Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (1): 49-59, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.49.59

Development of Low Bit Rate Speech Encoder based on Vector Quantization and Compressive Sensing

[1]L.A.S. Kassim, [1]T.S. Gunawan, [1]O.O. Khalifa, [2]M. Kartiwi, [1]A. Sulong,
[1]K. Abdullah and [1]N.F. Hasbullah
[1]Department of Electrical and Computer Engineering,
[2]Department of Information System,
International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Malaysia

Abstract: Speech coding is a representation of a digitized speech signal using as few bits as possible, while maintaining reasonable level of speech quality. Due to growing need for bandwidth conservation in wireless communication, the research in speech coding has increased. Recently, Compressive Sensing (CS) is gaining a great interest because of its ability to recover original signals by taking only few measurements. CS is a new approach that goes against the common data acquisition methods. In this research, a new system of speech encoding system is developed using compressive sensing. Since CS performs well in sparse signals, different sparsifying transforms are analyzed and compared using Gini coefficient. The quality of the speech coder is evaluated using Perceptual Evaluation of Speech Quality (PESQ), Signal-to-Noise Ratio (SNR) and subjective listening tests. Results show that the speech coders have achieved a PESQ score of 3.16 at 4 kbps which is a good quality as confirmed by listening tests. Furthermore, the coder is also compared with Code Excited Linear Prediction (CELP) coder.

Key words: Low bit rate coding, compressive sensing, Gini index, subjective evaluation, objective evaluation

INTRODUCTION

Due to the growing need for bandwidth conservation and enhanced quality in wireless cellular and satellite communication, the research of low bit rate speech coder with acceptable quality is becoming increasingly important. Applications like digital cellular and satellite telephony, video conferencing and internet voice communications, all have an increasing demand for efficient use of bandwidth without compromising the quality (Spanias, 1994). The process of obtaining compressed digital representations of voice signals for purpose of efficient transmission or storage is called speech coding. The analog signal is changed into a sequence of bits by the sampling process. The sequence is processed by an encoder to construct the coded representations. The coded representation is either sent to the receiver or stored. The receiver reconstructs an approximation of the original signal (Goldberg and Riek, 2000).

In speech coding, the primary objective is to realize a low bit rate speech signal at a high perceived quality. Bit rate have a significant importance in applications such as mobile communications. In digital speech communications, the bandwidth of speech is limited to 4 kHz and sampled at 8 kHz. Typically, speech samples are quantized 8-16 bits. The quantization can be either uniform or non-uniform (Chu, 2003). The simplest coding technique is the Pulse Code Modulation (PCM) which is coded at 64 kbit sec^{-1}. Nevertheless, the PCM signal is considered as uncompressed signal.

Advanced speech coding methods make use of techniques that eliminate redundancy and irrelevant information; this enabled to achieve a high quality at lower bit rates. Linear Prediction Coding (LPC) is used to model the speech signal which can realize a coding rate of 4 kbit sec^{-1}. The most common scheme used today is Code Excited Linear Prediction (CELP) which is based on analysis-by-synthesis coding (Vasuki and Vanathi, 2006).

In general, speech coding is a process to represent a digitized speech signal using as few bits as possible, while maintaining a reasonable level of speech quality (Chu, 2003). Sometimes it is also called speech compression. Figure 1 shows the block diagram of a typical speech coding system. The speech signal from a given source is digitized by a standard connection of anti-aliasing filter, a sampler which converts into a discrete time signal and an analog to

Corresponding Author: T.S. Gunawan, Department of Electrical and Computer Engineering,
International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Malaysia

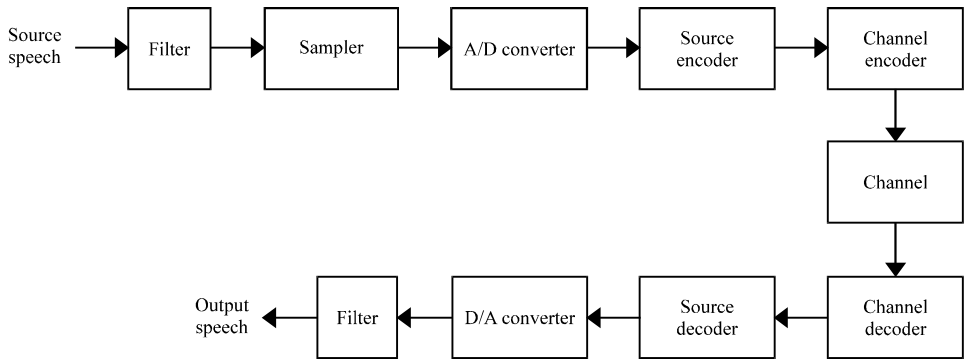

Fig. 1: Typical speech coding system (Chu, 2003)

digital converter which quantizes the speech signal. The output is this stage is called a digital signal.

The larger number of bits for accurate reproduction of speech samples makes systems complex and expensive. Increasing demand of wireless applications which requires an efficient use of bandwidth while maintain the quality has led to the research of developing low bit rate coders to gain a great interest. The reproduced speech quality is limited by the accuracy of the hypothetical models and parametric estimations. For the past decade, research focused on producing low bit rate coders but limitations still exists, mainly the quality of the synthesized speech. Therefore, the objective of this research was to develop and improved low bit rate speech coder based on vector quantization and compressive sensing and evaluate the system using objective and subjective evaluations.

SPEECH CODING ALGORITHMS

There has been a substantial progress towards the application of speech coders to communications as well as computer related voice applications. Central to this progress has been the development of new speech coders capable of producing high quality speech at low data rates. Because of the high demand for speech communication, speech coding technology has received high levels of interest from various researchers and businesses. The advances in microelectronics design and low-cost programmable processors have enabled conversion of research into a product development. This has encouraged researchers to find out more alternative schemes for speech coding, with the objective of reducing deficiencies and limitations (Chu, 2003).

Speech coding is carried out using a number of steps or processes called an algorithm. Algorithm is a well-defined computational procedure that takes a set of input and produces a set of output. These procedures can be translated in a code to be executed by a processor. The

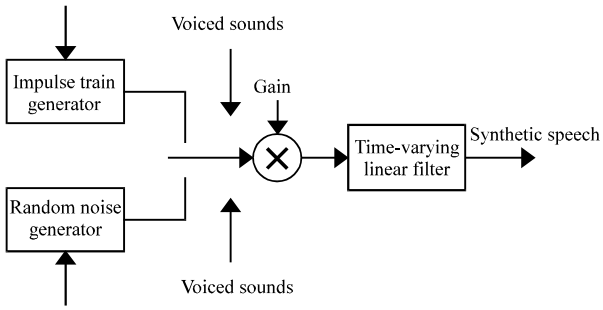

Fig. 2: LPC speech model (Spanias, 1994)

development of low bit rate coders emphasizes on parametric coders (Spanias, 1994). The speech signal is generated from a model controlled by some parameters. During the encoding, parameters of the model are estimated from the input speech signal and transmitted. By this type of coding, no original waveform is preserved (Chu, 2003). The basics of most parametric coding techniques originates from Linear Predictive Coding (LPC), Mixed Excitation Linear Predictive Coding (MELP) and Code Excited Linear Predictive Coding (CELP).

Linear prediction coding (LPC): One of the earliest standardized coders is based on Linear Prediction Coding (LPC) which works at low bit rate. The FS1015 LPC is an example of intelligent coder at 2.4 kbps (Spanias, 1994; Vasuki and Vanathi, 2006). In LPC, filter coefficients are predicted that is used for representing speech for low bit rate transmission or storage. In other words, speech samples can be approximated from linear combination of past samples (Spanias, 1994; Goldberg and Riek, 2000). By minimizing the squared error between past sample and current sample (predicted), coefficients can be determined. Figure 2 shows the LPC model of speech production.

This is the source-filter model of the speech production, inspired by observations of the basic

properties of speech signals. The combined spectral contributions of the glottal flow, the vocal tract and the radiation of the lips are represented by the synthesis filter. The driving output of the filter or excitation signal is modelled as either an impulse train (voiced speech) or random noise (unvoiced speech). The parameter estimation is repeated for each frame; so instead of transmitting the PCM samples, the parameters of the model are sent. By allocating bits for each parameter, an effective compression can be achieved. The encoder estimates the parameters of the model and the decoder takes the estimated parameters and uses the speech production model to synthesize speech (McCree and Barnwell III, 1995).

Mixed excitation linear predictive (MELP) coding: Traditional pitch excited LPC is a fully parametric technique to encode the important information of the speech so that a lower bit rate representation is achieved. However, with the presence of acoustic background noise it produces unneeded artefacts such as buzzes and tonal noises. So, it became unacceptable for many applications. Mixed excitation LPC preserves the low bit rate of fully parametric model but it adds more parameters so that the synthesizer can produce similar characteristics of natural human speech (McCree and Barnwell III, 1995). The mixed Excitation algorithm is based on LPC vocoder but contains four additional features (Chu, 2003; McCree and Barnwell III, 1995):

- Mixed pulse and noise excitation
- Periodic or aperiodic pulses
- Adaptive spectral enhancement
- Pulse dispersion filter

These features allow the mixed excitation LPC to mimic more of the characteristics of natural human speech.

Mixed excitation: Pulse train and noise sequence are combined together to give a full excitation. In each frame the frequency shaping filter coefficients are generated by a weighted sum of fixed band pass filters. By combining these two excitations, the buzzy quality is removed from LPC speech input. The presence of isolated tones in the synthesized speech can be reduced by eliminating the periodicity in the voiced speech by varying each pitch period length (aperiodic pulses). Strong voicing is defined by periodicity and easily detected from the strength of normalized correlation coefficient of the pitch search algorithm. Jittery voicing corresponds to erratic glottal pulses and it can be detected by either marginal correlation or peakiness in input speech (Goldberg and Riek, 2000).

Adaptive spectral enhancement: It helps the band-pass filtered synthesis of synthetic speech to match natural speech waveforms in formant regions. The waveforms of synthetic speech reach a lower valley between peaks than natural waveforms do. This could be inability of the poles in the LPC synthesis to reproduce the features of formant resonances in natural human speech. The adaptive spectral enhancement provides simple solution of matching formant waveforms by systematically varying the synthesis pole bandwidths within each period (Goldberg and Riek, 2000).

Pulse dispersion: The pulse dispersion filter improves the match of band-pass filtered synthetic speech and natural waveforms in frequency bands which do not contain a formant resonance. It is a fixed FIR filter based on spectrally flattened synthetic glottal pulse which introduces time-domain spread to the synthetic speech. This dispersion filter decreases the peakiness in frequencies away from the formants and results in a more natural sounding LPC speech output.

The addition of these four features to the LPC vocoder improved the quality at the expense of few extra bits per frame. The overall voicing decision is based on the strength of pitch periodicity. Strong pitch correlation results as strongly voiced while high peakiness in the residual signal is classified as jittery voiced. Unvoiced frame is declared if none of these are met (Goldberg and Riek, 2000; Ming, 2004).

Code excited linear prediction (CELP) coding: Code Excited Linear Prediction (CELP) is a speech coding algorithm originally proposed in Schroeder and Atal (1985). It provided better quality than existing low bit rate algorithms. It is used in MPEG-4 audio speech coding. In CELP, each waveform is synthesized by passing it through a two part cascade synthesis filter. The first is the pitch synthesis filter and the second is formant synthesis filter (Schroeder and Atal, 1985; Kumar, 2000; Wang, 2000).

The excitation waveform is chosen from a dictionary of waveforms. Each waveform is passed through a synthesis filter to determine which waveform is the best matches the input speech. The optimality criterion is based on the same frequency weighed mean square error criterion. The index of best waveform is transmitted to the decoder, both formant and pitch filters are updated periodically. These parameters are sent to the decoder to form the appropriate synthesis filters (Wang, 2000).

COMPRESSIVE SENSING

Compressive sensing, also known as compressive sampling or CS, is a method that goes against the common

data acquisition. It provides a way of recovering of the original signal by taking only far fewer samples or measurements than the traditional methods use (Candes and Wakin, 2008). Compression is obtained by storing only the largest basis coefficients. While in the recovery process, the non-stored coefficients are set to zero.

Compressive sensing uses fewer samples to approximate the original signal lower the Nyquist rate, where the sampling rate must be at least twice the maximum frequency of the signal. To make this possible, CS makes use of two principles: sparsity which is related to the signals of interest and incoherence which is related to sensing model (Candes and Wakin, 2008), CS concerns mainly the pairs with low-coherence. In other words, in order for CS work, number of measurement required is proportional to sparse factor. Although we acquire signals following the Shannon theory, most of the unnecessary data is thrown away by subsequent processes, like compression or coding (Donoho, 2006).

Sparsity: The most important theory of compressive sensing is that signal recovery from random projection is possible provided that the signal is a sparse in vector space (Gunawan *et al*., 2011). The sparse property is a signal redundancy measure which the CS make use of it at the acquisition stage. When the signal is represented in a convenient transform, such like Wavelet Transform (WT), Discrete Cosine Transform (DCT) or Fast Fourier Transform (FFT), most of the coefficients are very small and can be ignored and relatively few large coefficients contain most of the information. By zeroing small coefficients, construction of the original data are possible with the difference is hardly noticeable as stated (Gunawan *et al*., 2011). Figure 3 shows an example of a speech signal and its Fourier transform. As we can see there are two nonzero components.

Let x be the signal and let $\psi = \{\psi_1, \psi_1, \ldots, \psi_N\}$ be the basis vectors (such as wavelet basis). The signal is said to be "sparse" if:

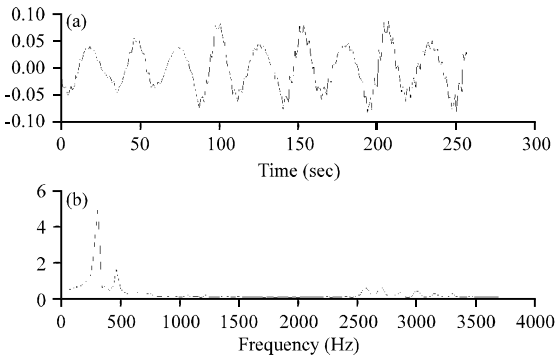

Fig. 3(a-b): (a) Speech signal and (b) its Fourier transform

$$x = \sum_{i=1}^{T} s_{n_i} \psi_{n_i}, \quad \{n_1, n_2, \ldots, n_T\} \subset \{1, \ldots, N\} \tag{1}$$

where, s_{n_i} are the scalar coefficients and T<<N. ψ is our knowledge about x that provides the key to compressive sensing. So, $x = \psi$, where s is the sparse vector with only T non-zero elements (Candes and Wakin, 2008; Gunawan *et al*., 2011; Christensen *et al*., 2009). The sampling (sensing) measurements are:

$$y_m = \sum \phi_m x(i), \; 1 \le m \le M \le N \tag{2}$$

or $y = \Phi.x$, where Φ is M×N measurement matrix. The Φ made up of orthonormal random basis vectors ϕ_m. Coherence is the measure of the correlation between any two elements of Φ (sensing basis) and ψ (representation basis). Therefore, if the incoherency condition of Φ and ψ are met, then x can be reconstructed from y with high probability if M>Tlog(N) (Candes and Wakin, 2008; Sreenivas and Kleijn, 2009). The basic objective in CS is to find out the minimal number of linear non-adaptive measurements that allows the reconstruction of the signal (Rauhut *et al*., 2008). The reconstruction method proposed is through convex optimization:

$$\hat{s} = \arg\min \|s\|_1, \text{ subject to } y = \Phi.\psi.s \text{ and } \hat{x} = \psi.s \tag{3}$$

where, the $\|.\|_1$ is the l_1 norm. The computational is the sensor is quiet low while at the receiver is complex iterative.

There are many solutions proposed for sparse approximation, such as Matching Pursuit (MP), Least Absolute Shrinkage and Selection Operator (LASSO), Basis Pursuit (BP), Gradient Pursuit (GP), where the performance depends on the number of measurements, signal sparsity and the reconstruction algorithm (Rauhut *et al*., 2008; Figueiredo *et al*., 2007).

DESIGN AND IMPLEMENTATION

The proposed system is based on compressive sensing algorithm. The input speech is divided into a frame size of 32 msec with a Gammatone filter. Each frame of the speech is applied to a Discrete Cosine Transform (DCT) to make signal sparse. Then Gradient Projection for Sparse Reconstruction (GPSR) is applied. The Gammatone filter output is quantized to reduce the bit rate using a codebook and the codebook indices are sent to receiver side for decoding. Figure 4 shows the proposed speech encoder.

Codebook indices are received at the decoder, the signal is retrieved from the look up table or codebook (Vasuki and Vanathi, 2006). The reconstructed signal is

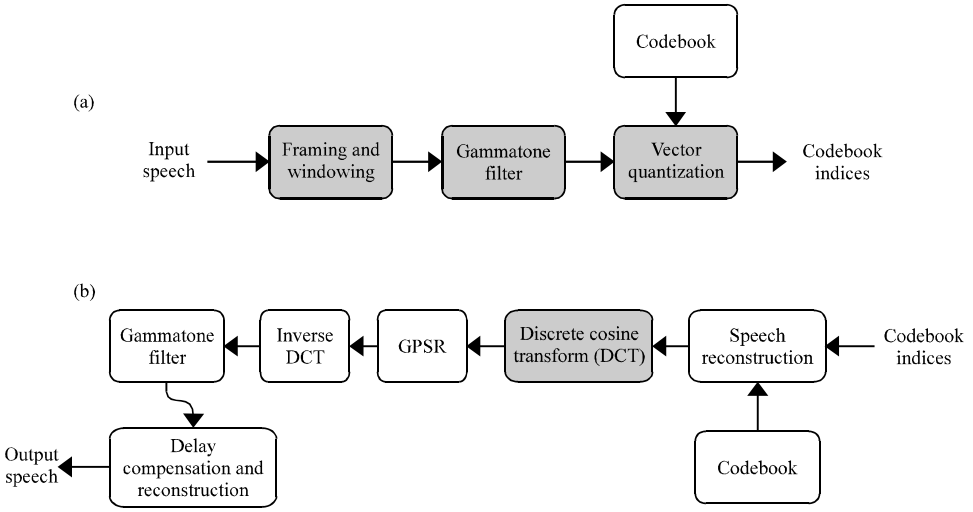

Fig. 4(a-b): Proposed speech coding algorithm (a) Encoder and (b) Decoder

further applied compressive sensing using GPSR algorithm followed by Gammatone filter and delay compensation for reconstruction of the original input speech. The compressive sensing is preceded by Discrete cosine transform to introduce sparsity to the speech signal.

Gammatone filter bank analysis: Gammatone filters are a popular linear approximation to the filtering performed by the ear. The speech signal $x(n)$ is decomposed into M sub-bands using Gammatone filter banks, i.e., $x_m = h_m(n)*x(n)$. Gammatone filter banks are implemented using FIR filters to achieve linear phase filters with identical delay in each critical band. The synthesis filter $g_m(n)$ is the time reverse of its analysis filters $h_m(n)$. Figure 5 shows front end processing for speech coding applications (Ambikairajah *et al*., 2001). The analysis filter is followed by a half-wave rectifier to simulate behaviour of the inner hair cell (Kubin and Kleijn, 1999). The input to the PESQ score is the reference speech signal $x(n)$ and the reconstructed speech signal $\hat{x}(n)$. The PESQ software then rates the speech quality between 1.0 (bad) to 4.5 (no distortion).

Discrete cosine transform (DCT): DCT is applied to each sub-band in order to introduce sparsity to the signal. Several sparsity algorithms are evaluated using Gini index. The Gini coefficient is a measure of the inequality of a distribution, a value of 0 expressing total equality and a value of 1 maximal inequality. Those algorithms are Discrete Fourier Transform (DFT) and Wavelet Transform (WT). Table 1 shows the Gini index for the three algorithms. The result shows that DCT achieves more sparsity than the other two transforms. The DCT signal has sparse peaks compared to the original signal. Compressive sensing works best with sparse signals.

Table 1: Gini index for various sparsifying transform

Transform	Gini index
Discrete cosine transform (DCT)	0.5245
Fast Fourier transform (FFT)	0.4734
Wavelet	0.0488

Gradient projection for sparse reconstruction (GPSR): Many algorithms have been proposed for solving the convex unconstrained optimization problem (Figueiredo *et al*., 2007):

$$\min \tfrac{1}{2}\|y - Ax\|_2^2 + \tau\|x\|_1 \quad (4)$$

where, $x \in R^n$, $y \in R^k$, A is k×n matrix, τ is a nonnegative parameter, $\|v\|_2$ refers the Euclidean norm of v and $\|v\|_1 = \Sigma_i|v_i|$ is the l_1 norm of v.

The Eq. 4 is related to the following convex constrained optimization problems:

$$\min \|x\|_1 \text{ subject to } \|y - Ax\|_2^2 \le \varepsilon \quad (5)$$

and

$$\min \|y - Ax\|_2^2 \text{ subject to } \|x\|_1 \le t \quad (6)$$

where, ε and t are nonnegative real parameters.

Many approaches have been used to minimize the objective function. These approaches include the standard one with quadratic (squared l_2) error term

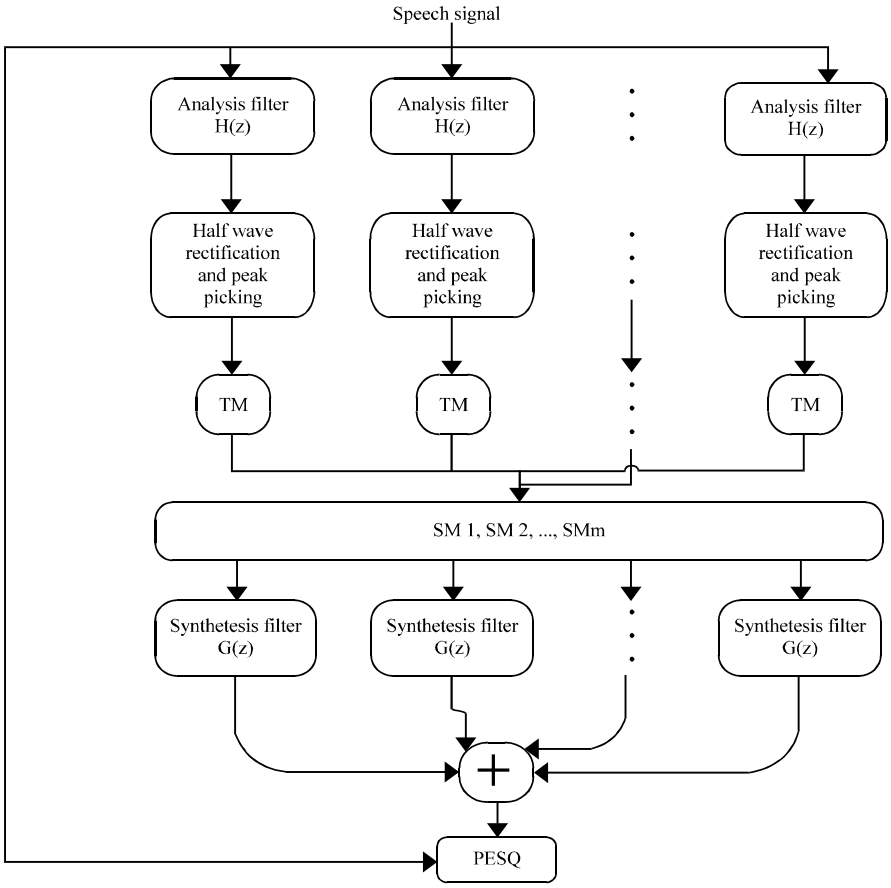

Fig. 5: Gammatone analysis and synthesis banks system (Ambikairajah *et al*., 2001)

combined with sparseness-inducing (l_1) regularization term. Other approaches are basis pursuit, the Least Absolute Shrinkage and Selection Operator (LASSO), wavelet based deconvolution. The GPSR method is faster and scales more favourably than other approaches like BB (Barzilai and Borwein) as shown in Table 2.

So in this project we use GPSR-Basic since it is the faster way to solve the convex problem in compressive sensing. If the signal is approximately sparse, then accurate reconstruction can be achieved from random projections which suggests potentially powerful alternative to conventional Shannon-Nyquist sampling. GPSR is a gradient projection algorithm applied to a quadratic programming formulation of in which the search path from each iterate is found by projecting the negative-gradient onto the feasible set.

Vector quantization: Vector Quantization (VQ) is a data compression technique, reconstructing a signal with small distortion as possible. The quality of the reconstruction depends on the amount of data that is discarded. The samples of the source output is divided into a k-dimensional vector which is the input to the vector quantizer. The vector quantizer consists of a codebook that contains a set of vectors called code vectors. The codebook is designed using LBG algorithm. The code vectors are the quantized values of the input samples. A copy of the codebook is maintained at the receiver. Figure 6 shows the block diagram of the vector quantizer (Vasuki and Vanathi, 2006).

An input vector is compared to the codebook vectors; the index of the codebook vector which achieves minimum distortion error is selected. The address of the index in binary form is transmitted to the receiver. At the

Table 2: CPU times of several GPSR types

Algorithm	CPU time (sec)
GPSR-BB monotone	0.59
GPSR-BB non-monotone	0.51
GPSR-basic	0.69
GPSR-BB monotone+debias	0.89
GPSR-BB non-monotone +debias	0.82
GPSR-basic debias	0.98
l1_ls	6.56
IST	2.76

GPSR: Gradient projection for sparse reconstruction, BB: Barzilai and Borwein, IST: Iterative Shrinkage and thresholding (Figueiredo *et al*., 2007)

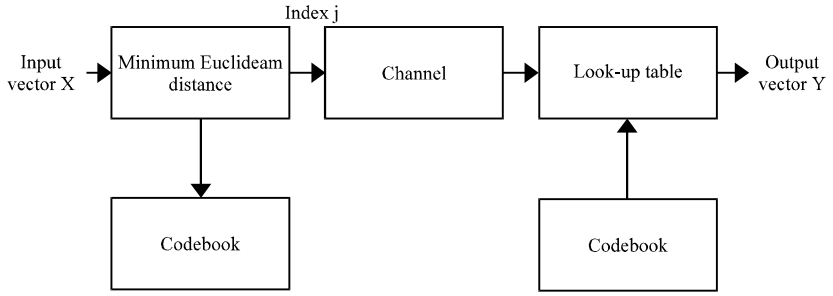

Fig. 6: Vector quantizer system

receiver, the codebook vector corresponds the index transmitted is selected which is approximation of the input vector. The number of bits needed to represent a code vector is kR, where k is the dimension of the vector and R is the rate. To inform the decoder which code vector is selected, $\log_2 k$ bits is used. The compression ratio increases with the dimension k of the vector but size of the codebook grows exponentially (Vasuki and Vanathi, 2006). The encoder complexity depends on the size of the codebook, dimension of the code vector.

Encoder and decoder: The input speech is divided into frames of 32 msec using Gammatone filter. Speech signal x(n) is decomposed into M sub-bands, where M is the number of filters. Each sub-band is coded separately. Each sub-band signal is processed separately. The Euclidean distance measure is used to search the codebook. The index which yields the minimum of the Euclidean distance is selected and transmitted to the receiver.

The codeword with the index received is selected. Discrete cosine transform is applied to the sub-band to introduce sparsity to the signal. The reason that we need sparsity is that compressive sensing algorithm works better with sparse signals. After DCT, Gradient Projection for Sparse Signal Reconstruction (GPSR) is applied to the signal. It provides a way of recovering the signal by taking only fewer samples or measurements. Delay compensation is followed, since the amount of filter delay accumulated by each sub-band is different, without compensating for this delay, the reconstruction of sub-band signal will lead to an incoherent signal.

ANALYSIS AND RESULTS

Here, objective and subjective evaluations of the speech coder are presented. A set of speech files are chosen to train the codebook and another set is used to test it. The speech coder is designed and simulated using MATLAB software.

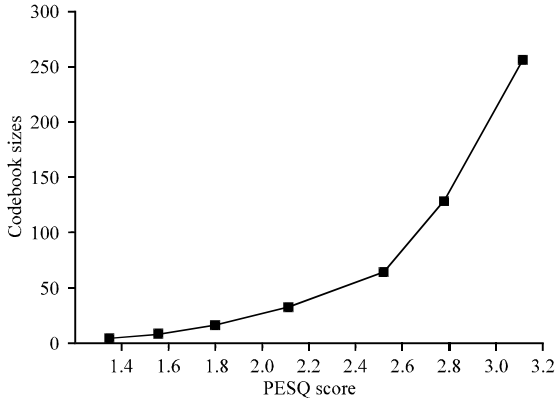

Fig. 7: Various codebook sizes (samples) and their perceptual quality

Training datasets: To evaluate the performance of the algorithms NOIZEUS speech datasets were used. Thirty sentences form IEEE sentence database were recorded in a sound-proof booth (Hu and Loizou, 2007). Three male and three female speakers produced the sentences and originally sampled at 25 kHz then downsampled to 8 kHz. The NOIZEUS datasets were used since it contains all phonemes in the American English language.

Various codebook sizes: A codebook of different sizes is trained for 20 speech files and tested on 10 speech files. The speech files are from the dataset NOIZEUS. The sizes of the codebook are 256, 128, 64, 32 and 16.10 of the 20 files are male speech voices, while the other 10 are female speech voices. The perceptual evaluation of speech quality is measured for each codebook. In this simulation, we take file No. 21 in the dataset and test it in different codebooks. Figure 7 shows the graph of the codebook size and their perceptual quality (PESQ) measurements. We can say that there is always a trade-off between the quality and the codebook sizes which determines the bit rate. The quality increases as we increase the codebook size which leads to increase of bit rate. From Fig. 7, it can be concluded that if we reduce the codebook size the quality will decrease also.

Objective evaluation for various files: The codebook is tested with 10 files (5 male voices, 5 female voices) from the datasets which are the non-trained files. The average PESQ score is 3.163 which is a good quality as confirmed by the subjective listening tests. Table 3 shows PESQ score for the file 1 to 10 and Fig. 8a-j compares the original speech and their reconstructed version with their PESQ. Figure 9 shows the Signal-to-Noise Ratio (SNR) for the speech files. Both SNR and Segmental SNR are measured.

Table 3: Perceptual evaluation of speech quality (PESQ) score for various speech files

Files	PESQ
1	3.1140
2	3.3070
3	3.3060
4	3.2080
5	3.3750
6	3.0510
7	3.1350
8	2.8670
9	3.1330
10	3.1430
Average	3.1639

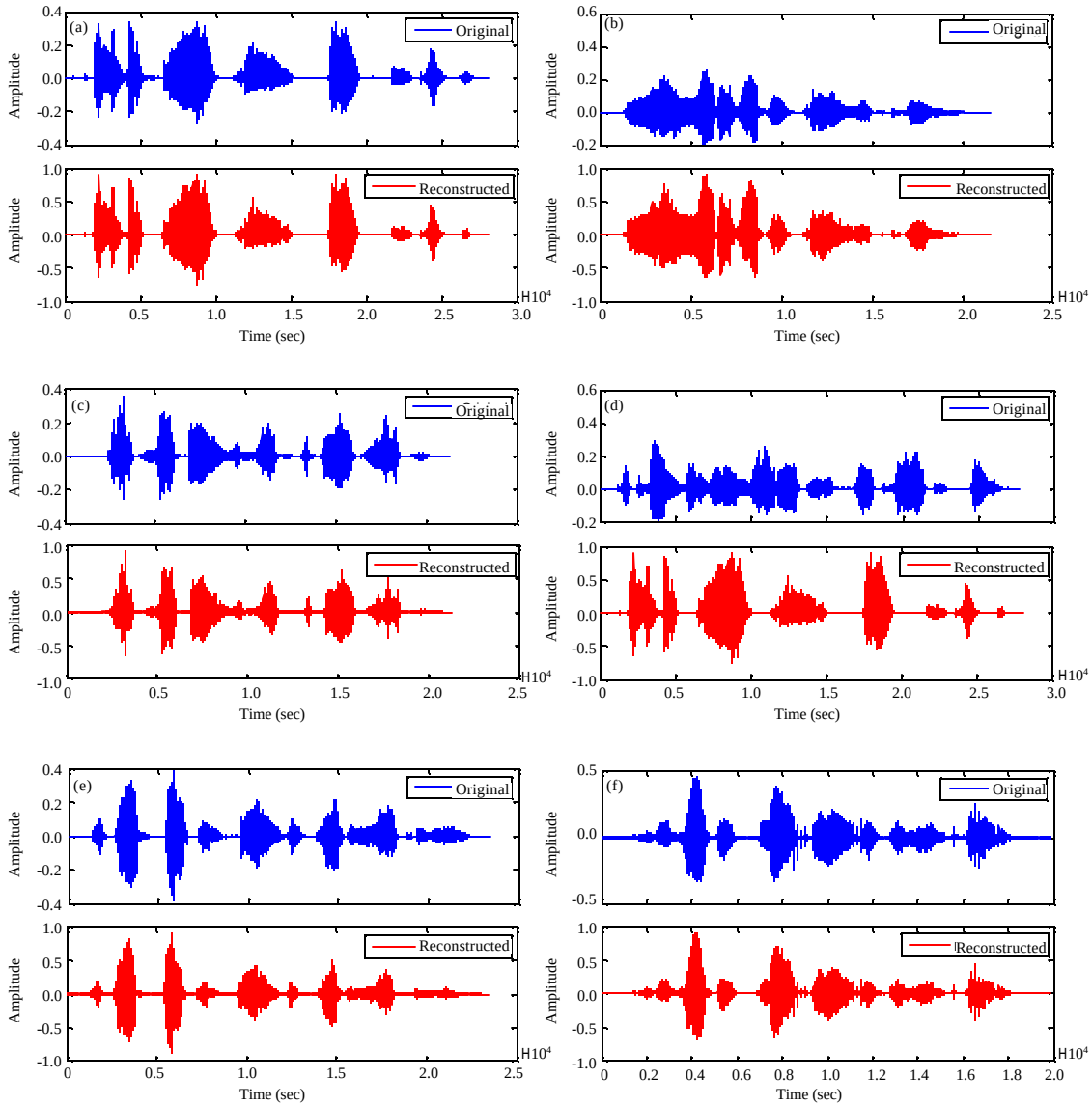

Fig. 8: Continue

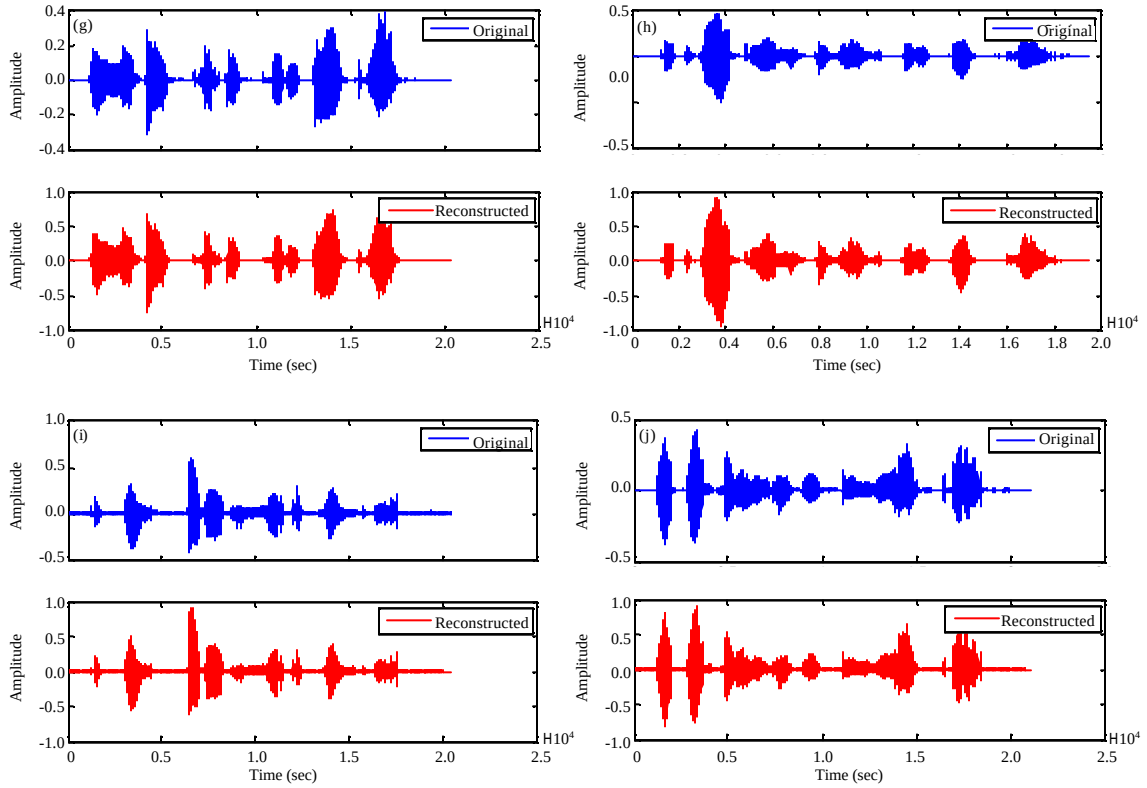

Fig. 8(a-j): PESQ score for speech files (original and reconstructed), (a) PESQ score (3.114) for speech file 1, (b) PESQ score (3.307) for speech file 2, (c) PESQ score (3.306) for speech file 3, (d) PESQ score (3.208) for speech file 4, (e) PESQ score (3.375) for speech file 5, (f) PESQ score (3.051) for speech file 6, (g) PESQ score (3.135) for speech file 7, (h) PESQ score (2.867) for speech file 8, (i) PESQ score (3.133) for speech file 9 and (j) PESQ score (3.143) for speech file 10

Figure 8a-j compares the original speech signal and the reconstructed signal for various speech files. The reconstructed signal is on top of the original signal. Therefore, it the reconstructed signal match the original, you will not see any blue colours in the figure. The reconstructed signal quality is quiet good so that it matches the original signal with very few errors.

In Fig. 9, the segmental SNR examines in detail the comparison of the original signal and the test signal in short segments (frames), while the Global SNR calculates the overall SNR. Therefore, the global SNR is higher than the segmental SNR.

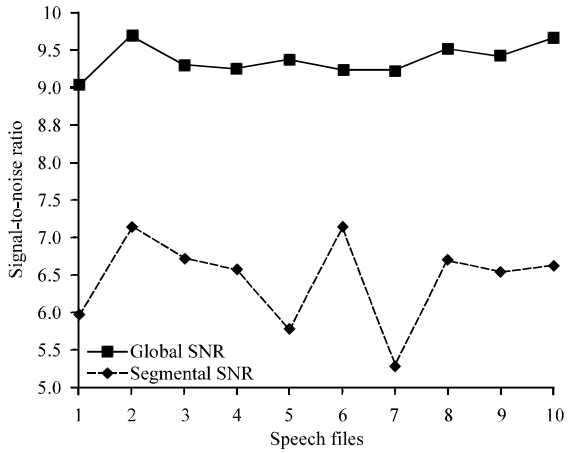

Fig. 9: Comparison of global and segmental signal to noise ratio (SNR)

Subjective evaluations: Around 25 listeners evaluated 10 speech files (5 male and 5 female). Each listener evaluates all the speech files based on his perception of the quality. The average of each file from the evaluations of all listeners will be taken. The Mean Opinion Score (MOS) is found for each file. The results yield average score of 3.712 which shows a very good quality. The subjective tests of the quality of the speech confirm the objective

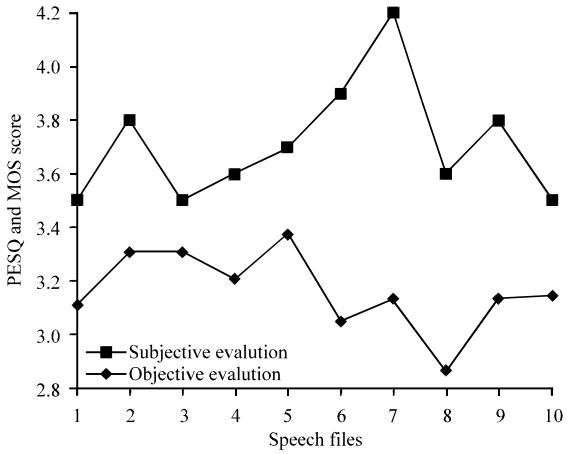

Fig. 10: Subjective and objective evaluations of speech files

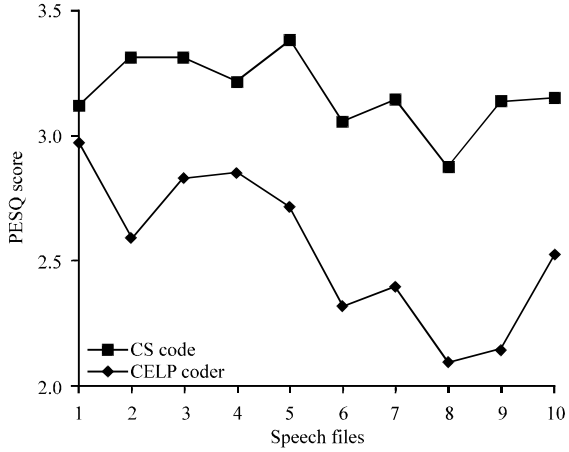

Fig. 11: Comparison of compressive sensing (CS) and CELP based coder

test with a higher average score of 3.712 comparing to objective test of 3.161. Figure 10 compares the objective and subjective tests.

Comparison with CELP coder: The algorithm is compared with other Code Excited Linear Prediction coding (CELP) encoders. Code Excited Linear Prediction (CELP) is a speech coding algorithm originally proposed by Schroeder and Atal (1985). It provided better quality than existing low bit rate algorithms. It is used in MPEG-4 audio speech coding. In CELP, each waveform is synthesized by passing it through a two part cascade synthesis filter. The first is the pitch synthesis filter and the second is formant synthesis filter (Vasuki and Vanathi, 2006; Kumar, 2000; Wang, 2000).

The average PESQ score for the CELP coder is 2.54 which is quiet lower than the compressive sensing-based coder. Both coders are tested on 10 speech files and the result shows that the compressive sensing based speech encoders out performs the CELP coder in terms of quality achieved. Figure 11 demonstrates the PESQ score for the both speech coders.

For the compression ratio, the CS coder uses a codebook of 256 and frame size of 32 msec, with 16 Gammatone filters (subbands); each sub-band is coded separately. Therefore the bit rate achieved is 4 kbps; while the CELP coder achieves a bit rate of 4.8 kbps. Bit rates that are below than 5 kbps are considered low bit rate.

CONCLUSIONS

Low bit rate speech coder using compressive sensing have been proposed in this study. The coder achieved a low bit rate of 4 kbps with an acceptable quality for everyday communications. The speech coder is designed and simulated using MATLAB software. Objective and subjective evaluations are performed to test the quality of the coder. The results shows a PESQ score of 3.16 and SNR ratio of 9.35 (SNR seg = 6.44) which demonstrated a good quality for a speech coder. The subjective listening tests confirm the objective results with an average MOS of 3.712. On the other hand, the CS coder achieved a higher PESQ score than CELP coder.

Compressive sensing technique finds sparse solutions which is kind of optimization algorithms, so it takes time to find the optimum solutions. Several algorithms have been proposed to solve the convex unconstrained optimization problem and so far GPSR achieves the highest. A faster algorithm for solving the convex optimization problem can help the time consumption problem. Another recommendation is that the proposed algorithm could be implemented in parallel processing or in embedded hardware to speed up the decoding time.

REFERENCES

Ambikairajah, E., J. Epps and L. Lin, 2001. Wideband speech and audio coding using Gammatone filter banks. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, May 7-11, 2001, Salt Lake City, UT., USA., pp: 773-776.

Candes, E.J. and M.B. Wakin, 2008. An introduction to compressive sampling. IEEE Signal Process. Mag., 25: 21-30.

Christensen, M.G., J. Stergaard and S.H. Jensen, 2009. On compressed sensing and its application to speech and audio signals. Proceedings of the Conference Record of the 43rd Asilomar Conference on Signals, Systems and Computers, November 1-4, 2009, Pacific Grove, CA., USA., pp: 356-360.

Chu, W.C., 2003. Speech Coding Algorithms: Foundations and Evolution of Standardized Coders. Wiley-Interscience Publications, New York, USA., ISBN-13: 9780471373124, Pages: 558.

Donoho, D.L., 2006. Compressed sensing. IEEE Trans. Inform. Theory, 52: 1289-1306.

Figueiredo, M.A.T., R.D. Nowak and S.J. Wright, 2007. Gradient projection for sparse reconstruction application to compressed sensing and other inverse problem. IEEE J. Sel. Top. Sign. Proces., 1: 586-597.

Goldberg, R.G. and L. Riek, 2000. A Practical Handbook of Speech Coders. CRC Press, Boca Raton, FL., USA., ISBN-13: 9780849385254, Pages: 256.

Gunawan, T.S., O.O. Khalifa, A.A. Shafie and E. Ambikairajah, 2011. Speech compression using compressive sensing on a multicore system. Proceedings of 4th International Conference on Mechatronics, May 17-19, 2011, Kuala Lumpur, pp: 1-4.

Hu, Y. and P.C. Loizou, 2007. Subjective evaluation and comparison of speech enhancement algorithms. Speech Commun., 49: 588-601.

Kubin, G. and W.B. Kleijn, 1999. On speech coding in a perceptual domain. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 1, March 15-19, 1999, Phoenix, AZ., USA., pp: 205-208.

Kumar, A., 2000. Low complexity ACELP coding of 7 kHz speech and audio at 16 kbps. Proceedings of the IEEE International Conference on Personal Wireless Communications, December 17-20, 2000, Hyderabad, India, pp: 368-372.

McCree, A.V. and T.P. Barnwell III, 1995. A mixed excitation LPC vocoder model for low bit rate speech coding. IEEE Trans. Speech Audio Process., 3: 242-250.

Ming, Y., 2004. Low bit rate speech coding. IEEE Potentials, 23: 32-36.

Rauhut, H., K. Schnass and P. Vandergheynst, 2008. Compressed sensing and redundant dictionaries. IEEE Trans. Inform. Theory, 54: 2210-2219.

Schroeder, M. and B. Atal, 1985. Code-Excited Linear Prediction (CELP): High-quality speech at very low bit rates. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 10, April 26-29, 1985, Tampa, FL., USA., pp: 937-940.

Spanias, A.S., 1994. Speech coding: A tutorial review. Proc. IEEE, 82: 1541-1582.

Sreenivas, T.V. and W.B. Kleijn, 2009. Compressive sensing for sparsely excited speech signals. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, April 9-24, 2009, Taipei, Taiwan, pp: 4125-4128.

Vasuki, A. and P.T. Vanathi, 2006. A review of vector quantization techniques. Potentials IEEE., 25: 39-47.

Wang, S., 2000. Variable rate multi-mode excitation coding of speech at 2.4 kbps. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 3, June 5-9, 2000, Istanbul, Turkey, pp: 1395-1398.