

# Adaptive Speech Spectrogram Approximation for Enhancement of Speech Signal

Manju Ramrao Bhosle, Nagesh K N, Ravi Chaurasia

**Abstract:** *The process of speech enhancement tends to decrease the noise with keeping undistorted speech signal amplitude. There are several benefits of speech processing systems which comes with the some challenges. In this paper, we proposed ASSA technique that used to tackle the de-noising and dereverberation in a single channel speech signal. The model is processed using sparse representation prototype in order to perform the de-noising process, where it remove the noise that present in speech signal more thoroughly. Where matrix factorization and SIFT is used to model the speech signal spectrogram, a time-varying filter is used to minimize the noise more effectively. The noise adaptive model is implemented via iterative updating parameters in order to approximate the lower reverberant speech signal in a SIFT domain. Afterwards, the proposed ASSA technique compute the variation in estimated speech signal in order to decrease the noise components and to predict the final speech magnitude. In order to evaluate the performance of proposed system it is compared with state-of-art techniques using some performance metrics.*

**Index Terms:** *Adaptive Speech Spectrogram Approximation (ASSA), Short-Interval Fourier Transform (SIFT), Matrix Factorization, Noise, Dereverberation*

## I. INTRODUCTION

In general, the speech signals are corrupted via the noise, which create challenges for the researchers in order to reduce the noise in speech signal. It is very important and challenging field when it comes to make several communication devices such as mobile phones and hearing aids, where these devices should work reliably in contrary acoustic scenarios (i.e. busy road or crowded mall). The captured speech signals using different microphone have acoustic noise that provide detrimental impact on the speech signal performance. Overcoming these noise effect has been trending topic of research in several recent years as per the increasing requirement for efficient technologies of the speech communication under the challenging scenarios. However, some development has made in both multi and single channel speech processing, the process of single channel enhancement is tends to be the noise robust technique but inappropriate for processing speech signal in many real time speech datasets.

**Revised Manuscript Received on May 21, 2019.**

**Manju Ramrao Bhosle** currently working as assistant professor at Government Engineering college, Raichur and PhD scholar in VTU Belgavi

**Dr. Nagesh K. N** obtained his PhD in wireless communication from JNTUCE, Anantapur, India. Master degree in Digital electronics and communication from Visvesvaraya Technological University India

**Ravi Chaurasia** obtained his Bachelor degree in Electronics & Communication from Visvesvaraya Technological University

Commonly, the techniques of single-channel speech are classified into the following types; nonlinear mapping [1], inverse filtering [2], [3], and probabilistic based procedures [4]–[6] and spectral enhancement [7], [8], [9]. The approaches of non-linear mapping not consider any different prototype for noise reverberation, instead this make use of parallel training data computation to acquire the function of nonlinear mapping using spectrogram of reverberant speech in order to clean speech signal. In [1], a fully end-to-end Deep Neural Network (DNN) is considered, where mean squared error (MSE) of speech signal between noise free speech log-power spectrum and the outcome of DNN is computed to minimize the noise. In addition, the obtained outcome has also improved through taking 1<sup>st</sup> and 2<sup>nd</sup> order interval derivatives of speech input features, these technique of speech enhancement can provide decrement in overall quality of speech [10].

The techniques of inverse filtering are used to reconstruct the original speech signal through developing an inverse filter tends for room impulse response, depending upon this type of observation of clean speech linear prediction residual that fourth order has less residually than reverberant speech. In [2], the estimation of the impulse response is computed by inverse filter using maximization of fourth order moment under linear prediction of inverse filter. In [3], similar approach has been considered in order to maximize normalized third-order moment at linear prediction residual of inverse filter. However, these types of methods reimburse only for the effect of coloration that happened due to early reflections, which required to be used as the conjunction with preceding reverberation conquest techniques to obtain better performance of noise reduction. The scenario when room impulse response is estimated or known, then inverse filtering method can directly apply by using homomorphic methods [11], [12] or using frequency or time domain methods [13]. The noise reduction have major limitation that can provide negative on quality of speech and intelligibility. In [14], a single-channel filter using minimum-variance distortion-less-response is consider in a case of single channel in order to avoid speech distortion.

In [15], a framework is consider of higher-order sub-band filters for speech-distortion weighted using inter-frame Wiener filter (SDW-IFWF), where filters are used to utilize the  $\mu$  parameters that sets a relation in between

speech distortion and noise reduction. In addition, traditional technique has used for the multi-channel applications at SDW multi-channel Wiener filter (SDW-MCWF), also a several type of SDW-IFWFs from the SDW-MCWF principle is derived in order to form a first rank SDW-IFWF. The modeling of acoustic channel is considered in [4] with respect to time dependent linear combination, where the speech signal and all-pole filters codebook is modeled by a block wise time based autoregressive technique. The channel source parameters and probability density function (PDF) is estimated using Bayesian inference, it is applied effectively on data simulation under some range of frequency. Though, the problem comes when the source models and assumed channel are not follow the data successfully. An extension of multi-channel linear probability method has used in [5] using the Bayesian inference in order to optimize spectrograms in the case of single-channel case. In [16], a positive auto-regressive noise model is proposed using data-driven manner where, the approach assumed to be under a noise-free scenario that is impractical in real-time scenarios. In [17], a convolute transfer protocol is proposed for the modeling of speech spectrogram and room impulse response via the non-negative matrix factorization to apprehension the spectral feature of speech signal. Moreover, two approaches has considered to get optimize solution for clean speech signal and room impulse response parameters, which simultaneously computed via the iterative apprise rules.

The FIR (Finite Impulse Response) filter under a composite short-interval Fourier transform (SIFT) domain is considered in [18], where the processing of each sub-band has done independently, also a recursive approach of expectation-maximization is used. Whereas, two steps has performed, first step to get the coefficient of clean speech with using Kalman filter and second step to updates the comprising vector parameters in order to get the variances of noise and speech signal. In this paper, we proposed an approach for enhancement of single channel speech signal, where we majorly concentrated of de-noising and dereverberation performance that executed using our proposed adaptive speech spectrogram approximation (ASSA) technique. The model is processed using sparse representation prototype in order to perform the de-noising process, where remove the noise present in speech signal more thoroughly. The magnitude of SIFT coefficient is used in spectrogram domain, also in each frequency the reverberant signal is computed through convolving the magnitudes of SIFT and make use of difference between the noise and clean speech signal. Initially, the noise adaptive model is implemented via iterative updating parameters in order to approximate the lower reverberant speech signal in a SIFT domain. Afterwards, using the obtained information of noise and speech signal the gain is computed to update the iterative parameters for better clean speech signal. The considered dataset is taken from TIMIT database [19] and proposed model is compared with state-of-art techniques.

## II. PROCEDURE FOR PAPER SUBMISSION

Here, we have proposed an adaptive speech spectrogram approximation approach in order to overcome the de-noising and dereverberation for single-channel speech signal, fig 1 shows the block diagram of proposed scheme. Where matrix factorization and SIFT is used to model the speech signal spectrogram, the de-noising process is considered to minimize the noise more effectively. Initially the clean speech signal with the noise is taken as the input, then it forwarded for matrix factorization in order to learn the patterns associated with signal. Furthermore, the proposed model is implemented through the iterative updating parameters to approximate the lesser reverberant speech signal.

The detailed formulation of our proposed scheme is given below, where  $g(a)$  shows for the clean speech signal and  $d(a)$  denotes for the impulsive response of  $B$  length in SIFT magnitude domain. Therefore, the resulting signal is get through by convolving impulsive response with the speech signal as;

$$f_1(a) = d(a) * g(a) \quad (1)$$

$$f_1(a) = \sum_{b=0}^{B-1} d(b)g(a-b) \quad (2)$$

Where, the convolution is denoted by  $*$  and  $\alpha$  shows for sample index, the  $h$ -th frequency at  $k$  frame in the SIFT domain is approximated by;

$$F_1(h, k) = \sum_{e=0}^{B_d-1} D(h, e)G(H, k - e) \quad (3)$$

$D$ ,  $G$  and  $F_1$  signify the complex value SIFT of  $d$ ,  $g$  and  $f_1$  respectively, the length impulsive response is denoted by  $B_d$  under the SIFT space. While considering the symmetric transformation of SIFT, only half of positive magnitude is considered in  $F_1$  size by taking use of FFT and  $e$  denotes the column matrix.

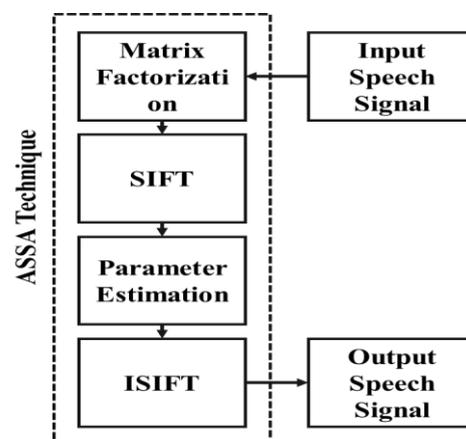


Fig 1: Block diagram of proposed model

Here, we aimed to enhance the speech signal at noisy reverberant scenario, where the speech signal is associated with the additive noises, then the resulting signal can be written as follows;

$$f(a) = \sum_{b=0}^{B-1} d(b)g(a-b) + p(a) \quad (4)$$

Where, noise signal is denoted by  $p(a)$ . The assumed sound is modeled with the speech signal and the noise reverberant parts are also considered as noise. This is considered to get reliable and better estimation of  $D$  because we have used fixed impulsive response at a period of time. The main aim of this study to optimize the speech from SIFT magnitude under the reverberant noisy signal. The speech enhancement process using sparsity model can be represented by the low-rank factorization matrix [20], where the non-negative matrix representation of noise and speech signal are denoted by  $M_p$  and  $M_g$  respectively, which also corresponds to  $N_p$  and  $N_g$  non-negative coefficients.

However, the temporal property of speech is generally used in several speech work such as automatic speech recognition, music transcription and speech enhancement. All of these work are taken into account to model the spectrogram. Moreover, the temporal signal continuity is utilizes, where the span frames  $J$  are considered, so the signal model can be written as;

$$F'_j \approx \sum_{e=0}^{B-1} [D'_j]_{e+1} \ominus M_{gj}^e N_g + M_{pj} N_p \quad (5)$$

Where, the  $j$  denotes for  $J$  consecutive frames that are considered to stack the temporal continuity, so the  $F'_j$  is a stacked form of noise reverberant signal, the frames 0.5 ( $J-1$ ) of SIFT magnitudes initial and after present frames are joined as single frames as stacked frame. Consecutively, the  $F'_j$  frame size is given as;

$$Z_{F'_j} = (0.5H + 1)J \times K \quad (6)$$

The  $M_{pj}$  and  $M_{gj}$  parameters are acknowledged from the stacked SIFT magnitudes, and the computation of  $D'_j$  is done by  $J$  times of the  $D$  in a direction of row. Then scheme a time-varying filter in order to improve the speech signal [21]. Several type process always have some type of limitation, while minimizing the noise to get enhanced speech signal, due to the high correlation between noise and speech and, the complex noise reverberant scenario. Therefore, we provide the de-noising process more thoroughly, where the speech is enhanced from the scenario of noisy reverberant via divided the impulsive response  $d$  to the  $d_1$  and  $d_2$  shorter impulsive responses. From our proposed model the signal model is extended as;

$$f(a) = d_1 * (d_2 * g) + p(a) \quad (7)$$

$$f(a) = \sum_{b_1=0}^{B_1-1} d_1 b_1 s(a-b_1) + p(a) \quad (8)$$

$$s(a-b_1) = \sum_{b_2=0}^{B_2-1} d_2 b_2 g((a-b_1)-b_2) \quad (9)$$

Where,  $B_1$  and  $B_2$  signify the length

of  $d_1$  and  $d_2$  impulsive responses under a time domain, the less reverberant signal is denoted by  $s(a)$ , which is the convolving outcome of clean speech signal at shorter impulsive responses. The  $S_j$  denotes for the less reverberant stacked speech signal in SIFT domain, so our proposed model has able to get enhance speech signal from the noisy environments. Initially, the values of  $N_p$  and  $S_j$  are computed then  $N_g$  is computed afterwards. Moreover, in order to calculate the sparsity properties of speech signals, the sparsity penalties [22] are used in SIFT domain.

The element wise is performed to divide the matrices, the initial stage parameters are computed by taking average of consecutive frames  $J$  at individual update, afterwards its repeated for time  $J$  in a direction of row in order to compute  $\tilde{S}_j$ . Consecutively,  $H_1$  iterations are performed for lesser noise speech magnitude  $\tilde{S}_j$ , so on the  $\tilde{N}_{p1}$  sparse coefficient and  $D_1$  impulsive response is computed. Here, we  $\tilde{S}_j$  is directly can be uses and mask  $\frac{S_j}{F_j}$  is modeled to compute the lesser noise speech signal output at the end of initial processing and  $C$  represent the approximate signal error. So whatever noise has left in initial processing, we will going to discard in second stage process, where we make use of mask to compute  $\tilde{S}_j$  lesser noisy signal. Therefore, the both noise reduction and dereverberation are used in this stage of process, so the considered cost function can be written as;

$$Y_2 = C(\hat{S}_j, \sum_{d_2=0}^{B_2-1} [D'_2 j]_{i+1} \ominus M_{gj}^e N_g) + L[F_n] \quad (10)$$

$$L[F_n] = L_{p2} \left\| N_{p2} \right\|_1 + L_g \left\| N_g \right\|_1 \quad (11)$$

Where,  $L_g$  and  $L_{p2}$  denotes for the trade-off parameters in between the error approximation, also the sparsity controls tends to have the different sparse coefficients such as;  $N_{p2}$  for the noisy signal and  $N_g$  for the speech signal. A predefined  $H_2$  iterations are done to estimate the  $\tilde{N}_g$  and  $\tilde{N}_{p2}$  sparse coefficient, and  $\tilde{D}'_2$  impulsive response.

$$\tilde{F}'_j \approx \sum_{h=0}^{B_1-1} [\tilde{D}'_1 j]_{h+1} \ominus \tilde{S}_j + M_{pj} N_{p1} \quad (12)$$

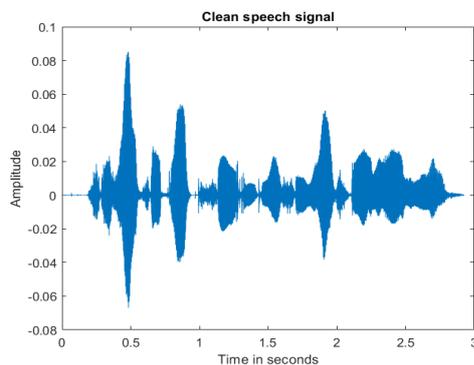
Therefore, the estimated final speech magnitude is  $\tilde{G}_1$ , which computed by taking convolution of unstacked mask and  $\tilde{F}'_j$ , the operation of unstacked is done by taking inverse transformation of the approximated stacking. Finally, the proposed ASSA technique computed the variation in estimated speech signal in order to decrease the noise components and enhancing the speech components from single-channel mixture to

predict the final speech magnitude.

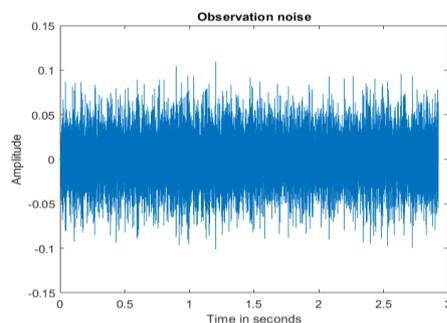
### III. RESULTS AND ANALYSIS

In this section, we are going to evaluate the proposed adaptive speech spectrogram approximation technique with some state-of-art techniques, where the received signals are acquired by convolving the considered speech signals with adding several types of different reverberant (that is signal to noise ratio (SNR)) and impulsive noise responses. Here, the considered dataset is taken from TIMIT database [19] and it have sample frequency of 16kHz, where the total length of the dataset is 3 sec and the speech signal has eleven number of words.

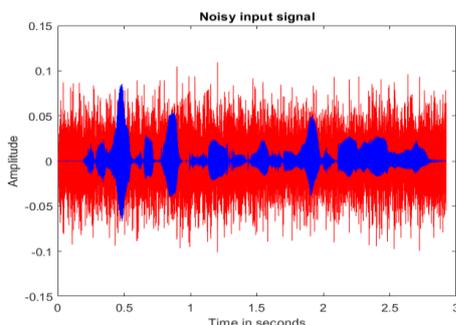
#### Scenario-A



**Fig 2:** The considered Clean Speech Signal



**Fig 3:** The Observed Noise Signal

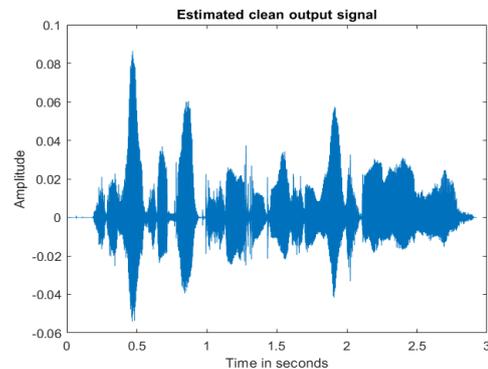


**Fig 4:** Noise Input Speech Signal

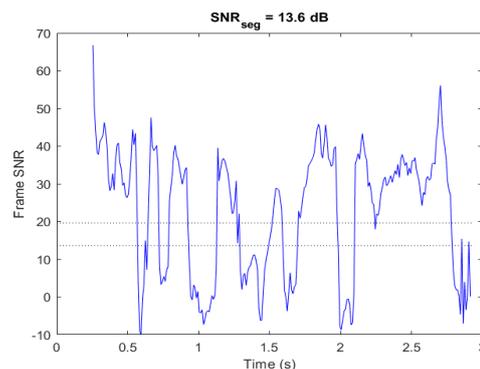
The Informed Wiener (InfW) [23] filtering approach is considered for the comparison purpose, which incorporates the power spectral density (PSD) of latent speech by minimum mean-square error (MMSE) approximation framework at complex speech spectral magnitudes. Therefore, it includes joint distribution of the speech PSD

and complex speech magnitude, which are conditioned on the noisy environment and it resolved by Bayesian filter and speech PSD priors. Similarly some other approach has also been consider such as P Bayes [23] histogram technique and SS Wiener [23] in order to evaluate the performance of our proposed model. Here, we rely on several performance metrics in order to enhancement of the noise speech signal, the segmented SNR (segSNR) [24] due its relationship with mean square error and the score of short-time objective intelligibility (STOI) [25]. The evaluation of results under some speech shaped noise is done in scenario A and using the white Gaussian noise is done in scenario B that has used in existing system.

Figure 2 shows the considered clean speech signal that is taken from TIMIT database [19], figure 3 shows the observed noise signal where the input SNR is considered of -10 dB and figure 4 shows the noise input speech signal. Finally the estimated clean speech signal is given in figure 5, and in order to evaluate the estimated clean speech signal we have computed segmented SNR that shows in figure 6, where the average segmented SNR value is 13.6dB.



**Fig 5:** Estimated Clean Speech Signal

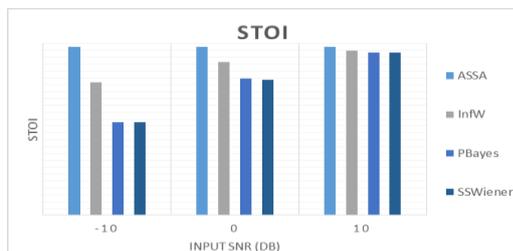


**Fig 6:** Segmented SNR (dB) as per the Sampling Time (s)

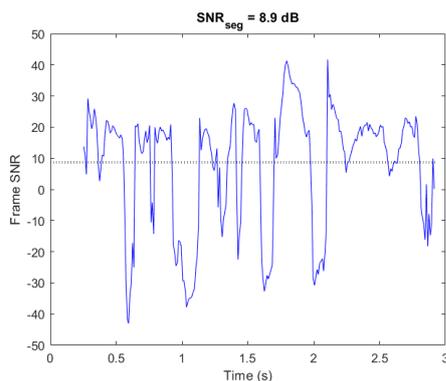
**Scenario-B**

**Table 1:** STOI values using several Approaches with White Noise

Input SNR (dB)	ASSA	InfW	PBayes	SSWiener
-10	0.9794	0.774	0.5431	0.5411
0	0.9793	0.891	0.7982	0.7901
10	0.9796	0.958	0.9467	0.9462



**Figure 7:** STOI Score Comparison



**Figure 8:** Segmented SNR (dB) as per the Sampling Time (s) under White Noise

In scenario-B, additive white Gaussian noise is considered, where table 1 shows the STOI score using several approaches and the graphical comparison is given in figure 7. In -10dB input SNR, the proposed ASSA approach has performed 20.97% as compared to InfW [23], in 0dB and 10dB SNR it performed 9% and 2% better STOI score than InfW [23]. Figure 8 shows for segmented SNR (dB) as per the sampling time (s) under white noise.

**IV. CONCLUSION**

In this paper, an ASSA method is proposed that based upon the sparse representation for the speech de-noising and dereverberation, where it make use of decomposition as the advantage. A long impulsive response is divided into the short impulsive responses in order to provide the speech de-noising and dereverberation more efficiently and save the time. Moreover, the temporal dynamics of speech property is considered through the temporal stacking process to obtain enhanced spectrogram signal model. The result analysis is

done using segmented SNR and STOI score under two different noise scenarios, where our proposed model shows significant improvement in STOI score while comparing with state-of-art techniques. Therefore, our proposed model scheme has obtained enhanced speech de-noising and dereverberation performance. In future work, some other type of room noise or outside noise can be consider, also some other new algorithms can be model to provide better recognition of speech structures.

**REFERENCES**

1. X. Xiao et al., "The NTU-ADSC systems for reverberation challenge 2014," in Proc. REVERB Challenge Workshop, 2014, pp. o2.2:1-8.
2. M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," IEEE Trans. Audio, Speech, Language Process., vol. 14, no. 3, pp. 774-784, May 2006.
3. S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, "Single-microphone LP residual skewness-based inverse filtering of the room impulse response," IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 5, pp. 1617-1632, Jul. 2012.
4. C. Evers and J. R. Hopgood, "Parametric modelling for single-channel blind dereverberation of speech from a moving speaker," IET Signal Process., vol. 2, no. 2, pp. 59-74, Jun. 2008.
5. A. Maezawa, K. Itoyama, K. Yoshii, and H. G. Okuno, "Nonparametric Bayesian dereverberation of power spectrograms based on infinite-order autoregressive processes," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 22, no. 12, pp. 1918-1930, Dec. 2014.
6. N. Mohammadiha and S. Doclo, "Speech dereverberation using nonnegative convolutive transfer function and spectro-temporal modeling," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 2, pp. 276-289, Feb. 2016.
7. E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Univ. Eindhoven, Eindhoven, the Netherlands, 2007.
8. K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," Acta Acoust., vol. 87, pp. 359-366, 2001.
9. B. Cauchi et al., "Combination of MVDR beamforming and singlechannel spectral processing for enhancing noisy and reverberant speech," EURASIP J. Adv. Signal Process., vol. 61, 2015, pp. 1-12.
10. K. Kinoshita et al., "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," EURASIP J. Adv. Signal Process., vol. 7, pp. 1-19, 2016.
11. J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 7, May 1982, pp. 1858-1861.
12. B.D.RadlovicandR.A.Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," IEEE Trans. Speech Audio Process., vol. 8, no. 6, pp. 728-737, Nov. 2000.
13. I.Kodrasi, T.Gerkmann, and S.Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2014, pp. 5177-5181.
14. J. Benesty and Y.Huang, "A single-channel noise reduction-MVDR-filter," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., May 2011, pp. 273-276.
15. K. T. Andersen and M. Moonen, "Robust Speech-Distortion Weighted Interframe Wiener Filters for Single-Channel Noise Reduction," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 97-107, Jan. 2018.
16. T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," Ann. Statist., vol. 1, no. 2, pp. 209-230, Mar. 1973.
17. H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparseness of speech spectrograms," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2009, pp. 45-48.
18. B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm,"



## Adaptive Speech Spectrogram Approximation for Enhancement of Speech Signal

- IEEE/ACM Trans. Audio, Speech, Language Process., vol. 23, no. 2, pp. 394–406, Feb. 2015.
19. J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic speech corpus LDC93S1," Web Download. Philadelphia: Linguistic Data Consortium. 1993.
  20. Mohammadiha, N., Smaragdis, P., Doclo, S., Joint acoustic and spectral modeling for speech dereverberation using non-negative representations. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP), pp. 4410–4414. 2015.
  21. Baby, D., Supervised speech dereverberation in noisy environments using exemplar-based sparse representations. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP), pp. 156–160. 2016.
  22. Zhang, L., Bao, G., Zhang, J., et al., Supervised single-channel speech enhancement using ratio mask with joint dictionary learning. Speech Commun. 82, 38–52. 2016.
  23. G. Enzner and P. Thüine, "Bayesian MMSE Filtering of Noisy Speech by SNR Marginalization With Global PSD Priors," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 12, pp. 2289–2304, Dec. 2018.
  24. S. Quackenbush, T. Barnwell, and M. Clements, Objectives Measures of Speech Quality. Englewood Cliffs, New Jersey: Prentice-Hall, 1988.
  25. C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 19, no. 7, pp. 2125–2136, 2011.

### AUTHORS PROFILE



**Manju Ramrao Bhosle** currently working as assistant professor at Government Engineering college, Raichur and PhD scholar in VTU Belgavi, Karnataka. Master degree in Digital electronics and communication from Visvesvaraya Technological University India, and Bachelor degree in Electronics & Communication from VTU India. Membership of ISTE and his areas of research interest are Speech

Processing.



**Dr. Nagesh K. N** obtained his PhD in wireless communication from JNTUCE, Anantapur, India. Master degree in Digital electronics and communication from Visvesvaraya Technological University India, and Bachelor degree in Electronics & Communication from Visvesvaraya Technological University India. His teaching interests are Communication systems, Digital communication. His areas of research interest are, Wireless communication.



**Ravi Chaurasia** obtained his Bachelor degree in Electronics & Communication from Visvesvaraya Technological University India. His areas of research interest are Digital signal processing, Wireless Communication and Image processing.