SERIL: Noise Adaptive Speech Enhancement using Regularization-based Incremental Learning

Chi-Chang Lee1,2, Yu-Chen Lin1,2, Hsuan-Tien Lin1, Hsin-Min Wang3, Yu Tsao2
1Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
2Research Center for Information Technology Innovation, Academia Sinica, Taiwan
3Institute of Information Science, Academia Sinica, Taiwan
r08922a27@csie.ntu.edu.tw, f04922077@csie.ntu.edu.tw, htlin@csie.ntu.edu.tw,
whm@iis.sinica.edu.tw, yu.tsao@citi.sinica.edu.tw

Abstract

Numerous noise adaptation techniques have been proposed to fine-tune deep-learning models in speech enhancement (SE) for mismatched noise environments. Nevertheless, adaptation to a new environment may lead to catastrophic forgetting of the previously learned environments. The catastrophic forgetting issue degrades the performance of SE in real-world embedded devices, which often revisit previous noise environments. The nature of embedded devices does not allow solving the issue with additional storage of all pre-trained models or earlier training data. In this paper, we propose a regularization-based incremental learning strategy to help noise adaptation issue. The proposed method is termed SERIL. SERIL exploits the advantages of two well-known incremental learning algorithms: (1) whole past optimization path information and (2) curvature-based strategy. We evaluated SERIL using two datasets: the Voice Corpus Bank corpus (VCB) and the TIMIT corpus. The experimental results show that SERIL and the direct fine-tuning approach both effectively adapt the SE model to new environments and improve SE performance. Moreover, compared to the direct fine-tuning approach, SERIL maintained high SE performance against all previously learned types of noise, thus effectively addressing the catastrophic forgetting problem.

The remainder of this paper is organized as follows. Section 2 presents some related work and explains the motivation of using incremental learning strategy to help noise adaptation issue on speech enhancement. In Section 3, we detail the philosophies of the proposed SERIL system. The experimental setup and results are then reported in Section 4. Finally, Section 5 presents some concluding remarks.
2. Related Work and Motivation

An intuitive SE method to overcome the mismatch problem is to collect as many types of noise as possible to increase the generalization ability [14]. However, it is impractical to cover the infinite types of noise that may be encountered in real situations. Several researches [20, 21] have been proposed to directly fine-tune a pre-trained model to improve the performance in a target domain. When entering a new circumstance, these algorithms only focus on the current noise domain, and ignore the memory of the previously learned noise types. In many applications, such as edge-devices, the type of noise changes frequently, and it is common to re-encounter learned types of noise. However, the adapted SE model cannot perform well in the previously learned noise types. This effect is called catastrophic forgetting [22, 23]. Although the SE model can be fine-tuned every time the environment is changed, the repeated model adaptation process will result in high computation and time costs.

The above limitations of adaptive methods based on direct fine-tuning motivated us to apply the incremental learning algorithm to SE. Incremental learning is also known as continuous learning or lifelong learning. Figure 1 illustrates the relationship between direct fine-tuning and incremental learning. Training trajectories are illustrated in a schematic parameter space, with parameter regions leading to good performance on the source (yellow region) and target (blue region), denoted as tasks $S$ and $T$, respectively. After learning in task $S$, the parameters are located in $\theta_S$. As shown by the dashed arrow in Figure 1 when the SE model is adapted by taking gradient steps to minimize the loss based on task $T$ alone, the resulting $\theta_T$ is beyond the good performance area of Task $S$, i.e., what is already learned in Task $T$ is forgotten. In contrast, in incremental learning, the SE model weights are updated to the target domain while retaining the knowledge learned from the source domain. This is often realized by finding the overlapping region of the source and target domains. The learning trajectory of incremental learning shown by the solid arrow in Figure 1 illustrates this concept. In this way, incremental learning can help the resulting model provide good SE results in the target domain while maintaining satisfactory performance in the source domain.

3. The SERIL System

3.1. Architecture and loss function of the SERIL system

The architecture of the SERIL system is depicted in Figure 2. The system performs SE in the spectral domain. Speech waveforms are first converted into time-frequency features using a 512-point short-time Fourier transform (STFT) with a hamming window size of 32 ms and a hop size of 16 ms. Each feature vector consists of 257 elements. The enhanced spectral features are then converted into the waveform domain by inverse STFT with an overlap-add method. In the SERIL system, the first 3 layers are LSTM layers (one-directional LSTM was used for achieving real-time inference). The hidden dimension of each LSTM is 257. A fully connected layer is concatenated to the output of the last LSTM layer for scaling.

![Figure 1: Relationship between fine-tuning and incremental learning from source noise domain to unseen target domain.](image)

Given the noisy spectral features, $Y$, the SE model aims to generate enhanced spectral features, $\hat{X}$. $\alpha$ is computed by $(X - \hat{X})/\|X\|^2$, where $X$ is the target clean spectral features. In addition, $f_0(\cdot)$ is equal to $X$; thus, we denote our loss function $-SDR^{STSA}(f_0(Y), X)$ as $l_0(Y)$.

3.2. Curvature-based regularization strategy

Considering the losses in the previous and new acoustic environments, $L_{old}$ and $L_{new}$, respectively, the total loss can be formulated as:

$$L(\theta) = L_{new}(\theta) + L_{old}(\theta).$$

Because the training data of the previous environment is usually not accessible online, we cannot calculate $L_{old}(\theta)$. Instead, we can assume that the loss of the previous environment can be revealed from the learned SE model, $\theta$. By approximating $L_{old}$ using the second-order Taylor expansion at $\theta = \theta^*$, we have

$$L_{old}(\theta) \approx L_{old}(\theta^*) + \delta\theta^T \nabla_{\theta} L_{old}(\theta^*) + \frac{1}{2} \delta\theta^T H(\theta^*) \delta\theta,$$

where $\delta\theta$ is $\theta - \theta^*$; $H(\theta^*)$ is the Hessian matrix of $L_{old}$ at $\theta = \theta^*$; and $L_{old}(\theta^*)$ is a constant. Because the elements in $\nabla_{\theta} L_{old}(\theta^*)$ are generally small enough to be ignored, we can obtain the approximate form as $L_{old}(\theta) \approx \frac{1}{2} \delta\theta^T H(\theta^*) \delta\theta$. Similar to the elastic weight consolidation [25, 26], we ignore the cross terms in $H(\theta^*)$ to improve computational efficiency. The approximate form becomes

$$H(\theta^*) \approx diag(\mathbb{E}_{Y \sim D_{old}}(\|\nabla_{\theta} L_{old}(Y) (\nabla_{\theta} L_{old}(Y)^T)\|_{\theta = \theta^*}),$$

where $Y$ is the speech sample from the previous environment $D_{old}$. Finally, substituting $L_{new}$ and $L_{old}$ into $L(\theta)$, we have

$$L(\theta) \approx L_{new}(\theta) + \lambda \sum_{i} F_{\theta_i} (\theta_i - \theta_i^*)^2,$$

where $\lambda$ is a hyperparameter; $i$ is the index of the parameters in the model; $\theta_i$ and $\theta_i^*$ are the $i$-th parameters in the current and
previous environments, respectively; and \( F_0 \) is the diagonal element of \( H(\theta') \). The intuitive interpretation of \( F_0 \) is the local curvature, which indicates the sensitivity that affects the performance of the previous acoustic environment.

Kolouri et al. [33] provided a different explanation for the geometric view of the regularization term, which can be applied to our scenario. As \( \theta \to \theta'^* + 1/2 \| \theta - \theta'^* \|_F^2 \) can be interpreted as the expectation of the squared difference of the loss values of the training samples of the previous environment, i.e., \( \sum_{t, i} F_0(\theta_i - \theta'^*_i)^2 \). Similarly to [3], the distance can be approximated by \( \sum F_0(\theta_i - \theta'^*_i)^2 \), which is also derived by the second-order Taylor expansion of \( \sum_{t, i} (1/2)(l_0(\theta_i(Y) - l_0(Y))^2) \). To refer to \([32, 33]\), we apply the interpolation approach to the case of multiple tasks. Given \( F_0 \) derived by all previous tasks, \( \tilde{F}_0 \) is updated as

\[
\tilde{F}_0 = \alpha F_0 + (1 - \alpha) \tilde{F}_0^{-1},
\]

where \( t \) is the index of the task; \( \alpha \) is a hyperparameter in \([0,1]\); \( F_0 \) denotes \( F_0 \) derived from the \((t-1)\)-th task; and \( \tilde{F}_0 \) is the interpolation result of \( F_0^{-1} \) and \( \tilde{F}_0 \), corresponding to the information of past accumulations and curvatures.

### 3.3. Path optimization augmenting approach

Although \( F_0 \) is equipped with rationality to avoid catastrophic forgetting, the commonly used curvature-based methods [32] of deriving \( F_0 \) rely on point estimation, which only capture local curvature information around \( \theta'^* \). In contrast, the path optimization-based method [24] considers the information over the optimization path on the loss surface. In particular, the importance score is determined by accumulating over the entire training trajectory, as illustrated in Figure 3.

![Figure 3: Relationship between the real loss (blue), curvature-based approximate loss (green), and path optimization-based approximate loss (red) while adapting the SE model.](image)

By using the first-order Taylor approximation and setting \( t_s \) and \( t_e \) as the start and end steps of the \( t \)-th task, the change in loss \( L \) over the time from \( t_s \) to \( t_e \) can be written as

\[
L(\theta(t_e)) - L(\theta(t_s)) \approx \int_{t_s}^{t_e} \left( \nabla_\theta L(\theta(t)) \cdot \frac{\partial L(\theta(t))}{\partial \theta} dt \right) dt = \int_{t_s}^{t_e} \left( \int_{t_s}^{t_e} \nabla_\theta L(\theta(t)) \cdot \frac{\partial L(\theta(t))}{\partial \theta} dt \right) dt,
\]

where \( i \) is the index of the SE model parameter. To simplify the description, we denote \( (t_s, \theta_i(t_s), \frac{\partial L(\theta(t))}{\partial \theta} dt) \) as \( -\Delta L_i \). Therefore, the change in the total loss of the \( \Delta L(t) \) associated with each parameter. We put a minus sign on the left side of \( \Delta L(t) \) to make the sign consistent with the regularization term. Practically, replace \( \int_{t_s}^{t_e} \frac{\partial L(\theta(t))}{\partial \theta} dt \) with \( \int_{t_s}^{t_e-1} \frac{\partial L(\theta(t+1)}{\partial \theta} dt \), where \( t \) is the index of iteration. From [24], the definition of importance scores as we begin to train the \( t \)-th task can be defined as

\[
S^{t}_t = \sum_{i, t'} \frac{\Delta L_i^{t'}}{\Delta L^{t'}_i^2 + \epsilon},
\]

where \( t' \) is the index of the task before the \( t \)-th task; \( \theta'^*_t \) is the \( t \)-th parameter of the SE model derived from training the \( t' \)-th task; \( \Delta L^{t'}_i \) is \( \theta'^*_t - \theta^{t-1}_t \); and \( \epsilon \) is a hyperparameter with a positive value.

Similar to [34], we combined the advantages of curvature-based [32, 34] and path optimization-based [24] approaches. The importance of parameter \( \theta \), when training the \( t \)-th task can be written as \((1 - \beta) \tilde{F}_0 + \beta S^{t}_t \). Therefore, the training loss is defined as:

\[
\tilde{L}^{t}(\theta) = L^{t}(\theta) + \lambda \sum_{i}(1 - \beta) \tilde{F}_0 + \beta S^{t}_t (\theta_i - \theta^{t-1}_i)^2,
\]

where \( t \) is the index of the task (if \( t \) is zero, \( \tilde{L}^{t}(\theta) \) is equivalent to \( L^{t}(\theta) \); \( \theta^{t-1} \) is the \( t \)-th parameter after training the \((t-1)\)-th task; and \( \beta \) is a scalar with the value in \([0,1]\), which determines the weight of the two strategies.

### 4. Experimental and Analysis

#### 4.1. Experimental Setup

We evaluated the proposed SERIL system on two speech corpora: VCB [25] and TIMIT [27]. Three data sets were prepared, namely, the training, adaptation, and testing sets. For the training set, 2,000 utterances were randomly selected from the VCB corpus. Each utterance was contaminated with one noise type at a particular SNR level (one out of the 6 SNR levels was randomly specified). For each of the testing sets, we selected 1,680 utterances, and each utterance was contaminated with one noise type at a particular SNR level (one out of the 6 SNR levels was randomly specified). Our implementation is publicly available for reproducibility.

Three standardized evaluation metrics were used to measure the performance: perceptual evaluation of speech quality (PESQ) [17], short-time objective intelligibility measure (STOI) [38], and extended STOI (eSTOI) [39]. PESQ was designed to evaluate the quality of processed speech. The higher the PESQ, the better the speech quality. Both STOI and eSTOI were designed to compute the speech intelligibility. The higher

[1] https://github.com/ChangLee0903/SERIL
STOI and eSTOI scores, the better the speech intelligibility. In addition, we also reported the SDR\textsuperscript{STSA} scores to illustrate the learning process. The higher the SDR\textsuperscript{STSA} score, the smaller the distortion of the spectral features.

4.2. Experimental Results

First, we compared SERIL and the direct fine-tuning approach in terms of the adaptation capability and the degree of catastrophic forgetting. We used the training set $T_0$ to train one baseline model, termed $M_0$. Then, based on the four adaptation sets, we sequentially adapted the model from $M_0$ to $M_4$ by using $T_1$, $M_1$ to $M_2$ by using $T_2$, $M_2$ to $M_3$ by using $T_3$, and $M_3$ to $M_4$ by using $T_4$. The five models ($M_0$ to $M_4$) were then tested on the five testing sets ($E_0$ to $E_4$). The SDR\textsuperscript{STSA} scores of the five models tested on the five testing sets are shown in Figure 4. The results of the baseline model without adaptation and the scores of unprocessed noisy speech are also given for comparison.

From the figure, we note that although the baseline model $M_0$ performs well on $E_0$, where the noise types and SNR levels are matched during the training and testing stages, notable degradation is observed for the mismatched conditions (cf. the gray lines on $E_1$ to $E_4$). Further, both SERIL and the direct fine-tuning approach effectively adapt the SE model to each target domain and achieve good performance. For example, in Figure 4(b), $M_1$ achieves the best performance on $E_1$ for both SERIL and the direct fine-tuning approach. The model trained by direct fine-tuning tends to forget the previously learned SE capability, whereas the model trained by SERIL can maintain good SE performance for previously learned noise types. For instance, in Figure 4(b), the performance of $M_1$ trained by direct fine-tuning is considerably reduced in $E_1$, showing that the adapted model has “forgotten” the SE capability for the previously learned noise type. This is because each noise type has different structural characteristics in different frequency bands, so direct fine-tuning without proper constraints can severely distort the modeling of previous noise environments. In contrast, the performance drop of the SERIL system for the same testing set is relatively minor. Consistent trends can be observed for all testing sets.

Table 1 shows the SDR\textsuperscript{STSA}, PESQ, STOI, and eSTOI scores of the final model ($M_4$) learned using the fine-tuning method and SERIL on the five testing sets. The scores of unprocessed noisy speech and the baseline model without adaptation ($M_0$) are also listed for comparison. Several observations can be drawn from the table. First, SERIL performs as well as direct fine-tuning in the current noisy environment in terms of all metrics (cf. the “clap” column in Table 1). Second, SERIL always outperforms direct fine-tuning for previous environments in terms of all metrics (cf. the “original” to “footsteps” columns in Table 1). Third, SERIL performs better than the baseline model in all testing environments except for “original”, which is under a matched training-testing condition for the baseline model. It is worth noting that compared with the direct fine-tuning approach, SERIL requires only a small amount of additional computational cost and storage to set the constraints when performing model adaptation. However, SERIL can produce performance comparable to the direct fine-tuning approach in each new environment while overcoming the catastrophic forgetting problem in old environments.

5. Concluding Remarks

When deploying an SE system in real-world applications, it is common to encounter a new noisy environment and re-visit to previous noisy environments. Although the direct fine-tuning approach can effectively adapt SE models to new environments, the adapted SE model may suffer from the catastrophic forgetting problem. The proposed SERIL model not only yields comparable performance to the direct fine-tuning approach but also effectively overcomes the catastrophic forgetting problem. To the best of our knowledge, this paper is the first work that incorporates incremental learning into SE tasks. Our experimental results confirmed the effectiveness of the proposed SERIL system for SE model adaptation and avoiding catastrophic forgetting. Based on the promising results, we believe that the proposed SERIL model can be used in various edge-computing devices, where the acoustic condition changes frequently and the cost of online retraining is high. In addition, we note that using an appropriate weight, $\lambda$, to combine the curvature-based and path optimization-based strategies can provide better SE performance in most tasks. Derivation of an algorithm that can automatically determine the optimal $\lambda$ is worthy of further study.

![Figure 4: SDR\textsuperscript{STSA} scores of incrementally learned models evaluated on five testing sets. The x-axis lists incrementally learned models $M_0$, $M_1$, $M_2$, $M_3$, and $M_4$. The y-axis presents the SDR\textsuperscript{STSA} score. The scores of the unprocessed noisy speech, baseline model, direct fine-tuning approach, and proposed SERIL are represented by yellow, gray, black, and blue lines, respectively.](image)
6. References


