n corresponding to the minimum $S_{amdf}(n)$. AMDF technique faces the normalization as well. In order to alleviate the influence introduced by the level changes, three normalized AMDFs are applied instead of the basic one. One is normalized by the square root of energy, the formula is shown below:

$$S_{amdfroot}(n) = \sum_{t=1}^{T} \left| \frac{x(t)}{\sqrt{\sum_{t=1}^{T}[x(t)]^2}} - \frac{y(n+t)}{\sqrt{\sum_{t=1}^{T}[y(n+t)]^2}} \right|, \quad n = 1,2,...,N-T \tag{3.6}$$

Second one is normalized by the sum of the absolute magnitudes of the samples, which is quite similar with the previous one.

$$S_{amdfabsolute}(n) = \sum_{t=1}^{T} \left| \frac{x(t)}{\sum_{t=1}^{T}|x(t)|} - \frac{y(n+t)}{\sum_{t=1}^{T}|y(n+t)|} \right|, \quad n = 1,2,...,N-T \tag{3.7}$$

Third one is normalized by dividing by the peak-to peak amplitude.

$$S_{amdfpeak}(n) = \sum_{t=1}^{T} \left| \frac{x(t)}{x_{max} - x_{min}} - \frac{y(n+t)}{y_{max} - y_{min}} \right|, \quad n = 1,2,...,N-T \tag{3.8}$$

where, x_{max} and y_{max} are the maximum value in x and y, respectively, and similarly for x_{min} and y_{min}.

In the simulation,compared with non-normalized methods, normalized methods made the search be sensitive to the waveform shape instead of amplitude change, so the reconstructed waveform was much more similar with the original. As for above four normalized methods, they produced almost equivalent results in our simulation.

- **Packet size, Template size and history buffer size**: In this section, we present the effects of packet size, template size and history buffer size on performance of reconstructions. To perform above evaluations, we generate a number of loss patterns by four-state Markov model (see appendix A) and recover losses by waveform similarity matching with various parameters. Composite objective measures (see

appendix C) is used to obtain quality scores. The packet size is changed from 2ms to 32ms, the template size is from 1ms to 30ms and the history buffer size is from 2ms to 512ms. The scores are average scores derived from different loss patterns and audio samples.

As shown in the Figure 3.12(a), compared with template size and history buffer size, the packet size is the major impact on the performance of reconstructions and smaller packet size shows higher quality score. The reasons is that smaller packet size always produces smaller loss gap, which is easier to be recovered by WSM algorithm because the missing segment is so short that the characters of the speech signal do not change significantly and the speech signal is still stationary. Apart from high quality scores, smaller loss gap saves resources (history buffer) used to recover losses. However, the packet size cannot be too small to lose the efficiency in packet transmissions. Considering all benefits and limits, the packet size is chosen to be 2ms or 4ms.

With 2ms packet size in our case, WSM always faces to a 4ms loss because of the odd-even interpolation, so we take the 4ms plate out of the Figure 3.12(a) and plot it in the Figure 3.12(b) in order to investigate the specific template size and history buffer size used in the WSM algorithm. It's easy to find that the highest scores concentrate on certain part in the figure. We cut two slices out of the surface. Figure 3.13(a) shows quality scores as a function of template size when the history buffer size is 32ms. From the figure, the appropriate template size is between 3ms and 5ms. If it is too small, the waveform information in the template is simply insufficient so that WSM may not find the ideal match. The quality also goes down when the template is too long, because it contains too many uncorrelated waveform components to find the lost segment. The choice of template size appears to be independent of packet size and history buffer size. Again, we eventually determine the template size as 4ms, which is a safe choice to implement this algorithm. Figure 3.13(b) plots quality scores as a function of history buffer size when the template size is 4ms. If the history buffer is too short, then it will omit the best substituted waveform, especially for some male voice with very low frequency. After a certain length, the history buffer size does not influence scores significantly as long as you have a perfect similarity matching algorithm. However, a long history buffer seems not necessary because it is resources-consuming. Additionally, longer history

(a) (b)

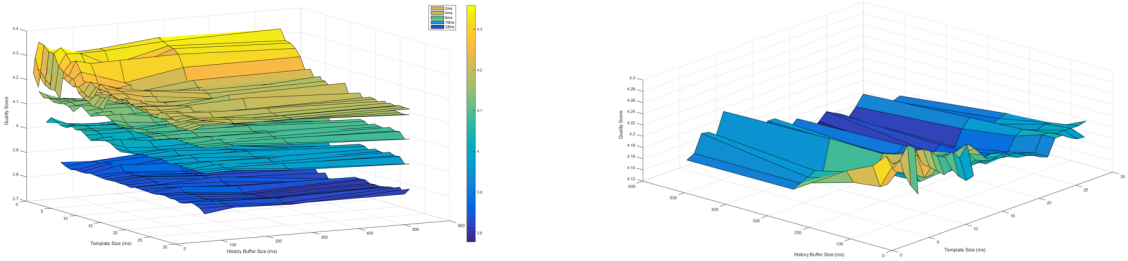


FIGURE 3.12. (a) Quality score with different packet size, template size and history buffer size. (b) Quality score with different template size and history buffer size when the packet size is 4ms.

(a) (b)

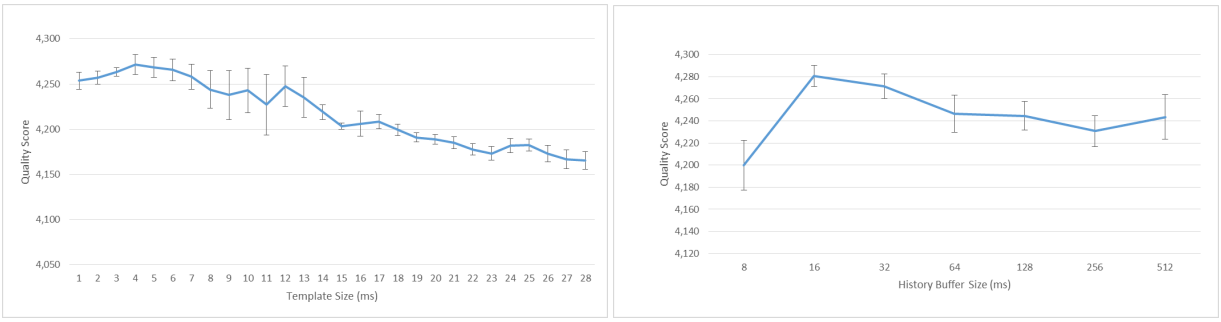


FIGURE 3.13. (a) Quality score versus template size when the history buffer is 32ms. (b) Quality score versus history buffer size when the template size is 4ms.

buffer is always equivalent with longer time to determine the best substituted waveform and introduces undesirable latency. As a result, we suggest that 30ms is an appropriate and safe choice for history buffer. Note that the history buffer appears to be dependent of packet size and template size, because it must contain at least one packet length for the recovery.

- **Packet Merging**: In waveform substitution, speech impairments can happen due to the discontinuities at boundaries between the successfully received speech segments and reconstructed speech segments. The discontinuities yield artificial effects, which is pretty annoying by listening. There are couple of methods used to smooth discontinuities. One method called phase matching algorithm is proposed in [41]. the reconstructed speech segment is time-scaled so that the ending phase

matches the following speech segment in a natural way without any discontinuity. However, a reliable phase estimator always imposes complexity into the algorithm. Another method is moving average, which creates series of averages of different parts of the speech segment. Moving average technique can smooth discontinuities with producing the distortions caused by averaging samples to the speech waveform. An overlap-based based packet merging technique was presented in [11]. The reconstructed waveform overlaps and adds with the received speech using window function which yields a weighted sum between the overlapped segments. This techniques, besides the fact of being easy to implement, the performance is also satisfying.

In this thesis, we use this overlap-add based packet merging technique to smooth discontinuities. In the similarity matching part, we take *Tm* more samples at each end of the substitution packet to become the Head (H) and Tail (T), respectively. As shown in the Figure 3.14, when a packet is indicated as a loss, *Received Packet 1* is delayed to be played-out, the Head of *Substitution Packet a* overlaps and adds with the last part of *Received Packet 1* by crossed Hanning window functions, while the Tail is not applied until the *received packet 2* arrives. In our simulation, the value of *Tm* was chosen to be 1ms. If *Tm* was too short, the discontinuity could not be erased perfectly, if *Tm* was too long, the overlap-add process introduced more latency because of the continuous update scheme had to delay received parts before and after the loss. In our simulation, a long *Tm* period seemed meaningless for smoothing discontinuities.

Waveform similarity matching (WSM) is effective when characters of speech do not change significantly during the missing part. Speech waveforms show quasi-stationary intervals which always fall into one of three distinct categories: voiced speech segments, unvoiced speech segments and silence. WSM will lead a wrong substitution if the transition from one category to another within the missing part. As shown in the Figure 3.15, there is a transition from unvoiced speech to voiced speech within the missing part. Based on previous received packets, WSM will simply copy the preceding packet since the missing part is predicted to be the unvoiced speech segment. This reconstruction leads a distortion as well as a loss of speech information. Some algorithms named transition protection were developed to protect the transition, however, such algorithms are very complex to implement and may introduce undesirable latency since they need to wait

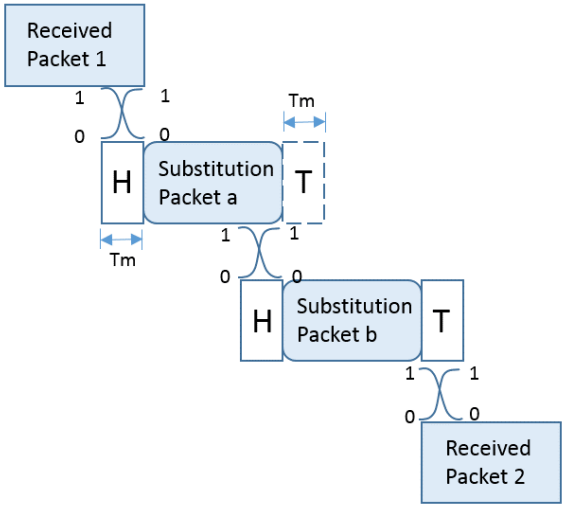


FIGURE 3.14. Packet merging technique

for other packets which contain transition information and then start recovering the waveform of that missing transition. As shown in the Figure 3.15, the use of odd-even interpolation may solve above problem in WSM because it can retrieve the speech information in the transition when only one of the packets containing the information of transition is lost.

Another reason makes WSM find an incorrect waveform substitution is that characters of the speech signal change within the missing part. As shown in the Figure 3.16, during the missing part, characters of the signal have changed and the signal cannot be assumed as a stationary signal with the prior part. Using the only preceding signal to reconstruct the missing part must lead a distortion. Again, two-side methods (fill-in the missing part from preceding and succeeding signals) cannot be used because the latency constraint doesn't allow the algorithm to take actions after receiving succeeding signals. To some extent, odd-even interpolation alleviates this kind of distortion, since the information of missing speech segment containing changing characters may be recovered by adjacent packets from both sides of the loss.

Above two problems in WSM can be alleviated by odd-even interpolation. These are also motivations to combine both of them together. Unfortunately, WSM sometimes produce

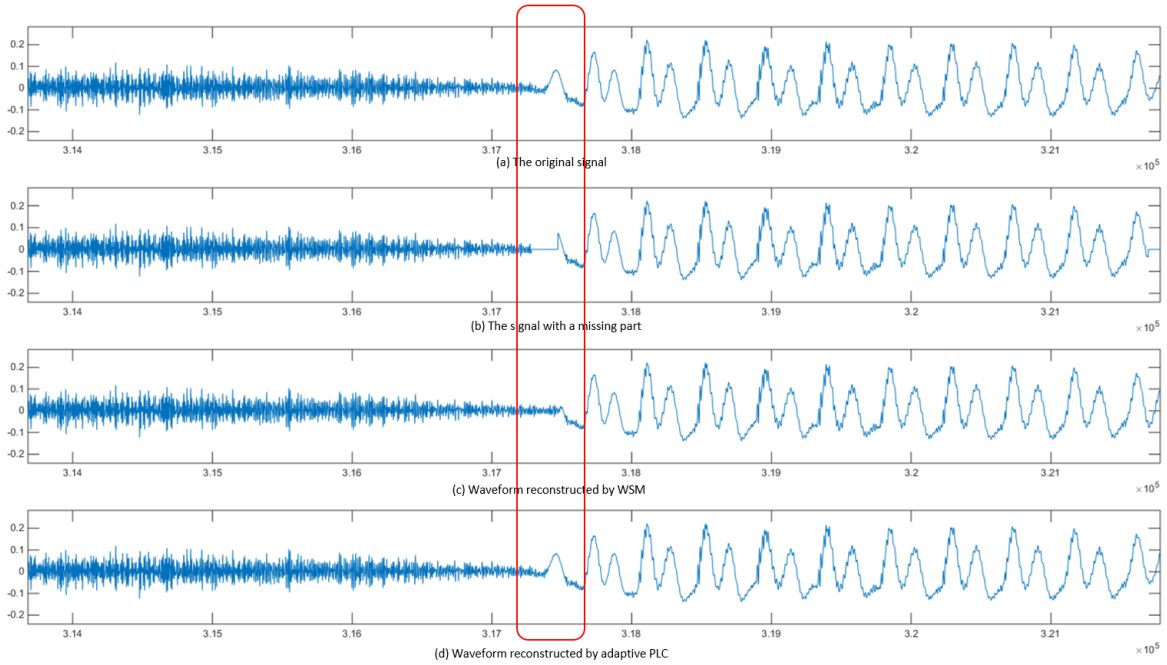


FIGURE 3.15. A transition from unvoiced speech to voiced speech within the missing part

periodic signal like PWR does, because the substitution may be a part of waveform in previous pitch period and produce artificial effects, however, this probability is small.

3.3.3 Attenuation and Silence Substitution

Packet transmissions sometimes suffer heavy congestion and disconnection with servers. The duration of burst loss may go up to hundreds milliseconds, or even longer. So far, no one PLC algorithm at receiver side can handle this without any assistance from codec information, transmitter part, etc. However, if the PLC algorithm is kept working locally in a standalone system, additional complexity and latency are not expected, especially in real-time interactive applications.

With long recoveries, it is necessary to attenuate the speech signal as the recovery process. As the reconstructed speech gets longer, the speech signal is more likely to diverge from the original signal. For the first 10ms of the recovery, reconstructed waveform would not attenuate significantly since the size of gap is unknown to the PLC and most of loss runs are shorter than 10ms (see measurements presented in the Figure 2.1). Slow attenuation

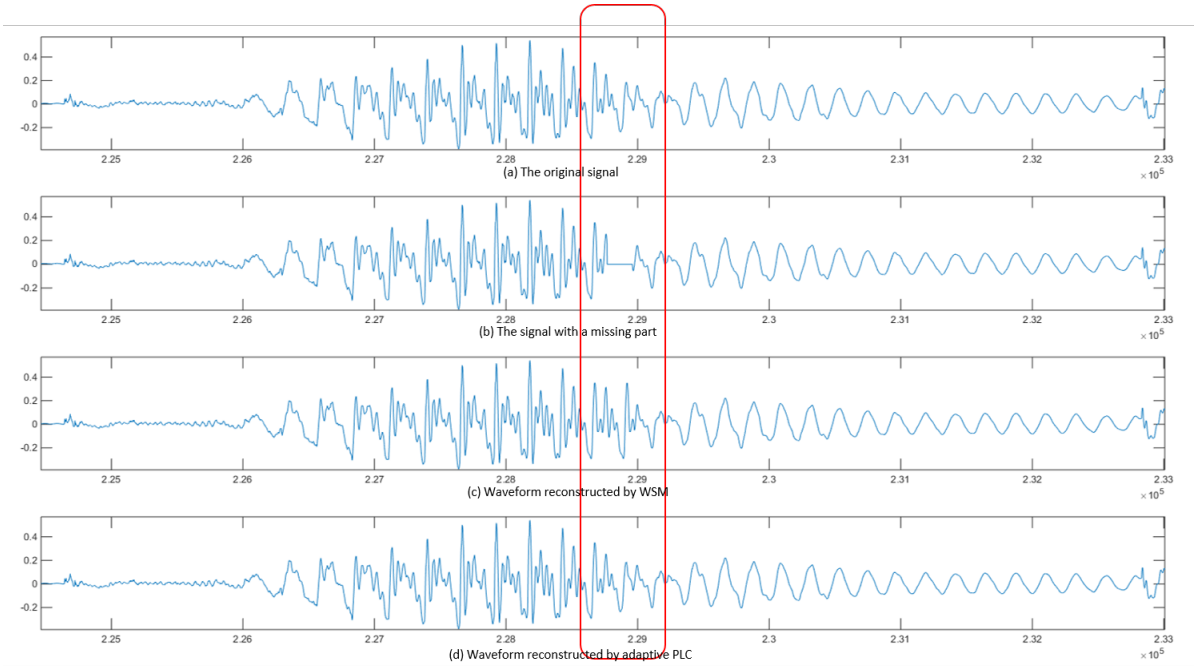


FIGURE 3.16. Characters of the speech signal change within the missing parts

at the beginning makes sure that the amplitudes of most of recovered waveform for the short burst loss are retained. At the start of coming recovery, the attenuation factor increase rapidly to zero. After 30 ms, the attenuation factor becomes zero, so the silence substitution will be applied instead, because the speech signals can only be assumed as stationary signals within 30ms. In another word, WSM reconstructs waveform based on waveform shape and characters of speech signals do not change significantly for a short duration. If the loss gap is longer than 30ms, it is unreasonable to recover the later waveform by preceding signals in the PLC buffer.

In order to attenuate reconstructions slowly in first 10 ms, while rapidly in the coming 20ms. Attenuation factors are obtained from the second half of the Hanning window function. Silence substitution stops when the incoming packet arrives. A fade-in function with 1ms duration will be applied to smooth the boundary between the silence substitution and the incoming packet.

3.4 Signal Processing Steps and Latency Consumptions

Signal processing steps:
Adaptive algorithm is composed of three PLC algorithms, all these three algorithms should be implemented in the system. However, three of them do not need to run at the same time. There is only one algorithm will be applied in every recovery. Steps of signal processing can, to a degree, be used to indicate how complex the algorithm is.

As for odd-even interpolation, the algorithm can be explained as a 2-step process: 1) creating a new sequence x_n consisting of the original sequence, separated by n-1 zeros. 2) if both of packets are received, replacing zeros by samples in the second packet. If not, smoothing the new sequence with a low-pass filter, which substitutes the zeros.

As for WSM, the first step is voice activity detection which contains two-steps: 1) zero-crossing rate calculation 2) energy computation. Then, if the segment is unvoiced/silence, a simple repetition is employed and samples in previous segment are simply copied. If the segment is voiced, 4-step process will be conducted: 1) normalized cross-correlation/AMDF is used to find the substituted samples. 2) the substituted samples with head and tail are copied. 3) attenuation function is applied to samples. 4) overlap-add happens at boundaries to smooth discontinuities.

As for silence substitution, only filling-in the gap with zeros is needed.

Latency consumptions:
Whole process produces extremely small latency because of the continuous update scheme. The latency existing in packet buffer parts really depends on the packet size. In our simulation, we chose the packet size to be 2ms. Figure 3.17 showed latency budget in the system, each buffer introduced 4ms (2 packets) delay by collecting speech samples. The reason of collecting 2 packets was that odd-even interpolation needed to make the decision after both even-sample and odd-sample packets were collected at the receiver side. The latency in WSM algorithm was imposed by packet merging. Since overlap-add parts existed at each end of the packet (see Figure 3.14), there were 2ms extra (2*Tm=2*1ms=2ms) latency. In total, the computational latency in the system was 10ms (2*4ms+2ms=10ms), which was acceptable for real-time applications like Bosch Dicentis

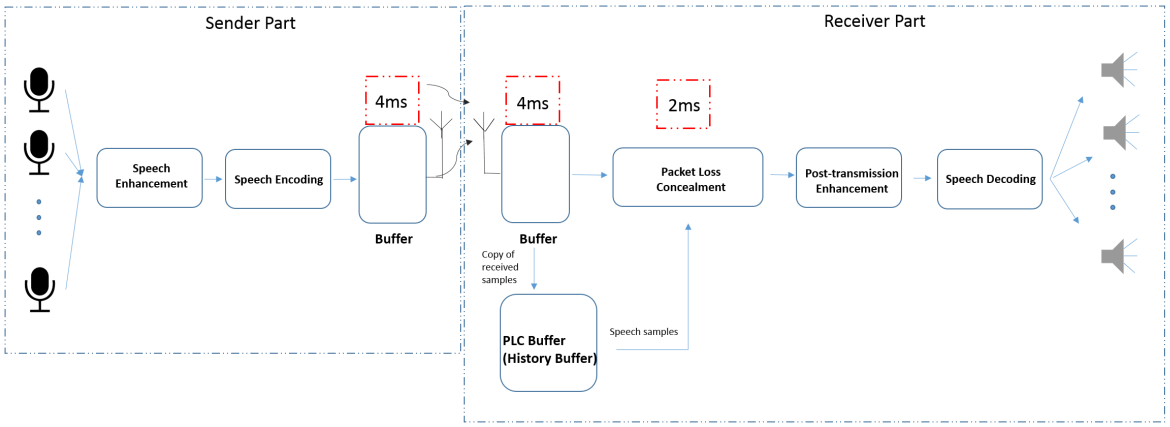


FIGURE 3.17. Latency budget in the system

conference system. Of course, some other parts, such as computation procedures, step transformations, etc., introduced delays to the system as well, however, the effects were so small that those latencies could, to some degree, be ignored.

CHAPTER 4

RESULTS, ANALYSIS AND DISCUSSION

In this chapter, I used MUSHRA listening test to evaluate the performance of several packet loss concealment algorithms, such as odd-even interpolation, waveform similarity matching (WSM), pitch waveform replication (PWR), under different loss conditions. First, details in experiments are presented and it is followed by evaluation results and corresponding discussions.

4.1 MUSHRA listening test

4.1.1 Subjective Quality Measures

Subjective quality measures are measured based on the subjective opinions from listeners on the quality of the speech and considered as being the most reliable method of evaluating the speech quality.

Multi-Stimulus Hidden Reference and Anchor (MUSHRA) is one of the advanced opinion-based measures defined by ITU-R recommendation BS.1534-1 [35]. Multi-Stimulus allows listeners to have instant random access to each of test items and the reference signal. Hidden reference means one of the test items is a copy of the reference signal (absolutely clean speech signal). Compared with other recommendations for listening test, MUSHRA uses multi-stimulus with hidden reference, continuous quality scale and loop function, which make sure listeners are able to deal with clearly audible differences

FIGURE 4.1. Relation between MUSHRA scores and speech quality

and give rational scores.

MUSHRA consists of two phases: training phase and evaluation phase. The former is required by the standard and lets listeners familiarise themselves and stimuli they are going to listen. Comparing all signals with the reference signal, listeners have to evaluate each signal belonging to different excepts by giving scores between 0 (really bad) and 100 (absolutely excellent). The numerical scores are related to speech qualities as shown in Figure 4.1. All the scores are statistical analysed by analysis of variance (ANOVA) and mean scores, the details will be presented in the next section.

4.1.2 Statistical Analysis to Results

In this thesis, one-way analysis of variance (ANOVA) is used to process the data collected from listening tests. ANOVA is applied to determine whether there are statistically significant differences between independent groups. Compare with other statistical analyses like t-test and comparison of mean values, ANOVA mainly focuses on the differences among several groups and clearly shows the degree of those differences. Also, ANOVA could show the range of absolute values given to the speech quality from listeners. The absolute values tell us the quality of speech after recoveries and how good the algorithm is. The illustration of ANOVA plot is described in the Figure 4.2. the Red line represents the median number (q2) in this data set, the horizontal blue lines which are higher and lower than median number represent the 75th (q3) and 25th (q1)

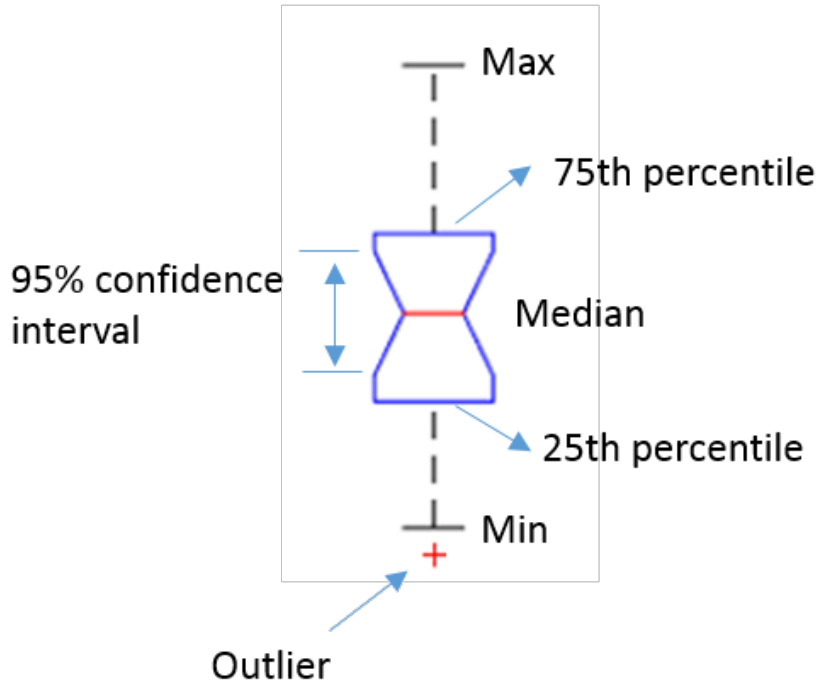


FIGURE 4.2. Illustration of the ANOVA plot

percentile of the data, respectively. The maximum and minimum values are shown with black lines on the top and bottom. The outlier is plotted individually using the red "+" symbol. The median number comes up with a 95% confidence interval, which reflects deviations among different excepts and listeners. Here, 95% confidence interval (CI) is presented in the following equation:

$$95\%CI = q2 \pm \frac{1.57(q3 - q1)}{\sqrt{NJ}} \tag{4.1}$$

where q2 is the median score of certain PLC algorithm, q1 and q3 are the 25th and 75th percentiles, respectively. N and J is the number of listeners and excerpts, respectively.

With 95% confidence interval, we're 95 percent confident that the true median value for the scores ranges from the upper confidence limit to the lower confidence limit. In another word, if the 95% confidence intervals of two groups do not overlap with each other, then two medians are significantly different at the 5% significance level. Apart from ANOVA, the mean score for each algorithm is calculated by averaging the scores from different excepts and listeners:

$$\overline{\mu}_i = \frac{1}{NJ} \sum_{n=1}^{N} \sum_{j=1}^{J} \mu_{ijn} \tag{4.2}$$

where μ_{ijn} is the score of PLC algorithm i for a given except j and a listener n.

The mean scores are used to refer to a central value of the scores given to different algorithms.

4.1.3 Test Set-up

The reference signals were sampled by 48kHz and divided in packets of 2ms lengths, they were processed by different PLC algorithms under different loss scenarios (see Table 4.1 and Table 4.2). The whole test consisted of two listening tests: test A and B.

In the **listening test A** (as shown in Table 4.1), 2 female (around 7 seconds long, 48kHz sampled speech sample with high pitch frequency) and 2 male (around 7 seconds long, 48kHz sampled speech sample with low pitch frequency) reference speech signals were corrupted by different loss patterns with fixed packet loss rate 4%. The insertion points of the different type of loss were chosen randomly in the sentence. As explained in the chapter 2, there were five types of loss patterns used to investigate performances of different PLC algorithms (adaptive PLC, WSM and PWR) under different packet loss conditions for each reference speech signal. First, single loss at a high rate in bursty periods meant single losses got close in a short time period (see chapter 2). Second to fourth, changing the burst length of losses: 1,2,4 packets. Fifth, generating losses which were 20 packets long (40ms). The corrupted signal without any repair was used to be the low anchor signal in MUSHRA test (see standards in MUSHRA). This test contained 100 scores in total (4 sentences x 5 loss patterns x (4 PLC algorithms +1 reference)).

In the **listening test B** (as shown in Table 4.2), only 1 female (23 seconds long, 48kHz sampled English speech sample) and 1 male (22 seconds long, 48kHz sampled English speech sample) reference speech signals were corrupted by changing the loss rate from 2% to 8%, which were generated by four-state Markov model (see appendix A). Compared with the audio samples used in the listening test A, audio samples in the listening test B are longer. With the same packet loss rate, longer sentences always have larger number of packet loss and more types of loss. This makes sentences corrupted by the model be

able to represent the one with packet loss in the realistic network. Each audio sample was processed by two algorithms: Adaptive PLC and State-of-Art PLC (see appendix B). The corrupted signal without any repair was used to be the low anchor signal in MUSHRA test (see standards in MUSHRA). This test contained 24 scores in total (2 sentences x 3 loss rates x (3 PLC algorithms +1 reference)).

Listening test A aimed to investigate which algorithm behaved the best in each loss scenario. Listening test B was the comparison between the whole adaptive PLC algorithm and STA PLC algorithm, to confirm that the adaptive one behaved better than State-of-Art PLC in realistic transmissions with different packet loss rate.

Considering the human concentration dropped down with the increasing of listening time, Listening test A and B were conducted separately and each one just costed listener about 20-30mins. This action ensured that results reflected the reliable opinions from listeners and were able to be used to represent algorithms' performances.

In a listening test, more listeners always give more reliable and convincing results, but the listening test is a time-consuming and cost-consuming process. Additionally, the quality of ears of listeners also influences the reliability of results. After a careful consideration, a total of 9 listeners with "good" ears were chosen to participate in the listening test A and B. The listening tests were conducted in a separate testing room and all material was rendered to listeners using a professional headphone. Each listener could freely access the MUSHRA software and evaluate different algorithms without any constraint.

4.2 Test Results and Analysis

In this section, the results of above two described listening tests are presented to evaluate different PLC algorithms. In order to easily and clearly show absolute scores and comparisons between different algorithms, the results are shown in the figure, where ANOVA plot is first presented, then a table with the mean score of each algorithm follows. The horizontal axe in each ANOVA plot represents PLC algorithms and the vertical axe represents a score in the scale from 0 (Bad) to 100 (Excellent). In the result analysis, I put results from female and male speakers under the same loss condition together and draw a little discussion from the figure.

Table 4.1: Listening test A

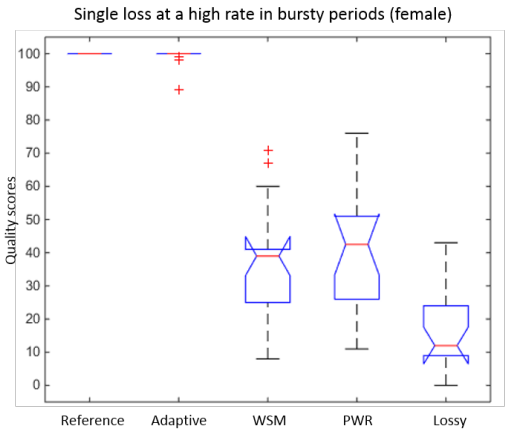
Listening Test A					
Gender	Samples	Excerpt	PLR(%)	Loss type	PLC algorithm
Female	Female 1	1	4	Single loss(high rate)	Adaptive,WSM,PWR,lossy
		2	4	Burst length 1	Adaptive,WSM,PWR,lossy
		3	4	Burst length 2	Adaptive,WSM,PWR,lossy
		4	4	Burst length 4	Adaptive,WSM,PWR,lossy
		5	4	Burst length 20	Adaptive,Adaptive(no silence),lossy
	Female 2	6	4	Single loss(high rate)	Adaptive,WSM,PWR,lossy
		7	4	Burst length 1	Adaptive,WSM,PWR,lossy
		8	4	Burst length 2	Adaptive,WSM,PWR,lossy
		9	4	Burst length 4	Adaptive,WSM,PWR,lossy
		10	4	Burst length 20	Adaptive,Adaptive(no silence),lossy
Male	Male 1	11	4	Single loss(high rate)	Adaptive,WSM,PWR,lossy
		12	4	Burst length 1	Adaptive,WSM,PWR,lossy
		13	4	Burst length 2	Adaptive,WSM,PWR,lossy
		14	4	Burst length 4	Adaptive,WSM,PWR,lossy
		15	4	Burst length 20	Adaptive,Adaptive(no silence),lossy
	Male 2	16	4	Single loss(high rate)	Adaptive,WSM,PWR,lossy
		17	4	Burst length 1	Adaptive,WSM,PWR,lossy
		18	4	Burst length 2	Adaptive,WSM,PWR,lossy
		19	4	Burst length 4	Adaptive,WSM,PWR,lossy
		20	4	Burst length 20	Adaptive,Adaptive(no silence),lossy

Table 4.2: Listening test B

Listening Test B			
Gender	Excerpt	PLR(%)	PLC algorithm
Female	1	2	Adaptive PLC,State-of-Art PLC,lossy
	2	4	Adaptive PLC,State-of-Art PLC,lossy
	3	8	Adaptive PLC,State-of-Art PLC,lossy
Male	4	2	Adaptive PLC,State-of-Art PLC,lossy
	5	4	Adaptive PLC,State-of-Art PLC,lossy
	6	8	Adaptive PLC,State-of-Art PLC,lossy

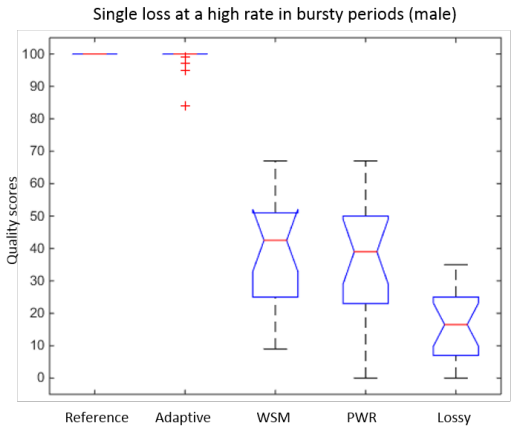
- **Listening test A – single loss at a high rate in bursty period**

(a)



	Refer.	Adaptive	WSM	PWR	Lossy
Mean	100	99	36	41	14

(b)



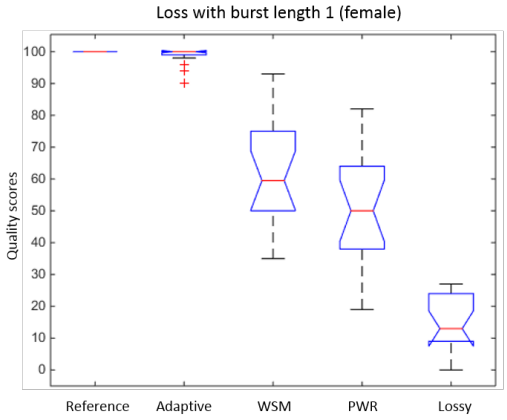
	Refer.	Adaptive	WSM	PWR	Lossy
Mean	100	99	38	37	16

FIGURE 4.3. (a) Scores for different algorithms from female sequences. (b) Scores for different algorithms from male sequences.

The plots in the Figure 4.3 showed that the adaptive method behaved way better than waveform similarity matching (WSM) and pitch waveform replication (PWR) in this scenario. The scores given to the adaptive method approached 100, corresponding to a excellent perceiving quality, because adaptive method in this case worked exactly the same with odd-even interpolation, which perfectly reconstructed the lost waveform based on odd-sample or even-sample in that packet. While WSM and PWR gave unsatisfying scores, since both of them relied on the waveform information in the history buffer to estimate the lost part, if there were so many losses in a bursty period and substitutions of early loss were updated to the history buffer in order to recover the later loss, the distortions introduced by recoveries were actually accumulated with the increasing of times of the reconstruction. In another word, as the history buffer used to recovery the next loss got worse, the reconstructed waveform was more likely to diverge from the original.

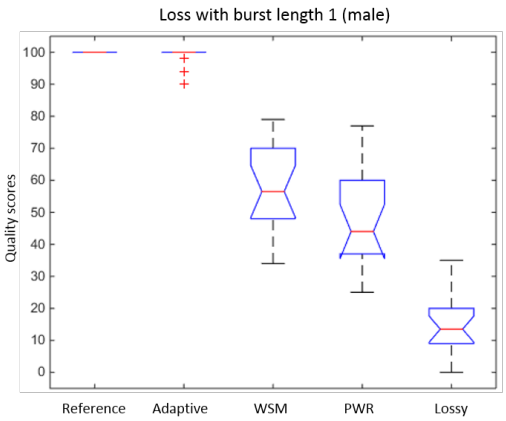
- **Listening test A – loss with burst length 1**

(a)



	Refer.	Adaptive	WSM	PWR	Lossy
Mean	100	99	64	52	14

(b)



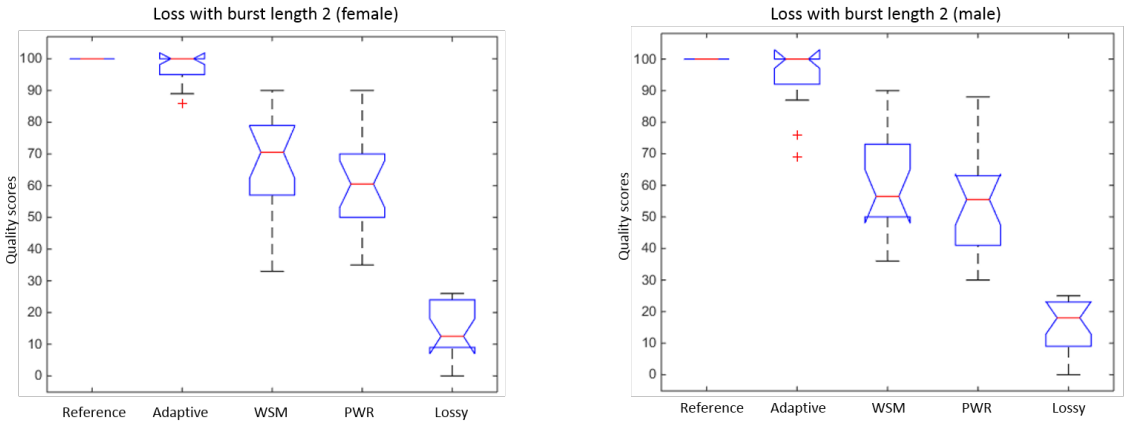
	Refer.	Adaptive	WSM	PWR	Lossy
Mean	100	99	57	48	14

FIGURE 4.4. (a) Scores for different algorithms from female sequences. (b) Scores for different algorithms from male sequences.

In Figure 4.4, for the loss with burst length 1, adaptive method showed extremely high scores again because of the contribution from the odd-even interpolation. Listeners hardly distinguished the difference between the one recovered by adaptive method and the original. As for WSM and PWR, they produced almost equivalent result, however WSM behaved slightly better than the PWR by looking at the median score and mean score.

- **Listening test A – loss with burst length 2**

(a) (b)



	Refer.	Adaptive	WSM	PWR	Lossy
Mean	100	98	62	61	14

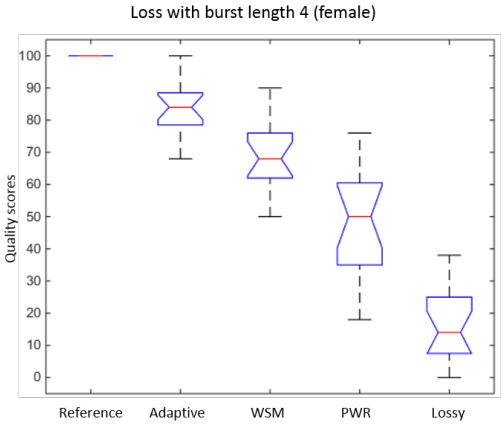
	Refer.	Adaptive	WSM	PWR	Lossy
Mean	100	95	61	58	15

FIGURE 4.5. (a) Scores for different algorithms from female sequences. (b) Scores for different algorithms from male sequences.

For the loss with burst length 2, compared with WSM and PWR, Figure 4.5 told us that adaptive method had really high scores, which were contributed by the odd-even interpolation again. If the even samples in one packet were lost and the odd samples in the next one were lost as well, they become a loss with burst length 2. However, this loss can be fully recovered by odd-even interpolation using adjacent packets of the loss. Taken to extreme, such losses existed with 50% probability in an infinite sentence. In another word, half of losses with burst length 2 could be perfectly reconstructed by odd-even interpolation instead of WSM. WSM and PWR did not produce significant difference in this case, since the 95% confidence intervals were overlapped.

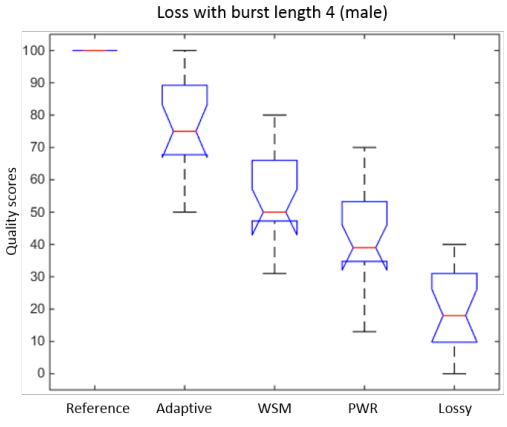
- **Listening test A – loss with burst length 4**

(a)



	Refer.	Adaptive	WSM	PWR	Lossy
Mean	100	84	70	49	15

(b)



	Refer.	Adaptive	WSM	PWR	Lossy
Mean	100	79	55	43	19

FIGURE 4.6. (a) Scores for different algorithms from female sequences. (b) Scores for different algorithms from male sequences.

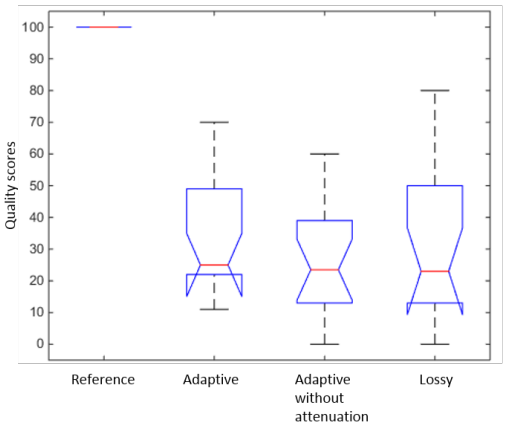
For the loss with burst length 4, Figure 4.6 (a) showed that, the adaptive method were preferred over the WSM, which worked significantly better than PWR. As explained in the last section, the odd-even interpolation was likely to shorten the loss to 2 consecutive packets long, which became much easier for WSM to repair. In another word, shorter gap always meant easier to repair. The reasons that WSM behaved better than PWR are twofold: first, as explained in the chapter 2, the PWR produced highly periodic speech signal because of the replication, which imposed "metal/tinny" sound and severely annoyed listeners. Second, PWR sometimes had wrong pitch detections which destroyed reconstructed speech signals significantly. Since the recovered waveform was repeated over the gap, repeated wrong substitutions led unbearable results after recoveries. For scores from the male side, as shown in the Figure 4.6 (b), the 95% confidence intervals of WSM and PWR were smally overlapped since the pitch periods for male speakers are normally longer than those for female speakers, replicated signals with 4-packet length might not that periodic. This made the difference between WSM and PWR become small.

However, in general, the WSM behaved better than PWR.

We also noticed that scores of the adaptive method were lower than its score in both loss with burst length 1 and loss with burst length 2 scenarios, because the adaptive method could not only rely on the odd-even interpolation any more, it must need WSM to repair the loss when facing to a loss with burst length 4. Additionally, the difference between WSM and PWR became larger with the increasing of the length of the burst loss, because PWR is based on replications, which occurred more times and generated more periodic signals when the loss got longer.

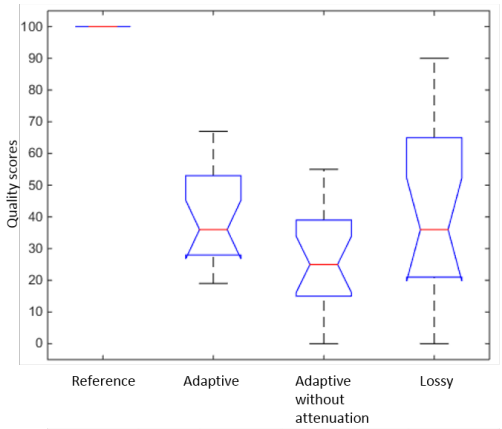
- **Listening test A – loss with burst length 20**

(a)



	Refer.	Adaptive	Adaptive without attenuation	Lossy
Mean	11	33	27	31

(b)



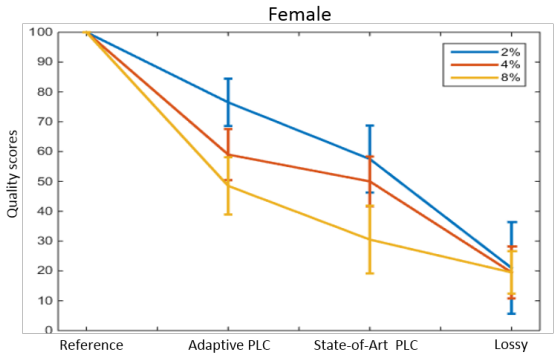
	Refer.	Adaptive	Adaptive without attenuation	Lossy
Mean	100	40	27	40

FIGURE 4.7. (a) Scores for different algorithms from female sequences. (b) Scores for different algorithms from male sequences.

Figure 4.7 showed that, in general, the adaptive method with attenuation, the adaptive method without attenuation and the lossy signal without any repair had almost equivalent performances when losses became long to 40ms. Normally, if the loss was very long, there was no single algorithm could handle that. As times of the repair got more, the reconstructed waveform was more likely to diverge from the original and produced unbearable artificial effects. Silence substitution seemed like a good way to reduce the artificial effects as well as complexities. Since we never know the size of gap because of the latency constraints and signal should attenuate to silence after a certain duration, attenuation factors had to be applied in order to smooth boundaries between reconstructed signals and silences. According to feedbacks from listeners, most of them preferred to listening missing speech information (silence) instead of annoying artificial effects. However, only 9 listeners participated in the listening test, more listeners should be involved to confirm that the silence is preferred over artificial effects caused by PLC.

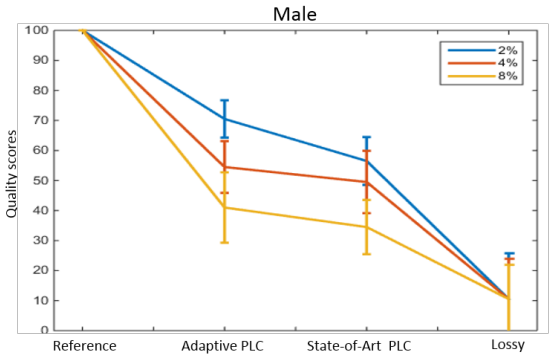
- **Listening test B**

(a)



	Refer. (mean score)	Adaptive (mean score)	STA (mean score)	Lossy (mean score)
2%	100	75	58	26
4%	100	62	51	20
8%	100	49	33	16

(b)



	Refer. (mean score)	Adaptive (mean score)	STA (mean score)	Lossy (mean score)
2%	100	72	60	24
4%	100	57	48	20
8%	100	47	36	18

FIGURE 4.8. (a) Scores for different algorithms with different packet loss rate from a female sequence. (b) Scores for different algorithms with different packet loss rate from a male sequence.

In the listening test B, I aimed to simulate the realistic packet loss in the network. Packet loss patterns with different packet loss rate in the listening test B were generated by four-state Markov model (see appendix A). Here, the plot are simplified by connecting median scores with 95% confidence interval by lines in order to clearly show comparisons between different algorithms and their performances when facing to different packet loss rates.

As shown in the Figure 4.8, the performances of both adaptive PLC and State-of-Art PLC (based on PWR, see appendix B) dropped down with the increasing of the packet loss rate. The adaptive PLC produced higher scores than State-of-Art PLC did by looking at median and mean scores. However, the overlapped confidence intervals told us that the difference, to some degree, was not that significant, since loss conditions in real network were complicated and packets might be heavily lost in a time period when a strong interference came to the network suddenly

or a heavy congestion suddenly happened in the transmission. Under conditions mentioned above, performances of both algorithms degraded.

CHAPTER 5

SUMMARY, CONCLUSION AND FUTURE WORK

Packet loss is still an inevitable problem in the use of the packet network. With the development of packet loss concealment, to some extent, packet loss can be made up. However, packet loss concealment with high performance is still considered as a research challenge as well as an indispensable part in many IP-based telecommunication applications.

There are so many constraints in designing a PLC algorithm in telecommunication systems, for instance, some applications do not have assistances from the sender and require that the PLC could work independently with the coding mechanism, in other words, the PLC can only deal with the speech, which exists in the PLC buffer. Another constraint is the latency requirement in many real-time telecommunication systems such as Bosch Dicentis conference system, which requires a really low latency to ensure that participants are able to communicate with each other in real-time. As a result, the PLC algorithm cannot introduce too many latencies to the system. Apart from latency requirements, the low complexity and low resource consumption are also essential in designing a feasible PLC algorithm in a practical system.

In this thesis, we first decided a scheme named continuous update, which repaired the loss immediately as long as a loss was indicated and imposed extremely small latency to the system. Then, we developed a PLC algorithm named adaptive algorithm, which changed the algorithm to use depending on the number of the consecutive loss. As results shown in the chapter 4, the adaptive method gave us satisfying scores. Additionally,

whole algorithm fully worked on the receiver side without any codec information and fulfilled the latency requirements.

This last chapter concludes the thesis by presenting the problems behind each algorithm and conducting some discussions about them. Eventually, some future works related to present work, in general, are proposed to improve the performance of PLC algorithm in the system.

5.1 Summary

The aim of this thesis was to design a packet loss concealment scheme which satisfied the constraints in Bosch Dicentis conference system and achieved a high performance. Beside, the algorithm could be implemented in any other system as well.

The logic behind this thesis started from investigating packet loss characteristics in realistic networks and existing PLC algorithms (see chapter 2), since understanding the type of loss we have in the network and problems in present algorithms was of importance to design an appropriate PLC algorithm which overcame the drawbacks of other methods and perfectly repaired the loss. Then, a scheme named continuous update was proposed to lower the latency of the system and a algorithm combining the odd-even interpolation, waveform similarity matching and silence substitution was designed under the continuous update (see chapter 3). Finally, the performance of each algorithm was evaluated by a subjective measure called MUSHRA test (see chapter 4).

Packet loss is actually a really simple conception so that many PLC algorithms only focus on finding a substitution when a packet is lost. Although these algorithms behave well in the simulation, performances of implementations in practical system are not satisfying, since the packet loss in realistic network is way more complicated and has to be taken into account when designing a packet loss concealment. Packet loss characteristics were investigated by analysing the data measured from Dicentis conference system. We derived run-length distributions of lost and received packets (loss run-lengths were the number of lost packets in a row, receive run-lengths were the number of received packets in a row). From the distributions, we found that the packet loss in network is the combination of single loss and burst loss, whose length is up to hundreds milliseconds. Most losses consisted of a single packet and individual loss runs were short, but fell

close together in some certain periods. Through the observation of Wi-Fi environment by Wi-Spy (a tool to monitor the load and occupancy in Wi-Fi channels), the channel used to transmit data packets might become very congested because of interferences or data packets with higher priority in certain time periods, thus many packets were lost in such time slots and impaired the speech quality significantly. As a result, the proposed PLC algorithm should work on dealing with single loss, burst loss (short burst and long burst) and loss at a high rate in bursty periods.

The researches on existing PLC algorithms aimed to investigate advantages and disadvantages in each algorithm and see how to improve them. The main problem of the use of WSOLA in real-time applications was the latency constraint. WSOLA technique needs to know the size of gap and the history buffer to be stretched in order to cover the gap, both of requests introduced unacceptable latency to the real-time applications like Bosch Dicentis conference system. However, the packet repetition and pitch waveform replication could be implemented. Packet repetition gave a bad estimation of speech information and could not make sure the substitution can connect itself in a synchronized way. Pitch waveform replication behaved better than packet repetition, but it generated highly periodic signals which produced artificial effects. Additionally, pitch waveform replication was sometimes implemented with wrong pitch information and resulted in a terrible reconstruction. From the problems in existing PLC algorithms, there were three components needed to be considered in designing the proposed PLC algorithm: First, the reconstructed waveform should be similar to the original and make artificial effects as small as possible. Second, the discontinuities between the repaired and received waveform should be smoothed. Third, PLC algorithm should meet low-latency and low-complexity requirements.

After understanding packet loss characteristics in real networks and knowing constraints as well as requirements in the PLC design, an adaptive PLC algorithm consisting of odd-even interpolation, waveform similarity matching and silence substitution was proposed to cope with the packet loss. Since the major loss was single loss and it might happen at a high rate in a short period, odd-even interpolation could be used to deal with single loss with perfect results. To some extent, the odd-even interpolation might even shorten the burst loss. Waveform similarity matching was implemented based on the speech signal is quasi-stationary signal. The substitution part was two-packet long waveform which can connect the previous packet in a natural way instead of the previous packet or

the segment in previous pitch period so the periodicity could be reduced. As shown in packet loss characteristics, the length of burst loss might exceed hundreds milliseconds, so the silence substitution seemed the best and easiest way to recover such long losses. Adaptive algorithm adjusted different PLC algorithms to use depending on the number of consecutively lost packets in a row and included a packet merging algorithm which applies cross-overlap-add function to the boundary. Adaptive method showed good results (see chapter 4) and met the low-latency and low complexity requirements.

In the chapter 4, a subjective measure called MUSHRA test was used to evaluate the performance of different PLC algorithms. So far, subjective measures seem to be the best way to judge the performance of PLC algorithms and MUSHRA is one of the most advanced methods to do the listening test. Scores were analysed by ANOVA.

5.2 Conclusion

Adaptive PLC scheme proposed in this thesis is preferred over single PLC, such as odd-even interpolation, WSM, PWR and silence substitution. I had argued that the odd-even interpolation can perfectly repair single loss and parts of short burst loss without producing any distinguishable artificial effect since the lost information could be fully recovered by up-sampling adjacent packets which still contained a half original speech information. As shown in the chapter 4, odd-even interpolation behaved significantly better than WSM and PWR in first three scenarios. However, it failed when both odd-sample and even-sample packets were lost. Then, WSM was proposed to recover the loss. Compared with the PWR, WSM did not only repeat the part in last pitch period and did not need to know the pitch period so that it reduced the periodicity of reconstructed signals and alleviated the artificial effects caused by periodicities or substituting the part with wrong pitch information. In our simulations (see chapter 4), the difference between WSM and PWR became large with the increasing of number of consecutively lost packets. Since the WSM reconstructed the waveform based on the speech signal could be assumed as a stationary signal for a short duration, the performance of WSM degraded if there was a transition from one voice activity to another within the missing part or characters of speech signal had changed during the missing part. The use of combining odd-even interpolation and WSM together solved above problems and made the adaptive PLC work significantly better than the WSM. If the loss was longer than a certain length, the reconstructed signals were likely to diverge from the original and

produced annoying artificial effects. I argued that the silence substitution should be applied to alleviate those artificial effects. Depending on the results from chapter 4 and feedbacks from listeners who participated in the listening test, the one recovered by adaptive PLC with an attenuation and silence substitution sounded the same with the one without any repair, and both of them sounded slightly better than the one repaired by adaptive PLC without any attenuation. As a result, the odd-even interpolation, WSM and silence substitution were combined together to become an adaptive PLC, and it showed better results than State-of-Art PLC which is based on PWR in the listening test. In addition, the adaptive PLC fully worked at the receiver side under the continuous update scheme which was designed for meeting the real-time systems' latency requirements.

5.3 Future Work

After finishing this thesis, I think some associated researches should be considered in the future. The following suggestions are listed:

- To add some specific coding mechanisms appears to be the next stage to conceal missing packets in the current system. So far, all PLC methods mentioned above focus on using local speech signal information to estimate substitutions for missing packets. However, it can still miss a large context of speech information and produce annoying artificial effects. For codec based on transform coding or linear prediction, decoder can interpolate between states. With codec information, the missing packet could be reconstructed by speech parameters estimated from previous codec information.

- The performance of PLC algorithm and speech quality decrease with the increasing of the packet loss rate. There are two ways to enhance the speech quality after a transmission, one is to have a stronger PLC algorithm, another one is to reduce the packet loss rate in the transmission. To achieve an excellent speech transmission and better results after recoveries, small packet loss rate appears to be valuable. The reason of dropping packets in a transmission is that the channel for transmission is occupied or congested. A easy way to solve this is to switch the present channel to another one. A channel switching algorithm used to decide when the present channel need to be left and which channel can be utilized seems necessary. In the current system, we have a simple channel switching algorithm which makes the decision only based on the packet loss rate. However, the distribution

of packet loss is also essential for channel switching algorithm to make decisions. In addition, collecting the qualities of other channels always consumes a period of time and the quality of certain channel may have already changed. This leads the channel switching algorithm to make a worse decision. A smarter channel switching algorithm should be considered in the future.

- Existing PLC algorithms focus on repairing losses in speech transmission based on properties in speech signal, such as periodicity, available pitch information, etc. However, telecommunication systems usually transmit other types of signal like music, which doesn't have many properties like speech signal. Designing a PLC algorithm which not only recovers losses in the speech transmission but also deals with losses in the music signal should be considered in the future.

- To minimize the large deviation in scores given by subjective measures like MUSHRA listening test, it's necessary to involve more listeners into listening tests. On the other hand, inviting people to do listening tests is a time-consuming and cost-consuming work. Another way to reduce the deviation in scores is to involve more sentences from different female and male speakers, but it may cost listeners more time. For a normal person, the concentration time of listening and rating sentences is around 20-30 minutes. If a single listening test is longer than 30 minutes, listeners may lose concentration and lead unreliable results. To avoid unnecessary cost as well as time and obtain convincing results, it's of interest to find a reliable objective measure to evaluate performances of algorithms instead. As shown in the appendix C, some objective measures are investigated and one algorithm named composite objective measure appears to be the best one. However, it showed unsatisfying result in our simulation, since it gave really small difference between different recovered sentences even they sounded quite different by ears. The reason behind is that the objective measures compared the whole recovered sentence with the original to get an average score. However, packet loss happened at certain time slots and impaired the speech quality. Scores from objective measures were averaged so that they indicated a small difference. For this, the objective measure could be applied to some specific frames with the packet loss, but scores cannot represent the overall quality of the sentence any more. In a word, an objective measure designed for evaluating performance of the packet loss concealment should be considered in the future.

APPENDIX A

Modelling Packet Loss in The Internet

Several models are made to model the packet loss in wireless networks in past years. The methods can be roughly categorized into two main classes: end-to-end packet loss models and link-layer specific packet loss models. Only the end-to-end models are taken into consideration because its performance is able to observed by applications, while the link-layer models have so complex loss pattern and sometimes are not available in practice [6].

Commonly use end-to-end models for packet loss are based on multi-state Markov models, or on complex hidden Markov models. Gilbert model is one of the straightforward and widely used statistical models for modelling the packet loss [10][16].

Packet loss in the Internet are, to a degree, temporally correlated [24], so that they usually happen with burst losses or alternative losses rather than random (i.e. Bernoulli) losses. As a result, the model used to generate the losses should accurately simulate losses in realistic network. The Gilbert model has been widely used in previous researches in order to generate temporal dependence of lost packets. To enhance the accuracy of modelling, the use of Markov chain model of k-th order can be applied with the expense of complexity [23][24].

As explained in the chapter 2, the short and long duration runs of both Good States (G_n) and Bad States (B_n) follow an exponential behaviour. [31] used random variables with a

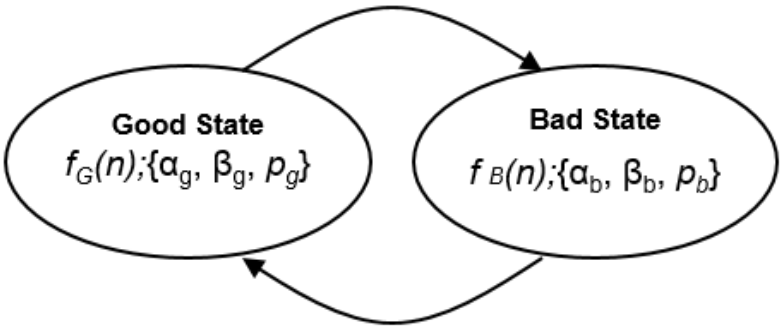


FIGURE A.1. Run length model for four-state Markov model

mixture of geometric distribution to approximate the exponential distributions,

$$f(n) = p(1-\alpha)\alpha^n + (1-p)(1-\beta)\beta^n \tag{A.1}$$

where α, β and p are parameters used to fit the desired run duration. One set of parameters $\{\alpha_b, \beta_b, p_b\}$ is used in bad run durations $f_B(n)$, and another one set $\{\alpha_g, \beta_g, p_g\}$ is for good run durations $f_G(n)$. Good State means the packet is successfully received, while the Bad State means the packet is lost. The model is illustrated in Figure A.1. The difference between two-state Gilbert model and four-state Markov model is that the latter one gives two different probabilities to both Good State and Bad State. Figure A.2 shows the four-state Markov model. The parameters used in this model to model the packet loss in wireless networks with different SNR values can be referred to the paper [31]. According to the experiments conducted in [31], the four-state Markov model provides a low-complexity channel model for precisely simulating the packet loss in wireless transmission.

In this thesis, the four-state Markov model was used to generate loss patterns in order to evaluate the performance of WSM with different packet size, template size and history buffer size. Additionally, it generated loss patterns with different packet loss to evaluate performances of different PLC algorithm (see chapter 4). This led to a more realistic and reliable evaluation.

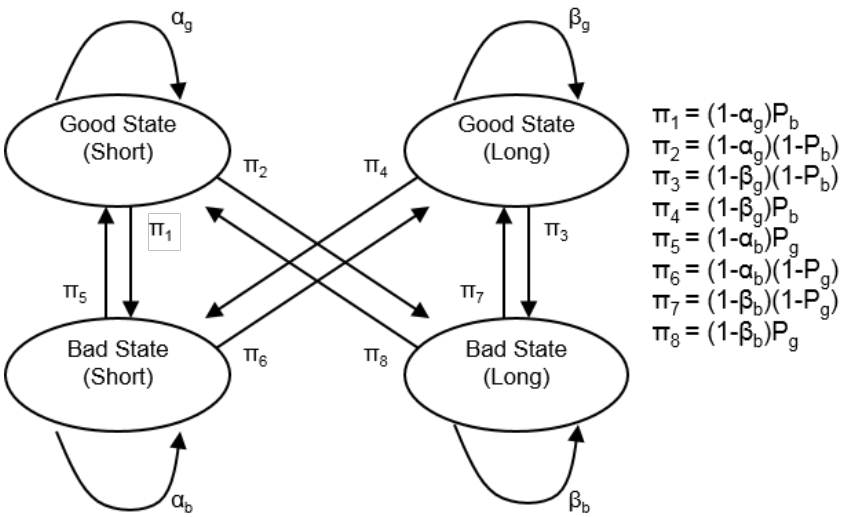


FIGURE A.2. Four-state Markov model

APPENDIX B

APPENDIX B

Packet Loss Concealment for Use with ITU-T Recommendation G.711

The American National Standard Institute (ANSI) presented a standardized PLC algorithm for use with the recommendation ITU-T G.711 [34]. This algorithm is based on linear prediction techniques and pitch waveform replication and fully implemented at the receiver side. The main part of the algorithm is presented in the Figure B.1.

The main features of this algorithm are listed below:

- The synthesis signal process is based on pitch detection of speech segments in the PLC buffer. The pitch detection algorithm is performed in the excitation domain using LP analysis filter. The synthesized excitation signal used to recovery the loss is created by going back one detected pitch period and copying the following segment.
- The reconstructed signal is derived by using LP synthesis filter to process the synthetic excitation signal.
- The speech signal attenuates with times of replication.
- Once a received packet arrives, the boundary between the synthesized speech signal and the received is smoothed by overlap-and-add technique.

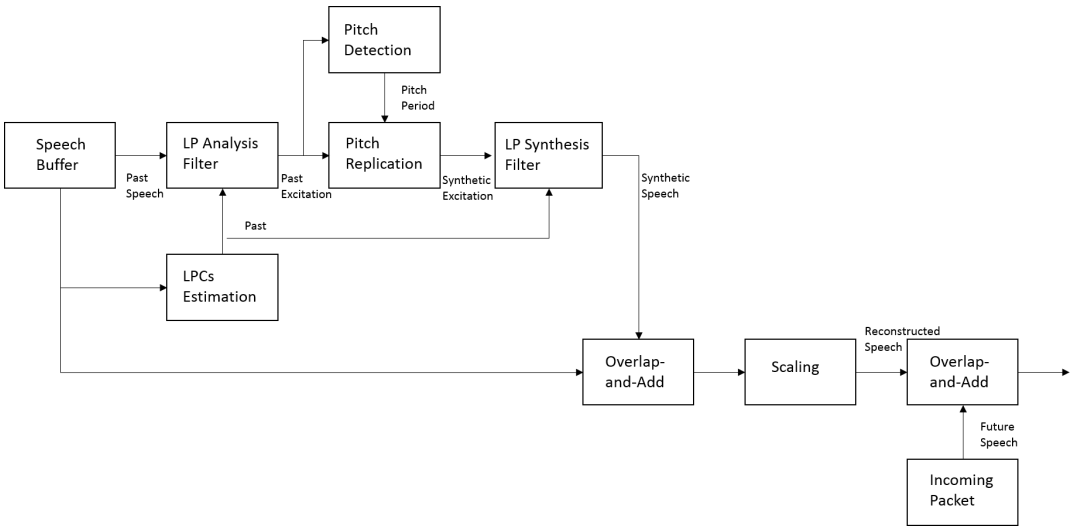


FIGURE B.1. Standardized PLC algorithm for use with the recommendation ITU-T G.711

The algorithm performs in two different domains. Pitch detection and pitch waveform replication are executed in the excitation domain, while the overlap-and-add technique is executed in the time domain, since the synthetic excitation signal is passed through the LP filter before the merging.

The LP coefficients are obtained by using Levinson-Durbin algorithm and stored to recover losses. The pitch period of preceding speech segments are estimated by computing the normalized autocorrelation function of the past excitation signal and then searching for the peak location. The pitch detector also gives voice activities to separate the voiced and unvoiced speech segments in order to apply different processing.

The reconstructed frame attenuates before it's played out to the speaker. The reconstruction stops until the gap is fully recovered by the replication. When the next packet is successfully received, the packet will be delayed and overlap-and-add with the last part of reconstructed speech signal. The coming signal will be scaled up before it's played out to the speaker.

APPENDIX C

APPENDIX C

Objective Measures

The most precise method for evaluating the speech quality is through subjective quality measures. However, subjective quality measure is always preformed under critical conditions (e.g., sizeable listener, objective impression from listeners, etc. [32]). For that reasons, an ideal objective quality measures which can predict the subjective quality speech is, to some degree, desirable [32]. Objective quality measures are based on physical measurements, such as acoustic pressure or its electrically converted level, or mathematically calculated values which come from the comparison between the original and the processed one. In the following part, several objective quality measures are listed.

C.1 Signal-to-Noise Ratio (SNR)

SNR measure is one of the widely used objective speech quality assessments. Basic SNR calculation is shown below and it's very easy to compute. However, SNR measure needs both the original speech signal and the processed speech signal, which are sometimes not available.

$$SNR = 10log_{10}\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{N}^{n=1}(x(n)-\hat{x}(n))^2}[dB] \tag{C.1}$$

where x(n) is the clean speech, $\hat{x}(n)$ is the processed speech, and N is the number of samples.

As we all known, speech is non-stationary through a long period, the classical SNR averages the SNR values over entire speech so that the average score does not correlate well with the speech quality. Thus, segmental SNR comes out and calculate the SNR value over very short frames, which are normally shorter than 30ms (speech is assumed to be stationary from 10ms to 30ms).

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1} log_{10}\frac{\sum_{n=Lm}^{Lm+L-1} x^2(n)}{\sum_{n=Lm}^{Lm+L-1}(x(n)-\hat{x}(n))^2}[dB] \tag{C.2}$$

where L is the length of frame, M is the number of frames in the entire speech (N=ML).

Segmental SNR performs better than classical SNR because the noisy parts stand-out and dominate the final speech quality assessment result, while the relative clean part can be, to a degree, ignored. However, the segmental SNR will be driven lower when the original speech contains silence, which impose large negative SNR values. To solve above problems, a range of SNR values (-10dB to 35dB) [15] can be set to alleviate the effect introduced by silence or apply voice activity detection (VAD) to separate the silence from speech signal.

Another advanced version of SNR measure called frequency-weighed SNR, which give segmental SNR weight factors proportional to the critical band. The $fwSNR_{seg}$ is defined as follows:

$$fwSNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=0}^{K-1} W(j,m) log_{10}\frac{X(j,m)^2}{(X(j,m)-\hat{X}(j,m))^2}}{\sum_{j=0}^{K-1} W(j,m)}[dB] \tag{C.3}$$

where W(j,m) is the weight on the jth subband in the mth frame, K is the number of subbands, X(j,m) is the spectrum magnitude of the jth subband in the mth frame, and $\hat{X}(j,m)$ is the spectrum magnitude of the jth subband in the mth frame.

Due to experiments, compared with SNR and SNR_{seg}, $fwSNR_{seg}$ shows the highest correlation with speech quality [19].

C.2 LPC Measures

Linear prediction can be used to model the speech production and linear prediction coefficients (LPC) can be derived from the model. The distance between the two sets of LPC, which come from clean speech and processed speech, respectively, could be used as an objective way to assess the speech quality. Of course, the LPC measures also need both clean speech and processed speech. As soon as LPC vectors are known, a couple of distance measures can be implemented, such as Log-Likelihood Ratio (LLR) measure [37], Itkura-Saito (IS) distance measure [33], Cepstrum Distance (CD) distance measure [26], etc. The equation (4),(5),(6) are used to described above three distances, respectively.

$$d_{LLR}(\mathbf{a}_p, \mathbf{a}_c) = log(\frac{\mathbf{a}_p \mathbf{R}_c \mathbf{a}_p^T}{\mathbf{a}_c \mathbf{R}_c \mathbf{a}_c^T}) \tag{C.4}$$

$$d_{IS}(\mathbf{a}_p, \mathbf{a}_c) = \left[\frac{\sigma_c^2}{\sigma_p^2}\right]\left[\frac{\mathbf{a}_p \mathbf{R}_c \mathbf{a}_p^T}{\mathbf{a}_c \mathbf{R}_c \mathbf{a}_c^T}\right] + log(\frac{\sigma_c^2}{\sigma_p^2}) - 1 \tag{C.5}$$

$$d_{CP}(c_p, c_c) = \frac{10}{log10}\sqrt{2\sum_{k=1}^{P}(c_c(k) - c_d(k))^2} \tag{C.6}$$

where, $\mathbf{a}_c$ is the LPC vector of clean speech, while $\mathbf{a}_p$ is the LPC vector of processed speech. $\mathbf{R}_c$ is the auto-correlation matrix of the clean speech, σ_c^2 and σ_p^2 are the all-pole gains of clean speech and processed speech, c_c and c_p are the Cepstrum vectors of clean speech and processed speech, and P is the order.

C.3 Spectral Distance Measures

Literately, spectral distance is calculated from the difference between clean speech spectrum and processed speech spectrum. Weighted spectral slop (WSS) measure is one kind of direct spectral distance measures. WSS uses the spectral slopes (spectral difference between the adjacent neighbouring bands) instead of the spectrum itself. the implementation of WSS can be described in the equation below:

$$d_{WSS} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)(S_c(j,m) - S_d(j,m))^2}{\sum_{j=1}^{K} W(j,m)} \tag{C.7}$$

where k is th number of bands, M is the number of frames , S(j,m) is the spectral slope of the jth band in the mth frame. W is the weigh factor, which can be found in [27].

C.4 Articulation Index (AI)

Articulation Index (AI) was first presented in [8] and used to evaluate the speech intelligibility. AI was renamed Speech intelligibility Index (SII) [17]. In the AI, the distortions are supposed uncorrelated in each critical frequency band and assumed to be additive noise or signal attenuation, etc. AI value is derived by computing SNR for each frequency band, and averaging them as follows:

$$AI = \frac{1}{N}\sum_{j=1}^{N}\frac{min(SNR(j),30)}{30} \tag{C.8}$$

where SNR(j) is the SNR in jth band, the number of subbands is set to be N, and the maximum subband SNR is set to be 30dB.

However, in many cases, the distortions in adjacent bands are convolutional or not completely uncorrelated, which degrades the performance of the AI measure.

C.5 Speech Transmission Index (STI)

Speech Transmission Index (STI) [18] is a widely used measure for evaluating the speech intelligibility. STI uses the modulation transfer function instead of SNR. STI assumes that the loss of intelligibility is relative with the loss in the modulation depth. By measuring the loss in modulation depth at the receiver, the intelligibility loss can be obtained. STI shows better performance than AI because it also considers the flattening of the information-carrying speech envelopes due to the reverberation.

C.6 PESQ

The Perceptual Evaluation of Speech Quality (PESQ) [36] is used for predicting the Mean Opinion Score (MOS) from both clean speech and processed speech. PESQ is commonly regraded as one of the most accurate and complicated methods to estimate the MOS today. PESQ uses a perceptual mode to convert and time-align both speeches

into internal representations. Then, a couple of parameters related to speech are applied to compare internal representations and put into the cognitive model to give the MOS.

A successor of PESQ is called Perceptual Objective Listening Quality Analysis (POLQA), which offers an advanced level of benchmarking accuracy and adds new capabilities for wideband and super-wideband (HD) speech signals. Compared with PESQ, POLQA maintains correct scoring also at high background noise levels and alleviates the effects introduced by different speech level in samples. Due to the relative measurements, POLQA gives more reliable and convincing results than PESQ does.

C.7 Composite Measures

By combining multiple objective measures, one can form the so-called composite measures. The motivations behind the use of composite measures is that different objective measures evaluate different characteristic of the speech signal, so combining them in a linear or non-linear way can produce significant gains in correlations. The overall score is obtained by linear combination of PESQ measure, log-likelihood ratio (LLR) measure and weighted spectral slop (WSS) measure. The equation is shown below:

$$C_{ovl} = 1.594 + 0.805 * pesq - 0.512 * llr_{mean} - 0.007 * wss_{dist} \tag{C.9}$$

The score varies from 1 to 5, which yields the same result with Mean Opinion Score (MOS). 5 means Excellent, while 1 means Bad.

In this thesis, we use composite measures to evaluate the performance of WSM (see chapter 3).

BIBLIOGRAPHY

[1] N. AOKI, *Voip packet loss concealment based on two-side pitch waveform replication technique using steganography*, in TENCON 2004. 2004 IEEE Region 10 Conference, vol. 100, IEEE, 2004, pp. 52–55.

[2] B. ATAL AND L. RABINER, *A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 24 (1976), pp. 201–212.

[3] C. G. BABU AND P. VANATHI, *Performance analysis of voice activity detection algorithms for robust speech recognition*, International Journal of Computing Science and Communication Technologies, 2 (2009), pp. 288–293.

[4] D. BOWKER AND C. DVORAK, *Speech transmission quality of wideband packet technology*, in IEEE GLOBECOM‚Äô87, 1987, pp. 1887–1889.

[5] S. CASNER AND S. DEERING, *First ietf internet audiocast*, ACM SIGCOMM Computer Communication Review, 22 (1992), pp. 92–97.

[6] M. ELLIS, D. P. PEZAROS, T. KYPRAIOS, AND C. PERKINS, *A two-level markov model for packet loss in udp/ip-based real-time video applications targeting residential users*, Computer Networks, 70 (2014), pp. 384–399.

[7] D. ENQING, L. GUIZHONG, Z. YATONG, AND C. YU, *Voice activity detection based on short-time energy and noise spectrum adaptation*, in Signal Processing, 2002 6th International Conference on, vol. 1, IEEE, 2002, pp. 464–467.

[8] N. R. FRENCH AND J. C. STEINBERG, *Factors governing the intelligibility of speech sounds*, The journal of the Acoustical society of America, 19 (1947), pp. 90–119.

[9] T. R. GAULT, *A voice operated musical instrument*, University of Louisville, 2007.

[10] E. N. GILBERT, *Capacity of a burst-noise channel*, Bell system technical journal, 39 (1960), pp. 1253–1265.

[11] D. GOODMAN, G. LOCKHART, O. WASEM, AND W.-C. WONG, *Waveform substitution techniques for recovering missing speech segments in packet voice communications*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 34 (1986), pp. 1440–1448.

[12] J. GRUBER AND N. LE, *Performance requirements for integrated voice/data networks*, IEEE Journal on Selected Areas in Communications, 1 (1983), pp. 981–1005.

[13] J. GRUBER AND L. STRAWCZYNSKI, *Subjective effects of variable delay and speech clipping in dynamically managed voice systems*, IEEE Transactions on Communications, 33 (1985), pp. 801–808.

[14] M. HANDLEY, J. CROWCROFT, C. BORMANN, AND J. OTT, *The internet multimedia conferencing architecture*, tech. rep., Internet Draft, Internet Engineering Task Force, 1997.

[15] J. H. HANSEN AND B. L. PELLOM, *An effective quality evaluation protocol for speech enhancement algorithms.*, in ICSLP, vol. 7, 1998, pp. 2819–2822.

[16] O. HOHLFELD, *Statistical error model to impair an H. 264 decoder*, PhD thesis, Technische Universität Darmstadt, 2008.

[17] B. W. HORNSBY, *The speech intelligibility index: What is it and what's it good for?*, The Hearing Journal, 57 (2004), pp. 10–17.

[18] T. HOUTGAST AND H. J. STEENEKEN, *Evaluation of speech transmission channels by using artificial signals*, Acta Acustica united with Acustica, 25 (1971), pp. 355–367.

[19] Y. HU AND P. C. LOIZOU, *Evaluation of objective quality measures for speech enhancement*, IEEE Transactions on audio, speech, and language processing, 16 (2008), pp. 229–238.

[20] M. JABER, *Voice activity detection method and apparatus for voiced/unvoiced decision and pitch estimation in a noisy speech feature extraction*, Feb. 7 2007. US Patent App. 11/672,106.

[21] N. JAYANT AND S. CHRISTENSEN, *Effects of packet losses in waveform coded speech and improvements due to an odd-even sample-interpolation procedure*, IEEE Transactions on Communications, 29 (1981), pp. 101–109.

[22] N. S. JAYANT AND P. NOLL, *Digital coding of waveforms: principles and applications to speech and video*, Englewood Cliffs, NJ, (1984), pp. 115–251.

[23] W. JIANG AND H. SCHULZRINNE, *Modeling of packet loss and delay and their effect on real-time multimedia service quality*, in PROCEEDINGS OF NOSSDAV'2000, Citeseer, 2000.

[24] ——, *Perceived quality of packet audio under bursty losses*, in IEEE INFOCOM, Citeseer, 2002, p. 1.

[25] S. KADAMBE AND G. F. BOUDREAUX-BARTELS, *Application of the wavelet transform for pitch detection of speech signals*, IEEE transactions on Information Theory, 38 (1992), pp. 917–924.

[26] N. KITAWAKI, H. NAGABUCHI, AND K. ITOH, *Objective quality evaluation for low-bit-rate speech coding systems*, IEEE Journal on Selected Areas in Communications, 6 (1988), pp. 242–248.

[27] D. KLATT, *Prediction of perceived phonetic distance from critical-band spectra: A first step*, in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82., vol. 7, IEEE, 1982, pp. 1278–1281.

[28] M. LEE, I. ZITOUNI, AND Q. ZHOU, *Prediction-based packet loss concealment for voice over ip: a statistical n-gram approach*, in Global Telecommunications Conference, 2004. GLOBECOM'04. IEEE, vol. 4, IEEE, 2004, pp. 2308–2312.

[29] B. M. LEINER, V. G. CERF, D. D. CLARK, R. E. KAHN, L. KLEINROCK, D. C. LYNCH, J. POSTEL, L. G. ROBERTS, AND S. WOLFF, *A brief history of the internet*, ACM SIGCOMM Computer Communication Review, 39 (2009), pp. 22–31.

[30] J. A. MARKS, *Real time speech classification and pitch detection*, in Communications and Signal Processing, 1988. Proceedings., COMSIG 88. Southern African Conference on, IEEE, 1988, pp. 1–6.

[31] J. MCDOUGALL, J. J. JOHN, Y. YU, AND S. L. MILLER, *An improved channel model for mobile and ad-hoc network simulations.*, in Communications, Internet, and Information Technology, 2004, pp. 352–357.

[32] S. QUACKENBUSH, T. BARNWELL, AND M. CLEMENTS, *Objective measures of speech qualityprentice-hall*, Englewood Cliffs, NJ, (1988).

[33] L. R. RABINER AND B.-H. JUANG, *Fundamentals of speech recognition*, (1993).

[34] A. RECOMMENDATION, *T1. 521a-2000 (annex b),„Äú*, Packet loss concealment for use with ITU-T recommendation G, 711 (2000).

[35] I. RECOMMENDATION, *1534-1: Method for the subjective assessment of intermediate quality level of coding systems*, International Telecommunication Union, (2003).

[36] A. RIX, J. BEERENDS, M. HOLLIER, AND A. HEKSTRA, *Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU-T Recommendation, 862 (2001).

[37] M. SAMBUR AND N. JAYANT, *Lpc analysis/synthesis from speech inputs containing quantizing noise or additive white noise*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 24 (1976), pp. 488–494.

[38] H. SANNECK, A. STENGER, K. B. YOUNES, AND B. GIROD, *A new technique for audio packet loss concealment*, in Global Telecommunications Conference, 1996. GLOBECOM'96.'Communications: The Key to Global Prosperity, IEEE, 1996, pp. 48–52.

[39] H. SCHULZRINNE, *Rtp profile for audio and video conferences with minimal control*, (2003).

[40] C. SEMERIA AND T. MAUFER, *Introduction to ip multicast routing*, 1998.

[41] R. VALENZUELA AND C. ANIMALU, *A new voice-packet reconstruction technique*, in Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on, IEEE, 1989, pp. 1334–1336.

[42] W. VERHELST AND M. ROELANDS, *An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech*, in Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on, vol. 2, IEEE, 1993, pp. 554–557.

[43] O. J. WASEM, D. J. GOODMAN, C. A. DVORAK, AND H. G. PAGE, *The effect of waveform substitution on the quality of pcm packet communications*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 36 (1988), pp. 342–348.

[44] C. WEINSTEIN AND J. FORGIE, *Experience with speech communication in packet networks*, IEEE Journal on Selected Areas in Communications, 1 (1983), pp. 963–980.