Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 10, October 2013, pg.243 – 266

RESEARCH ARTICLE

Analysis of Vocoder Technology on Male Voice

Enyi Francis[1], Chiadika Mario[2], Ekoko Ujerekre[2], Ifezulike N. Florence[3], Biobelemoye W. Roberts[4]

[1]Computer Science Department, Nnamdi Azikiwe Awka, Anambra State, Nigeria

[2]Computer Engineering Department, Delta State Polytechnic, Ogwashi-Uku, Delta State, Nigeria

[3]ICT Unit, Nnamdi Azikiwe University Teaching Hospital, Nnewi, Anambra state, Nigeria

[4]Heriot - Watt University, Edinburgh, United Kingdom

[1] frankenyi@gmail.com

Abstract- This study tries to reproduce a model of the male voice signal using vocoder technology and to demonstrate the various stages involved while processing the voice signal. The techniques employed to accomplish these goals include extracting the fundamental and formant frequencies from the vocal-tract of the speaker with the aid of the matlab tool. Cepstrum and Autocorrelation function techniques were used to determine the fundamental frequency in the frequency and time domains respectively. Linear prediction technique was also used to determine linear coefficients of the filter (vocal tract), the formant frequencies and then applying the autocorrelation function to calculate the equation to determine the reflection coefficients of the filter. These techniques were applied to samples of male voices. The test and results was able to show how to differentiate a voiced from an unvoiced speech; and how the original signal modeled can be reconstructed.

Keywords- cepstrum; linear prediction; spectrogram; vocoder; voice

I. INTRODUCTION

The vocal organs are responsible for the production of speech. Speech is produced by an air stream coming from the lung, which is forced through the glottis between the vocal cords and the larynx to the main cavities of the vocal tract which are the pharynx, the oral and nasal cavities. The air stream leaves the nasal and oral cavities through the nose and mouth respectively. This involves four processes: Initiation, phonation, oro-nasal process and articulation.

The initiation process is the moment air is expelled from the lungs. For English, speech sounds are the result of 'a pulmonic egressive air stream' (which was stated by Giegerich 1992) [1].

© 2013, IJCSMC All Rights Reserved

The phonation process takes place in the larynx. It has two horizontal folds of tissue in the passage of air which is known as the vocal folds. The gap between these folds is the glottis and it can be closed and opened. When it is closed, no air can pass through hence no sound. If it is opened narrowly enough to make the vocal folds vibrate, it produces the sound called 'voiced sounds'. If the glottis is opened wide, like in normal breathing, the vibration of the vocal fold is reduced producing the sound known as the 'voiceless sounds'.

The oro-nasal process is where we can differentiate between the nasal consonants (sounds produced from the exit of air through the nose) and other sounds. This selection process is done by the velum (soft palate) after the air has gone through the larynx and pharynx then into the nasal or oral cavity depending on the selection.

The final process is the articulation process. This takes place in the mouth and it is the process through which most speech sounds are differentiated from the other. The mouth helps in differentiating the oral cavity, acting as a resonator from the articulators which may be active or passive. In other words, the speech sounds are distinguished from one another in terms of manner, how and place where they are articulated.

Speech signal can be divided into three classes according to their mode of excitation [2]:

- Voiced sounds,
- Unvoiced sounds and
- Plosive sounds

A. Voiced Sounds:

These are sounds that are produced when the vocal cords are forced together and they vibrate in such a manner that as the air builds up, the glottis is forced open and it subsides as the air passes through [2]. The vibration of the cords produces an airflow wave which is almost triangular. Since it is periodic, it has a frequency spectrum of rich harmonics at multiples of fundamental frequency of vibration and decaying at rate of approximately 12Db/octave. The vocal tract plays the role of a resonant cavity that amplifies some of these harmonics and attenuates others to produced voiced sounds [2]. The range of pitch for a man is between 50Hz and 250Hz, with an average of 120Hz. For a woman, the upper limit could be as high as 500Hz. Examples of voiced sounds are produced when we say 'aah' or 'oh'.

B. Unvoiced Sounds:

Here, the vocal cords do not vibrate. They are two types of unvoiced sound: fricative sounds and aspirated sounds. For fricative sounds (example 's' and 'sh') a point of constriction is created at some points in the vocal tract. As air passes through it, turbulence occurs which results to random noise excitation [2]. Whereas in aspirated sounds (example 'h' or 'hello'), this turbulence occurs at the glottis as the vocal cords are held slightly apart.

C. Plosive Sounds:

For this type of speech signal, the vocal tract is closed at some point, and air is allowed to build up and then released suddenly. The rapid release of the pressure produces a transient excitation of the vocal tract. For example the 'puh' sound at the beginning of the word 'pin' or the 'duh' sound at the beginning of 'din'[2]. This excitation may occur with or without the vocal cord vibration to produce voiced or unvoiced plosives.

© 2013, IJCSMC All Rights Reserved

II. Speech Representation

The voiced and unvoiced speeches produce different sounds and spectra. Fundamental frequency and formant frequencies are one of the most important concepts in speech representation and analysis.

A. Voiced Sound:

Voiced sound as mentioned earlier, is produced when the glottis is opened narrowly enough to make the vocal folds vibrate. The vibration frequency has a range from 50 to 400Hz and resonance is formed at odd harmonics in the vocal track. This resonance Peaks are known as formants [3].

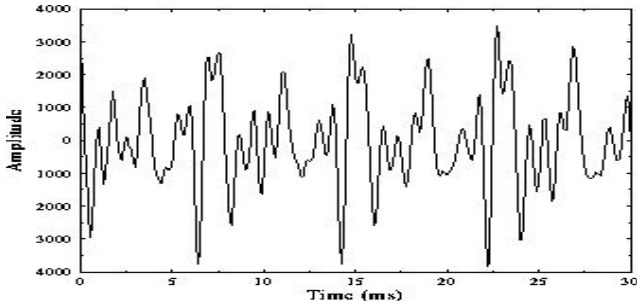


fig 1: Voiced Speech Sample [3]

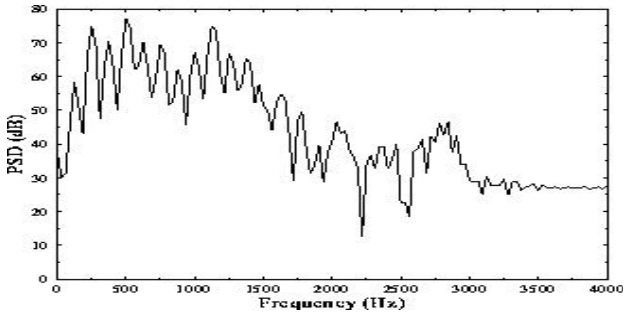


fig 2: Power Spectral Density for Voiced Speech [3]

B. Unvoiced Sounds:

They are also called fricatives. Since they are produced when the glottis is opened widely, there is less or no vibration in the vocal fold hence they don't produce as much periodic formant structure as in the voiced speech. They appear like noise and so the power spectral density does not display the resonant peaks clearly [3].

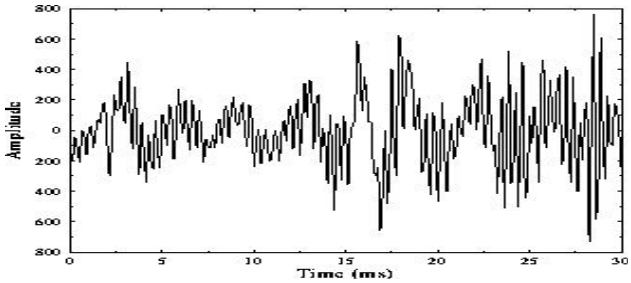


fig 3: Unvoiced Speech Sample [3]

© 2013, IJCSMC All Rights Reserved

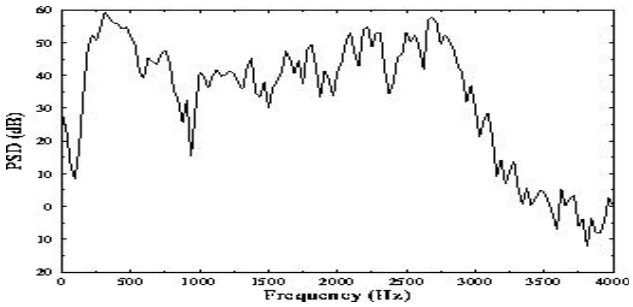


fig 4: Power Spectral Density for Unvoiced Speech [3]

III. Spectrogram

An effective way of observing the complete frequency spectrum is by calculating and displaying the spectrogram. This is a three-dimensional plot illustrating the way the frequency spectrum varies with time. The vertical axis represents time and the horizontal axis, frequency.

The diagram below shows the spectrogram and the time-domain waveform of the Finnish word 'kaksi...two'.

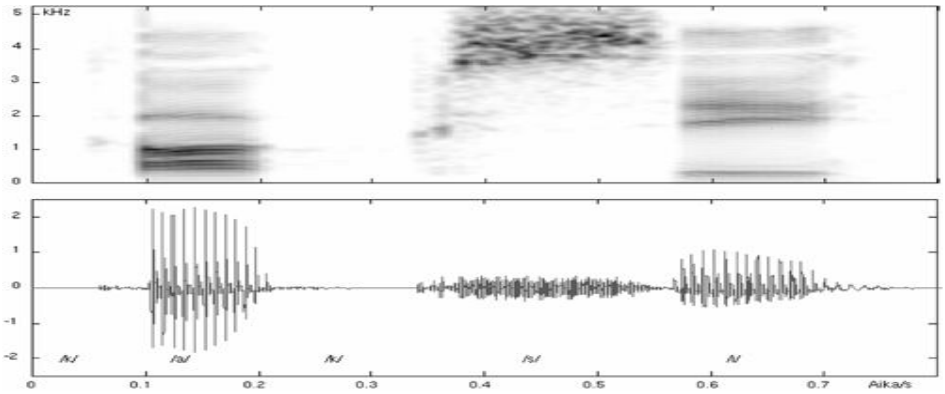


fig 5: Spectrogram and the time-domain waveform of the Finnish word 'kaksi'…….two'

In the fig 5, darker gray-levels shows the presence of higher amplitudes, in other that formant frequencies and trajectories are spotted easily. It also shows that vowels have more energy and it is focused at lower frequencies. Whereas, unvoiced consonants have less energy and it is focused at higher frequencies. Voiced consonants are placed in between of the two [1]. There are two types of spectrogram: Narrowband and Wideband.

A. Wideband Spectrogram:

This represents a window that has relatively short-time interval and is characterized by poor resolution in the frequency dimension and good resolution in the time-dimension [4].

B. Narrow Spectrogram:

In this case, a longer window is used to provide higher frequency resolution, which corresponds to a decrease in time resolution.

© 2013, IJCSMC All Rights Reserved

IV. Conversion from Analog to Digital Signal Using PCM

Pulse Code Modulation (PCM), is the standardized method used in telephone network to change an analog 'voice' signal to digital signal for transmission in a digital telecommunication network. The analog signal is sampled first at 8 kHz sampling rate, and then each sample is quantized into 1 of 256 levels and then encoded into 8-bit word. The overall data rate is 8000 X 8 = 64Kbps. In the United States, 1 bit of 8 in every sixth frame is 'robbed' for in-band signal hence reducing the channel capacity to 8000 X 7 = 56Kbps [5].

A. Sampling:

The amplitude of the input voice signal is first, sampled. The more the samples per second, the more representation of the voice signal samples formed. After sampling, the signal value is known only at discrete point called 'sampling instants'. The number of samples per second is the sampling frequency or sampling rate. It depends on the highest frequency component present in the voice signal. The relationship between sampling frequency and highest frequency of the signal to be sampled is the Nyquist Sampling Rate and it states : If the sampling frequency, fs, is higher than two times the highest frequency component of the analog signal, W, the original analog signal is completely described by these instantaneous samples alone; that is, fs >2W' [5]. The period is given as $Ts = 1/fs < (2W)$ [4] (1)

B. Quantizing:

For the sampled values to be transmitted in digital form in a digital system, each sample is represented in numerical form. This requires quantizing each sample to be rounded-off to the closest numerical number in a set of digital words. In this process (quantization), the information in the accurate signal is lost due to the round-off and the original signal cannot be reassembled exactly the same. The quality of the coding depends on the number of levels (quantum). The more the quantum level, the better the quality.

The number of quantum level is $q = 2n$ where q = number of quantum level; n = length of bit of binary word [5]

C. Speech Analysis:

Speech analysis is the process of estimating the relatively slow time-varying parameters which specify a filter from a speech signal that is assumed to be the output signal. [1] Fundamental frequency and formant frequencies are among the most important concepts in speech representation and analysis. [3] It can also be classified into Frequency-domain or time-domain approaches of speech analysis. In the frequency-domain, the Processing technique such as Cepstrum is used to estimate the frequency response of the vocal tract. Time-domain makes use of measures such as the autocorrelation function. Another very important technique is the linear prediction approach which is also an essential time-domain technique. It can also be used to produce an estimate of the frequency spectrum of the signal. [2]

V. Frequency-Domain Techniques

A. Cepstrum (Cepstral Analysis):

It is necessary that the duration of the analyzing window has to extend a few pitch-period of the speech signal in order to obtain good frequency resolution. In the cepstral analysis of voiced speech, the result is the frequency spectrum in which the discrete line spectrum of the periodic excitation is multiplied by the vocal tract spectral envelope. In course of this process, pitch ripples are produced and a technique known as Cepstral Truncation is used to remove them. [2]

© 2013, IJCSMC All Rights Reserved

B. Cepstral Truncation:

Let X(w) denote the spectrum of the voiced speech signal, P(w) the spectrum of the pitch impulses and H(w) the spectrum of the vocal tract which includes the effects of glottal waveform shape and radiation from the mouth. [2]

The relationship between the magnitudes of these three spectral can be expressed as follows:

$|X(w)| = |P(w)| \times |H(w)|$ [2] (2)

Taking the logarithm of this equation gives

$\text{Log}\{|X(w)|\} = \log\{|P(w)|\} + \log\{|H(w)|\}$ (3)

The contribution from |H(w)|, which is determined by the properties of the vocal tract tends to vary slowly with frequency, while the contribution of |P(w)|, the pitch tends to vary more rapidly and periodically with frequency [2]. These two components should now be separable by means of a linear filtering operation. The filtering operation is normally carried out by inverse fourier transforming "Log{|X(w)|}" to produce the cepstrum of the signal. The x-axis of the cepstrum has the dimension of time and is called quefrency of the signal [2]. The contribution due to pitch occurs at the multiples of the pitch period, and the contribution due to the vocal tract tends to occur near the origin since it varies much more slowly than the harmonic ripple. If only the values near the origin in the cepstrum are retained and these values are fourier-transformed, then the log-spectrum of the vocal tract ($\log\{|H(w)|\}$) is obtained.[2]

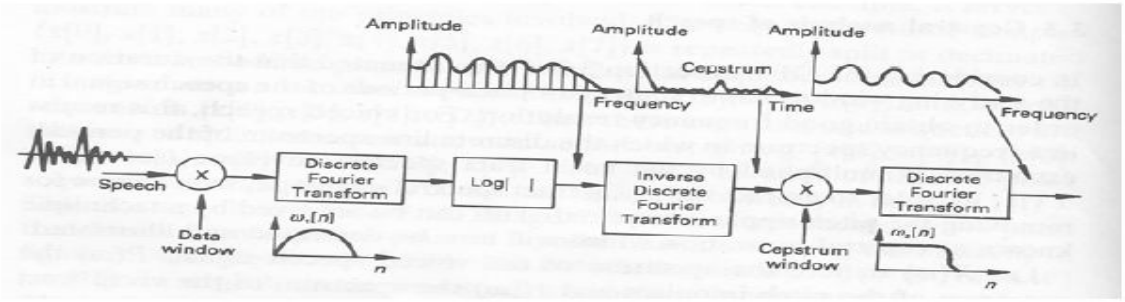


fig 6 – Block diagram of a Cepstral Processing System for Speech Analysis [2]

VI. Time-Domain Techniques

A. Autocorrelation Function (ACF):

The autocorrelation function gives a measure of the correlation of a signal with a delayed copy of itself [2]. The autocorrelation value, R(k), of a stationary signal x[n] for a time-shift of k samples is defined as:

$$R(k) = \sum_{n=-\infty}^{\infty} x[n].x[n+k] \quad (4)$$

For this to be computable the sample amplitude must be known at all time. For the simulation to be finite, the signal must have finite energy. As far as speech is concerned, the signal is quasistationary and the infinite energy is not known for all time. For a case like this, it is appropriate to calculate the short-time autocorrelation function by removing successive segment of the signal by multiplying it in a rectangular window W(n), of width N samples that has a value of 1 in the range (0, N-1) and in 0 outside. For the window, starting at the beginning of frame m the short-time autocorrelation function is defined as:

$$Rm(k) = \sum_{n=0}^{N-1} \{x[n].W[n]\} \, . \, \{x[n+k].W[n+k]\} \quad (5)$$

© 2013, IJCSMC All Rights Reserved

n=0

B. Linear Prediction Analysis:

This is an important and powerful speech processing technique. The basic idea in this method is that the sample values of speech, x[n] can be approximated as a linear combination of the past 'p' speech samples [1]. Linear prediction can be described mathematically by; [2]

$xp[n] = a1x[n-1] + a2x[n-2] + \ldots\ldots\ldots + apx[n-p] = \Sigma\, akx[n-k]$ (6)

where x_p = the predicted sample at instant 'n'.

$a_1, a_2, \ldots\ldots a_p$ = predictor coefficients

It is not possible to predict every signal sample exactly and this leads to "prediction error" e[n] at each sample instant:

$e[n] = x[n] - x_p[n]$ (7)

If the mean squared error between the actual speech samples and linearly predicted ones are minimized, the predictor coefficients can be determined by evaluating a set of linear equations. A set of predictor coefficients will be able to predict the speech signal reasonably accurately over stationary portions. For the time-varying properties of the speech signal to be matched, a new set of predictor coefficients are calculated every 10-30 ms.

Computing the error signal e[n],

$e[n] = x[n] - x_p[n]$ (8)

$e[n] = x[n] - a_1x[n-1] + a_2x[n-2] + \ldots\ldots\ldots + a_px[n-p]$ (9)

If the error signal e[n], is known, it is possible to reconstruct the original signal x[n] exactly from the predicted signal xp[n],

$x[n] = e[n] + x_p[n] = e[n] + \Sigma\, a_kx[n - k]$

Taking the z-transform,

$X(z) = E(z) + [\Sigma\, a_kz^{-k}]\, X(z)$

$X(z) = E(z) / (1 - \Sigma\, a_kz^{-k}) = H(z).E(z)$ (10)

Where X(z) and E(z) are the z-transform of x[n] and e[n].

Equation 1 shows that the speech signal x[n] can be seen as the output of this all-pole filter when the input is the error signal e[n] [2].

$H(z) = 1/(1 - \Sigma\, a_kz^{-k})$

The above equation is the transfer function of a digital system or filters which contains only powers of z in its denominator and hence referred to as an all-pole system. The filter H(z) models the vocal tract response and e[n] denotes the vocal tract excitation [2].

VII. Speech Compression and Coding

In the early 1970s, PCM coding scheme for digital voice communication was standardized by CCITT (now ITU-T). Before vocoders, PCM and early forms of Linear Predictive coders were used.

A. PCM Encoder:

Before encoding starts, the analog signal is filtered into the frequency range of 300 to 3400Hz. This filtering process is in order that Nyquist Sampling Theorem (which states that the analog signal frequency should not be higher than half the sampling frequency) is satisfied. The analog signal is then sampled at 8 kHz sampling frequency and coded into 8bits word by a

© 2013, IJCSMC All Rights Reserved

quantizer and encoded. The 8 bit word is then converted in serial form and multiplexed with other PCM-coded voice signal into a 2048kbps rate, which contains 30 voice channels (European standard).

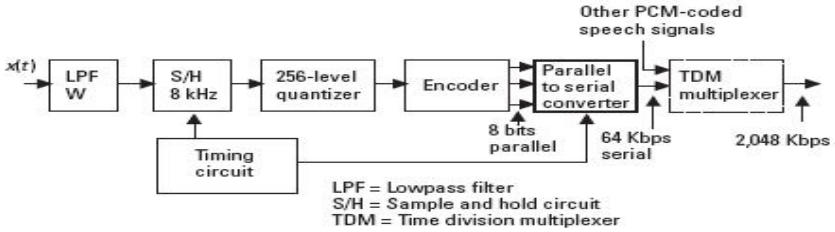


fig 7: PCM encoder [4]

B. Decoder:

At the receiver end, the de-multiplexer separates the incoming 64Kbps individual voice channel into the 8 bit sample with the aid of the Serial to Parallel converter. The sampled pulses are reconstructed to form the voice signal with close resemblance to the original.

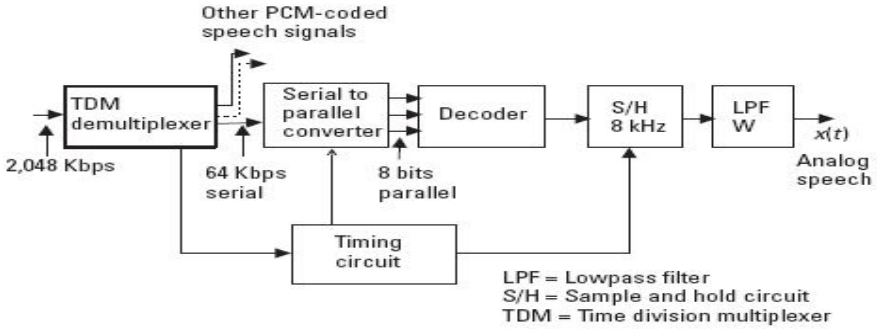


fig 8: PCM decoder [4]

Other speech coding methods are also in use with the PCM technique. They are:

C. Adaptive PCM:

In APCM, the signal strength information is transmitted at specific time intervals in addition to the sampled values. In this case, smaller number of bits is needed for samples which define the quantum level within a given scale. If the signal level is high, the quantizing error is high, since the same number of levels is used for all samples. This technique is used in GSM as part of voice coding process.

D. Differential PCM:

Here only the difference between a sample and the previous value is encoded. Since the difference is smaller than the overall value, fewer bits are needed compared to the original PCM with same accuracy; hence the bit rate is reduced. For better quality, DPCM uses several preceding samples to predict the next sample.

1. Differential Modulation (DM):

This is a simple type of DPCM. It transmits the binary value 1 if the sample is higher than the preceding one. 0 is transmitted if the value has decreased.

© 2013, IJCSMC All Rights Reserved

2. Adaptive DPCM:

This is a combination of APCM and DPCM. The qualities of commercial voice are coded at 32kbps or lower (24kbps or 16kbps).

All the above techniques operate directly on the time-domain signal. They achieve lower bit-rates by exploiting the sample-to-sample correlation in the speech signal. [2]

VIII. Vocoders

There are also frequency-domain coding techniques in which the speech is transformed from its time-domain state to frequency-domain, encoded and transmitted. The reverse process occurs in the receiver which converts the frequency-domain encoded signal back to the time-domain. Frequency-domain coders are classified as either Parametric or Non-parametric [2]. Most parametric coders are often referred to as Vocoders or analysis/synthesis coders. They contain an analysis stage for extracting a set of parameter which represents the vocal-tract excitation signal and the vocal-tract spectral envelope. These parameters are encoded and transmitted to the receiver which uses them to control a synthesis stage that reconstructs the original speech signal. [2] Some the parametric coders are Channel Vocoders, Formant Vocoders, Cepstral Vocoders and Linear Predictive Vocoders.

A. Channel Vocoders:

This consists of an analyzer and a synthesizer. The analyzer has a number of band-pass filters usually between 16 and 19, covering a speech frequency band of interest, and the outputs are smoothed. [2] The filter outputs are synchronously sampled every 10-30ms, hence deriving a short-time amplitude spectrum of the speech signal. It is multiplexed with a voiced/unvoiced decision and pitch frequency for voiced speech and transmitted to the receiver. [2]

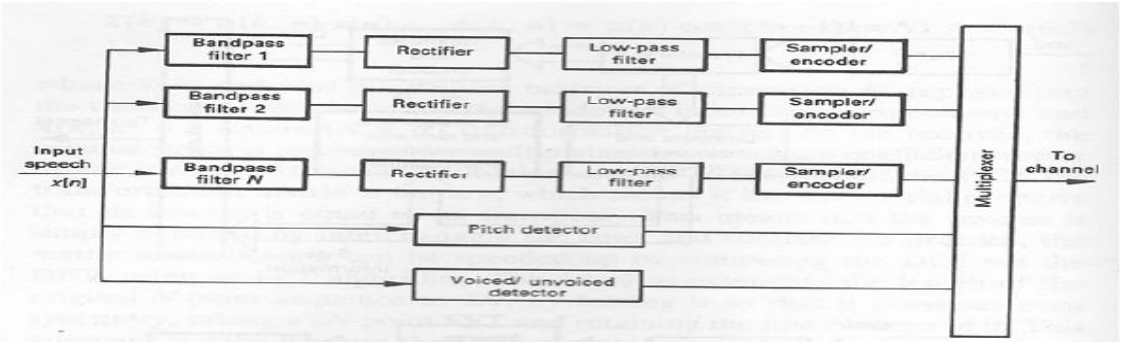


fig 9: – Block diagram of the Analyzer of a Channel Vocoder [2]

At the receiver, the vocal-tract and excitation information is multiplexed. The excitation information is used to switch in a random noise or pulse source and to set the frequency of the pulse source. [2] The signal received in each of the channels is used to modulate the amplitude of the excitation signal which excites the band-pass filter that corresponds to it. The band-pass filter outputs are added together to produced the synthesized signal. The frequencies of the filters used in the synthesizer are identical to those of the analyzer, but the bandwidths are less.[2]

© 2013, IJCSMC All Rights Reserved

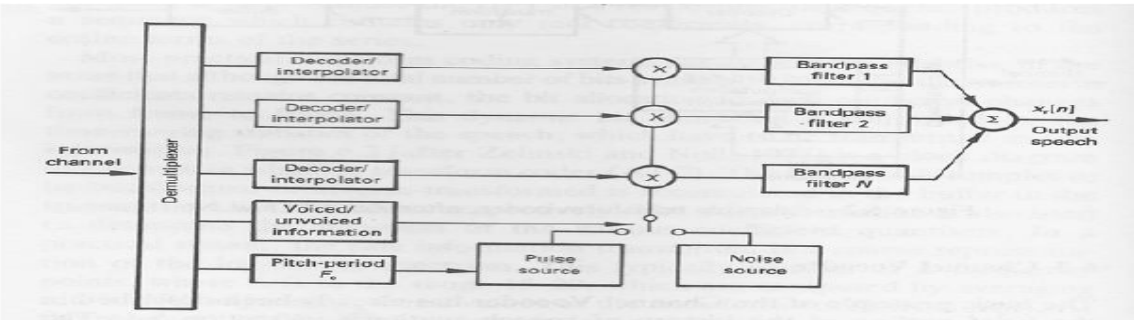


fig 10 – Block diagram of the Synthesizer of a Channel Vocoder [2]

B. Formant Vocoder:

A formant vocoder consists of an analyzer which extracts the time-varying formant and excitation parameters automatically from the speech signal. They are then encoded and passed to the synthesizer, where decoding takes place and used to control a formant synthesizer, which reconstructs the original signal. [2] Formant vocoder makes use of formant tracking algorithms. These have problems with the values of the formant frequency used, hence making them less successful. Recent test has shown that the performance of formant vocoders has improved through the use of neural networks which help to smooth the formant tracks.[2]

C. Cepstral Vocoder:

This technique involves separating the excitation and vocal-tract spectrum by inverse Fourier-transforming the log-magnitude spectrum which produces the cepstrum of the speech signal. [2] The low-frequency coefficients in the cepstrum all corresponds to the vocal-tract spectral envelope with the high-frequency excitation coefficients to form a periodic train at multiples of the sampling period. [2] The vocal-tract cepstral coefficients are separated from the excitation coefficients by linear filtering, which involve multiplying the cepstral coefficients by an appropriate window function. [2] Only the low-frequency excitation coefficients are retained after filtering. The time-position of the first pulse in the cepstrum gives the pitch period of the excitation signal and the availability of strong pulses which indicates whether the speech is voiced or unvoiced. [2]

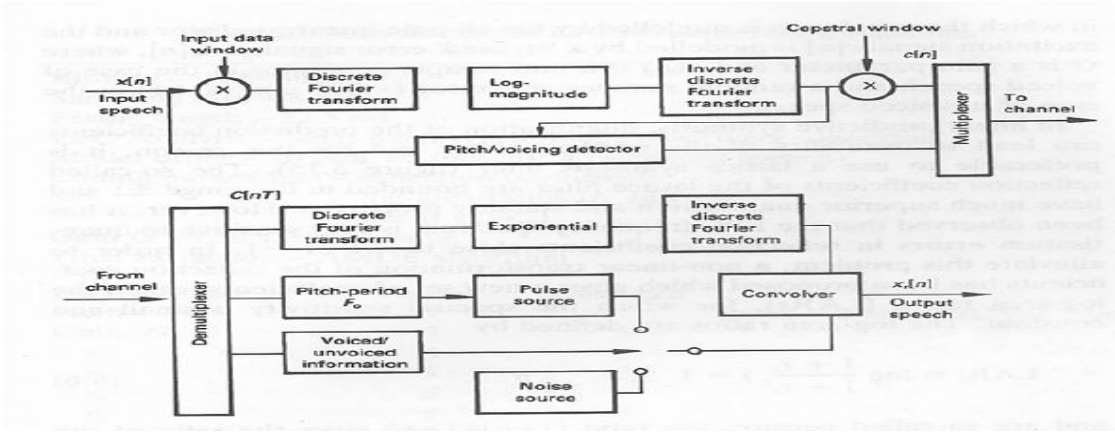


fig 11 - Cepstral Vocoder [2]

The above processes take place in the analyzer to derive the cepstral coefficients together with the pitch and voicing parameters, which are decoded and transmitted to the synthesizer.

© 2013, IJCSMC All Rights Reserved

At the receiver, the vocal-tract cepstral coefficients are Fourier-transformed, exponentiated and inverse Fourier-transformed to produce the vocal tract impulse response. The original speech is reconstructed by coevolving the impulse response with a synthetic excitation signal (random noise or a periodic pulse train). [2]

D. Linear Predictive Vocoders:

This is another method commonly used in coding speech signals and other dominating applications in the mobile communication. They are now embedded in several telephony and multimedia standards such G.729, G.723, IS-893, ISO 199 and so many others. As we already know, speech is produced by the interaction between the vocal tract and the vocal chords. From the time-representation of the speech signal we can model the slow varying spectral property of the upper tract by an all-pole filter and the prediction residual captures the voiced and unvoiced signal. The LP filter represented as A(z) is:

$$A(z) = 1 - \sum_{i=1}^{L} a_i z^{-i},$$

L = order of the linear predictor [6]

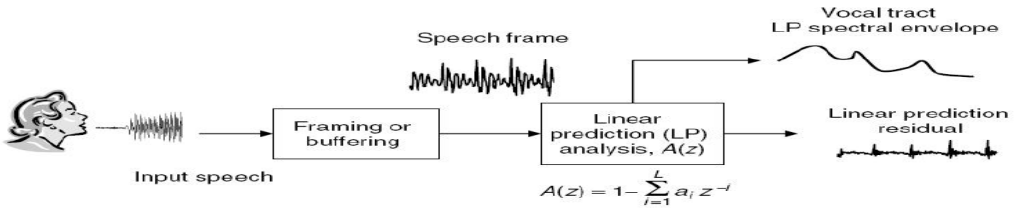


fig 12 - Linear Prediction [6]

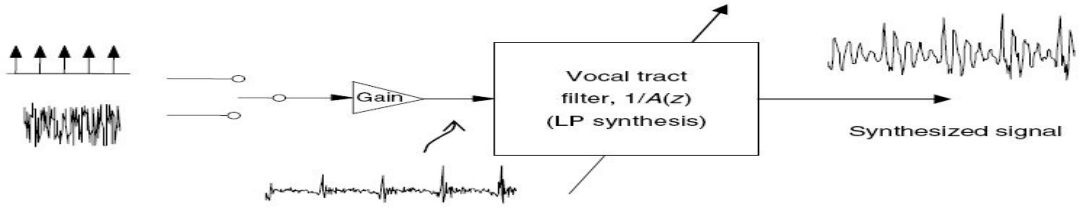


fig 13 - Engineering model for speech synthesis [6]

The figure above shows the simple speech synthesis model. A time varying digital filter is excited by quasi-periodic waveforms acting as the input (when the speech is voiced) and random waveform (when the speech is unvoiced). The inverse filter 1/A(z) is an all-pole linear prediction synthetic filter [6].

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{L} a_i z^{-i}}, \quad (11)$$

The frequency response related to the linear prediction synthesis filter which is the LPC spectrum represents the formants of the speech signal.

© 2013, IJCSMC All Rights Reserved

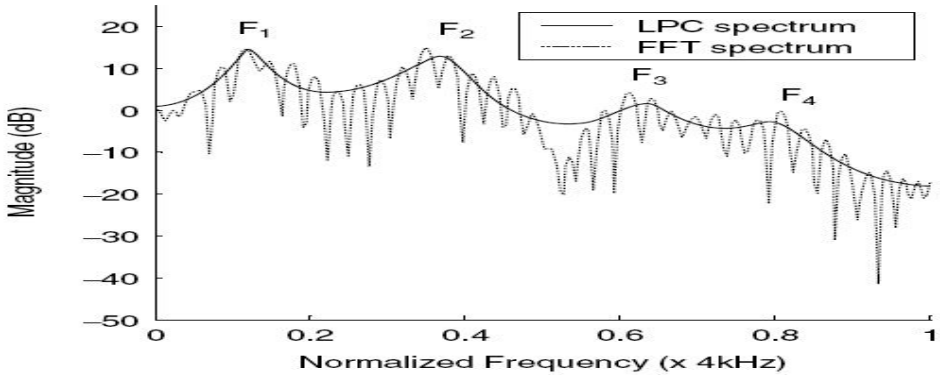


fig 14 - The LPC and FFT spectra (dotted line)

The formants represent the resonant modes of the vocal tract. The above figure shows that F1, F2, F3 and F4 represent four (4) formants.

1. Short-Term Linear Prediction:

Let say we have, an L-th order FIR linear predictor, we can calculate the linear combination of the previous samples, ŝ(n) by applying a forward linear prediction of s.

$$\hat{s}(n) = \sum_{i=1}^{L} a_i s(n-i) \quad (12)$$

where a_i are the LP coefficients.

The output of the LP analysis filter is called 'prediction residual', e(n) = s(n) – ŝ(n).

$$e(n) = s(n) - \sum_{i=1}^{L} a_i s(n-i). \quad (13)$$

[6]

In this case, since we are making use of only short delays, the linear predictor is called 'short-term linear predictor'. The linear predictor coefficients, ai are estimated using the least-square minimization of the error (prediction error) [6].

If we are using low-bit-rate coding, the prediction coefficients and the residual must be quantized efficiently because a_i do not have the adequate quantization property.

2. Long-Term Prediction (LTP):

This is the opposite of the short-term. This process captures the long-term correlation in the speech signal. LTP provides the means use to represent the periodicity in the speech signal; hence it also represents the harmonic structure in the short-term spectrum A(z). Using LTP for signal analysis requires the use of two parameters, the delay τ and the gain aτ . For strong voiced speech segments, the delay is usually an integer that approximates the pitch period.

Below is an equation used to represent the transfer function LTP filter given as Hτ (z). [6]

$$H_\tau(z) = \frac{1}{A_L(z)} = \frac{1}{1 - a_\tau z^{-\tau}} \quad (14)$$

© 2013, IJCSMC All Rights Reserved

3. ADPCM using Linear Prediction:

ADPCM is one of the simplest schemes that use the short-term linear prediction. Its algorithms encode the difference between the current and the predicted speech samples. In the ITU-T standards, ADPCM is also known as G.726 which works with 16, 24, 32 and 40kbps.

4. Open-Loop Analysis-Synthesis Linear Prediction:

In open-loop LP some of the parameters encoded and transmitted are the prediction coefficients, pitch information, frame energy and the voicing. At the receiver, the transmitted parameters are used to form the excitation e(n). This excitation is then used to excite the synthesis filter 1/a(z) to reconstruct the speech signal. Some standardized open-loop analysis-synthesis linear prediction algorithms are LPC10e also known as FS-1015 and the Mixed Excitation Linear Prediction (MELP). The diagram below is a set of LP open-loop synthesis used to capture the shape of the vocal tract and then model the excitation. [6]

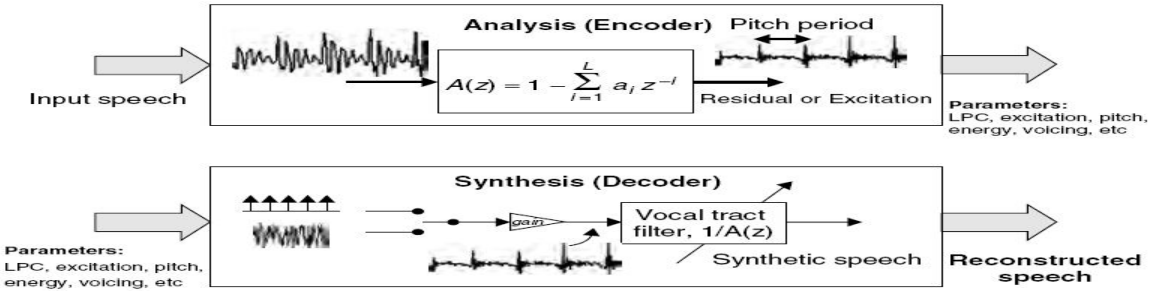


fig 15: Open-loop Analysis-Synthesis Linear Prediction [2]

5. Analysis-By-Synthesis Linear Prediction:

This is also called Closed-loop. Here the optimization process is used to determine the excitation process that minimizes the weighted mean-square-error (MSE) between the input speech signal and the reconstructed one. The closed-loop LP is the combination of spectral modeling properties of vocoders and the waveform of the coders, hence it is also called 'hybrid LP coder'. They consist of the short term

LP synthesis filter 1/A(z), a LTP synthesis filter $1/A_L$, perceptual weighting filter (PWF) and W(z) [6].

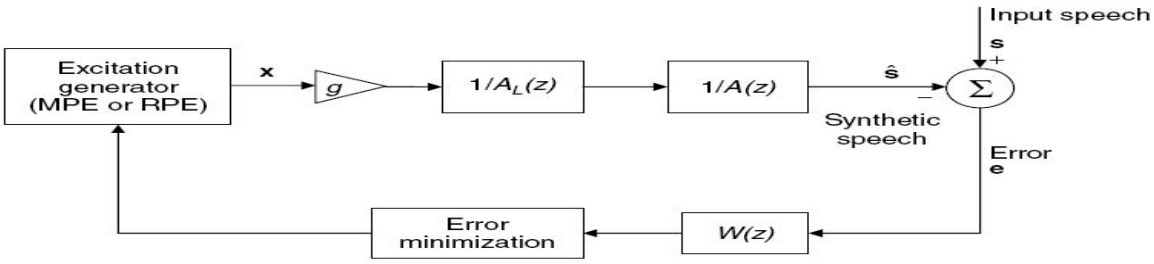


fig 16: Closed-Loop [2]

The three most common excitations models used as the excitation generator in a Closed-loop LP are Multi-pulse excitation (MPE), regular pulse excitation (RPE) and vector or code excited linear prediction (CELP). MPE has a rate of 9.6kbps and is used in Skyphone airline applications [6] and a 13kbps can be coded with RPE which is adopted by the European GSM full rate standard.

© 2013, IJCSMC All Rights Reserved

6. Code Excited Linear Prediction:

This is type of algorithm for closed-loop linear predictive coding is used for coding high-quality speech coding at low-rate. Since its proposal by Atal and Schroeder in the mid-1980s, it has undergone various evolutions as the technology improved.

- First-Generation (1986-1992):

The sets of algorithms operated at bit rates between 5.8kbps and 16kbps. They include the FS-1016 CELP (4.8kbps), IS-54 VSELP, ITU-T G.728 low delay CELP and the IS-96 Qualcomm CELP.

- Second-Generation (1993-1998):

This was targeted for TDMA and CDMA cellphones, internet audio streaming, VoIP and secure communications. Some of the algorithms include the ITU-T G.723.1 dual-rate speech codec, GSM EFR, IS-127 Relaxed CELP (RCELP) and the ITU-T G.729.

- Third-Generation (1999-present):

This is the standard that is been used in the 3G GSM network. The algorithms used are ETS198 (in Europe) and TIA (in the United States). They operate at multiple rates such as 12.2, 10.2, 7.95, 6.7, 5.9, 5.15 and 5.75kbps. The bit rate is adjustable depending on traffic conditions.

7. Linear Prediction in Wideband Coding:

Linear prediction can also be used in signal with bandwidth between 50Hz to 7 kHz. There are special standards of codec used to perform these coding in wideband bandwidth.

- ITU-T G.722:

This is the combination of both sub-band and ADPCM techniques; also called SB-ADPCM. Here, the input signal is sampled at 16 kHz and then it decomposes into two sub-bands of equal bandwidth using a Quadrature mirror filter (QMF). The low-frequency sub-band is quantized at 48kbps while the high frequency sub-band, at 16kbps.

- ITU-T G.722.2 AMR-WB:

This is an adaptive multi-rate wideband (AMR-WB) codec that operates at bit rates from 6.6 to 23.85kbps. The G.722 AMR□WB standard was develop for use in VoIP, 3G wireless communications, ISDN wideband telephony and audio/video teleconferencing. This standard is based on the ACELP coder and operates on audio frames of 20ms sampled at 16 kHz.

IX. Vocoders in Mobile Communication

When we speak through the speaker of a cell phone, the voice is sampled at the rate of 8 kHz with a resolution of 13bits, hence giving us a total bit rate of 104Kbps. This forms the input signal for the GSM speech codec. The current size has to be reduced to allow more subscribers make use of a given bandwidth of the GSM network. There are four different standard compression technologies used to analyze and compress the audio data. There are: Full□rate codec; Enhanced Full□rate (EFR); Adaptive Multi-rate (AMR); Half-rate codec.

© 2013, IJCSMC All Rights Reserved

A. Full-Rate Codec:

This is a modification of the linear predictive coder (LPC) known as a Regular Pulse Excitation and Long-term Prediction (RPE-LTP) that operates on a 20ms frame composed of 160 13 bit samples. It models the human vocal tract as a series of cylinders with different widths. Forcing air through these cylinders will generate speech sounds which are modeled into a set of simultaneous equations by the LPC coder. This equations show a slow decay in power as the frequency increases, thus the samples are fed through a pre-emphasis filter that improves the higher frequencies resulting in better transmission efficiency.

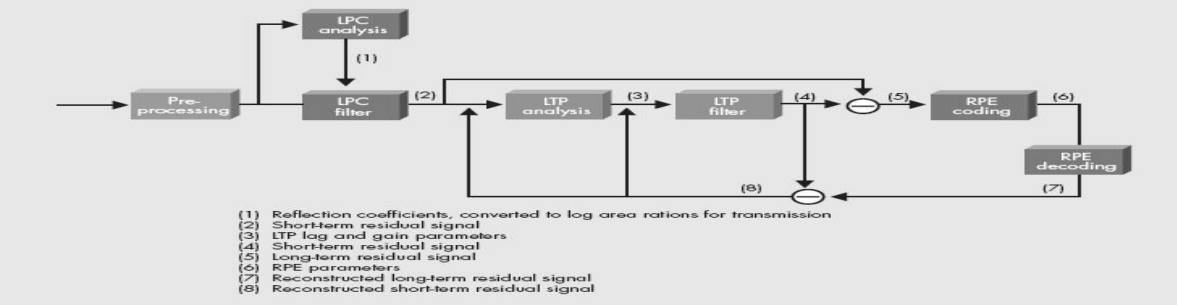


fig 17: Full-Rate Vocoder [7]

The diagram is a model of a full-rate vocoder. The model of the speech is generated as air is passing through the set of different size cylinders. The short-term analysis stage uses an approach called the autocorrelation to calculate a set of eight reflection coefficient that related to the eight cylinders in the model. A method known as Schur recursion is use to solve the set of equations efficiently. The parameters are then converted into 'log-area ratios' (LARs) that allows better quantizing into smaller number of bits. The coded LARs are then decoded back to the co-efficients and used to filter the input samples, resulting in 160 samples of residual signal. The 160 samples are then divided into 4 sub-windows of 40 samples each. The long-term predictor provides two parameters for each of the sub window called the 'lag' and the 'gain'. The lag is determined by the peak of the cross-correlation between the last two frames and current frame and the gain is determined by normalizing the cross-correlation co-efficient [4]. Both parameters are applied to a long-term filter and a prediction of the current short-term residual signal is made [8]. The RPE stage reduces the 40 long-term residual samples down to 13 parameters through a combination of sub-sampling (decimation) and interleaving. The resulting 13 parameters are coded using adaptive PCM where the maximum value is logarithmically coded into 6 bits and the 13 parameters represented as 3 bits each totaling 45bits. The last step is to update the short-term residual signal from the calculated long-term residual and analyzed signal. This data is used for calculating the next frame.

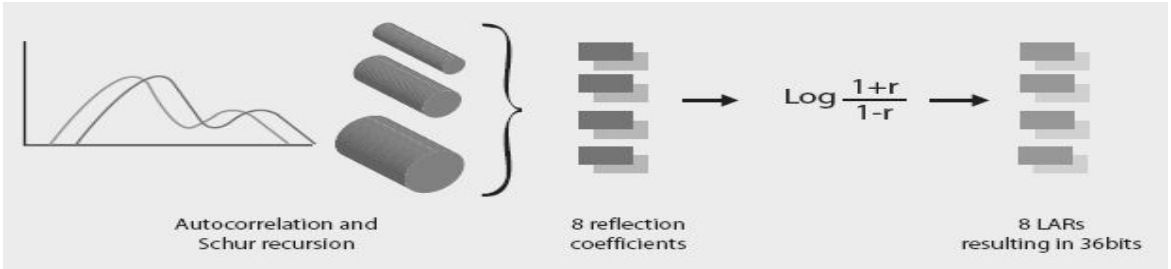


fig 18: Short-Term Analysis using the Autocorrelation and Schur recursion technique [7]

© 2013, IJCSMC All Rights Reserved

B. Enhanced Full-Rate Codec:

The EFR codec is an algebraic code excitation linear prediction (ACELP) codec [4]. It differs from the full-rate codec since it makes use of analysis-by-synthesis approach. The pre-processing stage consists of an 80Hz high-pass filter and some downscaling to make implementation less complex. Short-term analysis occurs twice per frame and consists of autocorrelation with two different asymmetric windows of 30ms in length concentrated around different sub-frames [8]. The resulting co-efficients are converted to line spectral pairs (for better transmission efficiency) and quantized to 38bits. The adaptive codebook contains excitation vectors that model the structure of the long-term speech [4]. The open-loop pitch analysis is performed to derive an estimate of the pitch lag for each frame. The open-loop result is then used to seed a closed-loop search for speed and reduced computational requirements [4][8]. The pitch lag values are applied to a synthesizer and compared against the non-synthesized input (analysis-by-synthesis). The minimum perceptually weighed error is found from the adaptive codebook and coded to 35 bits per sub-frame (lag) and 5 bits per Sub-frame (gain). The last stage is for the encoder to update the memory for the next frame.

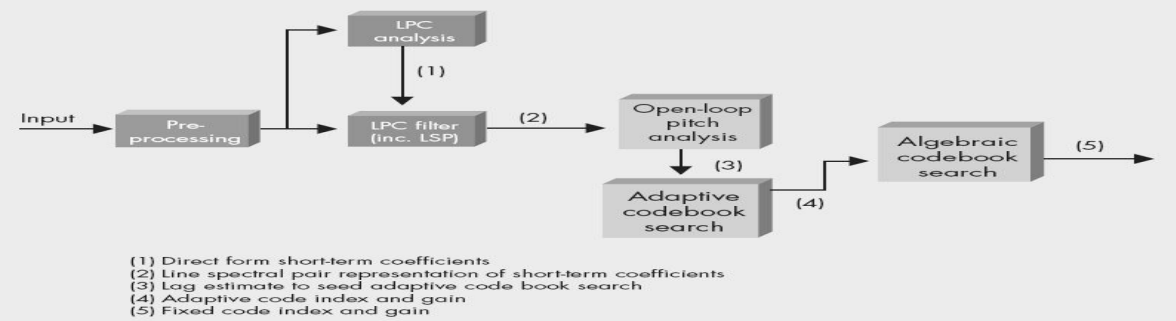


fig 19: Enhanced Full-Rate Vocoder Model [7]

C. Adaptive Multi-Rate Codec:

The AMR codec comprises of ACELP vocoders operating from 12.2Kbps down to 4.75Kbps [4]. Rates lower than 12.2Kbps, short-term analysis is performed only once per frame. Rates from 5.15Kbps and lower, open-loop pitch lag is estimated only once per frame [4]. In poor conditions (radio transmission), the 4.75Kbps codec will still be recoverable after the full-rate and EFR frames are lost.

D. Half-Rate Codec:

They make use of a 'Vector Sum Excitation Linear Prediction' (VSELP) codec which operates on an analysis-by-synthesis approach similar to the EFR and AMR codecs and resulting to an output of 5.7Kbps [4][8]. The output frame contains 2bits that is, whether the frame contain voice or no voice [4]. The operation is slightly different for each mode, allowing for the best quality of the audio data. Public view to half-rate speech is poor, so the technology is not in use today. However, due to its adaptive modes, the AMR vocoder lower rate (4.75Kbps to 7.95Kbps), fit within the available capacity of half-rate channel. The purpose is to use the half-rate channels with AMR to become much more widespread in high traffic areas.

The table below was obtained from the International Telecommunications Union, Radio Sector (ITU-R) from the result of various test performed. It also shows us when a particular kind of vocoder is implemented.

© 2013, IJCSMC All Rights Reserved

Table 1: Speech Coding Systems

System Type:	System Name:	Ref.:	Codec Type:	Codec Rate (kbit/s):	Forward Error Correction (kbit/s):	Codec Algorithmic Delay (mS):	Est. Quality (MOS):	Est. Processing:
Cellular TDMA	GSM/DCS/PCS Full Rate	ETSI/ETS 300580 ANSI J-STD-007	RPE-LTP	13	9.8	40	3.6-3.8	2.5 Mips
Cellular TDMA	GSM/DCS/PCS Half Rate	ETSI/ETS 300581	VSELP	5.6	5.8	40	3.5-3.7	17.5 Mips
Cellular TDMA	GSM EFR US PCS1900 EFR	ETSI/ETS 300723 ANSI J-STD-007A	ACELP	12.2	10.6	40	4.1	15.4 WMops
Cellular TDMA	D-AMPS Full Rate	TIA/EIA IS-85	VSELP	8.0	5.0	28	3.7	22 Wmops
Cellular TDMA	D-AMPS Enhanced	TIA/EIA IS-641	ACELP	7.4	5.6	25	4.1	14 Wmops
Cellular TDMA	PDC Full Rate	RCR-STD-27	VSELP	6.7	4.5	Unavailable	3.40	7.8 Mops
Cellular TDMA	PDC Half Rate	RCR-STD-27	PSI-CELP	3.45	2.15	Unavailable	3.34	18.7 Mops
Cellular CDMA	Composite CDMA/TDMA	TIA/EIA IS-661	CELP-Like	7.2	3.2	26	>4.0	11 Mips
Cellular CDMA	CDMA	TIA/EIA IS-96	CELP	8/4/2/0.8	19.2	27.5	3.3	22 Mops
Cellular CDMA	CDMA	TIA/EIA IS-127	RCELP	8/4/0.8	19.2	30	4.1	20 Mops
Cellular CDMA	W-CDMA	USA Unavailable	ADPCM	32	0	Unavailable	Unavailable	Unavailable
Digital Cordless	CT2	ITU-T-G.726	ADPCM	32	0	0.25	4.0	10 Mips
Digital Cordless	DECT	ITU-T-G.726	ADPCM	32	0	0.25	4.0	10 Mips
Digital Cordless	PHS	ITU-T-G.726	ADPCM	32	0	Unavailable	Unavailable	1.0 Mops
Dispatch	Project 25 U.S.	EIA/TIA IS-102.BABA	IMBE	4.4	2.8	80	3.4	6.9 Mips
Dispatch	TETRA Europe	ETSI/ETS 300395	ACELP	4.567	2.633	Unavailable	3.3-3.5	15 Mips
Dispatch	IDRA Japan	RCR STD-32A	CSELP	4.72	2.766	Unavailable	3.2	7.0 Mops
Dispatch	IDRA Japan	RCR STD-32A	VSELP	4.2	3.177	Unavailable	3.20/3.98	8.0 Mops
Dispatch	DIMRS Canada	Unavailable	VSELP	4.2/8.8	3.177/6.756			8.0 Mops

X. Matlab Function and Test.

Matlab can be used to illustrate how a vocoder processes the human voice. The vocoder is made up of an analyzer and synthesizer. The analyzer models the vocal-tract and encodes the parameters extracted from the vocal organ. The synthesizer decodes these parameters and reconstructs the original speech. This process is done in 3 stages on Matlab:

Stage 1

Input Speech Signal ⟶ Waveform ⟶ Spectrogram

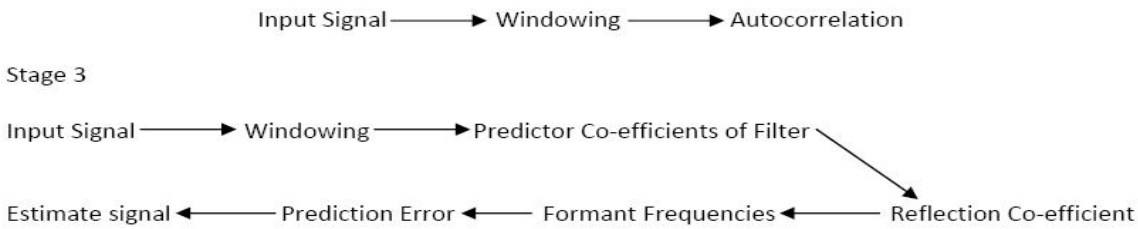
This is the first stage of analyzing the speech signal. The waveform is the graphical representation of the signal, showing the time and amplitude. The spectrogram is a three-dimensional plot illustrating how the frequency spectrum varies with time. The colour of the spectrogram also shows the amount of power on the speech at a particular time.

Stage 2

Input Signal ⟶ Windowing ⟶ DFT ⟶ Log ⟶ IDFT ⟶ Cepstrum

The cepstral analysis is used to determine the Fundamental frequency of the speech signal during analysis in the frequency-domain. The fundamental frequency can also be determined in the time-domain by using the Autocorrelation function.

© 2013, IJCSMC All Rights Reserved



The above process shows the steps to determine the formant frequencies and the estimated signal using the Linear Prediction Approach.

Using Matlab to demonstrate the steps above involves some functions that will be used in getting the result.

A. Matlab Functions:

1. "wavread" - Reads Microsoft wave (".wav") sound file
2. "specgram" ("spectrogram" for Matlab R14") – Time-dependent frequency analysis (Spectrogram)
3. "fft" – Discrete Fourier Transform. The function X = fft(x) implement the transform given for the vectors of length N by: $X(k) = \Sigma\, x(j) w_N^{(j-1)(k-1)}$ where $w_N = e^{(-2\pi i)/N}$
4. "hamming" – Hamming Window. w = hamming (n) returns an n-point symmetric hamming window in the column vector w. 'n' should be a positive integer. The coefficients of a hamming window are computed from the following equation:

$w[k+1] = 0.54 - 0.46\cos(2\pi k/n-1)$ k=0, ……., n-1

5. "fprintf" – Write Formatted data to file. Syntax: count = fprintf(fid, format, A, …) formats the data in the real part of matrix A (and in any additional matrix arguments) under control of the specified format string, and writes it to the file associated with the identifier.

6. "xcorr" – Cross-correlation Function. "xcorr" estimates the cross-correlation sequence of a random process. The true crosscorrelation sequence is:

$Rxy(m) = E\{x_{n+m} y^*_n\} = E\{x_n y^*_{n-m}\}$ where x_n and y_n are jointly stationary random processes, $-\infty<n<\infty$, and E{.} is the expected value operator. "xcorr" must estimate the sequence because in practice, only a finite segment of one realization of the infinite-length random process is available.

7. "lpc" – Linear Prediction Filter Coefficients: This determines the coefficients of a forward linear predictor by minimizing the prediction error in the least squares sense. It has applications in filter design and speech coding.

8. "freqz" – This computes the frequency response o discrete-time filters, adaptive filters and multi-rate filters).

9. "phasez" - This returns the unwrapped phase response for an adaptive filter, discrete filter or Multi-rate filter (Digital Filter Phase Response).

10. "schurrc" – This computes reflection coefficients from autocorrelation sequence.

B. Testing:

The vocoder as applied in communications is used to send the speech signals from the transmitter to the receiver by modeling the vocal-tract. The parameters from the model are used to create a synthetic signal which is a reconstruction of the

© 2013, IJCSMC All Rights Reserved

original signal sent. Some the parameters derived from the model are: Pitch-period; Fundamental Frequency and Formant Frequencies.

1. Waveform:

The waveform of the signal is the graphical representation of the speech. It is displayed on Matlab using the function:

[y,fs,nbit]=wavread('speech file'); The 'y' represents the speech file which in .wav format, 'fs' is the sampling frequency of 44100Hz and 'nbit' is the number of bit of the sample which in this case is 8. 'wavread' is the matlab command to 'speech file'.

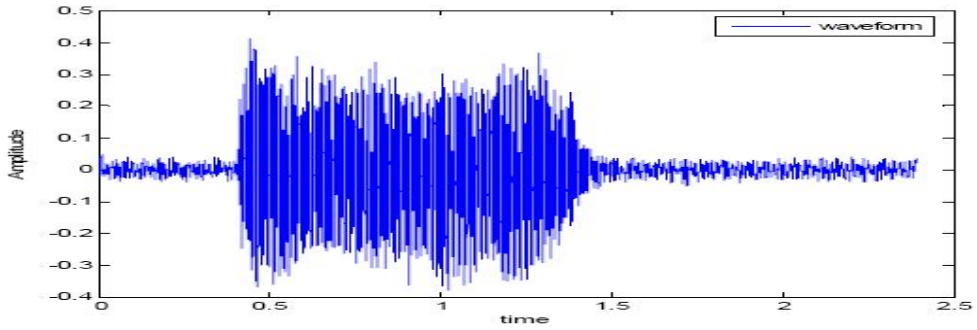


fig 20. – Waveform of the the speech signal

2. Spectrogram:

The spectrogram is used display important frequency content information by colours showing the regions of high content. This is represented in matlab by the following commands: specgram(y,256,fs,256,252); newer version of matlab no longer recognized the 'specgram' but instead makes use of 'spectrogram' and the arrangement of the parameters are also in a different order.

Spectrogram(y,256,252,256,fs);

The general form of the spectrogram code is as follows:

specgram(y,NFFT,fs,WINDOW,NOVERLAP);

The parameters y, NFFT, fs, WINDOW and NOVERLAP corresponds in the same other with y,256,fs,256,252

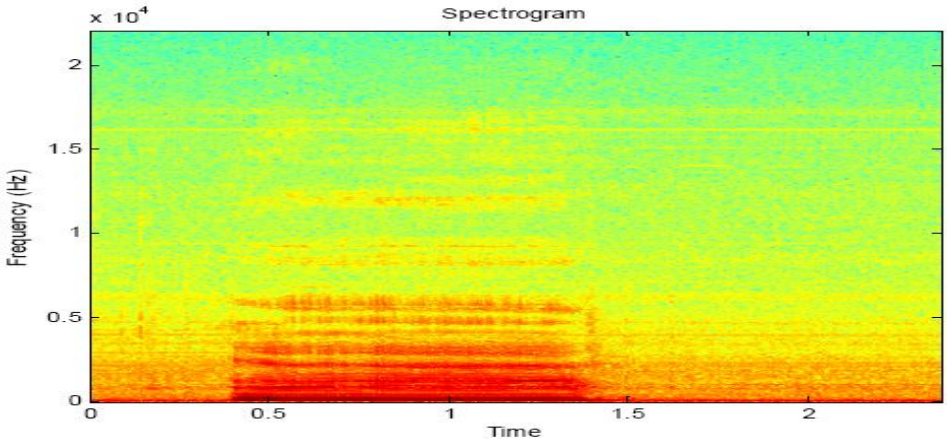


fig 21 Spectrogram

© 2013, IJCSMC All Rights Reserved

3. Cepstrum:

This is achieved by first calculating the DFT of the windowed speech segment and taking the logarithms which produces the log-magnitude spectrum. Thereafter the inverse DFT of the log-magnitude spectrum is calculated to produce the cepstrum. For a voiced speech signal, a strong peak is manifested on the graph which represents the fundamental frequency of the signal. Whereas for the unvoiced, there is hardly any peak point manifested hence from the cepstral calculations, you can state if the speech is voiced or unvoiced. In matlab, this can be done with codes in stages.

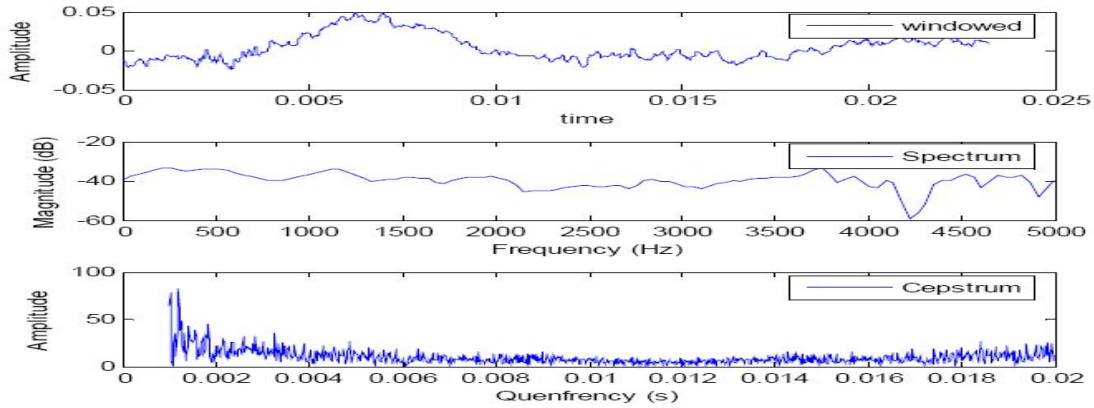


fig 22: Cepstral Analysis

4. Autocorrelation Function:

The Auto Correlation function shows the time domain alternative to the Cepstrum calculation. It also calculates the reflection coefficients from the auto correlation function, which are used in the GSM LTP RPE coders.

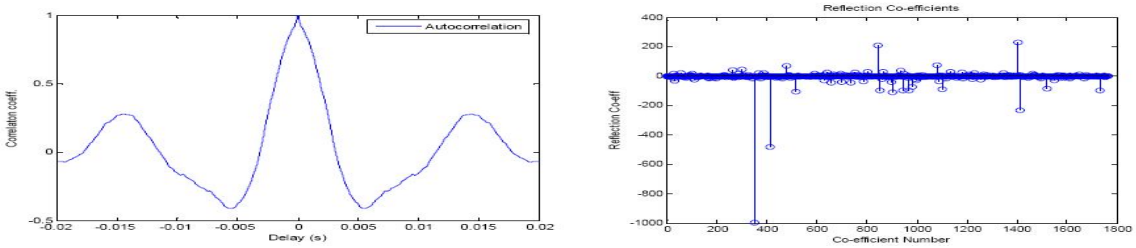


fig 23– Autocorrelation Function and Reflection Coefficients plot

5. Linear Prediction:

This is a very important technique used in analyzing speech signals. The first step is to determine the coefficient of the filter and determine the predictive polynomial of the filter. This is done with the matlab Codes with a set of predictors which are carried out every 10 – 30ms by solving the polynomial equations using autocorrelation. The result is presented as the frequency response and phase response of the filter (vocal tract). This also represents the formant frequencies of the filter.

© 2013, IJCSMC All Rights Reserved

XI. Result: Male Voice

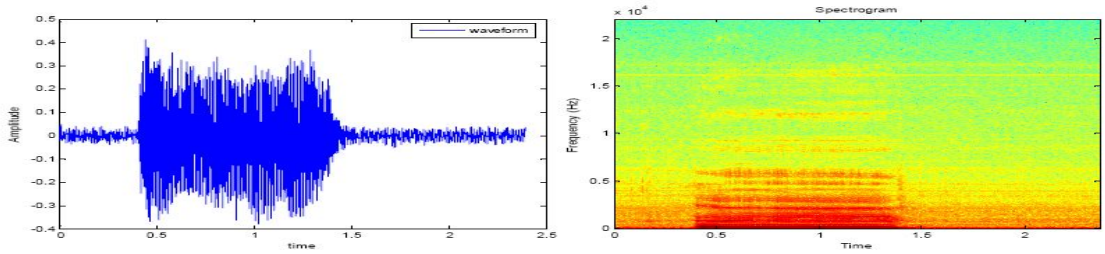


fig 24: Waveform and Spectrogram

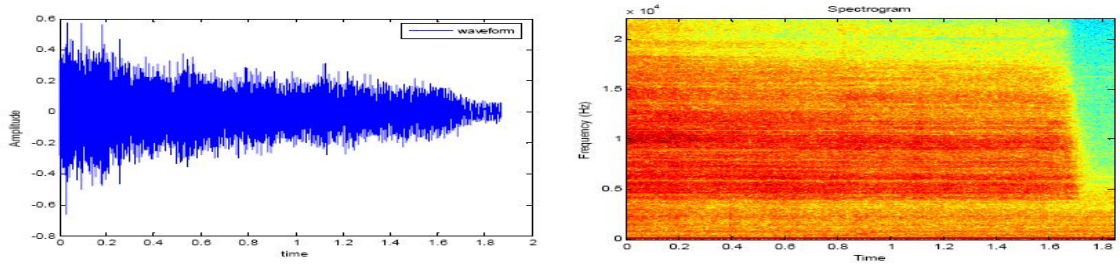


fig 25: Waveform and Spectrogram

(Fig 24 Waveform and Spectrogram of Male Voiced sound – 'a'. Fig.25 Waveform and Spectrogram of Male Unvoiced sounds – 'tt')

A. Fundamental Frequency:

In modeling the vocal tract, one of the major parameters needed is the fundamental frequency. This can be determined in the frequency and time domains respectively. The technique used in the frequency-domain is called Cepstrum and that used in the time-domain is called the Autocorrelation Function.

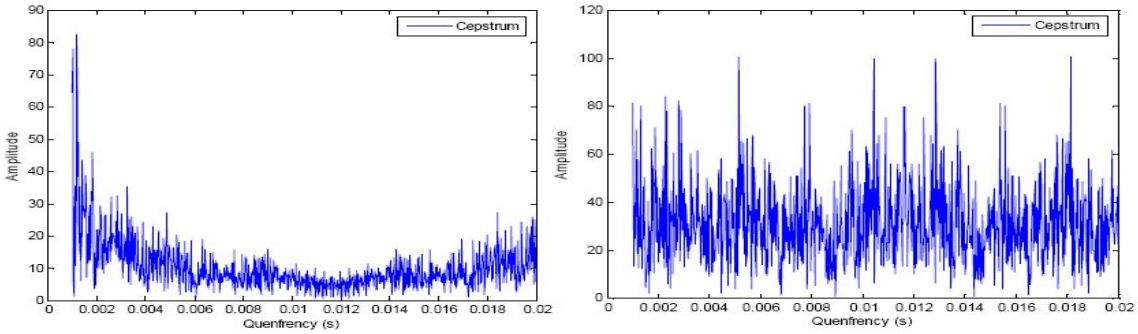


fig 26 – Fundamental Frequency of voiced sounds 'a' and 'tt' using Cepstrum

The table below shows the fundamental frequencies of the sounds 'a', 'e', 'I', 'ch', 'sh' and 'tt'. These are the maximum points on the cepstrum graphs shown above ('a' and 'tt')

© 2013, IJCSMC All Rights Reserved

Table 2: Fundamental frequencies of voiced and unvoiced sounds using the Cepstrum Technique

	Sounds	Fundamental Frequency (Hz)
1.	A	846.449
2.	E	772.329
3.	I	1000
4.	Ch	830.508
5.	Sh	733.777
6.	Tt	194.188

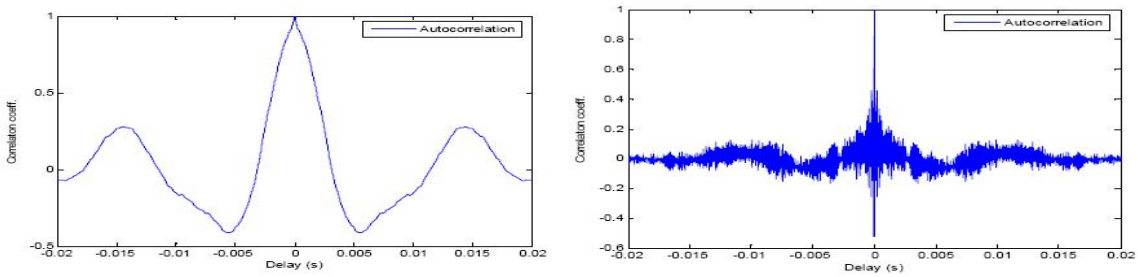


fig 27: Autocorrelation Peak for Voiced and Unvoiced Sounds 'a' and 'tt' using Autocorrelation Function

B. Cepstrum of Male Speaker:

Above is the graphical representation of the Cepstrum for the sounds 'a' and 'tt' from a female speaker. The fundamental frequency is the peak point on the Cepstrum graph. From the figures above, you will observe that the cepstrum of 'a' has one peak point, hence that is the fundamental frequency of 'a' and also, it is a voiced sound. On the other hand, the cepstrum of 'tt' the graph appears like that of a noise. Hence it is difficult to determine the fundamental frequency of 'tt'. This is an unvoiced sound.

C. Formant Frequency of Male Speaker:

This can be found by using the time-domain technique called Linear Prediction. Linear Prediction also uses the autocorrelation method to determine the predictor coefficients.

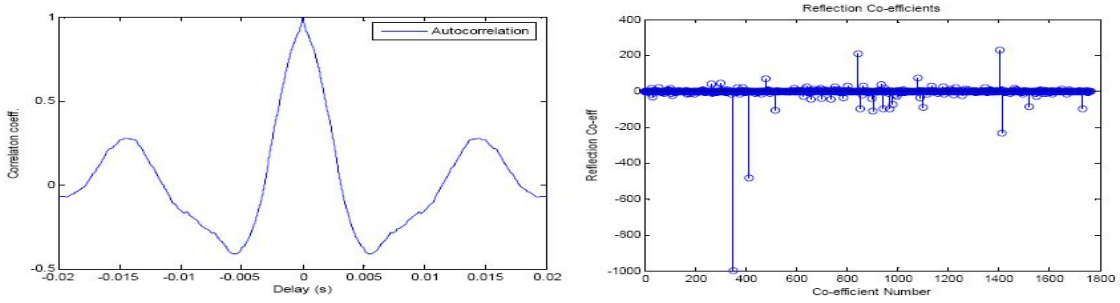


fig 28: Autocorrelation and Reflection Co-efficient of Voiced sound "a" (male)

© 2013, IJCSMC All Rights Reserved

The polynomial equations represent the filters of the vocoder used to analyze and synthesize the speech signal. The polynomial equations are shown in the table below:

Table 3: Polynomial Predictor Equations for a Male Speaker

sounds	Polynomial Equations
a	$xe^{46}-1.637xe^{45}+1.134xe^{44}-0.843xe^{43}+0.551xe^{42}-0.277xe^{41}+0.159xe^{40}$
e	$xe^{46}-1.733xe^{45}+0.876xe^{44}-0.405xe^{43}+0.519xe^{42}-0.363xe^{41}+0.216xe^{40}$
i	$xe^{46}-1.571xe^{45}+0.456xe^{44}+0.150xe^{43}+0.122xe^{42}-0.097xe^{41}-0.026xe^{40}$
sh	$xe^{46}-1.967xe^{45}+1.274xe^{44}-0.123xe^{43}-0.207xe^{42}+0.079xe^{41}+0.072xe^{40}$
ch	$xe^{46}-1.557xe^{45}+0.461xe^{44}+0.093xe^{43}+0.122xe^{42}-0.113xe^{41}+0.023xe^{40}$
tt	$xe^{46}-0.669xe^{45}+1.251xe^{44}-0.792xe^{43}+0.885xe^{42}-0.546xe^{41}+0.481xe^{40}$

The formant frequencies are extracted from the frequency response graph. They are the peak values of the graph.

Table 4: First Five Formant Frequencies for Male Speaker

	1st Formant	2nd Formant	3rd Formant	4th Formant	5th Formant
a	1173.6	2169.9	3092.6	4085.3	5245.2
e	422.0	1320.8	2160.5	3534.8	4075.3
i	1352.1	2395.8	3610.4	4603.9	4975.9
sh	156.1	1287.9	2405.9	2967.7	4095.2
ch	1116.7	2178.4	2876.3	3898.7	5009.5
tt	1011.8	2225.6	3814.2	4597.3	5328.2

XII. Conclusion.

From the tests carried out on the Matlab demonstration with several different sound samples from male speakers, it generally produced a high level of accuracy. A voiced speech can be differentiated from an unvoiced speech using the Cepstrum technique. Also the results show that, the original signal modeled can be reconstructed using the parameters such as the fundamental frequency which is obtained using either the frequency-domain approach (Cepstral Analysis) or the time-domain approach (Autocorrelation function) and formant frequencies was determined by the Linear Prediction Methods.

A more effective codebook demonstration, showing more interactively how a codebook is used in modern CELP coders; and a GUI based Vocal tract area function, obtaining the cross sectional area of the vocal tract from the speech sample are possible areas of improvement in vocoder technology.

© 2013, IJCSMC All Rights Reserved

Reference

1. L. Besacier, S. Grassi, A. Dufaux, M. Ansorge and F. Pellandini, "GSM Speech Coding and Speaker Recognition", IEEE Paper, 2000.

2. Owens, F.J., (1993). Signal Processing of Speech. 1st Edition, MacMillan Press LTD.

3. Fernando Trujillo, "English Phonetics and Phonology",http://www.ugr.es/~ftsaez/fonetica/stress.pdf.

4. Oppenheim A.V., Schafer R.W., and Buck J.R., (1999). Discrete-Time Signal Processing. 2nd Edition, Prentice Hall International Plc.

5. Tarmo Anttalainen, (2003) "Introduction to Telecommunications Network Engineering", Second Edition Artech House , pp. 92-114

6. Andreas Spanias, Ted Painter and Venkatraman Atti, (2007)"Audio Signal Processing and Coding", John Wiley and Sons, Inc., pp. 91-106

7. Owen, F.F.E., (1982). PCM and Digital Transmission Systems. 1st Edition, McGraw-Hill Book Company

8. Wade J.G., (1987). Signal Coding and Processing: An Introduction Based on Video Systems. 1st Edition, Ellis Horwood Limited.

© 2013, IJCSMC All Rights Reserved