A Hybrid Approach for Single Channel Speech Enhancement using Deep Neural Network and Harmonic Regeneration Noise Reduction

Norezmi Jamal[1], N. Fuad[2], MNAH. Sha'abani[3]
Faculty of Electrical and Electronic Engineering
Universiti Tun Hussein Onn Malaysia
Johor, Malaysia

***Abstract*—This paper presents a hybrid approach for single channel speech enhancement using deep neural network (DNN) and harmonic regeneration noise reduction (HRNR). The DNN was used as a supervised algorithm to predict new target mask such as constrained Wiener Filter (cWF) target mask from noisy mixture signal that was transformed into gammatone filter bank features. Meanwhile, HRNR algorithm was applied in the post-filtering strategy to eliminate residual noise. The DNN algorithm is an emerging supervised speech enhancement to overcome heavy nonstationary noise and low signal-to-noise ratio (SNR) issues. To validate the proposed algorithm with new target mask, 600 Malay utterances combining male and female speakers were used in a training session while 120 Malay utterances were used in a prediction session. The short time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) scores were calculated as the performance metrics. In this work, the proposed target mask outperformed other baseline target masks. Thus, PESQ and STOI scores for the hybrid speech enhancement algorithm is 1.17 and 0.79, respectively, at - 5 dB babble noise SNR.**

Keywords*—Speech enhancement; single channel microphone; deep neural network; constrained Wiener Filter; post-filtering

I. Introduction

Over the past few decades, automatic speech recognition (ASR) has gained a surge of interest among researchers in the speech processing research area. This is because ASR is widely used in mobile device applications, especially for navigation purposes. It allows human to talk to a computer, which responds to the given command. Basically, to capture human speech signals, a single microphone is practical enough to be used as an input device for speech data acquisition. Meanwhile, computer acts as a system to process the captured speech signal to realize human computer interaction (HCI) by applying mathematical algorithms and signal processing techniques. However, as the captured speech signals are in mobile remote condition [1], the speech signals could be easily contaminated by background noises, which may lead to complex computation and further processes. Thus, speech recognition may yield low speech recognition accuracy especially at low signal-to-noise ratio (SNR) and nonstationary noise, which the quality and intelligibility of speech are affected [2]. Thus, past studies proposed the Deep Neural Network (DNN)-based mask estimation approach to guarantee high intelligibility and quality of speech signal [3-9].

The first DNN-based mask estimation approach was pioneered by Wang, who demonstrated a DNN with promising results, owns good scalability and flexibility compared to other machine learning techniques [10], especially at high SNR. As mentioned in a study [4], DNN outperformed SVM algorithm during speech separation from noisy background as cited by [11]. The DNN algorithm became a great attention among researchers and most of them were interested to extend their work to improve in terms of features, target masks, and learning network [7, 12-15]. But when the mixture signal is at low SNR [10], the estimation between speech and noise signal is very challenging due to overlapping events. Excessive estimation of noise will cause speech distortion. Otherwise, when noise signal is underestimated, residual noise could be introduced after speech reconstruction [16]. For the first time, this study aims to propose a new target mask that could control speech distortion and noise distortion by using constrained Wiener Filter (cWF) in gammatone representation as the target mask for DNN algorithm with longer duration of speech utterance. The harmonic regeneration noise reduction (HRNR) was applied to reduce residual noise and generate new speech harmonic after the speech reconstruction.

This paper is organized as follows: Section II discusses the related works; Section III describes the methodology of research work; Section IV presents the experimental results and compares the results with the baseline approach and Section V summarizes the findings and proposes recommendations for improvement.

II. Related Works

Features, target masks and learning network are the three important elements in the DNN-based mask estimation approach. Previously, several studies on the effect of different features in a supervised DNN-based mask estimation were done [14, 15]. The findings revealed that gammatone features is one of the outperform features. Various speech dataset such as IEEE [10, 14, 15, 17] and TIMIT [18] were used to analyze the performance, which normally in short duration of speech utterance. A study in [10] also compared the training targets in DNN algorithm. From the results, the combination of features such as mel-frequency cepstral coefficient (MFCC), relative spectral transform perceptual linear prediction coefficient (RASTA- PLP), amplitude modulation spectrogram (AMS) and gammatone filter bank power spectra (GF) outperformed

with ideal ratio mask (IRM), which produced the STOI and PESQ scores of 0.72 and 1.92, respectively, during - 5dB SNR for duration length of 2 s using three hidden layers of DNN algorithm [10]. Otherwise, the authors studied the effect of different learning networks, which increasing the hidden layers by five [12].

A similar method of DNN-based mask estimation was done in past studies [19-23]. A study in [19] modified the feature and proposed adaptive masks in the DNN-based mask estimation with four hidden layers and 1024 hidden nodes. As a result, average PESQ and STOI scores of 2.12 and 0.78, respectively, were obtained at -5dB SNR [19]. Another study in [23] used the features fusion technique for the DNN input, while the phase-aware and magnitude mask were applied as the target mask. Five-layer structures, including one input layer, three hidden layers, and one output layer with 2048 rectified linear unit (ReLU) neurons were used for the network architecture. As a result, the STOI and PESQ scores of 0.74 and 2.01, respectively, were obtained. Some studies proposed a new target mask in the DNN such as less aggressive Wiener filtering, phase aware, and complex ratio mask [20-22]. Other studies also proposed an alternative approach using post-filtering techniques such as global variance equalization [17, 24] after the DNN-based mask estimation and speech reconstruction.

Another study also proposed a combination algorithm between the DNN-based approach and statistical-based approach to separate speech signals from background noise with TIMIT database [18]. The standard sparse non-negative matrix factorizations (SNMF) features were extracted from the noisy mixture. Five layers which consist of input layer, three hidden layers and one output layer were used. Then, 1024 number of neurons per layer and ReLU activation were applied in the network architecture. The SNMF-DNN with ideal ratio mask (IRM) target produced promising results compared to ideal binary mask (IBM) target. Even though the proposed approach outperformed, it suffered from features complexity. While another study in [5] investigated the generalization of DNN based mask estimation and contribution to modify a DNN model with permutation invariant training by [25]. This approach is time consuming and complex.

III. Research Methodology

Fig. 1 shows the DNN-based mask estimation framework in this study. The framework was proposed to overcome background noise issue. Firstly, the clean speech and noise signal were resampled simultaneously at fixed sampling frequency of 16 kHz, amplified, normalized and scaled with the equal length of data to generate the mixture noisy speech signal in the data preparation stage. The clean and noise speech signals were converted into a time-frequency domain, to be used in the target mask stage. Prior to the mixture signal processing to be parameterized in the time-frequency domain or known as spectrogram, the mixture signal was transformed into gammatone filter bank power spectra (GF) in the feature engineering stage. The parameterized mixture signal was used as the input features of DNN algorithm. The target output was predicted from the input features using the DNN algorithm. Specifically, cWF was used as the target mask in the gammatone time-frequency domain representation. Each frame of the proposed mask spectrogram corresponding to speech or noise for every audio samples was learnt by the DNN. The overall system involved similar process in the training and test sessions, excluding the post-filtering process. Lastly, a post-filtering strategy using HRNR was applied to overcome the residual noise issue after the speech synthesis process. The speech synthesis was done to reconstruct speech signals in time domain from the time-frequency domain using an inverse gammatone.

A. Data Preparation

MASS corpus dataset was used as the clean speech signal with sampling frequency of 22050 Hz [26]. Meanwhile, babble noise was used for background noise with the sampling frequency of 8000 Hz. The noise was artificially combined with the clean signal to produce mixture signal with different signal to noise ratio (SNR) of - 10 and - 5 dB. The mixture signal is as shown in Equation (1) below:

$$y(t) = x(t) + \alpha\, n(t) \quad (1)$$

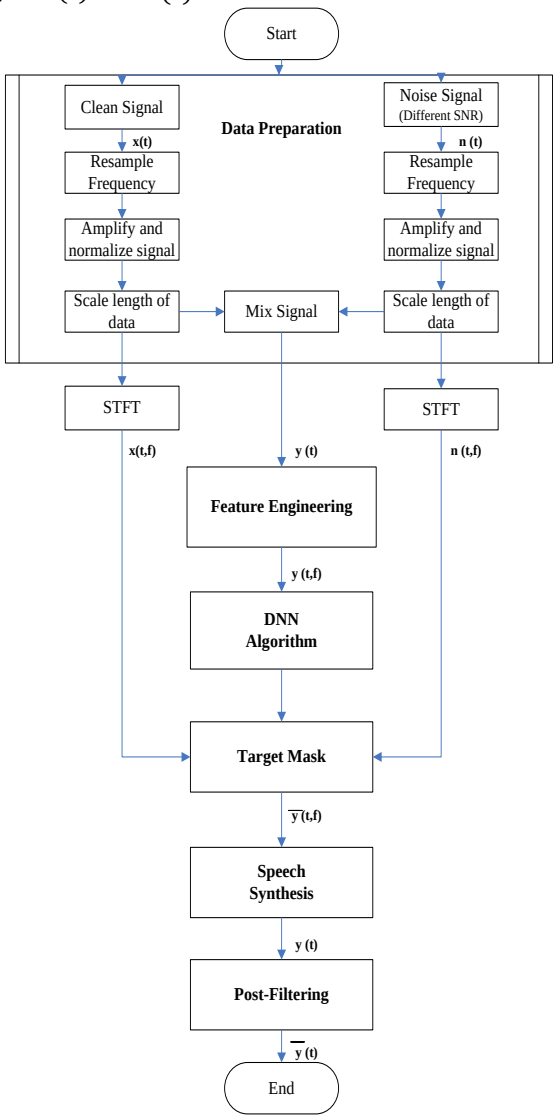


Fig. 1. Research Workflow.

where α is assumed as the gain for scale noise energy signal at different required SNR value as in Equation (2):

$$\alpha = \sqrt{\frac{\sum P_{x(t)}}{\sum P_{n(t)}.10^{SNR/10}}} \quad (2)$$

Then, 600 utterance samples were used for data training, while the remaining 120 samples were used for evaluation performance.

B. Feature Engineering

Gammatone filter bank power spectra (GF) was used because it can produce nonuniform time-frequency resolutions between the high and low frequency regions [27]. The 64-D GF vector was extracted by using 64 channels gammatone filter bank to produce an array of filtered responses with frequency range between 50 and 8000 Hz and loudness gain adjustment. A gammatone filter impulse response is simply defined in time-domain as the product of a gamma distribution and tone. Thus, the gammatone function can be defined as in Equation 3 [28].

$$g(t) = at^{n-1}e^{-2\pi bt}cos(2\pi f_c t + \varphi), \quad (3)$$

where a is the gain value based on loudness theory, $t^{n-1}$is the time onset, exponent function in negative term defines bandwidth, decay rate, f_c is the characteristic frequency in equivalent rectangular bandwidth (ERB), and φ is the initial phase. Typically, b is 1.019 ERB. The ERB of the filter is given in the Equation (4) [28]:

$$ERB = 24.7\left(\frac{4.37f_c}{1000} + 1\right) \quad (4)$$

C. Deep Neural Network (DNN)

The features were trained in a DNN architecture as shown in Fig. 2. The input and output nodes are represented by the blue colour with 640 input nodes and 64 output nodes. Meanwhile, the hidden nodes of four layers are represented by the orange colour. Next, the black arrow represents the forward network while the blue arrow represents the backpropagation network. 4 hidden layers DNN with ReLU hidden and output activation function were applied. Specifically, 430 hidden nodes in each hidden layer with adaptive gradient descent (AGD) and 64 output nodes were applied. Other hyper parameters of DNN to train the networks include 0.2 of dropout rate for backpropagation algorithm, 20 training epoch numbers, 0.5 of initial momentum rate, 0.9 of final momentum rate and loss function of mean squared error.

D. Target Mask

The cWF was proposed as the training target in DNN algorithm to control individual noise and speech distortion. The cWF and IBM target output masks were obtained by using 64 channels of gammatone filter bank, with 20-ms analysis window and 10-ms overlap. The equation of cWF is as follows [29]:

$$y = \widehat{W}(\omega) = \frac{1}{1+\sqrt{\frac{P_n(\omega)}{P_x(\omega)}}} \quad (5)$$

Next, the proposed target was evaluated and compared with other two target outputs such as ideal binary mask (IBM) and gammatone filter bank power spectrum of clean speech (GF-POW). The IBM is typically represented in the time-frequency domain. It is constructed from premixed signals between noise signal and speech signal based on the perceptual principles of auditory scene analysis [14]. The IBM is usually computed from a true signal-to-noise ratio (SNR) through thresholding with local SNR criterion (LC) as shown in Equation (6):

$$IBM(t,f) = \begin{cases} 1, & if\ SNR(t,f) > LC \\ 0, & otherwise \end{cases} \quad (6)$$

E. Post-Filtering Strategy

The HRNR technique is used to preserve the harmonics and to avoid speech distortion [30]. By applying the algorithm, the distorted speech signal could be processed by creating a fully harmonic signal where all the missing harmonics are regenerated. Then, the synthesized speech signal was used to compute a spectral gain and preserve the speech harmonics as shown in Equation (7) [30]:

$$G_{HRNR}(p,k) = \frac{\widehat{SNR}_{prio}^{HRNR}(p,k)}{1+\widehat{SNR}_{prio}^{HRNR}(p,k)} \quad (7)$$

Finally, the resulting speech spectrum was estimated using Equation (8) [30].

$$\hat{S}(p,k) = G_{HRNR}(p,k)X(p,k) \quad (8)$$

F. Performance Measure

Two objective performance measures such as perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) were applied to evaluate the performance between enhanced and clean speech signals after the speech enhancement process. To measure the quality of enhanced speech signal, PESQ is widely used compared to other performance metrics which is more useful even though it is complicated [16]. The PESQ estimates the enhanced speech quality by measuring the distortion difference between the clean speech and the enhanced speech signals. The PESQ score between 0.5 and 4.5 represents bad quality and good quality, respectively [16]. Meanwhile, STOI is widely used to evaluate the performance of enhanced speech signal in terms of speech intelligibility. It computes the correlation in time-frequency domain between the enhanced and clean speech signals without speech perception theory by representing higher correlation when the value of STOI score is greater than 0.9 [13].

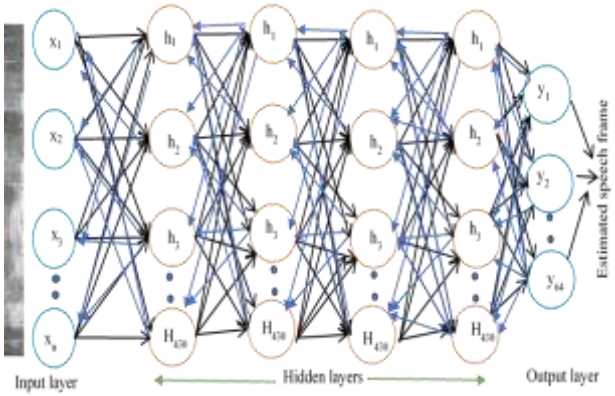


Fig. 2. DNN Architecture.

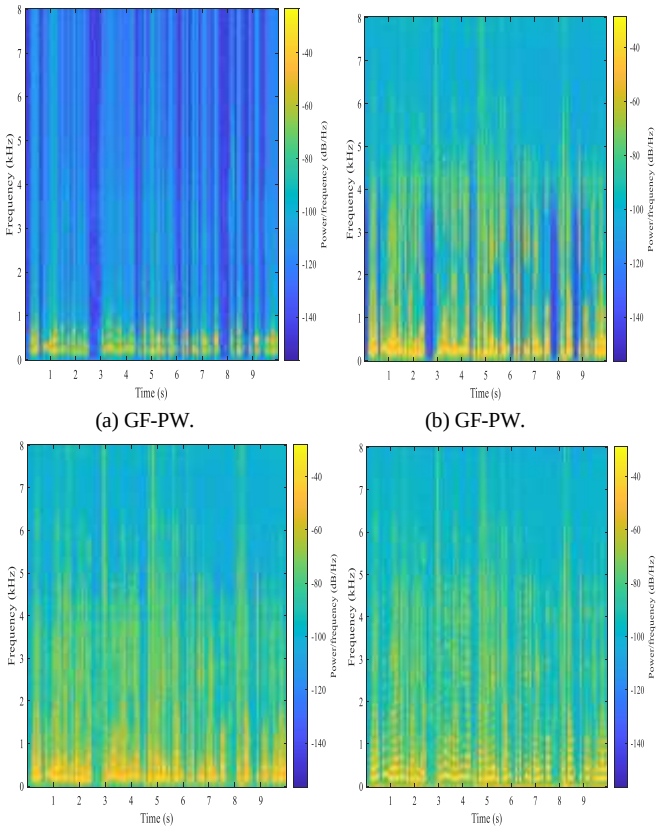
IV. Results and Discussion

Fig. 3 shows the illustration of spectrograms for different target masks and reference of clean signal spectrogram. The speech utterance activity is represented by the yellow colour. The gammatone filter bank power (GF-PW) target mask suffered several speech components loss from the speech harmonics as shown in Fig. 3(a), which may lead to low intelligible compared to Fig. 3(d). Fig. 3(b) shows that Ideal binary mask (IBM) mask with GF-PW was better than the GF-PW, where the background noise that overlapped with speech event and some harmonics of speech components were removed. Moreover, higher residual noise was also introduced in Fig. 3(b).

Fig. 3(c) shows that the proposed target mask for DNN algorithm, cWF with GF-PW did not fully eradicate the background noise that overlapped with speech event at the band frequency of 0 to 4 kHz. As a result, loss of useful speech components from excessive noise elimination could be avoided. The STOI and PESQ scores for the three target masks in seen noise conditions at – 5 dB babble noise of SNR, are illustrated in Fig. 4. Target mask B (IBM + GF-PW) performed better with the STOI score of 0.2 higher than that of Target mask A (GF-PW) at −5 dB of SNR for babble noise. This is because, Target mask B was constructed to retain time–frequency (T–F) units when the estimated speech is stronger than disturbing noise when SNR is greater than local criterion (LC) of -5 dB and remove the T–F units when disturbing noise is dominant when SNR less than LC. Meanwhile, Target mask A applied gammatone filter bank power in clean speech signal without prior noise signal information. Next, the proposed Target mask C (cWF+GF-PW) also performed better than the other two target masks. However, only a small STOI score difference was recorded between the highest STOI score for −5 dB babble noise by Target mask C compared to that of Target masks B and A.

Although Target mask A obtained worse STOI scores than either Target mask B or C, the PESQ scores were better than those of Target mask B. Hence, the proposed Target mask C provided a promising result in PESQ and STOI scores. For example, at babble noise condition, Target mask C obtained the highest PESQ scores of 1.17 for −5 SNR. Therefore, it shows that Target mask B tends to improve speech intelligibility but not speech quality, while Target mask A tends to improve speech quality but not speech intelligibility. In short, Target mask C prediction in DNN-based mask estimation approach seems to be especially beneficial on improving speech quality and speech intelligibility. Furthermore, the difference in performance between the three training targets also reduced when the SNR reduced. This is because, noise signal is more dominant compared to speech signal that may lead to complexity to be predicted.

Fig. 5 shows the spectrogram of enhanced speech with different speech enhancement algorithms: DNN algorithm, HRNR algorithm, DNN+HRNR algorithm and log-MMSE algorithm. The speech utterance was corrupted by babble noise at - 5 dB SNR. Among the four spectrograms, the DNN+HRNR algorithm in Fig. 5(c) was outperformed, the residual noise was reduced considerably without distorting the speech signal. Fig. 5(b) and Fig. 5(d) show that some harmonics in the enhanced speech signals were eliminated by HRNR and Log-MMSE algorithms. Therefore, both algorithms are not suitable for noisy speech signals at low SNR values. This happened when the fix threshold is applied, it may lead to excessive speech distortion and less noise distortion.



(a) GF-PW. (b) GF-PW.

(c) cWF+ GF-PW. (d) Clean (reference).

Fig. 3. Illustration of different Target Masks.

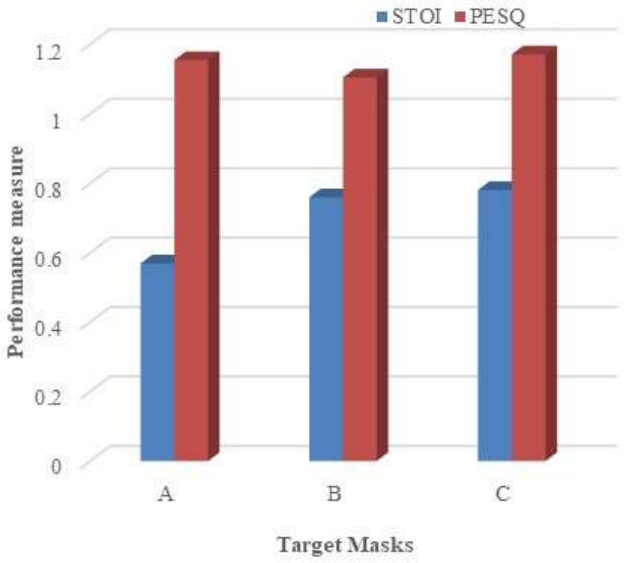
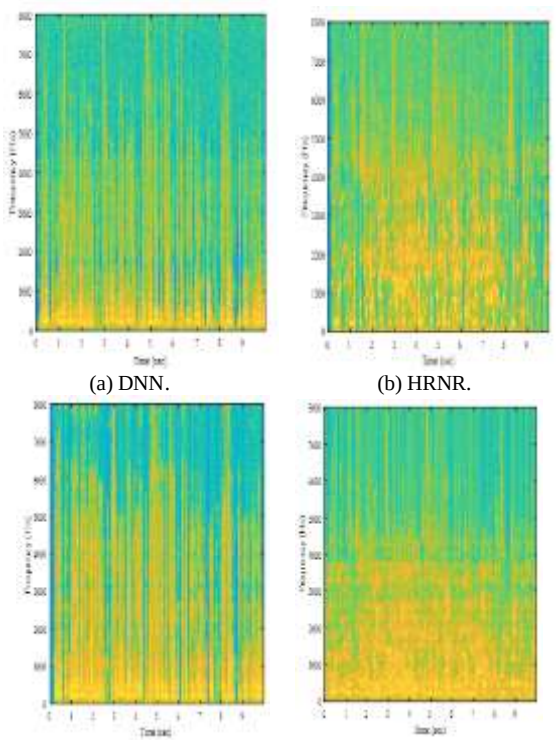


Fig. 4. STOI and PESQ Score for different Target Mask at -5 dB Babble Noise SNR.



(a) DNN. (b) HRNR.

(c) DNN+HRNR. (d) Log-MMSE.

Fig. 5. Spectrogram of Enhanced Speech Signal with different Speech Enhancement Algorithms.

However, to achieve performance like the human speech perception after the DNN-based mask estimation remains a challenging task because the PESQ value was still below 2.0 during low SNR due to the introduction of residual noise after the speech reconstruction. It is supported by illustration of Fig. 6 when measuring the magnitude coherence between estimated and clean speech signal using Welch's overlapped averaged periodogram method. The measured magnitude-squared coherence is a function of frequency with values between 0 and 1 by calculating the power spectral density correlation between estimated and clean speech signal. The values indicate the accuracy of speech signal estimation corresponds to clean speech signal at each frequency. It shows that higher correlation or cross power spectral density occurs after 4 kHz, while lower correlation or cross power spectral density occurs before 4 kHz due to the existence of residual noise.

The calculated STOI score of different algorithms at different SNR for long recorded speech signal length is shown in Fig. 7. The STOI score was slightly increased when the SNR value increased for different types of speech enhancement algorithms. DNN+HRNR improved the STOI score of DNN-based mask estimation approach by 0.01 at - 10 and - 5 dB SNR. The lowest STOI score was 0.2 at - 10 dB SNR produced by the HRNR algorithm, followed by Log-MMSE algorithm, which is 0.35. Fig. 8 shows the calculated PESQ score of different algorithms at different SNR for long recorded speech signal length. The PESQ score at - 10 and - 5 SNR for DNN+HRNR and DNN did not improve significantly. PESQ is used to calculate distortion in speech signal. It is determined that residual noise is remain issue in speech enhancement algorithm due to overlapping between noise and speech signal.

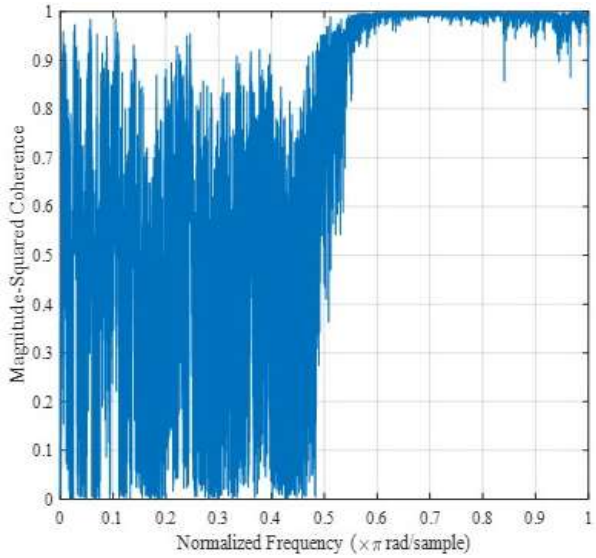


Fig. 6. Magnitude Coherence between Estimated and Clean Speech Signal.

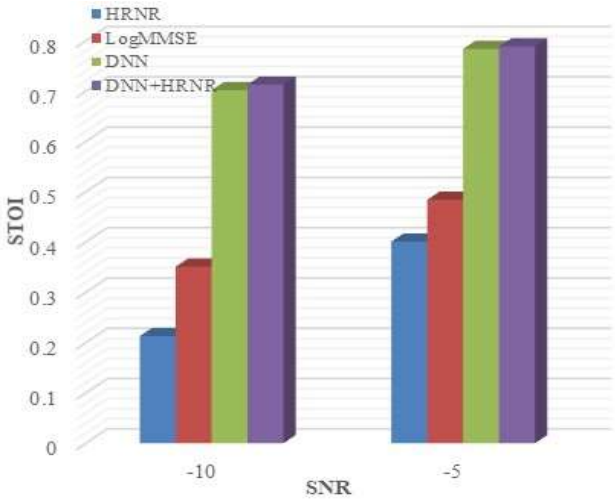


Fig. 7. Comparative STOI Score for different Speech Enhancement Algorithms and SNR Value.

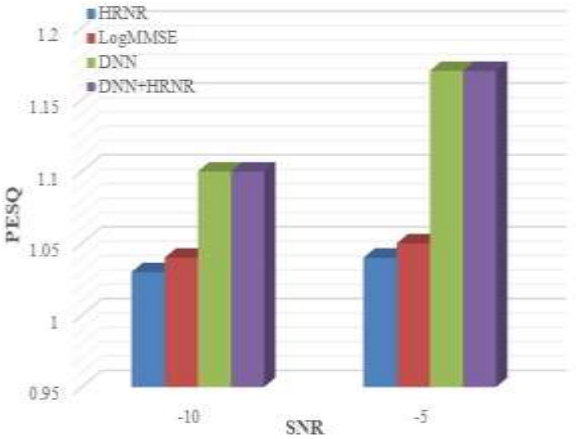


Fig. 8. PESQ Score for different Speech Enhancement Algorithms and SNR Value.

V. CONCLUSION

In conclusion, a supervised speech enhancement algorithm using Deep Neural Network (DNN)-based mask estimation approach has developed and analysed accordingly. A hybrid algorithm between the DNN-based cWF mask estimation and HRNR algorithm has proposed to overcome the heavy non-stationary noise case and low signal to noise ratio (SNR), especially for babble noise at - 5 dB SNR. The proposed mask provided promising results in speech intelligibility due to high STOI score. Moreover, the proposed target mask outperformed other baseline target masks and the hybrid approach has compared to the conventional approach. To be more robust, network architectures such convolution neural network (CNN) and recurrent neural network (RNN) are recommended to be designed.

ACKNOWLEDGMENT

Highly appreciation to Universiti Tun Hussein Onn Malaysia (UTHM) for funding this research work under GPPS (U712) and TIER1 (H268).

REFERENCES

[1] M. Wölfel and J. W. McDonough, Distant speech recognition. Wiley Online Library, 2009.

[2] R. Yao, Z. Zeng, and P. Zhu, "A priori SNR estimation and noise estimation for speech enhancement," EURASIP journal on advances in signal processing, vol. 2016, no. 1, p. 101, 2016.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702-1726, 2018.

[4] N. Saleem and M. I. Khattak, "A review of supervised learning algorithms for single channel speech enhancement," International Journal of Speech Technology, vol. 22, no. 4, pp. 1051-1075, 2019.

[5] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 153-167, 2016.

[6] H. Yu, W.-P. Zhu, Z. Ouyang, and B. Champagne, "A hybrid speech enhancement system with DNN based speech reconstruction and Kalman filtering," Multimedia Tools and Applications, pp. 1-21, 2020.

[7] P. P. Ingale and S. L. Nalbalwar, "Deep neural network based speech enhancement using mono channel mask," International Journal of Speech Technology, vol. 22, no. 3, pp. 841-850, 2019.

[8] X. Li, J. Li, and Y. Yan, "Ideal Ratio Mask Estimation Using Deep Neural Networks for Monaural Speech Segregation in Noisy Reverberant Conditions," in Interspeech, 2017, pp. 1203-1207.

[9] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7092-7096: IEEE.

[10] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 12, pp. 1849-1858, 2014.

[11] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 7, pp. 1381-1390, 2013.

[12] J. Chen and D. Wang, "Dnn based mask estimation for supervised speech separation," in Audio source separation: Springer, 2018, pp. 207-235.

[13] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," IEEE/ACM transactions on audio, speech, and language processing, vol. 24, no. 3, pp. 483-492, 2015.

[14] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1993-2002, 2014.

[15] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 2, pp. 270-279, 2012.

[16] P. C. Loizou, Speech enhancement: theory and practice. CRC press, 2013.

[17] N. Saleem, M. I. Khattak, and E. Perez, "Spectral Phase Estimation Based on Deep Neural Networks for Single Channel Speech Enhancement," Journal of Communications Technology and Electronics, vol. 64, no. 12, pp. 1372-1382, 2019.

[18] H.-W. Tseng, M. Hong, and Z.-Q. Luo, "Combining sparse NMF with deep neural network: A new classification-based approach for speech enhancement," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 2145-2149: IEEE.

[19] R. Li, X. Sun, Y. Liu, D. Yang, and L. Dong, "Multi-resolution auditory cepstral coefficient and adaptive mask for speech enhancement with deep neural network," EURASIP Journal on Advances in Signal Processing, vol. 2019, no. 1, p. 22, 2019.

[20] N. Saleem, M. Irfan Khattak, M. Y. Ali, and M. Shafi, "Deep neural network for supervised single-channel speech enhancement," Archives of Acoustics, vol. 44, 2019.

[21] F. Bao and W. H. Abdulla, "Noise masking method based on an effective ratio mask estimation in Gammatone channels," APSIPA Transactions on Signal and Information Processing, vol. 7, 2018.

[22] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 1, pp. 63-76, 2018.

[23] H. Lang and J. Yang, "Speech Enhancement Based on Fusion of Both Magnitude/Phase-Aware Features and Targets," Electronics, vol. 9, no. 7, p. 1125, 2020.

[24] Z. Huimin, J. Xupeng, and L. Dongmei, "An Iterative Post-processing Approach for Speech Enhancement," in Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing, 2019, pp. 130-134.

[25] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 241-245: IEEE.

[26] T.-P. Tan, X. Xiao, E. K. Tang, E. S. Chng, and H. Li, "MASS: A Malay language LVCSR corpus resource," in 2009 Oriental COCOSDA International Conference on Speech Database and Assessments, 2009, pp. 25-30: IEEE.

[27] B. Gao, W. L. Woo, and L. Khor, "Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation," The Journal of the Acoustical Society of America, vol. 135, no. 3, pp. 1171-1185, 2014.

[28] M. Russo, M. Stella, M. Sikora, and V. Pekić, "Robust cochlear-model-based speech recognition," Computers, vol. 8, no. 1, p. 5, 2019.

[29] M. Parchami, W.-P. Zhu, B. Champagne, and E. Plourde, "Recent developments in speech enhancement in the short-time Fourier transform domain," IEEE Circuits and Systems Magazine, vol. 16, no. 3, pp. 45-77, 2016.

[30] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 2098-2108, 2006.