International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 3, June 2023, pp. 2788~2800
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i3.pp2788-2800

Compressive speech enhancement using semi-soft thresholding and improved threshold estimation

Smriti Sahu, Neela Rayavarapu
Electronics and Telecommunication Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India

Article Info

Article history:

Received Jun 12, 2022
Revised Jul 19, 2022
Accepted Aug 18, 2022

Keywords:

Compressive sensing
Discrete wavelet transform
Normalized covariance measure
Thresholding

ABSTRACT

Compressive speech enhancement is based on the compressive sensing (CS) sampling theory and utilizes the sparsity of the signal for its enhancement. To improve the performance of the discrete wavelet transform (DWT) basis-function based compressive speech enhancement algorithm, this study presents a semi-soft thresholding approach suggesting improved threshold estimation and threshold rescaling parameters. The semi-soft thresholding approach utilizes two thresholds, one threshold value is an improved universal threshold and the other is calculated based on the initial-silence-region of the signal. This study suggests that thresholding should be applied to both detail coefficients and approximation coefficients to remove noise effectively. The performances of the hard, soft, garrote and semi-soft thresholding approaches are compared based on objective quality and speech intelligibility measures. The normalized covariance measure is introduced as an effective intelligibility measure as it has a strong correlation with the intelligibility of the speech signal. A visual inspection of the output signal is used to verify the results. Experiments were conducted on the noisy speech corpus (NOIZEUS) speech database. The experimental results indicate that the proposed method of semi-soft thresholding using improved threshold estimation provides better enhancement compared to the other thresholding approaches.

Corresponding Author:

Smriti Sahu
Electronics and Telecommunication Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University)
Lavale, Pune, 412115, India
Email: smritisahu13@gmail.com

1. INTRODUCTION

The quality and intelligibility of a speech signal are severely impaired in noisy environments and may increase listener fatigue. In this study, the focus is on enhancing the quality of a speech signal that is corrupted by common additive background noises coming from noisy environments such as streets, airports, and restaurants. Compressive sensing (CS) is a new sampling approach, which enables us to reconstruct the signal using significantly less samples compared to the Nyquist sampling process [1]–[4]. In this way, CS saves memory as well as processing time. The three important aspects of CS are sparsity [5], incoherence [6] and restricted isometry property [7]. Low *et al.* [8] suggests that CS enhances the speech signal by utilizing its sparse nature and the non-sparse nature of the noise, in the time-frequency domain. Thus, compressive speech enhancement allows us to utilize resources optimally [9]. Future communication systems would be required to sense signals rapidly. In this way, CS will be the revolution in the field of next-generation communication systems [10]–[14].

The effectiveness of compressive speech enhancement lies in the correct choice of sensing matrices, the transform domain to ensure signal sparsity and the recovery algorithm [15]–[17]. Gaussian random matrix was selected as the sensing matrix because it satisfies the restricted isometry property of CS, in the most probabilistic sense [18]. Pilastri *et al.* [19] presents the effectiveness of basis pursuit (l_1 minimization) reconstruction algorithm compared to the orthogonal matching pursuit-based algorithms in CS, as applied to images. Thus, l_1 minimization is chosen as the reconstruction algorithm [20], [21]. Our previous work has shown the effectiveness of the discrete wavelet transform (DWT) based basis function for compressive speech enhancement [22]–[24]. This study concluded that the Fejér-Korovkin wavelet, with a vanishing moment N=22 (fk22), was the optimal basis function, for compressive speech enhancement.

This work applies different threshold functions and analyzes the results for compressive speech enhancement using DWT as the basis function. Donoho [25] suggested two threshold functions; hard and soft. The hard threshold function creates signal discontinuity near the threshold points. The soft threshold function solves the signal discontinuity issue, but it causes the constant difference between the input signal and the output signal. To overcome the shortcomings of the hard and soft threshold functions, Gao [26] proposed the garrote threshold function. The denoising effect of the garrote threshold is better, but it does not resolve the constant deviation issue. Thus, the threshold functions should be improved to achieve a better denoising effect. Jing-Yi *et al.* [27] presents the basic principles and structure of wavelet threshold functions and their application for denoising. Jain and Tiwari [28] proposed an adaptive nonlinear mid-threshold function along with a new threshold estimation method for wavelet-based denoising of phonocardiogram signals.

In this paper, a semi-soft thresholding approach is proposed for compressive speech enhancement, which is based on the nonlinear mid-threshold function. In addition, improved threshold estimation and threshold rescaling parameters to achieve a better enhancement have been proposed. The effectiveness of the proposed semi-soft threshold function is compared with the hard, soft and garrote threshold function based on five performance measures: signal-to-noise ratio (SNR), segmental signal-to-noise ratio (SegSNR), root mean square error (RMSE), perceptual evaluation of speech quality (PESQ) and normalized covariance metrics measure (NCM) [29]. The obtained results show that, with the proposed method, better enhancement has been achieved compared to the other three threshold functions, when applied for compressive speech enhancement.

The remaining paper has been organized as: section 2 provides a brief description of the compressive speech enhancement process using the DWT basis function. Section 3 presents the wavelet threshold functions. Section 4 presents the semi-soft thresholding approach and the proposed methods. Section 5 describes the experimental settings, performance evaluation indices, results, and discussions. Section 6 summarizes the conclusions of this study.

2. COMPRESSIVE SPEECH ENHANCEMENT USING DWT BASIS FUNCTION

The compressive sensing approach involves four main steps: sparse representation, measurement of the signal using a sensing matrix, sparse recovery, and reverse sparsity [30]–[33]. The input signal is sparsified using the basis function. The sparse output is then measured into a small set of samples using the sensing matrices. Finally, the signal is reconstructed by reversing the sparsity followed by the sparse recovery algorithms.

Figure 1 presents the step-by-step process of compressive sensing. In the first step, namely 'sparse representation', the signal is projected onto a suitable basis function. During the second step 'measurement', the sparse signal $x \in R^N$ is multiplied with the sensing matrix $\phi \in R^{M \times N}$. Thus, only M number of measurements $y \in R^M$ ($M \ll N$) are taken from the sparse signal x. The set of under-sampled measurements y is called the observation vector. The third step, 'sparse recovery', is a problem of an underdetermined system of linear equations. But sparse recovery is possible as the signal is sufficiently sparse, and the sensing matrix complies with the restricted isometry property (RIP) [7]. The sparsest solution amongst all the possible solutions is to be found using l_1 minimization. In the last step, the clean signal is recovered by reversing the sparsity [34].

Previous work by the authors suggested that the DWT basis function is the optimal basis function for compressive speech enhancement [23]. Orthogonal wavelets support signal denoising and Fejér-Korovkin (fk22) wavelet was chosen as the optimal basis function. Earlier studies in the DWT basis-function based compressive speech enhancement process, suggest that a one-level wavelet decomposition be applied to the signal. Then the detail coefficients are processed using the CS approach and the approximation coefficients are used at the reconstruction stage.

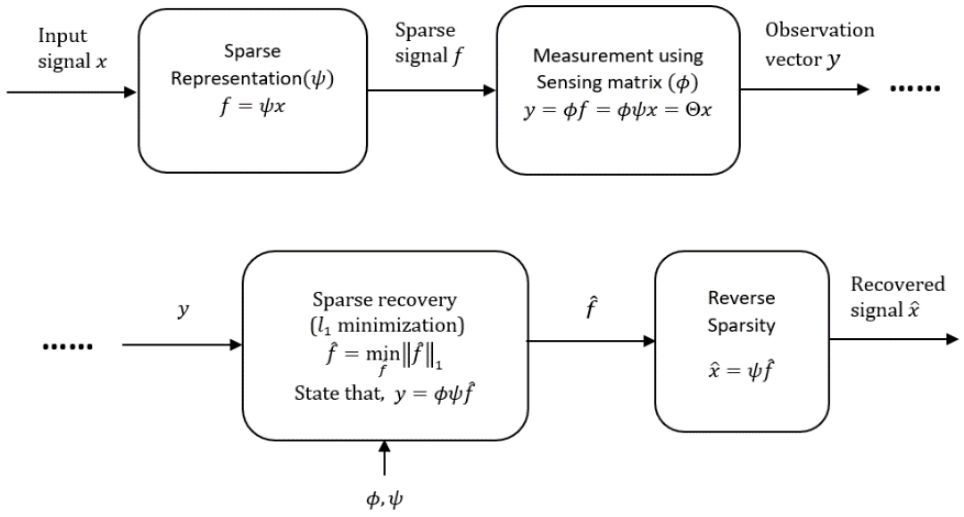

Figure 1. Compressive sensing process

3. WAVELET THRESHOLD FUNCTIONS

Wavelet denoising approaches diminish the noise by thresholding the wavelet decomposition coefficients according to the threshold functions. The two important factors for wavelet denoising are the threshold value and the threshold function, based on which the wavelet decomposition coefficients are shrunk or scaled. Donoho [35] suggested a universal threshold function as given in (1).

$$Th = \sigma\sqrt{2\log N} \tag{1}$$

where σ denotes the noise variance and N is the signal length. This threshold helps to reduce noise to a large extent and preserves the information in the input signal. σ is calculated as shown in (2).

$$\sigma = MAD\ (|x|)/0.6745 \tag{2}$$

where MAD is the median absolute deviation and is calculated as shown in (3):

$$MAD = median(|x - median(x)|) \tag{3}$$

The two most common threshold functions are the ones proposed by Donoho [25] and they are the hard threshold and soft threshold.

Hard threshold: this approach sets the value of the coefficients that are below the threshold value to zero, and the coefficients that are above the threshold value remain the same. The hard threshold function is given by (4).

$$THR_H(x, Th) = \begin{cases} x, & |x| \geq Th \\ 0, & |x| < Th \end{cases} \tag{4}$$

where x is the input signal of length N and Th denotes the threshold value. Hard thresholding does not affect the signal energy significantly, but it causes signal discontinuities.

Soft Threshold: This approach sets the value below the thresholds to zero and performs amplitude subtraction of the coefficients above the threshold [25]. The soft threshold function is given by (5).

$$THR_S(x, Th) = \begin{cases} Sgn(x)(|x| - Th), & |x| \geq Th \\ 0, & |x| < Th \end{cases} \tag{5}$$

where x is the input signal of length N, Th denotes the threshold value and $Sgn(.)$ represents the signum function and it gives the sign of x. The soft thresholding approach attenuates the high-frequency coefficients of the signal, which makes the signal very smooth. Due to this, high-frequency information is lost and there is a constant difference between the input signal and output signal.

Garrote Threshold: The garrote threshold is an intermediate threshold function between the hard and soft threshold functions as it performs hard thresholding of large data values and soft thresholding of small data values using the Garrotte shrinkage function given by (6).

$$THR_G(x,Th) = \begin{cases} (x - Th^2/x), & |x| \geq Th \\ 0, & |x| < Th \end{cases} \quad (6)$$

where x is the input signal of length N and Th denotes the threshold value. Garrote threshold also attenuates the signal and causes constant deviation in the output signal with respect to the input signal [36].

4. SEMI-SOFT THRESHOLDING AND PROPOSED METHODS

4.1. Semi-soft threshold function

In order to resolve the issues caused by the hard, soft and Garrotte thresholds, this work explored the non-linear mid function suggested by Jain and Tiwari [28]. A semi-soft thresholding approach, for compressive speech enhancement using the DWT basis function, is being proposed here. This approach performs thresholding in three parts: i) retaining the high-frequency coefficients, ii) setting small value coefficients to zero, and iii) shrinking the moderate value coefficients according to the nonlinear shrinkage function given by (7).

$$THR_{SS}(x,Th_L,Th_U) = \begin{cases} x, & |x| > Th_U \\ x^3/{Th_U}^2 & Th_L \leq |x| \leq Th_U \\ 0, & |x| < Th_L \end{cases} \quad (7)$$

where, Th_L and Th_U are lower and upper threshold values respectively. The output response for the given threshold functions, for a linear test input, was generated. The threshold values selected were: $Th_L = 4$ and $Th_U = 4.5$.

The performance of the threshold functions can be seen in Figure 2. As evident from Figure 2(a), there is a sharp discontinuity at the threshold point for the hard threshold function. Figure 2(b) shows that there is a constant deviation between the input and the output response in the case of the soft threshold function. Figure 2(c) shows that the garrote threshold tries to solve the discontinuity problem of the hard threshold as well as to decrease the difference between the output and input response. Figure 2(d) displays that the proposed semi-soft threshold function does not show sharp discontinuities and the output response gradually reaches zero at the threshold points. Thus, the semi-soft threshold function overcomes the signal discontinuity issue of the hard threshold function as well as the constant deviation issue of the soft and garrote threshold functions.

4.2. Wavelet decomposition and proposed method of thresholding the coefficients

The wavelet decomposition process decomposes the signal into low-frequency coefficients (approximation coefficients) and high-frequency coefficients (detail coefficients) [37], [38]. In the case of wavelet denoising multilevel wavelet decomposition, wavelet transform is applied to the noisy input, which generates the noisy wavelet coefficients to the level N. The detail coefficients are thresholded for each level from level 1 to N, while the approximation coefficients are used at the wavelet reconstruction stage [39].

Previous work was related to compressive speech enhancement and suggested the use of only one level wavelet decomposition. Thus, the detail, as well as the approximation coefficients, were quite noisy. If thresholding is only applied to detail coefficients, the noise present in the low-frequency region will not be reduced and a noisy output is obtained. Hence, this work suggests the application of the proposed semi-soft thresholding to both the detail and approximation coefficients, for compressive speech enhancement using the DWT basis function.

4.3. Threshold estimation

As thresholding is being applied to both the detail and the approximation coefficients, the effectiveness of the semi-soft thresholding is solely dependent on the selection of the threshold value. A large threshold value results in a noisy output and a low threshold value will not be effective for noise suppression. In this work, we propose two threshold estimation approaches based on the universal threshold suggested by Donoho and Johnstone [40]. The first one is an improved universal threshold Th_1 calculated according to (8).

$$Th_1 = s\sqrt{2\log N} \quad (8)$$

where s represents the standard deviation of the detail coefficients and N represents the length of the detail coefficients. The reason for this is that the standard deviation is useful for describing the variability of the coefficients and it remains on the same scale as the input coefficients. In the case of the universal threshold ($Th = \sigma\sqrt{2\log N}$) suggested in [40], σ is the noise variance of the signal.

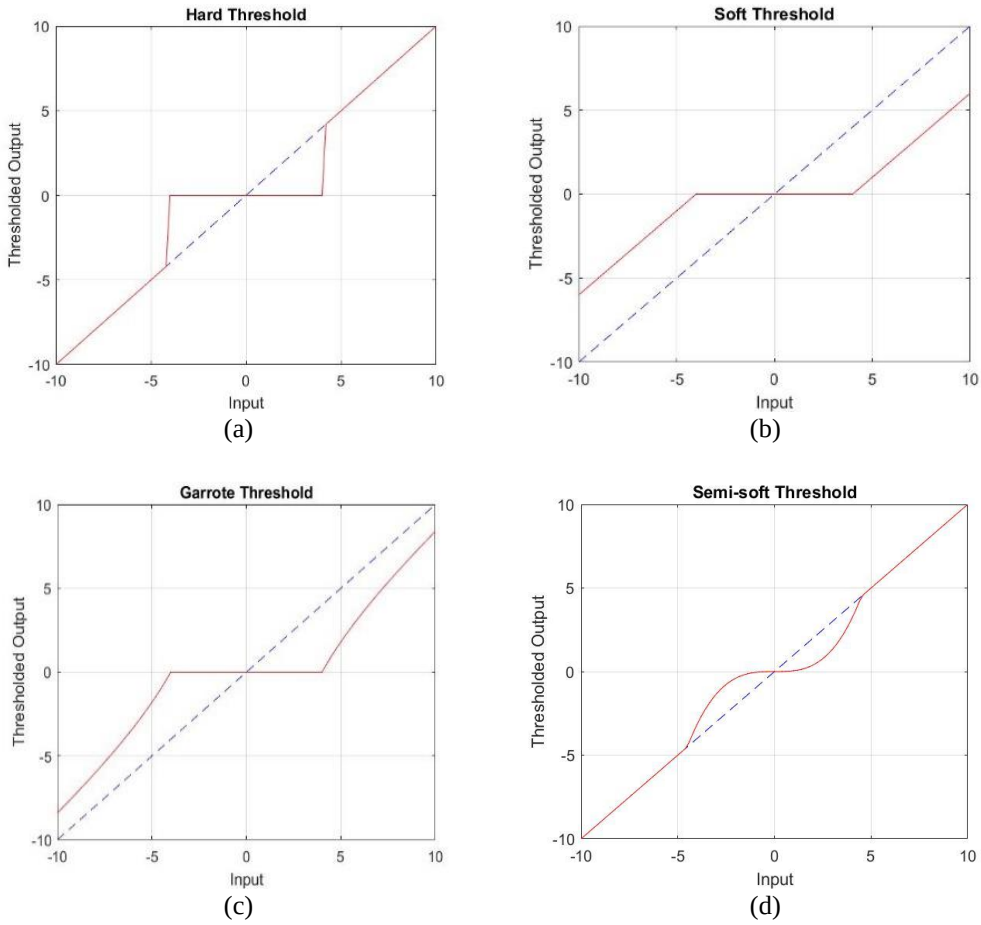

Figure 2. Output responses resulting from (a) hard, (b) soft, (c) garrote, and (d) semi-soft threshold functions, for linear test input

The second threshold is an initial-silence-region based universal threshold Th_2. When the speech signal is recorded or analysed, there is a period of silence before the utterance of a syllable. This region is called the initial-silence region, as the speaker takes some time to speak before uttering the first letter or word. The initial-silence region shows the signal having very little energy, which is equivalent to low noise. Thus, we can estimate the variance of the noise using this region. The initial-silence-region based universal threshold Th_2 is calculated as shown in (9).

$$Th_2 = \sigma_i\sqrt{2\log N_i} \tag{9}$$

where, σ_i and N_i are noise variance and length of the initial-silence-region respectively. Noise variance σ_i is calculated as:

$$\sigma_i = MAD\ (|x_i|)/0.6745 \tag{10}$$

where MAD is the median absolute deviation and is calculated as given in (11):

$$MAD = median(|x_i - median(x_i)|) \tag{11}$$

4.4. Threshold rescaling

Threshold rescaling refers to the optimization of threshold values as they affect the operating range of the threshold functions. It was found that if we applied the threshold directly, there was some noise residue present in the enhanced output signal. Thus, threshold rescaling is an important aspect to be considered for improving the performance of the semi-soft thresholding approach. The rescaled threshold values are given as (12) and (13):

$$T_1 = \alpha \times Th_1 \tag{12}$$

$$T_2 = \beta \times Th_2 \tag{13}$$

α and β are rescaling constants ranging from 1 to 1.5 and $\alpha > \beta$. Based on the performance comparison for different values of α and β, it was found that the value of α should be 1.5 so that the rescaled threshold allows the removal of a large range of noise residue. The optimum value of rescaling constant β was found to be 1.3. By rescaling the threshold values, the operating range of the non-linear mid function is increased. The higher of the two threshold values is considered as the upper threshold Th_{U} and the lower is considered as the lower threshold Th_{L}.

4.5. Method

The experimental steps for the DWT basis-function based compressive speech enhancement are as:

- The noisy speech signal x is taken as the input.
- Framing is essential as we do not process the entire signal in one go when using compressive sensing. The frame size should not be very small as the frames are further decomposed using DWT. The frame size should not be very large also as it may increase processing time. Thus, the input speech signal is divided into non-overlapping frames of 1024 samples each.
- Gaussian random matrix is defined as sensing matrix ϕ.
- One-level DWT ψ is applied to the framed input. We will get the high-frequency coefficients vector cD and low-frequency coefficients vector cA.
- The improved universal threshold Th_1 and initial-silence region-based threshold Th_2 are calculated. After threshold rescaling, the values of Th_{U} and Th_{L} are determined.
- Semi-soft thresholding is implemented on the detail as well as the approximation coefficients using the upper and lower thresholds, $Th_{\mathrm{U}}\ and\ Th_{\mathrm{L}}$. The thresholded coefficients cD_T and cA_T are obtained.
- The observation vector obtained is given in (14).

$$y = \phi * cD_T \tag{14}$$

- The reconstruction algorithm is applied to y to get cD'_T.
- One-level IDWT is applied to the reconstructed high-frequency coefficients vector cD'_T and the thresholded low-frequency coefficients vector cA_T to reverse sparsity.
- The processed frames are merged to form the enhanced speech signal x'.

5. EXPERIMENTAL RESULTS

sp04.wav (male), *sp07.wav* (male), *sp15.wav* (female) and *sp30.wav* (female) were taken from the noisy speech corpus (NOIZEUS) speech database [41]. These signals were distorted by babble, street, and airport noises; at 5 dB, 10 dB and 15 dB input SNR. The sampling frequency was 8 kHz. The input signal which is the original speech signal distorted by one of the above-mentioned noises was framed into non-overlapping frames of 1,024 samples for compressive speech enhancement process. For the initial-silence-region based threshold calculation an initial, 500 samples were taken. The compression ratio was chosen to be 50%. A Gaussian random matrix was selected for the sensing matrix. l_1 minimization was applied as a recovery algorithm, for sparse recovery. Objective quality measures, as well as speech intelligibility measures, were used for the performance assessment of the reconstructed signals.

5.1. Performance evaluation indices

Performance evaluation indices are the parameters which are used to test the effectiveness of the algorithms. Threshold function performance was assessed with the aid of three objective quality measures: SNR, SegSNR, RMSE. Two speech intelligibility measures: PESQ and NCM [29], [42].

SNR is the ratio between the power of a signal and the power of the background noise. SNR is expressed in decibels. Higher SNR represents that signal is more than the noise. It is expressed in (15), where x(n) represents the input signal, while x^' (n) represents the enhanced signal.

$$SNR = 10\,log_{10}\frac{\sum x^2(n)}{\sum(x(n)-x'(n))^2} \quad (15)$$

where $x(n)$ represents the input signal, while $x'(n)$ represents the enhanced signal.

Segmental SNR is the mean of SNRs of all frames of the speech signal. *SegSNR* is calculated using (16).

$$SegSNR = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}(x(n)-x'((n))^2} \quad (16)$$

where $x(n)$ represents the input signal, $x'(n)$ represents the enhanced signal, N gives the length of the frame and M gives the number of frames in the signal.

RMSE is the most commonly used objective measure, which indicates the difference between the clean signal and the enhanced signal. It is calculated using (17).

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n)-x'(n))^2} \quad (17)$$

where *N* is signal length, $x(n)$ represents the input signal and $x'(n)$ represents the enhanced signal.

The ITU-T P.862 standard defines the perceptual evaluation of speech quality (PESQ), which denotes the overall speech quality score. PESQ models the mean opinion score (MOS); it covers a signal quality scale from 1 (bad) to 5 (excellent) [43]–[45]. The NCM measure relies on the covariance between the envelope of the input and output signals [46]. It calculates the correlation coefficient equation between the input and output signal envelopes and maps it to the NCM index. This value of this parameter indicates the intelligibility of the speech signal. A higher value indicates higher intelligibility.

5.2. Results and discussions

In the case of the wavelet threshold denoising approach, the reconstructed signal quality is also related to the number of decomposition levels apart from the selection of threshold value and threshold function. More decomposition levels are required to achieve better denoising, which will increase the computation as well as the processing time. At each decomposition level, thresholding is applied to quantify the detail coefficients. This may lead to the reconstruction error being too large and may reduce signal quality. While in the case of the proposed compressive speech enhancement process, only one level of wavelet decomposition is required, and thresholding is applied to both the detail as well as approximation coefficients, which reduces the complexity of the process as well as provides better enhancement.

In the CS based speech enhancement process, threshold functions were applied to the sparsified signal, obtained using the DWT basis function, Experiments were performed on two male speech signals (*sp04.wav*, *sp07.wav*) and two female speech signals (*sp15.wav*, *sp30.wav*). The performance was observed for three input SNRs 5, 10, and 15 dB, for three background noises (babble, street, and airport). The effectiveness of the proposed algorithm has been presented here for one male speech signal (*sp04.wav*) and one female speech wave (*sp15.wav*) distorted by two noises babble and street.

5.2.1. Performance analysis of the threshold functions for compressive speech enhancement of signal '*sp04.wav*' distorted by babble and street noise

Table 1 shows the comparative assessment of the threshold functions for the CS based enhancement of speech signal *sp04.wav* (male) corrupted by babble noise and street noise. It may be noted from the table that the performance of the proposed method of semi-soft thresholding approach is consistently better than that of the hard, soft and garrote threshold functions in terms of all five performance measures, regardless of the input SNR and noise type.

Figures 3(a)-(f) to 5(a)-(f) show the comparative examination of threshold functions for compressive speech enhancement of signal *sp04.wav* distorted by babble noise and street noise at 5, 10, and 15 dB respectively. Figures 3(a), 4(a) and 5(a) shows the clean input signal *sp04.wav*, while the Figures 3(b), 4(b) and 5(b) shows the noisy speech signal. Figures 3(c), 4(c) and 5(c) shows the compressive speech enhanced signal using the hard threshold function. Similarly, subfigures (d), (e) and (f) shows the compressive speech enhanced signal for soft, garrote and semi-soft threshold functions respectively. By visual inspection of the figures, it is seen that signals were clipped off near the transition regions in the case of the hard threshold. For

the soft and garrote threshold, there is a constant deviation compared to the input signal. This makes the signal very smooth, due to which high-frequency information is lost. In the case of semi-soft thresholding, some noise residue was observed for low input SNR, but results were good in terms of performance parameter values and there was little or no signal clipping. Thus, the signal intelligibility is good and the same is suggested by the values obtained for parameters PESQ and NCM.

Table 1. Comparison of threshold functions for compressive speech enhancement of *sp04.wav* distorted by babble noise and street noise

Threshold Functions	Performance Measures	Babble Noise Input SNR (dB)			Street Noise Input SNR (dB)		
		5	10	15	5	10	15
Hard	SNR (dB)	6.9491	8.1201	9.1320	7.5599	8.1063	9.0085
	SegSNR(dB)	4.1646	3.7549	3.8061	4.8692	3.9199	3.8125
	RMSE	0.0233	0.0186	0.0160	0.0217	0.0186	0.0163
	PESQ	1.6769	0.9745	1.2795	1.2590	1.5589	1.7168
	NCM	0.7414	0.7786	0.8092	0.8249	0.8201	0.8130
Soft	SNR (dB)	3.9215	4.6263	5.0013	4.2819	4.7131	5.0352
	SegSNR(dB)	1.8617	1.7906	1.8104	2.2255	1.8954	1.8463
	RMSE	0.0330	0.0278	0.0258	0.0316	0.0275	0.0257
	PESQ	1.2387	1.0403	1.3169	1.5810	1.6334	1.7556
	NCM	0.6556	0.6889	0.7063	0.7408	0.7239	0.7180
Garrote	SNR (dB)	5.3171	6.3278	6.9272	5.7635	6.3590	6.9219
	SegSNR(dB)	2.7757	2.6424	2.6665	3.2921	2.7703	2.7041
	RMSE	0.0281	0.0229	0.0207	0.0267	0.0228	0.0207
	PESQ	1.3195	1.0564	1.3403	1.6056	1.6528	1.7632
	NCM	0.6771	0.7137	0.7351	0.7719	0.7549	0.7487
Semi-soft	SNR (dB)	8.7336	12.2407	16.4597	8.6220	11.0462	15.4204
	SegSNR(dB)	5.8638	7.3360	9.4517	5.8721	6.3690	8.5403
	RMSE	0.0190	0.0116	0.0069	0.0192	0.0133	0.0078
	PESQ	2.5864	2.8241	3.0337	1.8169	2.5004	2.8296
	NCM	0.8030	0.9036	0.9488	0.8543	0.9012	0.9452

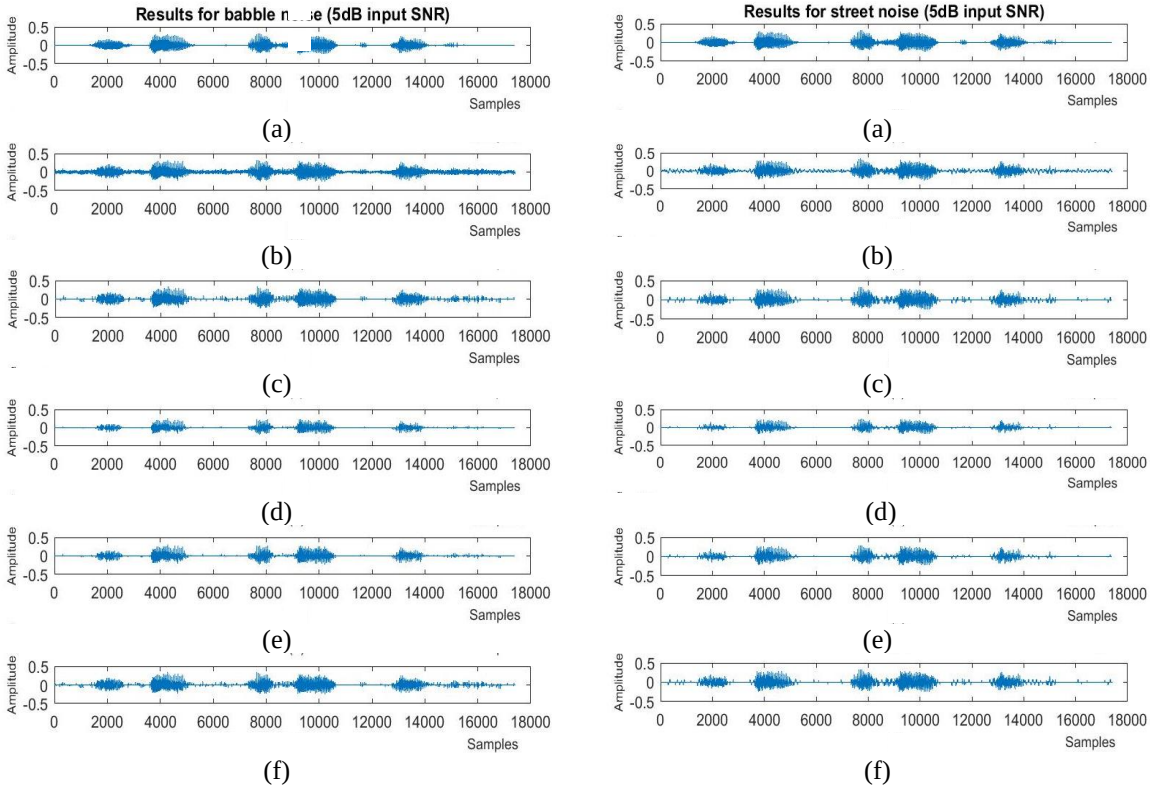

Figure 3. Compressive speech enhancement results using different threshold functions for *sp04.wav* distorted by babble and street noise at 5 dB input SNR (a) clean input signal (*sp04.wav*), (b) noisy signal (5 dB input SNR), (c) hard (d) soft, (e) garrote, and (f) semi-soft threshold function

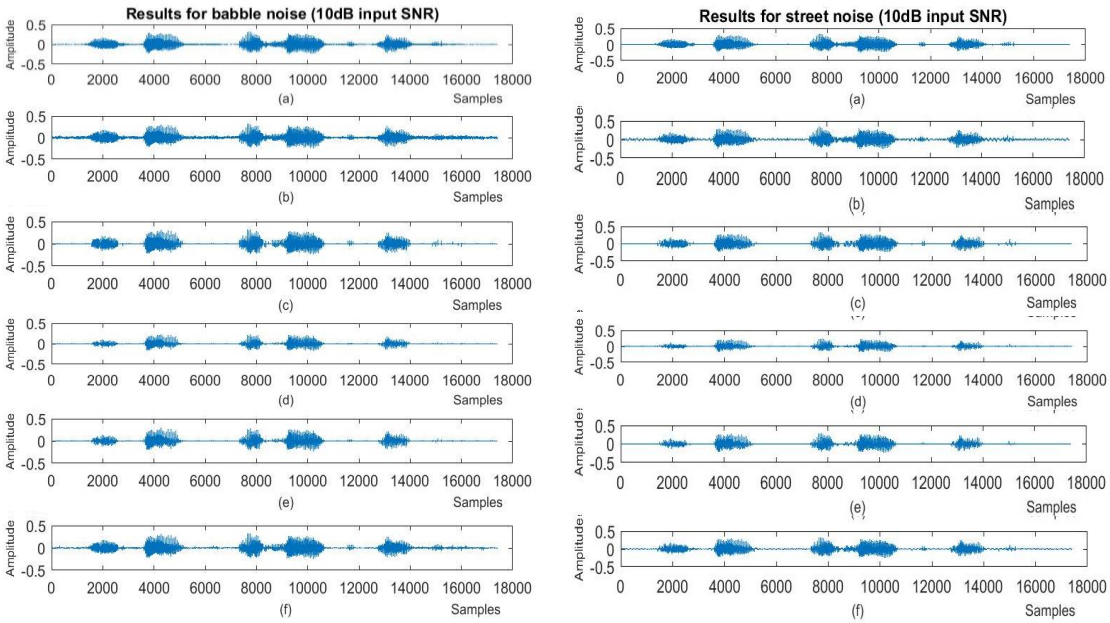

Figure 4. Compressive speech enhancement results using different threshold functions for *sp04.wav* distorted by babble and street noise at 10 dB input SNR (a) clean input signal (*sp04.wav*), (b) noisy signal (10 dB input SNR), (c) hard (d) soft, (e) garrote, and (f) semi-soft threshold function

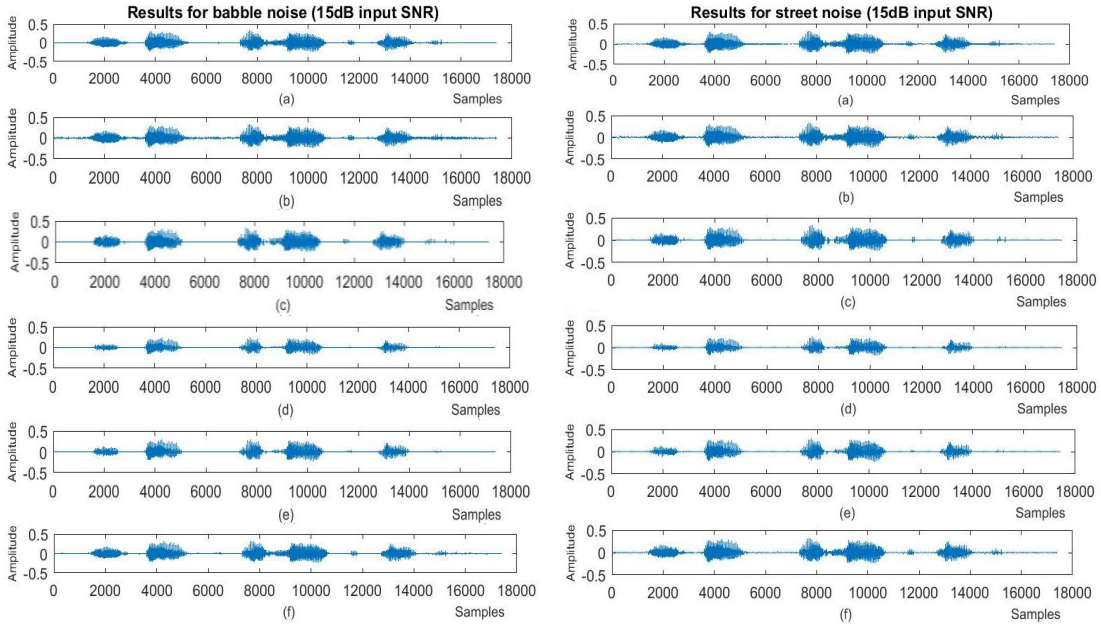

Figure 5. Compressive speech enhancement results using different threshold functions for *sp04.wav* distorted by babble and street noise at 15 dB input SNR (a) clean input signal (*sp04.wav*), (b) noisy signal (15 dB input SNR), (c) hard (d) soft, (e) garrote, and (f) semi-soft threshold function

5.2.2. Performance analysis of threshold functions for compressive speech enhancement of signal '*sp15.wav*' distorted by babble and street noise

Table 2 shows the comparative assessment of the threshold functions for the CS based enhancement of the speech signal, *sp15.wav* (female), corrupted by babble noise and street noise. It can be observed from the table that the proposed method of semi-soft thresholding, using the proposed threshold estimation, shows quantitative effectiveness in terms of the higher values of the SNR, SegSNR, PESQ and NCM parameters and lower MSE value.

Table 2. Comparison of threshold functions for compressive speech enhancement of *sp15.wav* distorted by babble noise and street noise

Threshold Functions	Performance Measures	Babble Noise Input SNR (dB)			Street Noise Input SNR (dB)		
		5	10	15	5	10	15
Hard	SNR (dB)	7.8635	8.8808	9.7073	9.4812	9.3760	9.8832
	SegSNR(dB)	5.6624	5.3251	5.2789	7.1728	5.7113	5.4088
	RMSE	0.0208	0.0169	0.0148	0.0172	0.0159	0.0145
	PESQ	2.2321	1.3001	1.4766	1.7410	1.6008	1.6650
	NCM	0.6825	0.7080	0.7525	0.8082	0.7917	0.7925
Soft	SNR (dB)	4.3485	5.1686	5.6220	5.3405	5.5184	5.7358
	SegSNR(dB)	2.4894	2.4740	2.4992	3.2411	2.6910	2.5653
	RMSE	0.0311	0.0259	0.0238	0.0277	0.0248	0.0234
	PESQ	1.7105	1.2487	1.3984	1.8705	1.6022	1.6415
	NCM	0.5756	0.6121	0.6460	0.6909	0.6844	0.6800
Garrote	SNR (dB)	5.9068	6.9653	7.6281	7.2461	7.4180	7.7768
	SegSNR(dB)	3.7312	3.6766	3.7040	4.8633	3.9898	3.7943
	RMSE	0.0260	0.0210	0.0189	0.0222	0.0199	0.0185
	PESQ	1.8389	1.2812	1.4363	1.9168	1.6350	1.6638
	NCM	0.6071	0.6426	0.6795	0.7292	0.7226	0.7175
Semi-soft	SNR (dB)	9.5422	12.8414	17.1266	10.1959	11.4555	15.5869
	SegSNR(dB)	7.3188	9.1442	12.0443	7.8726	7.6667	10.4260
	RMSE	0.0171	0.0107	0.0063	0.0158	0.0125	0.0075
	PESQ	2.7847	3.0169	3.1815	1.9917	2.2616	2.7882
	NCM	0.7667	0.8579	0.9198	0.8292	0.8690	0.9244

Figures 6(a)-(f) to 8(a)-(f) show the comparative results of threshold functions for compressive speech enhancement of the signal *sp15.wav* distorted due to babble noise and street noise at 5, 10, and 15 dB respectively. Figures 6(a), 7(a) and 8(a) shows the clean input signal *sp15.wav*. Figures 6(b), 7(b) and 8(b) shows the noisy speech signal. Figures 6(c), 7(c) and 8(c) shows the compressive speech enhanced signal using the hard threshold function. Similarly, Subfigures (d), (e) and (f) shows the compressive speech enhanced signal for soft, garrote and semi-soft threshold functions respectively. It can be observed from the figures that the proposed method of semi-soft thresholding, performs effective noise suppression without any significant signal loss. Thus, the proposed approach overcomes the signal discontinuity issue caused by hard thresholding and constant deviation issues caused by soft and garrote thresholding.

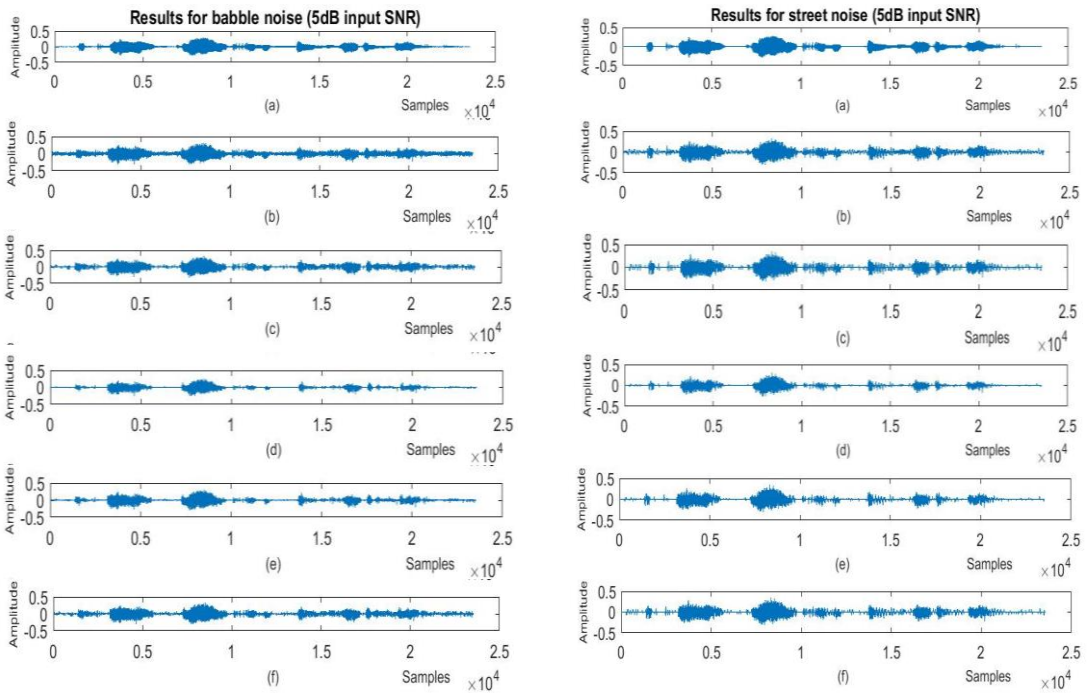

Figure 6. Compressive speech enhancement results using different threshold functions for *sp15.wav* distorted by babble and street noise at 5 dB input SNR (a) clean input signal (*sp15.wav*), (b) noisy signal (5 dB input SNR), (c) hard (d) soft, (e) garrote, and (f) semi-soft threshold function

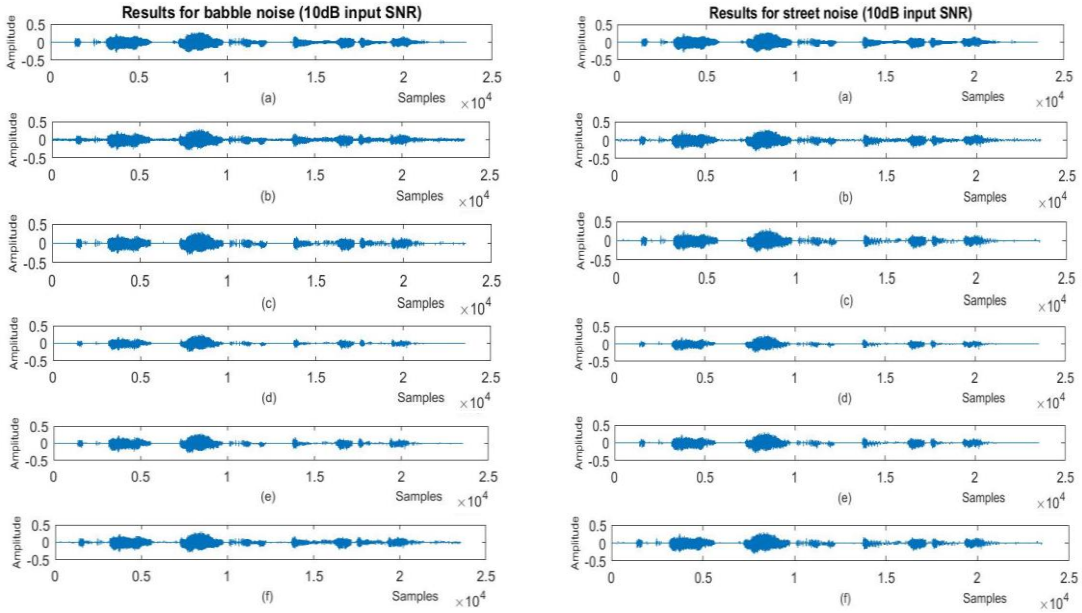

Figure 7. Compressive speech enhancement results using different threshold functions for *sp15.wav* distorted by babble and street noise at 10 dB input SNR (a) clean input signal (*sp15.wav*), (b) noisy signal (10 dB input SNR), (c) hard (d) soft, (e) garrote, and (f) semi-soft threshold function

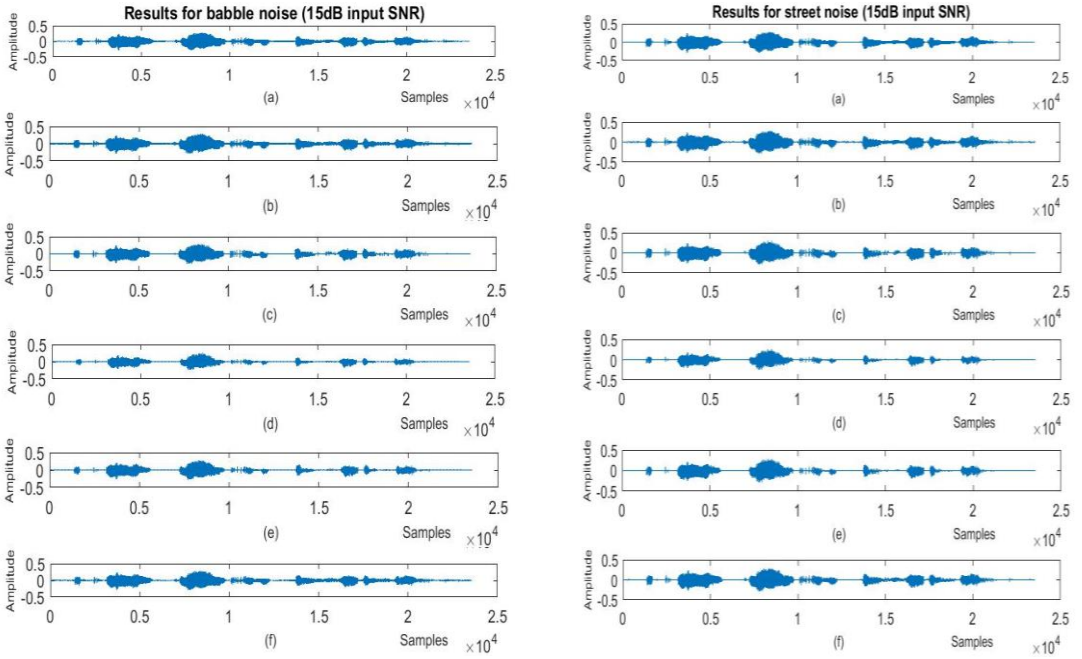

Figure 8. Compressive speech enhancement results using different threshold functions for *sp15.wav* distorted by babble and street noise at 15 dB input SNR (a) clean input signal (*sp15.wav*), (b) noisy signal (15 dB input SNR), (c) hard (d) soft, (e) garrote, and (f) semi-soft threshold function

6. CONCLUSION

To improve the performance of the compressive speech enhancement process, this paper analyzes the conventional threshold functions and introduces semi-soft thresholding with improved threshold estimation. The key contributions of the paper are as: i) this study highlights the problem associated with the traditional wavelet threshold functions and suggests a semi-soft threshold function, which is premised on the nonlinear mid-threshold function and utilizes two threshold values; ii) improved threshold estimation and

threshold rescaling parameters have also been proposed here. One threshold is an improved universal threshold, and the other threshold is estimated using the initial-silence-region of the signal; iii) this study suggests that thresholding should be applied to both the decomposition coefficients, to achieve effective noise reduction, in the case of the one-level wavelet decomposition; and iv) the proposed method of semi-soft thresholding with improved threshold estimation resolves the signal discontinuity issue of the hard threshold function by using two levels of thresholds. It also retains the large coefficients, thus reducing high-frequency information loss and constant signal deviation problems created by soft and garrote thresholding.

Experimental results and the visual analysis show that the proposed method improves signal quality and intelligibility. Thus, it is concluded that the proposed method gives a more accurate threshold estimation to achieve better denoising and is more effective and feasible when compared with the conventional threshold functions. A visual inspection of the enhanced speech signal indicates the presence of noise residue at low SNR, but the parametric analysis suggests that the signal intelligibility is quite good compared to the other thresholding approaches. Future work should be directed towards improving the algorithm to obtain effective noise reduction even at low SNRs.

REFERENCES

[1] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006, doi: 10.1109/TIT.2005.862083.

[2] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006, doi: 10.1109/TIT.2005.860430.

[3] G. Kutyniok, "Theory and applications of compressed sensing," *GAMM-Mitteilungen*, vol. 36, no. 1, pp. 79–101, Aug. 2013, doi: 10.1002/gamm.201310005.

[4] T. Strohmer, "Measure what should be measured: progress and challenges in compressive sensing," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 887–893, Dec. 2012, doi: 10.1109/LSP.2012.2224518.

[5] I. Selesnick, "Introduction to sparsity in signal processing," *A tutorial on sparsity-based methods in signal processing*, 2012. https://eeweb.engineering.nyu.edu/iselesni/teaching/lecture_notes/sparsity_intro/sparse_SP_intro.pdf (accessed Aug. 01, 2021).

[6] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, Jun. 2007, doi: 10.1088/0266-5611/23/3/008.

[7] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9–10, pp. 589–592, May 2008, doi: 10.1016/j.crma.2008.03.014.

[8] S. Y. Low, D. S. Pham, and S. Venkatesh, "Compressive speech enhancement," *Speech Communication*, vol. 55, no. 6, pp. 757–768, Jul. 2013, doi: 10.1016/j.specom.2013.03.003.

[9] H. Haneche, B. Boudraa, and A. Ouahabi, "Speech enhancement using compressed sensing-based method," in *2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, Oct. 2018, pp. 1–6, doi: 10.1109/CISTEM.2018.8613609.

[10] M. M. Abo-Zahhad, A. I. Hussein, and A. M. Mohamed, "Compressive sensing algorithms for signal processing applications: a survey," *International Journal of Communications, Network and System Sciences*, vol. 8, no. 5, pp. 197–216, 2015, doi: 10.4236/ijcns.2015.85021.

[11] M. G. Christensen, J. Ostergaard, and S. H. Jensen, "On compressed sensing and its application to speech and audio signals," in *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, 2009, pp. 356–360, doi: 10.1109/ACSSC.2009.5469828.

[12] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk, "Signal processing with compressive measurements," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 445–460, Apr. 2010, doi: 10.1109/JSTSP.2009.2039178.

[13] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to compressed sensing," in *Compressed Sensing*, Cambridge University Press, 2012, pp. 1–64.

[14] O. Holtz, "Compressive sensing: a paradigm shift in signal processing," *Arxiv.org/abs/0812.3137*, Dec. 2008.

[15] T. Savic and R. Albijanic, "CS reconstruction of the speech and musical signals," in *2015 4th Mediterranean Conference on Embedded Computing (MECO)*, Jun. 2015, pp. 299–302, doi: 10.1109/MECO.2015.7181927.

[16] T. V Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 4125–4128, doi: 10.1109/ICASSP.2009.4960536.

[17] "Compressive sensing resources," *Digital Signal Processing at Rice University*. http://dsp.rice.edu/cs/ (accessed Oct. 10, 2021).

[18] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, Dec. 2008, doi: 10.1007/s00365-007-9003-x.

[19] J. M. R. S. Pilastri, André Luiz and Tavares, "Reconstruction algorithms in compressive sensing: an overview," in *11th edition of the Doctoral Symposium in Informatics Engineering (DSIE-16)*, 2016, pp. 127–137.

[20] E. Candes, E. Candes, J. Romberg, and J. Romberg, "ℓ_1-magic: Recovery of sparse signals via convex programming," pp. 1–19, 2005.

[21] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, Jun. 2006, doi: 10.1002/cpa.20132.

[22] Y. V. Parkale and S. L. Nalbalwar, "Application of 1-D discrete wavelet transform based compressed sensing matrices for speech compression," *SpringerPlus*, vol. 5, no. 1, Dec. 2016, doi: 10.1186/s40064-016-3740-x.

[23] S. Sahu and N. Rayavarapu, "Performance comparison of sparsifying basis functions for compressive speech enhancement," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 769–783, Sep. 2019, doi: 10.1007/s10772-019-09622-9.

[24] S.-F. Xu and X.-B. Chen, "Speech signal acquisition methods based on compressive sensing," in *Systems and Computer Technology*, CRC Press, 2015, pp. 125–130.

[25] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995, doi: 10.1109/18.382009.

[26] H.-Y. Gao, "Wavelet shrinkage denoising using the non-negative garrote," *Journal of Computational and Graphical Statistics*, vol. 7, no. 4, pp. 469–488, Dec. 1998, doi: 10.1080/10618600.1998.10474789.

[27] L. Jing-yi, L. Hong, Y. Dong, and Z. Yan-sheng, "A new wavelet threshold function and denoising application," *Mathematical Problems in Engineering*, vol. 2016, pp. 1–8, 2016, doi: 10.1155/2016/3195492.
[28] P. K. Jain and A. K. Tiwari, "An adaptive thresholding method for the wavelet based denoising of phonocardiogram signal," *Biomedical Signal Processing and Control*, vol. 38, pp. 388–399, Sep. 2017, doi: 10.1016/j.bspc.2017.07.002.
[29] V. Abrol, P. Sharma, and S. Budhiraja, "Evaluating performance of compressed sensing for speech signals," in *2013 3rd IEEE International Advance Computing Conference (IACC)*, Feb. 2013, pp. 1159–1164, doi: 10.1109/IAdCC.2013.6514391.
[30] R. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, Jul. 2007, doi: 10.1109/MSP.2007.4286571.
[31] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, Mar. 2008, doi: 10.1109/MSP.2007.914731.
[32] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006, doi: 10.1109/TIT.2006.871582.
[33] S. Foucart and H. Rauhut, "A mathematical introduction to compressive sensing," *Applied and Numerical Harmonic Analysis*, no. 9780817649470, pp. 1–615, 2013.
[34] F. F. Firouzeh, S. Ghorshi, and S. Salsabili, "Compressed sensing based speech enhancement," in *2014 8th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Dec. 2014, pp. 1–6, doi: 10.1109/ICSPCS.2014.7021068.
[35] D. L. Donoho, "Progress in wavelet analysis and WVD: a ten-minute tour," *Progress in Wavelet Analysis and Applications*, pp. 109–128, 1993.
[36] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, Nov. 1995, doi: 10.2307/1269730.
[37] A. Graps, "An introduction to wavelets," *IEEE Computational Science and Engineering*, vol. 2, no. 2, pp. 50–61, 1995, doi: 10.1109/99.388960.
[38] M. Stéphane, *A wavelet tour of signal processing*. Elsevier, 2009.
[39] M. Misiti, Y. Misiti, G. Oppenheim, and J. M. Poggi, *Wavelet toolboxTM 4 user's guide*. The MathWorks, 2009.
[40] D. Donoho and I. Johnstone, "Ideal spatial adaptation via wavelet shringkage," *Biometrica*, pp. 425–455, 1994.
[41] P. Loizou, "NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms," University of Texas at Dallas. http://ecs.utdallas.edu/loizou/speech/noizeus/ (accessed Sep. 15, 2021).
[42] Z. Lin, L. Zhou, and X. Qiu, "A composite objective measure on subjective evaluation of speech enhancement algorithms," *Applied Acoustics*, vol. 145, pp. 144–148, Feb. 2019, doi: 10.1016/j.apacoust.2018.10.002.
[43] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008, doi: 10.1109/TASL.2007.911054.
[44] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, 2001, vol. 2, pp. 749–752, doi: 10.1109/ICASSP.2001.941023.
[45] P. C. Loizou and Philipos C. Loizou, *Speech enhancement: theory and practice*. Taylor and Francis, 2013.
[46] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009, doi: 10.1121/1.3097493.

BIOGRAPHIES OF AUTHORS

Smriti Sahu received her B.E. degree in Electronics and Telecommunication Engineering, from Chhattisgarh Swami Vivekanand Technical University (C.S.V.T.U.), Bhilai, Chhattisgarh in 2009 and M.E.(Communication) from C.S.V.T.U., Bhilai (C.G.) in 2014. She worked as a lecturer at Shri Shankaracharya College of Engineering and Technology (SSCET), Bhilai for 4 years in the Department of Electronics and Telecommunication Engineering. She received Research Fellowship from Symbiosis International (Deemed University) and pursuing PhD at Symbiosis Institute of Technology, Lavale, Pune. She has a research interest in digital signal processing, image processing and its applications. She can be contacted at email: smritisahu13@gmail.com.

Neela Rayavarapu received her B.E. degree in Electrical Engineering from Bangalore University, Bangalore, India in 1984. She received her MS Degree in Electrical and Computer Engineering from Rutgers, the State University of New Jersey, USA, in 1987, and a PhD degree in Electronics and Communication Engineering in 2012 from Panjab University, Chandigarh. She has been involved in teaching and research in Electrical, Electronics, and Communication Engineering since 1987. She worked as a professor at Symbiosis Institute of Technology, Pune for 9 years. Her areas of interest are digital signal processing and its applications and control systems. She can be contacted at email: neela.raya27@gmail.com.