Iterative Thresholding-Based Spectral Subtraction Algorithm for Speech Enhancement

Raj Kumar, Manoj Tripathy, and R. S. Anand

1 Introduction

Speech enhancement (SE) techniques find many applications such as automatic speech recognition (ASR) systems, speaker recognition, online conferencing, and voice-controlled devices for noise suppression and intelligibility improvement.

Spectral subtraction (SS) algorithm introduced by Boll [1] is still the most preferred speech enhancement algorithm because of simple and reliable design, which makes it the best choice for real-time application, e.g., online conferencing or audio calling. SS algorithm is easily implementable at moderate computing platforms, e.g., DSP processor [2] or FPGA [3], which makes this algorithm a better choice for hearing aids or speech assistive devices. In the SS algorithm, the noise component is subtracted from the spectrum of noisy speech. SS is still an active research area of speech enhancement, used in combination with other techniques like deep recurrent neural network [4], least mean square adaptive filter [5], statistical models [6], deep neural network [7, 8], orthogonal matching pursuit [9], etc.

If clean speech $x(n)$ is corrupted by additive noise $d(n)$, then noisy speech $y(n)$ is given by 1.

$$y(n) = x(n) + d(n) \tag{1}$$

The approximate relationship between spectral power of clean speech, noise, and noisy speech is given by 2.

R. Kumar (✉) · M. Tripathy · R. S. Anand
Electrical Engineering Department, Indian Institute of Technology Roorkee, Roorkee, India
e-mail: rkumar17@ee.iitr.ac.in

M. Tripathy
e-mail: manoj.tripathy@ee.iitr.ac.in

R. S. Anand
e-mail: r.anand@ee.iitr.ac.in

S. Kalya et al. (eds.), *Advances in VLSI, Signal Processing, Power Electronics, IoT, Communication and Embedded Systems*, Lecture Notes in Electrical Engineering 752,
https://doi.org/10.1007/978-981-16-0443-0_18

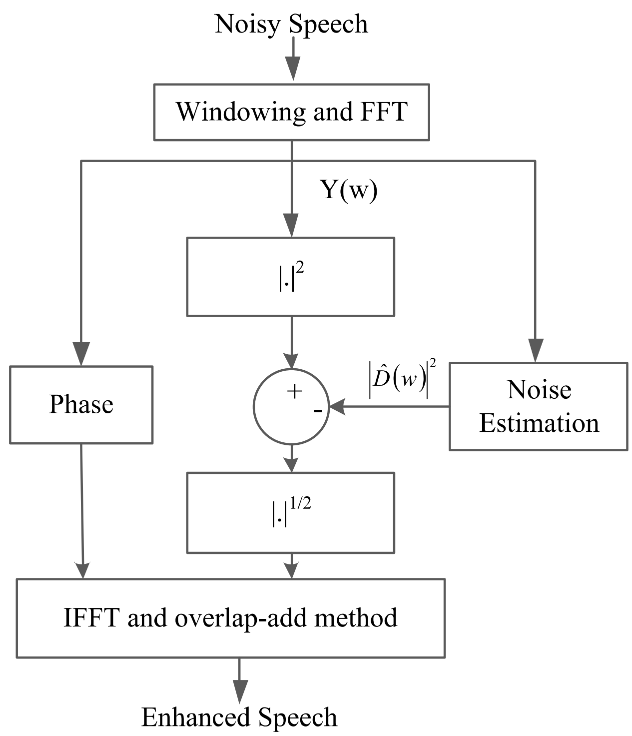

Fig. 1 General spectral subtraction algorithm

$$|Y(w)|^2 = |X(w)|^2 + |D(w)|^2 \tag{2}$$

where $Y(w)$ represents noisy spectrum, $X(w)$ represents clean speech spectrum, and $D(w)$ represents noise spectrum. A general methodology used in the SS algorithm is shown through the block diagram in Fig. 1. The term $\hat{D}(w)$ denotes estimated noise spectrum, which is estimated during silence intervals, i.e., when speech is absent.

In the simplest form, SS can be formulated as shown in 3 to get the estimated clean speech spectrum $\hat{X}(w)$.

$$\left|\hat{X}(w)\right|^2 = |Y(w)|^2 - \alpha\left|\hat{D}(w)\right|^2 \tag{3}$$

Here, $\alpha(\alpha \geq 1)$ ensures subtraction result non-negative value. However, this results in musical noise and removes residual noise in noise only region, which is a must for speech's naturalness. To remove the mentioned problem, Berouti et al. [10] have proposed the parametric spectral power subtraction algorithm, as described in 4.

$$\left|\hat{X}(w)\right|^2 = \begin{cases} |Y(w)|^2 - \alpha\left|\hat{D}(w)\right|^2 & \text{if } |Y(w)|^2 \geq (\alpha+\beta)\left|\hat{D}(w)\right|^2 \\ \beta\left|\hat{D}(w)\right|^2 & \text{else} \end{cases} \tag{4}$$

Here, $\beta(0 < \beta \ll 1)$ avoids isolated peaks in spectrum to suppress musical noise. The noisy phase, $\angle Y(w)$ is used in the final step because phase information is almost unchanged except at very low SNR [11] as shown in 5.

$$\hat{X}(w) = \left|\hat{X}(w)\right| e^{\angle Y(w)} \tag{5}$$

SS performance degrades at low SNR conditions [12]. Several modified SS algorithms have been proposed like multi-band SS [13], reduced delay convolution, adaptive averaging SS [14], and geometric SS [15] to deal with limitations of the SS algorithm.

From Fig. 1, it is clear that the performance of the SS algorithm greatly depends on noise estimation; hence, better the noise estimation, better will be performance [16, 17]. Other enhancement techniques also require prior noise information, e.g., the probability distribution of noise is assumed to be known in Wiener filter [18] and MMSE algorithm [19] or it assumes that noise and speech have independent spectral feature as in the case of subspace approach [20].

Martin and Cohen [21] has introduced the improved minima controlled recursive averaging (IMCRA) algorithm, which performs better than all methods mentioned earlier. It estimates noise even in speech-dominant frames and updates noise power recursively. It uses minima tracking of smoothed periodogram for noise estimation.

The algorithm presented in the paper use speech characterize to use as a clue to locate noise-dominant frame. Recently developed compressive sensing (CS) [22, 23] exploits signal characteristic. Equation 6 represents a compressible signal $x \in R^N$ in sparsifying basis $\psi = [\psi_1, \psi_{2,\ldots}\psi_N]$ while in 7, x_T represent reconstructed speech by keeping T largest coefficients of X where $T \ll N$.

$$x = \sum_{i=1}^{N} X(i)\psi(i) \tag{6}$$

$$x_T = \sum_{i=1}^{T} X_T(i)\psi_T(i) \tag{7}$$

The signal $x(n)$ will be sparse if reconstruction error (RE) as shown in 8 is exactly or nearly zero.

$$\text{R.E.} = \sum_{i=1}^{N} (x(i) - x_T(i))^2 \tag{8}$$

A CS-based signal recovery problem from noisy measurement y is shown in 9 which is referred as l_0 minimization problem.

$$x^{\text{opt}} = \arg\min_{x} \quad \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_0 \tag{9}$$

Here, $\|x\|_0$ represents number of non-negative elements in x. Above problem is basically a search problem to recover sparse vector x^{opt} which become exhaustive if dimension of x is large. The alternate solution method is to solve 9 by l_1 minimization as shown in 10.

$$x^{\text{opt}} = \arg\min_{x} \quad \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1 \tag{10}$$

where l_1-norm is defined as:

$$\|x\|_1 = \sum_i |x(i)| \tag{11}$$

It has been proved that if measurement is sufficient and A satisfy coherence property, then l_1 minimization problem will give same solution as l_0 minimization problem [24]. In the time domain, speech is not a sparse signal, but it shows some sparsity level in some other domain like wavelets, discrete Fourier, or discrete cosine transform domain [25]. The major advantage of CS-based SE is that it relies on signal characteristics rather than noise characteristics; hence, performance does not change whether the noise is stationary or non-stationary, which makes it useful for a real-world scenario. In paper [26], authors have described all available methods to solve the recovery problem defined in 9 and 10.

2 Iterative Soft Thresholding

Iterative thresholding is a technique commonly used to recover the signal from degraded or under-sampled signal [27, 28] based on the fact that the signal is sparse in some sparsifying domain. For solving l_1-regularized least square problem in 10, an algorithm, called split augmented Lagrangian shrinkage algorithm (SALSA) [29] based on the iterative thresholding, has been utilized in the proposed algorithm.

The proposed algorithm for spectral denoising of time domain noisy signal y_{fr} is as follow:

$$\begin{aligned}
&\textbf{Initialize}: y_{\text{fr}}, X = |A(y_{\text{fr}})|, \lambda,\ \mu > 0, d = 0\\
&\textbf{Repeat}\\
&v = \text{soft}(X + d,\ \lambda/\mu) - d\\
&d = 1/\mu A(y_{\text{fr}} - A^T(v))\\
&X = d + v\\
&\textbf{End}
\end{aligned} \tag{12}$$

where A and A^T represent Fourier and inverse Fourier transform, respectively, such that $AA^T = I$, I is identity matrix. The function soft(x, τ) in 12, attenuates input value x above threshold τ while values lower than threshold is made zero as shown in 13.

$$soft(x, \tau) = \begin{cases} x - \tau \ if \ \tau < x \\ 0 \qquad if \ -\tau < x < \tau \\ x + \tau \ if \ x < -\tau \end{cases} \tag{13}$$

While X in 12 represents the spectral feature of a frame after thresholding. If a particular frame contains a speech, then after thresholding, larger magnitude peaks will be obtained. Thus, if $|X|^2 > P_{\min}$, then that frame is a speech-dominant frame; otherwise, it is noise-dominant frame. Where $P_{\min}$ is the minimum power corresponding to residual noise. If a frame is a noise dominant, then noise power is updated recursively using 14.

$$\hat{D}_{\mathrm{fr}} = a\hat{D}_{\mathrm{fr}-1} + (1 - a)|A(y_{\mathrm{fr}})|^2, 0 < a < 1 \tag{14}$$

where a is smoothing coefficient, $\hat{D}_{\mathrm{fr}}$ is estimated noise for current frame and $\hat{D}_{\mathrm{fr}-1}$ is estimated noise previously. In the final step, estimated noise will be subtracted from noisy spectrum using 4.

The proposed algorithm is similar to the minima tracking of periodogram as in IMCRA; instead, it uses thresholding of spectra based on the sparsity of speech.

3 Experiment

A. *Experiment Setup*

In the experiment, NOIZEUS [30] data has been used, having 30 sentences spoken by three male and three female speakers. All speech samples are sampled at 8 kHz. Noises used in the experiment are babble, airport, exhibition, and car environment. Noises have been added to clean speech with SNR from −10 to 10 dB. The Hamming window has been used for windowing with 50% overlapping in all experiments.

B. *Performance Evaluation Measures*

For evaluation purposes, frequency weighted segmental SNR, fwSNR [31] is used to assess the gain in quality. Perceptual evaluation of speech quality (PESQ) [32] has been used to measure the overall quality of the enhanced speech. Short-time objective intelligibility (STOI) [33] measures speech intelligibility by computing the average correlation between the clean and enhanced speech temporal envelope in multiple frames and bands.

Table 1 Effect of smoothing coefficient on fwSNR performance of proposed algorithm in various noise environment

Smoothing coefficient (a)						
Noise type	0.4	0.5	0.6	0.7	0.8	0.9
Exhibition	4.8385	4.8544	4.8473	4.8540	4.8344	4.7881
Babble	4.8125	4.8114	4.8078	4.8062	4.7973	4.8039
Airport	4.7888	4.7944	4.8023	4.8052	4.8019	4.8086
Car	4.8939	4.9099	4.9098	4.9075	4.8901	4.8602

C. *Effect of Smoothing Coefficient and Frame Length*

Table 1 shows the effect of variation of smoothing coefficient, a on fwSNR gain under various noise condition at 0 dB with frame length of 80 ms and $P_{\min} = 5 \times 10 - 2$. The results shown in Table 1 indicate that, in most of the cases, the proposed algorithm performs good when $a\varepsilon[0.5,0.8]$ for the chosen database.

Results in Table 2 show the effect of varying frame size on the speech quality of the enhanced speech using the proposed algorithm for various noises under different SNR levels. The value of smoothing coefficient, a is 0.7 for whole experiment with $P_{\min} = 5 \times 10^{-5}$. The bold numerals represent the highest value in the row. As it is clear, fwSNR of enhanced speech increases as frame size increases from 20 to 80 ms in almost all cases. When frame size increases beyond 80 ms, there is a drop-in intelligibility performance in terms of STOI though quality improves. It happens because a larger frame contains both noise and voice activity components, which cause the removal of voice components from the subsequent frame, resulting drop in intelligibility.

D. *Comparison of Proposed Algorithm*

In this experiment, the proposed method with a frame size of 80 ms (adopted from the previous experiment) is compared with the statistical approach IMCRA [21] algorithms. All parameters in the proposed algorithm have kept constant except λ, which represents weightage given to l_1 regularization. For higher noise, λ is kept high and vice versa.

Figures 2 and 3 show the performance comparison of the proposed algorithm with the IMCRA algorithm in term of fwSNR and PESQ for quality gain. In Fig. 2, fwSNR of enhanced speech using the proposed algorithm is higher under all noisy conditions at SNR level from -10 to 10 dB. In terms of PESQ, the proposed algorithm performs better in all noise conditions except car noise at SNR level more than 0 dB (Fig. 3, bottom-right).

Figure 4 shows the intelligibility performance comparison of the proposed algorithm with the IMCRA algorithm in terms of STOI. It concludes that the STOI of enhanced speech using the proposed algorithm is higher under all noisy conditions at all SNR levels. The intelligibility improved significantly at negative SNR.

Table 2 Effect of frame length on fwSNR performance of proposed algorithm in various noise environment

Noise type	Input SNR (dB)	Frame length		
		20 ms	40 ms	80 ms
Exhibition	−10	2.1037	2.4046	**2.4913**
	−5	2.9003	3.2261	**3.3196**
	0	4.4816	4.6089	**4.7172**
	5	**6.7669**	6.7235	6.6216
	10	8.7186	8.9080	**8.9671**
Babble	−10	2.1121	2.1258	**2.1618**
	−5	2.9921	3.1548	**3.2000**
	0	4.3492	4.6991	**4.8108**
	5	6.5673	6.7860	**6.9251**
	10	8.7706	9.3968	**9.5200**
Airport	−10	1.8858	1.9671	**2.0662**
	−5	2.9102	3.0546	**3.1396**
	0	4.4212	4.6955	**4.7928**
	5	6.3888	6.8418	**6.9828**
	10	8.8742	9.4703	**9.6426**
Car	−10	1.6770	1.9211	**2.0912**
	−5	2.6459	2.9867	**3.1784**
	0	4.1308	4.6079	**4.8078**
	5	6.5275	6.6981	**6.8895**
	10	9.2413	**9.4334**	9.3804

4 Conclusion and Future Work

The proposed algorithm extracts non-speech or noise dominated frames effectively in various non-stationary noises like babble, exhibition, airport, and car noise compared to the statistical approaches. The proposed algorithm's performance does not depend on noise characteristics; hence, performance remains the same whether the noise is stationary or non-stationary.

The proposed method uses a larger smoothing coefficient ($0.5 < a < 0.8$) for noise update, i.e., higher weightage to current frame noise estimate compared to the previous frame; hence, it adapt to a large variation in noise power.

In this paper, the Fourier transform has been used as a sparsifying domain for speech. Various other transforms, e.g., wavelet transform, will be explored in which speech is more sparse than Fourier transforms in short time, thus, showing better performance in shorter frame time. A shorter frame time will reduce delay for real-time applications.

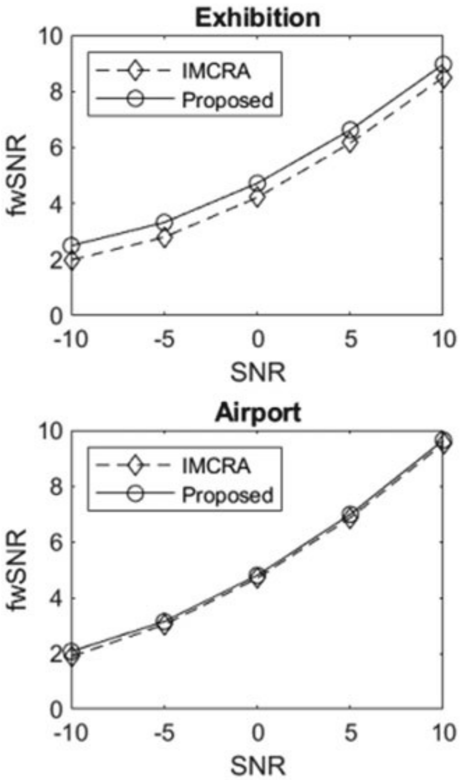

Fig. 2 fwSNR improvement performance

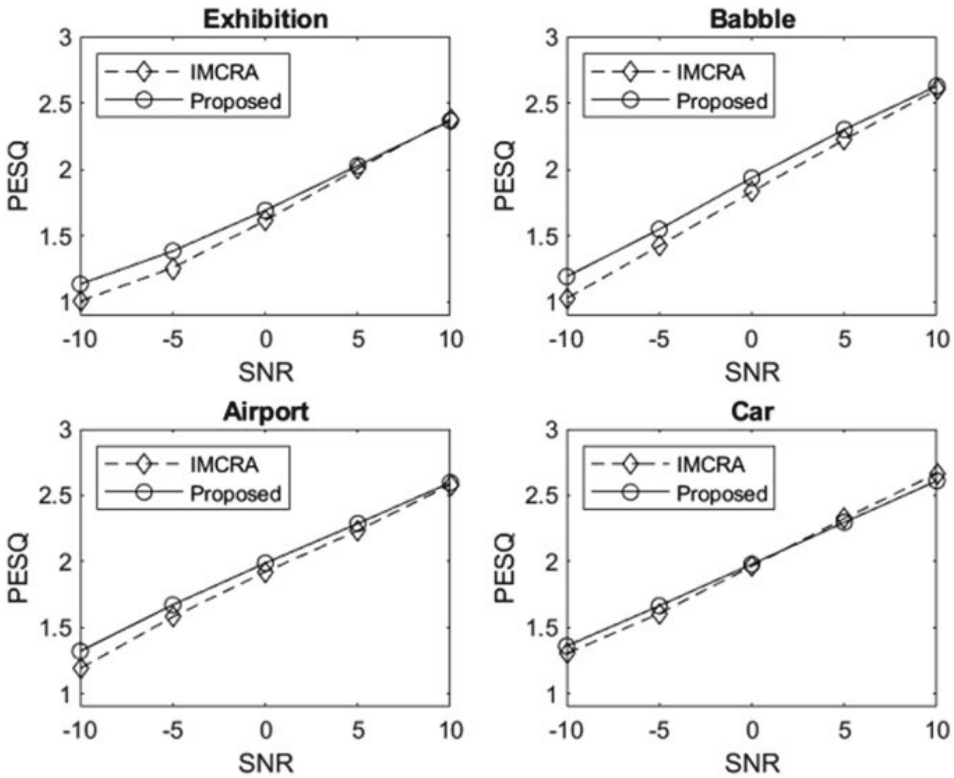

Fig. 3 PESQ improvement performance

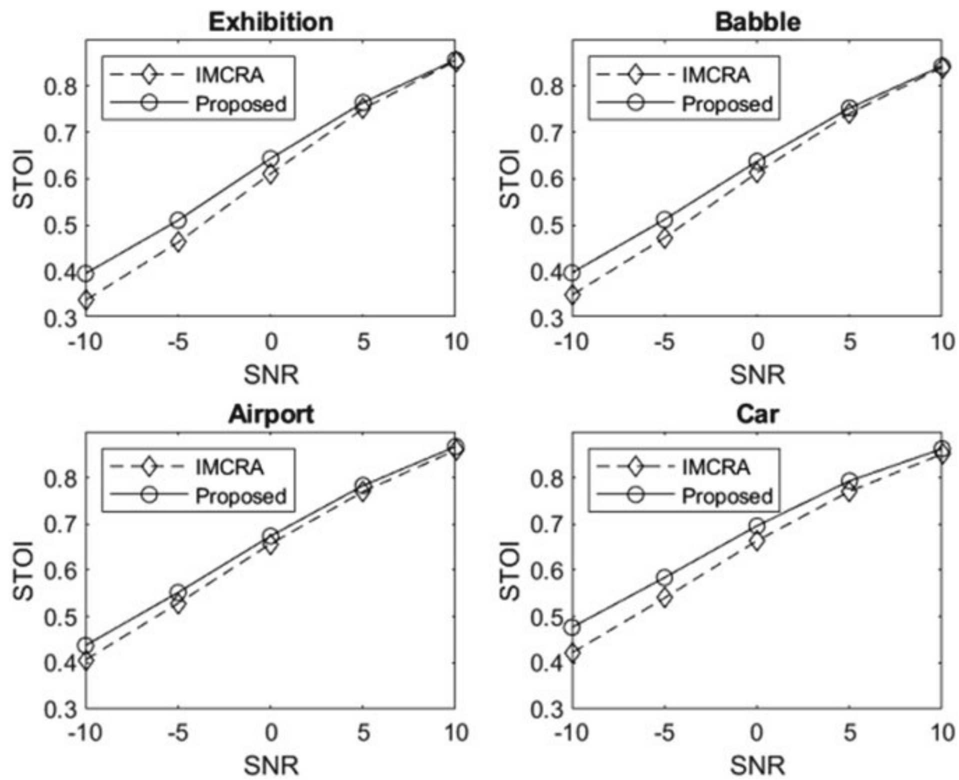

Fig. 4 Intelligibility improvement performance

References

1. S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. **27**(2), 113–120 (1979). https://doi.org/10.1109/TASSP.1979.1163209
2. U. Purushotham, K. Suresh, Implementation of spectral subtraction using sub-band filtering in DSP C6748 processor for enhancing speech signal, in *Advances in Intelligent Systems and Computing* (Springer, Singapore, 2018), pp. 259–267
3. M. Bahoura, FPGA implementation of multi-band spectral subtraction method for speech enhancement, in *Midwest Symposium on Circuits Systems*, vol. 2017-Augus (2017), pp. 1442–1445. https://doi.org/10.1109/mwscas.2017.8053204
4. M. Keshavarzi, T. Goehring, R.E. Turner, B.C.J. Moore, Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: a deep recurrent neural network and spectral subtraction. J. Acoust. Soc. Am. **145**(3), 1493–1503 (2019). https://doi.org/10.1121/1.5094765
5. D. Cao, Z. Chen, X. Gao, Research on noise reduction algorithm based on combination of LMS filter and spectral subtraction. J. Inf. Process. Syst. **15**(4), 748–764 (2019). https://doi.org/10.3745/JIPS.04.0123
6. V.R. Balaji, S. Maheswaran, M. Rajesh Babu, M. Kowsigan, E. Prabhu, K. Venkatachalam, Combining statistical models using modified spectral subtraction method for embedded system. Microprocess. Microsyst. **73**, 102957 (2020). https://doi.org/10.1016/j.micpro.2019.102957
7. T.K. Dash, S.S. Solanki, Speech intelligibility based enhancement system using modified deep neural network and adaptive multiband spectral subtraction. Wirel. Pers. Commun. **111**(2), 1073–1087 (2020). https://doi.org/10.1007/s11277-019-06902-0

8. Q. Zhou, Research on English speech enhancement algorithm based on improved spectral subtraction and deep neural network. Int. J. Innov. Comput. Inf. Control **16**(5), 1711–1723 (2020). https://doi.org/10.24507/ijicic.16.05.1711
9. H. Haneche, B. Boudraa, A. Ouahabi, A new way to enhance speech signal based on compressed sensing. Meas. J. Int. Meas. Confed. **151**, 107117 (2020). https://doi.org/10.1016/j.measurement.2019.107117
10. M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1 (1979), pp. 208–211. https://doi.org/10.1109/icassp.1979.1170788
11. Z. Chen, Y. Liu, G. Wang, S. Wang, W. Geng, Multiband spectral subtraction speech enhancement algorithm with phase spectrum compensation, in *Proceedings of the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference IAEAC*, vol. 20 (2019), pp. 2681–2685. https://doi.org/10.1109/iaeac47372.2019.8997837
12. T.K. Dash, S.S. Solanki, Comparative study of speech enhancement algorithms and their effect on speech intelligibility, in *Proceedings of the 2nd International Conference on Communication and Electronics Systems ICCES,* 2017, vol. 2018-Janua (2018), pp. 270–276. https://doi.org/10.1109/cesys.2017.8321280
13. S. Kamath, P. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in *ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, no. 2 (2002), p. 4164. https://doi.org/10.1109/icassp.2002.5745591
14. H. Gustafsson, S.E. Nordholm, I. Claesson, Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Trans. Speech Audio Process. **9**(8), 799–807 (2001). https://doi.org/10.1109/89.966083
15. Y. Lu, P.C. Loizou, A geometric approach to spectral subtraction. Speech Commun. **50**(6), 453–466 (2008). https://doi.org/10.1016/j.specom.2008.01.003
16. R. Dahlan, D. Krisnandi, A. Ramdan, H.F. Pardede, Unbiased noise estimator for Q-spectral subtraction based speech enhancement, in *Proceedings of the International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications ICRAMET*, no. 2 (2019), pp. 65–68. https://doi.org/10.1109/icramet47453.2019.8980396
17. K. Ozawa, M. Morise, S. Sakamoto, K. Watanabe, Sound source separation by spectral subtraction based on instantaneous estimation of noise spectrum, in *6th International Conference on Systems Informatics, ICSAI*, vol. 1 (2019), pp. 1137–1142. https://doi.org/10.1109/icsai48974.2019.9010477
18. I. Almajai, B. Milner, J. Darch, S. Vaseghi, Visually-derived Wiener filters for speech enhancement, in *ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4 (2007), pp. 2–5. https://doi.org/10.1109/icassp.2007.366980
19. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean-square error short-time spectral amplitude estimator. IEEE Trans. Audio, Speech Lang. Process. **32**(6), 1109–1121 (1984)
20. Y. Ephraim, H.L. Van Trees, A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. **3**(4), 251–266 (1995). https://doi.org/10.1109/89.397090
21. R. Martin, I. Cohen, Single-channel speech presence probability estimation and noise tracking, in *Audio Source Separation and Speech Enhancement* (Wiley, 2018), pp. 97–99
22. D.L. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006). https://doi.org/10.1109/TIT.2006.871582
23. R.G. Baraniuk, E. Candes, M. Elad, Y. Ma, Applications of sparse representation and compressive sensing. Proc. IEEE **98**(6), 906–909 (2010). https://doi.org/10.1109/JPROC.2010.2047424
24. M. Elad, *Sparse and redundant representations, from theory to applications in signal and image processing* (Springer, New York, 2010)
25. D. Wu, W.P. Zhu, M.N.S. Swamy, On sparsity issues in compressive sensing based speech enhancement, in *ISCAS 2012 IEEE International Symposium on Circuits and Systems* (2012), pp. 285–288. https://doi.org/10.1109/iscas.2012.6271907

26. M. Rani, S.B. Dhok, R.B. Deshmukh, A systematic review of compressive sensing: concepts, implementations and applications. IEEE Access **6**, 4875–4894 (2018). https://doi.org/10.1109/ACCESS.2018.2793851
27. I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004). https://doi.org/10.1002/cpa.20042
28. A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009). https://doi.org/10.1137/080716542
29. M.V. Afonso, J.M. Bioucas-Dias, M.A.T. Figueiredo, Fast image recovery using variable splitting and constrained optimization. IEEE Trans. Image Process. **19**(9), 2345–2356 (2010). https://doi.org/10.1109/TIP.2010.2047910
30. Y. Hu, P.C. Loizou, Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. **49**(7–8), 588–601 (2007). https://doi.org/10.1016/j.specom.2006.12.006
31. J.M. Tribolet, P. Noll, B.J. McDermott, R.E. Crochiere, A study of complexity and quality of speech waveform coders, in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (1978), pp. 586–590. https://doi.org/10.1109/icassp.1978.1170567
32. Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU, ITU-T Recomm. (2000), p. 862. [Online]. Available: https://www.itu.int/rec/T-REC-P.862-200102-I/en
33. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, A short-time objective intelligibility measure for time-frequency weighted noisy speech, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2010), pp. 4214–4217. https://doi.org/10.1109/icassp.2010.5495701