PAPER

A Multi-Task Scheme for Supervised DNN-Based Single-Channel Speech Enhancement by Using Speech Presence Probability as the Secondary Training Target

This paper was retracted because it was a duplicate submission with the following paper.

Lei WANG, Jie ZHU, and Ina Kodrasi, Multi-task Single Channel Speech Enhancement Using Speech Presence Probability As A Secondary Task Training Target, in Proc. EUSIPCO 2021, pp. 296-300, 2021.

https://eurasip.org/Proceedings/Eusipco/Eusipco2021/pdfs/0000296.pdf

Lei WANG[†], **Jie ZHU**[†a)], *Nonmembers, and* **Kangbo SUN**[†], *Member*

SUMMARY To cope with complicated interference scenarios in realistic acoustic environment, supervised deep neural networks (DNNs) are investigated to estimate different user-defined targets. Such techniques can be broadly categorized into magnitude estimation and time-frequency mask estimation techniques. Further, the mask such as the Wiener gain can be estimated directly or derived by the estimated interference power spectral density (PSD) or the estimated signal-to-interference ratio (SIR). In this paper, we propose to incorporate the multi-task learning in DNN-based single-channel speech enhancement by using the speech presence probability (SPP) as a secondary target to assist the target estimation in the main task. The domain-specific information is shared between two tasks to learn a more generalizable representation. Since the performance of multi-task network is sensitive to the weight parameters of loss function, the homoscedastic uncertainty is introduced to adaptively learn the weights, which is proven to outperform the fixed weighting method. Simulation results show the proposed multi-task scheme improves the speech enhancement performance overall compared to the conventional single-task methods. And the joint direct mask and SPP estimation yields the best performance among all the considered techniques.

key words: *multi-task learning, supervised deep neural network, speech presence probability, dereverberation, noise reduction*

1. Introduction

In the real world speech communication, the recorded speech signal is inevitably corrupted with reverberation or noise, which is detrimental to speech quality and intelligibility and the accuracy of speech recognition applications although the early reverberation can be advantageous [1], [2]. Many speech enhancement methods have been developed to recover the target signal and suppress the late reverberation and background noise, such as the spectral subtraction [3], wiener filtering [4], statistic model-based approach [5], blind probabilistic modeling-based method [6]. Overall the traditional methods are dependent on the prior assumption, parameter setting and manual experience, which limit the denoising and dereverberation performance.

In the last decades, deep neural networks (DNN) have been increasingly used in automatic speech recognition (ASR) and shown impressive performance [7], [8]. Consequently, DNN is also introduced for noise suppression and dereverberation. The supervised DNN is investigated to learn a mapping from reverberant and noisy input features to a user-defined target. According to the variety of targets, the supervised DNN-based technique can be broadly classified to two categories, i.e., magnitude approximation [9]–[11] and mask estimation techniques [12]–[14]. Magnitude approximation-based method uses DNN to learn spectral magnitude of desired signal from spectral magnitude of the recorded signal. The enhanced signal is then obtained by combining the estimated magnitude with the phase of the recorded microphone signal. On the other hand, the mask estimation technique aims at learning a time-frequency mask such as the Wiener gain. Then the enhanced feature is computed as the element-wise product of the estimated mask and the recorded signal feature. Other than directly estimating the time-frequency mask, recently other methods to obtain the mask have been also proposed, where the interference power spectral density (PSD) or the signal-to-interference ratio (SIR) are estimated by DNN [15], [16]. The estimated interference PSD or SIR can then be used to compute a time-frequency mask to recover the enhanced signal.

Multi-task learning scheme 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. In [17], the authors use the multi-task network to learn desired magnitude and noise magnitude simultaneously, then the network outputs are used to derive the time-frequency mask. In this paper, we propose a novel multi-task scheme for statistics estimation in speech enhancement, where speech presence probability (SPP) estimation serves as the secondary task to improve the estimation accuracy of the primary task (i.e., desired signal magnitude, time-frequency mask, interference PSD, or SIR estimation tasks). SPP is a helpful parameter in the traditional single-channel speech enhancement techniques [18]–[20] and highly related with the primary task. Hence, we consider SPP as the secondary target to assist the primary target in learning a more robust and generalizable representation. The loss of the multi-task learning network is the weighted sum of sub-task's losses, and the weights can be manually assigned. Since tuning the weights can be expensive, we propose to use the adaptive weighting method of

Manuscript received December 24, 2020.
Manuscript revised April 26, 2021.
Manuscript publicized August 5, 2021.

[†]The authors are with the Department of Electronic Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China.
a) E-mail: zhujie@sjtu.edu.cn

DOI: 10.1587/transinf.2020EDP7267

losses derived from the homoscedastic uncertainty of tasks in [21].

In the experiment, the performance of single-task techniques for estimating different targets is evaluated and compared. The direct mask estimation is proven to outperform others on both reverberant and noisy datasets. The performance of proposed multi-task scheme using SPP as a secondary target is also evaluated and compared. The adaptive weighting method for the loss of multi-task network shows superiority compared to the fixed weighting method. Additionally, simulation results prove that the proposed multi-task scheme improves the speech enhancement performance in most cases. And the joint direct mask and SPP estimation yields the best speech enhancement performance among all considered techniques.

2. DNN-Based Speech Enhancement

Assuming there is a single microphone which records a desired speech source interfered by reverberation and additive noise, the recorded microphone signal at time index t can be represented by,

$$\begin{aligned} y(t) &= h_e(t) * s(t) + h_l(t) * s(t) + n(t) \\ &= x(t) + r(t) + n(t), \end{aligned} \tag{1}$$

where $*$ indicates convolution operation, $s(t)$ is the clean speech signal, $n(t)$ is the additive noise signal, $h_e(t)$ is the impulse response for the direct sound and early reflection, $h_l(t)$ is the late reflection impulse response, $x(t) = h_e(t) * s(t)$ is the direct and early reverberation component and $r(t) = h_l(t) * s(t)$ is the late reverberation component. In the short-time Fourier transform (STFT) domain, the microphone signal is given by,

$$\begin{aligned} Y(k,l) &= X(k,l) + R(k,l) + N(k,l) \\ &= X(k,l) + I(k,l), \end{aligned} \tag{2}$$

where k is the frequency index, l is the time index, $X(k,l)$, $R(k,l)$ and $N(k,l)$ are the STFTs of $x(t)$, $r(t)$ and $n(t)$ respectively, and $I(k,l) = R(k,l) + N(k,l)$ represents the STFT of inteference $r(t) + n(t)$. Assuming that $X(k,l)$ and $I(k,l)$ are uncorrelated, the PSD of the microphone signal $Y(k,l)$ is given by

$$\Phi_y^2(k,l) = \mathcal{E}\{|Y(k,l)|^2\} = \Phi_x^2(k,l) + \Phi_i^2(k,l), \tag{3}$$

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.

Our goal in this work is to estimate $X(k,l)$ and suppress $I(k,l)$, as early reverberation component is beneficial to speech intelligibility improvement but the late reverberation combined with additive noise is detrimental to the speech intelligibility. Typical DNN-based techniques aiming to recover $X(k,l)$ use DNN learning 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 [9]–[11] and mask estimation techniques [12]–[16]. Mask estimation techniques can be additionally categorized into three subcategories, i.e., techniques that directly estimate a time-frequency mask [9]–[11], techniques that estimate the interference PSD required to compute a time-frequency mask [15], and techniques that estimate the a priori SIR required to compute a time-frequency mask [16]. It should be noted that 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 be able to provide a systematic review and to compare the performance for different targets in Sect. 4, 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). In the remainder of this section, a brief overview of the considered input and target definitions for such DNNs is provided.

2.1 Magnitude Approximation

Since the objective of speech enhancement is to estimate the direct and early reverberation component $X(k,l)$ from the noisy and reverberant observation $Y(k,l)$, one of the association solutions is to directly estimating the desired magnitude $|X(k,l)|$ from the recorded magnitude $|Y(k,l)|$. 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,

$$\boldsymbol{x}(l) = [|X(1,l)|, |X(2,l)|, |X(3,l)|, \ldots, |X(K,l)|]^T. \tag{4}$$

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

$$\begin{aligned} \boldsymbol{y}(l) = [&|Y(1,l-T)|, \ldots, |Y(K,l-T)|, \ldots \\ &\ldots, |Y(1,l+T)|, \ldots, |Y(K,l+T)|]^T. \end{aligned} \tag{5}$$

Using the estimated spectral magnitude $|\hat{X}(k,l)|$ and the phase information of the noisy signal, the enhanced signal is obtained as $\hat{X}_{\text{mag}}(k,l) = \frac{|\hat{X}(k,l)|}{|Y(k,l)|} Y(k,l)$.

2.2 Mask Approximation

2.2.1 Direct Mask Estimation

Noise and reverberation can be removed by directly applying a reference mask on the spectrum of recorded signal. Although different time-frequency masks have been investigated in the literature [13], [22], the commonly used Wiener gain is considered in this paper, which is represented as,

$$G(k,l) = \frac{\Phi_x^2(k,l)}{\Phi_x^2(k,l) + \Phi_i^2(k,l)}, \tag{6}$$

where $\Phi_x^2(k,l)$ is PSD of the target signal calculated from $X(k,l)$ as,

$$\Phi_x^2(k,l) = \beta\Phi_x^2(k,l-1) + (1-\beta)|X(k,l)|^2, \tag{7}$$

where β is a recursive smoothing parameter. Similarly, $\Phi_i^2(k,l)$ is the PSD of the interference signal computed from $I(k,l)$ using recursive averaging. The DNN target vector is the K–dimensional vector constructed using the gain $G(k,l)$ at time frame l across all frequency bin K,

$$\boldsymbol{G}(l) = [G(1,l),\ G(2,l),\ G(3,l), \ldots, G(K,l)]^T. \tag{8}$$

The DNN input vector is the $K(2T+1)$–dimensional vector in Eq. (5). Using the estimated Wiener gain $\hat{G}(k,l)$, the enhanced signal is obtained as $\hat{X}_{\text{gain}}(l) = \hat{G}(k,l)Y(k,l)$.

2.2.2 Interference PSD Estimation

Instead of directly estimating the gain in (6), in [15] it has been proposed to use a DNN for estimating the interference PSD $\Phi_i^2(k,l)$. The target vector is $\Phi_i^2(k,l)$ at time frame l across all frequency bin K,

$$\boldsymbol{\Phi}_i^2(l) = [\Phi_i^2(1,l),\ \Phi_i^2(2,l),\ \Phi_i^2(3,l), \ldots, \Phi_i^2(K,l)]^T. \tag{9}$$

Further, the DNN input vector can be defined as the $K(2T+1)$–dimensional vector constructed by concatenating the microphone signal PSD $\boldsymbol{\Phi}_y^2(l)$ from the past and future T time frames as in (5), i.e.,

$$\begin{aligned}\boldsymbol{\Phi}_y^2(l) = [&|\Phi_y^2(1,l-T)|, \ldots, |\Phi_y^2(K,l-T)|, \ldots \\ &\ldots,\ |\Phi_y^2(1,l+T)|, \ldots, |\Phi_y^2(K,l+T)|]^T.\end{aligned} \tag{10}$$

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], i.e.,

$$\hat{\xi}_{\text{psd}}(k,l) = \alpha\frac{|\hat{X}_{\text{psd}}(k,l-1)|^2}{\hat{\Phi}_i^2(k,l-1)} + (1-\alpha)\max\left[\frac{|Y(k,l)|^2}{\hat{\Phi}_i^2(k,l)} - 1, 0\right] \tag{11}$$

with α being the smoothing factor and $\hat{\Phi}_i^2(k,l)$ being the estimated interference PSD. The estimated a-priori SIR $\hat{\xi}_{\text{psd}}(k,l)$ is then exploited to compute the Wiener mask $\hat{G}_{\text{psd}}$,

$$\hat{G}_{\text{psd}} = \frac{\hat{\xi}_{\text{psd}}(k,l)}{\hat{\xi}_{\text{psd}}(k,l) + 1}. \tag{12}$$

The enhanced signal is obtained as $\hat{X}_{\text{psd}}(k,l) = \hat{G}_{\text{psd}}Y(k,l)$.

2.2.3 A-Prior SIR Estimation

Moreover, in [16] it has been proposed to use a DNN for estimating the a-prior SIR $\xi(k,l)$, which is defined as,

$$\xi(k,l) = \frac{\Phi_x^2(k,l)}{\Phi_i^2(k,l)}. \tag{13}$$

In this case, the target vector is $\xi(k,l)$ at time frame l across all frequency bin K,

$$\boldsymbol{\xi}(l) = [\xi(1,l),\ \xi(2,l),\ \xi(3,l),\ \ldots, \xi(K,l)]^T. \tag{14}$$

And the DNN input vector is the $K(2T+1)$–dimensional vector $\mathbf{y}(l)$ defined in (5). The estimated SIR $\hat{\xi}_{\text{sir}}(k,l)$ is used to compute the Wiener mask as

$$\hat{G}_{\text{sir}} = \frac{\hat{\xi}_{\text{sir}}(k,l)}{\hat{\xi}_{\text{sir}}(k,l) + 1}. \tag{15}$$

The enhanced signal is obtained as $\hat{X}_{\text{sir}}(k,l) = \hat{G}_{\text{sir}}Y(k,l)$.

3. Multi-Task Learning for Statistical Estimation

Instead of using a single-task DNN that only estimates one of the user-defined targets in Sect. 2 (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 [18]. 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 described in Sect. 2). Assuming that the desired signal and interference STFT coefficients are complex Gaussian distributed, the SPP can be computed as [18]

$$P(\mathcal{H}_1|y) = \left(1 + \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}(1+\xi_{\mathcal{H}_1})e^{-\frac{|y|^2}{\Phi_i^2}\frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}}\right)^{-1}, \tag{16}$$

where $P(\mathcal{H}_1)$ and $P(\mathcal{H}_0)$ are the prior probabilities of speech presence or absence respectively, $\xi_{\mathcal{H}_1}$ is the optimal fixed a-prior SNR, and Φ_i^2 is the interference PSD calculated by recursive smoothing. For notational convenience, the time and frequency indexes are omitted. In line with the target definitions in Sect. 2, the target vector for SPP estimation is given by

$$\textbf{SPP}(l) = [\text{SPP}(1,l),\ \text{SPP}(2,l),\ \ldots, \text{SPP}(K,l)]^T. \tag{17}$$

Figures 1 (c)–1(e) depict examples of the considered DNN architectures for jointly learning two tasks, with the first task being the estimation of a target vector as presented in Sect. 2 and the second task being the estimation of the SPP in (17). After obtaining the estimated target from the first task's output, the enhanced speech signals are obtained from the approaches described in Sect. 2, which are referred as $\hat{X}^{\text{M}}_{\text{mag}}(k,l)$, $\hat{X}^{\text{M}}_{\text{gain}}(k,l)$, $\hat{X}^{\text{M}}_{\text{psd}}(k,l)$, $\hat{X}^{\text{M}}_{\text{sir}}(k,l)$ respectively.

The loss function of multi-task network plays an important role in learning performance. The most common formulation is to sum the weighted loss of every sub-task. In this paper, we mainly concern the loss function of a multi-task learning network with two sub-tasks, i.e.,

$$\mathcal{L}_{\text{fixed}}(\mathbf{W}) = \lambda_1\mathcal{L}_1(\mathbf{W}) + \lambda_2\mathcal{L}_2(\mathbf{W}) \tag{18}$$

with $\mathcal{L}_1$ being the loss function for estimating a target vector from Sect. 2, $\mathcal{L}_2$ being the loss function for estimating the SPP in (17), λ_1, λ_2 being the user-defined weighting scalars, and $\mathbf{W}$ being the model parameters. When using the loss function in (18), 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 weigh the task-specific loss functions, i.e.,

$$\mathcal{L}_{\text{ada}}(\mathbf{W}, \sigma_1, \sigma_2) = \frac{1}{\sigma_1^2}\mathcal{L}_1(\mathbf{W}) + \frac{1}{\sigma_2^2}\mathcal{L}_2(\mathbf{W}) + \log \sigma_1 \sigma_2, \tag{19}$$

where σ_1, σ_2 are the observation noise parameter of each tasks. While minimizing the loss function, the value of σ_1, σ_2 would also be adjusted, which is regarded as learning the weight of losses $\mathcal{L}_1(\mathbf{W})$ and $\mathcal{L}_2(\mathbf{W})$.

In [17], the multi-task network is used to learn desired magnitude and noise magnitude simultaneously, then the time-frequency mask is derived, which obtains impressive performance. In this case, the equal weight parameters are adopted and good estimation for both targets are required to ensure the accuracy of mask estimation. In our proposed multi-task scheme, the SPP is incorporated as an auxiliary target in model training and only the the primary task is concerned with the speech enhancement performance, and the adaptive weighting method is considered to optimize the accuracy of primary target estimation.

4. Experimental Setup and Result

In this section, firstly the performance of all single-task techniques discussed in Sect. 2 is compared on the same datasets and DNN architectures. 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 [13], while the performance of the more recently proposed interference PSD and SIR estimation techniques has not been considered. Further, the performance of the proposed multi-task framework using SPP as a secondary task is investigated.

4.1 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. The clean utterances are from the TIMIT database [24].

For the reverberant dataset, 500 clean utterances are convolved with 16 room impulse responses (RIRs) to comprise 8000 training utterances totally. The validation dataset is generated by convolving 200 clean utterances with 8 RIRs, resulting in 1600 utterances totally. The test dataset is generated by convolving 200 clean utterances with 8 RIRs, resulting in 1600 utterances totally. There is no overlap between utterance files for different sets. The RIRs are selected from multiple databases measured in real environments [25]–[28]. The reverberation times of RIRs used in train, validation and test datasets are listed in Table 1, which range from 200 ms to 1 s.

Table 1 Reverberation times of training dataset, validation dataset and test dataset in the reverberant dataset.

Database	T_{60} (ms)
Train set	200, 250, 300, 390, 410, 440, 520, 580, 640, 680, 700, 749, 800, 880, 930, 1000
Validation set	220, 370, 450, 570, 650, 730, 850, 980
Test set	280, 360, 430, 560, 670, 760, 830, 910

To generate the noisy dataset, clean utterances are firstly convolved with the measured RIR and corrupted with different noise types from the DEMAND database [29]. For the training, validation, and test sets, we have used 250, 100, and [illegible] clean speeches convolved with an RIR with reverberation time 580 ms, 570 ms, and 560 ms, respectively. There is no overlap between the clean speech files and the RIRs for different datasets. Further, for the training, validation, and test sets, 5 different noise types (DKITCHEN, NPARK, TBUS, PCAFETER, OMEETING) at 3 different broadband signal-to-noise ratio (SNR) are added to the reverberant signals, with SNR $\in \{-5\text{dB}, 0\text{dB}, 5\text{dB}\}$. Every noise signal is divided into 3 parts for training, validation, and test sets respectively, and for every utterance, a random [illegible] from the noisy signal is used. To analyze the generalization capabilities of the proposed models, an unseen noisy test dataset is also generated by adding 3 unseen noise types (DLIVING, OHALLWAY, PSTATION) at unseen broadband SNRs to the test reverberant signals, with SNR$\in\{-3\text{dB}, 3\text{dB}, 10\text{dB}\}$.

4.2 Algorithmic Settings, Network Settings and Metrics

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% at a sampling frequency $f_s = 16\,\text{kHz}$. 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 $T = 3$. The PSDs $\Phi_y^2(k,l)$, $\Phi_x^2(k,l)$ and $\Phi_i^2(k,l)$ are computed as in (7) using recursive averaging with a smoothing factor of $\beta = 0.85$. In the interference PSD estimation, to compute the estimated SIR, we use the decision-directed approach represented in (11) with a smoothing factor of $\alpha = 0.9$. The values of α, β and T have been fine tuned according to the effectiveness. For the Wiener gain mask in (12) and (15), a minimum gain of -12 dB is used. To compute the SPP in (16), we use $P(\mathcal{H}_1) = 0.5$, $P(\mathcal{H}_0) = 0.5$, and $10\log_{10}\xi_{\mathcal{H}_1} = 15$ dB according to [18]. Prior to training, the input vectors $\boldsymbol{y}(l)$, $\boldsymbol{\Phi}_y^2(l)$ and target vectors $\boldsymbol{x}(l)$, $\boldsymbol{\Phi}_i^2(l)$, $\boldsymbol{\xi}(l)$ are transformed to the log-domain and are globally normalized into zero mean and unit variance.

To exploit network's ability of statistical estimation, we adopt networks with 2 and 3 hidden layers respectively.

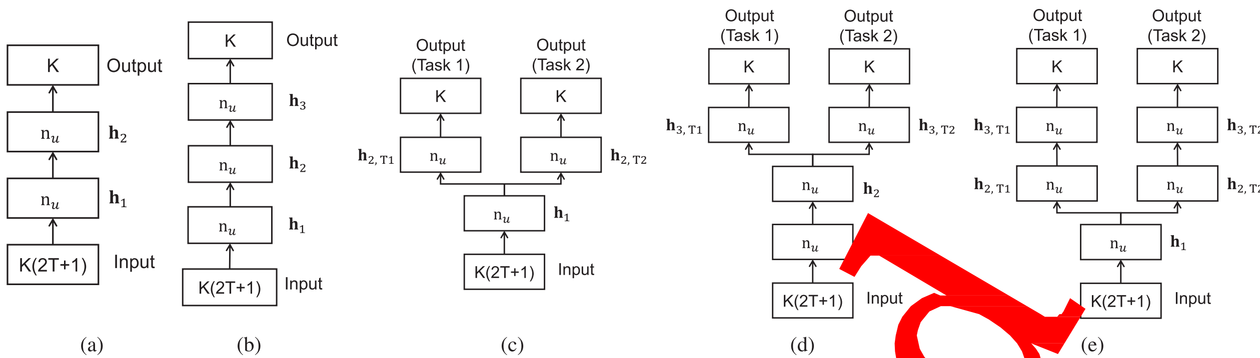


Fig. 1 Schematic illustration of the considered DNN architectures: (a) single-task estimation with two layers, (b) single-task estimation with three layers, (c) multi-task estimation with one shared layer followed by one task-specific layer, (d) multi-task estimation with two shared layers followed by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

Specifically, for single-task statistical estimation, the considered network architectures are shown in Figs. 1 (a), (b), where n_u is the unit number of hidden layers. For multi-task statistical estimation, the considered network architectures are shown in Figs. 1 (c)–(e). As previously mentioned, two different tasks are jointly learned, with the 1st task being the estimation of a target as presented in Sect. 2 and the 2nd task being the estimation of the SPP in (17). In Fig. 1 (c) both tasks share a 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 a hidden layer followed by two task-specific layers. For all architectures, we use rectifying linear unit (ReLU) as non-linearity on all hidden layers and input layers. For estimating an unbounded target (i.e., the desired 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 squared error is used as the loss function for training the single-task networks in Figs. 1 (a), (b) and as the loss function $\mathcal{L}_1$ for training the multi-task networks in Figs. 1 (c)–(e). Cross-entropy loss is used as the loss function $\mathcal{L}_2$ for training the multi-task networks. All considered architectures are trained for different n_u using the Adam optimizer with different hyper-parameters, i.e., learning rate l_r and weight decay w_d. 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.

The dereverberation and denoising performance is measured by the perceptual evaluation of speech quality (PESQ) [30] and frequency-weighted segmental signal to noise ratio (fwSSNR) [31].

4.3 Performance Evaluation of Multi-Task Loss Functions

For the multi-task DNN, the loss function is defined as a weighted sum of the task-specific loss functions of two tasks. To compare the performance of adaptive weighting in (19) and fixed weighting in (18) for the multi-task loss function, the two weighting methods is used for jointly direct mask estimation and SPP estimation on noisy dataset. In fixed weighting method, we set $\lambda_1 = 1$ and $\lambda_2 \in \{0.001, 0.01, 0.1, 1, 10, 100\}$. For all considered fixed weight parameters and the adaptive weighting, the two-layer network depicted in Fig. 1 (c) is trained for $n_u = 500$ and different hyper-parameters, i.e., learning rate $l_r \in \{0.001, 0.0001\}$ and weight decay $w_d \in \{0, 0.001\}$. The final network is selected as the one yielding the minimum validation loss. The average PESQ and fwSSNR scores obtained on the test noisy and unseen noisy datasets, as well as the scores of unprocessed input speech are presented in Table 2. From the table, adaptive weighting outperforms fixed weighting at any values of $\{\lambda_1, \lambda_2\}$ on all the considered test datasets and metrics. Hence the adaptive weighted loss in (19) is adopted in the following experiment.

Table 2 Performance of jointly direct mask estimation and SPP estimation using loss function in (18) (with $\lambda_1 = 1$) and (19) on the test noisy and unseen noisy datasets.

Weighting Method		Noisy Test Dataset fwSSNR	Noisy Test Dataset PESQ	Unseen Noisy Test Dataset fwSSNR	Unseen Noisy Test Dataset PESQ
Unprocessed		3.26	1.21	4.24	1.26
Fixed λ_2	0.001	5.82	1.43	6.41	1.48
	0.01	5.76	1.42	6.34	1.47
	0.1	6.10	1.44	6.73	1.49
	1	6.05	1.44	6.74	1.50
	10	6.12	**1.45**	6.70	**1.51**
	100	6.13	1.42	6.74	1.47
Adaptive		**6.26**	**1.45**	**6.88**	**1.51**

4.4 Performance Evaluation of Single-Task and Proposed Multi-Task Techniques

In this section, the performance of the single-task techniques presented in Sect. 2 is evaluated and compared. The single-task techniques are referred to as $\hat{X}_{\text{mag}}$, $\hat{X}_{\text{gain}}$, $\hat{X}_{\text{psd}}$ and $\hat{X}_{\text{sir}}$

Table 3 Performance comparison for single-task and proposed multi-task techniques of different targets on the test reverberant dataset.

fwSSNR		Unprocessed	$\hat{X}_{\rm mag}$	$\hat{X}_{\rm gain}$	$\hat{X}_{\rm psd}$	$\hat{X}_{\rm sir}$	$\hat{X}^{\rm M}_{\rm mag}$	$\hat{X}^{\rm M}_{\rm gain}$	$\hat{X}^{\rm M}_{\rm psd}$	$\hat{X}^{\rm M}_{\rm sir}$
T_{60} (ms)	280	9.69	11.30	12.29	10.24	10.62	11.37	12.46	10.13	10.82
	360	12.98	12.78	13.33	12.93	12.85	12.59	13.56	12.73	12.98
	430	8.37	11.33	11.75	9.71	10.07	11.08	12.02	9.99	10.36
	560	10.30	12.07	12.85	11.50	12.23	11.93	12.94	11.49	12.25
	670	10.61	12.46	13.29	10.69	12.14	12.37	13.58	11.08	12.56
	760	10.67	11.86	11.90	11.34	10.45	11.68	[illegible]	11.53	10.63
	830	2.15	4.86	4.90	4.67	4.75	4.78	[illegible]	[illegible]	4.90
	910	4.78	7.31	7.45	7.06	7.00	7.25	7.48	[illegible]	7.02
Average		8.70	10.50	10.97	9.77	10.01	10.38	[illegible]	9.85	[illegible]
PESQ		Unprocessed	$\hat{X}_{\rm mag}$	$\hat{X}_{\rm gain}$	$\hat{X}_{\rm psd}$	$\hat{X}_{\rm sir}$	$\hat{X}^{\rm M}_{\rm mag}$	$\hat{X}^{\rm M}_{\rm gain}$	$\hat{X}^{\rm M}_{\rm psd}$	$\hat{X}^{\rm M}_{\rm sir}$
T_{60} (ms)	280	1.59	1.73	1.82	1.63	1.68	1.72	1.84	[illegible]	1.70
	360	1.73	1.95	2.11	1.96	1.99	1.87	[illegible]	1.98	2.00
	430	1.40	1.63	1.60	1.46	1.49	[illegible]	[illegible]	1.47	1.50
	560	1.50	1.63	1.83	1.76	1.74	[illegible]	[illegible]	[illegible]	1.74
	670	1.63	1.76	1.97	1.81	1.84	[illegible]	[illegible]	[illegible]	1.85
	760	1.49	1.72	1.73	1.63	1.64	1.71	1.74	1.65	1.65
	830	1.32	1.34	1.36	1.35	[illegible]	1.32	[illegible]	1.36	1.37
	910	1.25	1.32	1.34	1.34	1.32	[illegible]	1.34	1.34	1.33
Average		1.49	1.63	1.72	[illegible]	1.63	[illegible]	[illegible]	1.63	1.64

Table 4 Performance comparison for single-task and proposed [illegible] task techniques of different targets on the test noisy dataset.

fwSSNR		Unprocessed	$\hat{X}_{\rm mag}$	$\hat{X}_{\rm gain}$	$\hat{X}_{\rm psd}$	$\hat{X}_{\rm sir}$	$\hat{X}^{\rm M}_{\rm mag}$	$\hat{X}^{\rm M}_{\rm gain}$	$\hat{X}^{\rm M}_{\rm psd}$	$\hat{X}^{\rm M}_{\rm sir}$
SNR (dB)	-5	1.58	3.96	[illegible]	[illegible]	[illegible]	4.65	4.98	3.12	3.41
	0	3.24	5.20	6.08	[illegible]	[illegible]	5.97	6.50	4.96	5.15
	5	4.94	5.93	7.26	[illegible]	[illegible]	6.83	7.61	6.31	6.43
Noise Type	DKITCHEN	3.37	5.88	[illegible]	4.92	4.85	6.78	7.05	5.02	5.67
	NPARK	2.17	4.46	5.00	3.95	3.66	5.13	5.53	3.93	4.16
	OMETTING	3.08	4.85	5.77	[illegible]	4.11	5.77	6.17	4.63	4.69
	PCAFETER	1.90	4.02	4.68	[illegible]	3.40	4.49	5.07	3.95	3.81
	TBUS	5.78	5.94	[illegible]	[illegible]	6.23	6.92	7.99	6.45	6.65
Average		3.26	[illegible]	5.94	4.90	4.45	5.82	6.36	4.80	5.00
PESQ		Unprocessed	$\hat{X}_{\rm mag}$	[illegible]	$\hat{X}_{\rm psd}$	$\hat{X}_{\rm sir}$	$\hat{X}^{\rm M}_{\rm mag}$	$\hat{X}^{\rm M}_{\rm gain}$	$\hat{X}^{\rm M}_{\rm psd}$	$\hat{X}^{\rm M}_{\rm sir}$
SNR (dB)	-5	1.12	1.19	[illegible]	1.22	1.19	1.24	1.33	1.22	1.21
	0	[illegible]	1.25	1.43	1.34	1.31	1.32	1.45	1.34	1.32
	5	1.29	[illegible]	1.54	1.46	1.42	1.37	1.56	1.46	1.43
Noise Type	DKITCHEN	1.23	1.33	1.50	1.38	1.33	1.43	1.53	1.39	1.37
	NPARK	[illegible]	1.23	1.38	1.30	1.27	1.31	1.40	1.30	1.29
	OMETTING	[illegible]	[illegible]	1.33	1.26	1.23	1.24	1.36	1.26	1.24
	PCAFETER	[illegible]	1.19	1.30	1.26	1.23	1.23	1.32	1.26	1.24
	TBUS	[illegible]	1.28	1.61	1.50	1.45	1.35	1.63	1.50	1.45
Average		[illegible]	1.25	1.43	1.34	1.30	1.31	1.45	1.34	1.32

respectively. The net[illegible]rks depicted in F[illegible]s. 1 (a), (b) are considered as the single[illegible]k networks. M[illegible]eover, to evaluate the performance of pr[illegible] multi-t[illegible]k scheme, the different targets presented in Se[illegible]ointly estimated with SPP as the secondary target. The multi-task techniques are referred to as $\hat{X}^{M}_{\rm mag}$, $\hat{X}^{M}_{\rm gain}$, $\hat{X}^{M}_{\rm psd}$ and $\hat{X}^{M}_{\rm sir}$ respectively. The networks depicted in Figs. 1 (c)–(e) are considered as the multi-task networks.

For each technique, the considered networks are trained for several numbers of hidden units $n_u \in \{500, 1000, 1500\}$ and different hyper-parameters, i.e., learning rate $l_r \in \{0.001, 0.0001\}$ and weight decay $w_d \in \{0, 0.001\}$. The final network is selected as the one yielding the minimum validation loss. Tables 3–5 presents the average fwSSNR and PESQ scores of the unprocessed input speech and the enhanced speech processed by single-task or proposed multi-task techniques on the test reverberant, noisy, and unseen noisy datasets respectively, where the results for each reverberation time, noise type, and SNR are also listed.

In single-task experiment, compared to the unprocessed speech, all considered techniques generally yield an improvement in PESQ and fwSSNR on all datasets, with the direct mask estimation technique (i.e., $\hat{X}_{\rm gain}$) yielding the best performance. The advantageous performance of the direct mask estimation technique in comparison to magnitude estimation was already established in [13]. How-

Table 5 Performance comparison for single-task and proposed multi-task techniques of different targets on the test unseen noisy dataset.

fwSSNR		Unprocessed	$\hat{X}_{mag}$	$\hat{X}_{gain}$	$\hat{X}_{psd}$	$\hat{X}_{sir}$	$\hat{X}^{M}_{mag}$	$\hat{X}^{M}_{gain}$	$\hat{X}^{M}_{psd}$	$\hat{X}^{M}_{sir}$
SNR (dB)	-3	2.08	4.12	4.94	4.09	3.53	4.95	5.45	3.99	4.01
	3	4.21	5.47	6.79	6.08	5.38	6.44	7.13	5.89	5.88
	10	6.42	6.29	8.11	7.66	6.72	7.21	8.34	7.29	7.26
Noise Type	DLIVING	4.16	5.31	6.61	5.86	5.09	6.22	6.95	5.64	5.63
	OHALLWAY	5.70	5.94	7.72	6.96	6.24	6.96	7.99	6.68	6.78
	PSTATION	2.85	4.64	5.50	5.00	4.30	[illegible]	[illegible]	4.84	4.73
Average		4.24	5.29	6.61	5.94	5.21	6.20	[illegible]	[illegible]	5.71
PESQ		Unprocessed	$\hat{X}_{mag}$	$\hat{X}_{gain}$	$\hat{X}_{psd}$	$\hat{X}_{sir}$	$\hat{X}^{M}_{mag}$	$\hat{X}^{M}_{gain}$	$\hat{X}^{M}_{psd}$	$\hat{X}^{M}_{sir}$
SNR (dB)	-3	1.15	1.20	1.35	1.28	1.24	1.26	[illegible]	1.28	1.25
	3	1.26	1.27	1.50	1.43	1.39	1.34	[illegible]	1.43	1.40
	10	1.37	1.31	1.63	1.56	1.51	1.39	1.64	[illegible]	1.52
Noise Type	DLIVING	1.24	1.25	1.47	1.40	1.35	[illegible]	[illegible]	1.40	1.36
	OHALLWAY	1.34	1.29	1.60	1.53	1.47	[illegible]	[illegible]	1.52	1.48
	PSTATION	1.20	1.24	1.41	1.35	1.31	[illegible]	1.44	1.35	1.32
Average		1.26	1.26	1.49	[illegible]	1.38	[illegible]	1.52	1.42	1.39

ever, 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. Additionally, except magnitude estimation on reverberant dataset and interference PSD estimation on the noisy dataset, the proposed multi-task scheme outperforms the traditional single-task scheme in most cases. And the score improvement for each reverberation time, noise type, and SNR is generally balanced. Among all the techniques using single-task and multi-task schemes, the technique of jointly direct mask estimation and SPP estimation obtains the best performance.

5. Conclusion

In this paper, multi-task learning using SPP as the secondary target has been proposed to improve the accuracy and generalization of supervised DNN-based single-channel speech enhancement techniques. Instead of only estimating a user-defined target (e.g., the desired signal [illegible], a time-frequency mask such as the Wiener gain derived directly or from the interference PSD, or the [illegible]), the SPP serves as the secondary task to [illegible] the domain [illegible] information for the main task. In the multi-task scheme, we have used a recently proposed adaptive weighting method of losses derived from the homoscedastic uncertainty of tasks. The simulation results result proves that the proposed multi-task learning framework outperforms single-task learning on most test [illegible]. And direct mask approximation jointly with SPP estimation outperforms other state-of-art techniques in all of the reverberant and noisy test datasets.

Acknowledgments

This work was 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.

References

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," IEEE Signal Processing Magazine, vol.29, no.[illegible], pp.114–126, Oct. 2012.

[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," Proc. International Workshop on Acoustic Signal Enhancement, Juan les Pins, France, pp.332–336, Nov. 2014.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on acoustics, speech, and signal processing, vol.27, no.2, pp.113–120, April 1979.

[4] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proceedings of the IEEE, vol.67, no.12, pp.1586–1604, Dec. 1979.

[5] 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.

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," IEEE Transactions on Audio, Speech, and Language Processing, vol.18, no.7, pp.1717–1731, Aug. 2010.

[7] 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, Oct. 2012.

[8] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol.20, no.1, pp.30–42, Jan. 2012.

[9] 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.

[10] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A study on target feature activation and normalization and their impacts on the performance

of DNN based speech dereverberation systems,” Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Jeju, South Korea, pp.1–4, Dec. 2016.

[11] B. Wu, K. Li, M. Yang, and C.-H. 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.

[12] K. Han and D.L. Wang, “A classification based approach to speech segregation,” Journal of the Acoustical Society of America, vol.132, no.5, p.3475, Nov. 2012.

[13] 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.

[14] Z.-Q. Wang, P. 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.

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

[16] 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.

[17] G.W. Lee and H.K. Hong, “Multi-task learning u-net for single-channel speech enhancement and mask-based voice activity detection,” Applied Sciences, vol.10, no.9, p.3230, May 2020.

[18] 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.

[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,” Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, [illegible], NY, USA, pp.145–148, Oct. 2011.

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

[22] 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.[illegible], no.[illegible], pp.[illegible]–1726, May 2018.

[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 Consorti[illegible] 1992.

[25] M. Jeub, M. Schafer, and P. [illegible] room impulse response database for the evaluation of dereverberation algorithms,” 16th International Conference on Digital Signal Processing, Santorini, Greece, pp.1–5, July 2009.

[26] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, “The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, pp.1–4, Oct. 2013.

[27] J. Eaton, N.D. Gaubitch, A.H. Moore, and P.A. Naylor, “The ace challenge – corpus description and performance evaluation,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, pp.1–5, Oct. 2015.

[28] “Echothief impulse response library,” www.echothief.com/downloads/, accessed: 2019-02-26.

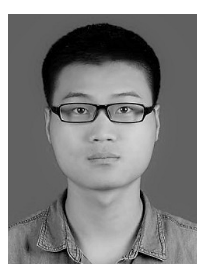
[29] J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings,” Journal of the Acoustical Society of America, vol.133, no.5, May 2013.

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

[31] Y. Hu and P.C. Loizou, “Evaluation of objective quality measures for speech enhancement,” IEEE Transactions on Audio, Speech, and Language Processing, [illegible] pp.229–238, Jan. 2008.

[illegible]ei Wang received the B.S. in Shandong University in 2016. She is currently a Ph.D. student in the School of Electronic Information and Electrical Engineering at Shanghai Jiao Tong University (SJTU). Her research interests are in microphone array signal processing, speech enhancement and dereverberation.

Jie Zhu received the Ph.D. Degree in engineering from Shanghai Jiao Tong University (SJTU), worked as professor and doctoral supervisor in SJTU from 1999. He used to be engaged in cooperative research in Bell Labs in US, Dresden University Technology (TUD) in Germany, Kwangju Institute of Science and Technology (GIST) in South Korea and Waseda University in Japan, led the research and development of Speech Signal Processing in a long term. Currently he serves as the Vice-Director of Institute of Signal Processing, SJTU, the Vice-chairman of the Shanghai branch of the International Society for Engineering and Technology (IET) and the President of the Shanghai branch of IEICE.



Kangbo Sun is a Ph.D. student at Shanghai Jiaotong University. He received a bachelor’s degree from Northwestern Polytechnical University in China in 2018. His research interests include computer vision, deep learning, image and video analysis, action recognition, and abnormal behavior detection.