CAMES, **29**(1–2): 71–85, 2022, doi: 10.24423/cames.397

This article belongs to the *Special Issue on Advanced Optimization Methods for Uncertainties in Intelligent Industrial Systems* edited by Dr. I.J. Jacob, Dr. B.H. bin Ahmad and Dr. Z.F. Khan

Computational Intelligence for Speech Enhancement using Deep Neural Network

Hepsiba D.[1),2)*], Judith JUSTIN[1)]

[1)] *Department of Biomedical Instrumentation Engineering*
Avinashilingam Institute for Home Science and Higher Education for Women
Coimbatore, Tamil Nadu, India, e-mail: hod_bmie@avinuty.ac.in

[2)] *Department of Biomedical Engineering*
Karunya Institute of Technology and Sciences
Coimbatore, Tamil Nadu, India
*Corresponding Author e-mail: hepsiba@karunya.edu

In real time, the speech signal received contains noise produced in the background and reverberations. These disturbances reduce the quality of speech; therefore, it is important to eliminate the noise and increase the intelligibility and quality of speech signal. Speech enhancement is the primary task in any real-time application that handles speech signals. In the proposed method, the most effective and challenging noise, i.e., babble noise, is removed, and the clean speech is recovered. The enhancement of the corrupted speech signal is done by applying a deep neural network-based denoising algorithm in which the ideal ratio mask is used to mask the noisy speech and separate the clean speech signal. In the proposed system, the speech signal corrupted by noise is enhanced. Evaluation of enhanced speech signal by performance metrics such as short time objective intelligibility and signal to noise ratio of the denoised speech show that the speech intelligibility and speech quality are improved by the proposed method.

Keywords: deep neural network, noisy speech, speech enhancement, feature extraction, speech quality, computational intelligence.

1. Introduction

Speech enhancement [21, 22] is very important for any speech signal suffering from distortions, reflections and background noise that varies from place to place. Therefore, the speech enhancement techniques are crucial and very important for improving the speech quality in applications such as speaker recognition, automatic speech recognition (ASR) [1, 2, 23, 40], speech coding [5, 6, 32] and hearing aids [3, 4, 32].

Speech enhancement algorithms [12, 34] help in reducing noise without disturbing the quality of target speech. When the speech quality and intelligibility are improved, it helps the listeners to listen to the speech without any restraints. The conventional algorithms of speech enhancement include minimum mean square error [7], spectral subtraction [8], Kalman filtering [11] and iterative Wiener filtering [10].

In the recent past, computational intelligence and machine learning have found wide applications in enhancing distorted speech and noise removal [38, 39]. The latest trend in speech enhancement uses deep learning [9, 37], which adopts the architecture of a deep neural network (DNN). A DNN is a feed-forward network capable of modeling relationships that are non-linear [36]. In order to model the DNN, the features such as relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP) [14, 15], amplitude modulation spectrogram [30, 13], gammatone frequency cepstral coefficients (GFCC) [9, 16] and mel frequency cepstral coefficients (MFCC) [18] are extracted.

The training data consists of the speech signal with different noise types and signal-to-noise ratio (SNR) for the non-linear DNN-based regression model. The performance of the DNN is restricted to varying real-time noisy situations. Therefore, to improve the network's generalization ability, the changing nature of the noise is given as input to the network for training [31]. This helps to enhance the efficiency of the network in detecting unseen noise types.

In the past few years, a DNN has played a vital role in separating noise from speech and improving speech quality [19]. To enhance the noisy-reverberant speech [19], spectral mapping [17] is done using a single DNN, which removes noise and reverberation. Basically, the background noise causes disturbance to the clean speech. Here, denoising is performed for speech signal mixed with babble noise.

The content of the paper is arranged as follows: Sec. 2 gives the detailed description about the similar works carried out for speech enhancement, Sec. 3 discusses the methodology of speech enhancement. Section 4 explains the various feature extraction methodologies. Section 5 discusses a DNN for denoising and Sec. 6 presents the obtained results and their discussion; finally, the conclusion is given in the last section.

2. Related work

In the past years, the related work carried out for the speech enhancement has dealt with unsupervised techniques such as spectral subtraction, Kalman filtering, Weiner filtering and many more. The problem occurring in these techniques is that the method adopted for analyzing the noise is just an assumption. The disadvantages occurring in these techniques are eliminated by the powerful super-

vised technique such as codebook vectors and the model-based techniques where the speech signal and noise are known *a priori*. For the distortion-independent acoustic model, the non-matrix factorization (NMF) is more powerful in the process of separating the source in the recording made in a single-channel microphone in the existence of additive noise. The NMF-based technique [41] helps in estimating the speech signal and noise in the frequency domain. Segment-based approach [47] is another method to identify longer speech segments with its full-length speech sentence matching to remove fast-varying noise.

The previous research clearly indicates the improvement in speech enhancement performance when the features are extracted from the speech signal. The prediction of the log-power spectra (LPS) feature of the clean speech signal can be made using multi-objective learning. A long short-term memory (LSTM) technique [43] is a powerful tool that helps in a consistent improvement of the speech quality and intelligibility. The encoding of features [44] helps in the voice conversion process, and the different encoders are more effective. The usage of the deep recurrent neural network [42] is also very helpful in identifying the speech denoising system model in which the time-frequency masking is applied to one of the layers in the network.

Speech enhancements with deep learning are based on mapping or masking [45]. In the mapping-based enhancement, the relationship between the features of the noisy speech and the clean speech is considered. In the masking-based scenario, the relationship between the features of the noisy speech and the time-frequency mask is considered. The estimated mask is used to obtain the features of the enhanced speech signal. The different ideal masks for speech enhancement are ideal binary masks, ideal ratio masks and complex ideal ratio masks (cIRMs). The studies indicate that the ideal ratio mask leads to better results compared to the ideal binary mask. The cIRM [46] takes both the real and imaginary components for estimating the target.

Due to the non-linear relationship between the input and the target of the speech signal features, the networks with multiple layers and non-linear activation functions are more effective than shallow networks for the enhancement of speech signal. In certain applications, when the speech signal needs to be masked, babble noise is utilized for security purposes. In this paper, the ideal ratio mask is incorporated to obtain the enhanced speech features for denoising the speech signal affected by babble noise.

3. Speech enhancement methodology

The clean speech is mixed with the babble noise to form the noisy speech signal, and the features are extracted and given to the DNN, as shown in Fig. 1.

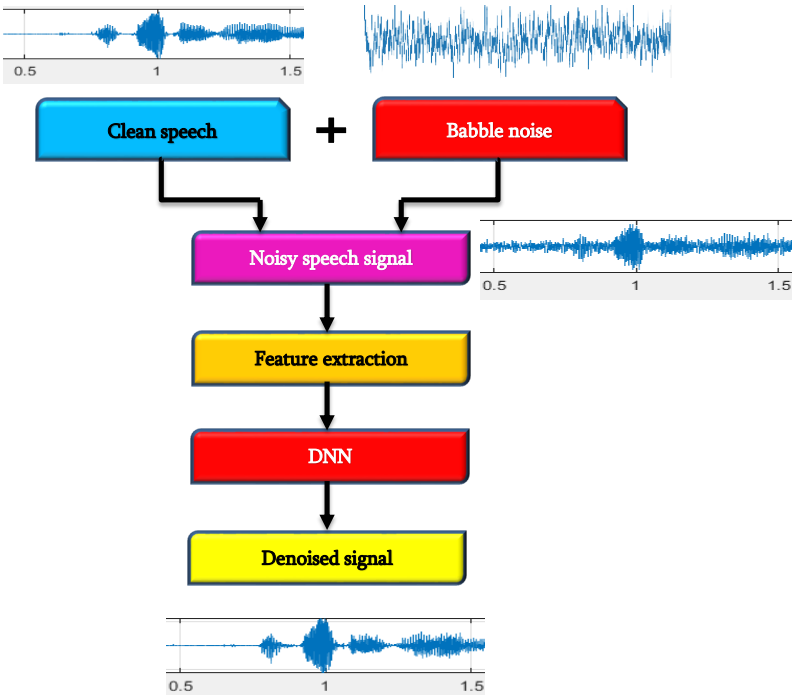

FIG. 1. Proposed speech enhancement system using a DNN.

3.1. Model of the speech signal

Let $c(t)$ and $b(t)$ represent the clean speech and babble noise, respectively. The noisy speech signal $n(t)$ is

$$n(t) = c(t) + b(t). \tag{1}$$

The babble noise signal $b(t)$ is usually not correlated with the desired signal $c(t)$; therefore, it is apparent that the noise can be removed first before recovering the clean speech. The target signal is the clean speech signal.

3.2. Process of denoising

The noisy utterance is given to the speech enhancement system, and the target signal is the noise-free clean speech. The noise is suppressed by using the time-frequency masking framework and removed by applying the time-frequency mask to the noisy speech signal. For the time-frequency masking, the ideal ratio mask is incorporated to remove the noise.

The ideal ratio mask is given by:

$$\mathrm{IRM}(t,f) = \left(\frac{C^2(t,f)}{C^2(t,f)+N^2(t,f)}\right)^{\beta}, \tag{2}$$

$$\mathrm{IRM}(t,f) = \left(\frac{\mathrm{SNR}(t,f)}{\mathrm{SNR}(t,f)+1}\right)^{\beta}, \tag{3}$$

where $C^2(t,\ f)$ shows the speech signal and $N^2(t,\ f)$ shows the noise signal, as a time-frequency (T-F) representation, and β acts as the tuning parameter for scaling the mask. At $\beta = 0.7$, the noisy signal is estimated and implemented using a DNN. After denoising, the signal is reconstructed in the time domain.

4. Feature extraction

The features extracted from the noisy speech signal are given below.

4.1. Mel frequency cepstral coefficients (MFCC)

The MFCC is the commonly used method in the feature extraction of speech signals. The speech signal is segmented into small duration blocks (windowed frames) and the fast Fourier transform (FFT) is applied to each frame sequence.

The signal is changed from the time domain signal into the frequency domain. The mel filter bank is applied to the power spectrum and energy is summed for all filter banks. The log filter bank energies are applied with a discrete cosine transform (DCT) [18, 29] to obtain the MFCC.

4.2. Relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP)

RASTA-PLP is a special methodology that implements band-pass filtering to the energy in each frequency sub-band. The high-pass filter portion in the band-pass filter reduces the convolutional noise [33]. The frame-to-frame spectral changes are smoothened by the low pass portion [15].

4.3. Amplitude modulation spectrogram (AMS)

The speech signal is converted to the frequency domain by applying a short-time Fourier transform (STFT). After decomposing the signal by the bark scale decomposition, the spectral analysis is made by a second STFT. Thus, the amplitude modulation coefficients such as acoustic frequencies, time and modulation frequencies are obtained [25].

4.4. Gammatone filter bank power spectra

The input speech signal is passed through a 64 channel gammatone filter bank [24]. In each channel, the filter response is fully rectified and decimated, which is similar to windowing. The absolute values taken specify the T-F representation. The cube root of the T-F representation is taken and the DCT is applied to the cepstral coefficients [16].

4.5. Autoregressive moving average model (ARMA)

The input speech signal is taken as long segments and converted using the DCT. The windowing function is applied to the DCT signal. The ARMA modeling is applied to sub-band DCT components of the sub-band envelope.

The power spectrum estimate is yielded by integrating the sub-band envelope with respect to time. The inverse fast Fourier transform (IFFT) is used to transform the power spectrum estimates into temporal autocorrelation estimates and further used based on linear prediction in the time domain. The output obtained gives a spectrally smoothed ARMA spectrogram [19].

5. DNN for denoising

The architecture of a DNN is a feed-forward neural network [9] and has the competence to map the features of the noisy speech signal to clean the speech signal [33]. The DNN model is trained with the features extracted [9]. Figure 2 shows the DNN architecture of the proposed model.

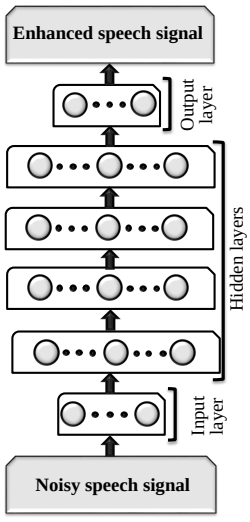

Fig. 2. The DNN architecture of the proposed model.

The sentence list is taken from the IEEE sentence database [26]. The clean and noisy audio file is taken from the Noizeus website [27]. A babble noise at 0 dB noise level is considered for mixing with the clean speech. The features extracted are 31-dimensional MFCC, 15-dimensional AMS, 64-dimensional gammatone filter bank power spectra, and 13-dimensional RASTA-PLP [35] and are taken as inputs to the DNN [19]. The DNN uses multilayer perceptron (MLP) as the discriminative learning machine, which shows good performance for speech separation. The DNN uses 4 hidden layers, each layer having 1024 rectified linear hidden units (ReLU). The number of hidden layers is taken as 4 in the process of tuning the hyperparameters as it reduces the MSE to 0.001. The network is trained with the back-propagation algorithm and the dropout rate considered is 0.2.

For the first 5 epochs, the momentum value is taken as 0.5, and after 5 epochs it is taken as 0.9. The increase of momentum rate from 0.5 to 0.9 does not fasten the training of the model, but it helps to increase the accuracy in training and testing the model. The DNN predicts the output for varying frequency ranges, and the cost function adopted is the mean squared error (MSE) [9].

For the targets in the range [0,1], the output layer uses a sigmoid activation function, and for the other layers, a linear activation function is employed. The input data given to the DNN are the features obtained from the 5-frame window for implementing the temporal context. The final estimate is obtained by finding the average of the multiple estimates of each frame [20].

6. Results and discussion

The proposed system uses sentences from the IEEE sentence database. Audio files are taken from the Noizeus website for the clean speech. The noise used for this work is a babble noise, which is the most challenging and it is considered to be the best noise for masking speech. The babble speech is generally the voice heard in the midst of the crowded ambience. The mixtures are obtained by mixing clean speech signal with babble noise with different SNR values.

The training data contains 600 sentences and the testing data consists of 120 sentences. The signal is sampled at 16 kHz and converted into frames using a 20 ms Hamming window with a 10 ms window shift for framing. For each frame, 320 frame FFT is applied, resulting in 161 frequency bins.

The SNR of the noisy speech signal shows that the noise power is greater than the signal power. After applying the denoising algorithm, the SNR is improved, which shows that the signal power has increased more than the noise power, as shown in Table 1. The noisy speech signal in the time domain, its periodogram and spectrogram are shown in Fig. 3. The spectrogram shows the intensity of noise present in the noisy speech signal. The periodogram displays the spectral density of the noisy speech signal.

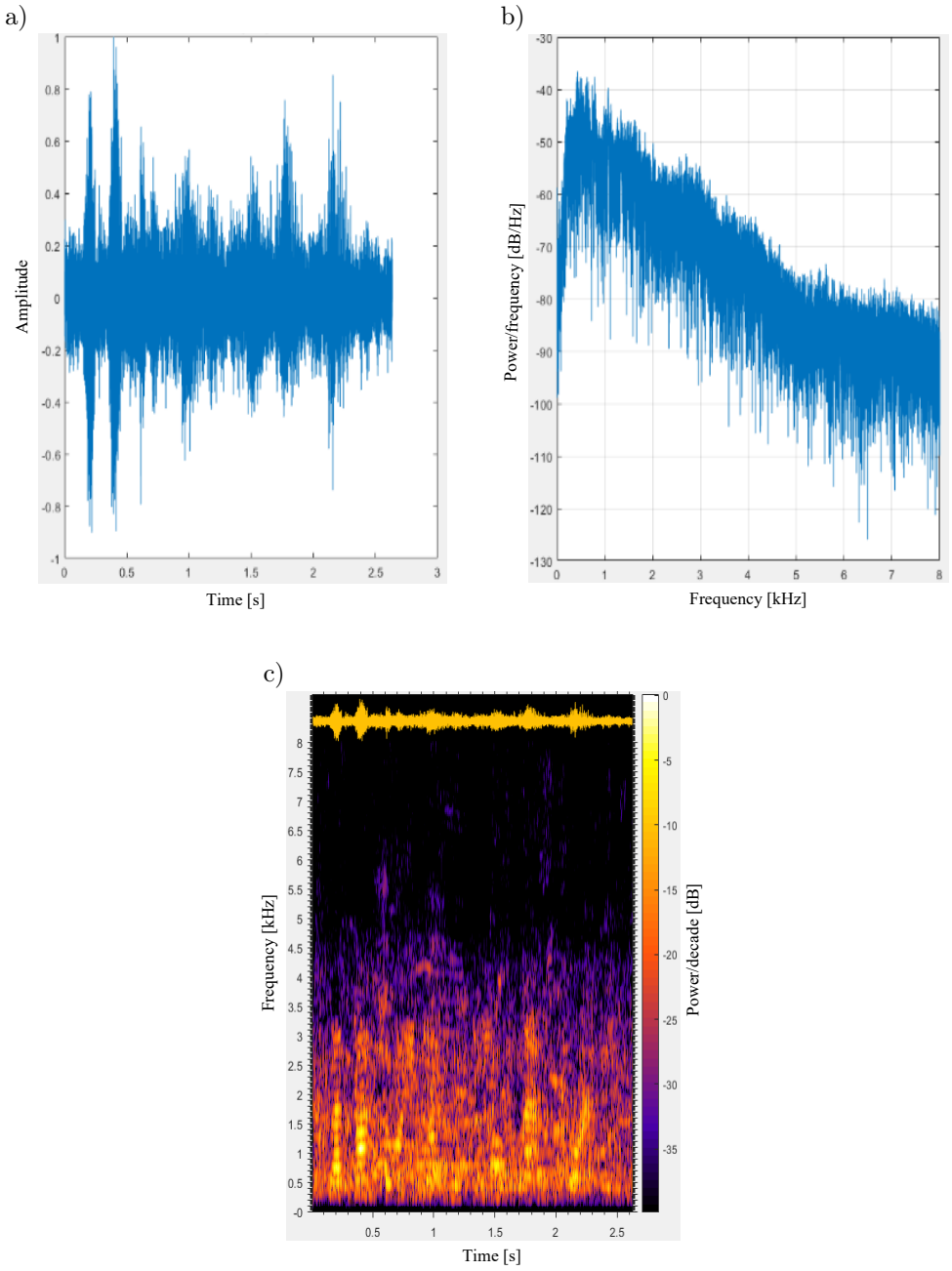

FIG. 3. Noisy speech: a) time domain, b) periodogram, and c) spectrogram.

After applying the DNN speech enhancement algorithm, the noise is removed, which can be observed in Fig. 4 that shows the denoised speech signal with

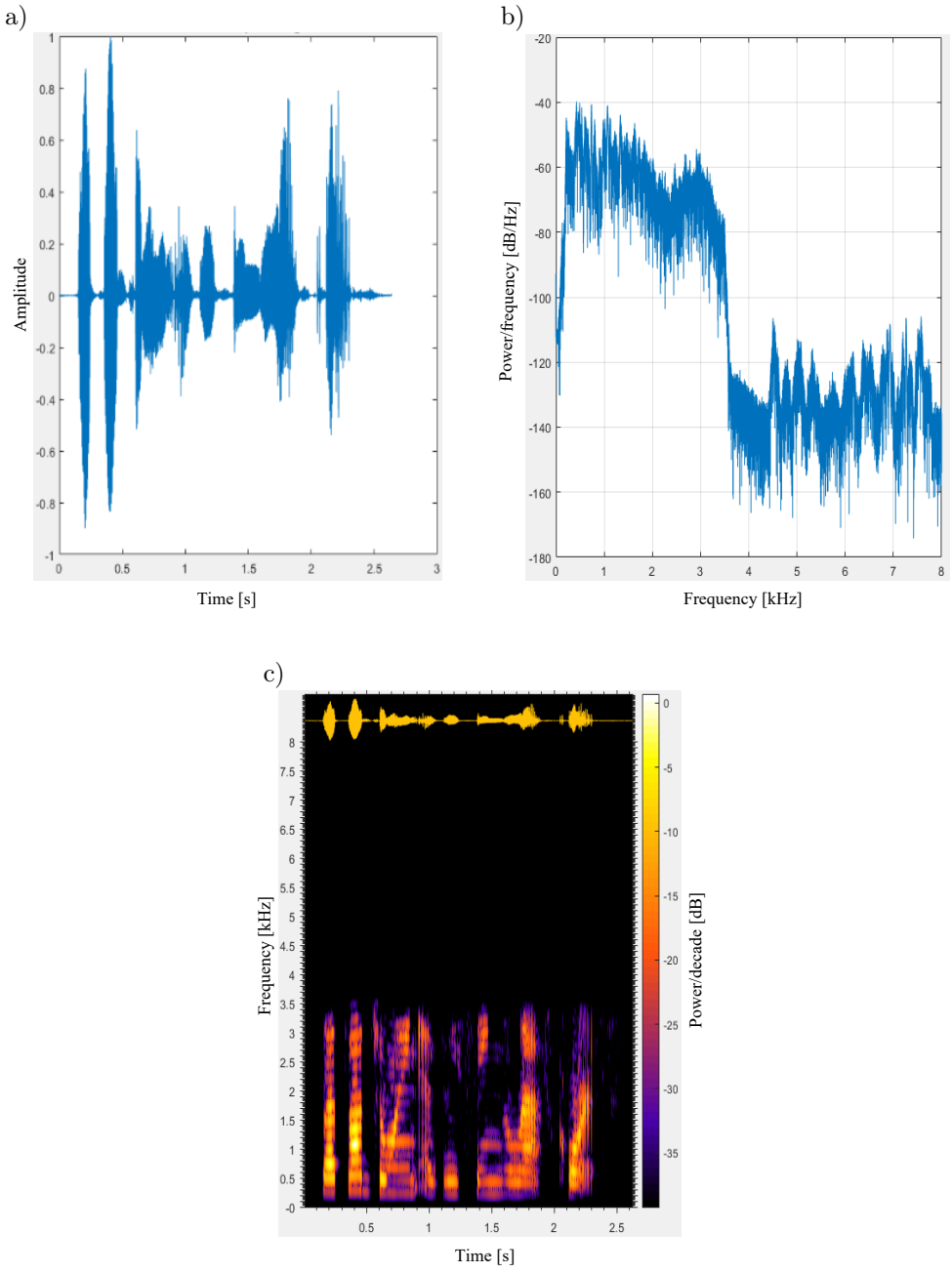

FIG. 4. Denoised speech: a) time domain, b) periodogram, and c) spectrogram.

its periodogram and spectrogram. Normalization of data helps to estimate the values between the minimum and maximum values so that it can be accessed

on a common scale. For the DNN training, the normalization of the features of the input speech signal is adjusted to zero mean and unit variance. All speech sentences are trained with the rectified linear unit. The denoised speech signal is tested for its performance based on the metrics such as the SNR and short-time objective intelligibility (STOI). STOI represents the similarity between reference and processed signal temporal envelopes for a short interval of time. STOI values are between 0 and 1, the higher values in this range indicate better intelligibility, as shown in Table 1.

Noise is removed from the noisy speech signal, and improved SNR and STOI values are shown in Table 1. Compared to the other methods adopted [48] for denoising the babble noise, the SNR is improved with this MLP DNN denoising model. The performance metrics are improved compared to the similar works adopted in speech enhancement. As the denoising system performs well for the babble noise, which is a non-stationary noise, the same methodology can be adapted for speech signals subjected to other noises for speech enhancement.

Table 1. SNR and STOI values of test sentences before and after denoising.

Input data	SNR before denoising [dB]	SNR after denoising [dB]	STOI before denoising	STOI after denoising
Test sentence 1	22.8878	27.0385	0.4915	0.6450
Test sentence 2	21.9277	27.0235	0.4014	0.5429
Test sentence 3	21.2534	26.9404	0.5544	0.5961
Test sentence 4	21.2279	27.0470	0.2677	0.4882
Test sentence 5	22.0107	27.0312	0.3192	0.5652
Test sentence 6	22.1754	26.9361	0.4983	0.6318
Test sentence 7	20.2182	26.7782	0.5444	0.6291
Test sentence 8	21.9834	26.9388	0.5175	0.6325
Test sentence 9	21.0655	27.0799	0.4337	0.6566
Test sentence 10	22.1023	27.4634	0.3488	0.5488

Figures 5 and 6 show the improvement of the denoised signal in terms of noise removal and intelligibility. The SNR of the denoised signal is more increased compared to the noisy signal, and the intelligibility of the speech signal indicates the increase in the clarity of the speech signal [48]. The denoising is very clearly observed when the denoised speech signal is listened as an audio output. The quality, as well as the intelligibility, is improved to a great extent. which shows the capability of the DNN in denoising the noisy speech and delivering the denoised signal equivalent to the clean speech signal.

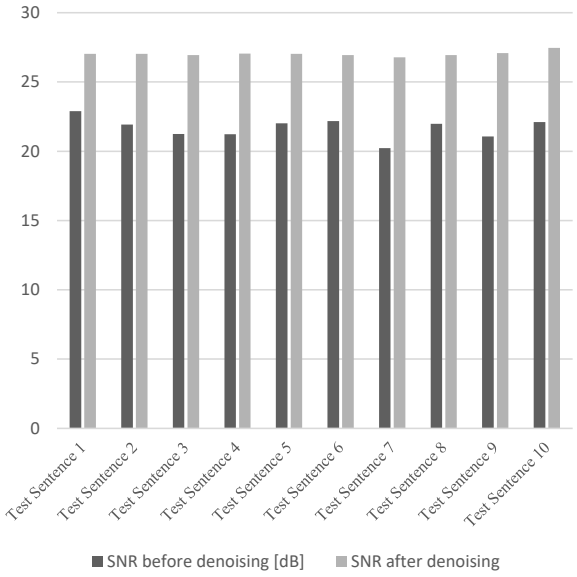

Fig. 5. The SNR of the denoised signal.

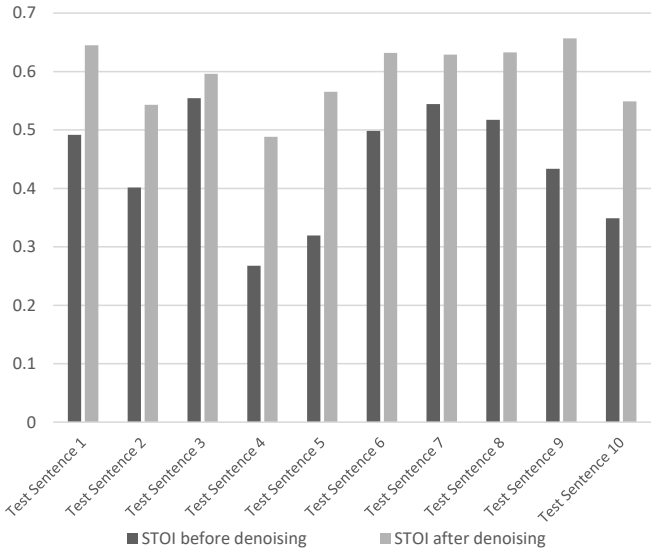

Fig. 6. STOI of noisy signal vs. denoised signal.

7. Conclusion

Background noise plays a major role in distorting the speech signal. The estimation of the ideal ratio mask yielded good results in estimating the noise and giving the denoised speech. The performance metrics such as SNR and STOI were

chosen to analyze the speech quality and intelligibility. The evaluations obtained in the performance metrics, STOI and SNR showed that the IRM-based deep learning algorithm excellently denoises the noisy speech signal and retrieves clean speech. The observations from the spectrogram also clearly indicate the removal of noise and display the denoised speech. Thus, the enhancement of speech signal was observed in the numerical values of the two-performance metrics.

References

1. J. Li, L. Deng, R. Haeb-Umbach, Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, 1st ed., Academic, Orlando, FL, USA, 2015.
2. B. Li, Y. Tsao, K.C. Sim, An investigation of spectral restoration algorithms for deep neural networks-based noise robust speech recognition, [in:] *Proceedings of Interspeech*, Lyon, France, pp. 3002–3006, 2013.
3. H. Levitt, Noise reduction in hearing aids: An overview, *Journal of Rehabilitation Research and Development*, **38**(1), 111–121, 2001.
4. A. Chern, Y.-H. Lai, Y.-P. Chang, Y. Tsao, R.Y. Chang, H.-W. Chang, A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom, *IEEE Access*, **5**: 10339–10351, 2017, doi: 10.1109/ACCESS.2017.2711489.
5. J. Li, L. Yang, J. Zhang, Y. Yan, Comparative intelligibility investigation of single-channel noise reduction algorithms for Chinese, Japanese and English, *Journal of the Acoustical Society of America*, **129**(5): 3291–3301, 2011, doi: 10.1121/1.3571422.
6. J. Li, S. Sakamoto, S. Hongo, M. Akagi, Y. Suzuki, Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication, *Speech Communication*, **53**(5): 677–689, 2011, doi: 10.1016/j.specom.2010.04.009.
7. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **33**(2): 443–445, 1985, doi: 10.1109/TASSP.1985.1164550.
8. S. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27**(2): 113–120, Apr. 1979, doi: 10.1109/TASSP.1979.1163209.
9. Hepsiba D., J. Justin, Role of deep neural network in speech enhancement: A review, [in:] J. Hemanth, T. Silva, A. Karunananda [Eds.], *Artificial Intelligence*, SLAAI-ICAI 2018. Communications in Computer and Information Science, Vol. 890, Springer, Singapore, 2019, doi: 10.1007/978-981-13-9129-3_8.
10. P. Scalart, J.V. Filho, speech enhancement based on a priori signal to noise estimation, [in:] *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 2, pp. 629–633, 1996, doi: 10.1109/ICASSP.1996.543199.
11. W. Xue, A.H. Moore, M. Brookes, P.A. Naylor, Modulation-domain multichannel Kalman filtering for speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(10): 1833–1847, 2018, doi: 10.1109/TASLP.2018.2845665.
12. J. Du, Q. Huo, A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions, [in:] *Proceedings of Interspeech*, pp. 569–572, Brisbane, Australia, 2008.

13. B. Kollmeier, R. Koch, Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction, *The Journal of the Acoustical Society of America*, **95**(3): 1593–1602, 1994, doi: 10.1121/1.408546.

14. H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America*, **87**(4): 1738–1752, 1990, doi: 10.1121/1.399423.

15. H. Hermansky, N. Morgan, RASTA processing of speech, *IEEE Transactions on Speech and Audio Processing*, **2**(4): 578–589, 1994, doi: 10.1109/89.326616.

16. T. Dau, D. Püschel, A quantitative model of the "effective" signal processing in the auditory system, *The Journal of the Acoustical Society of America*, **99**(6): 3615–3622, 1996, doi: 10.1121/1.414959.

17. K. Han, Y. Wang, D.L. Wang, W.S. Woods, I. Merks, T. Zhang, Learning spectral mapping for speech dereverberation and denoising, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(6): 982–992, 2015, doi: 10.1109/TASLP.2015.2416653.

18. S. Davis, P. Mermelstein, Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28**(4): 357–366, 1980, doi: 10.1109/TASSP.1980.1163420.

19. Y. Zhao, Z.-Q. Wang, D.L. Wang, Two-stage deep learning for noisy-reverberant speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(1): 53–62, 2019, doi: 10.1109/TASLP.2018.2870725.

20. Y. Wang, A. Narayanan, D.L. Wang, On training targets for supervised speech separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**(12): 1849–1858, 2014, doi: 10.1109/TASLP.2014.2352935.

21. J. Benesty, S. Makino, J.D. Chen, *Speech Enhancement*, Springer, New York, NY, USA, 2005.

22. P.C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, FL, USA, 2013, doi: 10.1201/9781420015836.

23. H.-Y. Lee, J.-W. Cho, M. Kim, H.-M. Park, DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition, *IEEE Signal Processing Letters*, **23**(8): 1091–1095, August 2016, doi: 10.1109/LSP.2016.2583658.

24. Y. Shao, S. Srinivasan, D.L. Wang, Incorporating auditory feature uncertainties in robust speaker identification, [in:] *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007*, pp. 277–280, 2007.

25. Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, A regression approach to speech enhancement based on deep neural networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(1): 7–19, 2015, doi: 10.1109/TASLP.2014.2364452.

26. IEEE, IEEE recommended practice for speech quality measurements, *IEEE Transactions on Audio and Electroacoustics*, **17**: 225–246, 1969.

27. Y. Hu, P. Loizou, Subjective evaluation and comparison of speech enhancement algorithms, *Speech Communication*, 2007, **49**: 588–601, https://ecs.utdallas.edu/loizou/speech/noizeus/.

28. K. Tan, D. Wang, Towards model compression for deep learning based speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**: 1785–1794, 2021, doi: 10.1109/TASLP.2021.3082282.

29. F. Bao, W. Abdulla, A new ratio mask representation for CASA-based speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(1): 7–19, 2018, doi: 10.1109/TASLP.2018.2868407.

30. Y. Liu, H. Zhang, X. Zhang, L. Yang, Supervised speech enhancement with real spectrum approximation, [in:] Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5746–5750, 2019, doi: 10.1109/ICASSP.2019.8683691.

31. C. Valentini-Botinhao, J. Yamagishi, Speech enhancement of noisy and reverberant speech for text-to-speech, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(8): 1420–1433, 2018, doi: 10.1109/TASLP.2018.2828980.

32. J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, H.-M. Wang, Audio-visual speech enhancement using multimodal deep convolutional neural networks, *IEEE Transactions on Emerging Topics in Computational Intelligence*, **2**(20): 117–128, 2018, doi: 10.1109/TETCI.2017.2784878.

33. P. Pujol, S. Pol, C. Nadeu, A. Hagen, H. Bourlard, Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system, *IEEE Transactions on Speech and Audio Processing*, **13**(1): 14–22, 2005, doi: 10.1109/TSA.2004.834466.

34. Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, Cross-language transfer learning for deep neural network-based speech enhancement, [in:] *Proceedings of the 9th International Symposium on Chinese Spoken Language Processing*, pp. 336–340, 2014, doi: 10.1109/ISCSLP.2014.6936608.

35. Z.-Q. Wang, D.L. Wang, Robust speech recognition from ratio masks, [in:] *Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5720–5724, 2016, doi: 10.1109/ICASSP.2016.7472773.

36. W. Yuan, A time–frequency smoothing neural network for speech enhancement, *Speech Communications*, **124**: 75–84, 2020, doi: 10.1016/j.specom.2020.09.002.

37. T. Lavanya, T. Nagarajan, P. Vijayalakshmi, Multi-level single channel speech enhancement using a unified framework for estimating magnitude and phase spectra, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**: 1315–1327, 2020, doi: 10.1109/TASLP.2020.2986877.

38. K. Sekiguchi, Y. Bando, A.A. Nugraha, K. Yoshii, T. Kawahara, Semi-supervised multichannel speech enhancement with a deep speech prior, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(12): 2197–2212, 2019, doi: 10.1109/TASLP.2019.2944348.

39. F.B. Gelderblom, T.V. Tronstad, E.M. Viggen, Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(3): 583–594, 2020, doi: 10.1109/TASLP.2018.2882738.

40. T. Kawase, M. Okamoto, T. Fukutomi, Y. Takahashi, Speech enhancement parameter adjustment to maximize accuracy of automatic speech recognition, *IEEE Transactions on Consumer Electronics*, **66**(2): 125–133, 2020, doi: 10.1109/TCE.2020.2986003.

41. D. Baby, T. Viratanen, J.F. Gemmeke, H. van Hamme, Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition, *IEEE/ACM Transactions*

on Audio, Speech, and Language Processing, **23**(11): 1788–1799, 2015, doi: 10.1109/TASLP.2015.2450491.

42. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Joint optimization of masks and deep recurrent neural networks for monaural source separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(12): 2136–2147, 2015, doi: 10.1109/TASLP.2015.2468583.

43. L. Sun, J. Du, L.-R. Dai, C.-H. Lee, Multiple-target deep learning for LSTM-RNN based speech enhancement, [in:] *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pp. 136–140, 2017, doi: 10.1109/HSCMA.2017.7895577.

44. W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, H.-M. Wang, Voice conversion based on cross-domain features using variational auto encoders, [in:] *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 51–55, 2018, doi: 10.1109/ISCSLP.2018.8706604.

45. W. Han, C. Wu, X. Zhang, Q. Zhang, S. Bai, Joint optimization of modified ideal ratio mask and deep neural networks for monaural speech enhancement, [in:] *Proceedings of 2017 9th International Conference on Communication Software and Networks (ICCSN)*, pp. 1070–1074, 2017, doi: 10.1109/ICCSN.2017.8230275.

46. D.S. Williamson, Y. Wang, D.L. Wang, Complex ratio masking for monaural speech separation,*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(3): 483–492, 2016, doi: 10.1109/TASLP.2015.2512042.

47. J. Ming, D. Crookes, Speech enhancement based on full-sentence correlation and clean speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25**(3): 531–543, 2017, doi: 10.1109/TASLP.2017.2651406.

48. R. Jaiswal, D. Romero, Implicit Wiener filtering for speech enhancement in non-stationary noise, [in:] *2021 11th International Conference on Information Science and Technology (ICIST)*, pp. 39–47, 2021, doi: 10.1109/ICIST52614.2021.9440639.

Received September 29, 2021; revised version December 15, 2021; accepted December 27, 2021.