Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

6th International Conference
on Electrical Engineering
ICEENG 2008

Speech signal reconstruction from modified STFT magnitude spectra using homomorphic analysis/synthesis

By

Mahmud E. Gadallah * Alaa Rohaiem * Ayman El Gezawy *

Abstract:

This work presents a low complexity method of reconstructing a speech signal from a homomorphic Cepstrum Coefficients. The homomorphic deconvolution has been the main core to explore the vocal tract envelope characteristics as well as the excitation signal Pitch frequency in order to simplify the source speech signal as minimum as possible. This work usually used for the purpose of speech coding in order to minimize the communication bandwidth. In this work we have avoided some analytical and reconstruction steps that used to cost extra processing time causing extra delay. Those steps are usually used in the most existing methods for this kind of work. Those steps include the samples overlapping lag and Iterative phase retaining algorithm. A set of harmonic sine waves based on the pitch frequency is used as an excitation signal to the vocal tract filter in order to regenerate the speech signal. The intelligibility of the reconstructed signal obtained is 3.5 in MOS of a subjective test, although the speech signal was still recognizable.

Keywords: Cepstrum Coefficients, Speech signal reconstruction, Pitch extraction

* Military Technical College

1. Introduction:

The acoustic model of speech (Fig.1) is the basis for most speech models and their analysis. Speech is produced by excitation of an acoustic tube (vocal tract) which is terminated on one end by the lips and on the other end by the glottis. There are three basic classes of speech, which are Voiced sounds, Fricative and Plosive. Fricative and plosive sounds together make unvoiced speech. For purposes of analysis, voiced speech is considered periodic within the analysis frame. The vocal tract is an acoustic transmission system characterized by its natural frequencies called formants. These formants correspond to the resonant frequencies in the frequency response of the vocal tract. In normal speech, the vocal tract changes relatively slow with time as the tongue and lips perform the gesture of speech. Thus it can be modeled as a slowly varying filter that imposes its frequency response properties on the spectrum of speech.

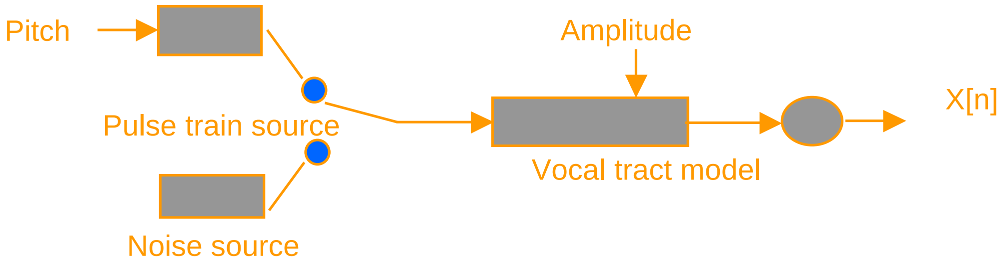

Figure (1):*The acoustic model of speech*

Speech communication has traditionally been made through the use of low bit-rate speech codecs. The low bit-rates at which these codecs operate causes a slight distortion to be introduced onto the speech signal. An improvement is to transform the codec vectors themselves into features for speech recognition without using the time domain signal. A number of techniques [1, 2] have been suggested for this. For example, in [1] codec vectors based on line spectral pairs (LSPs) are converted into linear predictive coding (LPC) coefficients. A Fourier transform then converts these into the magnitude spectrum, from which MFCCs can be estimated via the standard mel-filter bank analysis, logarithm and discrete cosine transform. Mel-frequency cepstral coefficients (MFCC) have been commonly used by speech recognition systems [3]. MFCC is based on a mel-scale transformation (among similar scales such as Bark and ERB) in order to model the sensitivity of the human ear more closely. A set of processing steps should be achieved in order to extract the MFCC vector of coefficients. The main building blocks of the MFCC are illustrated in Fig.2 [6], mel filter bank is shown in Fig.3 [5]. The MFCC vector of coefficients are between 13 and 23 coefficients, depending on the number of filters which depends on the sampling rate[4].

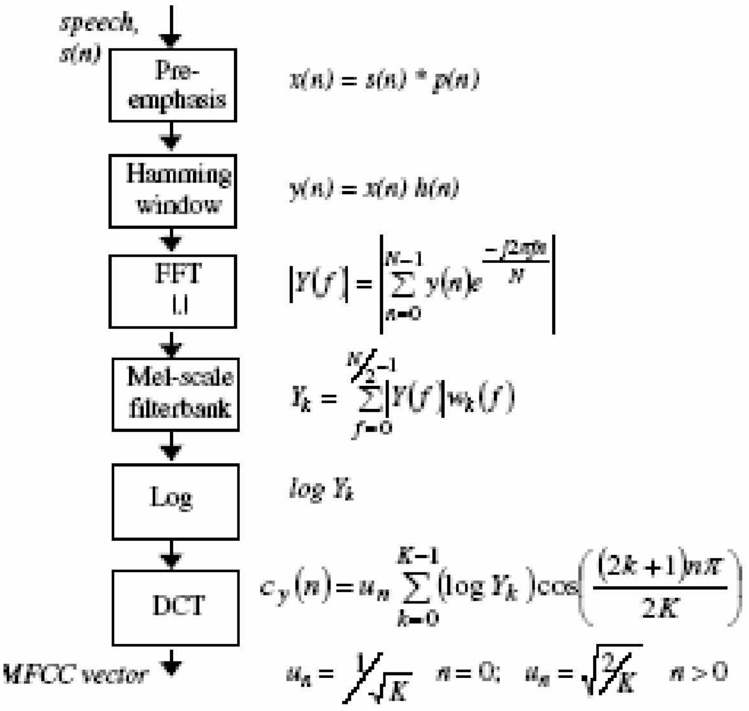
speech, s(n)
Pre-emphasis
x(n) = s(n) * p(n)
Hamming window
y(n) = x(n) h(n)
FFT |.|
$|Y(f)| = \left|\sum_{n=0}^{N-1} y(n) e^{\frac{-j2\pi fn}{N}}\right|$
Mel-scale filterbank
$Y_k = \sum_{f=0}^{N/2-1} |Y(f)| w_k(f)$
Log
log Y_k
DCT
$c_y(n) = u_n \sum_{k=0}^{K-1} (\log Y_k) \cos\left(\frac{(2k+1)n\pi}{2K}\right)$
MFCC vector
$u_n = 1/\sqrt{K}$ $n = 0$; $u_n = \sqrt{2/K}$ $n > 0$

Figure (2): Block *diagram of MEL-vector extraction*

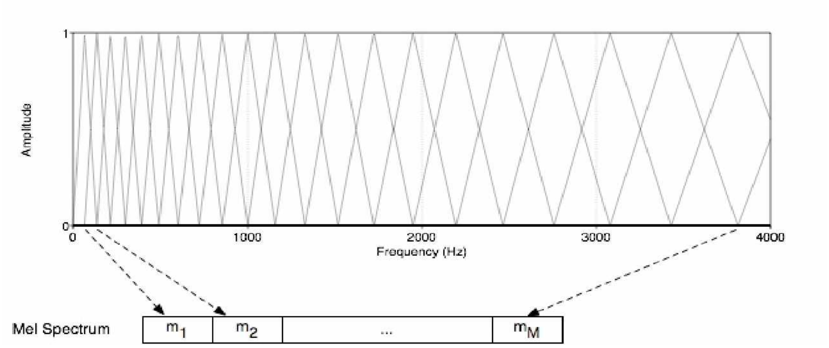

Figure (3): *MEL-filter bank*

Homomorphic analysis towards cepstrum coefficients extraction is similar, in away, to mel analysis. The following Fig.4 illustrates the block diagram of the vocal filter cepstrum representation using homomorphic analysis. Both MFCC and Homomorphic analysis lose the phase information during the action of the analysis steps. Estimating the phase once again depending on the magnitude or the pitch frequency component has a big role in reconstructing a high quality speech signal.

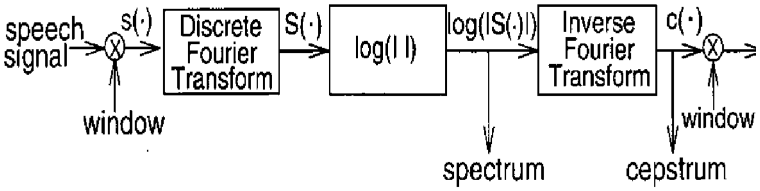

Figure (4): *Block diagram of Homomorphic analysis*

2. FEATURES EXTRACTION

Reconstruction of a time-domain signal from only the magnitude of the short-time Fourier transform (STFT) is a common problem in speech and signal processing. Many applications, including time-scale modification, speech morphing, and spectral signal enhancement involve manipulating the STFT magnitude, but do not clearly specify how to adjust the phase component of the STFT in order to invert back into the time domain. Indeed, for many STFT magnitude modifications, a valid inverse of the STFT does not exist and a reasonable guess must be made instead.

In this section we presented how we extracted the cepstral low time components that represents the vocal tract envelop. The speech signal was abstracted using a phone quality setting, 8 KHz sampling rate representing each sample with 8 bits. Fig.5 illustrates the block diagram of the algorithm done.

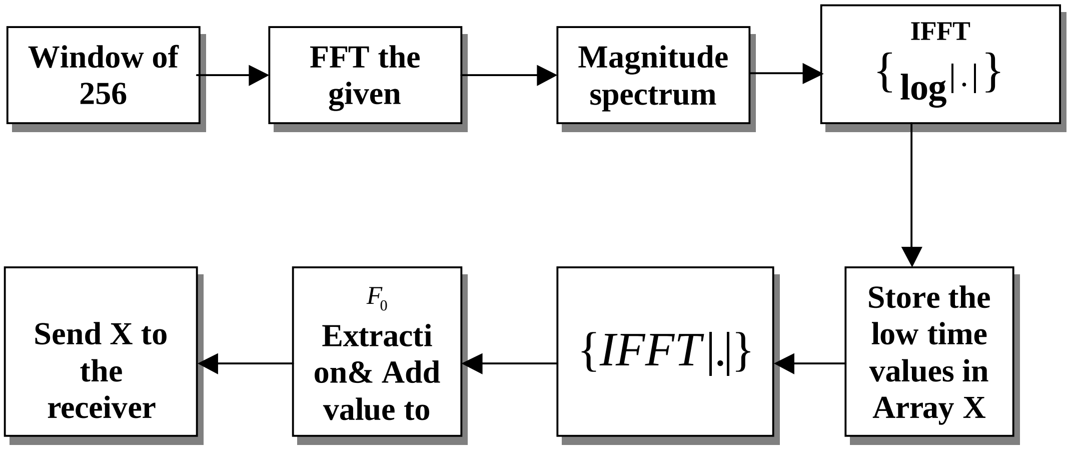

Figure (5): *Block diagram of Cepstral and fundamental frequency extraction*

Figure 6 presents the fine structure output of the inverse FFT of the magnitude (Cepstrum), while fig.7 presents the traditional pitch estimation from inverse log magnitude(notice the high peak in the high time)[7].

Figure 8 presents our pitch estimation from the squared IFFT for the magnitude value.

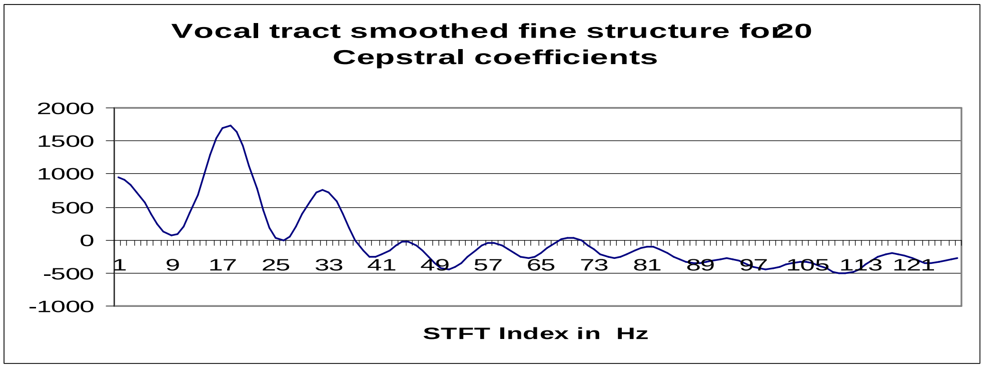

Figure (6)

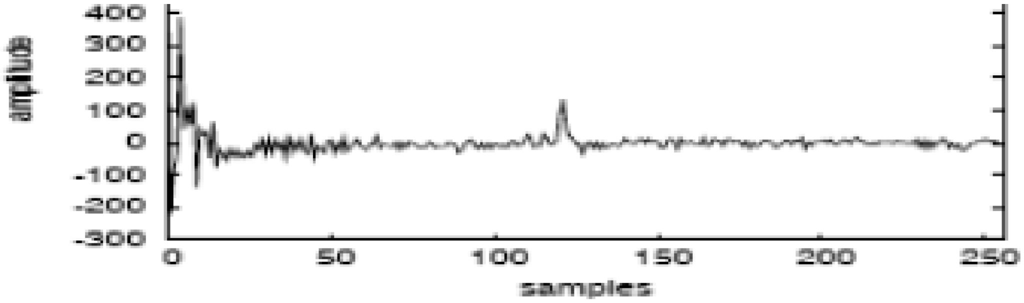

Figure (7): *Traditional Cepstrum pitch estimation*

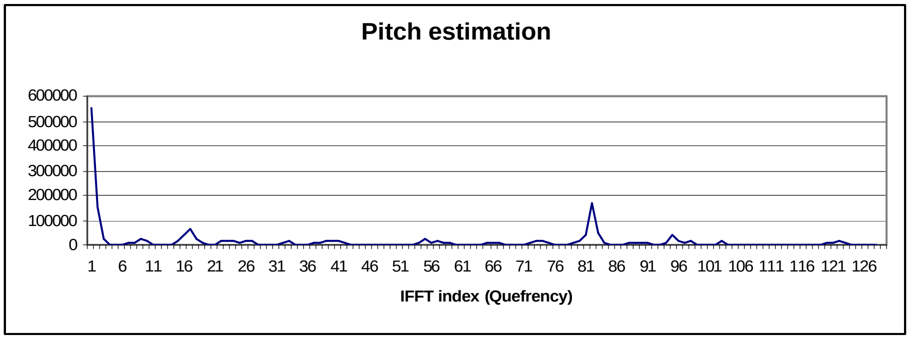

Figure (8)

3. SPEECH SIGNAL RECONSTRUCTION

The main task is done after receiving the sent array X at the receiver in which we start reconstructing the speech signal once again by preparing the fine structure array of that represents the vocal tract in order to multiply this filter to an excitation generated signal to result the Magnitude spectrum (modified) once again. At this step it is not possible to reconstruct the speech signal from only the magnitude spectrum, since the phase information is also needed for the estimation of the original sequence. We will discuss shortly the existing algorithms.

3.1 Excitation signal

Excitation source is independent of the vocal tract transfer function. A sine wave generator was used as an excitation source. For voiced speech, the frequencies of the sine waves generated correspond to the harmonic set, the fundamental being the pitch frequency (1/τ). In case of unvoiced speech, 120Hz was chosen as the pitch frequency. This value was arrived after observation of the effect of various fundamental frequencies between 70Hz & 200Hz on the reconstructed signal. The cumulative excitation signal is given by:

$$E(n) = \sum_{k=0}^{M} \sin(k\omega_0 n + \beta_k) \qquad (1)$$

While, M represents the FFT length which is 256 points in our case, ω0 is the fundamental frequency obtained and was sent previously within the array X, n=M. While β_k is the estimated phase of the kth harmonic.

3.2 The problem of phase estimation

It is not possible to reconstruct the original signal from only the magnitude spectrum, since the phase information is also needed for the estimation of the original sequence. As the phase spectrum is lost in the preprocessing stage, it is desired to estimate the phase function from the given data.

Phase retrieval Issue has been controversial from the middle of 70’s. Most of the work presented for the phase retrieval was depending on Iterative algorithm that is moving within certain constraints between time domain and frequency domain. T.F Quatieri [8] has proposed a set of algorithms to retain phase information from the magnitude iteratively. In 1984 an algorithm was introduced by Griffin and Lim[11] in attempt to reconstruct the phase information from magnitude STFT. The algorithm done by G&L is the widely used algorithm to retain the phase from the magnitude spectrum. Not less than 50 iterations between time domain and frequency domain is needed to retain an average quality signal. In a study Torsten [10] claimed that within 100 iterations a high quality reconstruction was obtained. The short-time phase spectrum for each segment is initially randomized. The algorithm does not converge toward the original speech informal listening tests; however, indicate that more overlap between frames (i.e., less shift) leads to the reconstructed speech sounding more like the original.

In a study, L. D. Alsteris and K. K. Paliwal [12] made the following observations: (i) intelligible signal reconstruction (albeit noisy) is possible from knowledge of only the phase spectrum sign information, (ii) an intelligible signal cannot be reconstructed from knowledge of only the phase spectrum frequency-derivative or only the phase spectrum time-derivative, and (iii) an intelligible signal can be reconstructed from the combined knowledge of both the phase spectrum frequency-derivative and time-derivative.

For a frame of 256 FFT frequency bins and frame shifting, such Iterations could cost a full duplex communication channel a considerable delay due to the extensive calculation at both terminals using the same system.

In this work we have used a non iterative algorithm depending on the formula proposed by Tokuda, Kobayashi and Imai [11]:

$$\beta(\omega) = \tan^{-1}((1-\alpha^2)\sin\omega/((1+\alpha^2)\cos\omega - 2\alpha)) \qquad (2)$$

They concluded that this formula gives a good approximation to auditory frequency scales within appropriate choice for α (for a sampling rate of 8 KHz α is proposed to be 0.31 using mel scale, 0.42 using Bark scale). In this work we took α = 0.31.

3.3 The reconstruction algorithm

The Array sent to the receiver after the analyzing steps represents only 30% of the original PCM speech signal band width over the communication channel using Floating point sample representation at the send time due to the accuracy of the calculations required for the vocal tract filter coefficients, if we use Byte representation the resultant array may reach up to only 8% of the original signal band width over the communication channel. More enhancements in the algorithm are needed if we would like to send the array in byte representation.

- At the receiver, we have used the coefficients to fill the first 20 bins of an IFFT array of complex with length N = 256 points, the rest of the array samples were zero padded. At this step, when we inverse this array into FFT, it will give a smoothed envelop representing the vocal tract filter fig. 6.
- The next step is to prepare the phase array using the pitch information from the array X then uses it within equation (2).
- After the phase spectrum is compromised, the excitation signal is to be generated depending on the phase spectrum information.
- At this point the modified STFT magnitude spectrum could be computed by multiplying the vocal tract filter $S(n)$ with by the excitation signal $E(n)$ giving the Magnitude spectrum $M(n)$.
- The inverse Fourier operation IFFT will give us a good intelligible signal spite of back ground noise that was created during the analysis/synthesis steps due to the modifications in the original magnitude spectrum and the lost of the original phase spectrum.
- The previous step is done after buffering 8000 samples into the playback buffer.

Figure 9 represents the block diagram of the reconstruction steps.

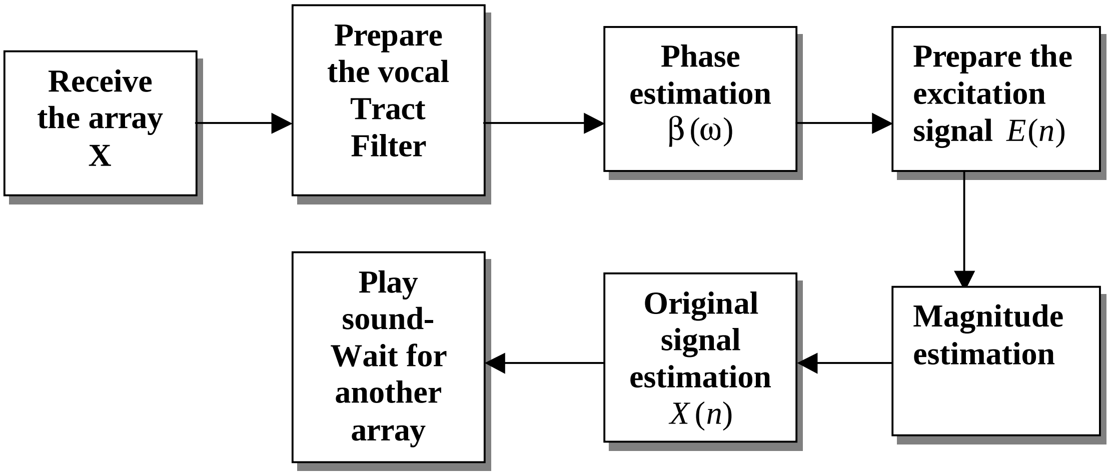
Receive the array X
Prepare the vocal Tract Filter
Phase estimation $\beta(\omega)$
Prepare the excitation signal $E(n)$
Magnitude estimation
Original signal estimation $X(n)$
Play sound- Wait for another array

Figure (9): The reconstruction steps

4. Results

In this work we used a lab network with PC's of high configuration: Pentium M 740 (2 MB L2 cache, 1, 73 GHz) – 1 GB DDRII (Dual channel) – High quality multimedia cards with a noise cancellation Microphone.

The results are mainly done using the subjective test MOS done by 8 persons testing at both ends (Transmitter-Receive) achieving a mean score of about 3.5 over 5.

The SNR (using Equation 3) gave a result of 23 to 28 dB in different frames.

$$SNR = 20\log10(|x_0|^2 / |x_0 - x_r|^2) \tag{3}$$

Where x_0 is the original signal, x_r is the reconstructed.

Figures 10 shows one second of original signal (left) and reconstructed one (right). Figures 11, 12 show the magnitude spectrum of an original Magnitude spectrum and the reconstructed one.

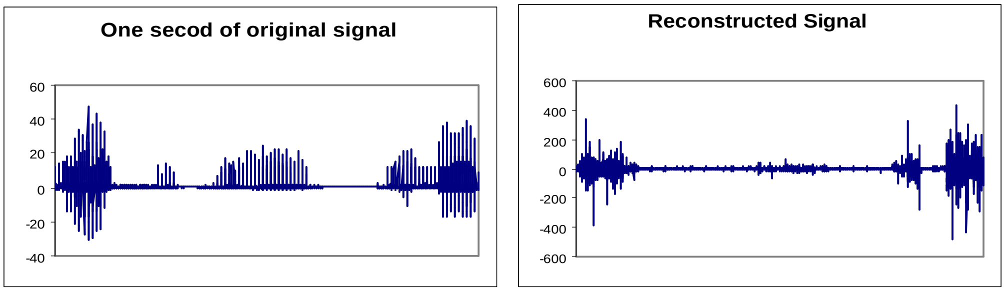

Figure (10): Original signal and the reconstructed one

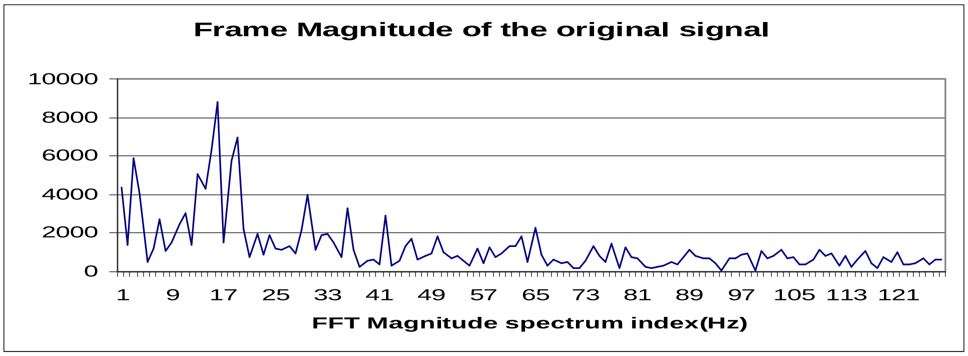

Figure (11):Magnitude spectrum of the original signal

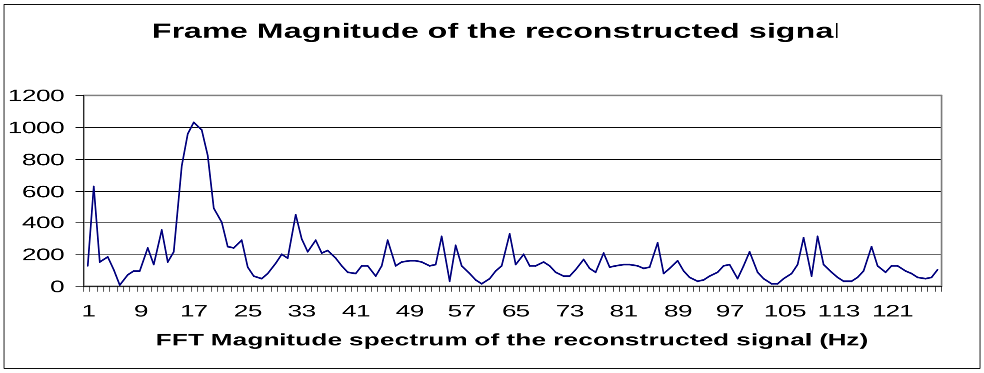

Figure (12):Magnitude spectrum of the reconstructed signal

6. Conclusions:

A short background for the previous work in the area of signal reconstruction is introduced in this paper; most of this work concentrates on Iterative phase reconstruction, which costs a considerable delay due to processing time.

An algorithm for speech signal analysis / synthesis has been explored to come up with a low bit-rate codec. Speech signal is split into its basic components – Cepstrum coefficients, Pitch, and pitch amplitude. A non iterative algorithm [9] was used to compromise the synthetic phase spectrum signal at the receiver that was originally lost at the transmitter. These components are the minimal parameters that could represent the original signal, although there is definitely a quality loss when the synthetic phases are used in place of the original phase values, particularly during sharp voice transitions, the synthetic speech signal derived from the algorithm presented in this work is of acceptable quality. An adaptive noise cancellation filter should be added to the synthesis steps in order to terminate the back ground noise generated due to the analysis / synthesis steps.

References:

[1] R. Tucker, T. Robinson and J. Christie Tucker, "Compression of acoustic features – are perceptual quality and recognition performance incompatible goals?", Proc. Euro speech'99, pp. 2155-2158, 1999.

[2] H. Kim and R. Cox, "A bit stream-based front-end for wireless speech recognition on IS-136 communication systems", IEEE Trans. Speech and Audio Processing, volume 9, number 5, pp. 558-568, 2001.

[3] S.Young," A review of large-vocabulary continuous speech recognition" IEEE signal processing,pp. 45-57, sept.1996.

[4] Mikael Nilsson and Marcus Ejnarsson, "Speech Recognition using Hidden Markov Model performance, evaluation in noisy environment", Degree of Master study, Blekinge Institute of Technology March, 2002.

[5] Dan Jurafsky," Feature Extraction and Acoustic modeling", LSA 352 summer 2007.

[6] ETSI standard document-ETSI ES 201 108 v1.1.1, Speech Processing, Transmission and Quality aspects (STQ), Distributed speech recognition, Front-end feature extraction algorithm, Compression Algorithm, 2000.

[7] Michael Noll,"Cepstrum Pitch Determination", JASA, pp. 293-309, August, 1966.

[8] T.F. Quatieri, "Phase estimation with application to speech analysis-synthesis", Technical report 491, Massachusetts Institute of technology, November 1979.

[9] Griffin and J. Lim "Signal estimation from short-time Fourier transform" IEEE Trans., 1984.

[10] K. Torsen,"Re-synthesis of speech from ASR features", Master degree project, KTH, Stockholm, Sweden 2006.

[11] K. Tokuda, T. Kobayashi and S. Imai "Recursive Calculation of Mel-Cepstrum from LP Coefficients" Tokyo Istitute of Technology, Yokohama, 227 Japan, April 1994.

[12] L. D. Alsteris and K. K. Paliwal,"Iterative reconstruction of speech from short-time Fourier transforms phase and magnitude spectra", School of Microelectronic Engineering, Griffith University, Brisbane, Qld 4111, Australia, April 2006.