The Voice Activity Detection (VAD) Recorder and VAD Network Recorder

A thesis presented in partial fulfilment of the requirements

for the degree of

Master of Science in Computer Science at Massey University

Feng Liu

2001

Acknowledgment

First I would like to thank Professor Chris Jesshope, my supervisor, for introducing me to the field of VoIP and computer telephony, for his guidance and comments during the evolution of this project, and especially for his patience on my English writing.

Second I would like to thank my father, HangZhang Liu, for his cultivation and encouragement.

I also would like to thank my colleagues, YongQiu Liu, and Phoebe Wang, for their helps.

Last I would like to thank my wife, Liping Cai. Without her continuous support, this research can never be completed.

Feng Liu, Bs (Bs)

Master of Science (Computer Science) candidate,
Massey University,
Computer Science,
Institute of Information Science and Technology,
Massey Unviersity,
Palmerston North,
New Zealand.

Abstract

The project is to provide a feasibility study for the AudioGraph tool, focusing on two application areas: the VAD (voice activity detector) recorder and the VAD network recorder. The first one achieves a low bit-rate speech recording on the fly, using a GSM compression coder with a simple VAD algorithm; and the second one provides two-way speech over IP, fulfilling echo cancellation with a simplex channel. The latter is required for implementing a synchronous AudioGraph. In the first chapter we introduce the background of this project, specifically, the VoIP technology, the AudioGraph tool, and the VAD algorithms. We also discuss the problems set for this project. The second chapter presents all the relevant techniques in detail, including sound representation, speech-coding schemes, sound file formats, PowerPlant and Macintosh programming issues, and the simple VAD algorithm we have developed. The third chapter discusses the implementation issues, including the systems' objective, architecture, the problems encountered and solutions used. The fourth chapter illustrates the results of the two applications. The user documentations for the applications are given, and after that, we analyse the parameters based on the results. We also present the default settings of the parameters, which could be used in the AudioGraph system. The last chapter provides conclusions and future work.

Table of Contents

List of Acronyms

ABS	:	Analysis-by-Synthesis
ADPCM	:	Adaptive differential PCM
ATM	:	Asynchronous Transfer Mode network
CCS7	:	Common Channel Signalling System number 7
CS-ACELP	:	Conjugate-Structure Algebraic Code Excited Linear Prediction
DSP	:	Digital Signal Processor
ESTI	:	European Telecommunications Standards Institute
ET	:	Energy Threshold
FEC	:	Forward Error Correction
FIFO	:	First In First Out
GAET	:	Geometrically Adaptive Energy Threshold
GSM	:	Global System for Mobile Communication
IP	:	Internet Protocol
ISO	:	Internetworking Operating System
ITU	:	International Telecommunication Union
LCD	:	Liquid Crystal Display
LCR	:	Least-Cost-Routing
LDAP	:	Lightweight Directory Access Protocol
LPC	:	Linear Prediction Coding
LSPE	:	Least-square Periodicity Estimator
MGCP	:	Media Gateway Control Protocol
MPLPC	:	Multi-pulse LPC
QoS	:	Quality of Service
PBX	:	Private Branch Exchange
PCM	:	Pulse Code Modulation
PSTN	:	Public Switched Telephone Network
RELP	:	Residual excited linear predictive coding
RPE-LTP	:	Regular pulse excitation long-term predictor
RSVP	:	Resource Reservation Protocol
RTP	:	Real-time Transport Protocol
SGCP	:	Simple Gateway Control Protocol
SIP	:	Session Initiation Protocol
SNMP	:	Simple Network Management Protocol
SNR	:	Signal-to-Noise Ratio
SS7	:	Signalling System number 7
TCP	:	Transport Control Protocol
UDP	:	User Datagram Protocol
VAD	:	Voice Activity Detection
VoIP	:	Voice over IP
WRED	:	Weighted Random Early Detection
WFQ	:	Weighted Fair Queuing
ZCR	:	Zero Crossing Rate

List of Figures

List of Tables

Chapter 1 Introduction

This project applies a number of speech capture and compression techniques to two application areas. The first is automatic voice capture for and application to record and subsequently playback, on demand and on the web, multimedia teaching material. The second is the ability to send voice over Internet again for educational purposes.

1.1 Background

1.1.1 VoIP

Voice over IP (VoIP) has become one of the fastest-growing technologies in telecommunications since 1995. It challenges the traditional technology of telephony, PSTN, which is the Public Switched Telephone Network and shows the trend towards substituting the PSTN in the future. As a result of the huge market demands, most of all computer vendors have dived into this "golden river" to develop leading edge products, and to provide services that we might have never heard of before, such as offering free toll call to your country from New Zealand. The sparkle behind this fact is that the Internet has been changed to support voice traffic, even though there are some issues of this technology that need to be resolved. But, eventually, the Internet and telephone network will be merged as one and the same [4].

After five years development, there are many of VoIP products around the world. But there is a common problem, which is how the VoIP system can provide the same quality of service (QoS) as PSTNs when the voice and data networks are combined as one. We have not got a perfect solution thus far, although people have been spending a lot of time and money into this research. In the following section, we will explore some of the

techniques used in this technology. We will outline the basic architecture of the VoIP, what advantages it has, and what kinds of issue have arisen when VoIP tries to satisfy the integration of voice and data networks.

Major components of VoIP system

Different vendors of VoIP technology can have their own variations of the overall VoIP network architecture and algorithms, but the backbone of the functionality should be the same for all of them.

VoIP Gateways – VoIP Gateways are a bridge between the local PSTN and the IP endpoint, performing such tasks as preparing data from analog to digital for the network, decompressing digital data back to analog signal when data is received from the network, etc [2]. Some of the VoIP Gateways might do a little bit more than those described above, such as connecting to the destination gateways, having the capability of demodulation and remodulation [4], etc.

VoIP Gatekeepers – VoIP Gatekeepers are used to manage all active clients within a network, used to provide real-time communication. The VoIP Gatekeepers' functionality can be implemented in two ways. One is that it is distributed among all VoIP Gateways and the other is that it is centralised at one or more locations. "When gatekeeper functions are embedded in each gateway, all gateways of the overall VoIP network act autonomously to coordinate their actions. With a centralised gatekeeper, all gateways of the network coordinate their actions with respect to the centralised gatekeeper rather than acting independently" [2].

A gatekeeper is required to perform the following functions [4][5]:

- Address Translation: Alias-address to transport-address translation must be provided.
- Admissions Control: Access to LAN is based on call authorisation, bandwidth or some other criteria.
- Bandwidth Control: Terminals send requests for network bandwidth to the gatekeeper.
- Zone Management: A gatekeeper is required to provide address translation, admissions control, and bandwidth control to all endpoint that have registered with it.

In VoIP it is needless to say, that, the analog voice signals are digitised, compressed and transmitted as a stream of packets over a digital data network, and it is the backbone of the technology. Some vendors may provide support for dynamic bandwidth allocation, packet loss recovery, adaptive echo cancellation, and speech processing to deliver voice quality as high as possible, etc.

In reality, VoIP can be implemented in many forms. The following five forms have covered most of products involved VoIP technology thus far:

- PC to PC service model
- PC to phone service model
- Phone to phone service model
- Network service model
- Service to service provider model

Benefits of VoIP technology

The reason why VoIP has become one of the hottest fields of research and development is that VoIP technology provides more significant benefits than those relevant technologies being widely applied today, such as PSTN. This can be summarised as four aspects described below:

- Cost reduction: The Public Switched Telephone Networks' toll services can be bypassed using the Internet backbone [4], which means slashing prices of the cost of data communication, such as the long distance calls and long distance fax.

- Integration of Voice and Data: The integration of voice and data traffic will be demanded by multi application software such as international telephone service provider or multimedia real-time education system, etc.

- Simplification: An integrated infrastructure, which covers several mature infrastructures and which supports all forms of communication, allows more standardization and simplifies equipment management [4]. The result is a fault tolerant design, and finally, the use of the same telephone line for voice and data will become a realization although it will take time.

- Bandwidth Saving and Network Efficiency: Various good compression solutions reduce the bandwidth demand, and data packetized over IP also releases the resource while voice is not being produced during the real-time communication, e.g. one is listening and not talking. In other words, reducing bandwidth increases network efficiency. Typically, some 50% of a conversation is silence. This provides a huge opportunity to remove the redundancy like silence in certain speech patterns so as to reduce the bandwidth. Therefore, the network efficiency can be increased significantly.

If the benefits described above are so attractive, why do we not directly move forward to fully adopt this technology? The answer is very straightforward -that is to say- there are many issues that need to be resolved, which will have a great impact on the performance of systems based on VoIP technology. This must be solved before we can move any further.

Issues

VoIP technologies are based on the infrastructure of the Internet network, but the original design of Internet network did not support real-time operations. Obviously, there is already a big puddle, which needs to be buried, so that we can walk through the Internet safely.

The main criterion for a successful real-time voice application is that the application can provide at least the same quality of voice when compared to results produced by PSTNs. So seven issues that may affect the result of quality of voice require considerable attention, and will be discussed respectively below:

The first issue is delay. Delay will be divided into two categories according to its predicability. Category one, including processing delay, buffering delay, delay of analog data converted to digital, delay of queuing for being sent or received, delay of digital data converted back to analog for playback, delay jitter and playback buffering, can be minimised under some good algorithms. Category two includes the delay of data across the IP network and private

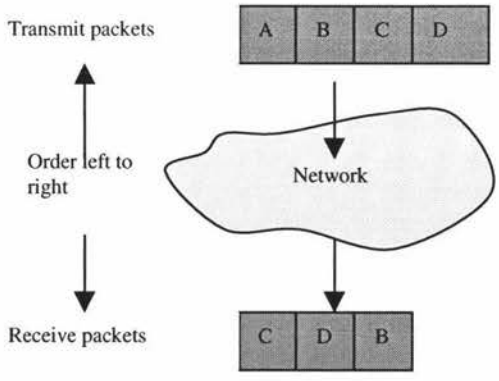

Figure 1.1.1-1 Delay jitter

networks and may not be predicable.

Within category one, most delays are caused by very straightforward reasons, so here we will not consider these easily understood delays. But delay jitter is more complicated and it is worth saying something about it.

Jitter delay is where the data across the network might not arrive in the same order as it was sent. A network does not act like a FIFO stack, where the element that comes in first will be the one that goes out first, so packets across the network will arrive in any order; In the worst case, some of the packets might be lost. Figure 1.1.1-1 shows the situation, where the packet A has been lost.

This kind of delay is really hard to handle for real-time applications. A common solution from most vendors is to use a playback buffer (adaptive jitter buffer) to tolerant this kind of delay. This is achieved by providing extra information in each packet such as time-stamping information, sequencing information, etc. At some stage if a packet such as the packet A in Figure 1.1.1-1 does not arrive we discard it and continue. This is based on the assumption that all packets are very small, so that discarding a packet would not cause serious problem during the speech reconstruction. But the solution also increases the sum of delay because of the buffering technique.

When the sum of the overall delay is greater than 250ms, real-time applications do not work quite so well and are unable to produce the same quality of sound as PSTNs. In order to remove stepping from the voice, which comes from the other end, people must avoid speaking at the same time at both ends. That is the reason why most IP phone vendors choose a protocol, which allows people to either speak or to listen at any one time.

As for category two, the delay comes from the network, such as that caused by congestion. There are some ways to reduce this kind of delay in order to achieve toll-

quality voice effects. Firstly, using an IP packet segmentation technique we can remove the delay caused by very large data packets. Secondly, prioritizing IP packets allows the network to maintain the highest voice quality over a congested network. Thirdly, using a Digital Signal Processor (DSP) architecture we can achieve high performance, especially when we need to apply some sophisticate algorithms, such as Weighted Random Early Detection (WRED), to reduce the delay caused by a congested network.

The second issue is the compression technique. As is well known, codec stands for encoder and decoder. The codec chosen for VoIP applications must provide good quality of voice and require as small a bit rate for transmission as possible.

To date, Pulse Code Modulation (PCM), which is a waveform coding technique based on a three-step process: sampling, quantisation, and coding, (at desired sample rate) is the most commonly used codec on a worldwide basic. Under PCM, an analog signal is sampled at say 8000 times per second. For 8 bit samples, this requires a transmission rate of 8000 samples per second $\times$ 8 bits/sample, or 64 Kbps. Obviously the PCM format does provide good quality of voice. However, it requires a large transmission rate. This means that the delay associated with transmission will be quite significant especially in the time sensitive case. Therefor we must choose another codec, which requires a smaller bit rate with transmission, to reduce the delay while keeping the quality of the voice as good as the quality provided by PCM.

The International Telecommunication Union's (ITU) officially recommended codec for all wide area networking applications is G.729, Conjugate-Structure Algebraic Code Excited Linear Prediction (CS-ACELP) (described in section 2.2) [3][7][8]. This codec uses a sample rate of 8000 samples per second. It also uses a 10-ms frame size plus a 5-ms look-ahead. To analyze the data properly it is sometimes necessary to analyze data beyond the frame boundary, and this is referred to as look-ahead. This results in a total of a 15-ms algorithmic delay (a delay results from buffering a frame's worth of data to analyze the speech), and provides near-toll quality. It only requires approximately 8 Kbps transmission rate. Recently some vendors have added a proprietary silence suppression

capability to the G.729 coding mechanism that reduces the demand of bandwidth required in a conversation down to 4 Kbps.

The third issue is echo cancellation. "In a traditional telephony network, echo is normally caused by a mismatch in impedance from the four-wire network switch conversion to the two-wire local loop and controlled by echo cancellers" [6]. But in voice-packet based applications such as multimedia applications, the playback of the voice received through the network might have a chance to be recorded as input again. If so, the background noise would be rapidly increased, and echo cancellation becomes a critical issue in that kind of application. Common solutions for echo cancellation are to use echo cancellers or to use protocols to ensure echo suppression. The Echo cancellers are built into the low bit-rate codecs and are operated on each DSP. The protocols for echo cancellation are widely used on commercial products nowadays such as IP phones.

The fourth issue is VoIP forward error correction. Most of the VoIP applications choose UDP as its IP protocol in order to satisfy the real-time demands. The disadvantage of using UDP/IP is that there is no guarantee that the destination end will receive all data sent by the source end. So data corruption and loss while transmitting through the network will need to be compensated for in order to obtain good quality voice. VoIP forward error correction (FEC) does this for us. Usually there are two kinds of FEC, one is Intra Packet FEC and the other is Extra Packet FEC.

- With Intra Packet FEC, extra bits are added into the packet so that the receiving end can determine whether the data received is correct at playback.

- With Extra Packet FEC, extra information is added to each packet that allows the receiving gateway to extrapolate from the previously received good packet and to reconstruct the missing or severely corrupted packet.

The fifth issue is bandwidth consumption. As is well known, receiving or sending data through a modem is the bottleneck of the Internet infrastructure. To reduce bandwidth demand so as to maximise the use of modem is one of our goals. Although choosing good codecs can significantly reduce the bandwidth requirement, silence suppression also can achieve that goal. Removing silence from a speech or even from words can reduce the bandwidth demand in order to get the maximum performance of the modem. The most commonly used technique to suppress silence within a conversation is called voice activity detection (VAD), which will be discussed later in this report. After having used VAD, when the sound is reconstructed, there might be a need to smooth the sound in order to make it sound as natural as it was. So comfort noise generation (CNG) is also required.

The sixth issue is IP protocol. The UDP/IP protocol is widely used in time-sensitive products, because the TCP/IP protocol will always retransmit the corrupted packets and this causes additional delays. Most of the VoIP products are real-time applications, which could not stand the delays caused by retransmitting. But in some applications, which are not time-sensitive, the TCP/IP protocol would be the better alternative, such as a Fax application [4]. However, UDP is only a best-effort protocol so that it does not provide reliable service. This could results in a situation that when a person is speaking, you cannot hear his/her speech in its original order. So another application layer level protocol such as RTP (discuss in section 2.3.2) could be employed. The RTP protocol provides mechanisms that can detect the loss of packets, and provide sufficient information for reconstructing the speech in its original order.

The final issue is security. To reduce cost and provide higher productivity, VoIP is widely integrated with the Internet. Therefore, security issue becomes critical. The issue involves access control, authentication, and encryption with respect to transmission over a public packet network [8].

Summary

VoIP is still in the relatively early stage of deployment, but the economics are compelling because of its significant bandwidth efficiencies over traditional PSTN. Unfortunately the toll quality provided by this technology depends on the implementation. Clearly, there is a need to standardise this technology. No matter when the standard shows up, VoIP technology is becoming one the fastest-developed technologies towards the rapid transition of all networks based on the digital/packet-based architecture.