Speech Coding Techniques for VoIP Applications: A Technical Review

Manas Ray, Mahesh Chandra and B.P. Patil

1,2Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Ranchi-835215, India
3Department of Electronics and Telecommunication, Army Institute of Technology, Dighi, Pune-411015, India

Abstract: Voice over Internet Protocol (VoIP) is a revolutionary technology which is acting as a platform for the development of latest trends in modern communication world. The speech signal quality in VoIP is governed by the speech coding techniques employed. Currently various standard coders like FS-1015(LPC-10), ITU-G.711, ITU-G.726, FS-1016(ITU-G.728), etc. are used to digitize the speech signal. Modern research trends in signal processing are concentrating on the wavelet based compression techniques. This paper analyzes the performance of the standard codecs with various wavelet based codecs by statistical analysis of the coding capabilities of the codecs for English and Hindi language sentences spoken by a male and a female speaker. The performance is measured in terms of Compression Ratio (CR), Signal to Noise Ratio (SNR), Normalized Root Mean Square Error (NRSME) and Mean Opinion Score (MOS)

Key words: Performance evaluation • Statistical Analysis • Standard speech codecs • Wavelets • VOIP

INTRODUCTION

Voice over IP (VoIP) lays the foundation of the concept of the voice signal delivery on an internet protocol based packet switched network rather than the conventional connection oriented public switched telephone network (PSTN). Due to its wide acceptance in the modern communication domain and numerous applications, VoIP has become the driving force behind the evaluation of telephony industry [1]. The latest example is the rolling out of the 4G technology, which according to International Mobile Telecommunications Advanced (IMT-A) standard defined by International Telecommunication Union-Radio communication (ITU-R), is based on all-internet protocol packet switched network having peak data rates up to 1 Gbps [2]. The advantages of VoIP includes, minimum initial deployment cost, lower cost of operation as compared to the conventional PSTN, ease of deployment, user friendliness, ease of launch of additional services [3] etc. This had led to rapid deployment of VoIP services over the recent years and had opened up various Quality of Service (QoS) challenges in the deployment [4]. The central component of the VoIP services delivery is the voice coder or the codec deployed for the application. There are different types of codec in use, based on the application, complexity, bandwidth requirement etc. As modern VoIP deployments spans across the continents, various ITU/FS standard codec are deployed for VoIP applications for seamless interconnection without any ambiguity. These codecs were initially designed and deployed in the legacy PSTN, however due to the inherent qualities of the VoIP application to adapt to the existing infrastructure; these codecs have been seamlessly integrated to VOIP systems and have become the central theme of VoIP, dictating the QoS requirements. The most recent codec deployments in the VoIP setup are FS-1015/LPC-10, ITU G.711, ITU G.728, ITU G.729 etc. The data rate generated by VoIP codecs differs based on the engineering trade-off between voice quality, available bandwidth and complexity of the Codec [4].

During latter half of the last decade, various non ITU standard codecs are deployed for VoIP applications. These codecs are being developed by various organizations and deployed by Internet Engineering Task Force (IETF). These codecs are in various stages of development, testing and standardization and defined as Request for Comments (RFC)[5]. These codecs include
Internet Low Bit-rate Coding (iLBC), SPEEX, SILK, OPUS etc. These codecs are being deployed for specific applications on internet and are generally based on the basic principles of the ITU standard codecs and are governed by free BSD license [6].

In recent years wavelet transform and its applications are being extensively used in various fields of engineering and research [7]. Wavelet transform provides excellent resolution of signals in both frequency as well as time domains. Discrete wavelet transform provides a platform for multi-resolution analysis of speech signals [8]. Research is being carried out to achieve speech compression and coding using various transform techniques [9][10].

This paper evaluates the speech processing capabilities of the wavelet transform and evaluated the performance against the ITU standard codecs employed in VoIP applications. Analysis of the working of the codecs is performed by testing their characteristics on various sentences of different speaker’s speech in Hindi and English languages.

The paper is organized in six sections. Section-I reviews the concept of VoIP. The different types of standard codecs deployed in the VoIP applications along with a brief background on their signal processing techniques are presented in Section-II. Introduction to the wavelet transform and its speech processing techniques are presented in Section-III. The parameters measured for the performance evaluation of codecs are defined in Section-IV. The simulation results of the codecs and their response to the various Hindi and English sentences are presented in Section-V and finally conclusions are drawn in Section-VI.

**Standard Codecs For VOIP Applications:** There are various standard codecs used for speech coding applications. Some of them are Linear Predictive Coding (LPC), Pulse Code Modulation (PCM), Adaptive Differential Pulse Code Modulation (ADPCM), Code Excited Linear Predictive Coding (CELP) etc. In this paper five different standard speech codecs are discussed for VoIP applications.

**Pulse Code Modulation (PCM) (ITU-G.711):** Pulse Code Modulation (PCM) [11], is speech codec where the input speech signal is filtered by a 4 KHz Low Pass Filter and the resultant output is sampled at 8 KHz to obtain a pulse train. The pulse train are then quantized and encoded with 8 bit binary code. These binary codes are then transmitted to the receiver. At the receiving end the binary codes are converted into the original speech with the help of digital to analog converter and a low pass filter. It is a waveform type speech coding technique where the waveform of the synthesized speech signal agrees with that of the original speech signal waveform. The operating bandwidth of the coder is 64Kbps. The reproduced speech signal has the highest quality known as Toll quality.

**Adaptive Differential Pulse Code Modulation (ITU-G.726):** Adaptive Differential Pulse Code Modulation (ADPCM) is a modified version of the Pulse Code Modulation where 4 bits per sample are used for encoding the pulse stream [12]. Due to lower bit rate representation; ADPCM codecs require appreciable less storage space. The encoder records only the difference between the adjacent samples and dynamically adjusts the coding scale to accommodate large and small differences [13].

**Linear Predictive Coding (LPC-10/FS-1015):** Linear Predictive Coding (LPC) is an auto-regressive method of speech coding, in which the speech signal at a particular instant is represented by a linear sum of the ’p’ previous samples [14]. It operates at a data rate of 2.4kbps. LPC is based on the concept of parametric coding; where in the human vocal tract is mathematically approximated as a tube with varying diameter. LPC is used to estimate the basic speech parameters like voice/ unvoiced decision, pitch, formant etc. In the analysis stage of the LPC coding, the speech samples are broken down into segments or blocks then each segment is analyzed to determine the following:

- Nature of signal as voiced or unvoiced?
- Pitch of the signal
- The Vocal tract filter is analyzed as per the equation

\[ y_n = \sum a_i y_{n-i} + G e_n \]  

(1)

Where \( y_n \) the output and \( e_n \) the input and \( a_i \) and \( G \) are the parameters of the Vocal tract filter which needs to be estimated. These parameters are passed by the encoder to the decoder of the synthesis stage to reconstruct the signal. At the decoder end the vocal tract filter is estimated from the parameters received from the analysis stage. The voice quality generated by the codec is of robotic nature and the trade-off is the low bit rate employed for the coding purpose.
Code Excited Linear Prediction (CELP) (ITU G-728/FS-1016): CELP is a hybrid coding technique which combines the advantages of both techniques (waveform and parametric) to provide a robust low bit speech coder [15]. In the analysis stage the speech signal is passed through a cascade of formant predictor filter and pitch predictor filter. The formant predictor filter removes sample to sample correlation and pitch filter removes the long term correlations. The residual signal resembles a noise like signal which is compared with the entries of the code book of signal available, the index of the best matched entry is selected. The parameters of the filters along with the index value of the codebook representing the residual filter are passed on to the synthesizer. In the synthesis stage the excitation waveform is chosen from a dictionary of waveforms as per the index transmitted; which in-turn drives a cascade of filters synthesized from the parameters received from the analysis stage to approximate the input speech signal.

Conjugate Structure-Algebraic Code Excited Linear Prediction (CS-ACELP) (ITU G-729): CS-ACELP is the most modern hybrid coding technique which is currently deployed in almost all the latest VoIP applications. It operates at 8Kbps and provides near Toll quality performance of the voice signal. The coder is based on code excited linear prediction model. It utilizes conjugate structure for the 2 D joint vector quantization of the sub-frame based adaptive codebook gain and fixed codebook gain. Here the linear combination of the code book vector from the two code books is a simple summation of the conjugate structure codebook [16]. This coder is robust against channel errors. The characteristics of the currently deployed codecs can be summarized as presented in Table 1.

Discrete Wavelet Transform and its Application in Speech Coding: The signal compression using wavelets is achieved by the inherent fact of relative scarceness of the wavelet domain representation of the signal [20]. Here the signal is de-composed into components that are not pure-sine waves and the information is condensed in both time and frequency domains. Wavelet is waveform of limited duration that has an average value of zero. It is finite in nature. Wavelets help in localization of signal in both frequency as well as time domain. The multi-resolution capability of the wavelet provides us with dilate and translate versions of the wavelet [21]. The wavelet prototype function used for analysis is called as the mother wavelet \( \Psi(x) \) [22]. This function is dilated and translated to achieve the basis function at different scales by following the basis set equation

\[
\Psi\left(\frac{x-b}{a}\right) 
\]

For dyadic case \( a=2^j \) and \( b=K \), where \( k \) and \( j \) are integers. The equation which defines the scaling function \( \Phi(x) \) is defined as

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Codec</th>
<th>Standard by</th>
<th>Description</th>
<th>Bit rate (Kbps)</th>
<th>Type</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>G.711</td>
<td>ITU-T</td>
<td>Pulse Code Modulation(PCM)</td>
<td>64</td>
<td>Waveform Codec</td>
<td>Superior Quality of speech encoding</td>
</tr>
<tr>
<td>2</td>
<td>G.726</td>
<td>ITU-T</td>
<td>Adaptive Differential Pulse Code Modulation</td>
<td>16</td>
<td>Waveform Codec</td>
<td>Near Superior Quality at lower bit rate</td>
</tr>
<tr>
<td>3</td>
<td>LPC-10</td>
<td>Govt. of USA</td>
<td>Linear Predictive Coding</td>
<td>2.4</td>
<td>Parametric Coding</td>
<td>Robotic Speech, Excellent Bit Rate for speech coding applications</td>
</tr>
<tr>
<td>4</td>
<td>G.728/FS1016</td>
<td>ITU-T/ Govt. of USA</td>
<td>Code Excited Linear Prediction</td>
<td>16</td>
<td>Hybrid Codec</td>
<td>Used as the building block of modern Speech codecs systems</td>
</tr>
<tr>
<td>5</td>
<td>G.729</td>
<td>ITU-T</td>
<td>Conjugate Structure-Algebraic Code-Excited Linear Prediction</td>
<td>8</td>
<td>Hybrid Codec</td>
<td>Latest Codec and widely used</td>
</tr>
<tr>
<td>6</td>
<td>iLBC[17]</td>
<td>RFC 3951</td>
<td>Internet Low Bit rate Codec</td>
<td>13.33 and 15.2</td>
<td>Waveform Codec based on PCM</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>SPEEX[18]</td>
<td>RFC 5574</td>
<td>Hybrid coding based on CELP</td>
<td>2-44</td>
<td>Hybrid coding based on CELP</td>
<td>The software is Open Source and Royalty Free</td>
</tr>
<tr>
<td>8</td>
<td>SILK</td>
<td>Not Standardized</td>
<td>Parametric coding based on LPC</td>
<td>6-40</td>
<td>Parametric coding based on LPC</td>
<td>Developed by Skype limited. The software is Open Source and Royalty Free</td>
</tr>
<tr>
<td>9</td>
<td>OPUS [19]</td>
<td>RFC 6716</td>
<td>Hybrid coding based on CELP</td>
<td>6-510</td>
<td>Hybrid coding based on CELP</td>
<td>Open Source and Royalty Free Software</td>
</tr>
</tbody>
</table>
\[ \Phi(x) = 2^{j/2} \Phi(2^{j} x - b) \]  
(3)

The mother wavelet \( \Psi(x) \) may be expressed as

\[ \Psi(x) = 2^{j/2} \Psi(2^{j} x - b) \]  
(4)

The wavelet function \( \Psi(x) \) is orthogonal to the scaling function \( \Phi(x) \) and both function follow the following conditions:

\[ \int_{-\infty}^{\infty} \phi(x) dx = A \]  
(5)

\[ \int_{-\infty}^{\infty} \psi(x) dx = 0 \]  
(6)

\[ \int_{-\infty}^{\infty} |\phi(x)|^2 dx < \infty \]  
(7)

\[ \int_{-\infty}^{\infty} |\psi(x)|^2 dx < \infty \]  
(8)

Where, ‘\( A \)’ is a constant. The given basis set for analysis contains only one scaling function and the rest of the elements are wavelets. Daughter wavelets are derived by scaling the Mother wavelet, these wavelets can cover the entire time axis by translation [22].

Using the scaling and the wavelet functions the wavelet transform of the original function \( f(x) \) may be obtained as:

\[ f(x) = \sum_{k=-\infty}^{\infty} c_k \phi_k(x) + \sum_{k=-\infty}^{\infty} d_{j,k} \psi_{j,k}(x) \]  
(9)

Where \( C_k \) are called the average coefficients and \( d_{j,k} \) is called detail coefficient.

**Haar Wavelet (Haar):** Haar wavelet is the simplest possible wavelet [23]. For an input represented by a list of 2^e numbers, Haar wavelet transform would pair up the input values, providing the difference and forwarding the sum to the next level. The process is repeated recursively, pairing up the sums to provide the next scale, resulting in 2^e-1 differences and one final sum. It is not continuous. It has orthogonal property. The Haar wavelet’s mother wavelet function \( \Psi(x) \) is defined as

\[ \psi(x) = \begin{cases} 
1 & 0 \leq t \leq \frac{1}{2} \\
-1 & \frac{1}{2} \leq t \leq 1 \\
0 & \text{otherwise} 
\end{cases} \]  
(10)

Haar scaling function is defined as

\[ \Phi(x) = \begin{cases} 
1 & 0 \leq t < 1 \\
0, & \text{Otherwise} 
\end{cases} \]  
(11)

**Daubechies Wavelet:** Daubechies wavelets are the family of orthogonal wavelets, characterized by a maximal number of vanishing moments for some support and are commonly used for analysis of a signal. The Daubechies wavelets are not defined in terms of the resulting scaling and wavelet functions[24]. The index number refers to the number \( N \) of coefficients. Each wavelet has a number of zero moments or vanishing moments equal to half the number of coefficients [22].

**Discrete Approximation of Meyer Wavelet (Dmey):** Dmey wavelet is the discrete format of meyer wavelet function and is defined as

\[ G_{\omega} \left( e^{j\omega} \right) \sqrt{2} \sum_{k} \Phi(2\omega + 4k\pi) \]  
(12)

Given the basis function \( \Phi \) for the approximation space Meyer employed the Fourier techniques to derive DTFT of the two scale coefficients [9].

**Coiflet Wavelet:** Coiflets are discrete wavelets designed to have scaling functions with vanishing moments. The wavelet is near symmetric and has \( N/3 \) vanishing moments and scaling function have \( N/3-1 \) vanishing moments [22]. If the number of taps \( N=6p \), then \( 2p \) number of vanishing moment conditions are imposed on wavelet function and \( 2p-1 \) on scaling function and the remaining on normality and orthogonality conditions.

Thus the conditions imposed are [25]:

\[ \int \phi(t) dt = 1 \]  
(13)

\[ \int \phi(t)\phi(t-k) dt = \delta_{0,k} \]  
(14)

\[ \int t^n \psi(t) dt = 0 \quad \text{for} \quad n = 0,1,2,\ldots,2p-1 \]  
(15)

\[ \int t^n \phi(t) dt = 0 \quad \text{for} \quad n = 0,1,2,\ldots,2p-1 \]  
(16)
Principle of Speech Coding Using Wavelet Transform Based Codec: The process of speech coding using wavelet transform based codec is explained in Fig.1. The speech quality requirements of the codecs dictate the choice of mother-wavelet function. The objective is to minimize reconstructed error variance and maximize Signal to Noise Ratio (SNR) [26]. Wavelets work by decomposing a signal into different resolutions or frequency bands. The signal compression is achieved by reconstructing the signal by selecting a small number of approximation coefficients and some details coefficients by the concept of thresholding. Generally 5-level decomposition is adequate for speech signals [27].

Thresholding is applied on the coefficients obtained after wavelet decomposition in order to have the required level of compression and can be varied as per the requirement. Inverse wavelet transform is applied on the coefficients obtained to obtain the synthetic speech signal.

Performance Evaluation of Speech Codecs

The performance evaluation of the codecs are carried out by subjective and objective testing for six speech samples containing both male and female voices of Hindi and English languages. As the voice signals generated by male and speakers differ in parameters like pitch, formant etc due to the dimensions of vocal tract [28],[29] testing the performance of the codecs against both the samples will provide a measure of robustness. The subjective test for each samples can also be carried out by Mean Opinion Score. The objective tests were carried out by evaluating the performance in terms of Compression Ratio (CR), SNR, PSNR and NRMSE[30],[31].The expressions of these parameters are given below.

\[
CR = \frac{\text{Length of } (x(n))}{\text{Length of } (r(n))} \quad (17)
\]

Where \(x(n)\) and \(r(n)\) are the original and reconstructed signals respectively

\[
SNR = 10 \log_{10} \left( \frac{\sigma_x^2}{\sigma_e^2} \right) \quad (18)
\]

Where \(\sigma_x^2\) and \(\sigma_e^2\) e mean square of the speech signal the mean square difference between the original and constructed signal respectively.

The Normalized root mean square error (NRMSE) is defined as:

\[
NRMSE = \sqrt{\frac{(x(n) - r(n))^2}{(x(n) - \mu x(n))^2}} \quad (19)
\]

Where, \(x(n)\) is the speech signal, \(r(n)\) is the synthetic signal and \(\mu x(n)\) is the mean of the speech signal.

RESULTS AND DISCUSSION

MATLAB simulation model is prepared for PCM [32], CELP [33], ADPCM[34], CS-ACELP[35]and LPC [36] as per the standard available. Further Discrete wavelet transform based codec is simulated based on the speech compression principle adapted in Wavelet transforms. The test sentences are iterated against each of the above codecs and the set of 4 different wavelet families viz. Haar, Daubechies, Discrete Meyer wavelet and Coieflets. The sentences used for testing the performance of Codecs are given in Table 2.

Comparative performance of LPC, CELP, PCM, ADPCM and the different wavelet based codecs are presented in Fig. 2, Fig. 3, Fig. 4 and Fig.5 in terms of compression ratio, SNR, NRMSE and MOS. It is observed from results that wavelet based codecs provides a greater degree of compression than the ITU standard codecs in

<p>| Table 2: Details of the Test Sentences |</p>
<table>
<thead>
<tr>
<th>Sample No</th>
<th>Language</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hindi</td>
<td>Male</td>
</tr>
<tr>
<td>2</td>
<td>Hindi</td>
<td>Male</td>
</tr>
<tr>
<td>3</td>
<td>English</td>
<td>Male</td>
</tr>
<tr>
<td>4</td>
<td>English</td>
<td>Female</td>
</tr>
<tr>
<td>5</td>
<td>Hindi</td>
<td>Female</td>
</tr>
<tr>
<td>6</td>
<td>Hindi</td>
<td>Female</td>
</tr>
</tbody>
</table>
general. The Discrete Approximation of Mayer Wavelet (dmey) provides a greater compression to the speech signals as compared to the other families of wavelet under study. This compression ratio can further be varied by adjusting the threshold values of the coefficients obtained during wavelet decomposition. ITU standard codecs like PCM, ADPCM, CS-ACELP, CELP shows excellent SNR and NRMSE measurements; In the wavelet domain, the Daubechies family wavelet (db10) provides near comparable results as compared to the standard codecs among other families of the wavelets. The quality of the reconstructed signal was tested as per the subjective analysis and found to be in compliance with the Mean Opinion Score standard requirements of the ITU standard [37][38]. The MOS of PCM is best followed by Daubechies family wavelet based codecs. Hence it can be inferred from the above results that wavelet based codecs provides a good alternative to the ITU standard Codecs employed in the VoIP applications.

CONCLUSIONS

The performance of various standard codecs used for VoIP applications has been evaluated along with the wavelet based codecs in this paper. It is observed that the wavelet based codecs produces comparable results to that of the standard codecs currently deployed in modern
VoIP applications. Wavelet based codecs provide a greater degree of compression of the speech signal as compared to the conventional ITU standard codecs. This compression can further be fine-tuned as per the requirements of the application. Further it is observed that the quality of the reconstructed signal obtained by the wavelet analysis closely resembles the ITU standard codecs both in terms of the subjective and the objective measurement of the speech signals. Hence it can be concluded that the wavelet based speech codecs provides a viable alternative to the currently deployed ITU standard codecs deployed in the modern VoIP communication setup, with greater degree of flexibility as required in the actual implementation of the setup.

REFERENCES

5. https://www.ietf.org/


