Multi-View Attention Transfer for Efficient Speech Enhancement

Wooseok Shin†, Hyun Joon Park†, Jin Sob Kim, Byung Hoon Lee, Sung Won Han∗

School of Industrial and Management Engineering, Korea University, Seoul, Republic of Korea

{wsshin95, winddori2002, jinsob, qlascm777, swhan}@korea.ac.kr

Abstract

Recent deep learning models have achieved high performance in speech enhancement; however, it is still challenging to obtain a fast and low-complexity model without significant performance degradation. Previous knowledge distillation studies on speech enhancement could not solve this problem because their output distillation methods do not fit the speech enhancement task in some aspects. In this study, we propose multi-view attention transfer (MV-AT), a feature-based distillation, to obtain efficient speech enhancement models in the time domain. Based on the multi-view features extraction model, MV-AT transfers multi-view knowledge of the teacher network to the student network without additional parameters. The experimental results show that the proposed method consistently improved the performance of student models of various sizes on the Valenti and deep noise suppression (DNS) datasets. MANNER-S-8.1GF with our proposed method, a lightweight model for efficient deployment, achieved 15.4× and 4.71× fewer parameters and floating-point operations (FLOPs), respectively, compared to the baseline model with similar performance.

Index Terms: speech enhancement, multi-view knowledge distillation, feature distillation, time domain, low complexity

1. Introduction

Speech enhancement (SE) involves the removal of background noise to improve the perceptual quality of noisy speech. SE is a fundamental task in various speech-processing applications, such as automatic speech recognition, hearing aids, and tele-conference systems. Recently, deep learning-based methods have been applied and have significantly improved the performance of SE. These methods can be divided into time domain and time-frequency (TF) domain approaches. TF domain methods use the spectral features obtained by applying the short-time Fourier transform to the signal to estimate clean speech [1, 2, 3, 4, 5]. In contrast, time domain methods use a raw waveform containing implicitly all of the signal’s information to predict clean speech [6, 7, 8, 9]. Although TF and time domain methods have achieved significant improvements in SE, the problem remains that they do not simultaneously satisfy the low computational complexity and model complexity required in various deployment environments while minimizing performance degradation.

To mitigate this issue, knowledge distillation (KD) methods, which transfer useful knowledge from a large teacher network to a compact student network, were first studied in image recognition [10, 11, 12, 13]. After verifying that KD can obtain a low-resource student model with satisfactory performance in the image domain, researchers have applied KD methods to obtain efficient student models for speech relevance tasks such as speech recognition [14, 15, 16], acoustic scene classification [17, 18], and acoustic event detection [19, 20].

In addition, a few studies have exploited the advantages of KD for SE. In the TF domain, [21] proposed a sub-band KD framework to improve a single student network using multiple teachers trained for each sub-band. [22] designed a two-stage training distillation method and a co-worker-based network to improve the performance of SE. In the time domain, to improve performance at both low and high signal-to-noise ratios (SNRs), [23] built multiple teachers trained under SNRs and then transferred knowledge to the student network. Further, [24] applied standard KD [10] to reduce the system latency while preventing performance degradation.

Although the previous KD methods in SE achieved improvements in the model efficiency and performance, they only applied standard KD methods confined to output distillation which is the method to use the soft labels of the teacher as additional supervision of the student. In contrast to standard KD for classification tasks, the disadvantages of standard KD for regression tasks such as SE include the following: i) It cannot take advantage of class distributions (soft labels) in which the student network can easily learn intra-class variations [10]. ii) Because the output of the teacher network is unbounded, it may provide confusing guidance, which may limit performance improvement [25]. Including the above problems of applying standard KD in SE, the insights that the intermediate features provide richer information than outputs [11, 13, 26] suggest the necessity of feature-based KD in SE. In addition, recent SE models [6, 8, 9] based on a dual-path method that extracts local and global contextual information from long-range speech waveforms motivated us to design feature-wise knowledge transfer.

In this study, we propose multi-view attention transfer (MV-AT) to obtain efficient speech enhancement models. To the best of our knowledge, this is the first feature-based KD method for time domain SE. We adopted attention transfer (AT) [12] applying the same transformation to the features of the teacher and student for feature-based distillation as the base feature transformation method. As studied in MANNER [6], which describes the importance of separated signal feature representations (i.e., channel, local, and global), MV-AT based on the MANNER backbone transfers the feature-wise knowledge of the teacher by utilizing each feature highlighted in the multi-view. By applying MV-AT, the student network can easily learn the teacher’s signal representation and mimic the matched representation from each perspective. Furthermore, MV-AT can not only compensate for the limitation of standard KD in SE as feature-based distillation but also make an efficient SE model without additional parameters. The results of experiments conducted on the Valenti and DNS datasets indicate that the proposed method achieves significant efficiency. While exhibiting comparable performance to the baseline model, the model generated by the proposed method required 15.4× and 4.71× fewer parameters and flops, respectively.

†These authors contributed equally to this work.
∗Corresponding author.
2. Method

This section reviews the MANNER [6] architecture used as the base architecture in this study and presents our MV-AT method based on feature distillation.

2.1. Base Architecture

We adopted a state-of-the-art time domain SE model, MANNER, as the base architecture of the teacher and student network for knowledge distillation. As shown in Figure 1, MANNER consists of a multi-layer ($L_T$, $L_S$) encoder and decoder with skip connections. The encoder and decoder contain Down and Up Conv layers, a residual conformer (ResCon) block, and a multi-view attention (MA) block. Given a noisy input $x$, a 1D convolution expands the channel size of $x$. The latent representations are extracted by the encoder applied to $x$, followed by a linear transformation, and are passed to the decoder. While forwarding $x$, the MA block in each encoder and decoder layer emphasizes each channel, global, and local signal representation to consider all the signal information. Finally, the enhanced output $\hat{y}$ is obtained by applying a mask gate followed by convolution. To obtain an efficient student network, we set the student network to be composed of smaller depth, channel dimension, and the number of MA blocks than the teacher network, reducing computational and model complexity.

2.2. Attention Transfer

To utilize AT in domain SE, we define an activation map $A^T \in \mathbb{R}^{C \times t}$ of a pre-trained teacher network (T) and the corresponding activation map $A^S \in \mathbb{R}^{C \times t}$ of a student network (S). Here, $C$ and $t$ denote the channel dimension and signal length. Subsequently, a mapping function $F$ receives the activation map (i.e., $A^T$ or $A^S$) as an input, aggregates it in the channel dimension, and outputs a temporal attention map (TAM) as follows:

$$F(A) = \sum_{i=1}^{C} |A_i|^p \in \mathbb{R}^{1 \times t}$$

where $A_i$ denotes the activation map for the $i^{th}$ channel and $p$, which adjusts the intensity of the TAM, is set to two, as in [12]. AT assumes that the absolute value of the hidden activation can be used as an indicator of importance. Based on this, as shown in Eq. 1, we extract the TAM for transferring knowledge by aggregating channel representations. Then, $F(A)$ is replaced by the $L_2$-normalized form $\frac{F(A)}{||F(A)||_2}$, as done in [12]. We define the AT loss for a pair of teacher-student attention maps using $l_1$ loss (p=1) as follows below. (It is also possible to use the $l_2$ (p=2) loss, but we used the $l_1$ loss in all experiments because $l_1$ is slightly better than $l_2$. See Section 3.3.)

$$L_{AT} = \left\| \frac{F(A^T)}{||F(A^T)||_2} - \frac{F(A^S)}{||F(A^S)||_2} \right\|_p$$

2.3. Multi-View Attention Transfer

As mentioned above, we adopt the MANNER architecture based on the MA block to transfer the diverse knowledge emphasized in the multi-view. Figure 2 presents the process of obtaining temporal attention maps (TAMs) from the MA blocks of the encoder and decoder layers. After each attention in the MA block, the mapping function $F$ is applied to the three representations, followed by normalization. The MV-AT loss is then defined for a teacher-student pair with three TAMs each.

$$L_{MV-AT} = \sum_{i \in \{c,g,l\}} L_{AT_{view}}$$

where $c$, $g$, and $l$ denote channel, global, and local, respectively. This allows the student network to mimic the manner in which the teacher network emphasizes each view of the signal representations. For the total distillation loss, we combine MV-AT with the standard KD loss distilling the network output. The loss is defined as follows:

$$L_{distill} = \sum_{i \in \mathcal{I}} L_{MV-AT}^i + L_{KD}(\hat{y}_T, \hat{y}_S)$$

where $\mathcal{I}$ indicates the index of all teacher-student activation layer pairs for which knowledge is transferred. Regarding $L_{KD}$, we adopt a weighted loss considering the estimated clean and noisy speech, where each loss consists of $l_1$ loss and multi-resolution STFT loss, as done by [6].
Table 1: Objective evaluation results of the proposed method on the Valentini test dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Params (M)</th>
<th>FLOPs (G)</th>
<th>PESQ</th>
<th>STOI(%)</th>
<th>CSIG</th>
<th>CBAK</th>
<th>COVL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANNER-B-Teacher</td>
<td>24.07</td>
<td>61.92</td>
<td>3.21</td>
<td>95</td>
<td>4.53</td>
<td>3.65</td>
<td>3.91</td>
</tr>
<tr>
<td>MANNER-S-1.5GF (Baseline)</td>
<td>0.26</td>
<td>1.46</td>
<td>2.88</td>
<td>94</td>
<td>4.32</td>
<td>3.47</td>
<td>3.62</td>
</tr>
<tr>
<td>MANNER-S-1.5GF + MV-AT</td>
<td>0.26</td>
<td>1.46</td>
<td>2.94</td>
<td>94</td>
<td>4.34</td>
<td>3.49</td>
<td>3.67</td>
</tr>
<tr>
<td>MANNER-S-1.5GF + MV-AT + KD</td>
<td>0.26</td>
<td>1.46</td>
<td>2.95</td>
<td>94</td>
<td>4.33</td>
<td>3.50</td>
<td>3.66</td>
</tr>
<tr>
<td>DEMUCS-S [7]</td>
<td>18.87</td>
<td>34.56</td>
<td>2.93</td>
<td>95</td>
<td>4.22</td>
<td>3.25</td>
<td>3.52</td>
</tr>
<tr>
<td>TSTNN [8]</td>
<td>0.92</td>
<td>166.54</td>
<td>2.96</td>
<td>95</td>
<td>4.33</td>
<td>3.53</td>
<td>3.67</td>
</tr>
<tr>
<td>MANNER-S-5.3GF (Baseline)</td>
<td>0.90</td>
<td>5.32</td>
<td>2.98</td>
<td>95</td>
<td>4.38</td>
<td>3.54</td>
<td>3.71</td>
</tr>
<tr>
<td>MANNER-S-5.3GF + MV-AT</td>
<td>0.90</td>
<td>5.32</td>
<td>3.04</td>
<td>95</td>
<td>4.42</td>
<td>3.56</td>
<td>3.76</td>
</tr>
<tr>
<td>MANNER-S-5.3GF + MV-AT + KD</td>
<td>0.90</td>
<td>5.32</td>
<td>3.06</td>
<td>95</td>
<td>4.42</td>
<td>3.58</td>
<td>3.77</td>
</tr>
<tr>
<td>MANNER-S-8.1GF (Baseline)</td>
<td>1.38</td>
<td>8.14</td>
<td>3.01</td>
<td>95</td>
<td>4.40</td>
<td>3.55</td>
<td>3.74</td>
</tr>
<tr>
<td>MANNER-S-8.1GF + MV-AT</td>
<td>1.38</td>
<td>8.14</td>
<td>3.07</td>
<td>95</td>
<td>4.43</td>
<td>3.58</td>
<td>3.78</td>
</tr>
<tr>
<td>MANNER-S-8.1GF + MV-AT + KD</td>
<td>1.38</td>
<td>8.14</td>
<td>3.12</td>
<td>95</td>
<td>4.45</td>
<td>3.61</td>
<td>3.82</td>
</tr>
<tr>
<td>MANNER-S-38.GF</td>
<td>21.25</td>
<td>38.34</td>
<td>3.12</td>
<td>95</td>
<td>4.45</td>
<td>3.60</td>
<td>3.81</td>
</tr>
</tbody>
</table>

2.4. Dual-depth Matching

We assume that different depths contain different knowledge and feature distillation of the same location allows the student network to easily mimic the teacher network. If $L_S$ and $L_T$ are the same, the teacher-student features of the same depth position are matched one-to-one when transferring the feature knowledge. By contrast, for efficiency, when $L_S$ is smaller than $L_T$, we utilize a dual-depth knowledge transfer strategy. The basis for this is that a large teacher network can extract sufficiently detailed representations, whereas a student network only learns representative representations in a compact space. For example, if $L_T^{(1-4)}$ and $L_S^{(1-3)}$ are 4 and 3 layer encoder-decoder, the student’s third layer $L_S^3$ contains partial representations of the teacher’s third and fourth layers ($L_T^3$ and $L_T^4$). Thus, we distill both the third- and fourth-layer features of the teacher adjacent to the third-layer feature of the student, called dual-depth knowledge transfer. We adopt interpolation to match the signal lengths of teacher-student attention maps under different layer pairs when we utilize the dual-depth knowledge transfer.

3. Experiments

3.1. Experimental Setup

Implementation details In experiments conducted on the Valentini and DNS datasets, we used MANNER-Base (B) and MANNER-Small (S), respectively, as powerful teacher networks. The depth and channel dimensions of the teacher networks were set to 4 and 60 to transfer the knowledge of enriched representations. We employed MANNER-S, which includes an MA block only in the $L^5$ layer, as a compact student network for efficiency. In addition, small depths and channel dimensions were used to reduce the complexity without changing the structure of the base architecture. To compare student models at various complexities, we suffixed the flop regime to the models (e.g., MANNER-S-1.5GF). Specifically, the MANNER-S-1.5GF, 5.3GF, and 8.1GF models had depths of 3 and channel dimensions of 12, 24, and 30, respectively. Here, GF indicates $10^9$ flops. For training settings, we set a batch size of 4 and 350 epochs to train the student networks. Except for the smaller depth, channel dimension, and the number of MA blocks than those of the teacher networks for achieving high efficiency of the student networks, we adopted the other parameters and learning strategies the same as those of MANNER.

Datasets To compare the proposed method with previous methods, we used the Valentini dataset [27], for which the total time is approximately 10 h. The dataset consists of 30 speakers; we used 28 speakers in the train and validation sets, and the remainder in the test set. The train set comprised 11,572 utterances with four SNRs (15, 10, 5, and 0 dB) of noise. The test set comprised 824 utterances with four SNRs (17.5, 12.5, 7.5, and 2.5 dB) of unseen noise.

To evaluate the proposed method on a larger dataset, we adopted the DNS dataset [28]. For training, we generated 50 h of utterances from the clean speech set and mixed them with noise data at various SNRs. Among the given test sets, we selected the test set that did not contain reverberations for the evaluation. The SNRs of the noise in the test set were distributed between 0 dB and 20 dB. All datasets used were downsampled from 48 to 16 kHz.

Evaluation metrics We evaluated the models in terms of objective measures for the enhanced speech. For objective measures that are used to evaluate speech quality and intelligibility, we adopted the perceptual evaluation of speech quality (PESQ) [29] and short-time objective intelligibility (STOI) [30]. The PESQ score ranges from -0.5 to 4.5, and that of STOI ranges from 0 to 100. As composite measures, we adopted CSIG, CBAK, and COVL [31], which are used to evaluate signal distortion, noise intrusiveness, and overall signal quality, respectively. The scores on the three measures range from 1 to 5. Furthermore, we recorded the floating-point operations (FLOPs) and the number of model parameters to evaluate model efficiency.

3.2. Results

To verify the efficacy of the proposed method, we conducted experiments using different base network configurations on the Valentini and DNS datasets. We compared the performance of three training setups: Base model, Base model+MV-AT, and Base model+MV-AT+KD, where “MV-AT” is the proposed method and “KD” is the standard KD [10] that distills the output. The top row of each table represents the teacher model. As demonstrated in Table 1, the proposed method achieved significant performance improvements without increasing the parameters and flops in the evaluation metrics compared to the baseline. The PESQ scores of the model to which MV-AT was
Table 2: Objective evaluation results of the proposed method on the DNS no reverb test dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Params (M)</th>
<th>FLOPs (G)</th>
<th>PESQ</th>
<th>STOI(%)</th>
<th>CSIG</th>
<th>CBAK</th>
<th>COVL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANNER-S-Teacher</td>
<td>21.25</td>
<td>38.34</td>
<td>3.07</td>
<td>97.1</td>
<td>4.52</td>
<td>3.81</td>
<td>3.84</td>
</tr>
<tr>
<td>MANNER-S-1.5GF (Baseline)</td>
<td>0.26</td>
<td>1.46</td>
<td>2.47</td>
<td>94.8</td>
<td>4.07</td>
<td>3.35</td>
<td>3.27</td>
</tr>
<tr>
<td>MANNER-S-1.5GF + MV-AT</td>
<td>0.26</td>
<td>1.46</td>
<td>2.47</td>
<td>94.9</td>
<td>4.07</td>
<td>3.34</td>
<td>3.27</td>
</tr>
<tr>
<td>MANNER-S-1.5GF + MV-AT + KD</td>
<td>0.26</td>
<td>1.46</td>
<td>2.47</td>
<td>94.9</td>
<td>4.04</td>
<td>3.34</td>
<td>3.25</td>
</tr>
<tr>
<td>MANNER-S-5.3GF (Baseline)</td>
<td>0.90</td>
<td>5.32</td>
<td>2.70</td>
<td>95.9</td>
<td>4.25</td>
<td>3.53</td>
<td>3.48</td>
</tr>
<tr>
<td>MANNER-S-5.3GF + MV-AT</td>
<td>0.90</td>
<td>5.32</td>
<td>2.75</td>
<td>96.1</td>
<td>4.30</td>
<td>3.58</td>
<td>3.55</td>
</tr>
<tr>
<td>MANNER-S-5.3GF + MV-AT + KD</td>
<td>0.90</td>
<td>5.32</td>
<td>2.72</td>
<td>96.0</td>
<td>4.25</td>
<td>3.53</td>
<td>3.49</td>
</tr>
</tbody>
</table>

In this study, we proposed multi-view attention transfer (MV-AT), which distills features highlighted from three perspectives in speech enhancement tasks that require low complexity for deployment in various environments. Experimental results show that the proposed method achieves significant efficiencies by reducing the model parameters and computational complexity by $15.4 \times$ and $4.71 \times$, respectively, compared to the baseline. An ablation study suggests that transferring diverse knowledge, including multi-view features and output, significantly improves performance. Finally, further experiments on a large DNS dataset suggest that MV-AT is a suitable knowledge distillation method for speech enhancement tasks.

5. Acknowledgements

This research was supported by Brain Korea 21 FOUR. This research was also supported by Korea University Grant (K2202151).
6. References


