Multi-task Single Channel Speech Enhancement Using Speech Presence Probability As A Secondary Task Training Target

1st Lei Wang
Dept. of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
wang_lei@sjtu.edu.cn

2nd Jie Zhu
Dept. of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
zhujie@sjtu.edu.cn

3rd Ina Kodrasi
Speech and Audio Processing Group
Idiap Research Institute
Martigny, Switzerland
ina.kodrasi@idiap.ch

***Abstract*—To cope with reverberation and noise in single channel acoustic scenarios, typical supervised deep neural network (DNN)-based techniques learn a mapping from reverberant and noisy input features to a user-defined target. Commonly used targets are the desired signal magnitude, a time-frequency mask such as the Wiener gain, or the interference power spectral density and signal-to-interference ratio that can be used to compute a time-frequency mask. In this paper, we propose to incorporate multi-task learning in such DNN-based enhancement techniques by using speech presence probability (SPP) estimation as a secondary task assisting the target estimation in the main task. The advantage of multi-task learning lies in sharing domain-specific information between the two tasks (i.e., target and SPP estimation) and learning more generalizable and robust representations. To simultaneously learn both tasks, we propose to use the adaptive weighting method of losses derived from the homoscedastic uncertainty of tasks. Simulation results show that the dereverberation and noise reduction performance of a single-task DNN trained to directly estimate the Wiener gain is higher than the performance of single-task DNNs trained to estimate the desired signal magnitude, the interference power spectral density, or the signal-to-interference ratio. Incorporating the proposed multi-task learning scheme to jointly estimate the Wiener gain and the SPP increases the dereverberation and noise reduction further.**

***Index Terms*—multi-task learning, supervised deep neural network, speech presence probability, dereverberation, noise reduction**

I. Introduction

In many speech communication applications, the recorded microphone signal is inevitably corrupted with late reverberation and noise, which can be detrimental to speech quality and intelligibility and to the accuracy of speech recognition systems [1], [2]. The goal of single channel speech enhancement is to recover the desired signal while suppressing the interference, i.e., late reverberation and noise. Single channel speech enhancement has been traditionally approached using spectral enhancement techniques [3], [4] or probabilistic modeling-based techniques [5], [6]. In recent years however, successful contributions based on data-driven approaches such as deep neural networks (DNNs) have been proposed [7].

This work is supported by the National Key Research Project of China under Grant No. 2017YFF0210903 and the National Natural Science Foundation of China under Grant No. 61371147.

Typical supervised DNN-based techniques for single channel speech enhancement learn a mapping from reverberant and noisy input features to a user-defined target [7]. Depending on the definition of the target, such techniques can be broadly categorized into magnitude estimation [8]–[10] and mask estimation techniques [11]–[13]. Magnitude estimation techniques aim at estimating the desired signal spectral magnitude. The enhanced signal is then obtained by combining the estimated magnitude with the phase of the recorded microphone signal. Mask estimation techniques on the other hand aim at estimating a time-frequency mask such as the Wiener gain. The enhanced signal is then obtained by applying the estimated time-frequency mask to the recorded microphone signal. Instead of directly estimating the time-frequency mask, indirect mask estimation techniques have been recently proposed in [14], [15], where the interference power spectral density (PSD) or the signal-to-interference ratio (SIR) are estimated. The estimated interference PSD or SIR can then used to define a time-frequency mask to recover the enhanced signal.

To improve the generalization performance of such DNN-based enhancement techniques, in this paper we propose to incorporate multi-task learning [16], which means using one network to estimate multiple targets simultaneously. Multi-task learning has been successfully applied in various areas such as computer vision [17] or natural language processing [18], in this paper it is incorporated for DNN-based single channel speech enhancement. Multi-task learning improves learning efficiency and generalization performance by using shared representations to jointly learn multiple related tasks, such that what is learned from one task can help learning and generalization in another task. To incorporate multi-task learning in supervised DNN-based single channel speech enhancement techniques, we propose to use speech presence probability (SPP) estimation as a secondary task. SPP is a useful parameter in traditional single channel speech enhancement techniques for accurately estimating the interference PSD, and hence, for improving the speech enhancement performance [3], [19], [20]. Consequently, we expect that the incorporation of SPP estimation as a secondary task results in learning more robust representations for the primary target (i.e., desired signal magnitude, time-frequency mask, interference PSD, or SIR) estimation task. To simultaneously learn both tasks, we

propose to use the adaptive weighting method of losses derived from the homoscedastic uncertainty of tasks in [21].

II. DNN-BASED SINGLE CHANNEL ENHANCEMENT

We consider a reverberant and noisy microphone system with a single speech source and a single microphone. In the short-time Fourier transform (STFT) domain, the received microphone signal $Y(k,l)$ at frequency bin k and time frame index l can be written as

$$Y(k,l) = X(k,l) + \underbrace{R(k,l) + N(k,l)}_{I(k,l)}, \tag{1}$$

with $X(k,l)$ being the direct and early reverberation component, $R(k,l)$ being the late reverberation component, $N(k,l)$ being the additive noise component, and $I(k,l)$ denoting the total interference component (i.e., late reverberation and noise). Assuming that $X(k,l)$ and $I(k,l)$ are uncorrelated, the PSD of the microphone signal $Y(k,l)$ is given by

$$\Phi_y^2(k,l) = \mathcal{E}\{|Y(k,l)|^2\} = \Phi_x^2(k,l) + \Phi_i^2(k,l), \tag{2}$$

with $\mathcal{E}$ denoting the expected value operator and $\Phi_x^2(k,l)$ and $\Phi_i^2(k,l)$ denoting the PSDs of $X(k,l)$ and $I(k,l)$, respectively.

Since early reverberation is desirable [22], the objective of speech enhancement is to recover an estimate of the direct and early reverberation component $X(k,l)$. Typical DNN-based techniques aiming to recover $X(k,l)$ are trained to learn a mapping from reverberant and noisy input features to a user-defined target. Depending on the target definition, such techniques can be broadly categorized into magnitude estimation [8]–[10] and mask estimation techniques [11]–[15]. Mask estimation techniques can be additionally categorized into 3 subcategories, i.e., directly time-frequency mask estimation [8]–[10], interference PSD estimation required to compute a time-frequency mask [14], a priori SIR estimation required to compute a time-frequency mask [15]. These techniques differ not only in terms of the target definition, but also in terms of the used input features and DNN architectures. However, to provide a systematic review and compare the performance for different targets in Section IV, in this paper we consider only different target definitions for standard feed-forward DNN architectures with temporal context depicted in Figs. 1(a) and 1(b). Next, a brief overview of the considered input and target definitions for such DNNs is provided.

A. Magnitude estimation

When estimating the desired signal magnitude, the DNN target vector can be defined as the K–dimensional vector constructed using the spectral magnitude of $X(k,l)$ at time frame l across all frequency bins K, i.e.,

$$\mathbf{x}(l) = [|X(1,l)|\ |X(2,l)|\ \dots\ |X(K,l)|]^T. \tag{3}$$

To incorporate temporal context, the DNN input vector can be defined as the $K(2T+1)$–dimensional vector constructed by concatenating the spectral magnitude of $Y(k,l)$ from the past and future T time frames across all frequency bins K, i.e.,

$$\begin{aligned}\mathbf{y}(l) \quad &= [|Y(1,l-T)|\ \dots\ |Y(K,l-T)|\dots \\ &\quad \dots|Y(1,l+T)|\dots|Y(K,l+T)|]^T.\end{aligned} \tag{4}$$

Using the estimated spectral magnitude $|\hat{X}(k,l)|$, the enhanced signal can be obtained as $\hat{X}_{\text{mag}}(k,l) = \frac{|\hat{X}(k,l)|}{|Y(k,l)|}Y(k,l)$.

B. Mask estimation

Although different time-frequency masks have been investigated in the literature [7], [12], the commonly used Wiener gain is considered in this paper. With the a priori SIR $\xi(k,l)$ defined as $\xi(k,l) = \frac{\Phi_x^2(k,l)}{\Phi_i^2(k,l)}$, the Wiener gain can be computed as

$$G(k,l) = \frac{\xi(k,l)}{\xi(k,l)+1}. \tag{5}$$

1) Direct mask estimation: When directly estimating the Wiener gain, the DNN target vector can be defined as the K–dimensional vector constructed using the gain $G(k,l)$ at time frame l across all frequency bins K, i.e.,

$$\mathbf{G}(l) = [G(1,l)\ G(2,l)\ \dots\ G(K,l)]^T, \tag{6}$$

whereas the DNN input vector can be defined as the $K(2T+1)$–dimensional vector $\mathbf{y}(l)$ in (4). Using the estimated Wiener gain $\hat{G}(k,l)$, the enhanced signal can be obtained as $\hat{X}_{\text{gain}}(k,l) = \hat{G}(k,l)Y(k,l)$.

2) Interference PSD estimation: Instead of directly estimating the gain in (5), in [14] it has been proposed to use a DNN for estimating the interference PSD $\Phi_i^2(k,l)$. Hence, the DNN target vector can be defined as the K–dimensional vector constructed using the interference PSD $\Phi_i^2(k,l)$ at time frame l across all frequency bins K, i.e.,

$$\mathbf{\Phi}_i^2(l) = [\Phi_i^2(1,l)\ \Phi_i^2(2,l)\ \dots\ \Phi_i^2(K,l)]^T. \tag{7}$$

Further, the DNN input vector can be defined as the $K(2T+1)$–dimensional vector constructed by concatenating the microphone signal PSD $\Phi_y^2(l)$ from the past and future T time frames as in (4), i.e.,

$$\begin{aligned}\mathbf{\Phi}_y^2(l) = \quad &[|\Phi_y^2(1,l-T)|\ \dots\ |\Phi_y^2(K,l-T)|\dots \\ &\dots|\Phi_y^2(1,l+T)|\ \dots\ |\Phi_y^2(K,l+T)|]^T.\end{aligned} \tag{8}$$

To compute the enhanced signal, first the estimated interference PSD $\hat{\Phi}_i^2(k,l)$ is used to obtain an estimate of the a priori SIR $\hat{\xi}_{\text{psd}}(k,l)$ based on the decision directed approach [23]. The estimated a priori SIR $\hat{\xi}_{\text{psd}}(k,l)$ is then exploited to compute the Wiener gain $\hat{G}_{\text{psd}}$ as in (5), yielding the enhanced signal $\hat{X}_{\text{psd}}(k,l) = \hat{G}_{\text{psd}}Y(k,l)$.

3) SIR estimation: Instead of directly estimating the Wiener gain in (5), in [15] it has been proposed to use a DNN for estimating the SIR $\xi(k,l)$. Hence, the DNN target vector can be constructed as

$$\boldsymbol{\xi}(l) = [\xi(1,l)\ \xi(2,l)\ \dots\ \xi(K,l)]^T, \tag{9}$$

whereas the DNN input vector is the $K(2T+1)$–dimensional vector $\mathbf{y}(l)$ defined in (4). To compute the enhanced signal,

the estimated a priori SIR is used to compute the Wiener gain $\hat{G}_{\text{sir}}$ as in (5), yielding $\hat{X}_{\text{sir}}(k,l) = \hat{G}_{\text{sir}} Y(k,l)$.

III. Multi-task learning for DNN-based speech enhancement

In this section, we propose to increase the generalization performance of the DNN-based speech enhancement techniques reviewed in Section II by incorporating multi-task learning, where multiple targets are learned simultaneously in one network. Instead of using a single-task DNN that only estimates the user-defined target (i.e., desired signal magnitude, time-frequency mask, interference PSD, or SIR), we propose to use a multi-task DNN that additionally estimates the SPP. The SPP is a useful parameter in single channel speech enhancement for accurately tracking the interference PSD, and hence, for improving the speech enhancement performance [3]. We hypothesize that jointly learning to estimate the user-defined target and the SPP through shared DNN layers within a multi-task learning framework yields more robust and generalizable representations for the primary task (i.e., estimating the user-defined target). Assuming that the desired signal and interference STFT coefficients are complex Gaussian distributed, the SPP can be computed as [3]

$$\text{SPP}(k,l) = \left(1 + \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}(1+\xi_{\mathcal{H}_1})\text{e}^{-\frac{|Y(k,l)|^2}{\Phi_{\text{i}}^2(k,l)}\frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}}\right)^{-1}, \tag{10}$$

where $P(\mathcal{H}_1)$ and $P(\mathcal{H}_0)$ are the prior probabilities of speech presence and absence and $\xi_{\mathcal{H}_1}$ denotes the typical a priori SIR when speech is present. In line with the target definitions in Section II, the target vector for SPP estimation is given by

$$\mathbf{SPP}(l) = [\text{SPP}(1,l)\ \text{SPP}(2,l)\ \ldots\ \text{SPP}(K,l)]^T. \tag{11}$$

Figs. 1(c)–1(e) depict examples of the considered DNN architectures for jointly learning two different tasks, with the first task being the estimation of a target vector as presented in Section II and the second task being the estimation of the SPP in (11). In Fig. 1(c) both tasks share one hidden layer followed by a task-specific layer, in Fig. 1(d) both tasks share two hidden layers followed by a task-specific layer, whereas in Fig. 1(e) both tasks share one hidden layer followed by two task-specific layers. To train these architectures, the loss function can be defined as a weighted sum of the task-specific loss functions [21], i.e.,

$$\mathcal{L}_{\text{fixed}}(\mathbf{W}) = \lambda_1 \mathcal{L}_1(\mathbf{W}) + \lambda_2 \mathcal{L}_2(\mathbf{W}), \tag{12}$$

with $\mathcal{L}_1$ being the loss function for estimating a target vector from Section II, $\mathcal{L}_2$ being the loss function for estimating the SPP in (11), λ_1 and λ_2 being user-defined weighting scalars, and $\mathbf{W}$ being the model parameters. When using the loss function in (12), the performance of the model can be sensitive to the values of λ_1 and λ_2 and finding optimal values can be expensive [21]. To avoid tuning λ_1 and λ_2, we propose to use the adaptive loss function derived in [21] to automatically weight the task-specific loss functions, i.e.,

$$\mathcal{L}_{\text{ada}}(\mathbf{W}, \sigma_1, \sigma_2) = \frac{1}{\sigma_1^2}\mathcal{L}_1(\mathbf{W}) + \frac{1}{\sigma_2^2}\mathcal{L}_2(\mathbf{W}) + \log \sigma_1 \sigma_2, \tag{13}$$

where σ_1 and σ_2 are scalars jointly learned with the model parameters $\mathbf{W}$. Although not presented in this paper due to space constraints, using (13) yields a better performance than using (12) for several user-defined λ_1 and λ_2 for the reverberant and noisy acoustic scenarios considered in Section IV.

IV. Simulation results

In this section, the performance of all single-task techniques discussed in Section II is first compared on the same datasets and DNN architectures.[1] Further, the performance of the proposed multi-task framework for joint direct mask and SPP estimation is investigated.

A. Datasets

Two datasets are considered, i.e., a reverberant dataset where the interference consists of different reverberation levels and a reverberant and noisy dataset (referred to as a noisy dataset) where the interference consists of a fixed reverberation level and varying levels and types of noise. As clean speech material, we have used the TIMIT database [24].

To generate the reverberant dataset, clean speech files are convolved with measured room impulse responses (RIRs) with reverberation times ranging from 200 ms to 1 s. For the reverberant training, validation, and test sets we have used 500, 200, and 200 clean speech files and 16, 8, and 8 RIRs, respectively, with no overlap between files for different sets.

To generate the noisy dataset, clean speech files are firstly convolved with one measured RIR and corrupted with different noise types from the DEMAND database [25]. For the training, validation, and test sets we have used 250, 100, and 100 clean speech files convolved with an RIR with reverberation time 580 ms, 570 ms, and 560 ms, respectively. As before, there is no overlap between the clean speech files and the RIRs for different sets. Further, for the training, validation, and test sets, 5 different noise types at 3 different broadband signal-to-noise ratio (SNR) are added to the reverberant signals, with SNR $\in \{-5\text{dB}, 0\text{dB}, 5\text{dB}\}$. To analyze the generalization capabilities of the proposed models, an unseen noisy test set is also generated by adding 3 unseen noise types at unseen broadband SNRs to the test reverberant signals, with SNR$\in \{-3\text{dB}, 3\text{dB}, 10\text{dB}\}$.

B. Parameters, network architectures, and measures

Parameters. Signals are processed in the STFT domain using a weighted overlap-add framework with a tight analysis window of 256 samples and an overlap of 50%. Considering only half of the spectrum, the number of frequency bins is $K = 129$. Further, the number of time frames used for temporal context is $T = 3$. To compute the PSDs required in (5)–(10), we use recursive averaging with a smoothing factor of 0.85. To compute the SPP in (10) we use $P(\mathcal{H}_1) = 0.5$, $P(\mathcal{H}_0) = 0.5$, and $10\log_{10}\xi_{\mathcal{H}_1} = 15$ dB.

[1] To the best of our knowledge, only the performance of magnitude and direct mask estimation techniques has been compared on the same datasets and DNN architectures in [12], while the performance of the more recently proposed interference PSD and SIR estimation techniques has not been considered.

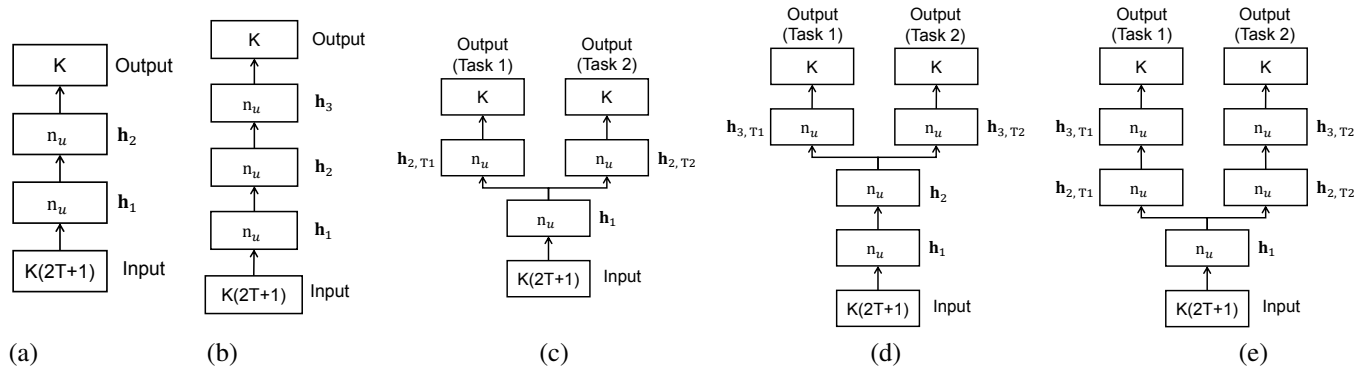

Fig. 1. Schematic illustration of the considered DNN architectures: (a) single-task estimation with two layers, (b) single-task estimation with three layers, (c) multi-task estimation with one shared layer followed by one task-specific layer, (d) multi-task estimation with two shared layers followed by one task-specific layer, and (e) multi-task estimation with one shared layer followed by two task-specific layers.

TABLE I
PERFORMANCE OF SINGLE-TASK ESTIMATION OF DIFFERENT TARGETS ON THE TEST REVERBERANT, NOISY, AND UNSEEN NOISY DATASETS.

	Reverberant				Noisy				Unseen Noisy			
Measure	$\hat{X}_{\text{mag}}$	$\hat{X}_{\text{gain}}$	$\hat{X}_{\text{psd}}$	$\hat{X}_{\text{sir}}$	$\hat{X}_{\text{mag}}$	$\hat{X}_{\text{gain}}$	$\hat{X}_{\text{psd}}$	$\hat{X}_{\text{sir}}$	$\hat{X}_{\text{mag}}$	$\hat{X}_{\text{gain}}$	$\hat{X}_{\text{psd}}$	$\hat{X}_{\text{sir}}$
ΔPESQ	0.14	**0.23**	0.13	0.14	0.04	**0.22**	0.13	0.10	0.00	**0.23**	0.16	0.12
ΔfwSSNR	1.80	**2.27**	1.07	1.31	1.77	**2.68**	1.64	1.19	1.06	**2.38**	1.71	0.98

Network architectures. As previously mentioned, the network architectures considered for the single- and multi-task techniques are depicted in Fig. 1. For all architectures, we use rectifying linear unit (ReLU) as non-linearity on all hidden layers. For estimating an unbounded target (i.e., the desired signal magnitude, the interference PSD, or the SIR), there is no non-linearity on the output layer. For estimating the Wiener gain or the SPP which are bounded between 0 and 1, a sigmoid non-linearity is used on the output layer. Mean square error is used as the loss function for training the single-task architectures in Figs. 1(a), (b) and as the loss function $\mathcal{L}_1$ for training the multi-task architectures in Figs. 1(c)–1(e). Cross-entropy loss is used as the loss function $\mathcal{L}_2$ for training the multi-task architectures in Figs. 1(c)–1(e). All considered architectures are trained for different number of hidden units $n_u \in \{500, 1000, 1500\}$ using the Adam optimizer with different hyper-parameters, i.e., learning rate $l_r \in \{0.001, 0.0001\}$ and weight decay $w_d \in \{0, 0.001\}$. After training for 200 epochs, the model parameters corresponding to the epoch with the lowest validation error (out of all considered architectures, n_u, l_r, and w_d) are used as the final model parameters.

Measures. The dereverberation and denoising performance is measured by the improvement in perceptual evaluation of speech quality (ΔPESQ) [26] and frequency-weighted segmental signal to noise ratio (ΔfwSSNR) [27] between the processed and recorded microphone signals.

C. Single-task performance

The performance of the techniques in Section II is compared on the test reverberant, noisy, and unseen noisy datasets. As previously mentioned, for each technique, the two- and three-layer networks in Figs. 1(a) and 1(b) are trained for all considered hyper-parameters and the final network is selected as the one yielding the minimum validation loss.

Table I presents the performance on all considered test datasets, with the presented performance measures averaged over all utterances in the respective datasets. It can be observed that the considered techniques generally yield an improvement in PESQ and fwSSNR on all datasets, with the direct mask estimation technique (i.e., $\hat{X}_{\text{gain}}$) yielding the best performance. The advantageous performance of the direct mask estimation technique in comparison to magnitude estimation was already established in [12]. However, also the more recently proposed interference PSD and SIR estimation techniques show a lower dereverberation and noise reduction performance than the direct mask estimation technique on all datasets.

D. Multi-task performance

The results presented in Section IV-C confirm the advantageous performance of the direct mask estimation technique in comparison to other state-of-the-art techniques. Hence, in the following, this technique is jointly used with SPP estimation within the proposed multi-task learning scheme. As previously mentioned, the two- and three-layer networks depicted in Figs. 1(c)– 1(e) are trained for all considered hyper-parameters and the final network is selected as the one yielding the minimum validation loss. The PESQ and fwSSNR improvement obtained on all considered datasets are shown in Table II. When comparing the presented ΔPESQ and ΔfwSSNR to the values in Table I (for $\hat{X}_{\text{gain}}(k,l)$), it can be seen that the proposed multi-task scheme improves the performance over single-task training on all datasets. While a small difference can be observed in ΔPESQ, a larger difference is observed in the presented ΔfwSSNR values.

TABLE II
PERFORMANCE OF THE PROPOSED MULTI-TASK FRAMEWORK FOR JOINTLY ESTIMATING THE WIENER GAIN ($\hat{X}_{\text{GAIN}}$) AND THE SPP ON THE REVERBERANT, NOISY, AND UNSEEN NOISY TEST DATASETS.

	Reverberant	Noisy	Unseen Noisy
ΔPESQ	0.24	0.24	0.25
ΔfwSSNR	2.40	3.11	2.74

In summary, the presented results confirm that using a multi-task learning framework with SPP estimation improves the dereverberation and noise reduction performance of conventional single-task DNN-based enhancement techniques. In the future, we will investigate the potential of incorporating parameters other than the SPP within the proposed multi-task learning framework.

V. CONCLUSION

In this paper, multi-task learning has been proposed to improve the performance of supervised DNN-based single channel speech enhancement techniques. Instead of only estimating a user-defined target (e.g., the desired signal magnitude, a time-frequency mask such as the Wiener gain, the interference PSD, or the SIR), it has been proposed to also jointly estimate the SPP through shared DNN layers. To simultaneously learn both tasks, we have used a recently proposed adaptive weighting method of losses derived from the homoscedastic uncertainty of tasks. Simulation results on reverberant and noisy datasets show that jointly estimating the Wiener gain and the SPP within the proposed multi-task learning framework outperforms other state-of-the-art techniques.

REFERENCES

[1] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, July 2006.

[2] A. Warzybok, I. Kodrasi, J. O. Jungmann, E. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective speech quality and speech intelligibility evaluation of single-channel dereverberation algorithms," in *Proc. International Workshop on Acoustic Signal Enhancement*, Juan les Pins, France, Nov. 2014, pp. 332–336.

[3] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2011.

[4] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Juki, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 61, July 2015.

[5] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., pp. 758–764. MIT Press, Cambridge, MA, USA, Feb. 2001.

[6] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[7] D. L. 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, May 2018.

[8] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, June 2015.

[9] B. Wu, K. Li, M. Yang, and C. Lee, "A study on target feature activation and normalization and their impacts on the performance of DNN based speech dereverberation systems," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Jeju, South Korea, Dec. 2016, pp. 1–4.

[10] B. Wu, K. Li, M. Yang, and C. Lee, "A reverberation-time-aware approach to speech dereverberation based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, Jan. 2017.

[11] K. Han and D. L. Wang, "A classification based approach to speech segregation," *Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475, Nov. 2012.

[12] 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.

[13] Z. Q. Wang, P. D. Wang, and D. L. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 99, pp. 1778–1787, May 2020.

[14] I. Kodrasi and H. Bourlard, "Single-channel late reverberation power spectral density estimation using denoising autoencoders.," in *Proc. Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018, pp. 1319–1323.

[15] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, Aug. 2019.

[16] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, July 1997.

[17] X. G. Wang, C. Zhang, and Z. Zhang, "Boosted multi-task learning for face verification with applications to web image and video search," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, Aug. 2009, pp. 142–149.

[18] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep Neural Networks with multitask learning," in *Proc. International Conference on Machine Learning*, New York, NY, USA, Jan. 2008, pp. 160–167.

[19] A. Abramson and I. Cohen, "Simultaneous detection and estimation approach for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2348–2359, Nov. 2007.

[20] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2011, pp. 145–148.

[21] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Dec. 2018, pp. 7482–7491.

[22] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, July 2003.

[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[24] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, Nov. 1992.

[25] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *Journal of the Acoustical Society of America*, vol. 133, no. 5, May 2013.

[26] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862*, International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[27] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.