Wideband Speech Codec Implementation

Radha Krishnan Swamynathan, 300183918
Department of Electrical and Computer Engineering
University of Ottawa

ABSTRACT

Speech compression primarily aims to remove the redundancy in the digital representation of speech to reduce transmission bandwidth and storage space. Speech Codecs have evolved from operating in Narrowband (200 – 3400 Hz) to Wideband (50 – 7000 Hz) mode of operation. There are multiple aspects into designing a speech codec that works robustly across various modes of communication such as wired telephony, wireless telephony, voice over IP (VoIP) etc. Parameter based encoding of speech signals has significance in speech compression, as a parameterized human vocal model is used which is available at both the transmitter and the receiver. The encoder analyzes the input speech signal and extracts the model parameters which are transmitted to the decoder, which synthesizes the voice from the model and the received parameters. This project focuses on using these parameter-based coding techniques for wideband speech codec implementation, which replaced the narrowband speech services in both wireless and wireline telephone services, because of the superior speech quality and naturalness due the wideband mode of operation. The codec consists of subdivision of the speech signal intro frames, extraction of the model parameters (Linear Prediction model and Voice/Code Excited Linear Prediction of parameters) of each speech frame, which are quantized and transmitted to the receiver. Such a codec was standardized by ITU-T in G.722.2, can operate in bitrates ranging from 6.6 to 23.85 kb/s, hence referred to as Adaptive Multirate Wideband (AMR-WB) codec.

1. INTRODUCTION

Speech coding can be described as lossy mode of coding which involves transformation of the speech signals to a compact representation which can be transmitted with considerably smaller bandwidth consumption and/or reduced storage costs. Speech codecs have become an important component in the infrastructure of multiple devices, due to the multitude of applications requiring speech codecs. Compression is an important factor as the data bandwidth might vary over time. Human speech or vocal signals has the average frequency range of 500 – 2000 Hertz, and the human auditory frequency range is 20 – 20000 Hertz (Considering the average frequency range across various age groups).

Given the continuous or analog nature of the speech, the digital representation of requires sampling and quantization done by either 8-bit or 16-bit quantizers. Digitized speech follows the Nyquist theorem which requires the sampling rate to be twice the highest frequency of the signals. In this project's context focuses on speech signals which are already sampled and quantized (Digitized speech). In terms of compression, speech coding is the process of creating a minimally redundant representation of the speech signal with the best possible perceptual quality. There are different methods to compress the digitized speech like modeling the speech, exploiting the redundancy, neglecting the irrelevant information etc., while preserving the quality of speech. The decoded speech should ideally be indistinguishable from the original signal. Typically, the speech codecs are measured by the bitrate of the compressed speech, reduced bit-rate reduction leads to power reduction of the transmitted signal, making it less immune to noise. To explore speech compression, a minimal knowledge on human auditory and vocal system would be useful.

1.1 Human Speech Model

Human speech model could be considered as a Physical model, wherein the physics of the source output are mathematically approximated. Speech signals are produced by forcing air through an elastic opening, the vocal cords and then through cylindrical tubes with nonuniform diameter (larynx and pharynx passages), and finally through nasal and mouth cavities with changing boundaries [1]. As a person speaks, the air from the lungs push out through the vocal tract (larynx and pharynx) and out of the mouth to produce a sound. The vibration of vocal cords could be open or closed, correspondingly producing unvoiced and voiced signals. This model is shown in figure 1. The variation in output (voiced and unvoiced)sound is relatively slow due to the physics, thus could be thought of as a Quasi-Stationary Process over 10-100ms interval, referred to as framing (regardless of the speaker) [2].

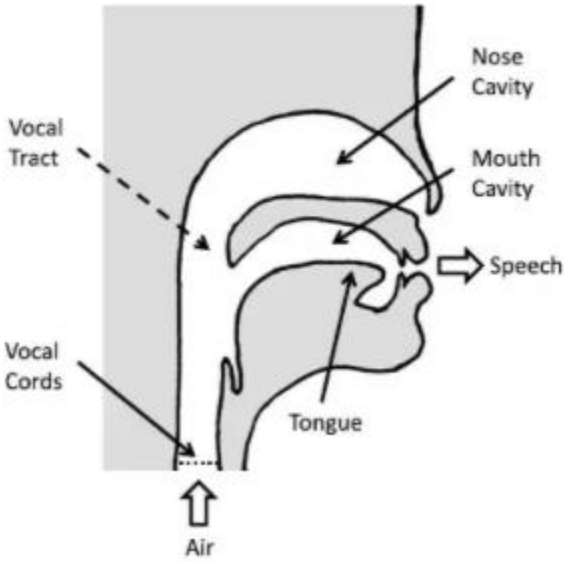


Figure 1. Mechanics of Human Speech Production [2]

Voiced speech segments or frames usually have repetitive patterns which are visible in the spectrum, wherein a dominant frequency value (pitch) and its corresponding harmonic frequencies are present. The unvoiced segments resemble noise signals with no repeatable patterns. The voiced speech sounds are usually vowel sounds (oscillatory pronunciations resulting in repeatable patterns). Unvoiced speech are the consonant sounds (fricatives and plosives). Figure 2 shows sample voiced signal (a) and unvoiced signal (b). Given the short-term stationary nature of the speech signal (frames), most of the compression techniques work on the frame-level, hence referred to as frame-based speech coding.

1.2 Speech Coding Techniques

Speech compression is a special case of audio compression, wherein the codecs can be categorized according to the bandwidth occupied by the input and the reproduced source as Narrowband coding, Wideband coding and Fullband coding. Narrowband or telephone bandwidth codecs operate in 200 – 3400 Hertz bandwidth, Wideband coding operates in 50 – 7000 Hertz bandwidth and the Fullband codecs operate in 20 – 20000 Hertz bandwidth and are typically used for Audio Compression. Speech compression can be efficiently done in narrowband and wideband modes of operation as the human speech signals do not have high frequency values.

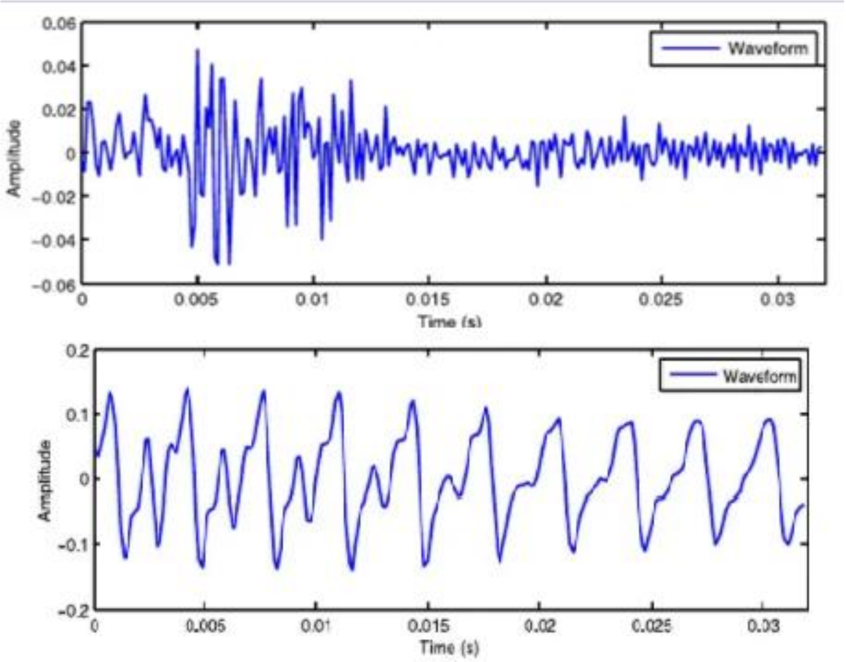


Figure 2. (a) Voiced speech and (b) Unvoiced speech

The most common approaches used by narrowband and wideband speech coding are waveform-based compression, parameter-based compression and analysis-by-synthesis compression.

1.2.1 Waveform based Compression: Waveform based compression techniques usually attempt to reproduce the speech wave (in time domain) as accurately as possible. In mathematical terms, the error between the reconstructed signal and the original signals needs to be minimized , which can be achieved quantizing the sampled signal using smaller quantization steps. Historically, the most famous waveform codec is the Pulse Code Modulation (PCM). The reference for any speech codec discussed hereafter is the PCM codec which operates at 64 kb/s [2]. An improved version of PCM, Adaptive Differential PCM which was used in traditional wired telephony, uses non-uniform quantization (Jayant Quantizer) to reduce the bitrate to 32 kb/s if each sample is coded with 4bits or 16kb/s if 2 bits/sample [3].

1.2.2 Parametric Compression: Waveform based compression techniques can achieve a maximum compression ratio of 4 (PCM is the raw channel rate for digital speech codecs). Another approach for achieving better compression is establishing the physical model of speech described in the section 1.1, and transmitting the parameters of the model, instead of the whole signal resulting in better compression. This approach is referred as Parametric Compression or Vocoders. The Linear Prediction Coding (LPC) proposed by Atal in 1971 at Bell Labs as the base for all the speech codecs available thereafter. LPC based codecs can achieve bitrates of 8 kb/s which can be further reduced to 1.2 kb/s resulting in compression ratio of [8 to 60). Given the achievement of high compression ratio, these codecs are generally not used due to the deteriorated sound quality and mechanical or robotic sounding output (no naturalness of speech).

1.2.3 Analysis-Synthesis Compression: LPC parametric coding has several problems associated with and Waveform coding has smaller compression ratio. Hybrid codecs or Analysis-Synthesis codecs combine the required characteristics of both the approaches to create codecs with better quality-bit rate trade off, shown in figure 3. The most successful hybrid compression technique is Code-Excited Linear Prediction and its variants.

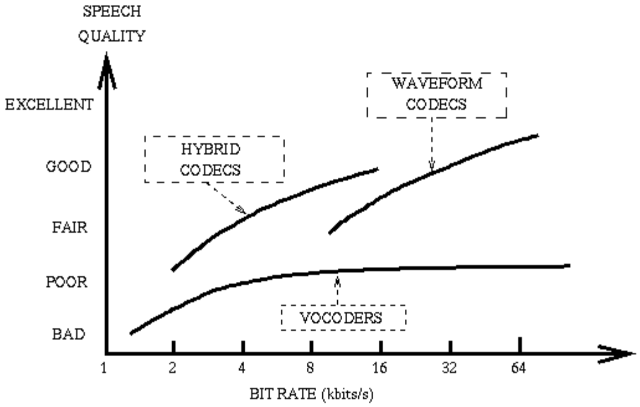


Figure 3. Quality-Bitrate tradeoff of speech coding techniques [4]

This project report focuses on compression of wideband speech signals by extracting LPC parameters (LPC Vocoder) and Analysis-Synthesis approach (CELP). The input speech is sampled at 16 KHz and divided into frames. The framed input speech is pre-processed, and the Linear Prediction Coefficients are extracted for every frame. Such a wideband codec was first standardized in 2002 by ITU-T (International Telecommunications Union – Telecommunication sector) as the recommendation G.722.2. The G.722.2 wideband speech codec has six different operation rates 23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85 and 6.6 kb/s, with the last two modes intended to be used only during severe channel conditions or network congestion [3].

2. CODEC DESCRIPTION AND METHODOLOGY

2.1 Mathematical model for Linear Prediction Coding and its variants

Given the speech production discussed in 1.1, a mathematical model of it is shown in figure 4. The speech signal “x(n)” is switched between voiced signal which contains pulse trains controlled by pitch coefficients, and unvoiced signal which simply is white noise. The signal generated (x(n)) is amplified by a gain parameter G and then sent to vocal tract filter or LPC filter. The output signal is established as:

$$y_n = \sum_{i=1}^{K} a_i y_{n-i} + G x_n$$

Where y_n is the output of the nth frame, which is given the weighted sum of previous K number outputs and the nth frame input x_n multiplied by the gain value G. The coefficients a_i are the linear prediction coefficients. The principle behind LP coefficients a_i it to minimize the sum of the squared differences between the original signal and the estimated signal over a finite duration i.e., a frame.

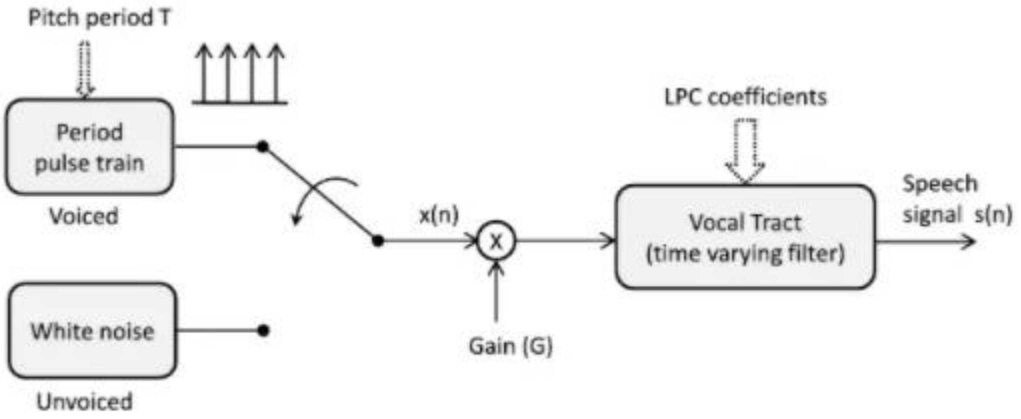


Figure 4. Mathematical model for speech production [2]

The LPC filter or Vocal tract filter is a linear all pole filter (IIR), with the transfer function:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_k z^{-k}} \quad \rightarrow \text{Equation 1}$$

Where p is the number of poles a_p are the parameters that determine the poles and G is the filter gain. To estimate the LP parameters for the frames for which the gain is usually ignored to allow the parameterization to be independent of the intensity of the signal. The idea here is given a speech frame at time n, x(n), can be approximated as a linear combination of past p speech samples:

$x(n) = \sum_{k=1}^{p} \alpha_k x(n-k)$ wherein α_k are the prediction coefficients. The prediction error is evaluated as:

$$e(n) = x(n) - \sum_{k=1}^{p} \alpha_k x(n-k)$$

The approach to find the coefficients $\alpha_1 - \alpha_p$ is to minimize the mean-square prediction error over the complete signal (all the frames).

$$E = \sum_{n} e(n)^2 = \sum_{n} \left(x(n) - \sum_{k=1}^{P} \alpha_k x(n-k) \right)^2$$

Evaluation of p (number of coefficients) parameters ($\alpha_1 - \alpha_p$) over n equations is a daunting task for which there are multiple approaches which on few minor assumptions. Two commonly used methods to solve for the coefficients are the autocorrelation method and the covariance method. In this project's context, autocorrelation method is used, as the roots of the polynomial in the denominator of the above equation 1 is always guaranteed to be inside the unit circle, hence guaranteeing the stability of the system H(z). Levinson – Durbin recursion will be utilized to compute the required parameters of the autocorrelation method [3]. The codec implementations discussed below will follow the model described in this section as the base for speech analysis and synthesis.

2.2 Plain LPC Codec

The block diagram of a LPC codec, which is commonly referred to as LPC vocoder is shown below. As the project focuses on Wideband codec implementation, the input signal is sampled at 16 KHz and divided into frames of length 20ms. These frames are input to LPC analyzer, Pitch detector, Voice Activity Detection which compute the LP coefficients and pitch frequency.

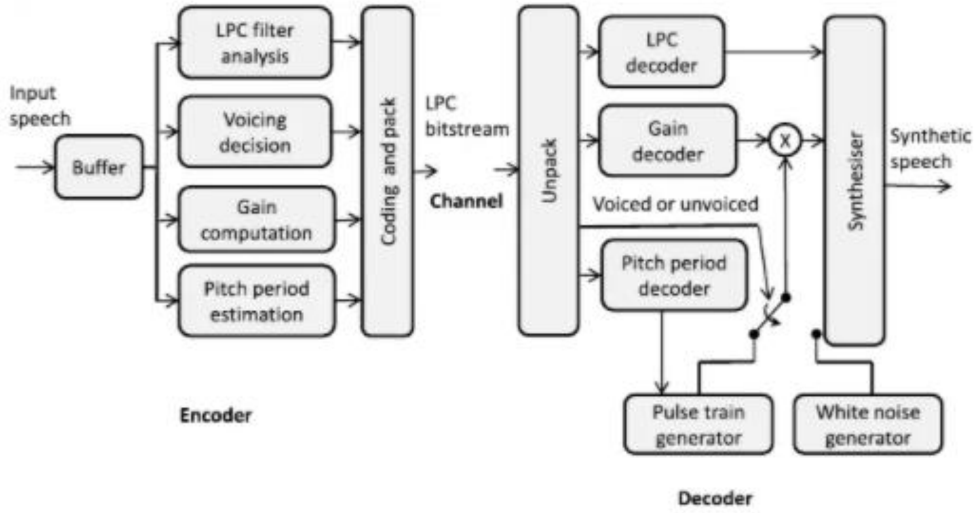


Figure 5. Plain LPC Codec block diagram [2]

The LPC encoder consists of pre-emphasis filtering and Levinson-Durbin recursion blocks. Generally, the speech signal's energy levels are concentrated at the low frequency level, some of the higher frequency values would be lost while encoding, as the codec's mode of operation is wideband. To counteract this pre-emphasis filtering is done which boosts higher frequency values to flatten the spectrum, leading to better estimation of LPC coefficients using Levinson Durbin autocorrelation method. The coder quantizes the coefficients for transmission through the channel. Direct quantization of the coefficients would lead to aliasing as LPC coefficients contain information about the Formant frequencies. Instead, the partial coefficients (PARCOR) generated by the Levinson-Durbin recursion are transmitted.

The parameters extracted and sent to the receiver are:

- Voice Activity Detection Flag (A single bit per frame to differentiate the voiced/unvoiced frames)
- Pitch Frequency (1-60 different values quantization using a-law or u-law companding).
- LPC Coefficients (10 coefficients for voiced speech and 4 coefficients for unvoiced speech per frame).
- Gain or Signal power: 5 bits per frame

The decoder segregates the input bitstream to retrieve the LPC coefficients, pitch frequency and gain. The voice activity flag controls the LPC speech model to generate the speech segments, referred to as the synthesizer.

2.2.1 Implementation results: The LPC Vocoder operating with the sampling rate of 16000 Hz and window length of 20ms results in 320 samples per frame. For perfect reconstruction, the frames are overlapped every

10ms, resulting in 480 samples per frame or 50 frames per second. Including the voice activity flag and the other parameters results in bitrate close to 8 kb/s, achieving the compression ratio of 8 [5].

2.3 Voice-Excited LPC Codec

There are many problems associated to LPC Vocoder such as the poor speech quality, synthetic speech output, poor pitch estimation etc. Also, any errors caused in the transmission across channel has harsh impact on the output speech. For example, if the channel noise somehow modifies voice activity flag, the frame might be completely skipped. Voice-excited LPC codec is an alternative that transmits the residual signal generated by the LPC filter, along with the vocoder parameters. This residual signal is mathematically referred to as short term autocorrelated values, which contain the pitch and voice activity information which are utilized by the synthesizer at the decoder to generate output signal with similar characteristics as the input, leading to better speech quality. As the input speech's excitation details exist in the residual signals the LPC vocoder's pitch and voice activity information need not be transmitted. However, the inclusion of residue signal results in higher bitrate of almost 16 kb/s. The block diagram of voice excited LPC codec is shown in figure 6.

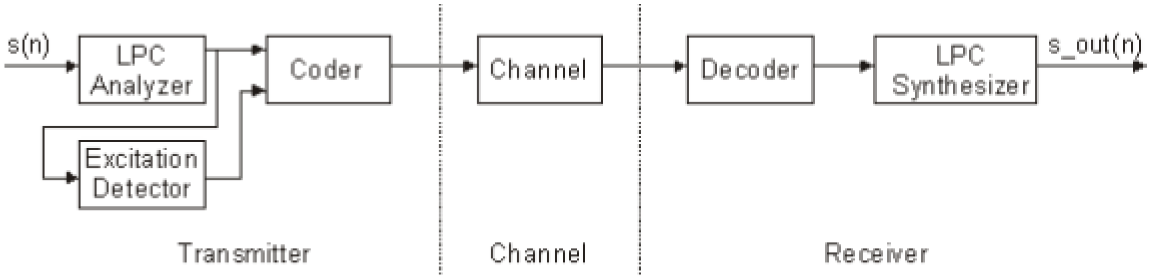


Figure 6. Voice-Excited LPC Codec block diagram [5]

To achieve higher compression rate, discrete cosine transform (DCT) in applied to the residual signal is done as DCT concentrates most of the energy on the initial few values, which are enough for the synthesizer to predict the speech signal's excitation sequence [5]. The DCT output of the residual signal is quantized with 4 bits. The parameters sent to the receiver are:

- LPC Coefficients
- Gain or Energy value (5 bits per frame)
- Residual signal's DCT coefficients

The decoder performs inverse DCT to get the voice excitation information from the residue sequence, which is used by the speech synthesis or mathematical speech model designed earlier to construct the output speech signal.

2.3.1 Implementation Results: As the residue information is included along with the LPC codec parameters, the overall bitrate is close to 16 kb/s, resulting in compression ratio of 4, along with better speech quality.

2.4 Code-Excited Linear Prediction (CELP)

CELP is the most common and successful compression technique wherein the framed-speech signal is further subdivided into segments (or subframes) of 5ms. The linear prediction analysis is done on the subframes along with a long-term redundancy predictor for all the possible excitations in what is referred as a codebook [6]. This codebook has usually 256 – 1024 entries wherein each entry is an excitation signal. This codebook is available at both sender and receiver, and the entries are waveform-like excitations instead of the pulse excitations, hence CELP can be referred to as Hybrid or Analysis-Synthesis codec[2].

The input speech of a subframe, x(n) signal is represented as an autoregressive model given as:

$x(n) = \sum x(n-k) + e(n)$ wherein each frame's LPC coefficients $(\alpha_1 - \alpha_p)$ are computed using Levinson-Durbin algorithm. The transformed equation is given as: $\frac{X(z)}{E(z)} = \frac{1}{A(z)} = \frac{1}{1-(\alpha_1 z^{-1}+\cdots)}$

2.4.1 CELP Encoder: The block diagram of the CELP encoder shown in Figure 7(a). Initially the codebook contains random values for the entries (or) code vectors (independent state distribution). The excitation signals or the code vectors are usually ADPCM values, which are passed on to the perceptual weighting filter (block shown as 1/(1-A(z)). Perceptual weighting is the key to obtain good speech coding performance, wherein the error in waveform encoding is shaped in the frequency domain to work with the wideband mode of operation. This weighted output is masked with the speech frame, to evaluate the weighted error, which is minimized over the long-term processing of the entire speech signal

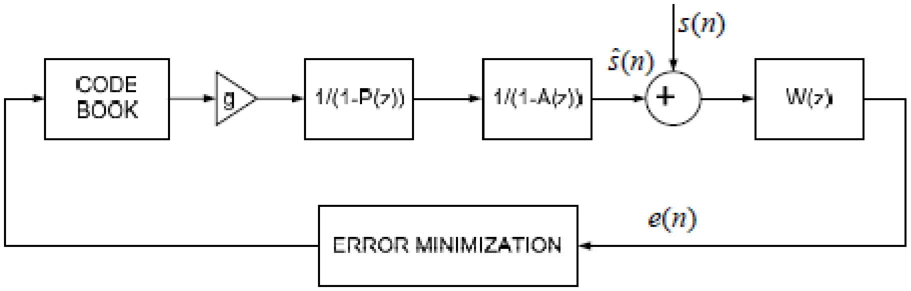


Figure 7 CELP Encoder [6]

Thus, the steps in encoding the speech signal using CELP in the proposed encoder can be summarized as:

- Codebook with independently chosen code vectors is generated.
- The code vectors are filtered to extract the LPC parameters (block shown as 1/(1-P(z)) and perceptual weighted (block shown as 1/(1-A(z))
- The filtered output is compared to the input subframe and the match with least mean-square error is chosen to represent the sub frame.
- The closed-loop process of identifying the bast matches is done for all the subframes and for each frame the corresponding code vector match will be transmitted to the decoder at the receiver. [2]

2.4.2 CELP Decoder: CELP decoder's operation is very similar to the encoder wherein the code vectors in the codebook are excited into forming short duration speech estimates, which are passed through LPC and Perceptual weighting filtering to get the decoded output speech signal. The block diagram of the CELP Decoder is shown in figure 8.

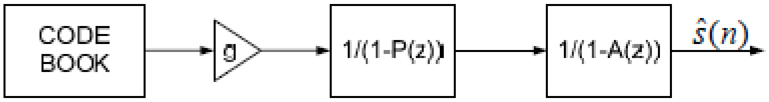


Figure 7 CELP Decoder

2.4.3 Implementation Results: The CELP codec implemented is an extension of LPC codec with a codebook size of 1024. Both narrowband and wideband mode of operation is possible by the codec by providing corresponding values which are listed below:

Parameter Name	Narrowband (8 KHz value)	Wideband (16 KHz value)
Frame Length	160	20
Subframe Length	40	5
Order of LPC Analysis	12	16
Perceptual Weighed Filter value	0.85	0.9
Pitch Lag Range	[16 160]	[34 231]

2.5 Algebraic Code Excited Linear Prediction

An improvement observable on CELP codec is to split the codebook into fixed codebook and adaptive codebook. The fixed codebook contains an algebraic structures of fixed pulse trains and the adaptive codebook contains the set of past excitation signals. These two codebooks are combined other blocks in CELP codec is referred to as Algebraic Code Excited Linear Prediction (ACELP), which has been standardized by ITU-T in 2002 as G.722.2. The G.722.2 ACELP codec is referred to as Adaptive-Multirate Wideband speech codec. The encoder, decoder and the bit allocation data of this codec is shown in the below figures 8 (a) (b) and (c).

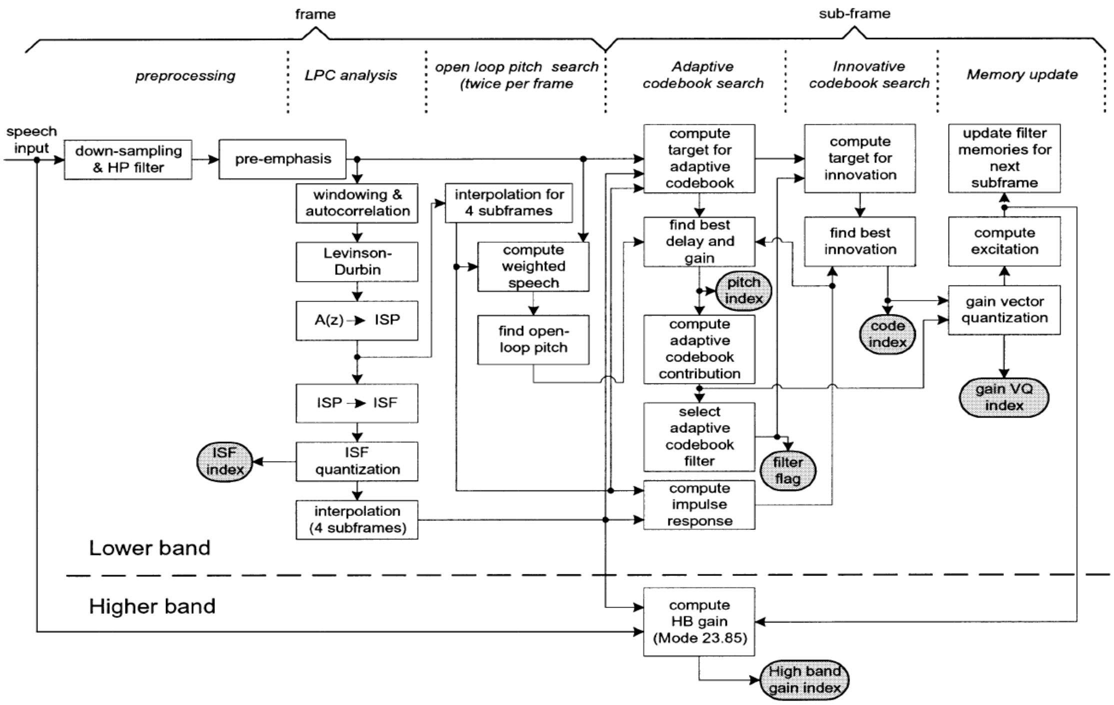


Figure 8(a). ITU-T G.722.2 encoder block diagram [3]

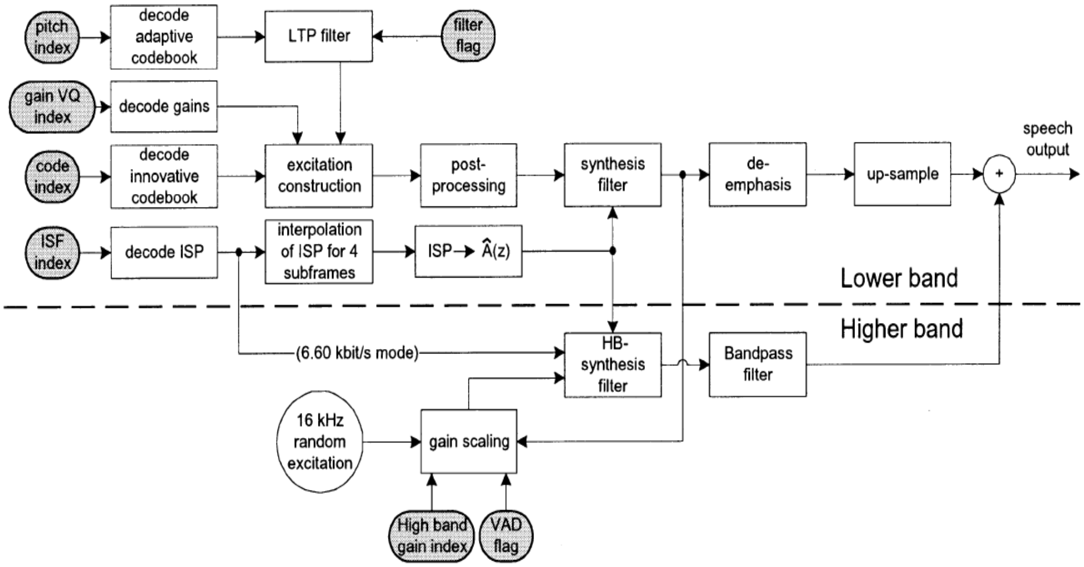
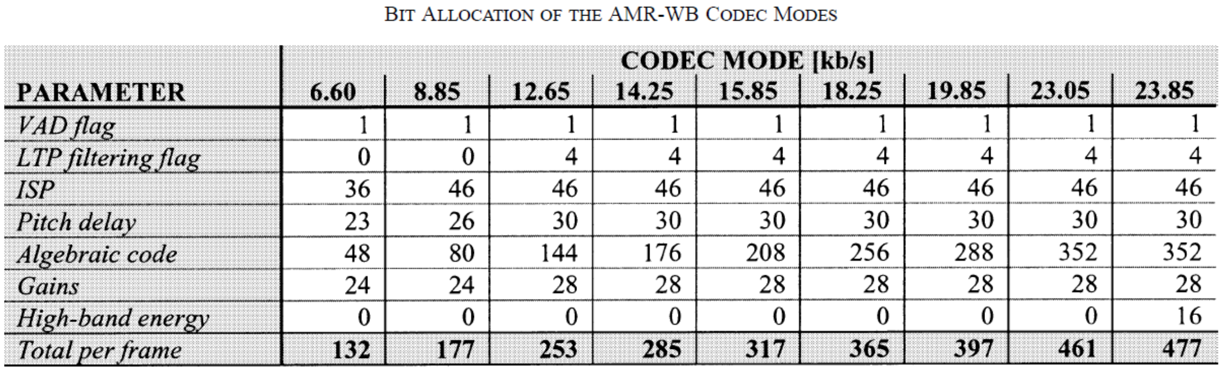


Figure 8(b). ITU-T G.722.2 decoder block diagram [3]



BIT ALLOCATION OF THE AMR-WB CODEC MODES

PARAMETER	CODEC MODE [kb/s] 6.60	8.85	12.65	14.25	15.85	18.25	19.85	23.05	23.85
VAD flag	1	1	1	1	1	1	1	1	1
LTP filtering flag	0	0	4	4	4	4	4	4	4
ISP	36	46	46	46	46	46	46	46	46
Pitch delay	23	26	30	30	30	30	30	30	30
Algebraic code	48	80	144	176	208	256	288	352	352
Gains	24	24	28	28	28	28	28	28	28
High-band energy	0	0	0	0	0	0	0	0	16
Total per frame	**132**	**177**	**253**	**285**	**317**	**365**	**397**	**461**	**477**

Figure 8(c). Bit allocation across various codec modes [3]

3. PERFOMANCE EVALUATION OF CODECS

Speech codecs have the drawback of not having objective measures of speech quality. As subjective measures vary from person to person, a codec is evaluated based on combination of the following parameters:

- **Bitrate**: Bitrates for codecs designed have been correspondingly mentioned. An important point here is bitrate alone would not be sufficient to evaluate a codec as the underlying speech could be poor
- **Overall Delay:** Overall delay of a codec can be defined as the time taken for the first speech sample is taken as the input by the encoder to the first sample of synthesized speech available at the decoder. This delay has the impact of channel coding and conditions, thus is important [6]. The codecs corresponding frame length is the ideal delay value. However there tend to be a few milliseconds of tradeoff added to the frame length.
- **Computational Complexity:** This deals with the time taken by the computer to process the codec operations, which is measured in floating point operations per second (FLOPS).
- **Mean Opinion Score:** Speech rating done by humans on a scale of 1 – 5 (1 is bad and 5 is excellent).

4. IMPLEMENTATION DETAILS AND RESULTS

MATLAB was used to implement the LPC vocoder, Voice-excited vocoder and the Code-excited vocoder (CELP). ACELP codec implementation requires tweaking of the CELP codecs to include two codebooks and transmission of Intermittent Spectral Pairs (ISP), which has not been implemented and remains the future scope. Bitrates of the output generated by the LPC vocoder is 8 kb/s, Voice-excited vocoder is 16 kb/s and CELP is also 16kb/s. Clearly the CELP codec has better speech quality. Compression ratio (measured with respect to 64 kb/s PCM technique) is higher for LPC vocoder (8), but the speech quality is poor. The other two codecs have almost same compression ratio of approximately 4. Following figures show the comparison of the input speech signal and the respective reconstructed signal.

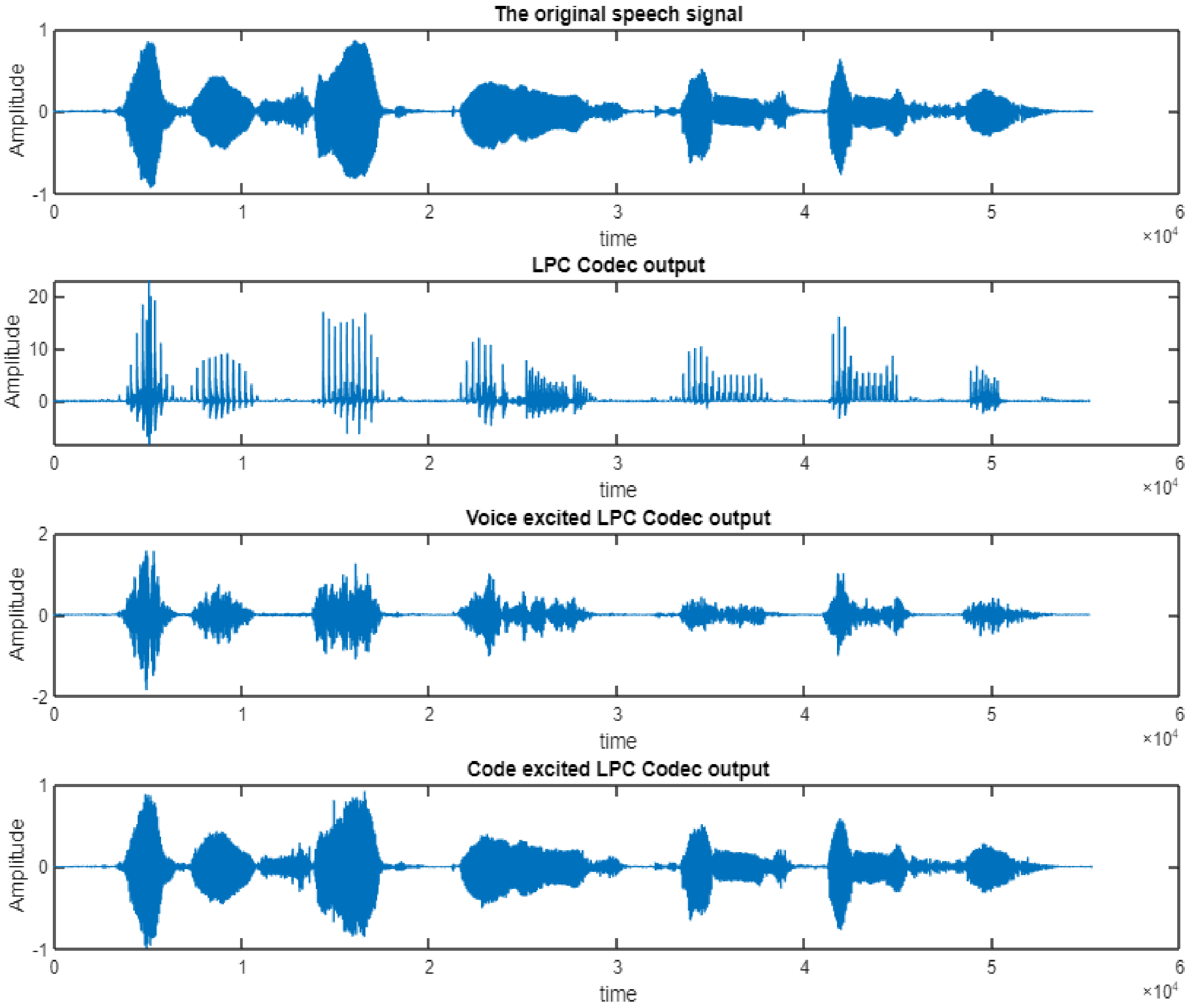


Figure 9. Comparison of various wideband codecs implemented in this project

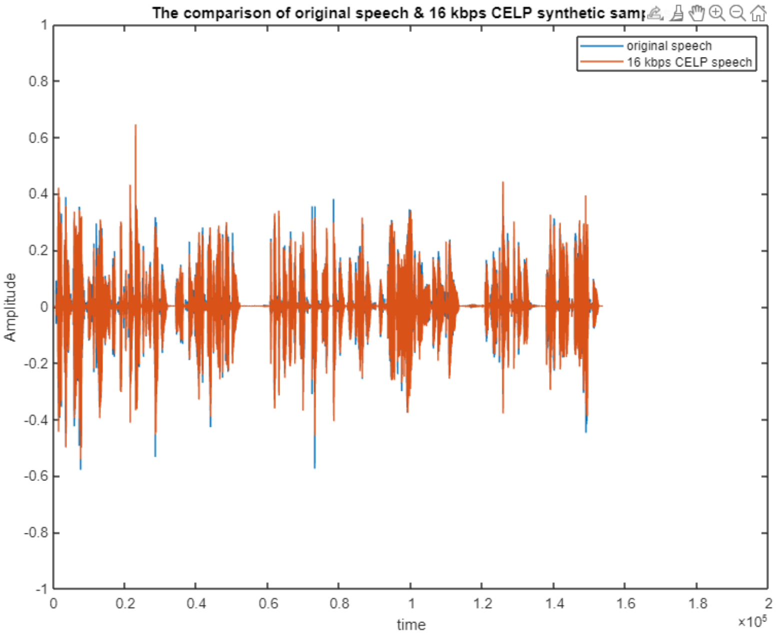


Figure 10. Various outputs of Code-Excited Linear Prediction

5. CONCLUSION AND FUTURE SCOPE

This project explored the different approaches/techniques involved in speech compression and focused on creating speech codecs for wideband mode of operation. LPC Vocoder, Voice-Excited LPC codec and Code-Excited LPC codec are explored, studied and arranged in MATLAB. While LPC Vocoder has poor speech quality and is susceptible to information loss, variants of LPC vocoder wherein the LPC coefficients (both short term and long term) are extracted and passed on to the decoder in both Voice-Excited LPC codec and CELP codec. CELP is the best out of the implemented codecs as the input-output variation is very minimal resulting in approximately similar reproduction of input speech signal. Given its superior audio quality and robustness against any transmission errors, CELP based Analysis-Synthesis codecs are widely used for the modern speech coding applications.

The extension of CELP to Algebraic CELP, which involves splitting of codebooks into fixed and adaptive codebooks and transmission of extracted coefficients as Line Spectral Pairs resulting multirate mode of operation and better performance over CELP codecs. Implementation of ACELP based Wideband speech codec can be considered as the future scope of this project.

REFERENCES

[1] K. Sayood, Introduction to Data Compression, fourth edition, Morgan Kaufmann Publishers, 2012.

[2] Arulmozhi Elango, Speech Compression, (https://www.academia.edu/6713164/2_Speech_Compression).

[3] B. Bessette et al., "The adaptive multirate wideband speech codec (AMR-WB)," in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 8, pp. 620-636, Nov. 2002, doi: 10.1109/TSA.2002.804299.

[4] http://www-mobile.ecs.soton.ac.uk

[5] http://www.seas.ucla.edu/spapl/projects/ee214aW2002/1/report.html

[6] J. D. Gibson, "Speech coding methods, standards, and applications," in IEEE Circuits and Systems Magazine, vol. 5, no. 4, pp. 30-49, Fourth Quarter 2005, doi: 10.1109/MCAS.2005.1550167.

[7] B. S. Atal, M. R. Schroeder, and V. Stover, "Voice-Excited Predictive Coding Systetm for Low Bit-Rate Transmission of Speech", Proc. ICC, pp.30-37 to 30-40, 1975

[8] http://www-mobile.ecs.soton.ac.uk/

[9] http://www.data-compression.com/speech.html

[10] https://core.ac.uk/download/pdf/82143559.pdf

[11] Orsak, G.C. et al., "Collaborative SP education using the Internet and MATLAB" IEEE SIGNAL PROCESSING MAGAZINE Nov. 1995. vol.12, no.6, pp.23-32

[12] Sourav Mondal (2020). CELP codec (https://www.mathworks.com/matlabcentral/fileexchange/39038-celp-codec), MATLAB Central File Exchange.