



# Low SNR speech enhancement with DNN based phase estimation

Samba Raju Chiluveru<sup>1</sup> · Manoj Tripathy<sup>1</sup>

Received: 7 December 2018 / Accepted: 12 February 2019 / Published online: 23 February 2019  
© Springer Science+Business Media, LLC, part of Springer Nature 2019

## Abstract

In low Signal-to-Noise Ratio environment phase information is one of the important factor and therefore this article consider the importance of clean phase in single channel speech enhancement technique. The proposed method uses Deep Neural Network based regression model to estimate clean phase and clean amplitude for speech reconstruction. Experiments are conducted over five different noises such as factory, restaurant, car, airport and babble at different levels and result are evaluated using objective quality measures like Perceptual Evaluation of Speech Quality, Weighted Spectral Slope, Cepstrum Distance, frequency weighted segmented Signal-to-Noise Ratio and Log Likelihood Ratio. The overall quality of speech improved for factory noise by 12%, restaurant noise by 8%, car noise by 13%, airport noise by 10% and babble noise by 14% respectively.

**Keywords** Speech enhancement · Deep neural network · Regression

## 1 Introduction

Single channel speech enhancement is active research topic since last few decades because of rapidly increasing real world applications such as mobile speech communication, robust speech recognition (Loizou 2013), and speaker verification (Reynolds et al. 2000) etc. All applications require clean speech signal for processing and generally speech enhancement technique is used for estimation of clean speech signal from noisy speech signal. Speech enhancement is performed either in time domain or in frequency domain with the aim of increasing quality and intelligibility of speech in noisy environment by suppressing noise signal from noisy speech signal (Benesty et al. 2005). There are various methods developed for speech enhancement over the last few decades such as spectral subtraction, Wiener method, and gain based method like MMSE-STSA, log-MMSE (Kamath and Loizou 2002; Ephraim and Malah 1984, 1985; Scalart et al. 1996; Cohen 2003; Surendran and Kumar 2015; Bouzid et al. 2016) etc. The spectral subtraction method has been proposed by Berouti et al. (1979) and

became popular due to its ability of reducing background (additive) noise but limitation is generation of annoying musical noise artifact (Scalart et al. 1996). Ephraim and Malah (1984) addressed musical noise problem using statistical model-based method and its updates are reported in Ephraim and Malah (1985) and Cohen (2003) respectively.

Another group of researchers were using recently developed algorithms based on Wavelet Packet Decomposition and DNN for speech enhancement (Bouzid et al. 2016; Dahl et al. 2012). The aforementioned algorithms works well at high Signal-to-Noise Ratio (SNR) like higher than 6 dB ( $\text{SNR} \geq 6$  dB) where as in low SNR i.e.  $\text{SNR} \leq 6$  dB environment quality and intelligibility of noisy speech signal even deteriorate further from original clean speech hence speech enhancement in low SNR environment becomes difficult (Loizou and Kim 2011).

Xu et al. (2015) proposed DNN based regression model for speech enhancement and to generalize the regression model large training data set at different SNR levels were used. The recommended regression model in Xu et al. (2015) is only applicable for high SNR signals. It claim improvement in musical noise by comparing minimum mean square error (MMSE) and the conventional speech enhancement techniques but it uses noisy phase for reconstruction of speech along with estimated amplitude of noisy speech signal hence the amplitude of estimated speech signal will degrade and it will continue as SNR value decreases (Xu et al. 2015).

✉ Samba Raju Chiluveru  
samba.dee2017@iitr.ac.in

Manoj Tripathy  
manojfee@iitr.ac.in

<sup>1</sup> Department of Electrical Engineering, Indian Institute of Technology, Roorkee, India

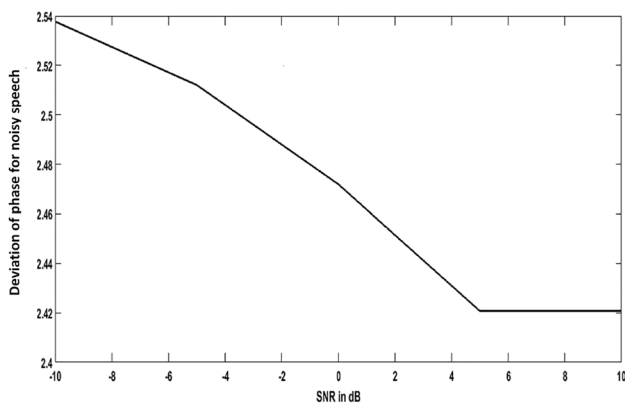


Fig. 1 Variation of phase with respect to SNR for noisy speech

Figure 1 shows phase deviation curve with respect to different values of SNR for noisy speech. The difference between clean phase and noisy phase is considered as phase difference and standard deviation of phase difference is plotted w. r. t. different values of SNR for noisy speech. From Fig. 1 it is clear that as SNR value decreases the phase deviation between clean phase and noisy phase increases and it will degrade speech quality. Moreover, Vary and Eurasip (1985) also proved the importance of phase in low SNR environment. Phase information is one of the important factor in human speech recognition and human ears are sensitive to small phase distortions or global spectral shifts in low SNR environment. Therefore this article presents DNN-based regression model to estimate clean amplitude and clean phase for reconstruction of speech signal to improve the quality and intelligibility for single channel at low SNR. The obtained result are compared with Xu et al. (2015) and it shows good improvement in its evaluation metrics and it is proved to be a good speech enhancer to any noise environments.

DNN is a member of feed forward neural network with many hidden layers. It has shown its excellent learning performance in many applications like pattern recognition, speech recognition, speech enhancement and voice analysis (Deng 2012) etc. Compared to shallow neural networks (SNNs) deep architecture has robust feature extraction capability and better learning ability i.e. ability to establish relationship between noisy speech and clean speech (LeCun et al. 2015). Earlier speech enhancement methods (Kamath and Loizou 2002; Ephraim and Malah 1984, 1985; Scarlart et al. 1996; Cohen 2003; Surendran and Kumar 2015; Bouzid et al. 2016) uses Mel Frequency Cepstral Co-efficient and Linear Predictive Coding co-efficient as speech features, while reconstructing speech signal from estimated samples approximation models has been used but this leads to reduce in speech quality and intelligibility (Chazan et al. 2000). Feature selection in neural network based algorithm

is very important and if features like Mel Frequency Cepstral Co-efficient or Linear Predictive Coding are used than it causes reduction in quality and intelligibility while reconstructing speech signal so in this proposed regression model direct log amplitude and phase values is used to reduce the speech reconstruction loss hence the overall speech quality and intelligibility improves (Xu et al. 2015).

Figure 2 shows typical architecture of DNN, it has N number of inputs ( $x_n$ ), M number of outputs ( $y_m$ ) and l number of hidden layers, though in the literature different types of DNN such as Deep Auto encoder (Hinton and Salakhutdinov 2006), Recurrent Neural Network (Graves et al. 2013), Long Short Term Memory (Hochreiter and Schmidhuber 1997) and Deep Belief Network (DBN) (Lee et al. 2009) etc. are reported, in this article DBN is considered.

Deep Belief Network (DBN) is a generative neural network model with deep architecture introduced by Hinton and Salakhutdinov (2006) along with a greedy layer wise learning algorithm. DBNs are built with stacked probabilistic model called Restricted Boltzmann Machine (RBM) (Bengio et al. 2009). A RBM is a two layer neural network in which inputs are connected stochastically with outputs and having symmetrical weighted connections. The training of DBN based regression model has been performed in two stages. First stage is unsupervised pre-training and second stage is supervised fine tuning. Pre-training is performed in deep network to avoid getting stuck in local minima of DBN. Discriminative fine-tuning of DBN is performed by adding a final layer of variables that represent the desired output and using back propagation Stochastic Gradient Descent (SGD) algorithm (Bottou 2010). The back propagation SGD algorithm with MMSE based objective function between the logarithm of noisy absolute values to logarithm of clean absolute values are adapted to train DBN (Xu et al. 2015) and similarly for phase the DBN regression

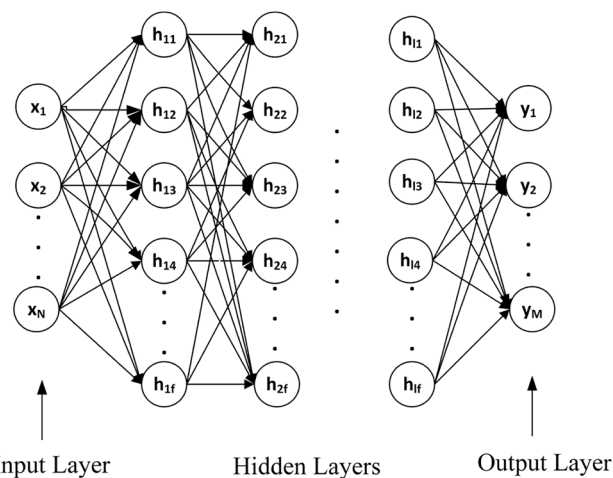


Fig. 2 Typical architecture of deep neural network

model is trained with noisy phase as input and clean phase as output.

The paper is organized as follows: Sect. 2 discusses proposed method, in Sect. 3 experimental result and analysis are explained, and at the end in Sect. 4 conclusion and future scope is presented.

## 2 Speech enhancement with phase estimation

Let  $x(n)$  is clean speech signal and  $d(n)$  is noise signal, then noisy signal  $y(n)$  is represented as  $y(n) = x(n) + d(n)$ , where,  $n$  is time index. Speech is assumed to be constant during short period of time (20–50 ms). Speech features are generated by applying Short Time Fourier Transform to the noisy speech signal which results in a complex values and then complex values of the noisy signal are obtained which are denoted by  $Y^c(k, l) = X^c(k, l) + D^c(k, l)$ , where  $Y(k, l) = |Y^c(k, l)|$ ,  $X(k, l) = |X^c(k, l)|$ ,  $D(k, l) = |D^c(k, l)|$  are the amplitude spectra for noisy signal, clean and noise speech respectively, while  $\phi_X(k, l) = \angle X^c(k, l)$ ,  $\phi_Y(k, l) = \angle Y^c(k, l)$ ,  $\phi_D(k, l) = \angle D^c(k, l)$  are phase angles of clean, noisy and noise speech signal respectively. Each sample of the complex number is expressed in terms of magnitude and phase value as shown below

$$|X^c| = \sqrt{X_r^2 + X_i^2} \tag{1}$$

$$\phi_X = \tan^{-1} \frac{X_i}{X_r} \tag{2}$$

where  $X_r, X_i$  is real and imaginary parts of speech sample respectively. The proposed algorithm framework is shown in Fig. 3. DBN based regression model is adapted to map between clean and noisy speech sample. The system is looked in two stages i.e. training stage and enhancement stage. The feature extracted from noisy speech signal for magnitude training is logarithm of the absolute amplitude whereas feature extracted from noisy speech signal for training of phase is linear phase. The estimation of amplitude and phase of speech signal are performed by using two similar type of DNN architecture and architectures are pre-trained using Bernouli–Bernouli RBM. The RBMs are stacked-up to produce multi layered structure and they are trained layer-by-layer in an unsupervised greedy fashion to maximize the likelihood over training samples. To improve the convergence rate, the SGD based back-propagation algorithm is used (Bottou 2010). The noisy speech is considered as input signal for DNN which is normalized so that the DNN is prevented from entering into saturation region.

Figure 3 illustrate block diagram of proposed speech enhancement method which has two stages i.e. training stage and enhancement stage. In training stage both magnitude training and phase training was performed simultaneously with the help of two similar type of DNNs. Training data were prepared using clean and noisy speech corpus and speech features are generated for both DNN’s training (magnitude training and phase training) simultaneously. The feature extraction procedure is same at training and enhancement stages. Prepared noisy speech corpus were downsampled to 8 kHz and it is applied to pre-emphasis block which will enhance high frequency components and the resulting signal was given for framing. Framing of signal

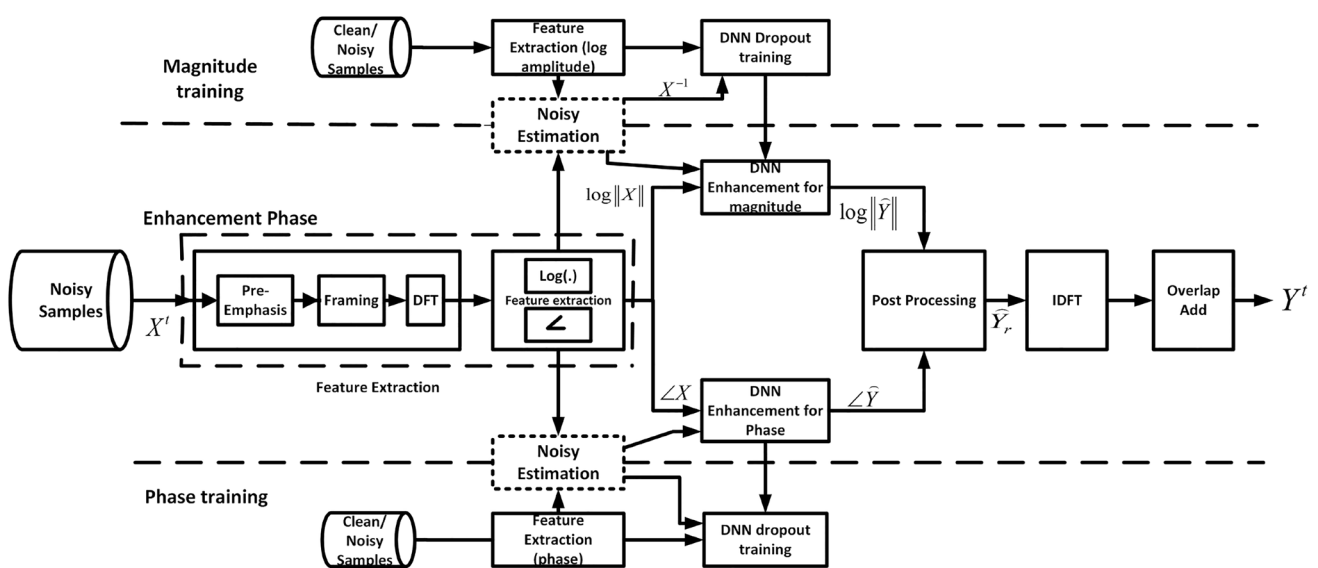


Fig. 3 Block diagram of proposed DNN based speech enhancement algorithm for low SNR speech signal with DNN estimated phase

has been performed by multiplying original signal with hanning window of 25 ms width and the resulted frame consist of 200 samples further framing losses can overcome by 50% (i.e.100 samples) overlapping of frame size. Suppose a signal has S number of sample points and further has a window size of m with an overlapping of d sample points then corresponding number of frames F is given by Eq. 3 (Mukherjee et al. 2018).

$$F = \left\lceil \frac{S - m}{d} + 1 \right\rceil \quad (3)$$

Short Time Fourier Transform is applied to the framed data and which result in complex data, further speech features are extracted from complex data i.e. logarithmic amplitude data and phase data. It is observed that phase shows odd symmetry and magnitude shows even symmetry so the half part of each data is rid of, the final resulted features used for training of respective DNNs.

Proposed algorithm was trained using SGD based back propagation algorithm and 50 inputs are considered for each epoch whereas batch size was kept at 100. The same architecture is used for both DNN-based amplitude regression model and DNN-based phase regression model. In DBN model hidden layers have sigmoid type activation function and output layer has ReLU type activation function (He et al. 2015).

In this work the baseline system mentioned in Xu et al. (2015) is used as regression model, it has three hidden layer and each layer consist of 500 neurons. Initialization of weights and biases has been done during pre-training with learning rate of 0.01 after that greedy layer unsupervised training was performed. Then back propagation of error derivatives are employed to fine tune the weights for optimal reconstruction of DBN. Output of DBN is calculated using following equation:

$$\hat{Y}_n = H(X_n, V_k^t) + b_n \quad (4)$$

While training DBN the error function is given by following equation:

$$E_r = \frac{1}{N} \sum_{n=1}^N \left\| \hat{Y}_n(V, b) - Y \right\|^2 \quad (5)$$

where  $E_r$  is mean square error,  $\hat{Y}_n(V, b)$  indicate estimated value at sample index n, Y represent reference value at sample index n, X represents input samples, H represent activation function and N is mini batch size, for each mini batch weights are updated, t be number of hidden layer, for tth layer iterative equation is given by following equation:

$$\begin{aligned} \Delta(\mathbf{V}_{n+1}^t, \mathbf{b}_{n+1}^t) = & -\lambda \frac{\partial E_r}{\partial (\mathbf{V}_n^t, \mathbf{b}_n^t)} - k\lambda(\mathbf{V}_n^t, \mathbf{b}_n^t) \\ & + \omega \Delta(\mathbf{V}_n^t, \mathbf{b}_n^t) \quad \text{for } 1 \leq t \leq L + 1 \end{aligned} \quad (6)$$

where  $\lambda$  is the learning rate,  $\omega$  is the momentum, k is weight decay coefficient, L represent total number of hidden layers. During training the DNN will learn non-linear mapping between the noisy speech and clean speech, no assumptions are made about the feature selection from noisy speech to the clean speech. It will automatically learn the complicated relationship to separate clean speech from the noisy signals given the sufficient training sets. To improve the generalization capability of DNN dropout training is used, dropout mechanism omits 20% of neurons in each layer for each back propagation step which is known as model averaging to avoid the over fitting problem. Drop out may decrease performance but it will improve the robustness in non-stationary training data (Hinton and Salakhutdinov 2006).

In enhancement stage, trained DNN-based regression model was tested with noisy speech corpus to predict clean speech features i.e. logarithmic amplitude feature and phase feature. Furthermore real and imaginary part of estimated complex sample value is calculated using following equations:

$$X_r = \left| \hat{X} \right| \cos \hat{\phi}_x \quad (7)$$

$$X_i = \left| \hat{X} \right| \sin \hat{\phi}_x \quad (8)$$

where  $\hat{X}$ ,  $\hat{\phi}$  are estimated amplitude and estimated phase respectively. The calculated real and imaginary values are converted into complex samples. The Inverse Short Time Fourier Transform is applied on complex samples to obtain an overlapped version of estimated/reconstructed speech signal and finally original signal is reconstructed using add and overlap method.

## 3 Objective evaluation

### 3.1 Datasets

The speech corpus is prepared by adding noise data with clean speech data at different SNR levels. Noise data was taken from Aurora2 database. Noise signals are selected to represent the most probable application scenarios for telecommunication terminals the framework was prepared as a contribution to the ETSI STQ-AURORA DSR Working Group (Pearce et al. 2000). Clean speech data was taken from TIMIT database which contains ten sentences spoken

by each of 630 speakers from eight major dialect regions of the United States. 70% of the speakers are male and 30% are female. Each speech and noise signals are sample down to 8 kHz (Garofolo et al. 1988). In this experiment 50 types of real time noise signal are considered for training and clean speech of 150 randomly chosen utterances (which have both male and female utterances) are considered for preparing the multi-condition training data set. Clean speech of 150 clean utterances are added to 50 type of noises at different levels i.e. SNR level varying from  $-25$  to  $+25$  dB with a step size of 5 dB. Total speech corpus becomes 82,500 utterances and each utterance has an average of 2 s, total speech corpus is divided into training set and validation set respectively. The training set has 80% of prepared speech corpus and validation set has 20% of speech data. In enhancement stage for evaluation of proposed algorithm clean speech is added to car, restaurant, factory, airport and babble noises at different levels such as  $-25$  to  $+25$  dB with a step size of 5 dB. Noisy speech data set was prepared with MATLAB software (Brainard and Vision 1997) and DNN-based regression model is implemented in python.

### 3.2 Performance parameters

Performance of proposed speech enhancement algorithm has been evaluated in two ways viz. speech quality and speech intelligibility. In this experiment speech quality measures considered are SNR (Steeneken and Houtgast 1980), segmental SNR (SNRseg) (Loizou 2013), frequency-weighted segmental SNR (fwSNRseg) (Hu and Loizou 2008), Log-Likelihood Ratio (LLR), Cepstral mean (Hansen and Pellom 1998), Weighted Spectral Slope (WSS) (Klatt 1982), Composite Objective Measure (COM) (Hu and Loizou 2008) and Perceptual Evaluation of Speech Quality (PESQ) (Rix et al. 2001) whereas speech intelligibility is measured using Short Time Objective Intelligibility (STOI) measure. In this work speech quality and musical noise of noisy speech was observed in low SNR environments.

### 3.3 Results and discussion

PESQ is highly correlated measurement proposed by ITU-T P.835 standards, it measures Mean Opinion Score (MOS), PESQ produces measured value in the range between  $-1$  and  $4.5$ , where, the maximum value  $4.5$  indicate best quality and  $-1$  indicate worst quality. Figures 4 and 5 shows the PESQ result for babble and restaurant noise respectively, bar charts are plotted at  $-15$  dB to  $0$  dB at a step size of 5 dB for low SNR and these results are plotted for proposed method, regression model with noisy phase, amplitude based regression model and logMMSE. It is found that for babble noise proposed method shows an improved average PESQ value

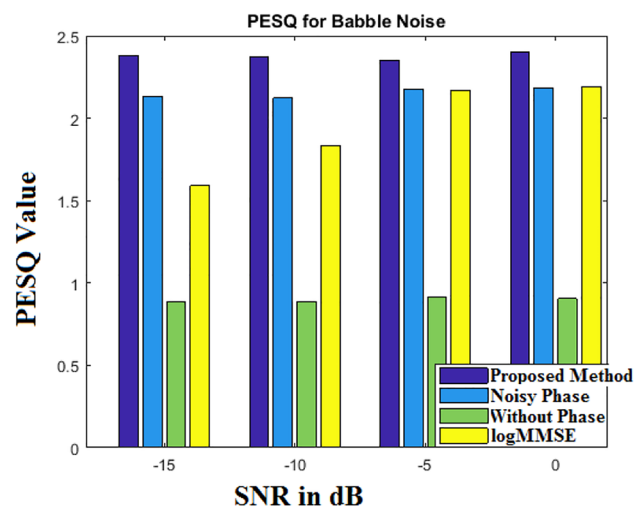


Fig. 4 Average PESQ results using different SNR levels for babble noise

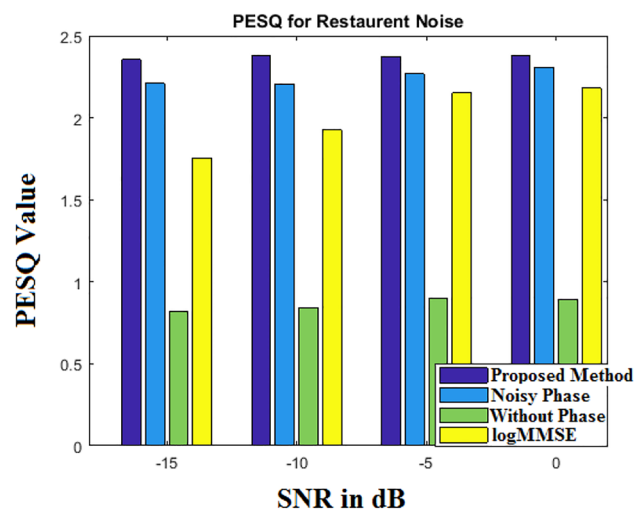


Fig. 5 Average PESQ results using different SNR levels for restaurant noise

of 9% compared to regression model with noisy phase, 190% compared to amplitude based regression model and 49% compared to logMMSE model. Similarly for restaurant noise improved average PESQ value of 7% compared to regression model with noisy phase, 185% compared to amplitude based regression model and 25% compared to logMMSE model.

COM is a speech quality measure and it consists of three ITU-T P.835 standard measures like signal distortion (Csd), noise distortion (Cnd), and overall signal quality (Covl). Any one of COM ( $C_x$ ) is evaluated by

$$C_x = \gamma_0 + \sum_{k=1}^5 \gamma_k O_k \tag{9}$$



where  $O_k$  is the individual objective measure and  $\gamma_k$  is the corresponding regression coefficient computed by multi variable adaptive regression splines technique (Hu and Loizou 2008). The regression coefficients are given in Table 1 (Samui et al. 2016) which shows the signal distortion (Csd), background distortion (Cnd) and overall signal quality (Covl), the detailed result of COM is given in Table 2, these result obtained for Factory, Restaurant, Car, Airport and Babble noise at different SNR levels i.e. 0 dB, – 5 dB, – 10 dB and – 15dB. The proposed algorithm is compared with regression model with noisy phase and proposed method shows improvement in an average Covl of 12% for factory noise, 8% for restaurant noise, 13% for car noise, 10% for airport noise and 14% for Babble noise respectively.

Table 3 shows result of speech quality measures like Frequency weighted Segmental Signal to Noise Ratio (fwSNRseg), Cepstral distance measure and WSS respectively.

fwSNRseg is extended version of SNRseg in frequency domain and it is given by

$$fwSNRseg = \frac{10}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^K B_j \log_{10} \left[ \frac{F^2(m,j)}{(F(m,j) - \hat{F}(m,j))^2} \right]}{\sum_{j=1}^K B_j} \quad (10)$$

where  $B_j$  is the weight placed on the  $j$ th frequency band,  $K$  is the number of bands,  $M$  is the total number of frames in the signal,  $F(m,j)$  is the filter-bank amplitude (excitation spectrum) of the clean signal in the  $j$ th frequency band at the  $m$ th frame, and  $\hat{F}(m,j)$  is the filter-bank amplitude of the enhanced signal in the same band. In Table 3 proposed fwSNRseg measurements are compared with regression model with noisy phase and amplitude based regression algorithm. Result shows proposed method improves average fwSNRseg of 13% compared to regression model with noisy phase and 340% compared to amplitude based regression algorithm.

**Table 1** Regression coefficients used for COM evaluation

Composite measure metric	$\gamma_0$	IS( $\gamma_1$ )	PESQ( $\gamma_2$ )	CEP( $\gamma_3$ )	LLR( $\gamma_4$ )	WSS( $\gamma_5$ )	SNRseg( $\gamma_6$ )
Signal distortion (Csd)	3.093	– 0.02	0.603	– 0.494	– 1.029	– 0.009	–
Background distortion (Cnd)	1.634	–	0.478	– 0.319	–	– 0.007	0.063
Overall signal quality (Covl)	1.594	– 0.011	0.805	–	– 0.512	– 0.007	–

**Table 2** COM in various noises at low SNR levels

Noise type	Input SNR (dB)	Signal distortion		Noise distortion		Overall signal quality	
		Estimated phase	Noisy phase	Estimated phase	Noisy phase	Estimated phase	Noisy phase
Factory	0	3.6904	3.3250	2.2396	2.0478	2.9877	2.6881
	– 5	3.6788	3.3134	2.2227	2.0150	2.9764	2.6686
	– 10	3.6511	3.3122	2.1955	2.0006	2.9456	2.6712
	– 15	3.6629	3.3109	2.2056	2.0077	2.9645	2.6809
Restarant	0	3.6848	3.5131	2.2892	2.2034	2.9806	2.8435
	– 5	3.6729	3.4306	2.2605	2.1419	2.9729	2.7771
	– 10	3.6630	3.3535	2.2402	2.0822	2.9690	2.7034
	– 15	3.6996	3.3401	2.2267	2.0657	2.9491	2.6923
Car	0	3.6781	3.5184	2.2761	2.1967	2.9859	2.8688
	– 5	3.6650	3.4517	2.2503	2.1468	2.9749	2.8142
	– 10	3.5795	3.2890	2.1745	2.0306	2.8632	2.6311
	– 15	3.5760	3.1933	2.1695	1.9724	2.8574	2.5294
Airport	0	3.6151	3.3199	2.2197	2.0614	2.9080	2.6625
	– 5	3.5424	3.2364	2.1528	2.0065	2.8182	2.5816
	– 10	3.5254	3.1455	2.1364	1.9629	2.7999	2.4950
	– 15	3.4746	3.0803	2.0917	1.9341	2.7333	2.4354
Babble	0	3.6800	3.2860	2.2470	2.0220	3.6800	3.2860
	– 5	3.6459	3.2446	2.2169	2.0077	3.6459	3.2446
	– 10	3.6486	3.1834	2.2188	1.9813	3.6486	3.1834
	– 15	3.6574	3.1964	2.2271	1.9866	3.6574	3.1964

**Table 3** Mean WSS, average Cepstral mean (Cep\_mean) and fwSNRseg for Low SNR

Noise type	Input SNR (dB)	WSS_Mean			Cep_Mean			fwSNRseg		
		Estimated phase	Noisy phase	Without phase	Estimated phase	Noisy phase	Without phase	Estimated phase	Noisy phase	Without phase
Factory	0	42.1943	58.1976	145.6180	3.9295	4.7457	8.7515	8.9518	7.8906	1.9318
	-5	42.2613	57.7732	146.1555	3.9629	4.7115	8.7428	8.9328	7.8427	1.9780
	-10	42.7474	57.4367	146.1991	3.9902	4.7823	8.7336	8.9029	7.9171	1.9672
	-15	42.7733	56.6882	146.4147	4.0000	4.8493	8.7287	8.8696	7.8153	1.9493
Restarant	0	43.2670	49.5455	143.7534	3.8141	4.2253	8.7467	8.9431	8.2115	1.9591
	-5	43.3728	51.8782	145.3654	3.8763	4.4764	8.7250	8.8832	8.1503	1.9320
	-10	43.7947	53.4178	146.2272	3.9369	4.6167	8.7202	8.8452	7.8817	1.9231
	-15	43.9433	55.5389	146.0578	3.9165	4.6602	8.7393	8.8769	7.7991	1.8942
Car	0	42.3882	49.8264	144.2300	3.9610	4.2908	8.7552	8.9238	8.1872	1.8676
	-5	42.5737	51.8937	144.7960	4.0310	4.4796	8.7477	8.8718	8.0006	1.8187
	-10	43.0168	54.2113	144.7490	4.0254	4.6285	8.7551	8.8641	7.7530	1.8406
	-15	43.2750	56.2229	144.7347	4.0220	4.7811	8.7679	8.8353	7.5553	1.8341
Airport	0	43.4976	55.7012	142.1477	3.9561	4.4995	8.7438	8.8557	7.6806	1.8969
	-5	43.7146	56.8719	145.7028	4.0092	4.6981	8.7435	8.7945	7.5429	1.9536
	-10	43.9793	58.2803	145.4499	4.0304	4.9127	8.7462	8.7206	7.3746	1.9679
	-15	43.9473	58.6236	145.7875	4.0306	5.1042	8.7446	8.7469	7.2336	2.0305
Babble	0	42.9208	56.7157	146.0007	3.9998	4.8545	8.7589	8.9845	7.6594	1.9189
	-5	43.1807	58.6518	146.3449	3.9957	4.9382	8.7703	8.9755	7.5153	1.8988
	-10	43.5307	59.6429	146.6884	4.0275	4.9670	8.7603	8.9112	7.2495	1.9087
	-15	43.2651	59.6581	146.5088	4.0280	4.9327	8.7584	8.8824	7.2130	1.9007

Cepstral distance measure is distortion measure between input and output speech signal which is measured in frequency domain and the Cepstral distance value is shown in following equation:

$$Cep = \frac{10}{\log_{10} \sqrt{2 \sum_{i=1}^P \{y(k, t) - x(k, t)\}^2}} \quad (11)$$

where  $y(k,t)$  and  $x(k,t)$  are input and output speech signal respectively and  $P$  is maximum number of coefficients. Table 3 shows results of Cepstral mean for different values of SNR, result shows proposed method improves mean value of Cepstral distance by 12% compared to regression model with noisy phase and 240% compared to amplitude based regression algorithm.

Weighted Spectral Slope (WSS) is a speech quality measure in frequency domain, WSS simulated result are shown in Table 3. Generally WSS value decreases as accuracy of the system increases, from the Table 3 it is observed that proposed method shows an improved performance of 27% compared to regression model with noisy phase and 310% compared to amplitude based regression algorithm.

The Log Likelihood Ratio (LLR) (Hu and Loizou 2008) is distance quality measure which is weakly correlated with noise reduction and strongly correlated with speech distortion, LLR value increases as distortion increases, for distortion-less speech low value of LLR is required, Table 4 shows LLR values it is observed that the LLR values for estimated

phase method show 22% improvement compared to noisy phase regression model and 320% compared to amplitude based regression algorithm.

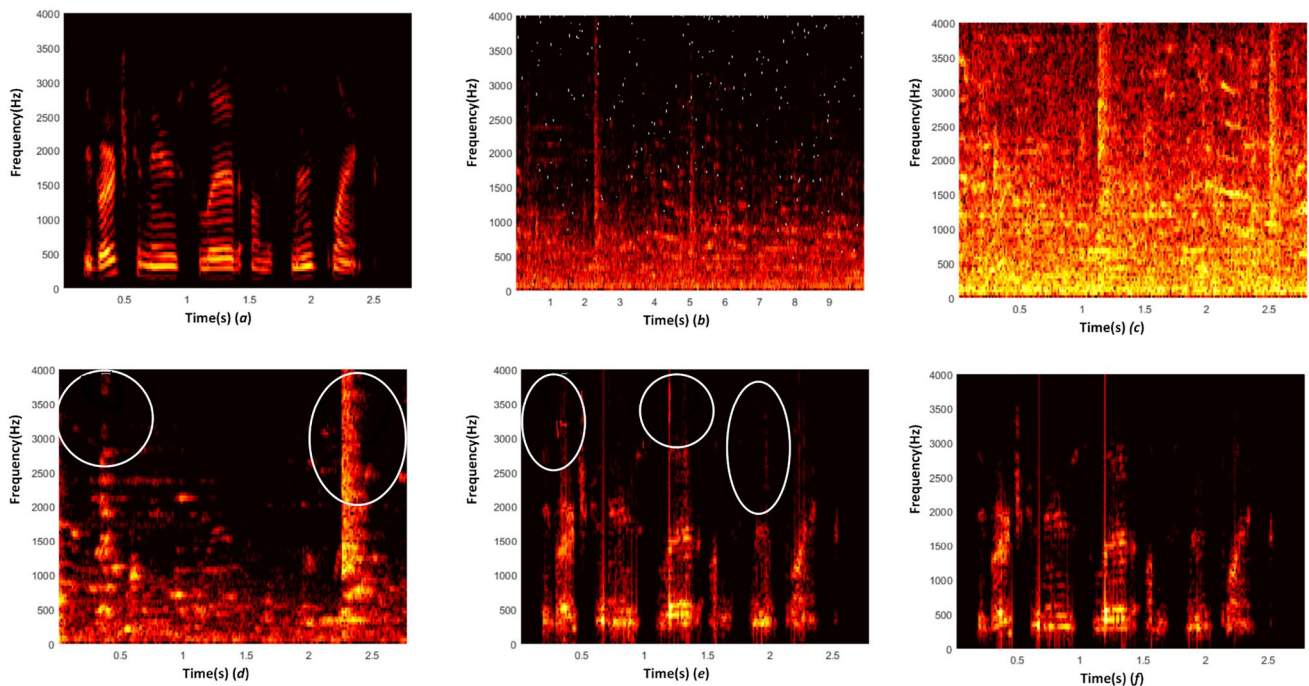
Spectrogram of utterance “the birch canoe slid on the smooth planks” is shown in Fig. 6. English clean speech utterance is of 3s width and it is sampled at 8 kHz. The clean speech shown in Fig. 6a, noise audio of airport is shown in Fig. 6b, and mixture of both clean speech and airport noise at  $-15$  dB SNR shown in Fig. 6c. The enhanced speech has been obtained using logMMSE, noisy phase constructed regression model and proposed method, Fig. 6d shows spectrogram of logMMSE reconstructed speech signal, Fig. 6e shows noisy phase constructed speech spectrogram and Fig. 6f shows the proposed method speech spectrogram. From Fig. 6 it is observed that Fig. 6d, e (i.e. logMMSE and noisy phase reconstructed spectrogram) shows high frequency noisy components after reconstruction whereas proposed method shows (i.e. Fig. 6f) better filtering of high frequency components so logMMSE and noisy phase regression model may not filter high frequency components which are responsible for musical noise and musical noise is highlighted in spectrogram with circles, whereas proposed algorithm filtered high frequency components hence the proposed method reduces musical noise compare to other two methods.

The proposed algorithm is evaluated with speech quality measure like PESQ, COM, WSS and LLR. The result of proposed method compared with noisy phase based regression model, without phase based regression model (amplitude

**Table 4** LLR measure values

Noise type	Input SNR (dB)	Estimated phase	Noisy phase	Without phase
Factory	0	0.4456	0.5604	1.6846
	-5	0.4508	0.5714	1.6846
	-10	0.4552	0.5743	1.6847
	-15	0.4586	0.5850	1.6844
Restarant	0	0.4405	0.5108	1.6807
	-5	0.4494	0.5480	1.6797
	-10	0.4577	0.5730	1.6820
	-15	0.4546	0.5685	1.6815
Car	0	0.4621	0.6549	1.6844
	-5	0.4686	0.5582	1.6843
	-10	0.4688	0.5844	1.6854
	-15	0.4661	0.6029	1.6865
Airport	0	0.4629	0.5644	1.6770
	-5	0.4701	0.5931	1.6792
	-10	0.4736	0.6252	1.6794
	-15	0.4753	0.6549	1.6777
Babble	0	0.4600	0.5964	1.6888
	-5	0.4620	0.6167	1.6909
	-10	0.4677	0.6338	1.6892
	-15	0.4689	0.6266	1.6879





**Fig. 6** English utterance “the birch canoe slid on the smooth planks”, by a male speaker spectrogram shown in (a), airport noise in (b), noisy speech (airport noise utterance at  $-15$  dB) in (c), logMMSE based reconstruction (d), noisy phase reconstruction (e) and proposed method in (f)

based model) and logMMSE model and the comparative result shows good improvement interms quality measure for proposed method. Among all compared methods proposed method shows constant improvement for all SNR values whereas remaining methods shows sharp falloff in speech quality as SNR value decreases. The proposed method shows good improvement over amplitude based regression model interms of all parameters. Further it can be found that the improvement between phase estimated regression method and noisy phase regression method is less as SNR value of noisy speech increases (shown in Fig. 5). Spectra illustrated in Fig. 6f which is obtained from the proposed method shows the removal of musical noise compared with noisy phase regression model as shown in Fig. 6d and logMMSE model shown in Fig. 6c respectively.

#### 4 Conclusion and future scope

In this contribution we presented importance of phase estimation in low SNR environment for single channel noisy speech signal. DNN-based regression model used for estimation of clean phase and clean logarithm of amplitude in low SNR environment and it is proved that the importance of phase estimation in low SNR (negative SNR) environment. Proposed method compared with noisy phase regression model, amplitude based regression model and logMMSE

model and it is observed that proposed method shows improved result. A clear distinction in the reconstructed speech signal is observed in terms of musical noise along with improvement in various objective measures such as PESQ, LLR, WSS, fwSNRseg and Cep-mean.

In future the proposed method may be implemented with live data on Field Programmable Gate Array so that it can be tested in real time environment. The live data can be handled by active learning (Bouguelia et al. 2018).

#### References

- Benesty, J., Makino, S., & Chen, J. (2005). *Speech enhancement, signals and communication technology*. Berlin: Springer.
- Bengio, Y., et al. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1), 1–127. <https://doi.org/10.1561/2200000006>.
- Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In: *Acoustics, Speech, and Signal Processing, International Conference on ICASSP, IEEE* (Vol. 4, pp. 208–211).
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (pp. 177–186). Springer. <https://doi.org/10.1007/978-3-7908-2604-316>.
- Bouguelia, M. R., Nowaczyk, S., Santosh, K., & Verikas, A. (2018). Agreeing to disagree: Active learning with noisy labels without crowdsourcing. *International Journal of Machine Learning and Cybernetics*, 9(8), 1307–1319.
- Bouزيد, A., Ellouze, N., et al. (2016). Speech enhancement based on wavelet packet of an improved principal component analysis.

- Computer Speech & Language, Elsevier*, 35, 58–72. <https://doi.org/10.1016/j.csl.2015.06.001>.
- Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision*, 10, 433–436.
- Chazan, D., Hoory, R., Cohen, G., & Zibulski, M. (2000). Speech reconstruction from mel frequency cepstral coefficients and pitch frequency. In *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings* (Vol. 3, pp. 1299–1302). IEEE. <https://doi.org/10.1109/ICASSP.2000.861816>.
- Cohen, I. (2003). Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5), 466–475. <https://doi.org/10.1109/TSA.2003.811544>.
- Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 30–42. <https://doi.org/10.1109/TASL.2011.2134090>.
- Deng, L. (2012). Three classes of deep learning architectures and their applications: a tutorial survey. *APSIPA Transactions on Signal and Information Processing*, 1, 60–88.
- Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6), 1109–1121. <https://doi.org/10.1109/TASSP.1984.1164453>.
- Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2), 443–445. <https://doi.org/10.1109/TASSP.1985.1164550>.
- Garofolo, J. S., et al. (1988). Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. National Institute of Standards and Technology (NIST), Gaithersburg, MD 107.
- Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6645–6649). IEEE. <https://doi.org/10.1109/ICASSP.2013.6638947>.
- Hansen, J. H., & Pellom, B. L. (1998). An effective quality evaluation protocol for speech enhancement algorithms. In *Fifth international conference on spoken language processing*. Sydney, Australia.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504–507. <https://doi.org/10.1126/science.1127647>.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Hu, Y., & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1), 229–238. <https://doi.org/10.1109/TASL.2007.911054>.
- Kamath, S., & Loizou, P. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *ICASSP, Citeseer* (Vol. 4, pp. 44164–44164).
- Klatt, D. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. In *Acoustics, speech, and signal processing, IEEE international conference on ICASSP'82* (Vol. 7, pp. 1278–1281). IEEE. <https://doi.org/10.1109/ICASSP.1982.1171512>.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>.
- Lee, H., Grosse, R., Ranganath, R., Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning* (pp. 609–616). ACM. <https://doi.org/10.1145/1553374.1553453>.
- Loizou, P. C. (2013). *Speech enhancement: Theory and practice*. Boca Raton: CRC Press.
- Loizou, P. C., & Kim, G. (2011). Reasons why current speech enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1), 47–56. <https://doi.org/10.1109/TASL.2010.2045180>.
- Mukherjee, H., Obaidullah, S. M., Santosh, K., Phadikar, S., & Roy, K. (2018). Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. *International Journal of Speech Technology*, 21, 753–760.
- Pearce, D., & Hirsch H. G. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000* (pp. 29–32). Paris, France.
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3), 19–41. <https://doi.org/10.1006/dspr.1999.0361>.
- Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings*. (Vol. 2, pp. 749–752). IEEE. <https://doi.org/10.1109/ICASSP.2001.941023>.
- Samui, S., Chakrabarti, I., & Ghosh, S. K. (2016). Improved single channel phase-aware speech enhancement technique for low signal-to-noise ratio signal. *IET Signal Processing*, 10(6), 641–650. <https://doi.org/10.1049/iet-spr.2015.0182>.
- Scalart, P., et al. (1996). Speech enhancement based on a priori signal to noise estimation. In *Acoustics, speech, and signal processing, 1996. ICASSP-96. Conference proceedings* (Vol. 2, pp. 629–632). IEEE. <https://doi.org/10.1109/ICASSP.1996.543199>.
- Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1), 318–326. <https://doi.org/10.1121/1.384464>.
- Surendran, S., & Kumar, T. K. (2015). Perceptual subspace speech enhancement with variance normalization. *Procedia Computer Science*, 54, 818–828. <https://doi.org/10.1016/j.procs.2015.06.096>.
- Vary, P., & Eurasip, M. (1985). Noise suppression by spectral magnitude estimation mechanism and theoretical limits. *Signal Processing*, 8(4), 387–400. [https://doi.org/10.1016/0165-1684\(85\)90002-7](https://doi.org/10.1016/0165-1684(85)90002-7).
- Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(1), 7–19. <https://doi.org/10.1109/TASLP.2014.2364452>.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.