Space-Based Voice over IP Networks [1,2]

Sam Nguyen, Clayton Okino, William Walsh, Loren Clare
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109
{Sam.P.Nguyen, Clayton.M.Okino, Loren.P.Clare }@jpl.nasa.gov

Abstract— In human space exploration missions (e.g. a return to the Moon and for future missions to Mars), there will be a need to provide voice communications services. In this work we focus on the performance of Voice over IP (VoIP) techniques applied to space networks, where long range latencies, simplex links, and significant bit error rates occur. Link layer and network layer overhead issues are examined. Finally, we provide some discussion on issues related to voice conferencing in the space network environment.

TABLE OF CONTENTS

1. INTRODUCTION

Communication is vital for any manned mission to the outer planet (e.g. a return to the moon and for future mission to mars). The functional needs for the voice range from highly robust emergency communications to operation voice requiring excellent intelligibility for flight safety seasons and excellent quality for public involvement. Currently, many system used M-law compressing and expanding scheme, because of its excellent intelligibility. However, M-law codec is very sensitive to error. In other words, when data is corrupted due to transmission, the quality of the speech is greatly diminished. Furthermore, M-law codec requires bandwidth at least 64Kbps, which is not feasible in emergency situation because we want to have the lowest possible useable data rate in those situations, but at the same time produces a robust voice.

Fortunately, within the past decade, many audio compression and decompression scheme had been developed that reduce bit rate, increase error resistant, and at the same time produce a speech intelligibility that is similar to the M-Law compression/decompression scheme. So by exploring, many new features that are present in a newer codec such as error concealment, Voice Activate Detection, variable bit rate, etc, we can design a more optimal system that can tolerate higher bit error and higher percent of packet drops, but produces more or at least the same robust level as M-law codec. Therefore, in this study, first we want to see how well various voice compression/decompression perform under the impacted of long range latencies, simplex links, and corrupted bit errors, these are the effects that are expected in wireless/RF environments. Then, we want to study what is the optimal frame size we can packet the data that minimize the effects of the overhead, but maximize the speech intelligibility. Last, we want to see whether we are better off dropping the packet or keeping the packet when it contains at least an error in the payload for all of the codecs. By understand all of these, we are able to make a decision whether to use the dynamic codec switching or find the best codec that can do the job. That way, our new system takes the advantage of the newer feature in order to optimize its performance.

2. TYPE OF CODECS

There are many audio codecs available today and each of the codec has it own unique characteristic. By exploring and understanding those characteristics, we can optimize their usefulness. However, the criteria that we are specifically looking for in the space application are: 1) resistant to bit error; 2) small bit rate (the smaller the better); 3) excellent intelligibility; and 4) complexity. With all these requirements, we narrow our study to 9 codecs, which are G711, G729, G726, iLBC, G723.1, AMR-GSM, G728, CVSD, and MELP.

- G.711 was developed in 1972 and it's an ITU-T standard, which can be downloaded through www.itu.com. G.711 comes with two main algorithms defined in the

[1] 1-4244-0525-4/07/$20.00 ©2007 IEEE.
[2] IEEEAC paper #1526, Version 1, Updated October 20, 2006

standard: mu-law algorithm (mainly used in North America and Japan) and a-law algorithm (used in Europe and the rest of the world). Both takes in 16, 14, or 13 bits data sampled at 8KH and compresses into an 8 bits data stream. Thus it produces a 64K bits/second bitstream. Since the current system uses G.711 we let G.711 be the benchmark of our study and see how well other codecs can outperform or underperform it.

• G729 is an audio data compression algorithm for voice that compresses voice audio in "chunks" of 10 milliseconds. Standard G729 operates at 8Kbs, which we used for this study, however there are extensions, which provide also 6.4kbps and 11.8 kbps rates for marginally worse and better speech quality respectively. Music or tones such as Dual-tone Multi-Frequency (DTMF) or fax tones cannot be transported reliably with this codec. [17]

• G.726 is an ITU-T standard that uses the Adaptive Differential Pulse Code Modulation (ADPCM) scheme. The bit rates it covers are 16, 24, 32, and 40 kbps. This codec initially was introduced to supersede G.711 with a 32 kbps rate. The four bits rates associated with G726 are often referred to by the bit size of a sample, which are 2-bits, 3-bits, 4-bits, and 5-bits respectively. In the study, however, we used 4-bits, which is 32 kbps. [7]

• iLBC (Internet Low Bit Rate) is developed by Global IP Sound (GIPS). It is designed for narrow band speech and results in a payload bit rate of 13.33kbps for 30 ms frames and 15.2kbp for 20 ms frames. Both are used in this study. [16]

• G.723.1 is a linear predictive analysis by synthesis coding. The excitation signal for the high rate coder is multipulse Maximum Likelihood Quantization (MP-MLQ) and for the low rate coder is Algebraic Code-Excited Linear Prediction (ACELP). This codec comes with two bit rate 6.3 and 5.6 Kbps. Both rates are used this study. [9]

• AMR-GSM-EFR – An ETSI standard, Adaptive Multi-Rate Global System for Mobile Communications Enhanced Full-Rate is used for digital cellular communications. Based on the codec-excited linear predictive (CELP) coding model, this codec has eight basic bit rates 12.2kbps, 10.2kbps, 7.95kbps, 7.4kbps, 6.7kbps, 5.9kbps, 5.15kbps, and 4.75kbps. 12.2 Kbps was used in this study. [5]

• G728 is a ITU-T standard for speech coding operating at 16kbps. It used LD-CELP (Low Delay Code Excited Liner Prediction). Delay of the codec is only 5 samples. The linear prediction is calculated backwards with a 50th order LPC filter. The excitation is generated with gain scaled VQ (vector quantization). The results were poor because of unequal bit protection where some bits are extremely important such that decoding can not be performed from that point forward if one of the key bits is corrupted. [18]

• CVSD (Continuously Variable Slope Delta Modulation), like Adaptive Differential Pulse Code Modulation (ADPCM), generates a code difference between the current input sample and a predicted value from past output. The general algorithm process is also similar to ADPCM, except that the input to the algorithm is analog. Since the actual CODECs were not obtainable, in order to emulate the process of the CVSD curves for the generation of the PESQ curve, the Sound Exchange (sox) Unix program was used to map Wav files to CVSD which were then corrupted with bit errors, CVSD to Wav remapped and then evaluated for PESQ performance. [3]

• MELP (Mixed-Excitation Linear Predictive) is the new 2.4kbps Federal Standard speech coder. MELP is robust in difficult background noise environments such as those frequently encountered in commercial and military communication systems. For each, 22.5 ms frame of input speech, it produces 54 bits (54bits/22.5ms = 2.4kbps).

3. CODEC COMPLEXITY

There is no consensus on how we can measure the complexity of the codecs without delving into the code and trying to calculate how many adding and multiplying operations there are in each codec. The reason for this is that many codecs were invented in the late 1970's and it was a common practice back then to use mips (million instructions per second, also commonly referred to as meaningless indicator of performance) to calculate the complexity. However, because different instructions require more or less time than the others, there is no standard way of measuring mips. Therefore, mips is not a good indication of performance. Nevertheless, we still want to know the complexity between codecs so instead of looking at them individually, we want to know how they are compare to each other. Therefore, we ran all the codecs on two different machines: Duo Core Pentium 4 @ 2.0 GHz with 2 G memory and normal Pentium 4 @ 2.0 GHz with .5 G memory. We ran the same 53 seconds clip test file for each codec 10 times on two different machines and calculate the time that it takes each codec to encode and decode. Here is the result.

Codecs	Encode Duo	Decode Duo	Encode normal	Decode normal
MELP	27.451 s	11.903 s	15.500 s	11.903 s
G711	.0919 s	.712 s	.740 s	1.089 s
iLBC 15.2	23.357 s	10.323 s	27.213 s	9.729 s
iLBC 13.3	27.388s	13.465 s	32.340 s	10.020 s
G728	36.794 s	33.607 s	39.430 s	34.633 s
G729	.469 s	.563 s	.430 s	.420 s
G723.1 6.3	59.902 s	7.497 s	1m20.820s	6.180 s
G723.1 5.4	46.094 s	8.218 s	54.710 s	6.100 s
CVSD 24K	.438 s	.564 s	.380 s	.470 s
CVSD 32K	.454 s	.533 s	.420 s	.533 s
AMR 12.2	1 m 6.482 s	9.731 s	1m30.322s	13.390 s
G726	8.967 s	8.452 s	15.300 s	13.760 s

As one can see the G.711 and G.726 takes the least amount of time to encode as well as decode, therefore it must be the least complex. However, AMR 12.2 and G.723.1 takes the most time, so these codecs must be the most complex.

4. BANDWIDTH REQUIREMENTS

One of the constraints that we are dealing with is the bandwidth limitation. Due to the minimal amount of bandwidth that is available on the network, in the time of crisis, it's important to utilize the codec with the least amount of bandwidth that way we do not congest our limited network. However, we have to sacrifice the speech quality as the result. Before getting into how much speech quality we have, or must sacrifice; we must understand what the minimum bandwidth requirement for each of the codec. The list below is amount of bandwidth it must have in order for the codec to function properly.

CODEC	Raw CODEC Data Rate
G.711	64kbps
AMR-EFR	12.2kbps
G.723	6.3 kbps 5.6 kbps
G.726	32kbps
G.729	8kbps
Ilbc	15.2kbps 13.3kbps
MELP	2.4 kbps
G.728	16kbps
CVSD	24kbps 36kbps

There is an additional overhead that the network requires such as IP, RTP, and UDP header, which are about 40 octets. In our system, the network also requires an additional header such as AOS, EP, HDLC, IPSEC which add another 31 more octets. So with 71 additional octets, this is what the minimum bandwidth looks like for each codec:

CODEC	CODEC Frame Size (octet rounded) per 20ms	AOS/EP/HDLC/IPSEC / IP/UDP/RTP
ITU-T G.711 (64kbps)	160 octets	92.4kbps
GSM-AMR-EFR (12.2kbps)	31octets	40.8kbps
G.723 (6.3kbps)	24 octets	25.3kbps
G.723 (5.6kbps)	20 octets	24.267kbps
G.726 (32kbps)	80 octets	60.4kbps
G.729 (8kbps)	20 octets	36.4kbps
Ilbc (13.3kbps)	50 octets	32.267kbps
Ilbc (15.2kbps)	38 octets	43.6kbps
MELP (2.4kbps)	7 octets	31.2kbps
G.728 (16kbps)	40 octets	44.4kbps
CVSD-sox (24kbps)	60 octets	52.4kbps
CVSD-sox (36kbps)	80 octets	60.4kbps

In a network, we define latency as the amount of time that it takes a packet to travel from one designated point to another. Consequently the lower the latency is, the more natural the conversation will sound. For VoIP systems, a one-way latency of up to 200 ms is considered acceptable, but ITU G.114 [13], however, recommends a maximum one way delay 150 ms, or 300 ms round-trip. It's best, however, to keep the round trip delay to no more than 250 ms because at that point conversation quality declines. Furthermore, the network and the gateways at either end of the call contributes a huge amount of the delay in a VoIP system, the speed of light adds at least 5 ms of delay for every 1,000 miles of fiber. This implies that coast to cost connection of 3,000 miles will add at least 30 ms delay for a round trip connection. The congestion of the network also adds additional delay so therefore, it's best to keep the frame size to be at 20 or 30 ms. This limitation, however, does not apply to the space environment. Often, it takes at least couple seconds to send the message from space to the earth, so natural sounds would never occur, therefore, we can eliminate the limitation of latency, which means that we can append multiple frames into a packet to reduce the effect of overhead. For instance, normally if we transmit G.711 with a header of 71 octets, the minimum bandwidth it requires is 92.4Kpbs on a 20 ms frame.

AOS/EP/HDLC/IPSEC 31	IP 20	UDP 12	RTP 8	20ms data

However, if we combine 5 frames into a single packet, we reduce the bandwidth to 69.680 Kpbs.

AOS/EP/HDLC/IPSEC 31	IP 20	UDP 12	RTP 8	20ms	20ms	20 ms	20ms	20ms

The table below shows frames/packet and the corresponding bits/sec.

	Frame size	frame duration	Frames/	packets/	payload size	bits/	packet size	bits/
Codec	Bytes	Ms	Packet	Sec	Bytes	sec	Bytes	sec
	160	20	1	50.00	160	64000	231	92400
G.711	160	20	2	25.00	320	64000	391	78200
G.711	160	20	3	16.67	480	64000	551	73467
G.711	160	20	4	12.50	640	64000	711	71100
G.711	160	20	5	10.00	800	64000	871	69680
G.711	160	20	6	8.33	960	64000	1031	68733

But the downside of multiple frames per packet is that if there is an error and the network drop the packet; the entire packet is lost, which means instead of losing one frame of data, now we lost multiple frames of data. This begs the question of what is the optimal number of frames per packet that we can transmit without significantly degrading our data.

5. SIMULATION

We simulate all the codecs with three sets of data from female and three from male speaker and average the PESQ score on several different frame sizes: 20, 30, 40, 60, 80, 90, 100, and 120 ms.

Data Corpus

In this experiment, we are using the data taken from the open speech Repository (OSR) [1]. OSR is the project that "provides freely usable speech files in multiple languages for use in Voice over IP testing and other applications." In this study, we used the male and female American English speech. We used three sets of male and three sets of female speeches. We selected this corpus because it contains speeches with clear intelligibility with no discernible white noise.

Evaluation Method

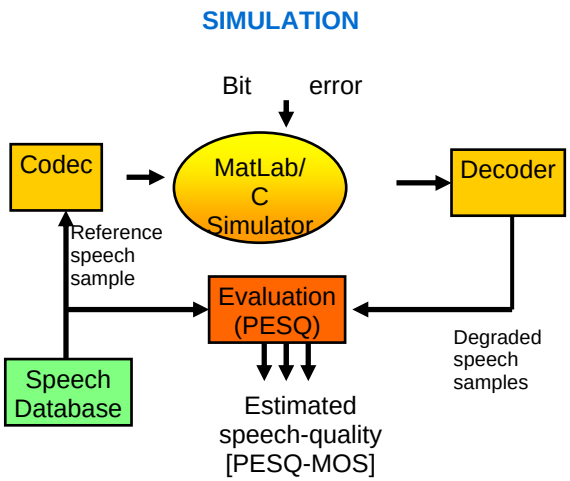
In general, it would be optimal with any speech experiment to have two different kinds of speech evaluating methods. One is subjective, in which the evaluator(s) listen to the speech and rate based on their perception of the speech. The other is objective, which is based on the physical parameters of the transmission channel. Ideally, we would like to have both, but due to a large volume of data that we processed, subjective testing method is not feasible. Therefore, we must rely on the objective testing method to evaluate how well the codecs perform under our constraint.

However, when we are talking about the objective testing method there are two criteria that are important for the evaluation: performance and quality. Even though there are many different ways that we can measure speech performance the most useful is the automatic speech recognition ("ASR"). Since ASR discriminates at the phoneme level, it can distinguish the difference between the rhyme words, which is similar to the speech intelligibility that is conducted using the human listener in the diagnostic rhyme test. This, however, had been studied by Quatieri [11]. The paper looked at what kind of affect the codecs can have on the ASR using GSM 12.2, G.729, G.723.1 at 5.3 kbps. However, the study did not address how BER on general transmission would affect the recognition. Hopefully, we would be able to elaborate on this issue on a future study.

As for quality, it has been established as an ITU standard to use the Perceptual Evaluation of Speech Quality algorithm ("PESQ", ITU-T P.862 standard) [4] which provides an objective measure of speech quality. PESQ uses a sensory model to compare the original, unprocessed signal with the degraded version at the output of the communications system. The result of comparing the reference and degraded signals is a quality score. This score is analogous to the subjective mean opinion score ("MOS") measured using panel tests according to ITU-T P.800 [12]. So, instead of looking at the speech intelligibility we can look at how much degradation each codec acquires through the transmission by looking at the PESQ score. Therefore, in this study, we use the PESQ score as the guideline to evaluate how well each codec performs.

Simulation Environment

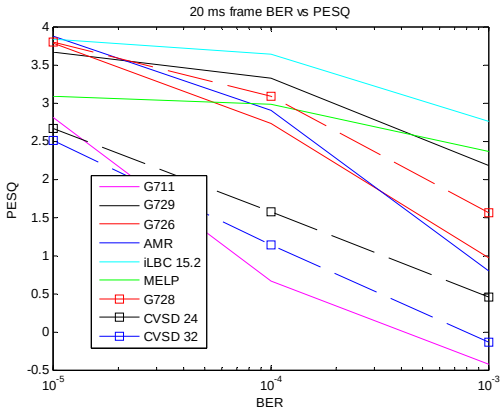
The simulation environment that we used is taken from the Florian Hammer paper [2], and the diagram is depicted in the figure below.

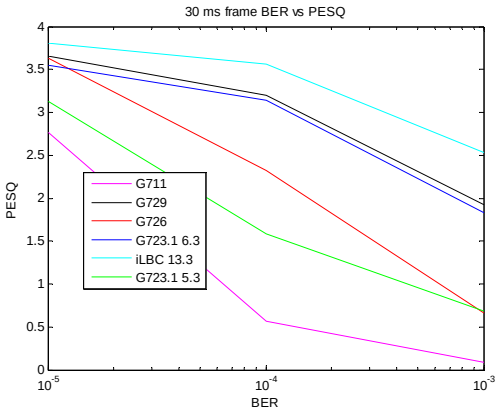


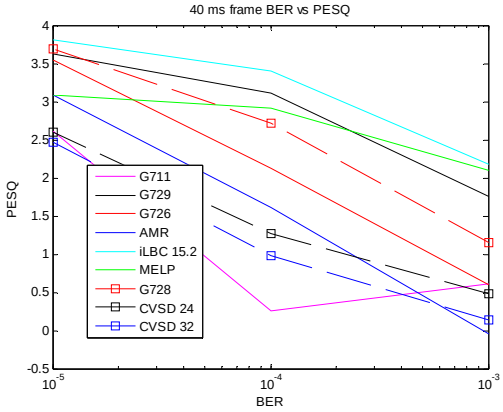
The three different types of bit errors rate ("BER") that we are looking at are 10-3, 10-4, and 10-5. We assume the error occurs in each bit is independent of each other and as a result of this we can use the binomial distribution to calculate the number of bit errors in each packet. Then, we distributed the errors uniformly throughout the packet. However, this is the simplistic error model, because our assumption is the special case where the channel coding at the physical link produces uniformly distributed bit errors. Furthermore, "we also assume that the lower system layers provide support for UDP-Lite by transferring the error data to the upper layers" [4]. In reality, this might not be the case; a more complex model is needed on future work.

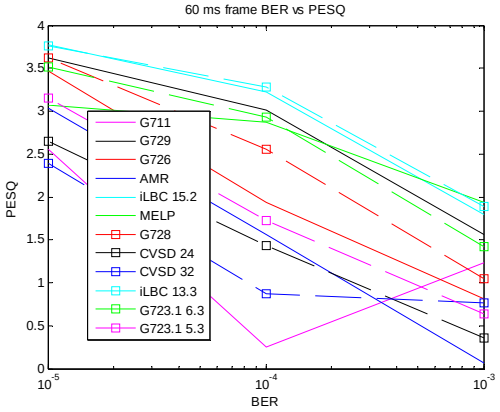
For each BER, we simulate the data files on all of the codecs listed above and each with the packet size of 20, 30, 40, 60, 80, 90, 100, and 120 ms. If there is an error within the packet we perform two kinds task: zeroing out the frame and flipping only the error bit. We ran each simulation 50 times and average the PESQ score.

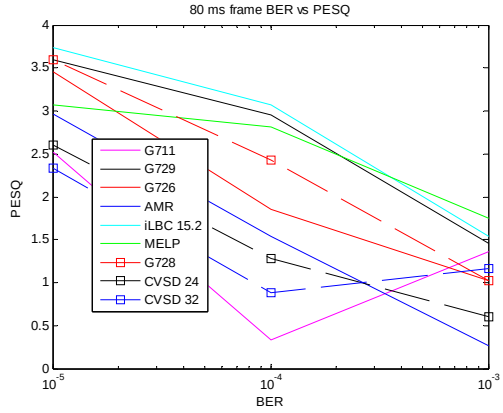
Results

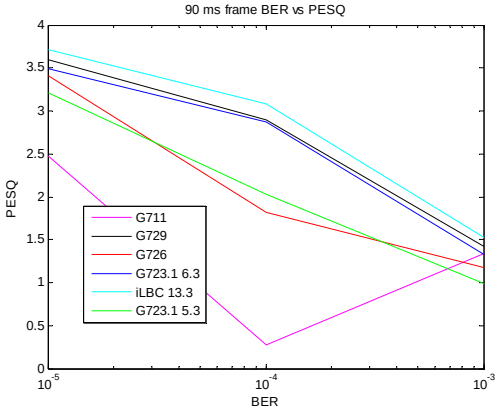


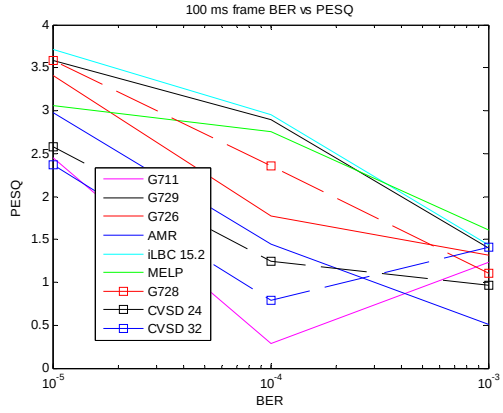


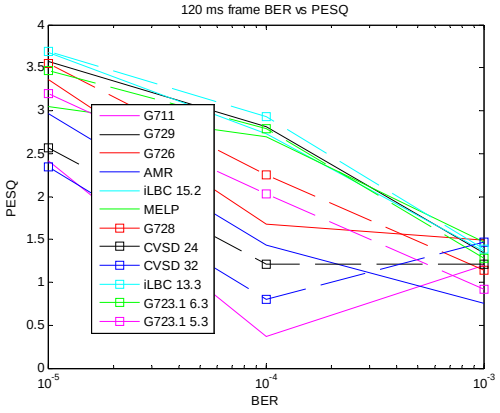








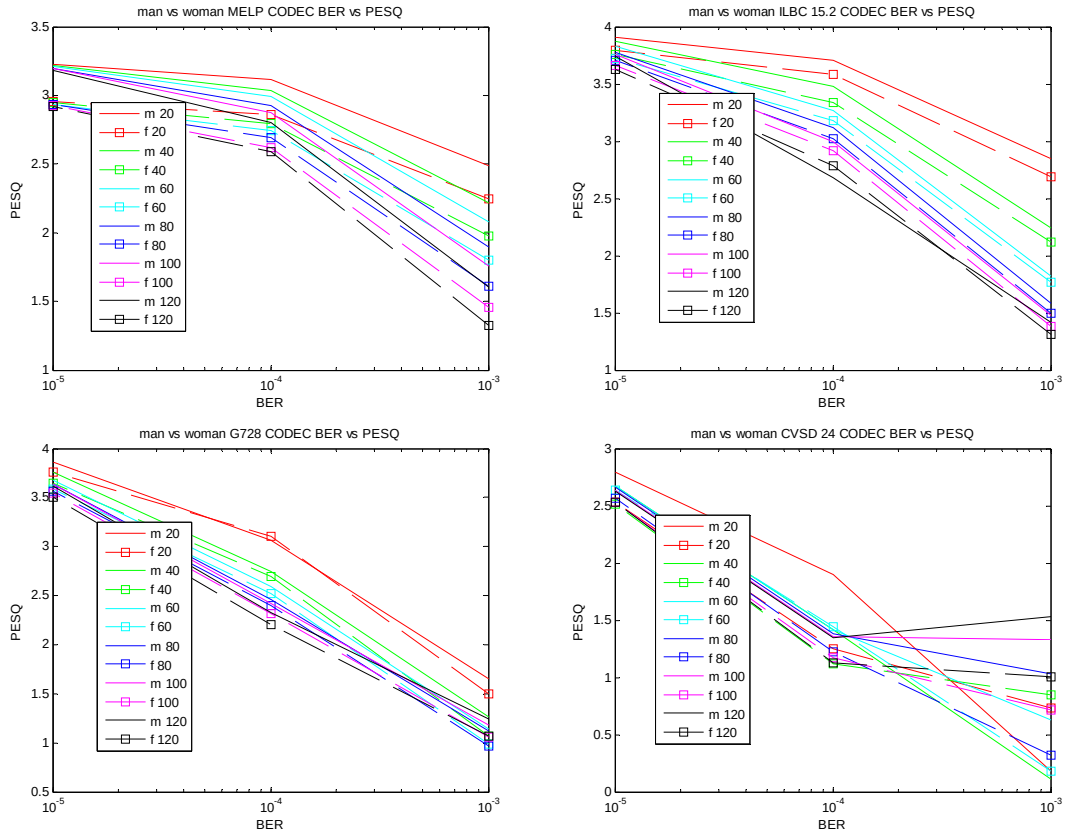


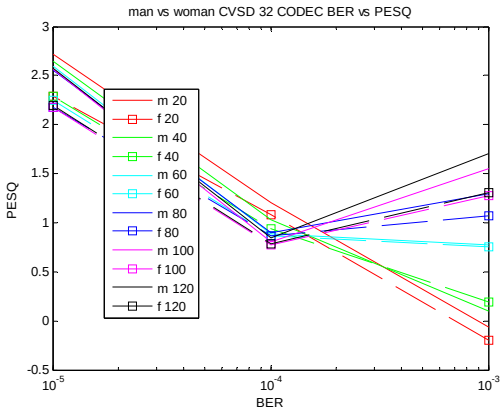


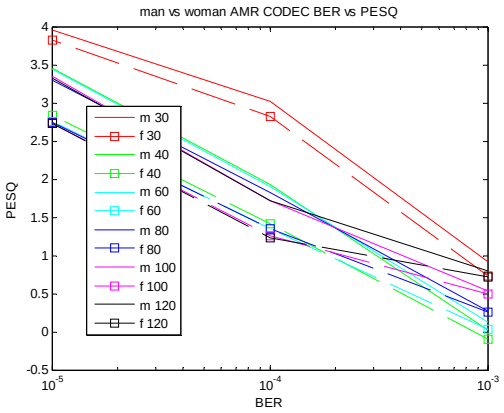
As we can see from the above graphs, even though we packet multiple frames into a packet, the degradation is still

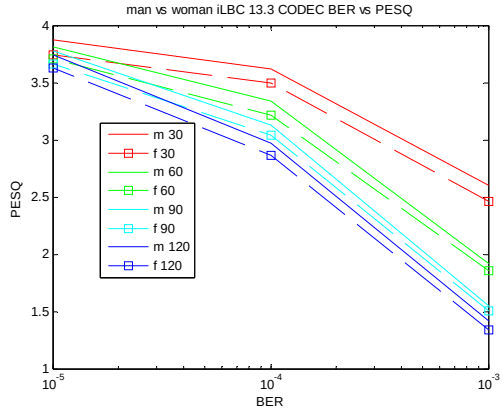
acceptable for BER 10^-5 and 10^-4 because the PESQ score is above 2.7 for some codecs. However, for BER 10^-3, at most we can packet is about 40 ms. If we look at codecs such as CVSD or G711, for example, the curve goes down then it goes up. The reason for this is if the all the values in the degraded speech are zero then the PESQ score is 1.5 comparing to the original speech. Meanwhile, the degraded speech sample, with a lot of noises or discontinuities, has negative PESQ. This is due the nature of PESQ measurement gives the zero value degraded speech data higher score than noise or discontinuity. Nevertheless, any PESQ below 2.7 indicates that it badly degrades and shouldn't be used so we don't really have to focus on it.

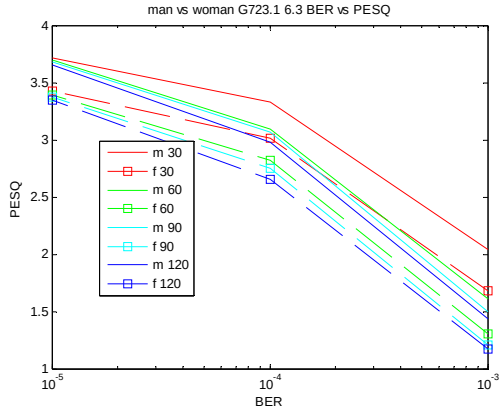
Next we try to see whether or not male and female speakers have any bearing on the quality of the speech. Florian has shown that at low bit rates, male voice rated higher than female speaker [2]. Using various codecs with different frame sizes, we confirm Florian's findings that male speaker are rated higher than female speaker in PESQ.

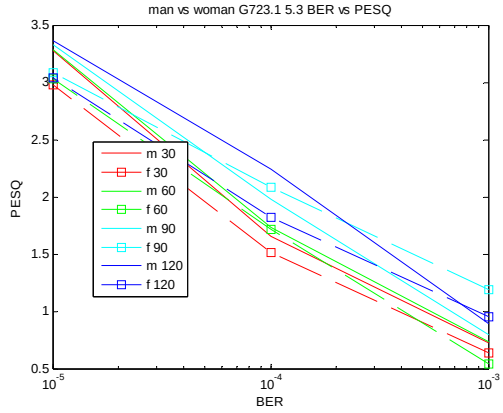


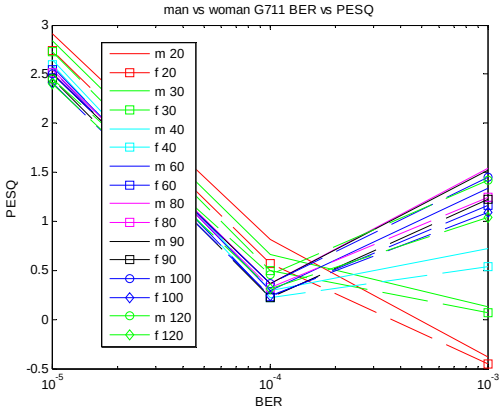


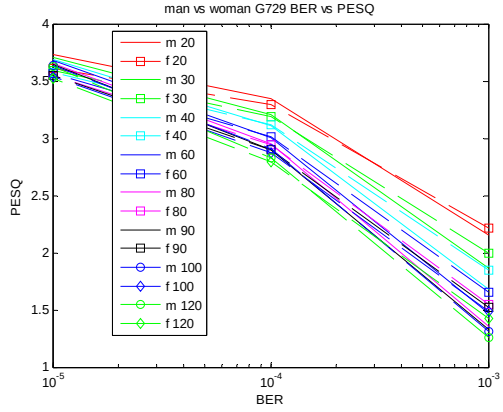


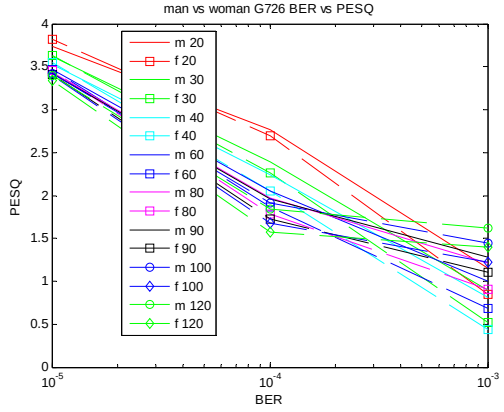








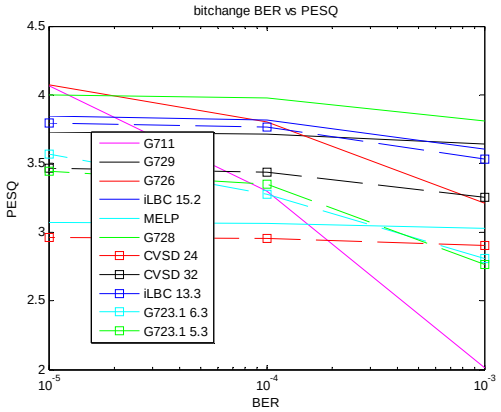




So if we want better score rating for female, we should use TOSQA instead.

The other characteristic of all the codecs we use is that they are lossy data compression. We do not need the receiving

data to be the exact replica of the sending data. It is to our advantage to turn off the cyclic redundancy check (CRC), because even with the bit flip, the result would not be devastating (like sending a command 'yes' and receive a command 'no'). In this experiment we are trying to see how well the codecs performs under flipping bit rather than dropping the whole packet.



When we are dealing with the bit flip, it doesn't matter whether we are using 20 ms or 120 ms frame size, the result is almost identical in most codecs. To our surprised, however, except for G711 with BER of 10^-3, every single codec is in the acceptable PESQ scored. This means it's better for us to turn off the CRC speech data sending from space to the earth because we can appending multiples frame into a packet to reduce the affect of overhead without degrading the source too much. Despite this, we still have to be careful because the PESQ scored only gives us the quality aspect of speech, but doesn't give us the intelligibility. So, it's necessary to continue working on how bit flip affects speech recognition. That way, we can have a comprehensive view of how BER affects speech in term of quality as well as intelligibility.

6. Conclusions and Future Work

There are many differences between space communication systems and day to day wireless networks. One of the main differences is the fact that space communication is not constrained by latency. As we mentioned above, taking this advantage, we can compact multiple frames into a package. As a result, we find that with BER of 10^-5 and 10^-4, the frame size does not severely degrade the original speech. However, with BER of 10^-3, we can only compact at most 40 ms frame to be acceptable. Furthermore, the codec seems to work best in maintaining the quality is iLBC 15.2Kbps, but G.729 is almost as good as iLBC 15.2, but with less complexity. So G.729 is an optimal choice when we are constrained by energy or resources. In addition, if we turn off the CRC, then the quality improves tremendously for low BER such 10^-3 and it doesn't have any constraints on how many frames we can packet. However, we do not know how much of the content has been altered. This leads us to our next study using speech recognition to see if we turn off the CRC, is the content severely altered.

7. Acknowledgements

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology and the National Aeronautics and Space Administration. Thanks to Jackson Pang and Philip Tsao for providing insightful comments.

References

[1] Open Speech Repository
http://www.voiptroubleshooter.com/open_speech/index.html

[2] Florian Hammer, Corrupted Speech Data Considered Useful: Improving Perceived Speech Quality of VoIP over Error-Prone Channels, Acta Acustica United with Acustica Vol. 90 (2004) 1052-1060

[3] Sox
sox.sourceforge.net/

[4] International Telecommunication Union: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T Recommendation P.862 (2001).

[5] European Telecommunications Standards Institute: Universal mobile telecommunications system (UMTS); Mandatory speech codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec frame structure (3GPP TS 26.101 version 5.0.0 Release 5). ETSI TS 126 101 v5.0.0 (2002).

[6] 3rd Generation Partnership Project (3GPP).
http://www.3gpp.org/.

[7] International Telecommunication Union: 40, 32, 24, 16 kbit/s ADAPTIVE
DIFFERENTIAL PULSE CODE MODULATION (ADPCM). ITU-T Recommendation G.726 (1990)

[8] Jagadeesh Balam and Jerry Gibson, Multiple Description Coding and Path Diversity for voice Communication over MANETS, International Wireless Communication and Mobile Computing Conference (IWCMC), Vancouver, Canada, July 3-6, 2006

[9] International Telecommunication Union: Dual rate Speech Coder for Multimedia Communications Transmitting at 5.3 Kbps and 6.3 Kbps. ITU-T Recommendation G.723.1 (1996)

[10] J. Holub and M.D Street, Low Bit Rate Networks – A Challenge for Instrusive Speech Transmission Quality Measurements, Prague: CTU, 2003, p. 47-49. ISBN 80-01-02822-4.

[11] T.E Quateri and E. singer, Speaker And Language Recognition Using Speech Codec parameters in Proc. Eurospeech'99, 1999, vol. 2, pp. 787–790.

[12] International Telecommunication Union: Methods for

subjective determination of transmission quality. ITU-T Recommendation P.800 (1996).
[13] International Telecommunication Union: G.114 (1996).
[14] I. Chakeres, H. Dong, E. Belding-Royer, A. Gersho, and J. D. Gibson Allowing Bit Errors in Speech over Wireless LANs Elsevier Computer Communications Journal, 28, pp. 1643-1657, March 2005.
[15] Data Compression Download
http://www.data-compression.com/download.shtml
[16] Internet Low Bit Rate
http://www.ilbcfreeware.org/software.html
[17] Voice Age
www.voiceage.com
[18] International Telecommunication Union: The 16 kb/s Low-Delay CELP algorithm.
(ITU-T Recommendation G.728).

BIOGRAPHIES

Sam Nguyen *is an academic part time member of the Communications Networks Group at JPL. He received a BS in Biomedical Electrical Engineering and BA in Mathematics from University of Southern California in 2003 and Currently working on his MS in Electrical Engineering and Ph.D in Mathematics the University of Southern California.*
.

Clayton Okino *received a BS in Electrical Engineering at Oregon State University in 1989, a MS in Electrical Engineering at Santa Clara University in 1993, and a Ph.D. in Electrical and Computer Engineering from the University of California, San Diego in 1998. After receiving his Ph.D., Dr. Okino accepted a position as an assistant professor in Thayer School of Engineering at Dartmouth College, where he pursued research in communication and wireless networks, emphasizing on performance and security. In 2001, Dr. Okino accepted a position as a Senior Member of the Technical Staff in the Digital Signal Processing group at Jet Propulsion Laboratory where his current research is in wireless link layer, network, and access algorithms, reconfigurable sensors, wireless QoS and location based processing techniques.*

Loren Clare *is the supervisor for the Communications Networks Group at the Jet Propulsion Laboratory. He obtained the Ph.D. in System Science from the University of California, Los Angeles in 1983. His research interests include wireless communications protocols, self-organizing systems, network systems design, modeling and analysis, and distributed control systems. Prior to joining JPL in May 2000, he was a senior research scientist at the Rockwell Science Center, where he acquired extensive experience in distributed sensor networks, satellite networking, and communications protocols for realtime networks supporting industrial automation.*