Theoretical Computer Science 941 (2023) 14–28

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Speech enhancement with noise estimation and filtration using deep learning models

Sravanthi Kantamaneni [a,*], A. Charles [a], T. Ranga Babu [b]

[a] ECE, Annamalai University, Tamil Nadu, India
[b] ECE, R.V.R&J.C College of Engineering, Andhra Pradesh, India

ARTICLE INFO

Article history:
Received 20 April 2022
Received in revised form 11 June 2022
Accepted 19 August 2022
Available online 24 August 2022

Keywords:
Speech enhancement
Perceptual quality
Speech signal
RESNET-50
Denoising
Deep transfer learning model

ABSTRACT

Speech enhancement helps in eliminating the environmental noises from the communication signals. The main intention of the augmentation system is to develop the perceptual quality of communication or speech. For this purpose, various filtering schemes, spectral restoration models and speech models were implemented. In order to improve the odds of reducing noise and restoring the original signal, artificial intelligence (AI) and machine learning algorithms (MLA) were included into every sector. Deep transfer learning was used in this work to remove noise from the data and restore the original signals. This proposed approach includes a filtration scheme instead of using a convolution layer in the RESNET-50 architecture. The filters tested for speech enhanced deep learning models are modified Kalman filter and enhanced wiener filter. The performance metrics were calculated between various algorithms and proposed models to identify which approaches to follow the better way result obtained. The performance metrics compared PESQ, LSD and segSNR for different low signal to noise ratio conditions.

1. Introduction

Spectral extraction strategy [1] is probably the most common strategy among the various methods of noise reduction in the phantom realm. This process achieves the reduction of chaos by simply removing the pre-tested phantom sound from the visual acuity of the ghost, when the negative phase is not addressed. False withdrawal strategy is easy to implement and effectively reduces the constant noise [2]. However, it introduces a false sound, called melodic commotion, caused by speech errors. The phantom drew approach usually gives assessment mistakes because it is unconcerned about extraterrestrial information [3]. With PDF (Probability Density Function) and audio PDF has been suggested an MMSE-STSA (Minimum Mean Square Error–Short-Short-Short-Spectral Amplitude). There was a demonstration of the capacity of Rayleigh (and Gauss) to thickening in the text of [4]. The MMSE-STSA sensor receives a well-configured voice signal thanks to this approach (the system can be converted into a Wiener channel [5] in case we expect Gauss gain in both speech and PDF audio). This technique delivers an anticipated voice signal with low musical instabilities, but it necessitates the acquisition of modified Bessel functions, which may be perplexing. In addition, since has been called the attention of some commentators, the histograms of the actual speech do not correspond with Rayleigh's work [6]. A variety of speech augmentation techniques have been

Abbreviations: Standard Kalman Filter, KF; Back Propagated Kalman Filter, BKF; Weiner Filter, WF; Modified Weiner Filter, MWF.

* Corresponding author.
E-mail addresses: sravanthivasanth@gmail.com (S. Kantamaneni), charlesgceb@gmail.com (A. Charles), trbaburvr@gmail.com (T.R. Babu).

https://doi.org/10.1016/j.tcs.2022.08.017
0304-3975/

published. Spectral subtraction, Wiener and Kalman filtering, MMSE estimation, comb filtering, subspace approaches, and phase spectrum compensation are some of the techniques used. Voice enhancement is an algorithm that increases perceived speech quality, reduces hearing fatigue, and improves speech intelligibility in digital communications, speech preparation for hearing aids, and speech recognition.

The created a sophisticated technique based on large-scale posteriori (MAP) tests. Two state borders are depicted in PDF presentations by [7] and Vary of the parametric super-Gaussian volumes. The parametric super-Gaussian volume is created from a histogram generated using a lot of real speech information with a small Signal to Noise Ratio (SNR) span [8] [9]. The ability on the way to hide chaos of this process is better than the Wiener channel. Nevertheless, the constant sound SNR may be affected by PDF speech, as Andrianakis and White were aware [10]. They employed three scatter plots created from voice sounds in three solid SNR zones to assess the Gamma density function of the data sets. Modifying these three-word PDFs in accordance with SNR, as shown in [11], may enhance the capacity to minimize noise. In contrast to PDF of speech de-nosing can attain more improvement [12]. In [13] suggested de-noise of the parametric super-Gaussian and modified parameters and its form according to the SNR [14]. Since then [15] modified this approach by performing and testing multiple real-world programming programs produced using limited SNR intervals. As it turned out [16], this strategy has a strong degree of chaos compared to other common speech development strategies, and is therefore particularly effective in low SNR scenarios.

The paper was organized as filtration scheme with their mathematical representation, continued by proposed deep learning models and the results of performance metrics was compared with final marks.

Noise may be removed from a noisy voice source using a speech enhancement algorithm (SEA). A front-end tool designed for a variety of implementations, consists of voice communication techniques, hearing-aids, and voice identification. Many SEAs have been developed; including SS (spectral subtraction), MMSE, WF (Wiener Filter), and KF (Kalman filter).

The SS method relies heavily on the accuracy with which the noise PSD is estimated. The MMSE and WF-based SEAs rely heavily on a priori SNR estimates. A decision-directed (DD) approach was used to estimate the a priori SNR. A priori SNR prediction is challenging since the speech and noise power levels from the previous frame are utilized [16].

How precisely the noise variance and LPCs are predicted affects KF-based SEA efficiency. There is no practical use for LPCs since they are based on clean speech, which is not accessible. Aside from that, it is only capable of improving voice that has been unclear with the addition of white Gaussian noisy (AWGN). For improving speech in varied noise environments, a sub-band iterative KF has been developed. Decomposition into 16 sub-bands begins with the loud speech (SBs). The partly reconstructed high-frequency (HF) SBs are subsequently enhanced using an iterative KF [17]. The low-frequency (LF) SBs are expected to be less influenced by noise and are left unprocessed. A derivative-based technique is used to estimate the noise variance for the sub-band iterative KF [18].

For voice augmentation, deep neural networks (DNNs) are becoming popular. To compute the spectrum of clean speech, time-frequency masks are utilized in DNN-based SEAs. Masks for voice enhancement have to be tested to discover the optimum one, a six-mask comparison research was conducted. Noise and musical noise are often introduced into improved speech via the masking method.

It was first presented in as an FCNN-based SEA, or fully convolutional neural network. This technique is specifically developed to improve speech that has been distorted by the interference of squeaks and screeches. A FCNN-based raw waveform-based SEA was suggested. A raw waveform input and output means that the improved speech is not influenced by the phase difficulties that are inherent in magnitude spectrum-based speech enhancement algorithms. Phase-aware DNN for speech augmentation was presented by Zheng et al. With the use of a time-frequency mask, phase information is combined with a time-frequency mask.

Reconstruction of the improved speech is done through the expected mask and the IFD segment data. A neural network deep learning is used to estimate LPCs in a KF-based SEA developed by Yu et al. Noise covariance is calculated during speech gaps, which is ineffective in nonstationary noise circumstances. It was also unclear how quiet detection works. The largest difference among the modulated wave's instantaneous phase angle and the carrier's phase angle in phase shift. The phase variation of a sinusoidal modulating wave, given in radians, is equal to the modulation index.

The input (speech signal) considered for the model is given by $I_{speech}(t)$ where it is known for the series of nonlinear autoregressions function (f(.)) with the original and estimated noise signal

$$I_{speech}(t) = f\left(I_{speech}(t-1).....I_{speech}(t-M), W\right) + v(t-1) \tag{1}$$

Where, t is time step of the signal. The extended signal considered as input signal was a combination of $I_{speech}(t)$ and noisy signal with respect to time $n(t)$ is expressed as

$$\hat{I}_{speech}(t) = I_{speech}(t) + n(t) \tag{2}$$

2. Literature survey

The Significance of speech de-noising technique increases in lip-reading and speech augmentation, as measured by both speech quality and intelligibility, can be seen in comparison simulations findings. Deep learning-driven lip-reading models are being worked on to improve their accuracy and generalizability via the addition of AV cues and other context-aware signals [19].

The Proposed voice enhancement approach has better quality and intelligibility than other methods in a broad range of noise circumstances, according to experiments [20].

The Speech and noise LPC improvements have been proven to help the AKF reduce the remaining sound along with deformation inside the improved voice. For a broad range of SNR values, experimental findings make known that the improved voice generated through the projected technique has a superior quality and intelligibility [21].

The approach is superior than a traditional minimum mean squared error spectrum estimator for real-time 48 kHz operations on a low-power CPU [22].

Especially in comparison to the other techniques, DNN-based complex-noisy speech improvement was superior in terms of voice quality when compared to PESQ, SNRSeg, LLR, and weighted spectral slope (WSS)-based speech improvement (WSS). In addition, short-term objective intelligibility improved speech intelligibility (STOI) [23].

The DNN-based Kalman filter (KF) technique for voice augmentation is presented by Author, where DNN is used to estimate critical KF parameters, such as linear prediction coefficients (LPCs) (LPCs). As a result, our proposed DNN-KF method is able to estimate LPCs from noisy speech more reliably and robustly than previous KF based techniques in speech enhancement, resulting in an enhanced overall performance. Our DNN-KF technique surpasses two other KF-based speech augmentation techniques are voice quality and intelligibility, according to the findings of our experiments [24].

An acoustic model of a hybrid DNN-HMM system and a single channel speech augmentation model are examined in this paper for noise-resistant ASR. Examining two methods of boosting. A DNN-based noise estimator and frame-wise estimate of the filter parameters are utilised to mask the noisy voice signal. Using a DNN-based mask estimation, a speech mask is generated. These components are improved using the acoustic model of the ASR system. When employing a Wiener filter on the CHiME-4 data, a single-channel ASR system may be able to improve its performance, as the WER drops from 11.7 percent to 10.6 percent [25].

There are various deep learning models for voice improvement, and this model was tested using a combination of the speech collection quantity and DEMAND information's. The mixed dataset was used to perform ablation experiments, and the results suggest that all three techniques are viable. Testing shows that the suggested technique has state-of-the-art performance, exceeding earlier methods significantly [26].

In compared to the original method created for singing voice separation in music, we show that a smaller number of hidden layers is adequate for speech augmentation. Initial results show that voice augmentation in the time domain may be used both as an end goal and as a pre-processing measure used for voice identification techniques [27].

The suggested system's efficiency with two distinct approximated masks is assessed using simulated and measured room impulse responses. IRM's direct application leads to greater short-term objective intelligibility gains than IBM's use as a signal for PSD modifications. The suggested system's performance analysis also reveals the system's ability to withstand varying angular locations of the voice source [28].

The improved speech generated by the proposed technique displays superior quality and transparency than the target techniques in coloured & non-stationary sound environments over a broad collection of SNR values [29].

Additionally, 10 listeners rated DeepLPC's improved speech as their favourite. The AKF's enhanced speech quality and intelligibility is made possible by DeepLPC's less biassed clean speech and noise LPC estimates [30].

The noise and distortion in enhanced speech may be reduced because to the AKF's better LPCs. On the NOIZEUS corpus, objective and subjective tests demonstrate that the improved voice generated through the projected technique has superior excellence and understandability than the standard systems under varied sound settings over a broad variety of SNR stages [31].

The Sensitivity tests and seven objective measures of quality and intelligibility evaluation in the NOIZEUS corpus show that the AKF built with MHANet-LPC-PS driven speech and noise LPC parameters produced improved speech with higher quality and intelligibility than competing methods (CSIG, CBAK, COVL, PESQ, STOI, SegSNR, and SI-SDR) [32].

The Deep-LPC-MHA Net-improved AKF's speech was also the most popular among the study's 10 participants. For the first time, the Deep-LPC-MHA Net approach produces the most accurate LPC estimations ever, allowing the AKF to provide improved speech of the highest quality and intelligibility [33].

Compared to a model that only used audio, the comparative simulation results express that the AV DNN performance superior than the A-only method in terms of objective metrics such as PESQ, STOI, SI-SDR, and DBSTOI. Subsequently, subjective hearing tests on the actual noisy AV ASPIRE corpus reveal that the proposed AV DNN outperforms current techniques in terms of accuracy and speed [34].

The Use of random projections and extreme learning machines creates a vector space of commonsense knowledge. It is possible to better represent high-dimensional data by combining random multidimensional scaling with random initialization learning approaches, as well as more efficiently identify the semantics or emotionally correlations among that information [35].

3. Filtration schemes

3.1. Extended Kalman filtration

A KALMAN filter works on the principle of linearity, but speech signals are non-linear. Therefore, a non-linear filter that is linearized at every step of time was required. This property of non-linear to linear is possible only by combining two

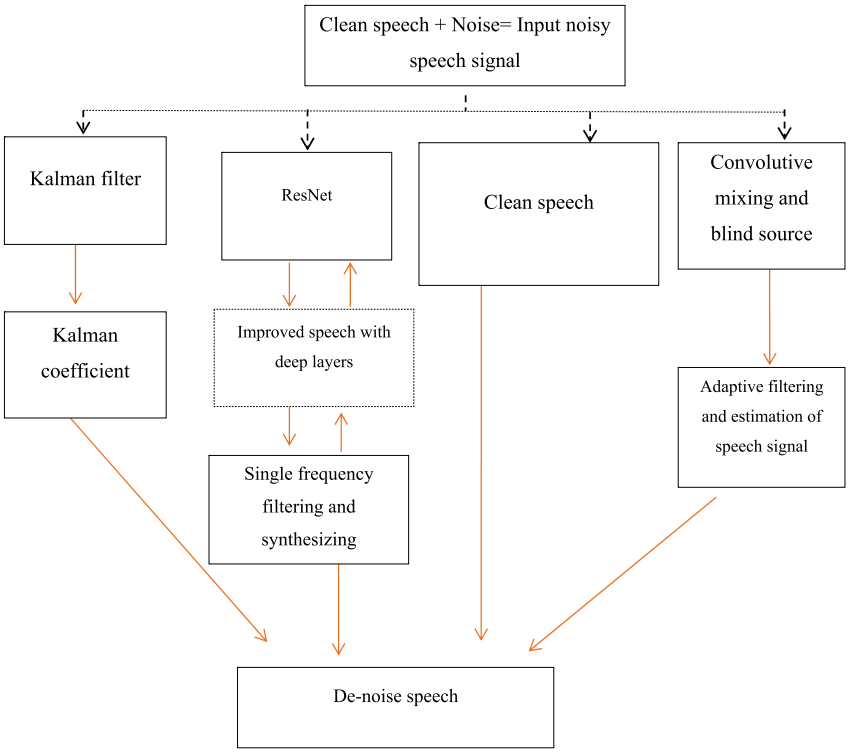

Fig. 1. Speech de-noise block diagram.

modes of Kalman filter a) Standard Kalman filter (KF) and b) Backward Propagated Kalman filter (BKF). The fusion of these two modes illustrates the smoothen description of the Kalman filter and is named the Extended Kalman Filter approach.

The Kalman presents a novel voice enhancement method based on perceptually correlated noise reduction. The suggested filters are based on applying two perceptual filtering models to noisy speech signals: the gammatone and gammachirp filter banks with nonlinear resolutions on the comparable rectangular bandwidth (CRB) scale. Perception filtering produces a number of sidebands, each of which is spectrally evaluated and changed using one of two noise suppression algorithms. The minimization of musical noise artefacts in unprocessed speech that arise after the standard subtractive process is linked to the need for accurate noise estimation. We employ continuous noise estimating techniques in this case. The suggested method is tested on voice signals that have been tainted by real-world sounds. It is demonstrated that our speech processing methodology using filter banks modelling the human auditory system outperforms the traditional spectral modifying algorithms in terms of quality and intelligibility of the improved speech using objective test cases based on the improvement measures PESQ score and the ratings of signal distortion (SIG), noise distortion (BAK), and overall quality (OVRL), as well as qualitative tests regarding the quality rating of voice recognition (VR) which is shown in Fig. 1.

Fig. 1, illustrates the overall block diagram of proposed methodology. It takes clean speech with some noise as input that is filtered using Kalman filter where Kalman coefficient is estimated, RESNET enhances the speech for analysis as well as synthesis, and adaptive filtering is done using convolutive mixing.

In KF, the maximum likelihood points in speech signals were traced, and in BKF, the covariance of the signal was estimated; this combination of estimating likelihood points and covariance leads to tracing the noise in the signal very effectively can be eliminated. The steps included in eliminating the noise from the signal are

a) Predicting non-linear likelihood points with respect to linearity in the speech signal, the predicted signal was given by $I^{-}_{speech}(t)$

$$I^{-}_{speech}(t) = F\left[I^{-}_{speech}(t-1).....I^{-}_{speech}(t-M), W(t-1)\right] \tag{3}$$

b) Estimating the noise covariance during the backward Kalman filter scheme that helps in predicting the noise variance of the signal ($P_I(t)$ and $P^{-}_{I}(t)$). During this step assuming noises $v(t)$ and $n(t)$ and represented variance variable σ^2 and this stage is mathematically represented as

$$P^{-}_{I}(t) = AP_I(t)A^T + B\sigma^2 B^T \tag{4}$$

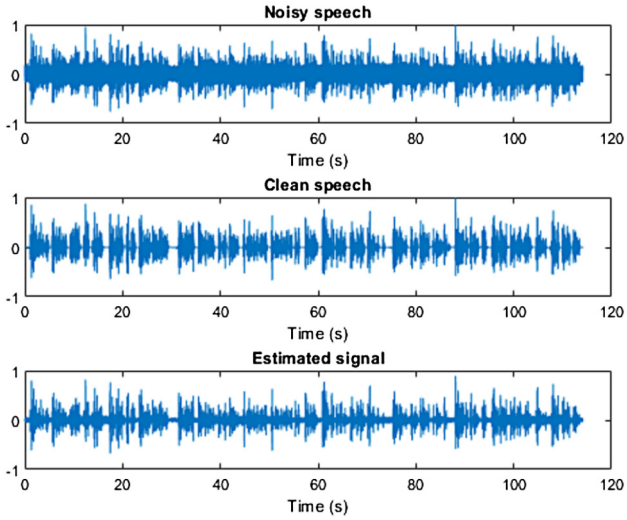

Fig. 2. Basic graphs before and after processing.

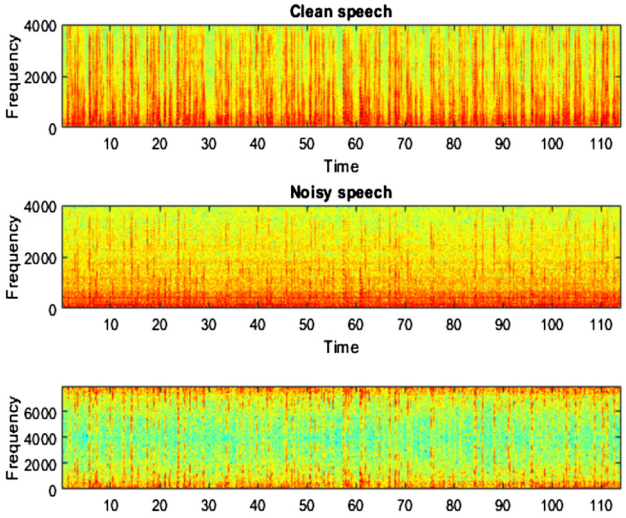

Fig. 3. EKFPSD.

c) As the signal needs to be verified at each step A and are given as the partial derivatives of the original signal and the transpose with multiplicand factor in equation (4) illustrates cancellation of the signal with change in prediction and is given by

$$A \triangleq \left.\frac{\partial F\left[I^{-}_{speech}, W^{-}\right]}{\partial I^{-}_{speech}}\right|_{I^{-}_{speech}(t-1)} \tag{5}$$

d) Extracting coefficients (K) that needs to eliminate from the signal in minimizing the noise effect of the signal that is given by

$$K(t) = P^{-}_{I}(t)\, C^{T}\left(C.P^{-}_{I}(t)\, C^{T} + \sigma^{2}\right)^{-1} \tag{6}$$

e) The final noise eliminated signal as updating the predicting signal $P_I(t)$ is expressed as

$$P_I(t) = \left(\hat{I}_{speech}(t) - K(t)\right) P^{-}_{I}(t) \tag{7}$$

The Figs. 2 and 3 clearly explain about before speech enhancement and after EKF PSD speech de-noise process. In this differentiation compared to earlier stage techniques proposed model can improve the speech quality and intelligibility.

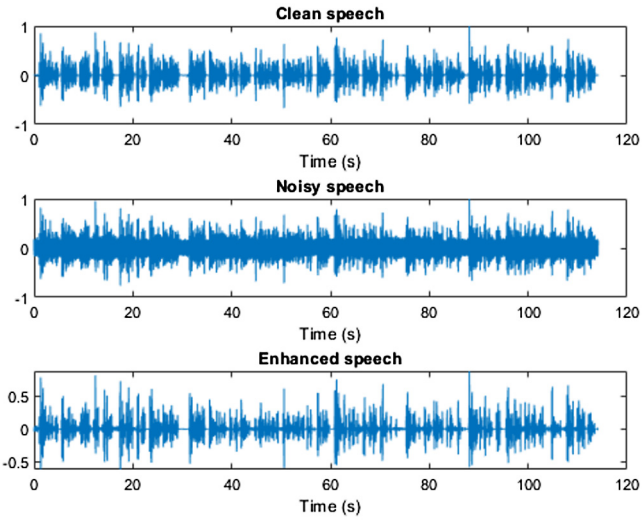

Fig. 4. Wave representation results of MWF scheme.

3.2. Modified Wiener filtration

Wiener Filter is known for frequency filtration approach and is not time variant there initially using FFT the speech converted into frequency then filtration was applied on the frequency converted signal. The conversion was applied equation (2) and is converted as

$$\hat{I}_{speech}(f) = I_{speech}(f) + n(f) \tag{8}$$

The multiplicative filtration approach converts the signal into frequency and filtering coefficients multiply the signal and is denoted as

$$I_{speech}(f) = \hat{I}_{speech}(f) * W(f) \tag{9}$$

The regular process includes identifying by estimating the coefficients as follows by accessing power spectral density (PSD)

$$w(f) = \frac{\left|I_{speech}(f)\right|^2}{\left|I_{speech}(f)\right|^2 + \left|\hat{I}_{speech}(f)\right|^2} \tag{10}$$

But this mode eliminates the portion of the original signal too and the reconstruction of the signal. Therefore equation (10) is modified by accessing the estimation of noise from the signal and eliminates the signal from the noised signals with respect to original will yield the exact signal compared with existing and the modified wiener coefficients acquired.

$$w(f) = \frac{\left|\hat{I}_{speech}(f)\right|^2 - \sigma^2}{\left|\hat{I}_{speech}(f)\right|^2} \tag{11}$$

Figs. 4 and 5 show the effects of speech enhancement before and after the MWF PSD speech de-noise method. The suggested approach can increase speech quality and intelligibility when compared to earlier stage techniques. The above figure shows the spectral form of speech.

4. Proposed models

4.1. Extended Kalman RESNET architecture

The residual filtering network is known as RESNET-50. It has neuron parameters accessible by Image data set with a minimum number of epochs and has a wide range of accuracy.

Fig. 6 illustrates the overall block diagram of EKF-RESNET architecture. ResNet filtering approaches are employed to create an effective EKF and EWF design, and modelling is used to improve the quality and intelligibility of noisy speech signals. The parameters are recovered from the noisy speech signal frames utilising the EKF on EWF to get the relevant elements

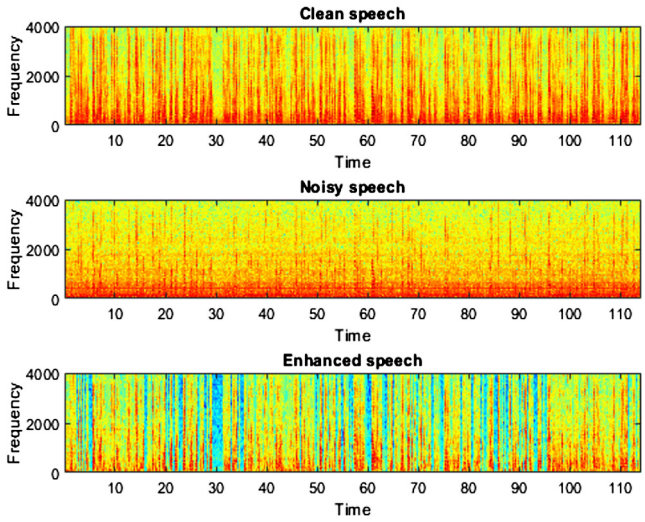

Fig. 5. Spectral waves for MWF scheme.

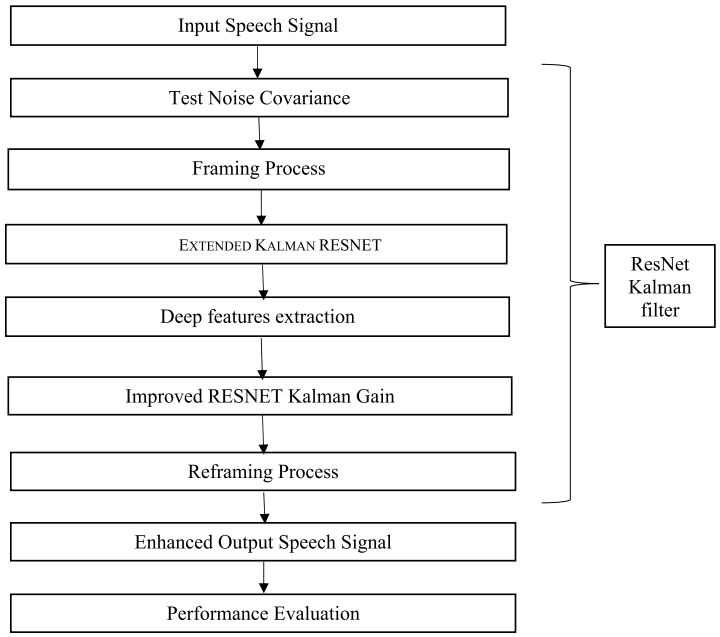

Fig. 6. Proposed EKF-ResNet block diagram.

required for speech denoising. The combination of EKF and ResNet allows for the creation of a customised E-Kalman gain value.

In this section, an effective EKF and EWF design is constructed using ResNet filtering methodologies, and modelling is used to improve the quality and intelligibility of noisy voice signals. The parameters are extracted from the noisy speech signal frames using the EKF on EWF to retrieve the essential parts necessary for denoising the speech signal. EKF with ResNet aids in attaining a tailored E-Kalman gain value. After the signal frames are rebuilt to generate a resulting speech signal, the tailored Kalman gain aids in attaining the appropriate noise reduction. The quality and intelligibility of the resulting voice signal has improved. For various additive noise levels, the performance of this approach is evaluated using three parameters: PESQ, STOI, and segmental SNR. When compared to the previous ones, the PESQ score improves by 4.2341 dB and 3.5422 dB in the 15 dB and 20 dB SNR ranges, respectively. The STOI score has increased by 0.0068 decibels. For the 10 dB and 20 dB SNR ranges, the Segmental SNR score is superior by 0.2515 dB and 0.562 dB, respectively. It exhibits a 1.5 percent improvement at lower SNR values. As a result, it can be argued that it is an effective strategy for improving speech quality and intelligibility. The E-Kalman ResNet filter is incapable of detecting large noise covariance. At the reconstruction

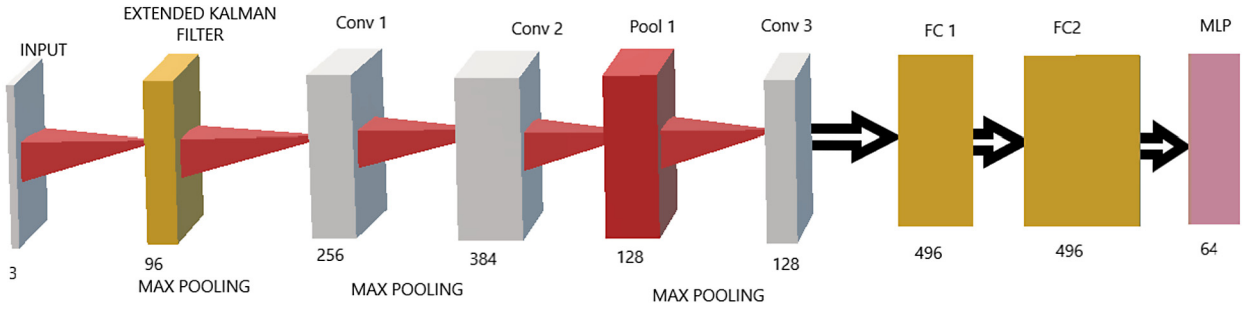

Fig. 7. Resnet transfer learning model.

Table 1
Resnet architecture operation.

Network layer	Resolution	Kernel size	Activation function	Drop out	Pooling layer	Output
Extended Kalman	1X256X1	3X3X1	ReLU	No	No	96
Conv2	117X117X3	3X3X1	ReLU	No	No	256
Conv3	64X64X3	3X3X1	ReLU	0.5	No	384
Pooling	32X32X3	3X3X1	ReLU	0.4	3X3X2	128
Conv5	32X32X3	3X3X1	ReLU	0.3	No	128
FC1	16X16X1		ReLU	0.3	No	496
FC2	8X8X1		SoftMax	0.3	No	496
SoftMax	64				No	64

step, high-quality encoding is necessary. The following restriction has been addressed in subsequent iterations, which will be explained in Fig. 6.

Fig. 7 illustrates the architecture of RESNET based transfer learning model. In RESNET 50, it uses filter from the range of 64 to 2048 that is based on the depth of the network and it utilizes the window size of 3*3.

In Fig. 7, Resnet model comprised of EKF Layer, three convolution layers, one pooling layer, two fully connected and one soft-max layer was present, and the architecture was driven ReLU activation function by accessing maximum point of the range. This model drops out gradient and tanh models to improve the training speed of the model and drops out the degree of overfitting.

Due to the significant spectral noise in bands in raw speech, de-noising is a little more complicated with this approach SegSNR, and the STOI score cannot be increased. With high density of noise, white noise, and spectral noise in speech samples, conventional speech de-noise models are ineffective. As a result, while speech de-noising is achievable, intelligibility and quality cannot be enhanced in terms of SegSNR and STOI scores. The CS frequency model is slightly better than previous models since high frequency noise is not reduced with filters or transformation techniques in this methodology. The wiener filter removes noise from speech signals, which is difficult to manage in this context due to the high density of noise. As a result, metrics like SegSNR and STOI haven't improved all that much. The LogMMSE approach is a speech de-noise model that works with amplitude estimation; however, it might degrade the SegSNR and STOI score owing to uncertainty in speech selection. The suggested approach, ResNet with Kalman filtering, overcomes the aforesaid restriction and effectively improves performance metrics.

In Table 1, Resnet operation is sequential with next stage connectivity and kernel function made to active. Here extended Kalman filter gets active in selection elimination of noise in speech signal portions that help training to understand the least level of detection, which enhances the chance of result prediction

Fig. 12 and Fig. 13 represent the results acquired using Resnet and graphs during the system's training phase and validation. In Figs. 3–12, it is evident that best-detected images were shown in a row factor with a 100-prediction rate. In Figs. 3–13, as the batch was considered for 0.6 and epoch, were tuned to below for a quick training process was represented that helps in the speed training process, and the SoftMax layer yields the result effectively. Gabor filter in the architecture tunes the results to be detected even in a blur.

In Fig. 8, acquired accuracy for the training and validation is given where the accuracy of training gradually increases over the validation accuracy across the epoch.

4.2. Modified WIENER RESNET architecture

The residual filtering network is known as RESNET-50. It has neuron parameters accessible by Image data set with a minimum number of epochs and has a wide range of accuracy.

In this section, ResNet filtering approaches are employed to build an effective EKF and EWF design, and modelling is applied to enhance the quality and intelligibility of noisy speech signals. The parameters are recovered from the noisy speech signal frames utilising the EKF on EWF to get the relevant elements required for speech denoising. The combination of EKF

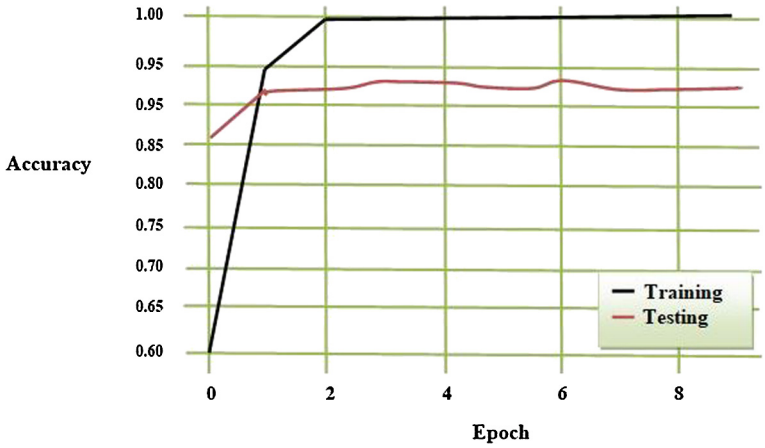

Fig. 8. Training and validation performance of Resnet.

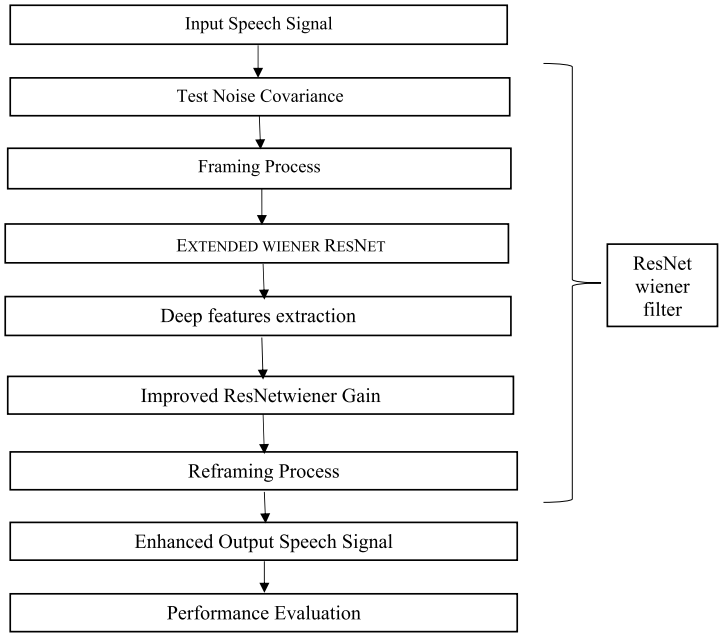

Fig. 9. Proposed EWF ResNet filtering.

and ResNet allows for the creation of a customised E-Kalman gain value. The adjusted Kalman gain assists in achieving the proper noise reduction when the signal frames are rebuilt to provide a resultant voice signal. The resultant speech signal has increased in quality and intelligibility. The performance of this technique is tested using three parameters: PESQ, STOI, and segmental SNR, for varied additive noise levels. In the 15 dB and 20 dB SNR levels, the PESQ score increases by 4.2341 dB and 3.5422 dB, respectively, when compared to the previous ones. By 0.0068 decibels, the STOI score has improved. The Segmental SNR score is 0.2515 dB and 0.562 dB higher in the 10 dB and 20 dB SNR bands, respectively. At lower SNR levels, it shows a 1.5 percent improvement. As a consequence, it may be claimed that it is a viable method for enhancing speech quality and comprehension. Large noise covariance is undetectable by the E-wiener ResNet filter. High-quality encoding is required throughout the reconstruction process. In future revisions, the following constraint was addressed, as will be discussed in the next chapters.

In Fig. 10, Resnet model comprised of Gabor Layer, three convolution layers, one pooling layer, two fully connected and one soft-max layer was present, and the architecture was driven ReLU activation function by accessing maximum point of the range. This model drops out gradient and tanh models to improve the training speed of the model and drops out the degree of overfitting.

In Table 2, Resnet operation is sequential with next stage connectivity and kernel function made to active. Here WKF gets active in selection denoised speech portions that help training to understand the least level of detection, which enhances the chance of result prediction shown in Fig. 11.

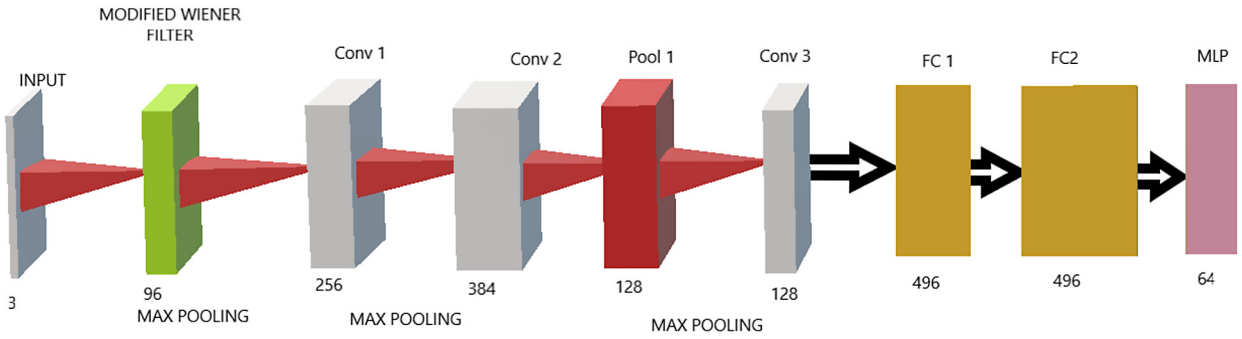

Fig. 10. Resnet transfer learning model.

Table 2
Resnet architecture operation.

Network layer	Resolution	Kernel size	Activation function	Drop out	Pooling layer	Output
Modified Wiener	117X117X3	3X3X1	ReLU	No	No	96
Conv2	117X117X3	3X3X1	ReLU	No	No	256
Conv3	64X64X3	3X3X1	ReLU	0.5	No	384
Pooling	32X32X3	3X3X1	ReLU	0.4	3X3X2	128
Conv5	32X32X3	3X3X1	ReLU	0.3	No	128
FC1	16X16X1		ReLU	0.3	No	496
FC2	8X8X1		SoftMax	0.3	No	496
SoftMax	64				No	64

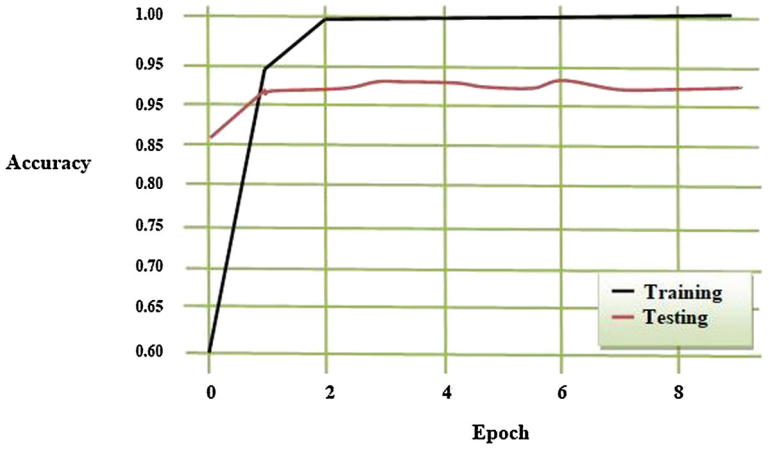

Fig. 11. Training and validation performance of Resnet.

5. Results

The overall performance metrics PESQ, LSD and SegSNR [1] was observed at 5 different noise conditions i.e., −15 dB, −10 dB, −5 dB, 0 dB, 5 dB and plotted below.

Fig. 12 and Fig. 13 represent the results acquired using Resnet and graphs during the system's training phase and validation. It is evident that best-detected images were shown in a row factor with 100 prediction rates. The batch normalization has been considered for 0.6 and epoch and were tuned to below for a quick training process was represented that helps in the speed training process, and the SoftMax layer yields the result effectively. Modified Wiener filter in the architecture tunes the results to be detected even in a blur.

Training on single CPU. Initializing image normalization.

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	1	00:00:03	75.00%	0.6928	0.0020
4	50	00:00:05	25.00%	0.7098	0.0020
7	100	00:00:07	25.00%	0.7172	0.0020

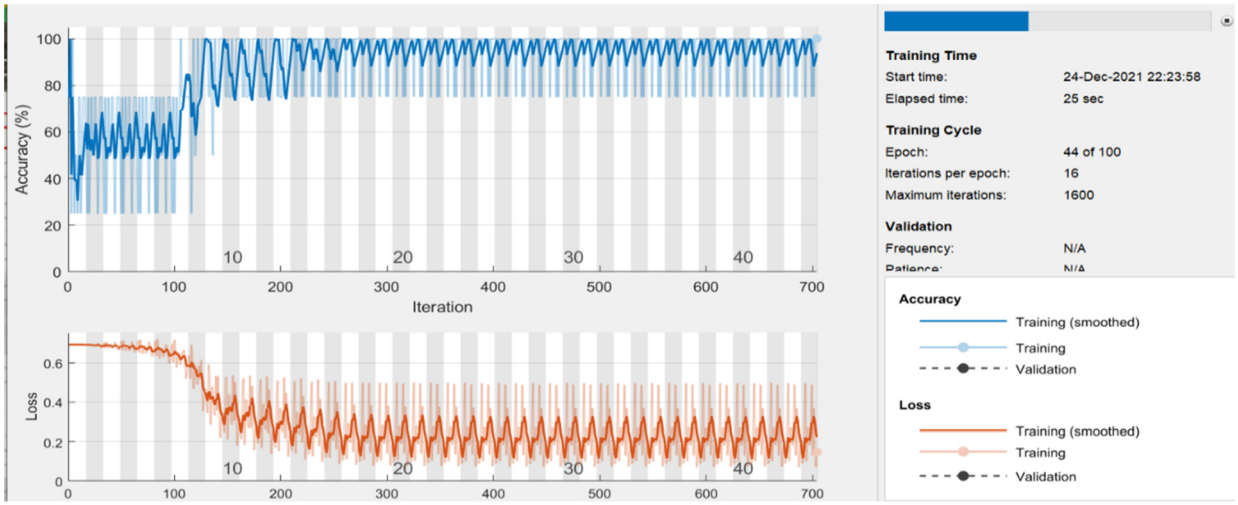

Fig. 12. Training process at 600 epochs.

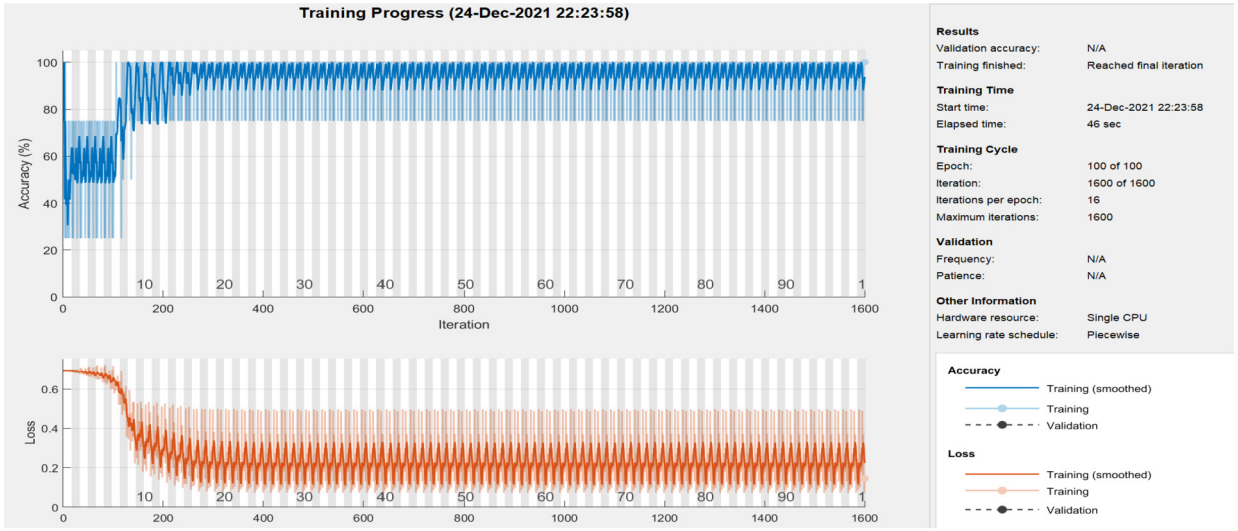

Fig. 13. Total Epochs training.

10	150	00:00:08	100.00%	0.2129	0.0002
13	200	00:00:09	75.00%	0.3641	0.0002
16	250	00:00:10	100.00%	0.2338	0.0002
19	300	00:00:11	100.00%	0.1419	2.0000e-05
22	350	00:00:12	100.00%	0.2004	2.0000e-05
25	400	00:00:14	100.00%	0.1482	2.0000e-06
29	450	00:00:15	100.00%	0.0885	2.0000e-06
32	500	00:00:20	100.00%	0.2881	2.0000e-06
35	550	00:00:21	100.00%	0.1012	2.0000e-07
38	600	00:00:23	100.00%	0.2705	2.0000e-07
41	650	00:00:24	100.00%	0.2213	2.0000e-08
44	700	00:00:25	100.00%	0.1314	2.0000e-08
47	750	00:00:26	100.00%	0.1967	2.0000e-08
50	800	00:00:27	100.00%	0.1471	2.0000e-09
54	850	00:00:29	100.00%	0.0880	2.0000e-09
57	900	00:00:30	100.00%	0.2877	2.0000e-10
60	950	00:00:31	100.00%	0.1011	2.0000e-10
63	1000	00:00:32	100.00%	0.2704	2.0000e-10

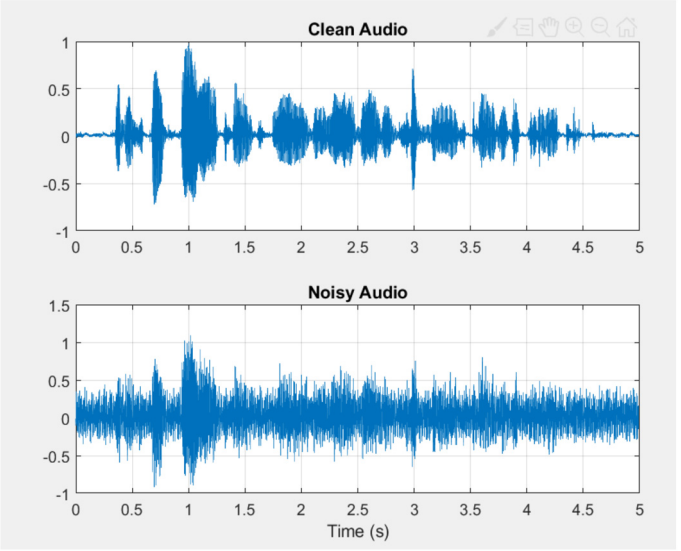

Fig. 14. Clean and noisy video.

```
|     66 |     1050 |     00:00:33 |     100.00% |     0.2213 |     2.0000e-11 |
|     69 |     1100 |     00:00:34 |     100.00% |     0.1314 |     2.0000e-11 |
|     72 |     1150 |     00:00:36 |     100.00% |     0.1967 |     2.0000e-11 |
|     75 |     1200 |     00:00:37 |     100.00% |     0.1471 |     2.0000e-12 |
|     79 |     1250 |     00:00:39 |     100.00% |     0.0880 |     2.0000e-12 |
|     82 |     1300 |     00:00:40 |     100.00% |     0.2877 |     2.0000e-13 |
|     85 |     1350 |     00:00:41 |     100.00% |     0.1011 |     2.0000e-13 |
|     88 |     1400 |     00:00:42 |     100.00% |     0.2704 |     2.0000e-13 |
|     91 |     1450 |     00:00:43 |     100.00% |     0.2213 |     2.0000e-14 |
|     94 |     1500 |     00:00:44 |     100.00% |     0.1314 |     2.0000e-14 |
|     97 |     1550 |     00:00:45 |     100.00% |     0.1967 |     2.0000e-15 |
|    100 |     1600 |     00:00:46 |     100.00% |     0.1471 |     2.0000e-15 |
=========================================================================================
PredictedLabels = categorical normal
c Elapsed time is 0.189601 seconds.
```

STOI is another measure that has a strong association with speech intelligibility and is used to assess the proposed system's intelligibility. The comparison of current approaches with the proposed method in terms of STOI score in (percent) for different noise kinds and SNR levels is shown in Fig. 14 clean and noisy video.

To conclude, the proposed method's PESQ scores vary from 2.256 to 3.368 for white noise, 2.284 to 3.352 for babbling noise, 2.395 to 3.374 for F16 noise, and 2.248 to 3.378 for industrial noise for different SNR levels in dB shown in Fig. 15. The suggested method's STOI score varies from 81.56 percent to 96.12 percent for white noise, 78.84 percent to 92.65 percent for babbling noise, 76.24 percent to 93.57 percent for F16 noise, and 76.23 percent to 94.21 percent for factory noise for different SNR levels in dB shown in Fig. 16. The suggested method's SSNR values vary from 12.28 to 17.26 dB for white noise, 11.94 to 16.88 dB for babbling noise, 11.83 to 17.01 dB for F16 noise, and 11.72 to 15.92 dB for industrial noise for different SNR levels in dB shown in Fig. 17.

The quality and intelligibility of the resulting voice signal has improved. For various additive noise levels, the performance of this approach is evaluated using three parameters: LSD, segmental SNR, and segmental SNR. When compared to the previous ones, the LSD score improves by 4.2341 dB and 3.5422 dB in the 15 dB and 20 dB SNR ranges, respectively. The LSD score has improved by 0.0068 decibels. For the 10 dB and 20 dB SNR ranges, the Segmental SNR score is superior by 0.2515 dB and 0.562 dB, respectively. It exhibits a 1.5 percent improvement at lower SNR values. As a result, it can be argued that it is an effective strategy for improving speech quality and intelligibility.

6. Conclusion

In this paper various modified deep learning schemes with filtration approaches were studied and implemented with comparing by replacing first convolution layer of original RESNET-50 architecture and modified the architecture using scratch code for implementation in MATLAB. The performance metrics represents best outcome for Weiner based deep learning

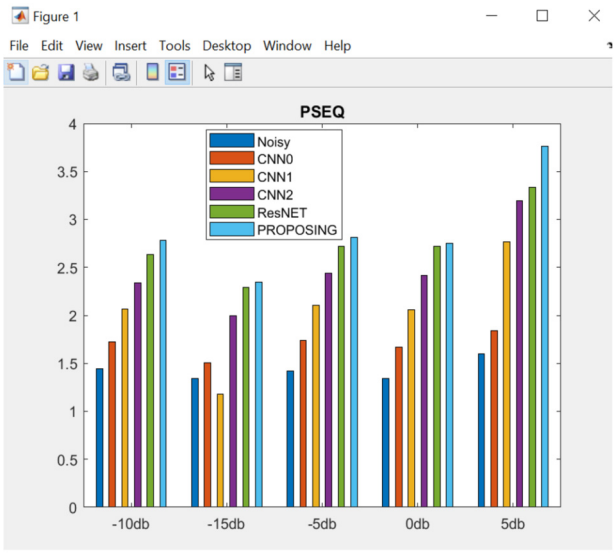

Fig. 15. Perceptual valuation of Speech excellence.

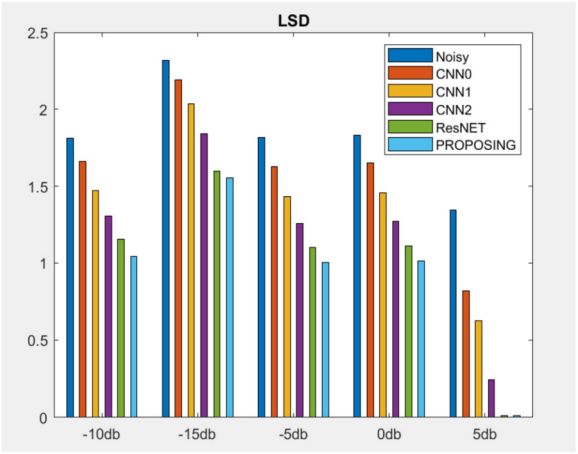

Fig. 16. Log Spectral Distance (LSD).

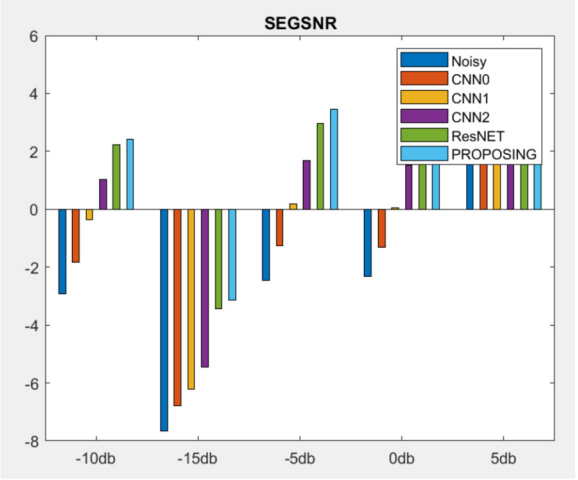

Fig. 17. SSNR representation (Segmental Signal-To-Noise Ratio).

filtration approach compared with other transfer learning, CNN and DNN approaches. Speech processing aids in the removal of background noise from communication signals. The augmentation system's primary goal is to improve the perceived quality of communication or speech. Various filtering algorithms, spectrum restorative models, and speech designs were used to achieve good improvement at de-noising. The accuracy of the approach is enhanced and the attained numerical outcome shows the effectiveness of the proposed scheme.

Ethics approval and consent to participate

No participation of humans takes place in this implementation process.

Human and animal rights

No violation of human and animal rights is involved.

Funding

No funding is involved in this work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Xue, G. Quan, C. Zhang, G. Ding, X. He, B. Zhou, Neural Kalman filtering for speech enhancement, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 7108–7112.
[2] Madam Aravind Kumar, Kamsali Manjunatha Chari, Noise reduction using modified Wiener filter in digital hearing aid for speech signal enhancement, J. Intell. Syst. 29 (1) (2020) 1360–1378, https://doi.org/10.1515/jisys-2017-0509.
[3] Y. Shi, W. Rong, N. Zheng, Speech enhancement using convolutional neural network with skip connections, in: 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei City, Taiwan, 2018, pp. 6–10.
[4] A. Kawamura, Y. Iiguni, Y. Itoh, A noise reduction method based on linear prediction with variable step-size, IEICE Trans. Fundam. Electron. Commun. Comput. Sci. E88-A (4) (2005) 855–861.
[5] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process. 27 (2) (1979) 113–120.
[6] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, L. Xie, DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement, arXiv preprint, arXiv:2008.00264, 2020.
[7] B. Widrow, J.G.R. Glover Jr., J.M. Mccool, et al., Adaptive noise cancelling: principles and applications, Proc. IEEE 63 (12) (1975) 1692–1716.
[8] P.J. Wolfe, S.J. Godsill, Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement, EURASIP J. Appl. Signal Process. 2003 (10) (2003) 1043–1051.
[9] K. Tan, D. Wang, A convolutional recurrent neural network for real-time speech enhancement, in: Interspeech, vol. 2018, September 2018, pp. 3229–3233.
[10] A. Karthik, J.L. MazherIqbal, Efficient speech enhancement using recurrent convolution encoder and decoder, Wirel. Pers. Commun. 119 (3) (2021) 1959–1973.
[11] R. Martin, Speech enhancement based on minimum mean-square error estimation and supergaussian priors, IEEE Trans. Speech Audio Process. 13 (5) (2005) 845–856.
[12] S. Gazor, W. Zhang, Speech enhancement employing Laplacian-Gaussian mixture, IEEE Trans. Speech Audio Process. 13 (5) (2005) 896–904.
[13] T. Lotter, P. Vary, Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model, EURASIP J. Appl. Signal Process. 2005 (7) (2005) 1110–1126.
[14] I. Andrianakis, P.R. White, Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors, Speech Commun. 51 (1) (2009) 1–14.
[15] J.H. Hansen, M.A. Clements, Constrained iterative speech enhancement with application to speech recognition, IEEE Trans. Signal Process. 39 (4) (1991) 795–805.
[16] A. Kawamura, W. Thanhikam, Y. Iiguni, A speech spectral estimator using adaptive speech probability density function, in: Proceedings of the EUSIPCO 2010, August 2010, pp. 1549–1552.
[17] S. Leglaive, L. Girin, R. Horaud, Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, May 2019, pp. 101–105.
[18] Y. Xu, J. Du, L.R. Dai, C.H. Lee, A regression approach to speech enhancement based on deep neural networks, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (1) (2014) 7–19.
[19] A. Adeel, M. Gogate, A. Hussain, W.M. Whitmer, Lip-reading driven deep learning approach for speech enhancement, IEEE Trans. Emerg. Top. Comput. Intell. (2019).
[20] S.K. Roy, A. Nicolson, K.K. Paliwal, A deep learning-based Kalman filter for speech enhancement, in: INTERSPEECH, October 2020, pp. 2692–2696.
[21] N. Koppula, K. Sarada, I. Patel, R. Aamani, K. Saikumar, Identification and recognition of speaker voice using a neural network-based algorithm: deep learning, in: Handbook of Research on Innovations and Applications of AI, IoT, and Cognitive Technologies, IGI Global, 2021, pp. 278–289.
[22] J.M. Valin, A hybrid DSP/deep learning approach to real-time full-band speech enhancement, in: 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), IEEE, August 2018, pp. 1–5.
[23] N. Saleem, M.I. Khattak, Deep neural networks for speech enhancement in complex-noisy environments, Int. J. Interact. Multim. Artif. Intell. 6 (1) (2020) 84–90.
[24] H. Yu, Z. Ouyang, W.P. Zhu, B. Champagne, Y. Ji, A deep neural network based Kalman filter for time domain speech enhancement, in: 2019 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, May 2019, pp. 1–5.

[25] M. Hasannezhad, H. Yu, W.P. Zhu, B. Champagne, PACDNN: a phase-aware composite deep neural network for speech enhancement, Speech Commun. 136 (2022) 1–13.
[26] H.S. Choi, J.H. Kim, J. Huh, A. Kim, J.W. Ha, K. Lee, Phase-aware speech enhancement with deep complex u-net, in: International Conference on Learning Representations, September 2018.
[27] C. Macartney, T. Weyde, Improved speech enhancement with the wave-u-net, arXiv preprint, arXiv:1811.11307, 2018.
[28] S. Chakrabarty, E.A. Habets, Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks, IEEE J. Sel. Top. Signal Process. 13 (4) (2019) 787–799.
[29] D. Yin, C. Luo, Z. Xiong, W. Zeng, Phasen: a phase-and-harmonics-aware speech enhancement network, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34(05), April 2020, pp. 9458–9465.
[30] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R.A. Saurous, R.F. Lyon, Exploring tradeoffs in models for low-latency speech enhancement, in: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), IEEE, September 2018, pp. 366–370.
[31] Z.Q. Wang, D. Wang, All-neural multi-channel speech enhancement, in: Interspeech, September 2018, pp. 3234–3238.
[32] T. Nakahara, K. Fukuyama, M. Hamada, K. Matsui, Y. Nakatoh, Y.O. Kato, J.M. Corchado, Mobile device-based speech enhancement system using lip-reading, in: International Symposium on Distributed Computing and Artificial Intelligence, Springer, Cham, June 2020, pp. 159–167.
[33] D. Michelsanti, Z.H. Tan, S. Sigurdsson, J. Jensen, Deep-learning-based audio-visual speech enhancement in presence of Lombard effect, Speech Commun. 115 (2019) 38–50.
[34] G. Yu, Y. Wang, H. Wang, Q. Zhang, C. Zheng, A two-stage complex network using cycle-consistent generative adversarial networks for speech enhancement, Speech Commun. 134 (2021) 42–54.
[35] T. Ajay, K.N. Reddy, D.A. Reddy, P.S. Kumar, K. Saikumar, Analysis on SAR signal processing for high-performance flexible system design using signal processing, in: 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, December 2021, pp. 30–34.