VOICE SIGNALS ANALYSIS TOOL

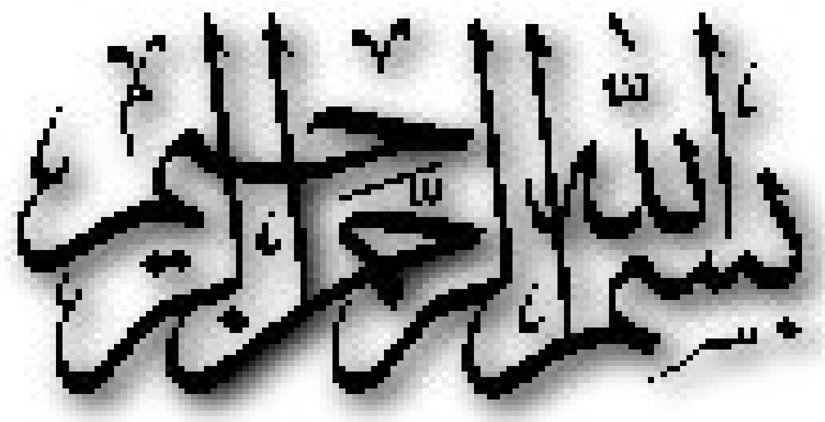
BY

ZAFAR IQBAL

Department of Computer Science

Military College Of Signals
National University Of Science and Technology
Rawalpindi

2003


بسم الله الرحمن الرحيم

ACKNOWLEDGEMENT

I am humbly great full to “Almighty Allah” for his blessing upon me, and giving me the wisdom, enlightenment and understanding, without which I would not have been able to downright this thesis.

I wish to explicit my deepest gratitude to my thesis advisor, Lieutenant Colonel Ahsen Saeed Zaidi for his supervision, devotion and commitments with this project and made this effort fruit full. I pay sincere gratitude to National University Of Science and Technology and Military College Of Signals for the provision of facilities for this project.

I also thank all those, whose names have not been mentioned here, for helping me on every step on my way and making this project a reality.

DEDICATED

TO

MY FAMILY

ABSTRACT

Last few years had witnessed an explosive growth in the field of digital communication. One of the principle techniques enabling this growth is voice/speech coding in which an analog signal from a microphone is digitally sampled via an A-to–D converter and then efficiently compressed into a digital bit stream for transmission or storage. A corresponding voice decoder receives these bit stream samples, which are suitable for playback through D-to-A converter and a loudspeaker.

Speech coders are of different forms each of which differ in terms of bit rate (degree of compression), complexity (MIPS and Memory) and voice quality. There are a large number of voice coding techniques which are being used presently e.g. RELP, CELP, STC, MBE, ADM, ADPCM, LPC, VQ etc.

The theme of thesis is to implement ADPCM voice coding technique and develop and implement software to identify the different versions of ADPCM technique like G.711, G.721, and G.723 at various compression rates. These codes were studied, analyzed, implemented and appropriate code was made for the respective identifications and conversions. Matlab was also used for their verifications.

TABLE OF CONTENTS

CHAPTER 1:INTRODUCTION

CHAPTER 2:SPEECH COMPRESSION

CHAPTER 3:VOICE COMPRESSION TECHNIQUES

CHAPTER 4: ADAPTIVE DELTA PCM

LISTOF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 GENERAL

Last few years had witnessed explosive growth in the field of digital communication. One of the principal technologies enabling this growth is compression. Compression is the art of representing information in a compact form. An early example of compression is Morse code, developed by Samuel Morse in the mid-19th century. Letters sent were encoded with dots and dashes. Morse noticed that certain letters existed more often than others. In order to reduce the average time required to send a message, he assigned shorter sequences to letters that occur more frequently such as *e* (·) and *a (*· -), and longer sequences to letters that occur less frequently such as *q* (- - ·-) and *j* (·- - -). An early version of compression approach, called the *vocoder (voice coder)* was developed by Homer Dudley at Bell Laboratories in 1936. The vocoder was demonstrated at the New York World's Fair in 1939 where it was a major attraction [8].

1.2 WHAT IS ‘COMPRESSION’?

Data compression is the removal of redundant data. It reduces the number of binary bits necessary to represent the information contained within the data. To achieve the best possible compression it requires not only an understanding of the nature of data in its binary representation but also how humans interpret the information that the data represents.

1.3 WHY IS IT NECESSARY?

"Compression is the key to the future expansion of the Web; it is certainly the key to emerging multimedia and 3-D technology." (Brown, Honeycutt, et al 1998)[2]

Since every *bit* incurs a cost when being transmitted or stored, any technology that can be introduced into existing systems that can reduce these costs is essential. When considering raw data that may contain over 50% redundancy, it raises the question – why pay for that redundant information?

1.4 WHY REDUNDANCY WITHIN BINARY DATA?

The nature of computing technology represents and processes pieces of data (symbols) of a similar type using equal numbers of bits. For example, the same number of bits represents the symbols ‘e’ and ‘z’. Statistically, however, it can be shown that ‘e’ appears within English text more frequently than ‘z’. It would seem that by representing the more-frequent symbols with fewer bits than the less-frequent symbols the total number of bits necessary to represent the information could be reduced.

1.5 COMPRESSION EXAMPLES

Compression plays an important role in the ability to transmit digital television signals. Consider High Definition Television (**HDTV**) in which to transmit a HDTV signal without any compression it would require about 884 Mbytes per second transmission rate. In order to transmit this much data a channel bandwidth of about 220 MHz is required. With data compression less than 20 Mbytes per second (along with audio information) only 6 MHz of transmission bandwidth is sufficient.

The 1980s saw tremendous growth in the use of the fax machine. This would not have been possible without compression technologies that permit speedy transmission. Without compression, it would take 24 hours to send one ISO-page document.

Compression is now very much a part of everyday life. Most modems now have compression capabilities, which allow transmitting data many times faster than otherwise possible. File compression utilities, which permit to store more on disks, are now commonly being used.

1.6 COMPRESSION THEORY TERMINOLOGY

Following are different types of terminologies which will be used at various places for necessary reference: -

Table1.1 Terminologies

COMPRESSOR	Software or hardware device that compresses the data
DECOMPRESSOR	Software or hardware device that decompresses the data
CODER	Software or hardware device that compresses/decompresses the data
ALGORITHMS	The logic that governs the compression/decompression process

Almost all the basic compression methods function by considering individual symbols within the context of their surrounding symbols. In a piece of text this symbol may be a character (or a sequence of characters), in a picture it may be a pixel and it could even just be a sequence of bits.

1.7 COMPRESSION TECHNIQUES

Compression technique or compression algorithm actually refers to two algorithms. One compression algorithm that takes an input X and generates a representation X_C that requires fewer bits, and other is a reconstruction algorithm that operates on the compressed representation X_C to generate the reconstruction Y. These operations as shown schematically (compression and reconstruction algorithms) mean the compression algorithm, [12].

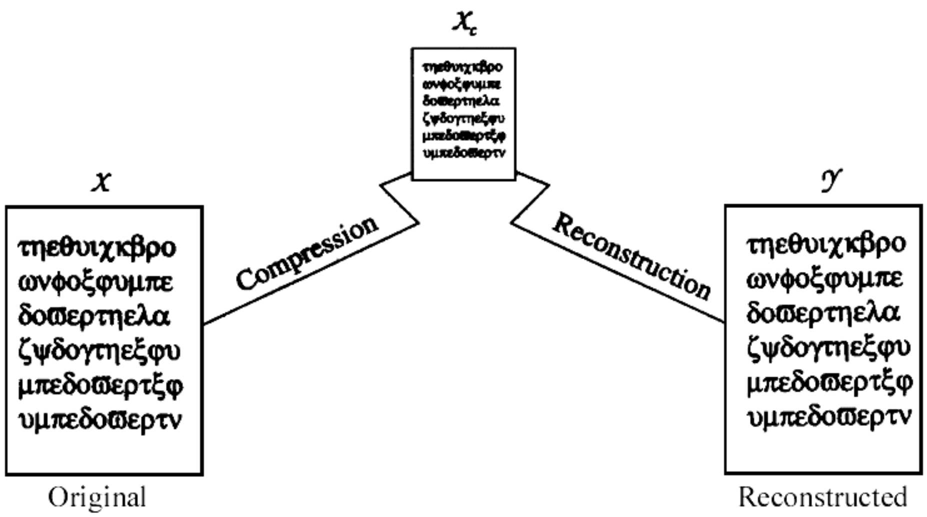


Figure1.1 Compression and reconstruction

Based on the requirements of reconstruction, data compression schemes can be divided into two broad classes: Loss less compression scheme, in which Y is identical to X and Lossy compression scheme, which generally provides much higher compression than loss less compression but allows Y to be different from X.

1.8 LOSSLESS COMPRESSION

Loss less compression technique as its name implies involve no loss of information. If data has been compressed by this technique the original data can be recovered exactly from the compressed data. Loss less compression is generally used for "discrete" data, such as text, computer-generated data, and some kinds of image and video information.

Text compression is an important area for loss less compression. It is very important that the reconstruction is identical to the original text as very small differences can result in statements with different meanings. Consider the sentences "Do *not* send money" and "Do *now* send money." A similar argument holds true for computer files and for certain types of data such as bank records.

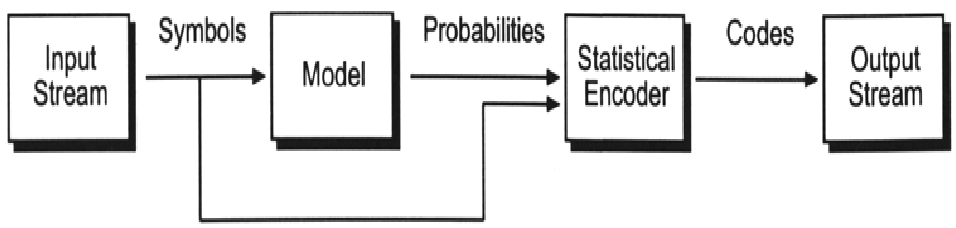


Figure 1.2 Loss less compressor

If data of any kind is to be processed or enhanced to yield more information later on, it is important that the integrity be preserved. For example if a radiological image is compressed in a lossy fashion and the difference between the reconstruction Y and original *X* are visually undetectable. If this image was enhanced later the previously undetectable differences may cause the appearance of artifacts, which could seriously mislead the radiologist. Because the price for this kind of mishap may be a human life so it makes sense to be very careful about using a compression scheme that generates a reconstruction that is different from the original.

Data obtained from satellites are often processed later to obtain different numerical indicators of vegetation, deforestation, and so on. If the reconstructed data are not identical to the original data, processing may result in enhancement of the differences. It may not be possible to go back and obtain the same data over again. Therefore, it is not advisable to allow for any differences to appear in the compression process. There are many situations that require compression where the reconstruction should be identical to the original. There are also a number of situations in which it is possible to relax this requirement in order to get more compression.

1.9 LOSSY COMPRESSION

Lossy compression technique involves some loss of information and data that has been compressed using lossy techniques generally cannot be recovered or reconstructed exactly. However in many applications this lack of exact reconstruction is not a problem. For example when storing or transmitting speech the reconstruction of exact value of each sample of speech is not necessary. Depending on the quality required of the reconstructed speech, varying amounts of loss of information about the value of each sample can be tolerated. If the quality of the reconstructed speech is to be similar to that heard on the telephone, a significant loss of information could be tolerated. However, if the reconstructed speech needs to be of the quality heard on a compact disc the amount of information loss that can be tolerated is relatively low. Similarly when viewing a reconstruction of a video the fact that reconstruction is different from the original is generally not important as long as the differences do not result in annoying artifacts. Thus video is generally compressed using lossy compression, [3].

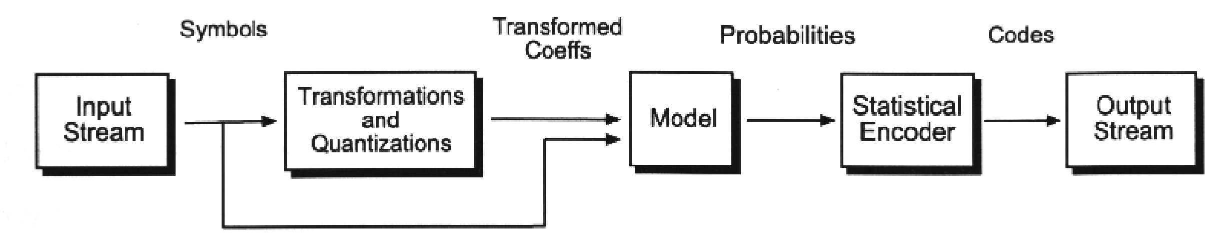


Figure1.3 Lossy Compressor

Once a data compression scheme is developed its performance is to be measured. Because of the number of different areas of applications different terms have been developed to describe and measure this performance.

1.10 MEASURES OF PERFORMANCE

A compression algorithm can be evaluated in a number of different ways like measuring the relative **complexity** of the algorithm, the **memory** required to implement the algorithm, how **fast** the algorithm performs on a given machine, the **amount** of compression and how closely the reconstruction **resembles** to the original, [1].

A very logical way of measuring how well a compression algorithm compresses a given set of data is to look at the ratio of the number of bits required to represent the data before compression to the number of bits required to represent the data after the compression. This ratio is called the **compression ratio**. Suppose to store an image made up of a square array of 256 X 256 pixels requires 65,536 bytes. Once this image is compressed and it requires 16,384 bytes it is said that the compression ratio is **4:1.**

Another way of measuring compression performance is to count the average number of bits required to represent a single sample. This is generally referred to as the **rate**. For example, in case of compressed image described above if eight bits per byte (or pixel) are assumed then the average number of bits per pixel in the compressed representation is two. Thus the rate is two bits per pixel.

In lossy compression, the reconstruction differs from the original data. Therefore, in order to determine the efficiency of a compression algorithm some way of quantifying the difference should be defined. The difference between the original and the reconstructed data is often called **distortion**. Lossy techniques are generally used for the compression of data that originates as analog signals, such as speech and video. Since analog signals are often called **waveforms** therefore compression of analog signals is referred to as *waveform coding.* In waveform coding of speech and video, the final arbiter of quality is human. Because human's responses are difficult to model mathematically many approximate measures of distortion are used to determine the quality of the reconstructed waveforms.

Other terms which are also used when talking about differences between the reconstruction and the original data are **fidelity and quality**. When the fidelity or quality of a reconstruction is high it means that the difference between the reconstruction and the original is small. Whether this difference is a mathematical difference or a perceptual difference.

1.11 LOSSLESS AND LOSSY SPEECH COMPARISON

In various digital speech applications which require speech compression, some are concerned with the intelligence of the speech and do not allow error in reconstructed speech (e.g. military application) and thus employ loss less compression techniques. But as in loss less digital data compression there is a trade-off of compression ratio for the error-free reconstruction. Nevertheless for most of commercial applications the speech signal is allowed to be distorted at various levels depending on different applications.

So if the loss is allowed, lossy speech compression can be used. The level of distortion varies from application to application. For example, speech application in toys requires less fidelity in reconstructed speech and conferencing application would certainly requires higher intelligence of speech as a result these can afford some loss of information during the coding process. In addition since most of the time the final receiver is human which cannot receive all informations due to the human auditory system characteristics therefore due to the human-based nature of lossy speech compression technique lossy compression performance usually outperforms the loss less compression.

The theoretical possibility of image compression (no matter loss less or lossy) is primarily based on the redundancy inside image. Beside this redundancy in speech signal lossy speech compression can also utilize the characteristics of Human Auditory System (**HAS**). These characteristics actually divided information in the speech into two types: *relevant* to HAS and *irrelevant* to HAS. By excluding information irrelevant to HAS a better compression ratio can be obtained. Some HAS characteristics such as spatial masking and equal loudness curve help predicting what information is relevant to our hearing.

CHAPTER 2

SPEECH COMPRESSION

2.1 INTRODUCTION

Speech compression algorithms seek to minimize the bit rate in digital representation of a signal without an objectionable loss of signal quality in the process. High quality is attained at low bit rates by exploiting signal redundancy. Models of signal redundancy and distortion masking are becoming increasingly more sophisticated, leading to continuous improvement in the quality of low bit rate signal.

Speech compression can be divided into two methods; one is through the frequency domain and the other through the time domain. The types of speech compression algorithms used depend on the functionality required. A measure of speech quality is MOS (Mean Opinion Score) that is rated on a 5-point scale. Other factor, which affects the selection of algorithms used, is the computational complexity of the algorithm. This is normally measured in MIPS (Million of Instruction per second). Compression for both speech and data is done through the removal of redundancy. In the case of speech compression, further compression can be obtained by removing irrelevancy, which is the imperceptible reconstruction error or distortion to the speech signal, [1].

For medium rate speech coding, an analysis-synthesis method is normally used. This means the speech is represented by a set of compact parameters, which are then coded. Prior to compression, the speech signal goes through an analysis stage, which can either be closed-looped or open-looped. In a closed-looped analysis, the parameters are extracted and coded and further reduced by a difference between the original speech signal and the reproduced signal. Hence this is also called an analysis-by-synthesis approach. The goal in most compression techniques is to reduce the transmitted data rate and the storage space required. The algorithms used to compress speech are also chosen basing upon the speech quality required. Certain algorithms are capable of high compression ratio of up to 24: 1.

2.2 BASIC PROPERTIES OF SPEECH

Before discussing the various types of speech compression techniques, a brief introduction of the properties of speech will be given to allow better understanding of the various speech coding methods.

Speech signals are non-stationary signals. If speech is split into segments, these can be treated as quasi-stationary over a time frame of 5-20 ms. Speech signals can either be voiced, unvoiced or a mixture of the two. Speech means the pronunciation of vowels and unvoiced is the pronunciation of other letters. Voiced signal energy is normally higher than unvoiced signals. Forcing air through a restriction in the vocal tract produces unvoiced speech. The acoustical coupling of the vocal tract of the nasal tract produces nasal sounds such as pronunciation of the letter “N”. Plosive sounds such as the pronunciation of the letter "P" is formed by releasing a build up of air pressure behind a closure in the vocal tract. Speech reproduction systems use these characteristics to reproduce speech, [3].

2.3 THE DIMENSIONS OF PERFORMANCE IN SPEECH COMPRESSION

Speech coders attempt to minimize the bit rate for transmission or storage of the signal while maintaining required levels of speech quality, communication delay, and complexity of implementation (power consumption). The brief description of the above parameters of performance, with particular reference to speech is as under: -

2.3.1 Speech Quality

Speech quality is usually evaluated on a five-point scale, known as the mean-opinion score (MOS) scale. The five points of quality are: *bad*, *poor*, *fair*, *good*, and *excellent.* Quality scores of 3.5 or higher generally imply high levels of intelligibility, speaker recognition and naturalness.

2.3.2 Bit Rate

The coding efficiency is expressed in bits per second (bps). Bit rates for different types of compressions are shown as under: -

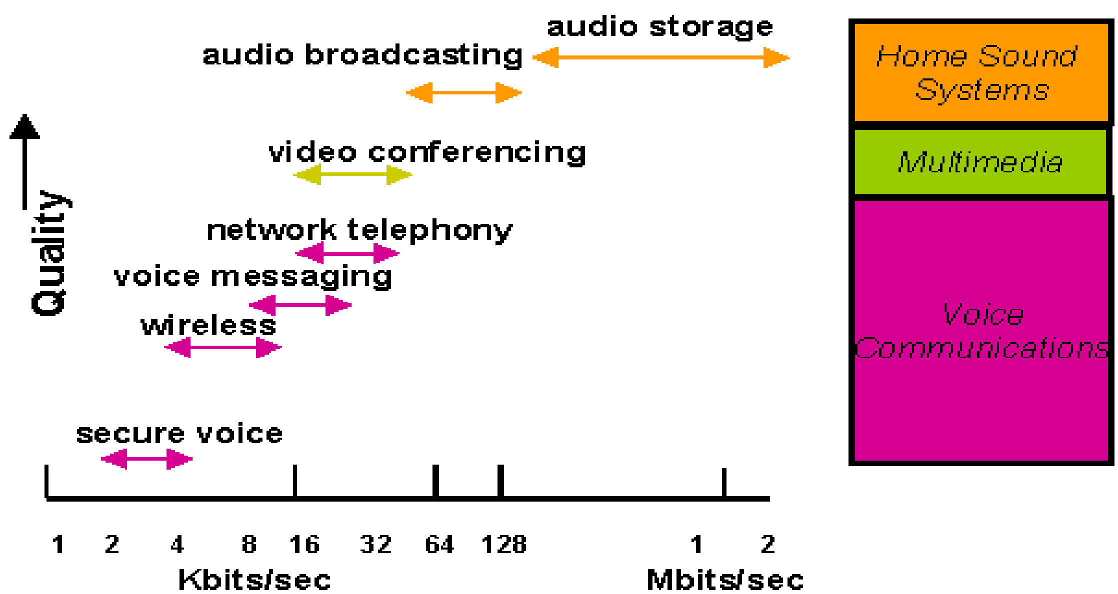


Figure2.1 Comparison Of Bit Rates

2.3.3 Communication Delay

Speech coders often process speech in blocks and such processing introduces communication delay. Depending on the application, the permissible total delay could be as low as 1 m sec as in network telephony or as high as 500 m sec as in video telephony. Communication delay is irrelevant for one-way communication, such as in voice mail.

2.3.4 Complexity

The complexity of a coding algorithm is the processing effort required to implement the algorithm and it is typically measured in terms of arithmetic capability and memory requirement. A large complexity can result in high power consumption in the hardware.

2.4 THE CONTINUING NEED FOR SPEECH COMPRESSION

The capability of speech compression is central to the technologies of robust long-distance communication, high-quality speech storage and message encryption. Compression is an important technology in communication in spite of the presence of optical transmission media of relatively unlimited bandwidth. This is because of continuous increasing need to use band-limited media such as radio and satellite links and bit-rate-limited storage media such as CD-ROMs and silicon memories. Storage and archival of large volumes of speech information make speech compression essential even in the context of significant increases in the capacity of optical and solid-state memories.

Low bit-rate speech technology is a key factor in meeting the increasing demand for new digital wireless communication services. Impressive progress has been made during recent years in coding speech with high quality at low bit rates and at low cost. Only few years ago once high quality speech could not be produced at bit rates below 24 kbps. Today high quality at 8 kbps is available for the new digital cellular service in North America. Using new techniques for channel coding and equalization it is possible to transmit the 8 kbps speech in a robust fashion over the mobile radio channel in spite of channel noise, signal fading and inter symbol interference, [13].

Applications of wideband speech coding include high quality audio conferencing, high-quality stereo conferencing and dual-language programming over a basic ISDN link. The compression creates new opportunities in audio transmission, networking, electronic publishing, tele teaching, multi location games, multimedia memos and database storage.

CHAPTER 3

VOICE COMPRESSSION TECHNIQUES

3.1 INTRODUCTION

Although with emergence of optical fibers bandwidth in wired communication has become inexpensive there is a growing need for bandwidth conservation and enhanced privacy in wireless, cellular and satellite communications. It is required that speech signal should be in digital form so that it can be processed, stored and transmitted under software control. The function of a speech codec is to convert an analogue speech signal into a digital form for efficient transmission over a digital path or to store on a digital storage medium .It also form the complementary function of converting a received digital signal back to the analogue form. It is normally the speech codec, which defines the basic speech quality for the whole system.

3.2 METHODS

Two major methods of compression are static and dynami*c also called as* adaptive methods. The static method is consistent in how it treats the data. For example, every occurrence of a symbol in the uncompressed data is represented by the same code word throughout the compressed data. Whereas the dynamic method (although consistent in how it compresses data) makes the decision, which affects the way, it compresses the data. For example the code word that represents a particular symbol may change because the algorithm has ascertained a possibly more-efficient code word, [14].

Speech coding techniques can be broadly divided into two classes: waveform coding that aims at reproducing the speech waveform as faithfully as possible and vocoders that preserve only the spectral properties of speech in the encoded signal. The waveform coders are able to produce high-quality speech at high bit rates whereas vocoders produce intelligible speech at much lower bit rates but the level of speech quality in terms of its naturalness and uniformity for different speakers is also much lower. So the applications of vocoders so far have been limited to low-bit-rate digital communication channels. The combination of the principles of waveform coding and vocoding has led to significant new capabilities in recent compression technology.

3.3 TECHNIQUES FOR SPEECH COMPRESSION

Digital encoding of voice-band speech has been a topic of research for over three decades and as a result of this intense activity many strategies and approaches have been developed for speech encoding. It is not possible to list and discuss all of them here. However some of the popular techniques in digital speech coding field are mentioned. Some of them are classical compression techniques such as ADPCM and others are modern advanced compression techniques such as CELP. Some even considered as state-of-the-art speech compression methods as they archive both good compression ratio and computational speed, [5].

Speech coding techniques can be broadly divided into different main sections as shown in Figure below. The figure only shows the most widely used algorithms and the general function of these coding techniques is to analyze the signal, remove the redundancies, and efficiently code the non-redundant parts of the signal in a perceptually acceptable manner. The quality versus bit rate for the three main coding strategies is also shown below. In this figure, the quality is represented by mean opinion scores (**MOS**) ranging from 1 to 5 which corresponds to 1 = bad, 2 = poor, 3 = fair, 4 = good and 5 = excellent.

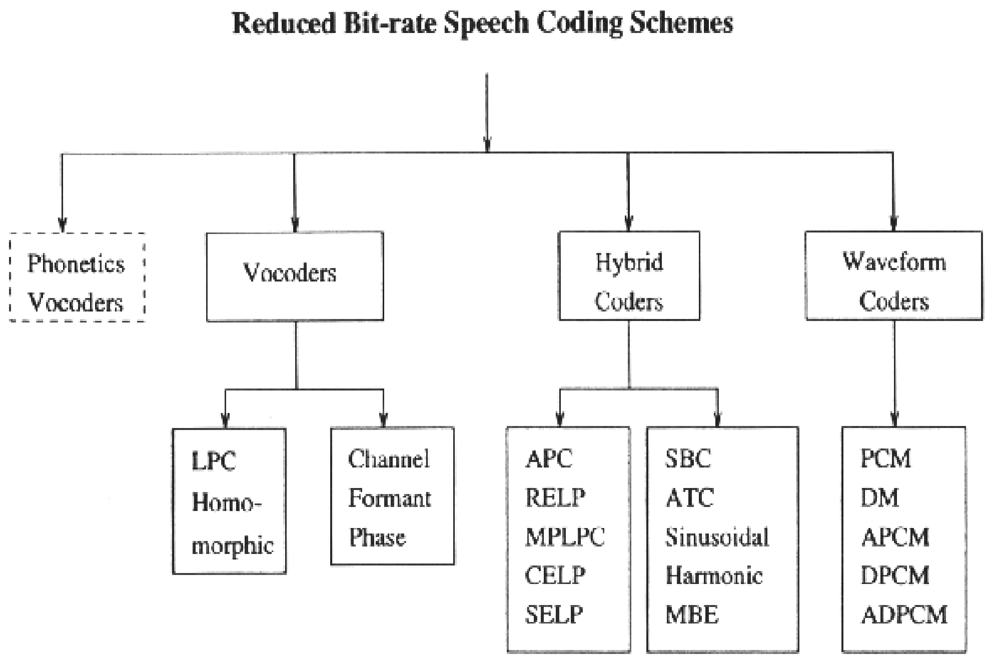


Figure 3.1 Classifications Of Speech Coding Techniques

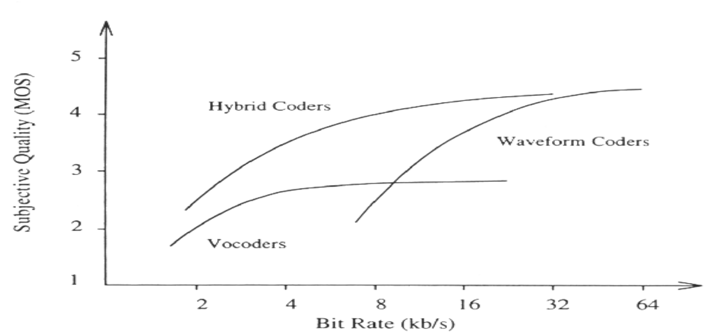


Figure 3.2 Quality Comparison Of Speech Coding Techniques

3.4 MEAN OPINION SCORE (MOS)

The quality of voice compression is usually measured in terms of Mean Opinion Score (MOS). This is a value between one and five that expresses how close the voice quality is to real-life communication. Figure shows this scale.

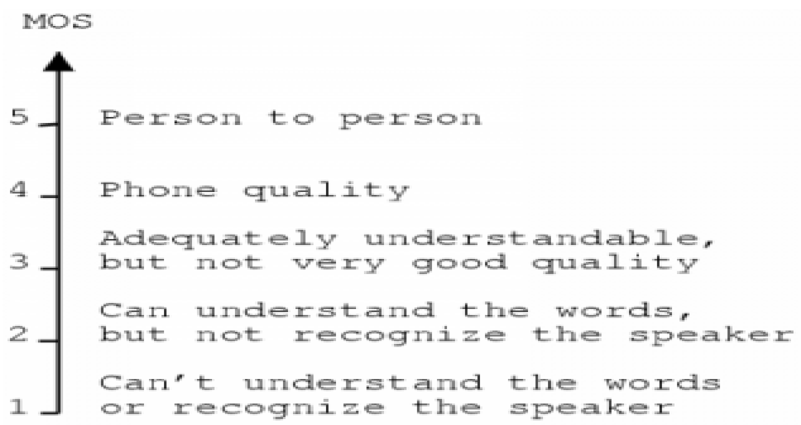


Figure 3.3 MOS

3.5 GENERAL COMPRESSION TECHNIQUES

3.5.1 WAVEFORM CODING

Waveform coding encodes the waveform itself in an efficient way. The signal is stored in such a way that upon decoding the resulting signal will have the same general shape as the original. Waveform coding techniques apply to audio signals in general and not just to speech as these techniques encode every aspect of the signal, [7].

The simplest form of waveform coding is PCM encoding technique. But a signal can be processed further to reduce the amount of storage needed for the waveform. In general, such techniques are lossy i.e. the decoded data can differ from the original data. Waveform coding techniques usually offer good quality speech requiring a bandwidth of 16 kbps or more.

3.5.1.1 PCM

Digitizing an audio signal is often referred to as pulse code modulation (PCM). Nowadays digitization and reconstruction of voice signals can be done by any PC soundcard.

3.5.1.1.1 Sampling And Quantization

A continuous signal (a voice signal for example) on a certain time interval has an infinite number of values with infinite precision. To be able to digitally store an approximation of the signal it is first sampled and then quantized. While sampling a signal infinite precision measures at regular intervals are taken. The rate at which the samples are taken is called the sampling rate. Next step is to quantize the sampled signal that means the infinite precision values are converted to values, which can be stored digitally, [4].

In general, the purpose of quantization is to represent a sample by an N-bit value. With uniform quantization the range of possible values is divided into 2N equally sized segments and with each segment, an N-bit value is associated. The width of such a segment is known as the step size .If the sampled value exceeds the range covered by the segments then this representation results in clipping.

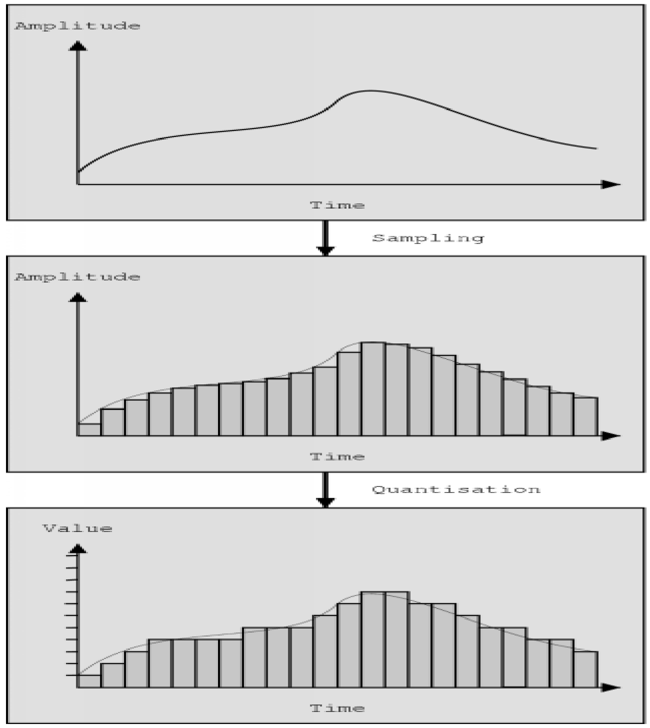


Figure 3.4 Sampling and quantization

With non-uniform quantization this step size is not constant. A common case of non-uniform quantization is logarithmic quantization. Here it is not the original input value that is quantized but in fact the log value of the sample. For audio signals this is particularly useful since humans tend to be more sensitive to changes at lower amplitudes than at higher ones.

Another non-uniform quantization method is adaptive quantization. With such methods the quantization step size is dynamically adapted in response to changes in the signal amplitude. PCM techniques which use adaptive quantization are referred to as adaptive PCM (APCM).

The sampling and uniform quantization steps are depicted in figure above. An important thing to note is that both steps introduce a certain amount of error. It is clear that a higher sampling rate and a smaller quantization step size will reduce the amount of error in the digitized signal, [8].

3.5.1.1.2 Reconstruction

Signal reconstruction is opposite to the digitization step. An inverse quantization is applied and from the samples a continuous signal is recreated. How much does the reconstructed signal resembles to the original signal depends on the sampling rate, quantization method and the reconstruction algorithm being used.

3.5.1.2 DIFFERENTIAL CODING

Differential coding exploits the fact that with the audio signals the value of one sample can be somewhat predicted by the values of the previous samples. Given a number of samples the algorithm will calculate a prediction of the next sampled value. Then it will only store the difference between this predicted value and the actual value. This difference is usually not very large and can therefore be stored with fewer bits than the actual sampled value, resulting in compression. Because of the use of a predicted value, differential coding is also referred to as predictive coding.

3.5.1.2.1 Differential PCM (DPCM)

Differential PCM merely calculates the difference between the predicted and actual values of a PCM signal and uses a fixed number of bits to store this difference. The number of bits used to store this difference determines the maximum slope that the signal can have if errors are to be avoided. If this slope is exceeded then the value of a sample can only be approximated, introducing an amount of error.

3.5.1.2.2 Adaptive DPCM (ADPCM)

An extension to DPCM is adaptive DPCM. With this encoding method there are a fixed number of bits used to store the difference. In contrast to the previous technique that simply used all of the bits to store the difference, ADPCM uses some of the bits to encode a quantization level. In this way, the resolution of the difference can be adjusted.

3.5.1.2.3 Delta modulation (DM)

Delta modulation can be said as a very simple form of DPCM. With this method only one bit is used to encode the difference. One value of the bit indicates an increase of the predicted value with a certain amount whereas the other indicates a decrease.

A variant of this scheme is called adaptive delta modulation (ADM). Here, the step size being used to increase or decrease the predicted value can be adapted. This way, the original signal can be approximated more closely.

3.5.1.2.4 Vector quantization

With vector quantization the input is divided into equally sized pieces called vectors. Essential to this type of encoding is the presence of a `codebook', an array of vectors. For each vector of the input, the closest match to a vector in the codebook is looked up. The index of this codebook entry is then used to encode the input vector, [8].

It is important to note that this principle can be applied to a wide variety of data and not only to PCM data. For example, vector quantization can be used to store an approximation of the error term of other compression techniques.

3.5.1.2.5 Transform coding

While considering PCM data the signal in time domain is being considered. With transform coding, the signal is transformed to its representation in another domain in which it can be compressed well than in its original form. When the signal is decompressed an inverse transformation is applied to restore an approximation of the original signal.

One of the domains to which a signal can be transformed is the frequency domain. Using information about human vocal and auditory systems a compression algorithm can decide which frequency components are most important. Those components can then be encoded with more precision than the others.

3.5.2 VOCODING

Waveform coding methods simply model the waveform as closely as possible. Vocoding techniques encode information about how the speech signal was produced by the human vocal system rather encoding the waveform itself.

The term vocoding is a combination of `voice' and `coding'. These techniques can produce intelligible communication at very low bit rates usually below 4.8 kbps.

3.5.2.1 Speech production

To understand how vocoding methods work a brief explanation of speech production is mentioned. Figure shows the human vocal system. To produce speech the lungs pump air through the trachea. For some sounds this stream of air is periodically interrupted by the vocal cords.

The resulting flow of air travels through the vocal tract. The vocal tract extends from the opening in the vocal cords to the mouth. A part of the stream travels through the nose cavity. The vocal tract has certain resonance characteristics. These characteristics can be altered by varying the shape of the vocal tract, for example by moving the position of the tongue. These resonance characteristics transform the flow of air originating from the vocal cords to create a specific sound. The resonance frequencies are called formants. The shape of vocal tract and the type of excitation (the flow of air coming out of the vocal cords) change relatively slowly. This means that for short time intervals, for example 20 ms, the speech production system can be considered to be almost stationary. Another important observation is that speech signals show a high degree of predictability. Sometimes due to the periodic signal created by the vocal cords and due to the resonance characteristics of the vocal tract, [6].

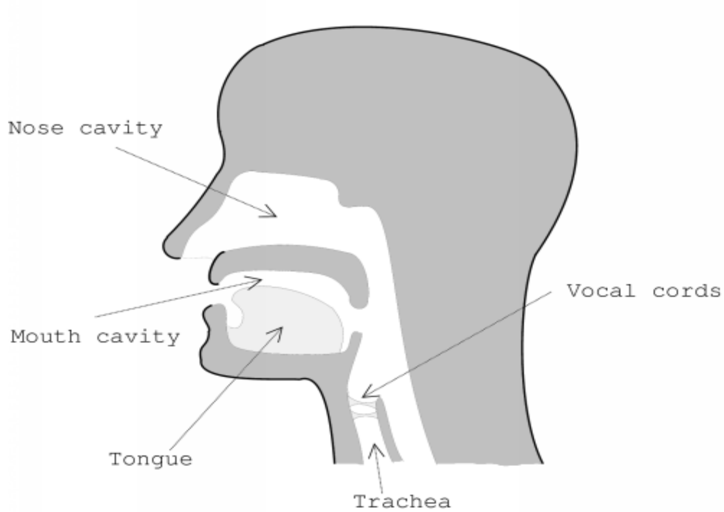


Figure 3.5 Human Vocal Systems

3.5.3 HYBRID CODING

Waveform coders in general do not perform well at data rates below 16 kbps. Vocoders on the other hand can perform at very low data rates, while still allowing intelligible speech. However the person producing the speech signal cannot be recognized and the algorithms usually have problems with background noises.

Hybrid coders exploit the advantages of both techniques as these encode speech in such a way which results in a low data rate while keeping the speech intelligible. Typical bandwidth requirements lie between 4.8 and 16 kbps. The basic problem with vocoders is simplistic representation of the excitation signal i.e. the signal is considered to be either voiced or unvoiced. The coders discussed below improve the representation of the excitation signal, each in their own way.

3.5.3.1 Residual Excited Linear Prediction (RELP)

The RELP coder works in almost the same way as the LPC coder. To analyze the signal the parameters for the vocal tract filter are determined and the inverse of the resulting filter is applied to the signal. This gives the residual signal.

Then the LPC coder checked if the signal was voiced or unvoiced and used it to model an excitation signal. In the RELP coder however the residual is not analyzed any further but used directly as the excitation for speech synthesis. The residual is compressed using waveform coding techniques to lower the bandwidth requirements. RELP coders can allow good speech quality at bit rates in the region of 9.6 kbps.

3.5.3.2 Codebook Excited Linear Prediction (CELP)

The CELP coder overcomes the synthetic sound of vocoders by allowing a wide variety of excitation signals which all are captured in the CELP codebook. To determine which excitation signal is to be used the coder performs an exhaustive search. For each entry in the codebook the resulting speech signal is synthesized and the entry that created the smallest error is then chosen. The excitation signal is then encoded by the index of the corresponding entry. So basically, the coder uses Vector Quantization to encode the excitation signal.

This technique is called an analysis-by-synthesis (ABS) technique because it analyses a signal by synthesizing several possibilities and choosing the one that caused the least amount of error. This exhaustive search is computationally very expensive. However fast algorithms have been developed to perform the search in real-time. CELP techniques allow bit rates of even 4.8 kbps.

3.5.3.3 Multipulse and Regular Pulse Excited coding (MPE and RPE)

Like the previous method MPE and RPE techniques improve the speech quality by giving a better representation of the excitation signal. With MPE the excitation signal is modeled as a series of pulses each with its own amplitude. The positions and amplitudes of the pulses are determined by an ABS procedure. The MPE method can produce high quality speech at rates around 9.6 kbps. The RPE technique works in a similar fashion only the pulses are regularly spaced as the name suggests. The GSM mobile telephone system uses a RPE variant that operates at approximately 13 kbps.

3.5.4 NEURAL NETWORK

The compression principles discussed above cover pretty much the whole speech compression domain. There is one more technique as neural networks for speech compression. There is not much information about this particular use of neural networks but there are documents that describe how neural networks can be used for lossy image compression. It is possible that similar techniques can be used for the compression of speech, [10].

To do this there are several ways in which artificial neural networks can be used. A neural net can be trained to predict the next sample from a given a number of previous samples. This way the network can perform the predictive function in differential coding schemes. If this prediction has been done more accurately than regular predictive techniques this would result in better compression.

3.6 DELAYS BY COMPRESSION

To preserve good communication quality the overall delay has to be kept as low as possible. This means to take the delay caused by compression and decompression into account, even if the signal is compressed in an in an excellent way. It has little use for real-time communication if it introduces an unacceptable amount of delay.

Delays during the compression stage can generally be divided into two categories. First of all there is always some delay due to the calculations that need to be done. This amount of delay depends much on the capabilities of the system performing the compression.

Some compression techniques introduce a second type of delay in order to compress a part of the speech signal. The amount of `look ahead' needed determines the amount of delay introduced. For a specific algorithm this delay is fixed and does not vary among systems. Decompressing the signal can usually be done much faster than compressing it. With faster computers and availability of specialized hardware the fixed delay during the compression stage is probably the most important to consider.

3.7 COMPRESSION STANDARDS FOR VARIOUS TYPES

Following table illustrates the complete picture of different types of ITU standards for various transmission systems in the world. It clearly describes the range of different series as well: -

Table 3.1 ITU G-SERIES RECOMMENDATIONS

INTERNATIONAL TELEPHONE CONNECTIONS AND CIRCUITS	G.100–G.199
INTERNATIONAL ANALOGUE CARRIER SYSTEM	
GENERAL CHARACTERISTICS COMMON TO ALL ANALOGUE CARRIER-TRANSMISSION SYSTEMS	G.200–G.299
INDIVIDUAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON METALLIC LINES	G.300–G.399
GENERAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON RADIO-RELAY OR SATELLITE LINKS AND INTERCONNECTION WITH METALLIC LINES	G.400–G.449
COORDINATION OF RADIOTELEPHONY AND LINE TELEPHONY	G.450–G.499
TRANSMISSION MEDIA CHARACTERISTICS	G.600–G.699
DIGITAL TRANSMISSION SYSTEMS	
TERMINAL EQUIPMENTS	G.700–G.799
General	G.700–G.709
Coding of analogue signals by pulse code modulation	G.710–G.719

Coding of analogue signals by methods other than PCM	**G.720–G.729**
Principal characteristics of primary multiplex equipment	G.730–G.739
Principal characteristics of second order multiplex equipment	G.740–G.749
Principal characteristics of higher order multiplex equipment	G.750–G.759
Principal characteristics of trans coder and digital multiplication equipment	G.760–G.769
Operations, administration and maintenance features of transmission equipment	G.770–G.779
Principal characteristics of multiplexing equipment for the synchronous digital hierarchy	G.780–G.789
Other terminal equipment	G.790–G.799
Digital Networks	G.800–G.899
General aspects	G.800–G.809
Design objectives for digital networks	G.810–G.819
Quality and availability targets	G.820–G.829
Network capabilities and functions	G.830–G.839
SDH network characteristics	G.840–G.899
DIGITAL SECTIONS AND DIGITAL LINE SYSTEM	G.900–G.999
General	G.900–G.909

Parameters for optical fibre cable systems	G.910–G.919
Digital sections at hierarchical bit rates based on a bit rate of 2048 kbit/s	G.920–G.929
Digital line transmission systems on cable at non-hierarchical bit rates	G.930–G.939
Digital line systems provided by FDM transmission bearers	G.940–G.949
Digital line systems	G.950–G.959
Digital section and digital transmission systems for customer access to ISDN	G.960–G.969
Optical fiber submarine cable systems	G.970–G.979
Optical line systems for local and access networks	G.980–G.999

Table 3.2 VOICE COMPRESSION STANDARDS

Standard	Description	Bit rate	MOS
G.711	Pulse Code Modulation using eight bits per sample, sampling at 8000 Hz	64 kbps	4.3

G.723.1	ech coder designed with low bit rate video mind [41]. The G.723.1 coder needs a 7.5 ms d used one of these coding schemes: • Multipulse Maximum Likelihood Quantization (MP-MLQ) • Algebraic CELP (ACELP)	6.3 and 5.3 kbps respectively	4.1
G.726	Coder using ADPCM. Contains obsolete standards G.721 and G.723	16,24,32 and 40 kbps	2-4.3
G.727	Five, four, three and two bits per sample embedded ADPCM. The encoding allows bit reductions at any point in the network without the need for coordination between sender and receiver [10].	16,24,32 and 40 kbps	2-4.3
G.728	Low Delay CELP (LD-CELP)	16 kbps	4.1
G.729	Conjugate Structure ACELP (CS-ACELP) • Annex A: Reduced complexity algorithm • Annex D: Low rate extension • Annex E: High rate extension These coders need a 5 ms lookahead.	8 kbps (CS-ACELP), 8 kbps (Annex A), 6.4 kbps (Annex D) and 11.8 kbps (Annex E)	4.1 (CS-ACELP) and 3.7 (Annex A)

GSM 06.10	Full rate speech transcoding using Regular Pulse Excitation-Long Term Prediction (RPE-LTP)	13 kbps	3.71
GSM 06.20	Half rate speech transcoding using Vector Sum Excited Linear Prediction (VSELP)	5.6 kbps	3.85
GSM 06.60	Enhanced full rate speech transcoding using ACELP	12.2 kbps	4.43

CHAPTER 4

ADAPTIVE DELTA PULSE CODE MODULATION

4.1 INTRODUCTION

ADPCM is more efficient way of storing waveforms than 16-bit or 8-bit PCM. It only uses 4 bits per sample taking up a quarter of the disk space of 16-bit PCM. However the sound quality is inferior. Since the windows sound system hardware only understands 8/16-bit PCM so the computer must compress and decompress the ADPCM into/from PCM form. For ADPCM the computer must have the Audio Compression Manager (ACM) installed. ADPCM stores the value differences between the two adjacent PCM samples and makes some assumption that allows data reduction. Because of these assumptions low frequencies are properly reproduced but any high frequency tends to get distorted. The distortion is easily audible in 11 kHz ADPCM files but becomes more difficult to discern with higher sampling rates and is virtually impossible to recognize with 44 kHz ADPCM files.

4.2 ADPCM

ADPCM stands for Adaptive Differential Pulse Code Modulation. This coding technique has been standardized under ITU G.721. This form of speech coding is similar to PCM. ADPCM is able to provide good quality speech for bit rates of 32 Kbits/s (MOS of 3.8). ADPCM has been standardized for bit rates of 16, 24, 32 and 40 Kbits/s.

4.3 PCM VS ADPCM

ADPCM algorithm is different from PCM because instead of quantizing the sampled voice data the difference between the quantized voice data and the predicted speech signal is quantized. With good prediction the difference between the actual voice data and the predicted data becomes small. This results in fewer bits required to code the quantized data. The adaptive quantization used is not uniform it can be changed to accommodate other sound characteristics besides the voice.

4.4 DECODING

Adding the quantized difference to the predicted speech signal decodes the signal. The combination of these signals allows the original speech data to be recovered.

The decoder function is to reconstruct the signal from the quantized bit stream. Although the overall compression ratio of ADPCM is only 2:1 at 32 Kbits/s but it can be used in conjunction with DSI (Digital Speech Interpolation) to achieve compression ratio of 4:1.

4.5 DIGITAL SPEECH INTERPOLATION (DSI)

Human speech is effectively half-duplex i.e. while one party is speaking the other is listening. Another property of a typical dialogue is the presence of irregular silence. From a statistical point of view the average usage of a voice channel is only 50%. A form of compression can be performed using these above properties. The use of Digital Speech

Interpolation enables the compression to 4:1. This is not actually "true" compression meaning that the actual bit rate of the voice channel is not reduced. The compression is actually done by multiplexing separate channels to a single output channel.

Determining that whether the channel is active or inactive speech interpolation is carried out. This is performed by setting a power threshold. The channel is active once the power is above the threshold and inactive once it is bellow the threshold. Since the threshold setting determines that whether the channel is active or not, it affects the performance of the overall compression.. In the event where high voice activity is encountered, the variable rate ADPCM (coupled with DSI) prevents the overloading of the channel by reducing the output bit rates.

Figure 4.1 Digital Speech Interpolations

4.6 ITU G.711

The mu-law is an international standard telephone encoding format and also known as ITU standard G.711. It packs each 16-bits sample into 8 bits by using a logarithmic table for encoding. The mu-law coding is a form of compression for audio signals including speech and is widely used in the telecommunication field because it improves the signals to noise ratio without increasing the amount of data typically. It is a companding technique i.e. it carries more information about the smaller signals than about the larger signals. There is a slight variation called A-law being used in European telephone systems that carries more information about the larger signals than about smaller signals.

4.7 ITU G.721

ADPCM encoders with pole zero decoder filters have proved to be versatile in speech applications. In fact the ADPCM 32 kbits/s algorithm for the G.721 CCITT standard uses a pole zero adaptive predictor. The algorithm consists of an adaptive quantizer and an adaptive pole zero predictor. The pole zero predictor estimates the input signal and the quantizer encodes it into a sequence of 4bits word. The performance of the coder is > 4 in terms of MOS.

4.8 ITU G.722

ITU G.722 standard for voice encoding in ISDN teleconferencing is based on two band separation and ADPCM. The low frequency sub-band is quantized at 48 kbits/s while the high frequency sub-band is quantized at 16 kbits/s. Lower bit rate is achieved by quantizing the lower frequency sub-band at 40 kbits/s or 32 kbits/s and it uses SB-ADPCM technique. The G.722 can be used for compression of speech and other audio signals at toll quality and can also be used as a part of H.320 and other high quality applications.

4.9 ITU G.723

It is a Dual rate speech coder designed with low bit rate video telephony. The G.723.1 coder needs a 7.5 ms look ahead and uses one of the coding schemes as Multi pulse Maximum Likelihood Quantization (MP-MLQ) or Algebraic CELP (ACELP) with a bit rate of 6.3 and 5.3 kbps respectively. The G.723 can be used for compression of speech and other audio signals being used in multimedia applications at a very low bit rate and also as a part of H.324 / H.323 or VON (voice over network) applications.

CHAPTER 5

PROJECT IMPLEMENTATION

5.1 GENERAL

The coders of CCITT (International Telegraph and Telephone Consultative Committee) for G.711, G.721, G.722 and G.723 for voice compression were implemented and tested on Sun Solaris Stations. Sun Microsystems support the CCITT audio formats in Solaris 2.0 system software. The source coders for CCITT conversion routines are:

g711.c ………………………………..………CCITT G.711 u-law and A-law compression

g72x.c ………………………Common denominator of G.721 and G.723 ADPCM codes

g721.c ………………………………CCITT G.721 32Kbps ADPCM coder (with g72x.c)

g723_24.c ………………………….CCITT G.723 24Kbps ADPCM coder (with g72x.c)

g723_40c …………………………...CCITT G.723 40Kbps ADPCM coder (with g72x.c)

encode.c ……………………………………………………………CCITT ADPCM encoder

decode.c ……………………………………………………………CCITT ADPCM decoder

The software calls various standard compression routines and packs/unpacks the bits. The sample programs contain examples of how to call the various compression routines and pack/unpack the bits. These programs read bytes streams from stdin (standard in) and write to stdout (standard out) as per the prescribed method below. The input/output data is raw data that means no file header or other identification information is embedded with the data. The sample programs are invoked as follows:

Encode [-3 | 4 | 5] [-a | u | l] < infile > outfile

Decode [-3 | 4 | 5] [-a | u | l] < infile > outfile

Where:

-3…………………. encode to (decode from) G.723 24kbps (3-bytes) data

-4…………………. encode to (decode from) G.721 32kbps (4-bytes) data

-5…………………. encode to (decode from) G.723 40kbps (5-bytes) data

-a…………………. encode from (decode to) A-law data

-u…………………. encode from (decode to) u-law data [the default]

-l………………….. encode from (decode to) 16-bit linear data

Examples:

Read 16-bit linear and output G.721

encode -4 -l < pcmfile > g721file

Read 40Kbps G.723 and output A-law

decode -5 -a < g723file > alawfile

Compress and then decompress u-law data using 24Kbps G.723

encode -3 < ulawin | deoced -3 > ulawout

5.2 VARIOUS EXAMPLES

An audio wave file named as SUN was recorded on sun solaris system in order to test/confirm the implementation of compression and decompression i.e. encoding and decoding, through all the prescribed techniques. Encoders are used to encode the audio file to specified format and then again decoded to check back that the original speech samples are recovered. Similarly a wave file named as PC was made in the Microsoft environment and implemented for all the prescribed techniques to analyze the coding and decoding techniques. All the possible combinations for compression and decompression were tested. The coded file was given the name with double letters as aa, bb, gg etc and the decoded file was given the name as combination of letter and figure as a1, d1, h1 etc. Table below describes all the encoded and decoded audio file formats along with their specified names. However if the codes are typed or entered wrongly or other than the prescribed method then the result will be garbage and the file will not be decoded or encoded.

Table5.1 Coders Implementation (Sun Wave File)

DESCRIPTION	ENCODING	DECODING
G.723 (24) TO A-LAW	- 3 – a < sun > jj	- 3 – a <jj > j1
G.723 (24) TO MU-LAW	- 3 – u < sun > kk	- 3 – u < kk > k1
G.723 (24) TO LINEAR	- 3 – l < sun > ll	- 3 – l < ll > l1
G.721 TO A-LAW	-4 – a < sun > mm	-4 – a < mm >m1
G.721 TO MU-LAW	-4 – u < sun > nn	-4 – u < nn > n1
G.721 TO LINEAR	- 4 – l < sun > oo	- 4 – l < oo > o1
G.723 (40) TO A-LAW	-5 – a < sun > pp	-5 – a < pp > p1
G.723 (40) TO MU-LAW	- 5 –u < sun > qq	- 5 –u < qq > q1
G.723 (40) TO LINEAR	- 5 –l < sun > r r	- 5 –l < r r > r1

Table5.2 Coders Implementation (Pc Wave File)

DESCRIPTION	ENCODING	DECODING
G.723 (24) TO A-LAW	- 3 – a < pc > aa	- 3 – a <aa > a1
G.723 (24) TO MU-LAW	- 3 – u < pc > bb	- 3 – u < bb > b1
G.723 (24) TO LINEAR	- 3 – l < pc > cc	- 3 – l < cc > c1
G.721 TO A-LAW	-4 – a < pc > dd	-4 – a < dd > d1
G.721 TO MU-LAW	-4 – u < pc > ee	-4 – u < ee > e1
G.721 TO LINEAR	- 4 – l < pc > ff	- 4 – l < ff > f1
G.723 (40) TO A-LAW	-5 – a < pc > gg	-5 – a < gg > g1
G.723 (40) TO MU-LAW	- 5 –u < pc > hh	- 5 –u < hh > h1
G.723 (40) TO LINEAR	- 5 –l < pc > ii	- 5 –l < ii > i1

5.3 MATLAB IMPLEMENTATION

To carry out the comparison between compressed and decompressed files of speeches Matlab was used. Amplitude, frequency, time scale and energy plot before and after the compression were compared and analyzed in Matlab. A directory named as VSAT (voice signal analysis tool) consisting of the MATLAB routines was made for handling, analyzing and displaying the different characteristics of speech wave files. Any audio file can be selected and processed by reading the extensions of the file.

Wave files can be compared basing on different characteristics to analyze the response before and after the compression being done. This includes Time Wave Form, which is used to display amplitude and time period for selected wave file in graphical shape. It provides a comparison between the compressed and decompressed audio wave files. Energy Plot option is used to display energy contours computed at every 20 msec intervals and expressed in db whereas the Power Spectral Density option is used to show an estimate of the power spectral density for each specified wave file. Following figures illustrate these characteristics:

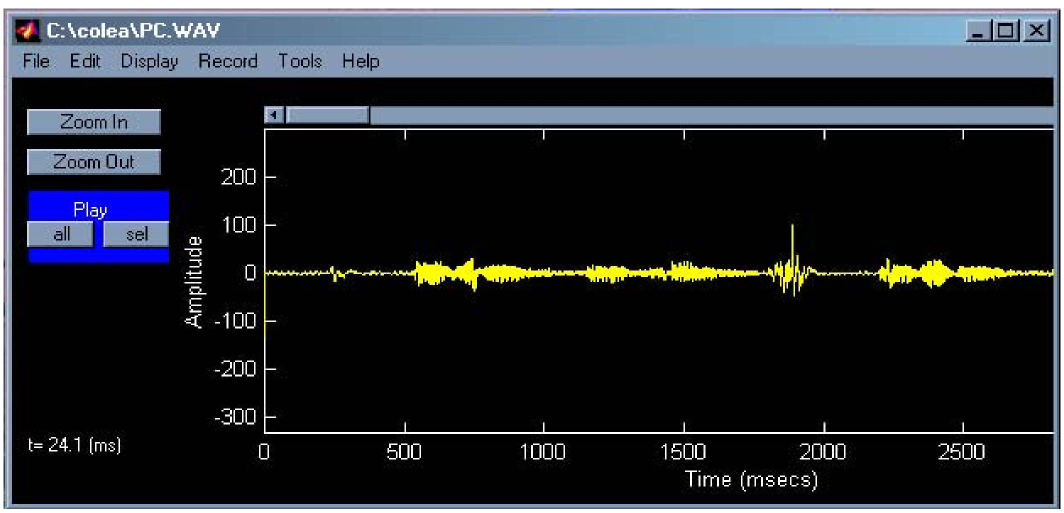


Figure 5.1Time Wave Form

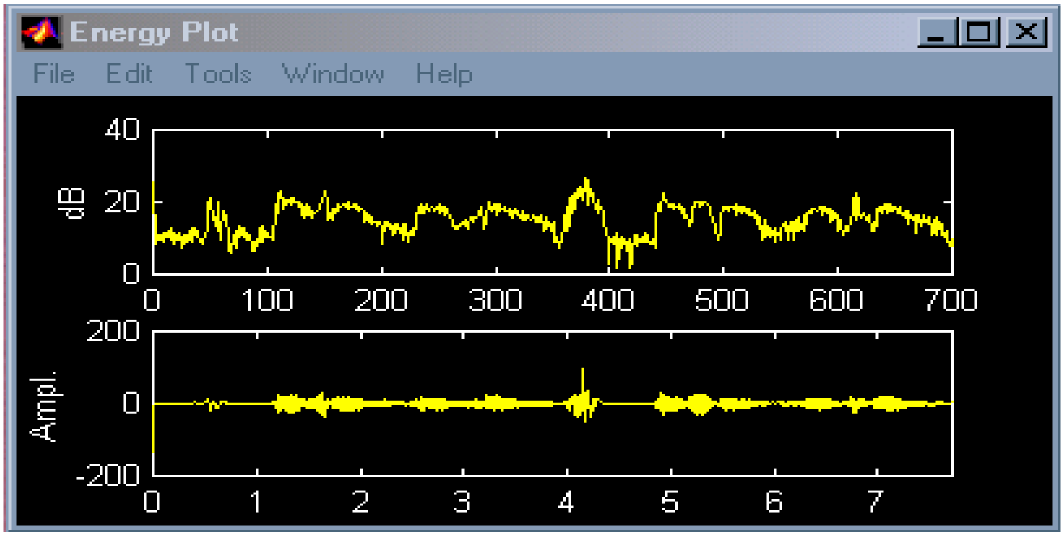


Figure 5.2 Energy Plot

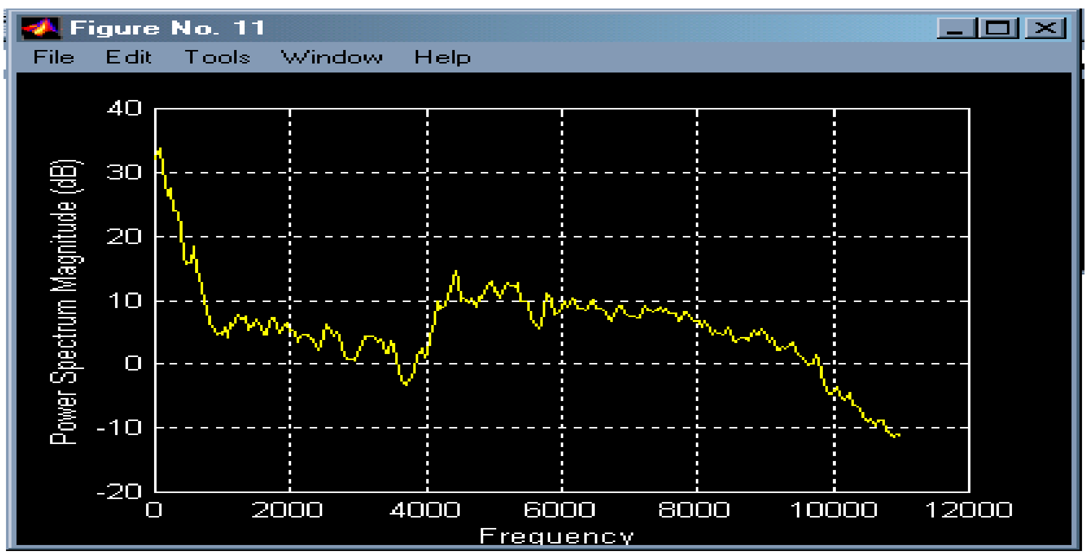


Figure5.3 Power Spectral Diagram

5.4 COMPARISONS ILLUSTRATIONS

Comparison between the coded and decoded speech files as already mentioned in the previous tables can be carried out to analyze the difference in their formats and amplitudes before and after the compression was carried out. Few of the files are illustrated below:

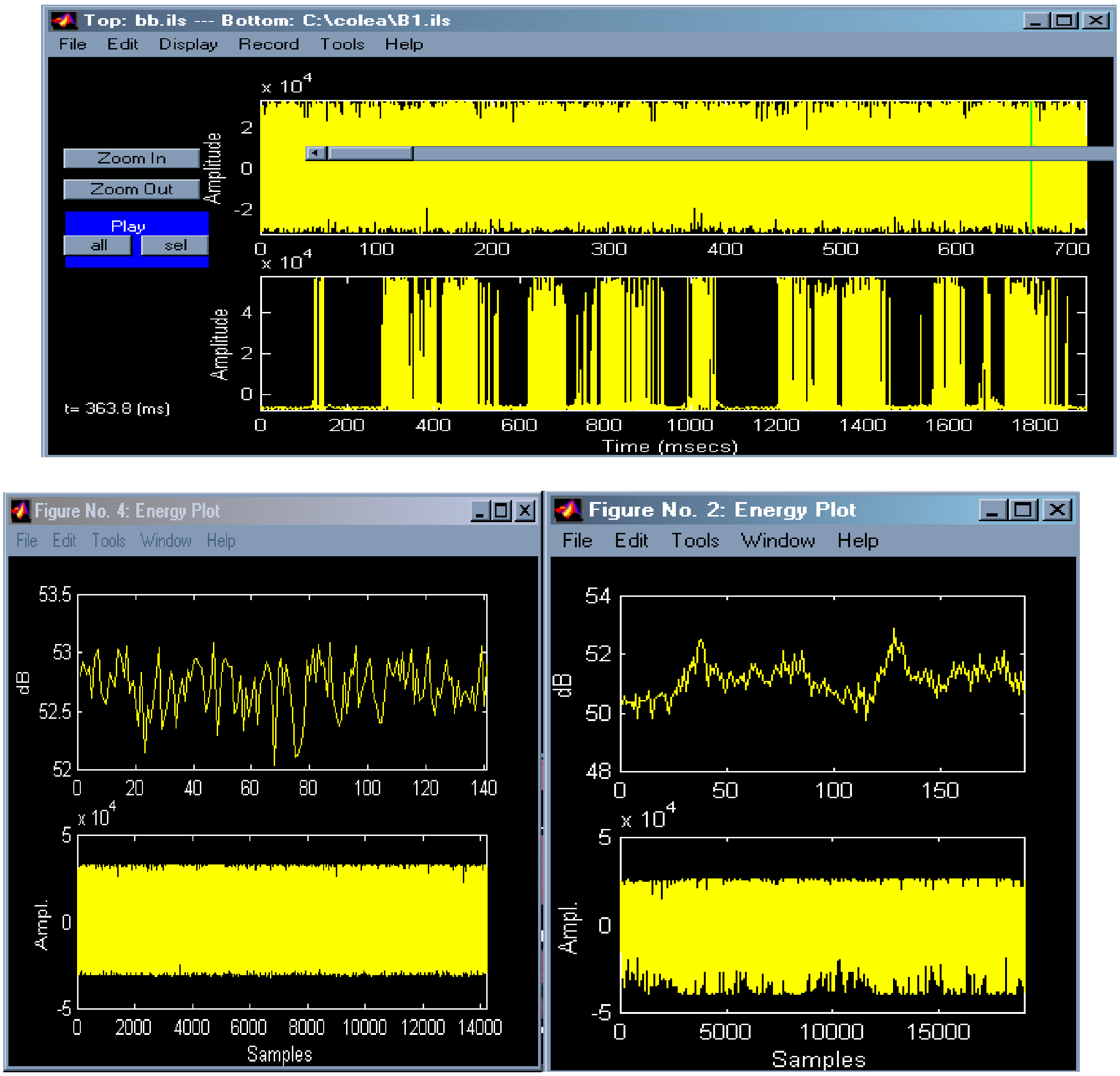


Figure5.4 Comparison Of bb And B1 Speech Files

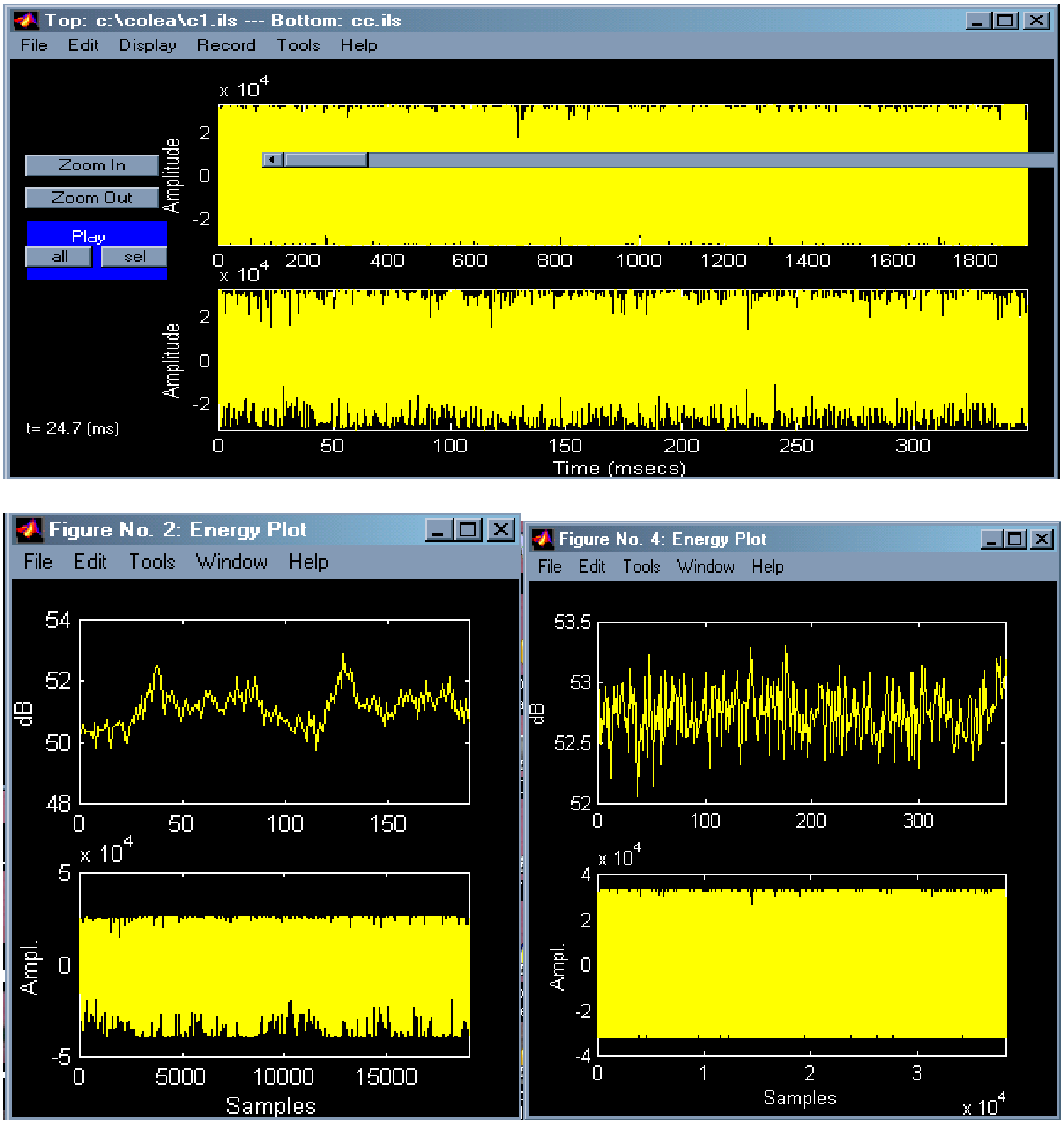


Figure 5.5 Comparison Of cc And c1 Speech Files

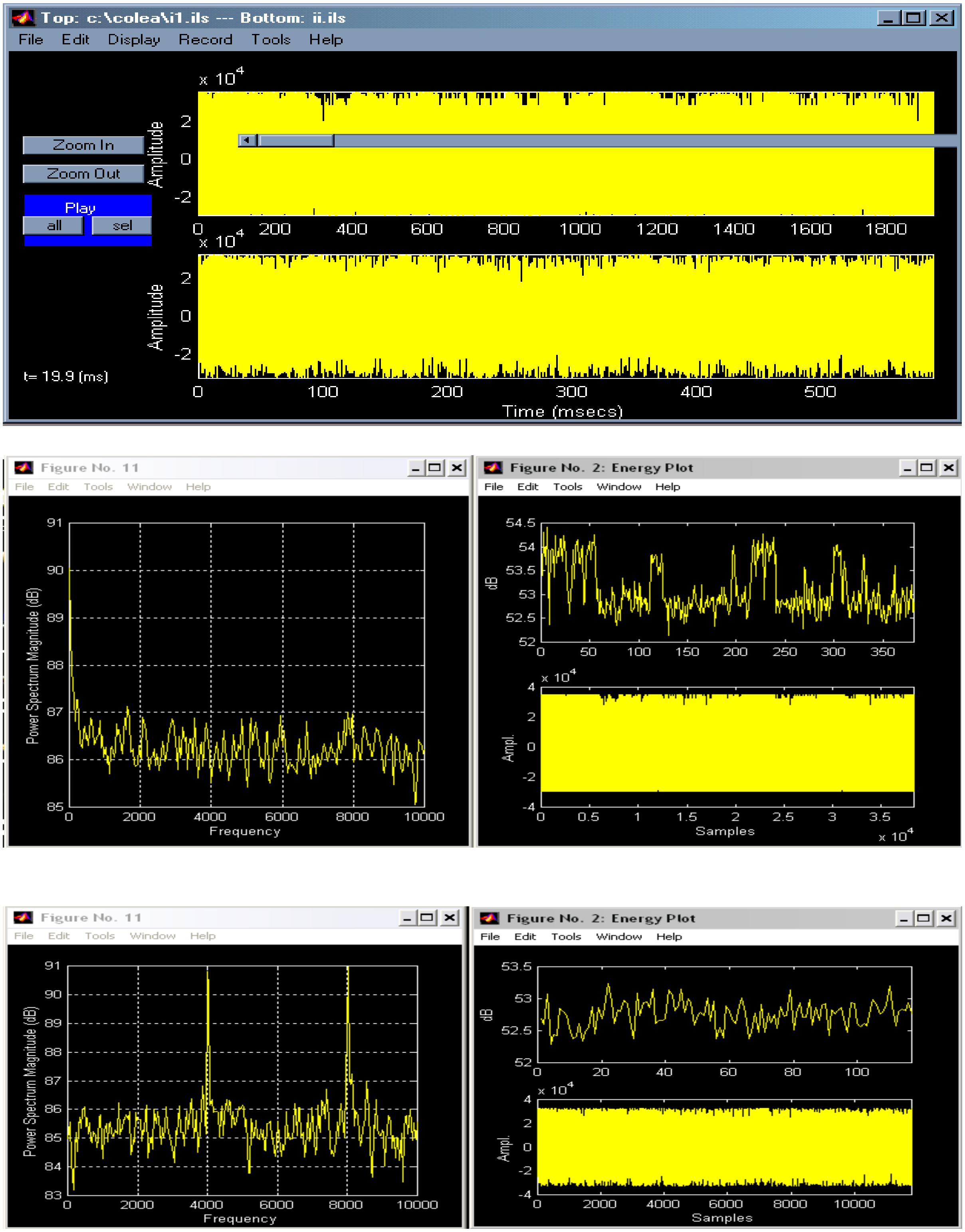


Figure5.6 Comparison Of ii And i1 Speech Files

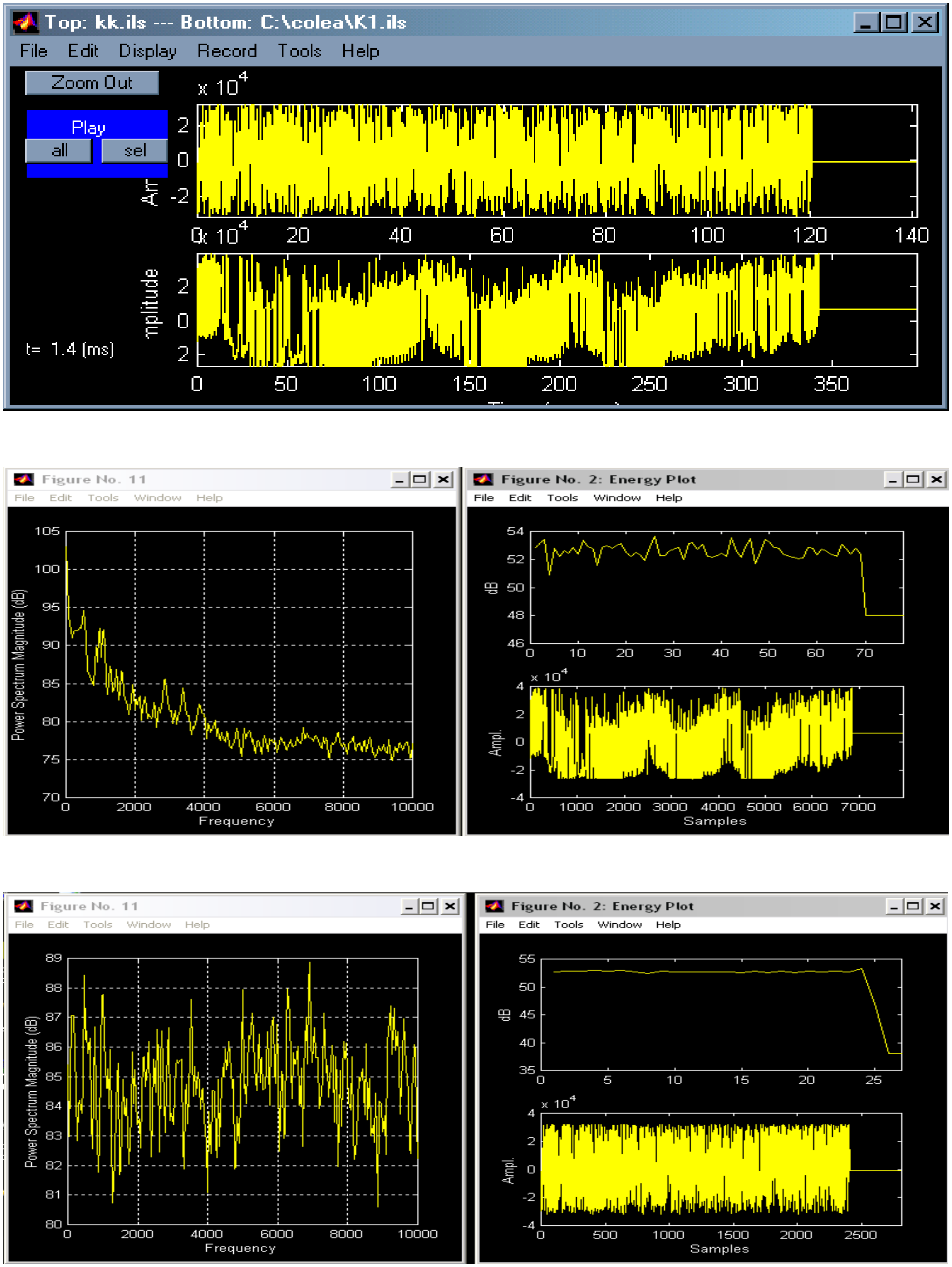


Figure5.7 Comparison Of kk And k1 Speech Files

5.5 FUTURE ENHANCEMENT

Apart from the objectives of this project still there is a room for the improvement in the present features. This project can be enhanced and updated by considering following aspects: -

- Encryption and decryption algorithms / techniques can be added with the coders for coding / decoding the crypted samples of speeches.
- Routines for conversion of raw to wave speech files and vice versa can be incorporated for efficient and enhanced applications.
- Coders for other compression techniques like CELP, RELP, MELP and LPC etc can be combined to increase the application area.

5.6 CONCLUSION

There is no end to the optimization and research in any of the system and in any of the field. Speech and audio compression techniques have advanced rapidly in recent years. The algorithms exploit models of speech production and auditory perception and offer a quality versus bit rate tradeoff that significantly exceeds most prior compression techniques. Nevertheless, there are considerable difficulties to surmount before these low rate coders can become telecommunication standards.

Speech and audio compressions are indeed very active areas for research and development and generally require a high level of specialization, which combines the strength in digital processing with a good understanding of human psychophysics and modern quantization methods.

BIBLIOGRAPHY

[1] N. Morgan And B. Gold, Speech And Audio Signal Processing

[2] M. R. Schroeder, Computer Speech: Recognition, Compression, Synthesis

[3] B. S. Atal, V. Cuperman, And A. Gersho, Advances In Speech Coding.

[4] Michael P. Chin, Karl Home. Incorporating Voice Telecommunication

[5] Spanias, A. S. Speech Coding: A Tutorial Review.

[6] Http://Www.Eas.Asu.Edu/~Spanias/Papers/Review.Ps

[7] Glenn K. Bindley. Voice Compression Compatibility And Deployment Issues.

[8] A. Gersho And R. M. Gray, Vector Quantization And Signal Compression.

[9] M. R. Schroeder, Computer Speech: Recognition, Compression, Synthesis.

[10] B. Atal And M. Schroeder, "*Stochastic Coding Of Speech At Very Low Rates*",

[11] D. W. Griffin And J. S. Lim, "*Multiband Excitation Vocoder*",

[12] T.Robinson, "Speech Analysis", Cambridge Speech Vision Robotics Group,

[13] Http://Theories Of Speech Compression.Htm

[14] Http://Voice Communication.Htm