Specialized Speech Enhancement Model Selection Based on Learned Non-intrusive Quality Assessment Metric

Ryandhimas E. Zezario[1], *Szu-Wei Fu*[12], *Xugang Lu*[3], *Hsin-Min Wang*[4], *Yu Tsao*[1]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[3]National Institute of Information and Communications Technology, Japan
[4]Institute of Information Science, Academia Sinica, Taiwan

{ryandhimas,jasonfu,yu.tsao}@citi.sinica.edu.tw, xugang.lu@nict.go.jp, whm@iis.sinica.edu.tw

Abstract

Previous studies have shown that a specialized speech enhancement model can outperform a general model when the test condition is matched to the training condition. Therefore, choosing the correct (matched) candidate model from a set of ensemble models is critical to achieve generalizability. Although the best decision criterion should be based directly on the evaluation metric, the need for a clean reference makes it impractical for employment. In this paper, we propose a novel specialized speech enhancement model selection (SSEMS) approach that applies a non-intrusive quality estimation model, termed Quality-Net, to solve this problem. Experimental results first confirm the effectiveness of the proposed SSEMS approach. Moreover, we observe that the correctness of Quality-Net in choosing the most suitable model increases as input noisy SNR increases, and thus the results of the proposed systems outperform another auto-encoder-based model selection and a general model, particularly under high SNR conditions.

Index Terms:speech enhancement, ensemble model, long-short-term memory model, non-intrusive quality assessment, PESQ

1. Introduction

Speech enhancement aims to generate cleaner speech from noisy speech. With the emergence of deep learning, many researchers have adopted this technique and had notable performances [1–10]. Recently, bidirectional long short term memory (BLSTM) [8–10] which allows capturing of long-term contextual information, has shown state of-the-art enhancement performance. However, generalizability in mismatched test and training data conditions remains a challenge.

The ensemble model is a feasible solution to increase the generalizability of the learned model [11–16]. In the field of speech processing, the integrated deep and ensemble-learning algorithm (IDEA), which incorporates multiple information from expert models into a unified fusion model, has shown notable dereverberation performances [12]. Similar to [12], several studies have improved the effect of ensemble learning on speech enhancement [13–16]. For instance, Kim [14] employed an auto-encoder to choose the most suitable candidates (from several expert models) using the reconstruction error. In addition, a phonetic-based mixture of experts (MoE) model consisting of phoneme-specific DNNs and a phoneme classifier also provided notable improvements [15]. Recently, Karjol et al. [16] estimated a clean speech spectrum by calculating the linear combination of the outputs of multiple DNNs, in a similar manner to [15]. Although most of current ensemble models have shown promising enhancement results, there remains room for further improvement by applying a novel-candidates decision criterion. We assume that the mismatch between the model selection criterion and the evaluation metrics may affect the performance of speech enhancement.

To reduce the mismatch between the model selection criterion and the final evaluation metrics, these two parameters should be associated with each other. However, most evaluation metrics [17–25] (e.g. perceptual evaluation of speech quality (PESQ) [26] and short-time objective intelligibility (STOI) [27]) need a clean reference so that they cannot be applied directly as the selection criterion. To solve this limitation, our previous paper [28] indicated that the learned model, termed Quality-Net, did not require clean references when computing the estimated scores (thus regarded as a nonintrusive quality estimation model) and could yield a high correlation to the PESQ scores. In this study, we employ Quality-Net to choose the proper candidates according to the estimated quality score.

Specialized speech enhancement model selection (SSEMS) is a novel approach in which Quality-Net is used to choose the best speech enhancement results from several ensemble models. Since collecting numerous possible noises types may not be a practical solution, in this study, rather than training several ensemble models based on the noise types, we intend to apply knowledge-based clustering to specifically capture acoustic information. In addition, Kolbk et al. [29] found that a specialized speech enhancement model can outperform a general model when the test condition is matched to the training condition. The training data is first clustered to male and female by gender information. Then, the gender-specific data are split based on the value of the signal-to-noise-ratio (SNR) into a male, high SNR (MHSNR); male, low SNR (MLSNR); female, high SNR (FHSNR); and female, low SNR (FLSNR). Each of these is then used to train a gender-SNR specific BLSTM enhancement model. Quality-Net is trained to non-intrusively predict the PESQ score by minimizing the MSE between the true PESQ score and the estimated one in a combined training set which includes enhanced, noisy, and clean speech. In the online stage, noisy speech is enhanced by the four ensemble models, and Quality-Net is then employed to choose the best candidates according to the estimated PESQ score.

Experimental results in unseen noise environments show that the proposed SSEMS can achieve consistent improvement in terms of PESQ and STOI. Thus, it confirms the effectiveness of the SSEMS approach in increasing the generalizability and improving the robustness of the speech-enhancement performance.

The remainder of this paper is organized as follows. We introduce the proposed SSEMS in Section II. In Section III, we describe the experimental setup and report the experimental results. Finally, we conclude our findings in Section IV.

2. Systems Description

The SSEMS follows a divide-and-conquer strategy to solve complicated regression tasks. Specifically, each model is trained with particular data, which allows it to be an expert at solving certain problems. Unlike previous ensemble models (e.g., collaborative deep learning [14] and mixture of experts (MoE) [14, 15]),]), our candidate-selection criterion is based on the learned Quality-Net [28]. This method aims to reduce the mismatch between the model selection criterion and the final evaluation metrics by performing a learned, non-intrusive quality assessment to estimate the PESQ score.

2.1. Ensemble model training stage

In this study, the tree structure of knowledge-based clustering is applied to partition the training data. Gender information is considered first to generate clustered training data, resulting in male (M) and female (F) data. Next, because the mismatched SNR condition between the training and test data may reduce the speech-enhancement performance, the SNR information is used to further split the training data, In our setting, the data are categorized as high SNR (HSNR) and low SNR (LSNR) with a threshold of 10 dB. This results in four classes of clustered training data, namely MHSNR, MLSNR, FHSNR, and FLSNR.

As shown in Fig. 1, the proposed SSEMS consists of four ensemble models, and Quality-Net is applied to choose the best speech enhancement results. In the training stage, each of the clustered training data is enhanced through different BLSTMs, resulting in K classes of ensemble models $\{EM_1, EM_2, ..., EM_{K-1}, EM_K\}$. The ensemble model equation can be derived as follows:

$$\hat{x}_n^k = EM_k(y_n) \tag{1}$$

where k ,n, $\hat{x}_n^k$ and y_n indicate k-th index of ensemble models, n-th utterance, enhanced speech, and noisy speech respectively. The training process of Quality-Net is then performed by first concatenating enhanced $\{\hat{x}_1^1...\hat{x}_N^1$, $\hat{x}_1^2...\hat{x}_N^2$, $\hat{x}_1^{K-1}...\hat{x}_N^{K-1}$, $\hat{x}_1^K...\hat{x}_N^K$ $\}$, noisy $\{$ $y_1...y_N\}$ and clean $\{$ $x_1...x_N\}$ into combination dataset C.

2.2. Quality-Net

In this study, Quality-Net is also based on BLSTM for modeling the context acoustic information. However, unlike the speech enhancement model, the true PESQ (Q_n) of C is set as a target to train the model. Furthermore, the conditional frame-wise constraint is introduced to obtain more accurate prediction results as in [28]. Accordingly, the objective function of Quality-Net is derived as follows:

$$O = \frac{1}{n}\sum_{n=1}^{N}[(Q_n - \hat{Q}_n)^2 + \frac{\alpha(Q_n)}{L(U_n)}\sum_{l=1}^{L(U_n)}(Q_n - q_n l)^2] \tag{2}$$

$$\alpha(Q_n) = 10^{(Q_n - Q_{MAX})} \tag{3}$$

where N indicates the total number of training utterances; $L(U_n)$ number of frames m of the n-th utterance; Q_n and $\hat{Q}_n$ the true and predicted PESQ score, respectively; and $q_{n,l}$ and Q_{MAX} the estimated frame level quality of l-th frame of utterance and the maximum score of PESQ, respectively. Finally, the Quality-Net equation can be derived as follows:

$$\hat{Q}_n = QualityNet(y_n) \tag{4}$$

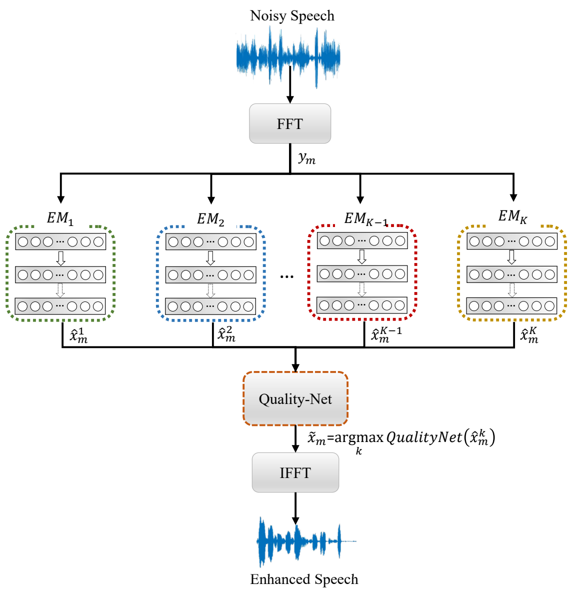

Figure 1: *Architecture of proposed SSEMS approach.*

2.3. Testing stage

In the testing stage, noisy speech is extracted to generate the speech features y_m where m corresponds m-th utterance. Later on, the magnitude spectrum of y_m is processed based on eq. (1) to generate several enhanced spectral. Quality-Net is employed to select the best enhanced spectral based on the following equation:

$$\tilde{x}_m = \underset{k}{\operatorname{argmax}}\, QualityNet(\hat{x}_m^k) \tag{5}$$

Finally, an inverse FFT (IFFT) is applied to reconstruct the selected enhanced spectral and phase features of y_m to obtain the enhanced speech.

3. Experiments

3.1. Experimental setup

We evaluated the proposed SSEMS algorithm on the Wall Street Journal (WSJ) [30] dataset, which consists of 37416 training and 330 test utterances recorded at a 16-Khz sampling rate. For the noisy training utterances, clean utterances were corrupted by 90 types of noises consisting of stationary and non-stationary noise at several SNR levels from 20 to -10 dB. In the test data, four types of noises, including car, pink, street and babble, are injected to generate noisy test data at seven SNR levels (-10, -5, 0, 5, 10, and 15 dB). Please note, the noise types used in the test data are not selected during the training stage, considering that the main purpose of this study is to improve the performances in unseen noise environments. Both the training and test utterances are extracted using a 512-point Short-time Fourier transform (STFT) with a Hamming window size of 32 ms and a hop size of 16 ms, resulting in 257-point STFT log-power-spectra (LPS) features.

In the baseline system, the BLSTM model, which consists of two bidirectional hidden layers with 300 nodes, is trained with all the training data. For the proposed SSEMS, knowledge-based clustering is first applied to create the MHSNR, MLSNR, FHSNR, and FLSNR training sets. These clustered training sets are then used to train ensemble models EM I, EM II, EM III, and EM IV, respectively. Each ensemble model has the same model architecture as the baseline. Next, Quality-Net which consists of one bidirectional hidden layer with 100 nodes, followed by two fully connected layers with 50 exponential linear units and one linear node is applied to estimate the PESQ score [28]. PESQ and STOI are employed to evaluate the performances of different speech-enhancement models.

3.2. Performance comparison between specialized and general enhancement model

We first conduct experiments to verify that a specialized speech-enhancement model can outperform a general model when the test condition (gender and SNR) is matched to the training condition. As shown in table 1, the best PESQ score is achieved when the training condition is matched to the test condition. In addition, although the specialized models (EM I, EM II, EM III, and EM IV) are trained by a relatively smaller dataset compared to the baseline, they can obtain better scores under the matched condition. Therefore, this experiment implies that choosing the correct specialized enhancement model is critical to further improve the results.

Table 1: MATCHED AND MISMATCHED EVALUATIONS.

	EM I	EM II	EM III	EM IV	Baseline
MHSNR	**3.26**	3.05	2.93	2.35	3.05
MLSNR	2.20	**2.28**	2.02	1.89	2.24
FHSNR	2.35	1.94	**3.15**	2.84	2.86
FLSNR	1.57	1.50	2.00	**2.02**	1.97

3.3. Objective evaluation results

We evaluate the proposed approach in several unseen noise environments including two stationary noises (e.g., car, pink) and two non-stationary noises (e.g., street, babble). Tables 2 and 3 show the PESQ and STOI scores of noisy, baseline (general model), and the proposed SSEMS under stationary and non-stationary noise conditions, respectively. These tables show that the PESQ scores of SSEMS can outperform the baseline by a large margin, especially under high SNR conditions. The improvement in the STOI score is less obvious because Quality-Net is trained to estimate PESQ score only. Therefore, these experiments imply that Quality-Net can select the correct specialized model with high accuracy. In the next section, we compare its performance with those of other model-selection methods.

3.4. Correctness comparison

In the previous section, we showed the effectiveness of SSEMS and the importance of selecting the correct model. Considering that the final enhancement model is selected based on the estimation of Quality-Net, we further analyze the performance of Quality-Net and compare its results with those of other model selection methods.

First, we evaluate the correctness score of model selection

Table 2: EVALUATION METRICS COMPARISON OF NOISY (STATIONARY NOISE), BASELINE, AND SSEMS

	Noisy		Baseline		SSEMS	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
15dB	3.13	**0.98**	3.04	0.91	**3.29**	0.93
10dB	2.68	**0.94**	2.93	0.90	**3.10**	0.91
5dB	2.26	**0.88**	2.74	0.87	**2.84**	**0.88**
0dB	1.90	0.80	2.47	0.83	**2.50**	**0.84**
-5dB	1.63	0.70	2.10	0.75	**2.10**	**0.76**
-10dB	1.45	0.61	1.72	**0.65**	**1.74**	**0.65**
ave	2.17	0.82	2.50	0.82	**2.60**	**0.83**

Table 3: EVALUATION METRICS COMPARISON OF NOISY (NON-STATIONARY NOISE), BASELINE, AND SSEMS

	Noisy		Baseline		SSEMS	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
15dB	2.93	**0.97**	3.04	0.91	**3.27**	0.93
10dB	2.49	**0.93**	2.87	0.90	**3.02**	0.91
5dB	2.10	0.86	2.60	0.87	**2.65**	**0.88**
0dB	1.79	0.76	2.21	0.80	**2.22**	**0.81**
-5dB	1.57	0.64	1.74	**0.68**	**1.76**	**0.68**
-10dB	1.51	0.54	1.43	**0.53**	**1.46**	0.52
ave	2.07	0.78	2.32	0.78	**2.40**	**0.79**

by Quality-Net at several SNR values. Correctness scores indicate the capability of an approach to select the most suitable candidate model, compared to the selected model generated by the true PESQ score. As shown in table 4, the correctness scores roughly increase as SNR increases. This explains why SSEMS can significantly outperform the baseline under high SNR conditions as shown in Tables 2 and 3. When dealing with low-SNR noisy speech, speech enhancement models may produce new artificial noises or distort speech components that may affect the judgment of Quality-Net in choosing the best model.

Table 4: CORRECTNESS SCORES OF QUALITY-NET AT SEVERAL SNR CONDITIONS

dB	15	10	5	0	-5	-10
%	94.67	85.74	66.74	67.27	61.19	51.35

Second, to determine how the noise types affect the correctness of Quality-Net, we calculate the correctness scores for several unseen test noises types including car, pink, street, and babble. As shown in table 5, Quality-Net obtains similar performances in the first three noise environments, regardless of if it is stationary or non-stationary noise. However, in the case of babble noise, the correctness drops by approximately 10% compared to others. We argue that this is mainly because Quality-Net cannot accurately distinguish between speech components and babble noise.

An auto-encoder based approach [14], termed DAE, is also employed to compare Quality-Net with other model-selection methods. DAE selects the candidates based on the reconstruction error of the auto-encoder, which is only trained on clean data. In addition, Oracle, which is based on the correct answer of the true PESQ score, is also compared. It indicates the highest performance that can be achieved if the most suitable speech-enhancement model is selected during the testing stage. In Figures 2 and 3, we compare the PESQ scores of DAE, Quality-Net, and Oracle under stationary and non-stationary noise conditions, respectively. From these two figures, it can be observed that the performance of our proposed Quality-Net is significantly better than that of the DAE baseline, especially under high SNR conditions. In addition, under all SNR conditions, Quality-Net is comparable to Oracle.

Table 5: CORRECTNESS SCORES OF QUALITY-NET AT DIFFERENT NOISE TYPES

noise types	Car	Pink	Street	Babble
%	72.97	72.87	74.22	64.56

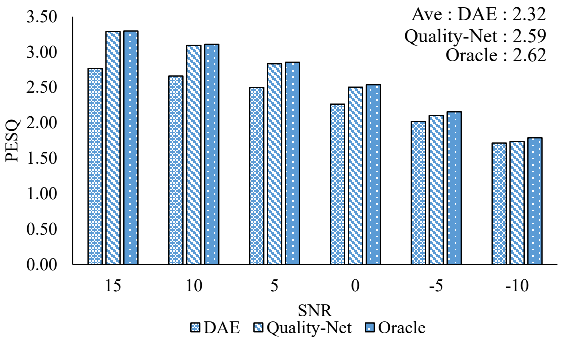

Figure 2: *PESQ comparison of DAE [14], Quality-Net and Oracle at stationary noise environments*

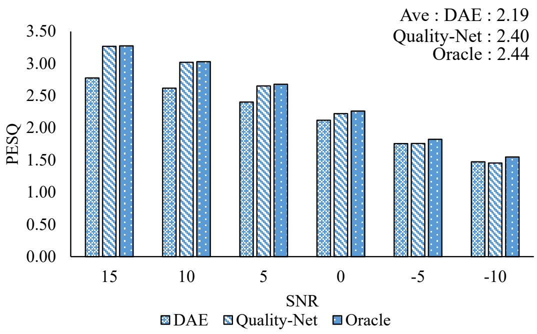

Figure 3: *PESQ comparison of DAE [14], Quality-Net and Oracle at non-stationary noise environments*

3.5. Spectrogram analysis

In addition to the objective evaluation and correctness comparison, we present the spectrogram to visually analyze the performances. Figure 4 shows the spectrograms of clean utterance and noisy utterance at 5 dB SNR under car-noise and enhanced-speech conditions with different models. From the figures, we can observe that the baseline system can effectively reduce the noise. However, the SSEMS can further improve the performance by effectively removing the noise and restoring more speech information as shown in the black box.

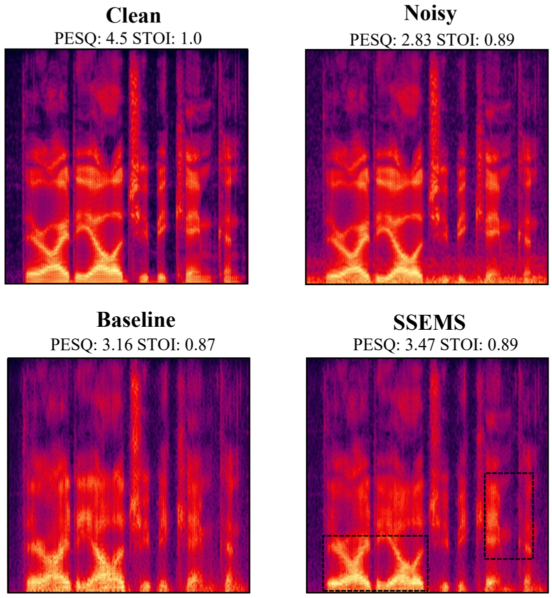

Figure 4: *Spectrograms of a clean utterance, with its noisy (car noise at 5dB SNR condition), baseline, and SSEMS.*

4. Conclusions

This study proposed a novel, specialized speech-enhancement model selection method based on a learned, non-intrusive quality-assessment metric. Because Quality-Net can estimate the speech quality without a corresponding clean reference, our proposed SSEMS can achieve notable improvement by choosing a matched model. Experimental results showed that Quality-Net can achieve a similar performance to the Oracle selection method. Our future works include applying SSEMS in different evaluation metrics and ensemble model strategies. Through SSEMS, we aim to eliminate the model mismatch and improve the ability to select the best speech-enhancement models.

5. Acknowledgements

The authors would like to thank the Ministry of Science and Technology for providing financial supports (107-2221-E-001-012-MY2, 106-2221-E-001-017-MY2).

6. References

[1] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," *in Proc. INTERSPEECH*, pp. 436–440, 2013.

[2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–9, 2015.

[3] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," *in Proc. ICASSP*, pp. 708–712, 2015.

[4] P.-S. Huang, M. Kim, P. Smaragdis, and M. Hasegawa-Johnson, "Deep learning for monaural speech separation," *in Proc. ICASSP*, pp. 1562–1566, 2014.

[5] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transac-*

tions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745–777, 2014.

[6] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv*, vol. arXiv:1609.07132, 2016.

[7] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement," *in Proc. INTERSPEECH*, pp. 3768–3772, 2016.

[8] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," *in Proc. INTERSPEECH*, pp. 3274–3278, 2015.

[9] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *in Proc. INTERSPEECH*, pp. 3314–3318, 2016.

[10] ——, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[11] Y. Tsao and C. H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 17, no. 5, pp. 1025–1037, 2009.

[12] W.-J. Lee, S.-S. Wang, F. Chen, X. Lu, S.-Y. Chien, and Y. Tsao, "Speech dereverberation based on integrated deep and ensemble learning algorithm," *in Proc. ICASSP*, pp. 5454–5458, 2018.

[13] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," *in Proc. INTERSPEECH*, pp. 885–889, 2014.

[14] M. Kim, "Collaborative deep learning for speech enhancement: A run-time model selection method using autoencoders," *in Proc. ICASSP*, pp. 76–80, 2017.

[15] S. E. Chazan, S. Gannot, and J. Goldberger, "A phoneme-based pre-training approach for deep neural network with application to speech enhancement," *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2016.

[16] P. Karjol, A. Kumar, and P. K. Ghosh, "Speech enhancement using multiple deep neural networks," *in Proc. ICASSP*, pp. 5049–5053, 2018.

[17] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[18] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Snr-based progressive learning of deep neural network for speech enhancement," *in Proc. INTERSPEECH*, pp. 3713–3717, 2016.

[19] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[20] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," *in Proc. ICASSP*, pp. 5220–5224, 2016.

[21] ——, "Complex ratio masking for monaural speech separation," *IEEE Transactions on acoustics, speech, and signal processing*, pp. 483–492, 2016.

[22] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *in Proc. INTERSPEECH*, pp. 1508–1512, 2015.

[23] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," *Proceedings of the Workshop on Hands-free Speech Communication and Microphone Arrays*, pp. 136–140, 2017.

[24] R. E. Zezario, J.-W. Huang, X. Lu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Deep denoising autoencoder based post filtering for speech enhancement," *APSIPA Annual Summit and Conference*, pp. 373–377, 2018.

[25] S. Wang, K. Li, Z. Huang, S. Siniscalchi, and C.-H. Lee, "A transfer learning and progressive stacking approach to reducing deep model sizes with an application to speech enhancement," *in Proc. ICASSP*, pp. 5575–5579, 2017.

[26] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *Int. Telecommun. Union, T Recommendation*, no. 862, p. 708712, 2001.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[28] S.-W. Fu., Y. Tsao, H.-T. Hwang, and H.-W. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *in Proc. INTERSPEECH*, pp. 1873–1877, 2018.

[29] M. Kolbk, 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, 2017.

[30] D. Paul and J. Baker, "The design for the wall street journal-based csr corpus," *ICSLP*, pp. 899–902, 1992.