DNN Classification Model-based Speech Enhancement Using Mask Selection Technique

Bong-Ki Lee
Artificial Intelligence Lab., CTO Division
LG Electronics Co., Ltd.
Seoul 06763, South Korea
bongki86.lee@lge.com

***Abstract*—This paper presents a speech enhancement algorithm using a DNN classification model combined with noise classification-based ensemble. Although various single-channel speech enhancement algorithms based on deep learning have been recently developed, since it is optimized for reducing the mean square error, it can not accurately estimate the actual target values in a regression task, resulting in muffled enhanced speech. Therefore, this paper proposes the DNN classification-based single-channel speech enhancement algorithm to overcome disadvantages of the existing DNN regression-based speech enhancement algorithms. To replace the DNN regression task into the classification task, gain mask templates are predefined using k-means clustering among the gain masks. The input feature vector extracted from the microphone input signal is fed into the DNN's input and then an optimal gain mask is selected from the gain mask templates. Furthermore, we define the gain mask templates for each noise environment using the DNN-based noise classification to cover various noise environments and use an ensemble structure based on a probability of the noise classification stage.**

***Index Terms*—speech enhancement, deep learning, classification, ideal ratio mask, log power spectrum, ensemble**

I. Introduction

In recent years, there is a growing need for a speech enhancement system that can be used for robust automatic speech recognition (ASR) and voice communication systems in a variety of noisy environments [1]. Although the performance of the speech enhancement algorithm continues to improve, its performance can not be guaranteed in a speech-like noise and very low signal-to-noise ratio (SNR) noisy environment. Therefore, it is very important to remove the noise or separate the speech from the noisy signal to improve the performance of the ASR and voice communication systems [2].

Over the past few decades, various single-channel speech enhancement algorithms using a noise power estimation based on a speech presence probability have been proposed [3]–[5]. Despite many developments over the years, these algorithms still have disadvantage in that they do not remove the speech-like noises well such as babble, TV, and music interferences. To overcome these drawbacks, deep learning-based single-channel speech enhancement algorithms have been developed in recent years. Firstly, in [6], [7], schemes for directly mapping of the noisy speech frames to the corresponding clean speech frames using a deep neural network (DNN) have been proposed. For the DNN training, the log-power spectrum (LPS) of noisy speech is used as an input feature and that of the clean speech is used as a target feature. Furthermore, a number of papers have been published related to the development of the various input features [8], target features including ideal binary mask (IBM) [9], ideal ratio mask (IRM) [10], and complex IRM (cIRM) [11], and deep learning networks including CNN [12] and RNN [13]. These algorithms corresponding to the regression task that directly estimates a target values such as the LPS and IRM optimize the DNN by the criterion where the mean square error (MSE) between the estimated value by the networks and the target value is reduced. As is well known, estimating a real value by the DNN leads to a smoothed result, so direct estimation of the LPS causes muffled sounds. Also, direct estimation of the masks including IBM, IRM and cIRM causes artifacts which leads to the annoying sounds, when the regression of the target values fails [14].

In this paper, the DNN classification-based single-channel speech enhancement algorithm is proposed to overcome the drawbacks of the conventional deep learning regression-based speech enhancement algorithms. In the proposed algorithm, the regression task of the DNN is replaced to the classification task where one of the gain mask templates predefined by unsupervised k-means clustering [15] among the target gain masks is selected. In order to consider diverse noise environments, we construct the gain mask templates for each noise environment and adopt an ensemble structure based on the probability value of the DNN-based noise classification stage.

II. Proposed Classification Model-based Speech Enhancement Algorithm

The proposed algorithm consists of the DNN classification-based speech enhancement and noise classification-based ensemble parts. Both parts are divided into off-line training and on-line speech enhancement phases since it is learning-based algorithms as shown in Fig. 1, which exhibits the feature extraction, noise classifier, DNN training, ensemble of the DNNs for speech enhancement, and signal synthesis. In Fig. 1, the IRM templates are generated by each noise type using k-means clustering in the off-line training phase. And then the DNN models are generated by each noise type using the input feature and IRM templates. In the on-line speech enhancement phase, the IRM template with the lowest MSE is selected for

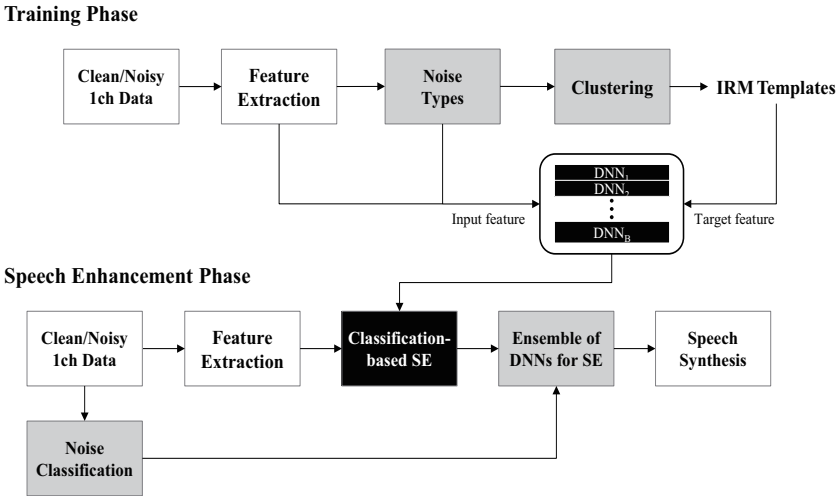

Fig. 1. Flow chart of the proposed speech enhancement algorithm.

the corresponding input feature. At this time, the proposed algorithm has an ensemble structure using the noise classifier to improve performance.

A. Feature Extraction

First, the feature extraction part is divided into the input and target feature vector extractions, respectively. In this paper, the LPS is used as the input feature and IRM is used as the target feature, which are widely used for speech enhancement system in many studies over the years [16]. For the LPS extraction, we perform the short-time Fourier transform (STFT) to obtain the DFT coefficients for each windowed frame such that

$$Y^f(k) = \sum_{m=0}^{M-1} y(m)h(m)e^{-j2\pi km/M}\,, k = 0, 1, \ldots, M-1 \quad (1)$$

where k and M are the frequency bin index and window length, respectively, and $h(m)$ and f denote the window function and frequency domain, respectively. After the STFT, the LPS are given as

$$Y^l(k) = \log|Y^f(k)|^2\,, \quad k = 0, 1, \ldots, K-1 \quad (2)$$

where $K = M/2 + 1$ and l denotes the LPS domain. For $k = K, \ldots, M-1$, $Y^l(k)$ is obtained using the symmetric property given by $Y^l(k) = Y^l(M-k)$; thus the dimension of the LPS is given as $M/2+1$. Next, the oracle IRM , $M(k)$, can be defined as follows [10]:

$$M(k) = \sqrt{\frac{|X(k)|^2}{|X(k)|^2 + |N(k)|^2}}, \quad 0 \le M(k) \le 1 \quad (3)$$

where $|X(k)|$ and $|N(k)|$ indicate the speech and noise magnitude spectra, respectively. In order to perform the speech enhancement based on the DNN classification model, a finite number of the IRM groups must first be defined, called IRM templates in this paper. Therefore, the oracle IRMs are extracted using speech and noise materials obtained from the equation (3) and then the IRM templates, T_a^b, are defined among the oracle IRMs using k-means clustering, where a denotes the index of the IRM in the IRM templates and b denotes the noise type since the IRM templates are obtained by each noise type. Finally, the DNN classification-based speech enhancement model is optimized to select the most accurate IRM in the IRM templates. It is noted that the number of IRM templates was determined as 48 through the experiments in section III.

B. DNN Training

After the feature extraction, the DNN training part is followed as a next step. For the DNN, we use a feed-forward neural network with many hidden layers and the input LPS feature is normalized to zero mean and unit variance. The target label of the DNN is defined in that the one of the IRM templates, T_o^b, having the lowest MSE between the predefined IRM templates and oracle IRM is selected as follows:

$$o = \arg\min_{\mathrm{a}}(T_a^b - M(p))^2, \quad a = 1, 2, \ldots, A \quad (4)$$

where T_a^b is the IRM templates and $M(p)$ denotes the oracle IRM corresponding to the input LPS feature, p. Furthermore, the number of the IRM templates is denoted as A, also represents the number of the DNN output nodes. Therefore, when the input LPS feature is fed into the DNN's input, the DNN is trained so that the optimal IRM T_o^b with index o is selected.

In the training process of the DNN classification, a conjugate gradient (CG)-based back-propagation algorithm to minimize a cross-entropy error [14]. Given a classification problem, the estimated DNN output $y_{a,b}^{\mathrm{se}}$ is fed into the softmax function to obtain the probabilistic soft output $q_{j,b}^{\mathrm{se}}$, as given by

$$q_{j,b}^{\mathrm{se}} = \frac{\exp(y_{j,b}^{\mathrm{se}})}{\sum_{a=1}^{A}\exp(y_{a,b}^{\mathrm{se}})}, \quad \sum_{j=1}^{A} q_{j,b}^{\mathrm{se}} = 1 \quad (5)$$

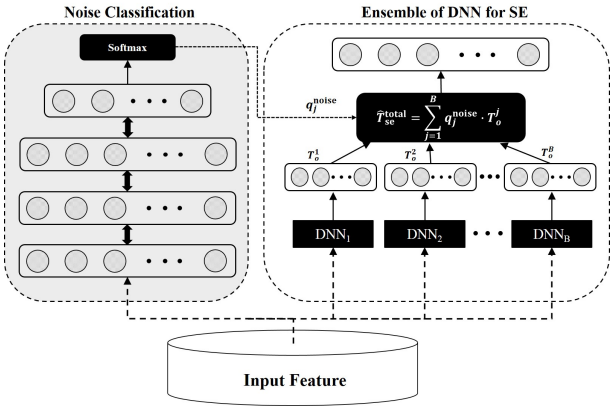

Fig. 2. The proposed DNN ensemble structure using the noise classification.

, where $y^{\text{se}}_{j,b}$ denotes the value of the output of the DNN speech enhancement model b in which $b = 1, 2, \ldots, B$. Finally, the optimal IRM is selected by following the formula $q^{\text{se}}_{o,b} = \max(q^{\text{se}}_{j,b})$, where $o = \arg\max_j q^{\text{se}}_{j,b}$. It is noted that the DNN model is trained for each noise type to construct the ensemble and the total number of the DNN model is B equal to the number of the noise types. Finally, each DNN model $\text{DNN}_\text{b} = \{\text{DNN}_1, \text{DNN}_2, \ldots, \text{DNN}_\text{B}\}$ trained using only the corresponding noise type b, respectively, is used for the DNN ensemble to be presented in the subsection D.

C. Noise Classifier

In order to improve the performance of the classification-based speech enhancement considering various noise coditions, the noise classification-based ensemble is proposed in this paper. For the noise classification, the LPS feature is used for the input feature vector and the index of the noise type is used for the output feature vector of the DNN-based noise classification model. Thus, when the noisy LPS is input, DNN is trained to select the appropriate noise type corresponding to the noisy frame, which is learned in the same way as the classification model of the speech enhancement as follows:

$$q^{\text{noise}}_j = \frac{\exp(y^{\text{noise}}_j)}{\sum_{b=1}^{B} \exp(y^{\text{noise}}_b)}, \quad \sum_{j=1}^{B} q^{\text{noise}}_j = 1 \tag{6}$$

, where B and y^{noise}_j denote the number of the DNN's output nodes and the value of DNN's output for the noise classification model. Finally, the DNN classification model is trained in that the optimal noise type is selected by following the formula $q^{\text{noise}}_o = \max(q^{\text{noise}}_j)$, where $o = \arg\max_j q^{\text{noise}}_j$. It is noted that the noise classification is performed on the first 20 frames at the beginning of the speech, which is thought to be completely noise-only frames. The performance of noise classification is described in section III.

D. Ensemble of DNNs for SE

The DNN model of the noise classification presented in the previous subsection is generated for each noise type DNN_b, as shown in Fig. 2. Then the final output of the DNN for speech enhancement ensemble, $\hat{T}^{\text{total}}_{\text{se}}$, is calculated as follows:

$$\begin{aligned} \hat{T}^{\text{total}}_{\text{se}} &= q^{\text{noise}}_1 \cdot T^1_o + q^{\text{noise}}_2 \cdot T^2_o +, \ldots, q^{\text{noise}}_B \cdot T^B_o \\ &= \sum_{j=1}^{B} q^{\text{noise}}_j \cdot T^j_o \end{aligned} \tag{7}$$

, where q^{noise}_j and T^j_o are the probability obtained from the noise classification DNN model and optimal IRM obtained from the each speech enhancement DNN model, respectively. Then, the final optimal IRM $\hat{T}^{\text{total}}_{\text{se}}$ is multiplied to the noisy magnitude spectra $|Y(k)|$ to obtain the estimated clean magnitude spectra $\hat{X}$, as follows:

$$\hat{X}(k) = \hat{T}^{\text{total}}_{\text{se}}(k) \cdot |Y(k)|, \quad k = 0, 1, \ldots, K-1 \tag{8}$$

In this way, the DNN ensemble for the speech enhancement system can improve the performance of the classification-based speech enhancement model while considering various noise conditions.

E. Signal Synthesis

From the above equations, the final enhanced signal of time-domain can be obtained using the inverse-STFT (ISTFT) as follows:

$$\hat{x}(m) = \frac{1}{M} \sum_{k=0}^{M-1} \hat{X}(k) e^{j2\pi km/M} \tag{9}$$

, where $\hat{x}$ denotes the time-domain signal in the proposed speech enhancement algorithm. In this case, the phase of the input noisy signal is used for the ISTFT.

III. Experimental Results

In this section, the simulation performed on various experiments including the number of the IRM templates, noise classifier, and objective speech quality measures to verify the superiority of the proposed algorithm.

We used native Korean speech databases which consist of 7,560 utterances (8.4 hours long in total) for a training set and 1,890 utterances (2.1 hours long in total) for an objective evaluation set. For algorithm implementation, we considered frame lengths of 20 ms with 50% overlap-add using the Hamming window and 512-point FFT. Therefore, the input dimension of the proposed DNNs for speech enhancement is 257 while the input dimension of the proposed DNN for noise classification is 257x20 frames = 5140. Both, the speech enhancement DNNs and noise classification DNN each have three hidden layers with 1024 hidden nodes and a rectified linear unit (ReLU) is used for the activation function of the hidden layer.

To consider both matched and mismatched noise environment cases, 6 types of noise including white, factory, office, babble, street, and wind sounds from NOISEX-92 [17] were used for the training ($B = 6$), and 3 types of noise including car, TV, vacuum sounds acquired from real environment were used for the test. We added the aforementioned noises to the clean speech signal at different SNRs of 0, 5, and 10 dB.

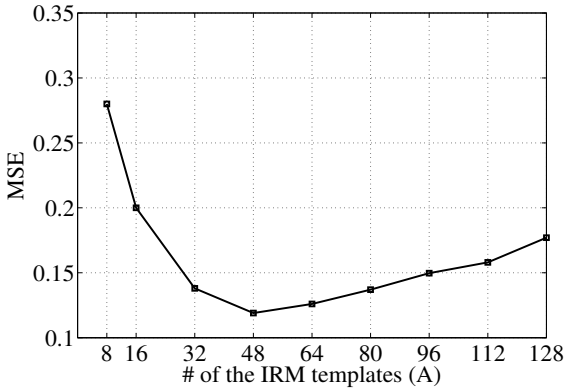

Fig. 3. MSE between oracle IRM and selected IRM in the IRM templates according to the number of the IRM templates.

TABLE I
RESULT OF THE NOISE CLASSIFICATION THROUGH A CONFUSION MATRIX AT SNR = 5 DB (AVERAGE ACCURACY OF THE NOISE CLASSIFICATION UNDER SEEN ENVIRONMENTS: 92.71%).

Accuracy	White	Factory	Office	Babble	Street	Wind
White	**97.86**	0	0	0	0	2.14
Factory	0	**95.59**	1.66	0	2.38	0
Office	0	1.25	**88.10**	10.65	0	0
Babble	0	0	7.61	**88.33**	4.06	0
Street	0	4.01	0	8.36	**87.63**	0
Wind	1.24	0	0	0	0	**98.76**
Car	6.32	0	0	0	4.31	**89.37**
TV	0	4.32	10.41	**76.53**	8.74	0
Vacuum	**80.36**	6.78	0	0	0	12.86

First, we investigated the optimal number of the IRM templates which are generated by using the k-means clustering within the oracle IRM. If the number of the IRM templates increases, the accuracy of the classification model decreases. On the contrary, if the number of the IRM templates decreases, the accuracy of the classification model may increase, but it has difficulty when removing noise because the number of IRMs that can be selected is small. Fig. 3 shows a graph of the MSE between the oracle IRM and selected IRM in the IRM templates according to the number of the IRM templates. As the number of the IRM templates increases to 48, the MSE decreases, but after 48, the MSE increases. Through this experimental result, the number of the IRM templates are set to 48 ($A = 48$) in this paper.

Next, the performance of the noise classifier is shown in Table I. For the reason of the paper volume, only the results with SNR of 5 dB are attached. Since we used 6 noises including white, factory, office, babble, street, and wind for training, it can be seen that the unseen noises including car, TV, vacuum are classified into the similar noises among the trained noises. The noise classification performance for seen noise is 92.71% on average.

In addition, to evaluate the performance of the proposed speech enhancement algorithm, we used the perceptual evaluation of speech quality (PESQ) [18] and short-time objective intelligibility (STOI) [19] which are objective measures widely used for evaluation of the speech enhancement algorithm.

TABLE II
PESQ RESULTS FROM THE CONVENTIONAL AND PROPOSED ALGORITHMS.

Noise	SNR (dB)	Method			
		DNN-LPS	DNN-IRM	MS	MS-Ens.
White	0	2.16	2.43	2.48	**2.54**
	5	2.79	2.93	2.96	**2.98**
	10	3.11	3.23	3.27	**3.31**
Babble	0	1.85	2.14	2.21	**2.24**
	5	2.43	2.66	2.71	**2.74**
	10	2.76	2.92	3.05	**3.10**
Car	0	1.97	2.36	2.39	**2.47**
	5	2.43	2.78	2.83	**2.92**
	10	2.87	3.13	3.15	**3.18**
TV	0	1.71	2.04	2.07	**2.12**
	5	2.24	2.49	2.54	**2.57**
	10	2.55	2.86	2.93	**2.95**

TABLE III
STOI RESULTS FROM THE CONVENTIONAL AND PROPOSED ALGORITHMS.

Noise	SNR (dB)	Method			
		DNN-LPS	DNN-IRM	MS	MS-Ens.
White	0	0.71	0.78	0.80	**0.83**
	5	0.79	0.88	0.88	**0.90**
	10	0.87	0.93	0.94	**0.95**
Babble	0	0.63	0.72	0.74	**0.75**
	5	0.75	0.82	0.83	**0.85**
	10	0.83	0.90	0.90	**0.91**
Car	0	0.65	0.73	0.74	**0.76**
	5	0.77	0.85	0.85	**0.88**
	10	0.84	0.89	0.91	**0.94**
TV	0	0.59	0.69	0.72	**0.76**
	5	0.72	0.77	0.83	**0.85**
	10	0.80	0.86	0.88	**0.90**

To show the superiority of the proposed algorithm, we compare it with the deep learning-based speech enhancement algorithms which use the LPS (DNN-LPS) [7] and IRM (DNN-IRM) [10] as the target feature, respectively. The DNN configuration of the conventional algorithms including DNN-LPS and DNN-IRM is the same with the proposed algirhthm in this paper. For a fair comparison, since the proposed method using ensembles (MS-Ens.) uses more DNNs than conventional algorithms, we also conducted a performance evaluation on the single mask selection DNN that do not use ensembles (MS), which uses the same number of the DNN parameters as the conventional algorithms.

The Table II and III give the results of the PESQ and STOI comparisons under various SNR conditions with seen

and unseen noisy environments. For the reason of the paper volume, only some of the results are attached. In the case of the MS method using the same number of the DNN parameters as the DNN-LPS and DNN-IRM, most of the PESQ and STOI results showed good performance. Also, in all cases, the performance was the best in MS-Ens. method using the ensemble structure, and the performance is guaranteed in the seen environments as well as the unseen environments. Overal experimental results show that the proposed algorithm is effective in suppressing noises especially in the low SNR noisy environments while minimizing the distortion of the speech than the conventional algorithms.

IV. Conclusion

In this paper, we propose a speech enhancement algorithm based on the DNN classification model adopting an ensemble structure to overcome the disadvantages of the DNN regression-based speech enhancement. Unlike conventional DNN regression-based speech enhancement algorithms that directly generate LPS or IRM, the DNN classification model is used to select the optimal IRM among the predefined IRM templates. The ensemble structure based on the noise classification is used to consider various noise environments. The simulation results show that the proposed algorithm is superior to the conventional algorithms.

In future works, we can try to construct more ensembles than the six ensembles used in this paper and use the complex IRM instead of the IRM as the target feature. Also, the performance verification using various DNN structures such as bidirectional LSTM and CNN can be carried out.

References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, April 2014.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

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

[7] Y. Xu, J. Du, L. Dai, and C. 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. 7–19, Jan 2015.

[8] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 826–835, April 2014.

[9] 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, Dec 2014.

[10] 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*, May 2013, pp. 7092–7096.

[11] 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, March 2016.

[12] S. Fu, T. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2017, pp. 1–6.

[13] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3709–3713.

[14] B. Lee, K. Noh, J. Chang, K. Choo, and E. Oh, "Sequential deep neural networks ensemble for speech bandwidth extension," *IEEE Access*, vol. 6, pp. 27 039–27 047, 2018.

[15] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm," in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, April 2010, pp. 63–67.

[16] 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, Dec 2014.

[17] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

[18] K. Kokkinakis and P. C. Loizou, "Evaluation of objective measures for quality assessment of reverberant speech," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 2420–2423.

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