



# Ultra Fast Speech Separation Model with Teacher Student Learning

Sanyuan Chen<sup>1</sup>, Yu Wu<sup>2</sup>, Zhuo Chen<sup>3</sup>, Jian Wu<sup>2</sup>, Takuya Yoshioka<sup>3</sup>, Shujie Liu<sup>2</sup>, Jinyu Li<sup>3</sup>, Xiangzhan Yu<sup>1</sup>

<sup>1</sup>Harbin Institute of Technology, China

<sup>2</sup>Microsoft, China

<sup>3</sup>Microsoft, USA

sychen@ir.hit.edu.cn

{yuwul, zhuc, wujian, tayoshio, shujliu, jinyuli}@microsoft.com, yxz@hit.edu.cn

## Abstract

Transformer has been successfully applied to speech separation recently with its strong long-dependency modeling capacity using a self-attention mechanism. However, Transformer tends to have heavy run-time costs due to the deep encoder layers, which hinders its deployment on edge devices. A small Transformer model with fewer encoder layers is preferred for computational efficiency, but it is prone to performance degradation. In this paper, an ultra fast speech separation Transformer model is proposed to achieve both better performance and efficiency with teacher student learning (T-S learning). We introduce layer-wise T-S learning and objective shifting mechanisms to guide the small student model to learn intermediate representations from the large teacher model. Compared with the small Transformer model trained from scratch, the proposed T-S learning method reduces the word error rate (WER) by more than 5% for both multi-channel and single-channel speech separation on LibriCSS dataset. Utilizing more unlabeled speech data, our ultra fast speech separation models achieve more than 10% relative WER reduction.

**Index Terms:** speech separation, Teacher Student Learning, Transformer, deep learning

## 1. Introduction

Speech separation plays a vital role in front-end speech processing, aiming to handle the cocktail party problem. Recently, with the success of Transformer model in speech community [1, 2], the Transformer [3, 4] and its variants [5] have successfully achieved superior performance on this task. However, these models tend to have heavy run-time costs due to the deep encoder layers, while the real-time inference is crucial for product deployment especially on resource limited edge devices.

Given the great demand of better computational efficiency, a small speech separation model is preferred for the deployment, with considerably fewer encoder layers and fast inference speed. Unfortunately, the use of the smaller model directly tends to degrade the separation performance and thus hurts performance of downstream tasks such as multi-speaker speech recognition [5].

To build a small model with both fast inference speed while maintaining the accuracy, teacher student learning (T-S learning) is a common strategy for model training, and has been shown effective in various tasks [6, 7]. With the T-S learning, a smaller Transformer based separation model (student) is trained to mimic the behavior of a large pretrained model (teacher). In this work, we apply the T-S learning to fast transformer based separation network training, and introduce three updates to further enhance the performance. Specifically, with the help of

*Layer-wise T-S learning*, not only the final prediction but also the intermediate feature maps of the teacher model are leveraged. Since the teacher model is not perfect and may generate results with noises and errors, we introduce an *Objective Shifting* mechanism to let the learning objective gradually shift from the teacher predictions to the golden predictions. Going beyond the limitation of the labelled training data, large-scale *unlabeled speech separation data* are used in our T-S learning, to allow the student to better capture teacher’s behaviours. Different from previous work applying T-S learning for speech enhancement and separation, which train student models in the same model size while operating at different input features [8, 9], this paper aims to distill the teacher model’s knowledge to create a smaller and faster student model. Besides, this paper is the first one to use a large amount of unlabeled data in the T-S learning for speech separation.

We conduct the experiment on the public LibriCSS dataset [10]. The experimental results show that our ultra fast Transformer model can achieve more than 5% average relative WER gains with our proposed T-S learning for both single-channel and multi-channel speech separation, and the improvements are more significant for the utterances with higher overlap ratio. Several ablation experiments show that both Layer-wise T-S learning and Objective Shifting mechanisms are crucial to the performance improvements. Moreover, since annotated data are not required for the Layer-wise T-S learning, pretraining on large-scale unlabeled data enables our ultra fast Transformer model achieve more than 10% average relative WER gains with the proposed T-S learning methods.

## 2. Background

### 2.1. Problem Formulation

Continuous speech separation (CSS) aims to estimate individual speaker signals from a continuous speech input where the source signals are fully or partially overlapped. Let  $y(t)$  denote the mixed signal and  $x_s(t)$  the  $s$ -th individual target signal, where  $t$  is the time index. The mixed signal is modeled as follows:

$$y(t) = \sum_{s=1}^S x_s(t). \quad (1)$$

Their short-time Fourier transforms (STFTs) are denoted as  $\mathbf{Y}(t, f)$  and  $\mathbf{X}_s(t, f)$ , respectively.  $f$  denotes frequency index.

Following [11, 12], instead of directly outputting the STFT of the individual signals  $[\mathbf{X}_1(t, f) \dots \mathbf{X}_S(t, f)]$ , we employ the mask learning to recover the clean speech, where a group of masks  $\mathbf{M}(t, f) = [\mathbf{M}_1(t, f) \dots \mathbf{M}_S(t, f)]$  are firstly estimated with a deep learning model  $F(\cdot)$ . Then, for the  $s$ -th individual signal,  $\mathbf{X}_s(t, f)$  is obtained either by mask-based beamforming

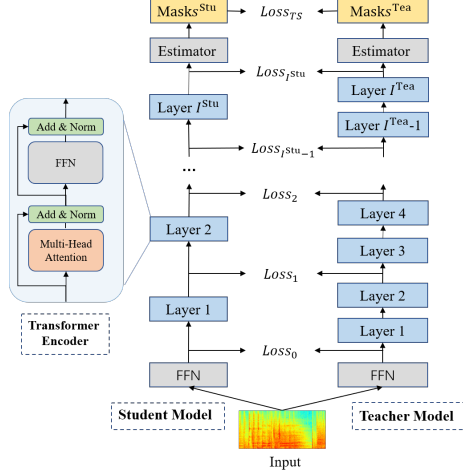


Figure 1: Layer-wise Teacher Student Learning of Transformer model.

or by direct masking, i.e.,  $\mathbf{M}_s(t, f) \odot \mathbf{Y}^1(t, f)$  where  $\odot$  is the element-wise product,  $\mathbf{Y}^1(t, f)$  is the first channel of  $\mathbf{Y}(t, f)$ .

## 2.2. Transformer Model

As shown in the Figure 1, The Transformer model [13] is composed of a stack of identical Transformer encoder layers, each of which consists of a multi-head self-attention module and a position-wise fully connected feed-forward module.

Before sending to Transformer encoder, for both single and multi channel separation network, the input feature  $\mathbf{Y}(t, f)$  is projected to representation  $\mathbf{h}_0$  with fixed dimension, by a feed-forward module  $\text{FFN}(\cdot)$ :

$$\mathbf{h}_0 = \text{FFN}(\mathbf{Y}(t, f)). \quad (2)$$

Given the input,  $\mathbf{h}_{i-1}$ , of the  $i$ -th layer, the output  $\mathbf{h}_i$  is calculated as

$$\mathbf{h}'_i = \text{layernorm}(\mathbf{h}_{i-1} + \text{MultiHeadAttention}(\mathbf{h}_{i-1})) \quad (3)$$

$$\mathbf{h}_i = \text{layernorm}(\mathbf{h}'_i + \text{FFN}(\mathbf{h}'_i)), \quad (4)$$

where  $\text{MultiHeadAttention}(\cdot)$  and  $\text{layernorm}(\cdot)$  denote the multi-head self-attention module and the layer normalization, respectively. The multi-head self-attention module is implemented with relative position embedding as [14, 5, 15].

Given  $\mathbf{h}_J$ , the output of the final layer, we obtain the masks  $\mathbf{M}(t, f)$  with  $\text{Estimator}(\cdot)$ , an estimator consisting of a feed-forward module and a sigmoid activation function, i.e.,

$$\mathbf{M}(t, f) = \text{Estimator}(\mathbf{h}_J) \quad (5)$$

$$= \text{sigmoid}(\text{FFN}(\mathbf{h}_J)). \quad (6)$$

## 2.3. Teacher Student Learning

Teacher student learning is a common training strategy for model compression, where a smaller and faster student model is trained to generate the same output as a more powerful teacher model. Specifically, in separation task, the T-S learning can be represented as the minimization of the mean square error (MSE) between the estimated signals of the student and the teacher model:

$$\mathcal{L}_{\text{TS}} = \frac{1}{T \times F \times S} \sum_{s=1}^S \|\mathbf{X}_s^{\text{Stu}}(t, f) - \mathbf{X}_s^{\text{Tea}}(t, f)\|^2 \quad (7)$$

where  $T$ ,  $F$  and  $S$  denote the number of the time frames, frequency bins and target signals, respectively. The estimated signals  $\mathbf{X}_s^{\text{Stu}}(t, f)$  and  $\mathbf{X}_s^{\text{Tea}}(t, f)$  is calculated as:

$$\mathbf{X}_s^{\text{Stu}}(t, f) = \mathbf{M}_s^{\text{Stu}}(t, f) \odot \mathbf{Y}^1(t, f) \quad (8)$$

$$\mathbf{X}_s^{\text{Tea}}(t, f) = \mathbf{M}_s^{\text{Tea}}(t, f) \odot \mathbf{Y}^1(t, f) \quad (9)$$

where  $\mathbf{M}_s^{\text{Stu}}(t, f)$  and  $\mathbf{M}_s^{\text{Tea}}(t, f)$  are the estimated masks of the student model and the teacher model.

## 3. Method

To further enhance the efficacy of knowledge distillation, two mechanisms are introduced to baseline T-S learning, namely **Layer-wise T-S Learning** and **Objective Shifting**, that allows the student model to also benefit from Teacher's intermediate representation and oracle training label. In addition, to further boost the performance of the student model, we leverage the **unlabeled data training** in our T-S learning framework.

### 3.1. Layer-wise T-S Learning

We introduce the layer-wise T-S Learning mechanism to train the student to reproduce not only the final prediction but also the intermediate outputs of the teacher model [16].

As Figure 1 shows, given the  $I^{\text{Stu}}$  layer student model and  $I^{\text{Tea}}$  layer teacher model, we minimize the mean square error (MSE) between the output of  $i$ -th layer of the student model and the corresponding  $g(i)$ -th output of the teacher model:

$$\mathcal{L}_i = \frac{1}{T \times F} \|\mathbf{h}_i^{\text{Stu}} - \mathbf{h}_{g(i)}^{\text{Tea}}\|^2 \quad (10)$$

where  $g(\cdot)$  is a uniform layer mapping function between indices from student layers to teacher layers.

Then the objective function of layer-wise T-S Learning is the weighted average function as:

$$\mathcal{L}_{\text{LTS}} = \frac{\sum_{i=0}^{I^{\text{Stu}}} (i+1) \cdot \mathcal{L}_i + (I^{\text{Stu}} + 1) \cdot \mathcal{L}_{\text{TS}}}{\sum_{i=0}^{I^{\text{Stu}}} (i+1) + (I^{\text{Stu}} + 1)} \quad (11)$$

where  $\frac{i+1}{\sum_{i=0}^{I^{\text{Stu}}} (i+1) + (I^{\text{Stu}} + 1)}$  is the weight for  $\mathcal{L}_i$ . The loss of a higher layer is assigned with a larger weight as [15].

### 3.2. Objective Shifting

Since the student model is trained to recover the predictions of the teacher model, the performance of T-S learning would be limited to the teacher's capability. To avoid this limitation, we introduce the Objective Shifting mechanism to train the student with both the teacher's prediction and training datasets [17, 18].

Specifically, an additional loss item  $\mathcal{L}_{\text{PIT}}$  in added to training objective, that minimizes the MSE between the estimated signals of the student model and the references in the training sets. The final loss function of layer-wise T-S learning with objective shifting is calculated as:

$$\mathcal{L} = \lambda(t) \mathcal{L}_{\text{PIT}} + (1 - \lambda(t)) \mathcal{L}_{\text{LTS}} \quad (12)$$

where  $t$  refers to the training timesteps,  $\lambda(t) = \text{sigmoid}(-k \cdot (t - t_0))$  is set to the sigmoid annealing function.

It should be noted in  $\mathcal{L}_{\text{PIT}}$ , we apply permutation invariant training (PIT) [19, 20] to remedy the source permutation problem, while the permutation in T-S  $\mathcal{L}_{\text{LTS}}$  loss is determined by the teacher model.

With objective shifting, at the beginning of the training process, the student model is solely guided with the teacher’s predictions, as soft label is believed to provide richer indication of teacher’s behavior, thus leading to more efficient starting. As training continues, the student gradually reduces the loss weight from the teacher, with more emphasis on clean reference, until the end of the training process, where the student completely learns from the clean target, to escape the limitation of the teacher’s knowledge.

### 3.3. Unlabeled Data training

Training data for speech separation is generally artificially synthesized, so it requires clean speech as well as various noises. However, real overlapped data is slightly different from the artificially synthesized data, and it is hard to obtain the ground-truth of the unmixing results. The gap between artificial training data and the test data in the real scenario is a potential issue for speech separation. In this paper, we aim to leverage large-scale unlabeled mixing data in T-S learning. In this way, the student model can approach the teacher model by mimicking the teacher’s behaviours, not only on the limited annotated data but also the large-scale unlabeled data.

Specifically, the student model is trained with T-S learning for two stages. In the first stage, we pretrain the student model with the layer-wise T-S learning mechanism (Eq. 11) on the large-scale unlabeled mixing data. The student model learns to reproduce the final prediction and intermediate outputs of the teacher’s model on the real overlapped data. In the second stage, we train the student model with the layer-wise T-S learning and objective shifting (Eq. 12) on the annotated training data. The student model begins with mimicking the teacher’s behaviours on the annotated data, and ends with learning from the golden predictions of the annotated training data.

## 4. Experiment

### 4.1. Datasets

In this work, except the unlabeled learning part, all models are trained with 219 hours of artificially reverberated and mixed speech signals sampled randomly from WSJ1 [21]. Following [22], we include four different mixture types in the training data. Each training mixture is generated by randomly picking one or two speakers from the WSJ1 dataset and convolving each with a 7 channel room impulse response (RIR) simulated with the image method [23]. Then, we rescale and combine them with a source energy ratio between -5 and 5 dB. Simulated isotropic noise [24] is also added at a 0–10 dB signal to noise ratio. The average overlap ratio of the training set is around 50%. For the unlabeled data training, we apply the LibriVox and a Microsoft in-house dataset. LibriVox contains over 60k hours of audio derived from open-source audio books [25]. The Microsoft in-house dataset contains 564 hours recording of discussion from Microsoft employees. We create 2k hours and 600 hours speech mixtures for LibriVox and the in-house dataset, by simply mixing two single speaker utterances. As in-house recording contains a noticeable amount of noise, there is no clean reference for mixtures derived from this dataset. We evaluate the models on the LibriCSS dataset [10], which consists of 10 hours of concatenated and mixed LibriSpeech utterances played and recorded in a meeting room. We test our model performance for both single channel and seven-channel setting, with word error rate (WER) as evaluation metric. We conducted both the utterance-wise evaluation and continuous input evaluation (re-

fer to [10] for the two evaluation schemes).

### 4.2. Implementation Details

The teacher model is the Conformer model from [5] which contains 16 encoder layers, 256 attention dimensions and 2048 FFN dimensions, resulting in 26.49M and 26.09M parameters for multi-channel and single-channel evaluation respectively. For multi-channel evaluation, the student Transformer model with 3.89M parameters consists of 6 encoder layers with 2 attention heads, 128 attention dimensions and 2048 FFN dimensions. The layer mapping function of the Layer-wise T-S learning is defined as  $g(i) = \max(3 \times i - 2, 0)$ . For single-channel evaluation, the student Transformer model with 7.25M parameters consists of 12 encoder layers with 4 attention heads, 128 attention dimensions and 2048 FFN dimensions. The layer mapping function is defined as  $g(i) = \min(2 \times i, i + 4)$ . The models are trained with the AdamW optimizer [26] where the weight decay is set to 1e-2, the learning rate is 1e-4. We use the warm-up learning schedule with linear decay where the warm-up step is 10k, and the training step is 260k. For Object Shifting, we set  $t_0$  to 150k, and select the best  $k$  in  $\{1e-4, 5e-4\}$ . For the unlabeled data training, we select the best  $t_0$  in  $\{10k, 20k\}$ . The small Transformer trained from scratch, denoted as Transformer-small<sub>Baseline</sub> is used as baseline system. The vanilla T-S learning, T-S learning with objective shifting, and layer-wise T-S learning are denoted as Transformer-small<sub>vanilla TS</sub>, Transformer-small<sub>OS</sub>, and Transformer-small<sub>LTS</sub> respectively.

We evaluate the speech separation accuracy with two ASR models. One is a hybrid system with a BLSTM based acoustic model and a 4-gram language model as used in the LibriCSS paper [10]. The other is one of the best open source end-to-end Transformer [27] based ASR models<sup>1</sup> which achieves WERs of 2.08% and 4.95% for LibriSpeech test-clean and test-other, respectively. We follow the sliding window-based CSS processing in continuous speech separation [10] where the window size is set to 2.4s. As with [10], we generate the individual target signals with spectral masking and mask-based adaptive minimum variance distortionless response (MVDR) beamforming for the single-channel and seven-channel cases, respectively.

### 4.3. Evaluation Results

The result for utterance-wise and continuous separation are shown in Table 1 and 2. We analyze the experiment results from three aspects: comparison with the teacher model, baseline small Transformer model, and models with unlabeled data training.

**A comparison with the teacher model.** Compared to the Conformer teacher model, the Transformer-small model with much less parameters can achieve an ultra faster speech separation speed. We can obtain  $21.5\times$  and  $11.4\times$  speed-up for seven-channel and single-channel continuous speech separation with 2.4s window size. Even if the runtime cost is largely reduced, we observe performance degradation in all experiments, but the seven channel degradation is not as serious as the single channel. We guess the MVDR component bridges the gap between different models. To prove our hypothesis, we remove the MVDR in seven channel, and observe the gap between teacher and student becomes larger as shown in Table 2.

**A comparison with training from scratch.** We can achieve significant improvements with the proposed T-S learn-

<sup>1</sup><https://github.com/MarkWuNLP/SemanticMask>

Table 1: *Utterance-wise evaluation for seven-channel and single-channel settings. Two numbers in a cell denote %WER of the hybrid ASR model used in LibriCSS [10] and E2E Transformer based ASR model [27]. OS and OL are utterances with short/long inter-utterance silence.*

System	Overlap ratio in %					Avg gains	
	OS	OL	10	20	30		40
Seven-channel Evaluation							
Conformer (Teacher)	7.0/3.1	7.2/3.2	8.9/3.6	11.1/4.6	13.6/5.8	15.1/6.3	13.2%/15.4%
Transformer-small <sub>Baseline</sub>	8.1/3.4	8.5/3.4	10.6/4.3	12.4/5.3	15.2/6.6	17.8/8.0	0.0%/0.0%
Transformer-small <sub>vanilla TS</sub>	7.6/3.3	7.9/3.4	10.0/3.9	12.3/5.2	15.0/6.8	17.2/7.4	3.3%/3.8%
Transformer-small <sub>LTS</sub>	7.3/3.3	7.7/3.2	9.6/4.0	12.2/5.1	14.8/6.8	17.2/7.6	5.0%/3.8%
Transformer-small <sub>OS</sub>	7.4/3.3	7.7/3.3	10.1/4.0	12.2/5.1	14.8/6.6	17.0/7.4	5.0%/3.8%
Transformer-small <sub>LTS + OS</sub>	7.7/3.4	8.0/3.3	10.0/3.9	11.9/5.0	14.6/6.4	16.3/7.4	5.8%/5.8%
Transformer-small <sub>unlabeled LTS + OS</sub>	7.2/3.2	7.4/3.3	9.2/3.8	11.5/4.9	14.2/6.2	16.0/6.9	9.9%/9.6%
Single-channel Evaluation							
Conformer (Teacher)	10.2/4.2	10.0/4.4	13.3/6.6	17.9/10.0	22.1/13.2	26.7/15.6	23.4%/28.6%
Transformer-small <sub>Baseline</sub>	12.5/4.8	12.0/4.4	17.0/8.5	23.8/13.7	30.0/19.3	35.5/25.1	0.0%/0.0%
Transformer-small <sub>vanilla TS</sub>	12.0/4.2	11.9/4.0	16.9/8.7	23.4/13.7	29.6/19.8	35.6/25.9	0.9%/−0.8%
Transformer-small <sub>LTS</sub>	12.0/4.1	11.8/4.1	16.7/8.6	22.6/13.8	29.0/19.7	35.0/25.3	2.8%/0.0%
Transformer-small <sub>OS</sub>	12.6/4.7	12.2/4.3	16.9/8.5	23.4/13.4	29.5/19.4	35.1/24.9	0.9%/0.8%
Transformer-small <sub>LTS + OS</sub>	12.2/4.4	12.0/4.5	16.1/8.3	22.0/13.1	27.7/18.1	33.4/23.1	5.5%/5.6%
Transformer-small <sub>unlabeled LTS + OS</sub>	11.0/4.3	10.8/4.5	15.0/7.6	20.6/11.6	25.6/16.3	31.1/20.2	12.8%/14.3%

Table 2: *Continuous speech separation evaluation for seven-channel and single-channel settings.*

System	Overlap ratio in %					Avg gains	
	OS	OL	10	20	30		40
Seven-channel Evaluation							
Conformer (Teacher)	11.8/5.7	9.0/4.1	13.2/6.3	14.1/7.1	18.6/9.8	20.3/10.8	9.9%/18.9%
Transformer-small <sub>Baseline</sub>	12.7/6.6	10.1/5.5	15.1/8.1	15.7/9.0	21.0/12.3	22.2/12.6	0.0%/0.0%
Transformer-small <sub>LTS + OS</sub>	12.3/6.6	9.6/5.1	14.6/7.2	15.5/8.7	20.1/11.6	22.7/12.9	1.9%/3.3%
Transformer-small <sub>unlabeled LTS + OS</sub>	12.2/6.1	9.2/4.6	14.1/7.2	14.7/7.8	20.1/11.1	21.0/12.3	5.6%/8.9%
Seven-channel Evaluation (w/o MVDR)							
Conformer (Teacher)	13.9/6.3	11.7/5.1	15.2/8.1	19.1/10.3	24.0/14.5	27.5/16.4	21.5%/30.3%
Transformer-small <sub>Baseline</sub>	18.0/9.4	15.3/8.5	20.5/11.6	24.4/14.9	30.0/19.1	33.9/23.2	0.0%/0.0%
Transformer-small <sub>LTS + OS</sub>	16.4/8.9	14.0/8.1	18.2/10.4	22.5/13.7	27.4/18.1	32.0/21.8	8.0%/6.9%
Transformer-small <sub>unlabeled LTS + OS</sub>	14.0/7.5	12.2/6.5	16.0/9.5	19.6/12.1	24.9/16.6	29.0/19.7	18.6%/17.2%
Single-channel Evaluation							
Conformer (Teacher)	16.4/9.6	15.0/9.0	19.3/12.1	24.3/15.6	29.1/20.5	32.4/23.5	36.1%/49.0%
Transformer-small <sub>Baseline</sub>	30.7/23.0	28.5/25.3	31.3/25.2	37.0/29.6	41.3/34.4	45.4/40.1	0.0%/0.0%
Transformer-small <sub>LTS + OS</sub>	28.5/23.2	25.5/22.4	28.9/23.5	34.5/28.4	38.0/32.6	42.5/36.6	7.6%/6.1%
Transformer-small <sub>unlabeled LTS + OS</sub>	22.9/17.8	21.0/20.0	24.0/19.0	28.8/22.1	33.1/26.6	37.2/29.5	22.1%/24.0%

ing method, compared to training from scratch, especially on the highly overlapped cases. For the seven-channel settings, we can obtain 5.8% average relative WER gains with both the hybrid and E2E ASR systems for the utterance-wise evaluation. If we remove either the Layer-wise T-S Learning or Objective Shifting mechanism, performance drops are witnessed. It shows that the student model can benefit from the intermediate knowledge from the teacher model and more knowledge from the training datasets.

For the single-channel settings, due to the limited input information, we experiment with the deeper Transformer-small model with more parameters. Similar to the seven-channel cases, our T-S learning method can consistently outperform the baseline by a large margin, and achieve over 5% relative WER gains for utterance-wise evaluation and over 6% relative WER gains for continuous evaluation on average.

**Leveraging more unlabeled data.** By leveraging more unlabeled data, we can further boost the performance improvements of our proposed T-S learning methods. For the single-channel settings, with the student model pretrained on the large-scale unlabeled data and shifted learning objective on the annotated training data, we can obtain 14.3% and 24.0% average

relative WER gains for utterance-wise evaluation and continuous evaluation with E2E ASR systems. For the seven-channel evaluation, utilizing more unlabeled data, we can obtain 9.6% and 8.9% average relative WER gains for utterance-wise evaluation and continuous evaluation with E2E ASR systems. If we remove MVDR in seven-channel settings, our T-S learning methods can bring more significant improvements and 17.2% average relative WER gains can be witnessed.

## 5. Conclusions

Because of the ultra fast inference speed, the small speech separation Transformer model is preferred for the deployment on devices. In this work, we elaborate Teacher Student learning for better training of the ultra fast speech separation model. The small student model is trained to reproduce the separation results of a large pretrained teacher model. We also introduce Layer-wise Teacher Student Learning and Objective Shifting mechanisms to benefit the Teacher Student learning with more transferred knowledge. The experimental results show the proposed methods can successfully improve the separation results of the small Transformer model. Furthermore, pretraining on unlabeled data can further enhance the improvement.

## 6. References

- [1] L. Dong, S. Xu, and B. Xu, “Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.
- [2] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, and S. Liu, “On the comparison of popular end-to-end models for large scale speech recognition,” in *Proc. Interspeech*, 2020.
- [3] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, “End-to-end multi-speaker speech recognition with transformer,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6134–6138.
- [4] J. Chen, Q. Mao, and D. Liu, “Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation,” *Proc. Interspeech 2020*, pp. 2642–2646, 2020.
- [5] S. Chen, Y. Wu, Z. Chen, J. Li, C. Wang, S. Liu, and M. Zhou, “Continuous speech separation with conformer,” *arXiv preprint arXiv:2008.05773*, 2020.
- [6] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, “Learning small-size DNN with output-distribution-based criteria,” in *Proc. Interspeech*, 2014, pp. 1910–1914.
- [7] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [8] S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, “Student-teacher network learning with enhanced features,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5275–5279.
- [9] A. S. Subramanian, S.-J. Chen, and S. Watanabe, “Student-teacher learning for blstm mask-based speech enhancement,” *Proc. Interspeech 2018*, pp. 3249–3253, 2018.
- [10] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, “Continuous speech separation: Dataset and analysis,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7284–7288.
- [11] 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, 2014.
- [12] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio,” in *New Era for Robust Speech Recognition*. Springer, 2017, pp. 165–186.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [14] P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-attention with relative position representations,” in *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2018, pp. 464–468.
- [15] S. Chen, Y. Wu, Z. Chen, T. Yoshioka, S. Liu, and J. Li, “Don’t shoot butterfly with rifles: Multi-channel continuous speech separation with early exit transformer,” *arXiv preprint arXiv:2010.12180*, 2020.
- [16] S. Sun, Y. Cheng, Z. Gan, and J. Liu, “Patient knowledge distillation for bert model compression,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4314–4323.
- [17] E. Kiperwasser and M. Ballesteros, “Scheduled multi-task learning: From syntax to translation,” *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 225–240, 2018.
- [18] S. Chen, Y. Hou, Y. Cui, W. Che, T. Liu, and X. Yu, “Recall and learn: Fine-tuning deep pretrained language models with less forgetting,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7870–7881.
- [19] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.
- [20] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.
- [21] L. D. C. Philadelphia, “CSR-II (WSJ1) Complete,” 1994, <http://catalog.ldc.upenn.edu/LDC94S13A>.
- [22] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, “Multi-microphone neural speech separation for far-field multi-talker speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5739–5743.
- [23] E. A. Habets, “Room impulse response generator,” *Technische Universiteit Eindhoven, Tech. Rep.*, vol. 2, no. 2.4, p. 1, 2006.
- [24] E. A. Habets and S. Gannot, “Generating sensor signals in isotropic noise fields,” *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.
- [25] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, “Libri-light: A benchmark for asr with limited or no supervision,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673, <https://github.com/facebookresearch/libri-light>.
- [26] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2018.
- [27] C. Wang, Y. Wu, Y. Du, J. Li, S. Liu, L. Lu, S. Ren, G. Ye, S. Zhao, and M. Zhou, “Semantic mask for transformer based end-to-end speech recognition,” *Proc. Interspeech 2020*, pp. 971–975, 2020.