Heliyon 5 (2019) e01820

Contents lists available at ScienceDirect

Heliyon

www.elsevier.com/locate/heliyon

A scalable speech coding scheme using compressive sensing and orthogonal mapping based vector quantization

M.S. Arun Sankar *, P.S. Sathidevi

Department of Electronics and Communication Engineering, National Institute of Technology Calicut, Kerala, India

ARTICLE INFO

Keywords:
Electrical engineering
Speech processing
Wavelet
Speech coding
CELP
Compressive sensing
Speech compression

ABSTRACT

A novel scalable speech coding scheme based on Compressive Sensing (CS), which can operate at bit rates from 3.275 to 7.275 kbps is designed and implemented in this paper. The CS based speech coding offers the benefit of combined compression and encryption with inherent de-noising and bit rate scalability. The non-stationary nature of speech signal causes the recovery process from CS measurements very complex due to the variation in sparsifying bases. In this work, the complexity of the recovery process is reduced by assigning a suitable basis to each frame of the speech signal based on its statistical properties. As the quality of the reconstructed speech depends on the sensing matrix used at the transmitter, a variant of Binary Permuted Block Diagonal (BPBD) matrix is also proposed here which offers a better performance than that of the commonly used Gaussian random matrix. To improve the coding efficiency, formant filter coefficients are quantized using the conventional Vector Quantization (VQ) and an orthogonal mapping based VQ is developed for the quantization of CS measurements. The proposed coding scheme offers the listening quality for reconstructed speech similar to that of Adaptive Multi rate - Narrowband (AMR-NB) codec at 6.7 kbps and Enhanced Voice Services (EVS) at 7.2 kbps. A separate de-noising block is not required in the proposed coding scheme due to the inherent de-noising property of CS. Scalability in bit rate is achieved in the proposed method by varying the number of random measurements and the number of levels for orthogonal mapping in the VQ stage of measurements.

1. Introduction

Speech is the fundamental means of communication among human beings. The constraints on transmission bandwidth arose the need for speech compression but without degrading quality. This is made possible by speech coding that aims at the representation of digital speech using fewer bits as possible with the minimum perceptual quality loss. Even though the coding parameters, quality and bit rate are inversely related, it is possible to augment the delivering voice quality by lowering the constraints of latency and complexity [1, 2]. The advancements in speech technology have made it a prime area of research in signal processing due to the increased demand for communication services [3, 4, 5]. Along with the technological advancements in digital domain, the need for efficient utilization of transmission network bandwidth and data storage is also increasing.

CS deals with the non-linear recovery of the signal by exploiting its sparsity in some domain. The popularity of CS paradigm in signal processing is due to its advantages; inherent compression, de-noising and encryption and less complexity at the transmitter [6]. The prime need in CS application is to find a sparsifying basis for the signal under consideration and to know the level of sparsity for proper reconstruction. Most of the naturally occurring signals exhibit sparsity in some domain and this opened the door for applying CS in image, video, speech, and bio-signals [7, 8, 9, 10].

In speech processing, CS is used for coding, enhancement and encryption [11, 12]. Daniels et al. [8] proposed a scalable Algebraic Code Excited Linear Prediction (ACELP) using the compressed framework which offers scalability by varying the dimensions of measurements. In [13], a hybrid dictionary of Linear predictive Coding (LPC) and Discrete Cosine Transform (DCT) is proposed for the representation of speech signal to account for the time-varying nature. In our previous work [12], we did a dynamic selection of basis as per the statistical properties of the signal. In [14], the CS concept is used for both speech coding and enhancement by exploiting its sparsity in the frequency domain. The CS based coding provides the inherent advantages of scalability by varying the number of measurements, enhancement due to sparse recovery as noise is non-sparse and encryption due to sensing matrix. In

* Corresponding author.
E-mail addresses: arun_p150036ec@nitc.ac.in (M.S.A. Sankar), sathi@nitc.ac.in (P.S. Sathidevi).

https://doi.org/10.1016/j.heliyon.2019.e01820
Received 8 October 2018; Received in revised form 21 March 2019; Accepted 22 May 2019

addition to the benefits offered by CS, there exist some drawbacks that cause technical difficulties which have been addressed in this paper.

1.1. Basics of CS frame work

Compressive sensing relies on the reconstruction of high dimensional signal, $X \in \mathbb{R}^N$ from its low dimensional linear projections, $Y \in \mathbb{R}^M (M \ll N)$ by exploiting sparsity K, of the signal in a certain basis ψ, where $K < M < N$.

The measurement vector, $Y = \phi X$, where ϕ is called the sensing matrix and X, the signal of interest is represented in ψ as, $X = \psi S$, where S is the sparse vector that contains only K non zero or significant values, the rest $(N - K)$ values can be discarded. So,

$$Y = \phi X = \phi \psi S = AS \tag{1}$$

Solving (1) is an NP-Hard combinatorial problem. Candes et al. [15] proposed l_0 minimization based solution for sparse recovery and is given below as

$$\min_{U \in \mathbb{R}^N} \| U \|_0 \quad such \quad that \quad AU = Y \tag{2}$$

To obtain a unique solution for (2), A should satisfy Restricted Isometry Property (RIP) of order K given as

$$(1 - \delta_K) \| U \|_2^2 \leq \| AU \|_2^2 \leq (1 + \delta_K) \| U \|_2^2 \quad where \quad \| U \|_0 \leq K \tag{3}$$

If A satisfies RIP, then there exists RIP constant δ_K in the interval [0 1] that satisfies (3). Solving (2) cannot be done in polynomial time and an approximate solution can be obtained by replacing l_0 norm with l_1 norm as

$$\min_{U \in \mathbb{R}^N} \| U \|_1 \quad such \quad that \quad AU = Y \tag{4}$$

Since RIP is difficult to determine [10], another condition is minimum correlation between ϕ and ψ which can be easily verified by evaluating coherence, $\mu(\phi, \psi)$ expressed as follows:

$$\mu(\phi, \psi) = \sqrt{N} \max_{i,j} \frac{| \langle \phi_i, \psi_j \rangle |}{\| \phi_i \|_2 \| \psi_j \|_2} \tag{5}$$

where ϕ_i (for $i = 1$ to M) represents the rows of ϕ and ψ_j (for $j = 1$ to N) represent the columns of ψ. Signal of interest X is estimated as $\hat{X} = \psi \hat{U}$ and the accuracy of estimation increases with decrease in value of $\mu(\phi, \psi)$ and if it is nearer to 1, then ϕ and ψ are incoherent [10].

1.2. Key problems addressed in this paper

1.2.1. Efficiency of sensing matrix

The most commonly used sensing matrices include random matrices obtained using independent and identically distributed (i.i.d.) Gaussian process, non binary orthogonal/non-orthogonal matrices, and sparse matrices. The Gaussian sensing matrix is most popular due to its incoherence with most of the bases. Due to its non zero non-integer values, the matrix is denser, and hence the computational complexity and storage requirements are very high [9]. This will also add extra cost to the hardware implementation.

For capturing measurements $Y \in \mathbb{R}^M$ from signal $X \in \mathbb{R}^N$, the dimension of sensing matrix will be $M \times N$. Using Gaussian random matrix for sensing measurements, the storage and computational requirements respectively are $M \times N$ and $M \times N$ multiplication and $M \times (N - 1)$ additions. Since multiplications are computationally expensive, this will increase the sensing time for measurements.

1.2.2. Sparsity of speech signal

Most of the naturally occurring signals show sparsity in some domain. Most of the CS based speech coding in literature has used DCT as the sparsifying domain. But for speech, due to its time-variant nature, neither the domain nor the level of sparsity is fixed. In [13], a hybrid basis consisting of DCT and LPC is used for sparse representation of frames. This choice of hybrid basis will increase the reconstruction accuracy but also the computational time for sparse recovery due the usage of greedy algorithms. Thus, the associated challenge is to identify a sparsifying basis that gives a compact representation of the speech signal for carrying out sparse recovery based reconstruction efficiently. But in the case of speech, which is highly non-stationary in nature, this is very difficult.

1.2.3. Quantization of transmission parameters

To bring down the transmission bit rate to medium, the quantization of parameters is required. The quantization error of measurements has a great impact on the reconstruction accuracy due to its significance in sparse recovery. In our previous works [12, 16], the probability distribution of speech is exploited for quantization of measurements but it requires both the mean and variance of each measurement vector to be transmitted along with indices. In [14], the quantization of measurements is done using Analysis-by-synthesis (AbS) approach which requires the reconstruction of signal at transmitter that increases its complexity. Due to a large number of measurements, scalar based quantization schemes will not give a significant reduction in bit rate.

1.3. Major contribution of this work

The following solutions are proposed in this paper for addressing the above issues;

1. Instead of the most widely used Gaussian random matrix, a new sparse binary matrix which is a variant of Binary Permuted Block Diagonal (BPBD) is proposed for sensing measurements that enhances the perceptual quality of reconstructed speech and also reduces the transmitter complexity.
2. Instead of finding a single sparsifying basis for speech signal, we put forward the solution of allocating separate basis for each category of speech signal that enhances the sparsity. For this purpose each speech frame is categorized as voiced, unvoiced or inactive using statistical features, Zero Crossing Rate (ZCR) and Average power that ensures the allocation of most appropriate basis for efficient coding. For each category, the following hybrid bases are experimentally determined which is beneficial for CS recovery.
   - Voiced - The hybrid basis of DCT and Daubechies wavelet of order 8.
   - Unvoiced - Hybrid basis consisting of LPC and Daubechies wavelet of order 3.
   - Inactive - Daubechies wavelet of order 3.
3. To reduce the bit rate of coding mechanism, a conventional Vector Quantization (VQ) is used for LPC coefficients and an orthogonal mapping based VQ scheme is proposed for the quantization of CS measurements. The orthogonal VQ provides the benefit of direct summation due to its property of zero correlation.

The rest of the paper is organized as follows; Section 2 describes the functionalities of various blocks in the encoding and decoding sections of the algorithm. The frame work of the proposed sparse binary sensing matrix is explained in section 3. The determination of various speech categories and the algorithm for identification of category is given in section 4. The selection of optimal basis that enhances sparsity for each category of the speech signal and the quantization of transmitting parameters is given respectively in sections 4 and 5. The validation of the proposed scheme and the comparison of its performance with other

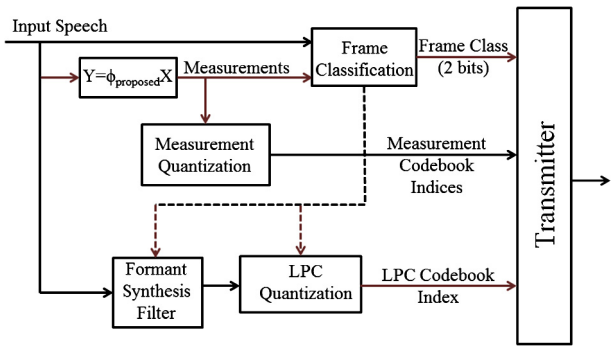

Fig. 1. Encoder block of CS based proposed coding scheme.

Table 1
Bit representation for frame categorization and required parameter quantization.

Category	Bit representation	Quantization
Voiced	11	Measurements
Unvoiced	10	Measurements and LPC coefficients
Inactive	00	Measurements

coding mechanisms are summarized in Section 7 followed by the conclusion and future work given in section 8.

2. Architecture of the proposed speech coding mechanism

A detailed description of the encoding and decoding sections of the proposed CS based speech coding is given in the following subsections.

2.1. Encoder

The incoming Narrow-Band (NB) signal sampled at 8 kHz is converted to frames of 20 ms duration. The functionalities of encoder include capturing measurements followed by frame categorization, quantization of measurements and LPC coefficients as shown in Fig. 1.

From the input speech frame consisting of 160 samples, the measurements that can be of various lengths (46, 64, 80, 96, 112, 128, 144) depending upon the reduction factor are obtained using the proposed sensing matrix. From the average power of measurements and ZCR of the frame, the speech category has been identified and a 2-bit representation shown in Table 1 is sent to the receiver for identification. The measurements are quantized using 10-bit orthogonal VQ with multiple levels of mapping. Along with this, if the frame is categorized as unvoiced, the LPC coefficients are also transmitted which have been quantized using 7-bit conventional VQ.

2.2. Decoder

From the transmitted parameters; frame category, measurement indices and LPC index, the decoder has to reconstruct the frame using the appropriate basis. The 2-bit frame category representation is used for the selection of basis for recovery as shown in Fig. 2. Using the transmitted indices, the measurements are obtained from the codebooks for the known number of levels of mapping. The selected basis ψ and the sensing matrix ϕ are key ingredients in the construction of sparse recovery matrix A, and this is used for reconstructing the frame of speech signal.

3. Proposed sparse binary sensing matrix

Sparse binary sensing matrix would be a good replacement for Gaussian iid random matrix. The major advantage of sparse matrix lies in the ease of implementation due to less number of non zero values that reduces memory needs, computational power, and sensing delay. In [9],

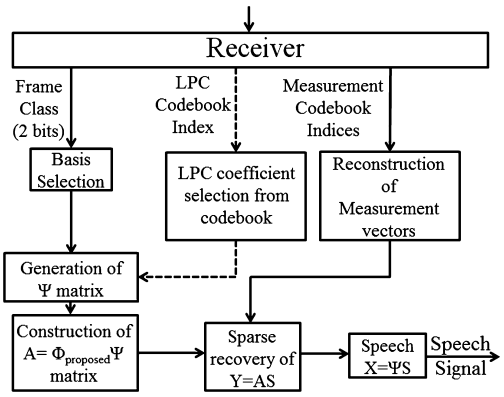

Fig. 2. Decoder block of CS based proposed coding scheme.

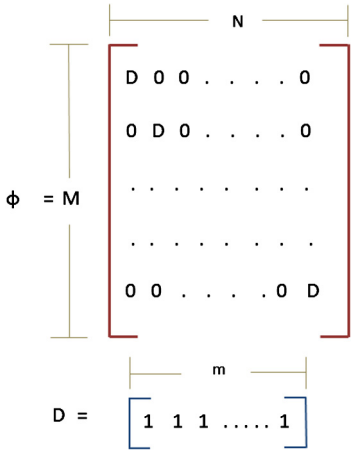

Fig. 3. Sparse binary sensing matrix proposed in literature.

Table 2
Values of m determined as a function of s for CS based coding scheme.

s	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
m	5	3.33	2.5	2	1.67	1.43	1.25	1.11

BPBD is used for sensing natural images, and a modified form of BPBD is proposed in [10] for sensing physiological signals. The major benefit obtained from using BPBD, which is a sparse binary sensing matrix, is its less complexity in comparison to sparse sensing matrix by the elimination of multiplication. This gave the motivation for replacing the Gaussian random matrix with sparse binary sensing matrix due to the similarities of properties that exist among natural signals that include image, speech and bio-signals.

The sensing matrix proposed in [10] for acquiring bio-signals is shown in Fig. 3. It consists of identical diagonal blocks D having m ones, where $m = N/M$ and rest all elements of the sensing matrix are zeros. The dimension of sensing matrix is $M \times N$ where M is the number of measurements to be acquired from signal of length N. In [10], the ratio of length of signal to measurements m, is always considered as integers. But for speech, $N = 160$ and it represents the length of a speech frame of duration 20 ms. The number of measurements M acquired from a speech frame is solely determined by the reduction factor s that spans the interval [0 1]. s is the ratio of number of measurements to length of speech signal denoted by $s = M/N$, and thus m and s are inversely related. m have mostly non-integer values for different values of s as shown in Table 2. Having non-integer values for m is not acceptable as it denotes the number of ones in the diagonal block and hence we modified the diagonal block D for fractional values of m.

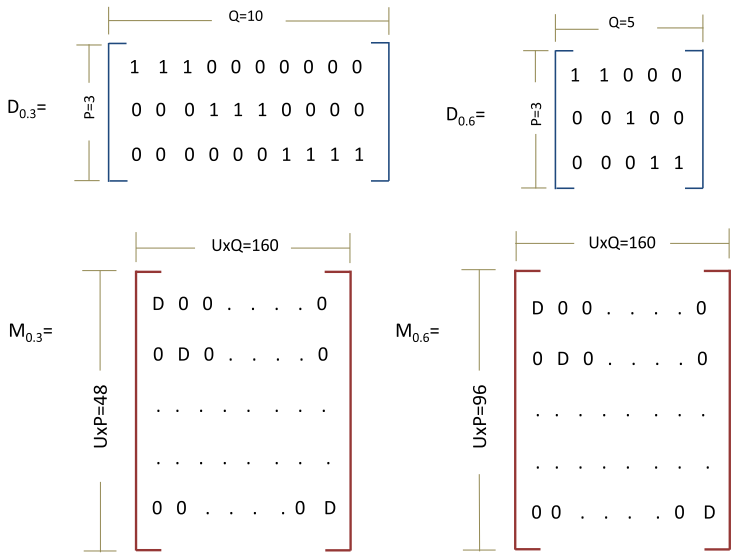

Fig. 4. The proposed sensing matrix M for $s = 0.3$ and 0.6 with block diagonal D. The subscript indicates the value of s.

In this paper, the sensing matrix for speech signal is obtained in either of two ways; (1) if m is an integer, the method proposed in [10] is used for the generation of sampling matrix. (2) For fractional values of m, customized diagonal blocks are generated in accordance with the description given in Algorithm 1.

From a speech frame of length N, the generation of sensing matrix using Algorithm 1 for a particular value of s first requires the evaluation of $m = 1/s$ and determine whether m is integer or fraction.

(i) If m is integer, diagonal block D is generated that contains m ones. First row of sensing matrix r_1 will have D followed by all zeros. The rest $(M-1)$ rows are obtained by circular shifting r_1, m times. For, e.g., let $s = 0.5$, then $m = 2$ which is integer and number of measurements $M = N \times s = 80$. So $D = [1 \quad 1]$ and r_1 will have D followed by 158 zeros. The second row of sensing matrix will be generated by circular shift of r_1 by 2 to right and in this similar manner, rest 78 rows are built and stacked to form the sensing matrix.

(ii) If m is non-integer, then obtain m in the form Q/P. Then, the dimension of diagonal block D is $P \times Q$. The number of diagonal blocks is $U = N/Q = M/P$. Next step is the generation of customized diagonal block D. Assign the integer values above and below m respectively to m_2 and m_1. D will be consisting of rows with either m_1 or m_2 ones. Number of rows with m_2 ones, $N_{m2} = Q - P \times m_1$, and the rest $P - N_{m2}$ rows will have m_1 ones. These rows with m_1 and m_2 ones are circular shifted and stacked to form the diagonal block D. By placing these diagonal block in a zero matrix of dimension $M \times N$, the sensing matrix is formed.

The sensing matrices developed using the proposed algorithm for $s = 0.3$ and 0.6 is illustrated in Fig. 4. For $s = 0.3$, $m = 3.33$ and the reduced integer representation is $10/3$. Hence the dimension of diagonal block $D_{0.3}$ is 3×10. The number of diagonal blocks is 16. Here, $m = 3.33$, so $m_1 = 3$ and $m_2 = 4$. The $D_{0.3}$ will be having one row with 4 ones and two rows with 3 ones. The sensing matrix is formed by placing $D_{0.3}$ diagonally. Similarly, for $s = 0.6$, $m = 1.6667$ and the reduced integer representation is $3/5$ that gives the number of diagonal blocks $U = 32$. The $D_{0.6}$ will be consisting of two rows with 2 ones and another row with only single one. By placing $D_{0.6}$ diagonally, the sensing matrix is generated.

The correlation of the proposed sensing matrix with that of DCT, one of the sparsifying bases of speech is computed using equation (5) and is shown in Fig. 5. The μ values for $\phi_{Guassian}$ has a little variation with s but for $\phi_{proposed}$, the value of μ decreases with increase in s. The μ for $\phi_{Gaussian}$ is less than that of $\phi_{proposed}$ only for $s \leq 0.1$ which corresponds to $M = 10$ which is not practically feasible for the optimum reconstruction due to high compression ratio. For $\phi_{proposed}$, the μ remains almost constant from $s = 0.5$ to 0.9, an optimum range for reconstruction.

Algorithm 1 Formation of sensing matrix.

Given: The length of frame, N and the number of measurements, M.
Find: The sensing matrix, M_s of dimension $M \times N$.

Initialize: $m = N/M$
if m is integer **then**
  $row_t = \{1_1 1_2 \cdot\cdot 1_m 0_1 0_2 \cdot\cdot 0_{(N-m)}\} \longleftarrow m$ ones and $(N-m)$zeros
  for $i = 1$ to M **do**
    $row_i = row_t$
    row_t = circular shift row_t right by m times
  end for
  $M_s = \{row_1^T row_2^T \cdot\cdot row_M^T\}^T$
else
  $m = Q/P$, reduced integer form and $U = N/Q$.
  Initialize: m_1, m_2 respectively integer value just below and above m.
  Number of rows with m_2 ones, $N_{m2} = (Q - P \times m_1)$.
  $N_{m1} = P - N_{m2}$.
  if $N_{m2} = 1$ **then**
    row position, R_{pm2} = first row(r_1)
  else if $N_{m2} = 2$ **then**
    $R_{pm2} = r_1$ and r_M
  else if $N_{m2} \geq 3$ **then**
    $R_{pm2} = r_1$, r_M and equally spaced rows between r_1 and r_M.
  end if
  $R_{pm1} = \{r_1, r_2, \cdots r_M\} - R_{pm2}$.
  $row_{t1} = \{1_1 1_2 \cdot\cdot 1_{m1} 0_1 0_2 \cdot\cdot 0_{(N-m1)}\} \longleftarrow m1$ ones and $(N-m1)$zeros
  $row_{t2} = \{1_1 1_2 \cdot\cdot 1_{m2} 0_1 0_2 \cdot\cdot 0_{(N-m2)}\} \longleftarrow m2$ ones and $(N-m2)$zeros
  for $k = 1$ to M **do**
    if $r_k \in R_{pm1}$ **then**
      $row_k = row_{t1}$
      row_{t1} = circular shift row_{t1} right by m_1 times
      row_{t2} = circular shift row_{t2} right by m_1 times
    else
      $row_k = row_{t2}$
      row_{t1} = circular shift row_{t1} right by m_2 times
      row_{t2} = circular shift row_{t2} right by m_2 times
    end if
  end for
  diagonal block, $D_b = \{row_1^T row_2^T \cdot\cdot row_M^T\}^T$.
  M_s = matrix with U number of D_bs.
end if

4. Determination of various categories for speech

Identification of a unique basis that provides a sparse representation for speech signal is not possible. Hence, we proceed to divide the speech signal into various categories and thereby to determine an appropriate basis that suits best for each category in terms of enhancing

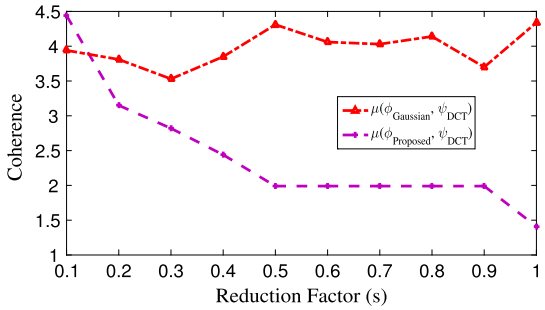

Fig. 5. The deviations in $\mu(\phi, \psi_{DCT})$ with values of $s = 1/m$ for $\phi_{Gaussian}$ and $\phi_{proposed}$.

sparsity, a primary factor that determines reconstruction accuracy. The variations in characteristics of the speech signal is prominently due to the differences in speech production mechanism.

The different categories of speech samples that are manually marked are illustrated in Fig. 6 using relevant parts of clean speech samples from NOIZEUS database [17]. The various types of speech that have similarity in the structure can be collectively grouped under same category. The various categories derived from the allocation of optimum basis are given below.

1. *Voiced.* From the various types of speech sounds, the categories that have periodicity due to vocal chord vibration are placed under voiced category for basis selection. This includes vowels, diphthongs and some from unvoiced category with certain periodicity due to simultaneous vibration and frication such as nasal consonants, voiced plosives, and stops.
2. *Unvoiced.* The components of the speech signal that are stochastic can be categorized as unvoiced. These components include fricatives, plosives or bursts, breath noise, noise in whispered speech and creaky voice that are of high-frequency components and are of noise like.
3. *Inactive.* The regions of speech signal between voice activity are placed under this category. These consist of silence regions and signal components that have randomness and are of very low energy. To have a fair reconstruction of the signal, the coding of these parts of the speech signal is a necessity.

4.1. Speech frame categorization

The TIMIT database [18] consisting of 6466 speech samples are split into two halves for performing experiments. The speech samples of first half (TIMIT A) is used for the development of frame categorization method and the second half (TIMIT B) is used for the test and evaluation of the proposed speech categorization method. Due to the differences associated with the production of above mentioned speech categories and the existence of some correlation among them, a multi-feature based categorization will give better accuracy. The most commonly used features extracted from speech signal for grouping are Mel-frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), energy of frames, pitch and LPC coefficients. The categorization is based upon the comparison of feature values with respect to a predefined threshold set for each feature. In [19], the author used ZCR, energy, correlation between adjacent samples, first LPC coefficient and energy of prediction error for categorization using pattern recognition approach. Qi and Hunt [20] suggested non-parametric methods based on multilayer feed forward network for the categorization of speech. Mojtaba et al. [21] used a multi-feature categorization with cepstral peak, ZCR and Auto-Correlation Function (ACF) parameters using clustering method.

We conducted experiments using various combination of features and selected ZCR and energy of frames for the categorization as proposed in our previous work due to its performance [12].

Zero Crossing Rate. It defines the average rate of sign change in a speech signal and indicates the frequency region where the energy is concentrated. The ZCR for speech signal $s(n)$ of length L is defined as

$$ZCR = \frac{1}{L-1} \sum_{n=1}^{L-1} I_{R<0}(s(n)s(n+1)) \tag{6}$$

where $I_{R<0}$ is an indicator function. Due to the quasi periodic nature of the voiced segments of speech signal, the ZCR is low and the unvoiced regions have a large ZCR. For silence or inactive regions, the ZCR lies between that of voiced and unvoiced but overlaps with both at boundaries.

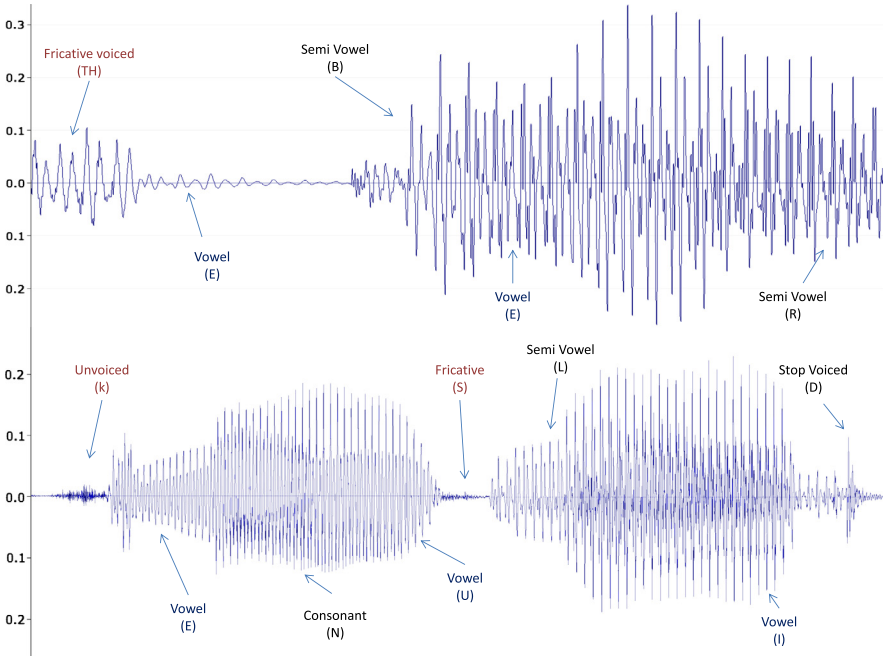

Fig. 6. The speech sample that shows various types of speech for words: “the birch” (upper) and “canoe slid” (down).

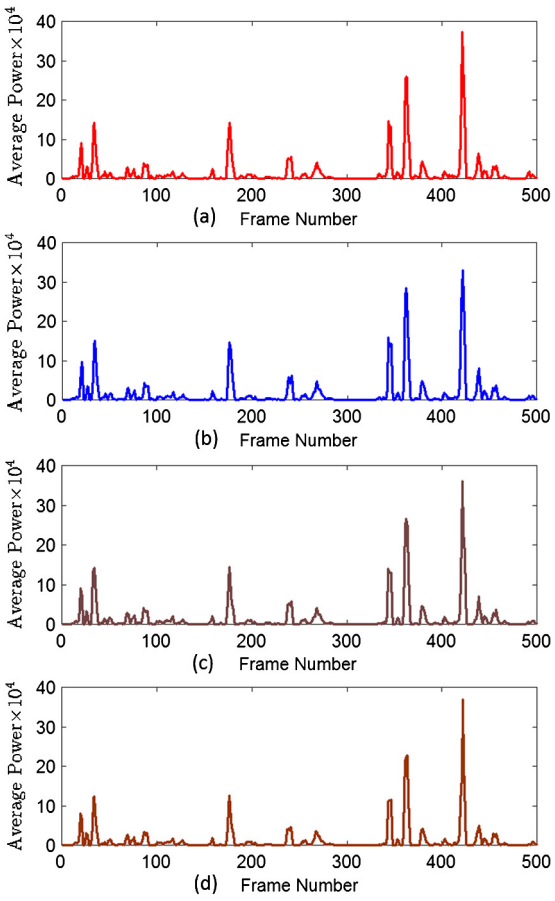

Fig. 7. Average power computation using frame (a) and its random measurements for $s = 0.8$ (b), 0.6 (c) and 0.4 (d).

Average Power. The voiced regions of speech have higher energy than the unvoiced regions of speech. The average power P_{Ax} of speech signal $x(n)$, of length L is

$$P_{Ax} = \frac{1}{L}\sum_{j=1}^{L} | x_j(n) |^2 \tag{7}$$

where x_j denotes the jth sample of $x(n)$. The silence or inactive regions of speech have the lowest average power.

Average power computation using measurements. The measurements $Y \in \mathbb{R}^M$ are obtained from $X \in \mathbb{R}^N$ for $m = N/M$ using the proposed sensing matrix as $Y = \phi_{proposed} X$. The average power of Y is computed using the equation as follows;

$$P_{Ay} = \frac{1}{mN}\sum_{j=1}^{M} | y_j(n) |^2 \tag{8}$$

The scaling of power by $1/(mN)$ instead of $1/M$ will compensate for the effect of $\phi_{proposed}$ that augments the value of Y. Each measurement y will be the sum of m adjacent speech samples x, and the summation of y^2 will be consisting of mN values of x^2. Since the average power of speech samples is required, the summation of y^2 is scaled by $1/(mN)$. The average power is evaluated using both X and Y with $N = 160$ for 500 frames of NB speech samples from TIMIT A database as a function of s and its values are plotted in Fig. 7 where the average power is scaled by 10^4 for better visibility. It is evident from the figure that the envelope of average power is preserved with a little variation while using measurements which is sufficient for the discrimination. This method reduces the computational burden in the classification of frames as the number of multiplications and additions, respectively are reduced from $N^2, N-1$ to $M^2, M-1$.

Table 3
Frame categorization method.

Speech category	ZCR	Average power	Category defined
Voiced, diphthongs	<0.1	≥ 0.1	Voiced
nasal components, Voiced plosives	0.1-0.27 >0.27	≥ 0.1 ≥ 1	
Plosives, bursts, breath noise, creaky voice	>0.27	<1	Unvoiced
Silence	<0.27	<0.1	Inactive

Table 4
Performance evaluation of proposed speech categorization method.

Category	Miss %	False %	Accuracy %
Voiced	2	1.5	96.5
Unvoiced	1.2	2.1	96.7
Inactive	0.8	0.9	98.3

4.2. Proposed method

The ZCR of the frame and average power of random measurements for $s = 0.5$ are evaluated for NB frames of both male and female speeches belonging to TIMIT A database that are manually categorized into voiced, unvoiced and inactive frames in order to obtain the correlation among the extracted parameters for different categories of speech. Here after, the average power means the average power scaled by 10^4 to have a better visibility in the discrimination. The categorization of the frames based on the ZCR and average power is summarized in Table 3. The thresholds 0.1 and 0.27 are found experimentally to get maximum fit with the manual categorization. The proposed method has been validated by experimentally categorizing 1000 speech frames from NB samples of TIMIT B database and its results are given in Table 4. The miss and false rates indicate respectively the ratio of true frames but not detected and false frames that are detected for a particular category using the categorization algorithm. The high accuracy values and low error values (miss and false rate) for voiced, unvoiced and inactive frames in Table 4 illustrates the capability of algorithm in speech categorization.

5. Identification of optimum basis for each category

For the above mentioned categories of speech, separate basis needs to be identified that suits best for each type in enhancing sparsity. Random wavelets are used for representing unvoiced sounds in [22], and in [23] super wavelets consisting of Daubechies wavelet of order three is used for unvoiced representation. In [24], the author proposed Morlet wavelet for the representation and categorization of voiced sounds. Kadambe and Srinivasan [25] used adaptive wavelets to model speech where the prominent coefficients are obtained using neural networks. A comparison of different wavelets for representation of speech revealed Daubechies as a suitable candidate [26]. Hence in this work, we have considered DCT, LPC and Daubechies family of wavelets from order 2 to 10 (db2 to db10) for the representation of speech.

5.1. Derivation of various bases

1. *LPC*

The representation of an arbitrary signal $s(n)$, using Linear Prediction (LP) model of order M is given as

$$s(n) = \Sigma_{m=1}^{M} a_m s(n-m) + e(n) \quad m = 1, 2.....M \tag{9}$$

where $e(n)$ is the prediction error and $a_m \in a = [a_1, ...a_M]$ are the LP coefficients. In matrix form, (9) can be rewritten as

$$s = Ae \tag{10}$$

where $A \in R^{N\times N}$ is the matrix that maps the error to synthesized signal. The matrix A can be obtained from the unit impulse response of $H(z)$ that denotes the transfer function of (9) expressed as

$$H(z) = \frac{1}{1 - a_1 z^{-1} - ..... - a_M z^{-M}} = \sum_{n=0}^{+\infty} h(n) z^{-n} \tag{11}$$

Every column of A matrix forms the basis vector and is given as

$$A \cong \begin{pmatrix} h(0) & 0 & ... & ... & 0 \\ h(1) & h(0) & ... & ... & 0 \\ h(2) & h(1) & ... & ... & 0 \\ ... & ... & ... & ... & ... \\ h(N-3) & h(N-4) & ... & ... & 0 \\ h(N-2) & h(N-3) & ... & h(0) & 0 \\ h(N-1) & h(N-2) & ... & h(1) & h(0) \end{pmatrix}$$

2. *Wavelet*

Any arbitrary real valued signal, $X \in \mathbb{R}^N$ can be represented in wavelet domain by identifying the daughter wavelets or nodes of wavelet decomposition tree that best suit the signal using entropy based best basis selection algorithm [27]. The signal X can be denoted in matrix form as $X = \Psi K$, where $\Psi = \begin{bmatrix} \psi_1 & \psi_2 & ... & ... & \psi_N \end{bmatrix}$ and $K^T = \begin{bmatrix} k_1^T & k_2^T & ... & ... & k_N^T \end{bmatrix}$. The ψ_j matrix that represents the basis corresponding to subspace of jth node is generated using the impulse response, $h_j = \left(h_{j1} h_{j2} ..... h_{jP}\right)$ obtained at jth node by passing an impulse down the tree from root node [28]. The rows of ψ_j matrix of dimension $P \times (2P-1)$ given in (12) is created by shifting the row matrix, h_j concatenated with P-1 zeros by one value at a time.

$$\psi_j = \begin{bmatrix} h_{j1} & h_{j2} & ... & h_{jP} & 0 & ... & 0 \\ 0 & h_{j1} & ... & h_{jP-1} & h_{jP} & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & h_{j2} & h_{j3} & ... & 0 \\ 0 & 0 & ... & h_{j1} & h_{j2} & ... & h_{jP} \end{bmatrix} \tag{12}$$

Every column of ψ_j matrix given in (12) is normalized and in a similar manner, the ψ_j matrices for the rest $N-1$ nodes are obtained. The final Ψ matrix is obtained by concatenating the ψ_j's of all N nodes.

5.2. Best basis selection for each speech category

The speech samples to determine the best basis for each category consists of both male and female speeches taken randomly from TIMIT A database that is band limited to 4 kHz followed by down sampling by a factor of 2. The optimum basis for a particular category is that basis which gives the minimum average basis count for the representation of speech signals belonging to that category and the procedure for basis selection is given below.

Frame work for basis selection

1. The frames of the speech signal, X_i for $i = 1, 2, ..., N$ are manually categorized by analysing the signal in Sonic Visualiser to obtain various categories $X = \{X_{c1}, X_{c2}, ..., X_{cT}\}$ where X_{ci} is the set of frame belonging to category ci. Here $T = 3$.
2. For all frames f_{ij} for $i = 1, 2, ...M$ belonging to speech category X_{cj}, the optimum wavelet tree structure $w_{cjk} = \{w_1, w_2, ...w_M\}$ is identified for WT_k wavelet using entropy based basis selection algorithm.
3. The structure of wavelet tree common to w_{cjk}, w_{cjkopt} is selected.
4. Steps 2 and 3 are repeated for each wavelet type WT_k, for $k = 2, 3.., P$ to obtain $w_{cj} = \{w_{cj1opt}, w_{cj2opt}, ..., w_{cjPopt}\}$.
5. The bases of wavelet families, $\psi_{cj} = \{\psi_{c1}, \psi_{cj1}, ..., \psi_{cjP}\}$ are obtained from w_{cj} using the method explained in the previous section.

Algorithm 2 Measurement quantization using orthogonal projection.

Given: The length of frame N, the measurement codebook, CB of dimension $L \times M$, the number of projection P and the measurement vector Y of length M.
Find: The orthogonal projections, $P_y = \{P_{y1}, P_{y2}, \cdot\cdot P_{yP}\}$.

```
Initialize: cb = CB
for j = 1 to P do
  for k = 1 to L do
    correlation_k = ⟨cb_k, Y⟩, cb_k is the kth row of cb
  end for
  correlation_opt = maximum of {correlation_1, correlation_2, ··· correlation_L}
  P_yj = cb_opt
  for m = 1 to L do
    cb1_m = cb_m − ⟨cb_m, cb_opt⟩
  end for
  cb = cb1
end for
P_y = {P_y1, P_y2, ·· P_yP}
```

6. The sparsity of X_{cj} in various bases, $\Psi_{cjtotal} = \{\psi_{cj}, \psi_{cjlpc}, \psi_{dct}\}$ is evaluated by sparse recovery using OMP greedy algorithm, $\hat{X}_c = \psi_{cj} b_j$ for $j = 1, 2, ..., P+2$ where $b_j = argmin \parallel X_c - \psi_{jc} b_j \parallel_2^2 + \gamma \parallel b_j \parallel_1^1$.
7. The evaluation is done by keeping an upper threshold for reconstruction error as $\varepsilon \leq \parallel X_c - \hat{X}_c \parallel_2$, $\varepsilon = 0.005$ and find $\parallel b_j \parallel_0$.
8. The basis ψ_{cjopt} from $\Psi_{cjtotal}$ that gives maximum sparsity or less reconstruction error is selected.
9. The steps from 2 to 8 are repeated for other categories cj to find out $\Psi_{opttotal} = \{\psi_{c1opt}, \psi_{c2opt}, \psi_{c3opt}\}$.

Optimum wavelet tree structure

The best wavelet packet tree structures using Daubechies wavelets from order 2 to 10 (db2 to db10) are obtained using entropy based best basis selection [27] and the structure common for all the frames is selected. It is experimentally observed that the common wavelet packet tree structures for all the frames considered are same even though bases are different. This common wavelet packet structure Tree1, has been shown by dashed line in the 5 level wavelet packet decomposition structure shown in Fig. 8. The average count of basis for different tree structures for a 5 level decomposition is shown in Fig. 9 for randomly chosen 50 male and female speeches. It can be found that the average basis count is minimum for Tree1 structure while reconstructing speech with an error $\leq = 0.005$ which indicates a higher level of sparsity for Tree1.

The optimum wavelet family is selected for the best tree structure Tree1 by finding the basis that gives minimum average basis count for 500 frames of speech category and the results of the experiment are summarized in Table 5 which also includes the hybrid basis consisting of wavelet with other basis.

(i) *Voiced*

Results show that db8 and db9 are good candidates for voiced frames due to their low average basis count. The voiced frames can be more sparsely represented by a sine waves due to its quasi-periodicity. Hence, the reconstruction of speech samples using the hybrid basis of wavelet db8 and DCT gives the minimum average basis count and has been selected as the basis for voiced speech components.

(ii) *Unvoiced*

Our previous work on dynamic basis selection [12, 16] used LPC basis for mapping unvoiced sounds. In [25], the author used Daubechies wavelet of order 3 due to the high-frequency nature for analysis and representation of unvoiced sounds. The average basis count is minimum for hybrid basis of db3 & LPC which is lower than that for db3 alone. So this hybrid basis of db3 & LPC of order 12 has been selected as the basis for unvoiced sounds.

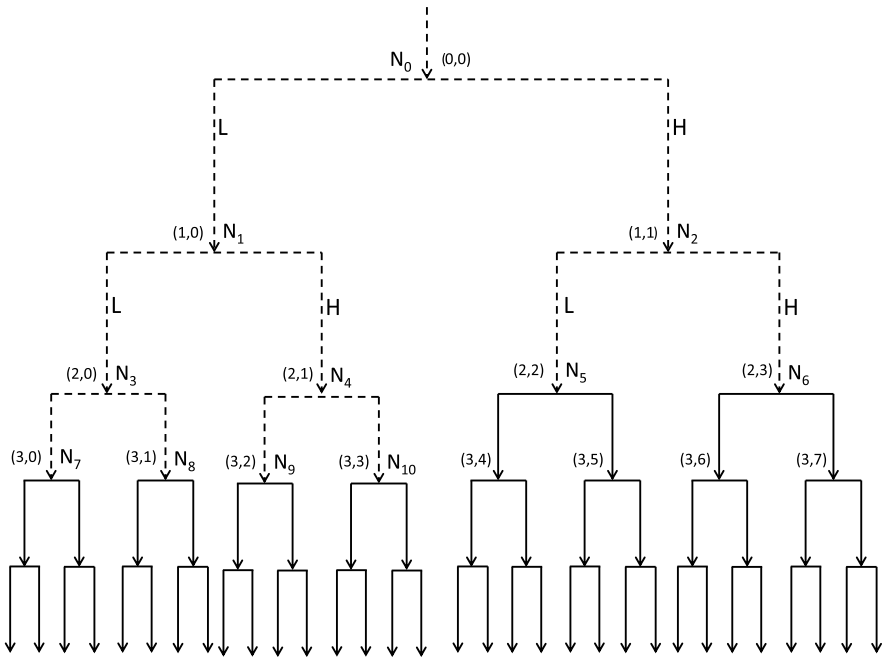

Fig. 8. Wavelet packet tree structure for 5 levels of decomposition and common tree structure $\{N_5N_6N_7N_8N_9N_{10}\}$ shown as dashed line.

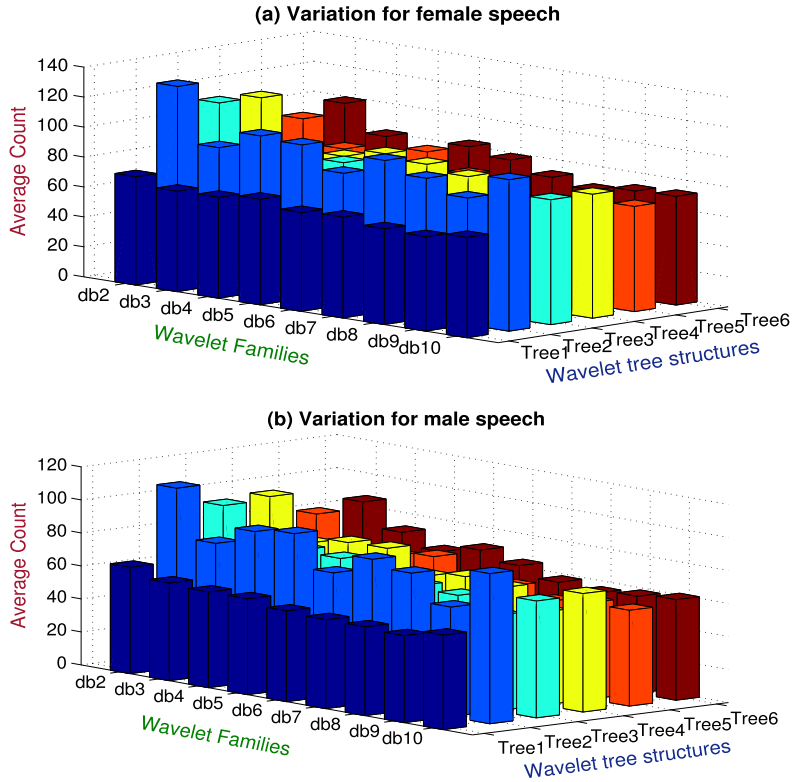

Fig. 9. The average basis count with different tree structures (Tree1: $N_5N_6N_7N_8N_9N_{10}$, Tree2: $N_{15}N_{16}N_{17}N_{18}N_9N_{10}N_2$, Tree3: $N_{15}N_{16}N_{17}N_{18}N_9N_{10}N_5N_6$, Tree4: $N_{31}N_{32}N_{16}N_{17}N_{18}N_9N_{10}N_5N_6$, Tree5: $N_7N_8N_9N_{10}N_{11}N_{12}N_{13}N_{14}$, Tree6: $N_{15}N_{16}N_{17}N_{18}N_9N_{10}N_{11}N_{12}N_{13}N_{14}$) using Daubechies wavelets for (a) 50 female speakers and (b) 50 male speakers.

(iii) *Inactive*

As displayed in Table 5, it is clear that the average basis count for db3 is nearer to that of the hybrid basis of LPC & DCT which gives a minimum value. As the inactive regions have little information, db3 will be a suitable candidate for representation as this gives a reduction in the bit rate.

6. Quantization of parameters

The success of speech coding mechanism lies not only on the quality of reconstruction but also on its efficiency in reducing bit rate without perceptional loss [29]. For the proposed frame work, the measurements obtained and the LPC coefficients of each frame need to be transmitted. As the LPC coefficients are required only for unvoiced components

Table 5
The average basis count for greedy reconstruction of various speech categories using Tree1 wavelet structure for Daubechies family db2 to db10 and its combinations using DCT and LPC.

Basis	Count	Basis	Count	Basis	Count	Basis	Count
				Voiced			
db2	63.05	db6	54.59	db10	56.28	db7&DCT	44.31
db3	57.62	db7	55.39	db4&DCT	44.72	**db8&DCT**	**43.90**
db4	57.90	db8	54.90	db5&DCT	44.87	db9&DCT	43.95
db5	57.07	db9	55.14	db6&DCT	44.52	db10&DCT	44.36
				Unvoiced			
db2	55.40	db8	55.22	db5	59.19	db3&DCT	52.65
db3	52.24	db9	54.96	db6	53.53	LPC	68.82
db4	55.97	db10	55.26	db7	55.13	**db3&LPC**	**48.35**
				Inactive			
db2	14.90	DCT	22.90	**db3**	**14.25**	LPC	17.17

which constitute only nearly 30% of the speech signal, the quantization of measurements of each frame that has to be transmitted plays a leading factor in determining the transmission bit rate.

6.1. LPC quantization

Standard vector quantization is applied to LPC coefficients of order 12 for unvoiced frames. The codebook is generated by clustering the LPC coefficients and finding the centroids using LBG algorithm. The LPC coefficients are derived for NB speech signals of TIMIT A and clean region of NOIZEUS databases for a frame size of 160 that has a collective duration of 323 minutes and sampled at 8 kHz. Due to the instability associated with LPC coefficients, they are converted to Line Spectral Frequencies (LSF) before quantization. The codebooks of size 256, 512 and 1024 are constructed to measure the impact of quantization. Using vector quantization, the bit rate can be reduced from 2400 bps (assuming 50% frames are unvoiced) allocating 8 bits/LPC coefficient to 250 bps for a codebook of length 1024. Since the LPC coefficients are obtained only for unvoiced frames, a further improvement is obtained by using codebooks that are generated using unvoiced frames of databases alone and its results are given in the next section.

6.2. Measurement quantization

The quantization error of measurements has a great impact on the reconstruction accuracy due to its significance in sparse recovery. In our previous works [12, 16], the probability distribution of speech is exploited for quantization of measurements which requires both the mean and variance of each measurement vector to be transmitted along with indices. Due to a large number of measurements, scalar based quantization schemes will not give a significant reduction in bit rate. The various schemes considered including our proposed scheme for the quantization of measurements are as follows;

(i) *Gaussian codebook based quantization*

The measurement vectors obtained are mapped to Gaussian codebook of length 1024. Due to the difference in the length of measurement vectors with reduction factor s, ranging from 0.9 to 0.2, various Gaussian codebooks of size $1024 \times (Ns)$ are generated. The measurement vectors are normalized before mapping by subtracting its mean and dividing by its variance. The index corresponding to the codebook entry that has maximum closeness to the measurement vector is selected by l_2 norm minimization for quantization error. These mean and variance are transmitted along with the codebook index to the receiver for recovery.

(ii) *Measurement codebook based quantization*

Measurement codebooks of length 1024 and 2048 are generated using LBG algorithm for various values of s, from the measurement vectors of training data using $\phi_{proposed}$ as the sensing matrix.

(iii) *Proposed orthogonal measurement quantization*

The measurement vectors are divided into orthogonal projections which have the most significant property of zero correlation among them. The orthogonal projections are directly added for the reconstruction of measurement vector. The quantization error decreases with increase in orthogonal projections. The initial codebook used for projection can be any of the measurement codebooks generated as explained before. The procedure for mapping the measurement vectors is explained in Algorithm 2. The subsequent codebooks used after the first mapping is generated from the residue of mapping the other codebook entries to the selected entry. The indices of the codebooks are transmitted, and they are used for reconstruction of measurement vectors by generating the codebooks in a similar manner.

The performance evaluation of the proposed coding methodology is carried out in the next section.

7. Results & discussion

The bit rate efficiency of the proposed coding method is compared with that of the conventional ACELP and standardized codecs such as Adaptive Multi-rate (AMR) [30], AMR-Wideband (AMR-WB) [31] and Enhanced Voice Services (EVS) [32]. The implementation details of 8 kbps ACELP are as follows;

1. Frame length 20 ms with Hamming window
2. 24-bit LPC indexing using split VQ
3. Allocation of 34 bits per subframe with 10 and 7 bits respectively for index and gain of each code book

The proposed method is also compared with CS based speech compression mechanisms proposed in [14] and [33]. The speech samples used for carrying out the experiments are selected at random from both male and female speakers of TIMIT B database that are band limited to 4 kHz and the results shown are average obtained for 50 speech samples. The coding performance is evaluated by the following metrics.

1. Signal-to-Noise-Ratio (SNR) for the original speech signal X and its reconstructed version $\hat{X}$, is defined as
$$SNR\ (\text{dB}) = 20log\frac{\| X \|_2}{\| X - \hat{X} \|_2}$$
2. Mean Square Error (MSE) given as
$$\frac{\sum_N \| X - \hat{X} \|_2^2}{N}$$
where N is the length of the signal.
3. Perceptual Evaluation of Speech Quality (PESQ) is an objective method to test the speech quality defined in the ITU-T P.862 standard [34].
4. MOS (Mean Opinion Score) is the average of informal listening quality results for 20 speakers using a 5 point scale described as follows;

Rating	5	4	3	2	1
Quality	Excellent	Good	Fair	Poor	Bad

To validate the improvements in the individual sections of the proposed coding scheme, only a few from the above mentioned quality measures are required.

Table 6
Reconstructed speech quality as a function of reduction factor using DCT basis for both $\phi_{proposed}$ and $\phi_{Gaussian}$.

Sensing matrix	SNR (dB)			
	$s = 0.8$	0.6	0.4	0.2
$\phi_{proposed}$	13.18	9.94	6.42	3.26
$\phi_{Gaussian}$	12.55	8.69	4.20	0.14

Table 7
Performance comparison as SNR of reconstructed speech for proposed frame categorization & basis selection with various bases used for CS recovery.

Reduction factor (s)	SNR (dB)			
	DCT	LPC	Dynamic LPC&DCT	Proposed
0.9	13.44	14.18	13.47	15.58
0.7	10.27	11.33	10.30	11.64
0.5	6.83	6.89	6.84	7.53
0.3	3.45	3.54	3.45	3.67

7.1. Evaluation of proposed sensing matrix

The performance of $\phi_{proposed}$ is compared with the most commonly used Gaussian random matrix $\phi_{Gaussian}$ by using DCT as the sparsifying basis, and the recovery is obtained using greedy OMP algorithm. From the summary depicted in Table 6, the degradation in SNR is higher for the Gaussian random matrix at lower values of s. In addition to that, the proposed sensing matrix also offers the advantage of lower complexity by eliminating the multiplications and reducing the additions due to the sparse nature.

7.2. Validation of frame categorization

We conducted experiments for speech reconstruction using the proposed frame categorization & basis selection and also using DCT basis, LPC basis, dynamic selection of basis from DCT and LPC [12] for the purpose of comparison. Reconstruction quality is evaluated for above mentioned bases by obtaining the SNR as a function of s and the results are summarized in Table 7. There is an augmentation in the quality of reconstructed speech using dynamic basis allocation in comparison to fixed basis and among all allocations, the proposed categorization & basis selection gives an improved quality for the reconstructed speech at the receiver due to more accurate representation of various categories of speech.

7.3. Quantization of parameters

The obtained measurements and LPC coefficients need to be transmitted for every speech frame. Various quantization schemes are compared in this section to validate the performance efficiency of the proposed method.

7.3.1. LPC quantization

Using LBG algorithm, LPC codebooks in the form of LSFs are constructed of length 256, 512 and 1024 to quantize the LPC vectors. Since the LPC vector needs to be transmitted only for the unvoiced frames in the proposed architecture, the impact of quantization is less due to less number of unvoiced frames in a speech signal. The unvoiced frames of NB clean speech signal belonging to NOIZEUS and TIMIT A databases are separated out manually to derive the LPC vectors that serve as the training data for obtaining unvoiced codebooks of various length. From the results summarized in Fig. 10 in terms of collective l_2-norm of quantization error for unvoiced speech samples, it is evident that the new developed codebook gives a better performance and the quality of quantization increases with the length of codebook. It has been found that due to the less number of unvoiced frames, the variation in quality of perceived speech is negligible with length of the unvoiced codebook and hence the bit allocation is fixed to 7 for LPC quantization.

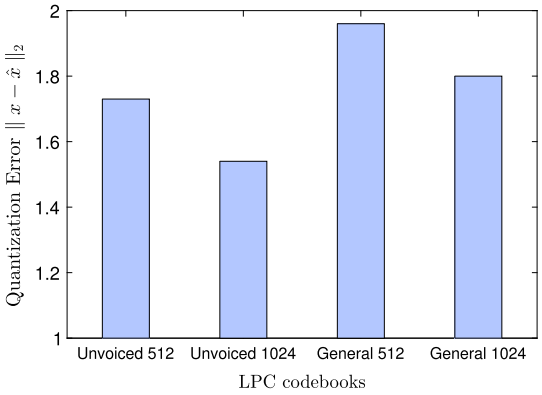

Fig. 10. LPC quantization error variation with length and category of codebook.

Table 8
Performance comparison of measurement quantization using Gaussian and Measurement codebooks of lengths 1024 and 2048 for $s = 0.5$. The values given in bracket for codebook dimension column indicates the number of codebooks.

Codebook dimension	Measurement codebook		Gaussian codebook	
	SNR (dB)	MSE	SNR (dB)	MSE
1024×80	2.75	0.046	0.25	0.082
2048×80	3.20	0.042	0.27	0.082
1024×10 (8)	6.75	0.015	1.23	0.065
2048×10 (8)	7.11	0.012	1.30	0.064
1024×20 (4)	5.63	0.024	0.68	0.074
2048×20 (4)	6.19	0.020	0.72	0.073
1024×40 (2)	3.82	0.036	0.41	0.079
2048×40 (2)	4.34	0.032	0.43	0.078

7.3.2. Measurement quantization

The quantization of measurements is significant as it occupies majority of the transmission bandwidth and plays a crucial role in sparse recovery. Gaussian codebooks are generated using Gaussian iid process and for measurement codebook based quantization, codebooks of dimension 1024 or $2048 \times (Ns)$ are constructed by using the obtained measurements from speech signals of TIMIT A and NOIZEUS databases as training data. The comparison between Gaussian and measurement codebooks is illustrated in Table 8 for $s = 0.5$ by measuring the distortion in terms of MSE and SNR using y and $\hat{y}$, respectively the unquantized and quantized measurements. The quantization performance using measurement codebook is better than the Gaussian codebook for all the cases. The split measurement quantization gives improvement over the other methods but at a higher bit rate due to the increase in the number of indices.

There is an augmentation in SNR with number of mappings and length of the codebook. For various reduction factors, the experiments have been done using LSF codebook of length 256 and measurement codebook of length 1024 to evaluate the performance of the coding scheme and its outcomes are shown in Table 9 which shows an augmentation in perceptual quality with increase in level of mapping.

7.4. Impact of proposed methodologies on codec performance

In addition to the previous experiments that analysed the individual performance of various proposals; (i) Proposed sensing matrix, (ii) Speech categorization based selection of basis and (iii) Orthogonal VQ for measurement quantization, the impact of these contributions on the developed coding scheme is measured and analysed in this section.

Experiments are done by replacing each proposal in the proposed coding scheme with a suitable candidate and the reconstructed quality of speech is measured. The outcomes of the experiments are illustrated in Table 10 for $s = 0.3$, 0.5 and 0.7. As supplementary materials, we have provided 4 speech files each of original and reconstructed using the proposed coding scheme. For each original speech file denoted as *originalX* with $X = 1/2/3/4$, there are 2 reconstructed speech files *reconstructedXs3* and *reconstructedXs5* that correspond to reduction factors 0.3 and 0.5 respectively. In order to evaluate the significance of proposed sensing matrix, the experiments are conducted by replacing it with Gaussian sensing matrix shown as Variant 1 in Table 10. Similarly, experiments are conducted by replacing speech category based basis with DCT basis (Variant 3) and orthogonal mapping with conventional VQ for measurement quantization (Variant 2).

For the proposed coding scheme and its variants, there is an augmentation in reconstructed speech quality with increase in reduction factor (s) from 0.3 to 0.7. The use of proposed binary sensing matrix with conventional VQ yields poor result. This is because, in conventional VQ, mapping is done by using Gaussian distribution for speech samples. But, the usage of proposed binary sensing matrix violates this Gaussian distribution due to the addition of m adjacent samples. The usage of a common DCT basis which is one of the sparsifying basis of speech gave better results but lower than that of proposed coding method. This is due to the usage of better representation for different categories of speech signal. The performance of Gaussian iid is poor due to the large mutual coherence that exists with the sparsifying basis. Thus, the proposed coding scheme gives better performance for various reduction factors.

7.5. Evaluation of the proposed coding methodology

The performance comparison of the proposed coding framework with ACELP implemented and the standardized codecs; AMR, AMR-WB and EVS is given in Table 11. The optimum performance for the proposed coding mechanism is chosen at five levels of mapping with $s = 0.5$ that give rise to a bit rate of 5.275 kbps with 7 bits for LPC coding (assuming a maximum of 50% as unvoiced frames), 100 bits for measurement quantization and 2 bits for categorization of each frame. At 5.275 and 7.275 kbps respectively, the MOS score of the proposed coding scheme is similar to that of 6.7 kbps AMR-NB and 7.2 kbps EVS.

The proposed speech coding scheme allows the flexibility of multi-level scalability by varying both the number of measurements and levels of orthogonal mapping for measurement quantization. It also offers the advantages of inherent de-noising and encryption that are benefited from compressed sensing. Noisy inputs having SNR ranging from 0 to 10 dB are given to proposed coding methodology and the results are summarized in Table 12 for different levels of measurement quantization.

Table 9
Quality of reconstructed speech as a function of reduction factor and levels number of mapping for orthogonal quantization using measurement codebook of length 1024 and LPC codebook of length 256.

Reduction factor	Levels of mapping	Bits per frame	SNR	PESQ	MOS
0.8	2	40	4.41	2.82	2.91
0.8	3	60	5.26	2.90	2.88
0.8	4	80	5.85	2.96	3.02
0.8	5	100	6.29	3.00	3.04
0.7	2	40	4.37	2.80	2.80
0.7	3	60	5.23	2.87	2.92
0.7	4	80	5.82	2.92	2.93
0.7	5	100	6.24	2.95	2.98
0.6	2	40	4.41	2.74	2.78
0.6	3	60	5.23	2.82	2.86
0.6	4	80	5.82	2.86	2.90
0.6	5	100	6.23	2.91	2.98
0.5	2	40	4.45	2.74	2.76
0.5	3	60	5.31	2.82	2.77
0.5	4	80	5.87	2.88	2.84
0.5	5	100	6.30	2.93	2.97
0.4	2	40	4.30	2.70	2.73
0.4	3	60	5.03	2.73	2.76
0.4	4	80	5.52	2.76	2.78
0.4	5	100	5.85	2.77	2.79
0.3	2	40	4.09	2.66	2.60
0.3	3	60	4.80	2.71	2.63
0.3	4	80	5.25	2.74	2.63
0.3	5	100	5.55	2.77	2.67
0.2	2	40	3.18	2.48	2.52
0.2	3	60	3.60	2.50	2.50
0.2	4	80	3.84	2.52	2.56
0.2	5	100	3.97	2.53	2.53

Table 11
Quality of proposed coding scheme at an optimum bit allocation with various standardized codecs.

Codec	Bit rate	MOS
AMR-NB	4.75 kbps	2.30
AMR-NB	6.7 kbps	3.01
AMR-NB	8 kbps	3.10
AMR-WB	6.6 kbps	2.90
EVS	5.9 kbps	3.20
EVS	7.2 kbps	3.30
EVS	8 kbps	3.40
Proposed	5.275 kbps	2.97
Proposed	7.275 kbps	3.17
ACELP	8 kbps	2.87

Table 12
Improvement in SNR of reconstructed speech using proposed coding algorithm for noisy inputs having different types of noise. L indicates the number of mapping levels for measurement quantization.

Noise type	Output SNR (dB)					
	$L = 4$			$L = 5$		
	0 dB	5 dB	10 dB	0 dB	5 dB	10 dB
Airport	3.35	4.98	8.34	2.35	5.05	8.42
Babble	3.33	4.76	7.84	3.40	4.80	7.99
Car	4.32	5.34	8.60	4.68	5.44	8.76
Exhibition	4.30	6.72	9.53	4.70	6.54	9.40
Restaurant	3.30	4.92	7.98	3.28	4.99	7.93
Station	3.48	5.60	8.86	3.48	5.51	8.89
Street	4.45	6.19	8.93	4.32	6.18	8.78

Table 10
The impact of various methodologies used for proposed coding scheme on the reconstructed speech quality.

Coding scheme	Sensing matrix	Measurement quantization	Basis	$s = 0.3$			$s = 0.5$			$s = 0.7$		
				SNR	PESQ	MOS	SNR	PESQ	MOS	SNR	PESQ	MOS
Proposed	Sparse binary proposed matrix	Orthogonal VQ	Proposed hybrid basis	5.55	2.77	2.67	6.30	2.93	2.97	6.24	2.95	2.98
Variant 1	Gaussian iid	Orthogonal VQ	Proposed hybrid basis	1.22	0.98	1.16	2.47	1.82	1.50	3.27	2.01	1.64
Variant 2	Sparse binary proposed matrix	Conventional VQ	Proposed hybrid basis	1.31	1.59	1.37	2.71	1.97	1.61	3.61	2.00	1.63
Variant 3	Sparse binary proposed matrix	Orthogonal VQ	DCT basis	4.87	2.31	1.92	6.07	2.49	2.12	6.10	2.59	2.28

Table 13
Comparison of proposed coding scheme with various compression methods based on compressed sensing. s indicates the reduction factor and L indicates the number of mappings for measurement quantization.

Coding method	Bit rate (kbps)	CR%	SNR (dB)	PESQ	Sensing matrix	Encoder operation	Complexity
Proposed $s=0.5$, $L=4$	4.275	6.68	5.87	2.93	Proposed sparse Binary sensing Matrix (low complex)	1. Sensing frames 2. Measurement quantization	Medium
Proposed $s=0.5$, $L=5$	5.275	8.24	6.30	2.88			Medium
Proposed $s=0.7$, $L=3$	3.275	5.11	5.23	2.87			Low
Proposed $s=0.8$, $L=2$	2.275	3.55	4.41	2.82			Low
Ramdas [14]	8	12.5	6.20	2.73	iid Gaussian Matrix (highly complex)	1. DCT transformation 2. Sensing frames 3. Measurement quantization	High
	12	18.75	6.72	2.79			
Maher [33]	25.6	40	11.83	2.73	Chaotic system based matrix	1. Contourlet transform 2. Sensing	Medium

The results show that without having a subsystem for noise elimination, the output SNR is improved by 2-5 dB which shows its inherent capability for noise reduction at low input SNRs. As the input SNR increases, the amount of noise, which is the non-sparse component reduces. But, at low input SNR, the noise is more and hence noise suppression performance is improved. This is evident from Table 12 which shows an improvement of 3-4 dB for 0 dB input signal.

A comparison of the proposed speech coding with CS based compression schemes given in [14] and [33] in terms of bit rate, Compression Ratio (CR), complexity and reconstruction accuracy has been carried out. Table 13 displays the summary of the results where the proposed coding scheme can operate at low or medium complexity based upon the number of levels of mapping for measurement quantization. In both [14] and [33], the signal transformation prior to sensing measurements increases complexity. In Ramdas et al. [14], the input speech frame is converted into DCT domain, and the measurement are quantized using Analysis by Synthesis approach which further increases the complexity. For $s=0.5$, Ramdas et al. [14] coding scheme can operate at a bit rate of 8 and 12 kbps by allocating 2 and 3 bits respectively for measurement quantization. Even though measurement quantization has been done in Ramdas et al. [14], the low bit rate of our proposed scheme is due to the vector quantization of measurements. In [33], the encoder complexity is due to the application of Contourlet Transform (CT) to the spectrogram of speech signal, and the higher bit rate is due to the direct transmission of measurements. The proposed method gives a similar performance as that of others but at a very low bit rate and at par with the standard codecs.

8. Conclusion

A novel CS based speech coding scheme by considering the signal features has been designed and implemented in this paper. Instead of Gaussian random matrix, a sparse binary sensing matrix is proposed for the measurement sensing that enhanced the performance of the speech coding. By categorizing the frames of speech, a proper basis beneficial for sparse recovery is allocated. The complexity of the proposed coding is reduced by utilizing the measurements for frame categorization and by using a reduced number of bases for sparse recovery. We used conventional VQ for LPC coefficients and an orthogonal mapping based VQ is devised for the measurement vectors without much degradation in the quality of perceived speech. In comparison to other CS based speech coding in literature, the measurements are obtained without doing signal transformation which further reduced the transmitter complexity. The performance of the coding scheme at 5.275 kbps is equivalent to that of 6.7 kbps AMR-NB and at par with 7.2 kbps EVS at a slightly higher bit rate of 7.275 kbps. In addition to all, the CS based coding offers the advantages of inherent encryption and de-noising of 2-4 dB at lower input SNRs. The complexity reduction of sparse recovery will enable the implementation of CS based speech coding for real time applications due to its low bit rate.

Declarations

Author contribution statement

Arun Sankar M.S.: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Sathidevi P.S.: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2019.e01820.

References

[1] J.D. Gibson, Challenges in speech coding research, in: Speech and Audio Processing for Coding, Enhancement and Recognition, Springer, 2015, pp. 19–39.
[2] J.D. Gibson, Speech coding methods, standards, and applications, IEEE Circuits Syst. Mag. 5 (4) (2005) 30–49.
[3] Tarek Mellahi, Rachid Hamdi, LPC-based formant enhancement method in Kalman filtering for speech enhancement, AEU, Int. J. Electron. Commun. 69 (2) (2015).
[4] Zoran N. Milivojevic, Milorad Dj. Mirkovic, Estimation of the fundamental frequency of the speech signal modeled by the SYMPES method, AEU, Int. J. Electron. Commun. 63 (3) (2009).
[5] W.C. Chu, Speech Coding Algorithms: Foundation and Evolution of Standardized Coders, John Wiley & Sons, 2004.
[6] D. Giacobello, M.G. Christensen, M.N. Murthi, S.H. Jensen, M. Moonen, Sparse linear prediction and its applications to speech processing, IEEE Trans. Audio Speech Lang. Process. 20 (5) (2012) 1644–1657.
[7] F. Beritelli, S. Casale, G. Ruggeri, Hybrid multimode/multirate cs-acelp speech coding for adaptive voice over ip, Speech Commun. 38 (3) (2002) 365–381.
[8] M.L. Daniels, B.D. Rao, Compressed sensing based scalable speech coders, in: 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers, ASILOMAR, IEEE, 2012, pp. 92–96.
[9] Z. He, T. Ogawa, M. Haseyama, The simplest measurement matrix for compressed sensing of natural images, in: 2010 17th IEEE International Conference on Image Processing, ICIP, IEEE, 2010, pp. 4301–4304.
[10] A. Ravelomanantsoa, H. Rabah, A. Rouane, Compressed sensing: a simple deterministic measurement matrix and a fast recovery algorithm, IEEE Trans. Instrum. Meas. 64 (12) (2015) 3405–3413.
[11] Siow Yong Low, Duc Son Pham, Svetha Venkatesh, Compressive speech enhancement, Speech Commun. 55 (6) (2013) 757–768.
[12] M.S.A. Sankar, P.S. Sathidevi, Compressive sensing based scalable speech coder using dynamic basis selection and vector quantization, in: IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET, IEEE, 2017, pp. 1–5.

[13] Y. Wang, Z. Xu, G. Li, L. Chang, C. Hong, Compressive sensing framework for speech signal synthesis using a hybrid dictionary, in: 2011 4th International Congress on Image and Signal Processing, CISP, vol. 5, IEEE, 2011, pp. 2400–2403.
[14] V. Ramdas, D. Mishra, S.S. Gorthi, Speech coding and enhancement using quantized compressive sensing measurements, in: 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems, SPICES, Kozhikode, 2015, pp. 1–5.
[15] E.J. Candes, T. Tao, Decoding by linear programming, IEEE Trans. Inf. Theory 51 (12) (2005) 4203–4215.
[16] M.S.A. Sankar, P.S. Sathidevi, Scalable low bit rate celp coder based on compressive sensing and vector quantization, in: 2016 IEEE Annual India Conference, INDICON, IEEE, 2016, pp. 1–5.
[17] Y. Hu, P. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Trans. Speech Audio Process. 16 (1) (2008) 229–238.
[18] John S. Garofolo, et al., TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1, Linguistic Data Consortium, Philadelphia, 1993.
[19] B. Atal, L. Rabiner, A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition, IEEE Trans. Acoust. Speech Signal Process. 24 (3) (1976) 201–212.
[20] Y. Qi, B.R. Hunt, Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier, IEEE Trans. Speech Audio Process. 1 (2) (1993) 250–255.
[21] M. Radmard, M. Hadavi, M.M. Nayebi, A new method of voiced/unvoiced classification based on clustering, J. Signal Inf. Process. 2 (04) (2011) 336.
[22] C. d'Alessandro, G. Richard, Random wavelet representation of unvoiced speech, in: Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, 1992, IEEE, 1992, pp. 41–44.
[23] S. Kadambe, P. Srinivasan, B.A. Telfer, H.H. Szu, Representation and classification of unvoiced sounds using adaptive wavelets, in: Visual Information Processing, 1993, pp. 324–335.
[24] H.H. Szu, B.A. Telfer, S.L. Kadambe, Neural network adaptive wavelets for signal representation and classification, Opt. Eng. 31 (9) (1992) 1907–1916.
[25] S. Kadambe, P. Srinivasan, Application of adaptive wavelets for speech coding, in: Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, 1994, IEEE, 1994, pp. 632–635.
[26] Yan Long, Liu Gang, Guo Jun, Selection of the best wavelet base for speech signal, in: Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004, Hong Kong, China, 2004, pp. 218–221.
[27] R.R. Coifman, M.V. Wickerhauser, Entropy-based algorithms for best basis selection, IEEE Trans. Inf. Theory 38 (2) (1992) 713–718.
[28] V. Meena, G. Abhilash, Sparse representation and recovery of a class of signals using information theoretic measures, in: 2013 Annual IEEE India Conference, INDICON, IEEE, 2013, pp. 1–6.
[29] K. Krishna, V. Murty, K. Ramakrishnan, Vector quantization of excitation gains in speech coding, in: Special Section on Markov Chain Monte Carlo (MCMC) Methods for Signal Processing, Signal Process. 81 (1) (2001) 203–209.
[30] ETSI TS 26.071: 3GPP mandatory speech CODEC speech processing functions; AMR speech codec; general description, 2000.
[31] ETSI TS 26.171: Adaptive multi-rate - wideband (AMR-WB) speech codec; general description, 2001.
[32] ETSI TS 26.445: EVS codec detailed algorithmic description, 2014.
[33] M.K.M. Al-Azawi, A.M. Gaze, Combined speech compression and encryption using chaotic compressive sensing with large key size, IET Signal Process. 12 (2) (2018) 214–218.
[34] A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs, in: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, vol. 2, 2001, pp. 749–752.