INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

OTF: Optimal Transport based Fusion of Supervised and Self-Supervised Learning Models for Automatic Speech Recognition

Li Fu, Siqi Li, Qingtao Li, Fangzhu Li, Liping Deng, Lu Fan, Meng Chen, Youzheng Wu, Xiaodong He

JD AI Research, Beijing, China

{fuli3,lisiqi26,liqingtao8,lifangzhu1,dengliping6,fanlu}@jd.com
{chenmeng20,wuyouzheng1,hexiaodong}@jd.com

Abstract

Self-Supervised Learning (SSL) Automatic Speech Recognition (ASR) models have shown great promise over Supervised Learning (SL) ones in low-resource settings. However, the advantages of SSL are gradually weakened when the amount of labeled data increases in many industrial applications. To further improve the ASR performance when abundant labels are available, we first explore the potential of combining SL and SSL ASR models via analyzing their complementarity in recognition accuracy and optimization property. Then, we propose a novel Optimal Transport based Fusion (OTF) method for SL and SSL models without incurring extra computation cost in inference. Specifically, optimal transport is adopted to softly align the layer-wise weights to unify the two different networks into a single one. Experimental results on the public 1k-hour English LibriSpeech dataset and our in-house 2.6k-hour Chinese dataset show that OTF largely outperforms the individual models with lower error rates.

Index Terms: Automatic speech recognition, model fusion, optimal transport, self-supervised learning

1. Introduction

Recently, Self-Supervised Learning (SSL) has emerged as a successful paradigm to address the issue of label scarcity in low-resource Automatic Speech Recognition (ASR) tasks, e.g. multiple languages [1–3] and domain shift [4, 5]. It usually pre-trains a representation model on numerous unlabeled utterances, and then fine-tunes the model with a relatively small amount of labeled speech [6–12]. Nevertheless, the gains achieved by the pre-training might diminish when the amount of downstream labeled dataset increases [13–15]. Thus, there would be a dilemma in many resource-rich industry applications – Shall we train a Supervised Learning (SL) ASR model from scratch or fine-tune the pre-trained representation model?

In general, SL models are optimized to perform well when large amounts of labeled and in-domain speech are available [16–18]; and SSL models are considered to have good generalization via pre-training on numerous unlabeled utterances [19–21]. While the performance gaps in Word/Character Error Rate (WER/CER) between the SL and SSL ASR models become small in resource-rich settings, we believe that empirically, the two models would contain diverse or even complementary abilities due to the training process being quite different. Based on this idea, as analyzed in Sec. 3.1, we first verify the potential of fusing SL and SSL models according to 1) Recognition accuracy: We roughly estimate the *upper bound* of model fusion via picking out the best hypothesis (HYP) of different models for each test sample, which implies a large improvement room on fusing these models; and 2) Optimization property: We study the optimization process of the two models via loss landscape [22], which qualitatively indicates the SSL and SL models' advantages in generalization and in-domain task learning, respectively.

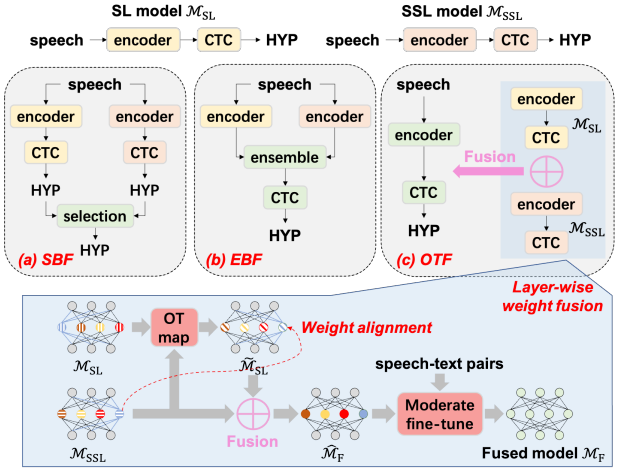


Figure 1: *Overview of the proposed (c) OTF over the existing methods (a) SBF and (b) EBF: each layer's weights of SL model $\mathcal{M}_{SL}$ are aligned with SSL model $\mathcal{M}_{SSL}$ via Optimal Transport (OT) maps to yield aligned model $\widetilde{\mathcal{M}}_{SL}$, which is added with $\mathcal{M}_{SSL}$ and then moderately fine-tuned on labeled speech to obtain the fused model $\mathcal{M}_F$ with the same architecture.*

Existing model fusion methods mainly focus on how to aggregate the constituent models' output hypotheses or latent features. For example, Selection-Based Fusion (SBF) method (see Fig. 1(a)) was proposed to select the best hypothesis from different models via the designed criteria, e.g. confidence score [23]. Ensemble-Based Fusion (EBF) method (see Fig. 1(b)) was explored to combine the latent features of multiple encoders pre-trained in different frameworks [24–27]. However, both the SBF and EBF methods would inevitably suffer high computational cost at test time since each utterance is processed through all of these encoders. A naive solution might be aggregating all the models into a single one by directly averaging the model parameters. However, since the weight parameters of SL and SSL models are not one-to-one corresponded, such direct averaging is ineffective and may even damage the well-trained models. To circumvent this issue, matching weight before averaging was investigated in the computer vision domain [28, 29]. For example, Yurochkin et al. [28] proposed the Bayesian nonparametric matching method to align and average the image classification models on different edge devices for federated learning. Singh

10.21437/Interspeech.2023-1609

et al. [29] explored the fusion of image classification models by leveraging optimal transport to align each layer's weight. However, discussion about the weight-based fusion for ASR models is quite rare, which might be more challenging than the image classification tasks [28, 29] since speech signals are sequential values [7].

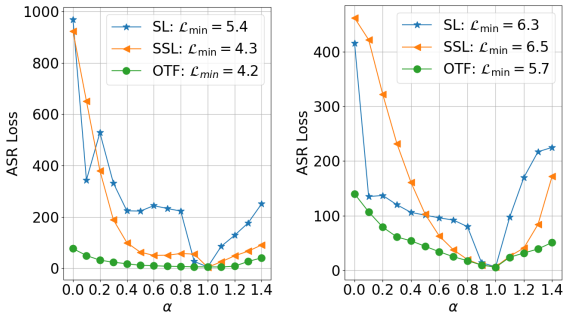
To combine the SL and SSL models without increasing the inference cost, we propose an Optimal Transport based Fusion (OTF) method, which fuses the two ASR models into the same architecture and improves the performance in the following way (see Fig. 1(c)). First, inspired by the work of [29], we adopt optimal transport to softly align each layer's weights of the SL and SSL models. Specifically, a layer-wise transport map is estimated via minimizing the cost of transferring the distribution of the SL model's weight to the SSL model. Then, the input and output parameters of the SL model's weight are aligned by multiplying with the transport maps of preceding and current layers, respectively. Finally, to enhance the ASR performance, the aligned SL model is averaged with the SSL model and moderately fine-tuned on labeled data to obtain the fused model. Extensive experiments on different datasets show that OTF effectively fuses the SL and SSL models with obvious WER/CER reductions. Our main contributions are summarized as follows:

- To the best of our knowledge, this is the first work exploring the fusion of SL and SSL models for speech recognition.
- We propose a novel approach, named OTF, which unifies SL and SSL ASR models efficiently without incurring an extra computational cost in inference.
- We verify the effectiveness of OTF, with discussion, on English and Chinese datasets with large WER/CER reductions compared with the individual models and baseline methods.

2. Related Work

Model fusion. In recent years, it has been shown favorable to fuse different models to enhance the ASR performance. Although the SBF method was explored to pick the better hypothesis of two ASR models without re-training, it would result in a suboptimal solution since the criteria used for hypothesis selection might not be accurate for each utterance [23, 30]. More recently, the EBF method was proposed to tightly couple two pre-trained models. Arunkumar et al. [24] investigated an ensemble model to combine the outputs of the last layer of HuBERT [7] and wav2vec2 [6] fine-tuned ASR models. Tang et al. [25] explored the combination of the multi-layer latent features of wav2vec2 [6] and data2vec [31]. Wu et al. [26] proposed an ensemble framework, with a combination of ensemble techniques to fuse SSL speech models' embedding. However, both the SBF and EBF methods would suffer high computational cost at test time with aggregating all the models' parameters. In contrast, our proposed OTF fuses the two ASR models into the same architecture via weight alignment, which will not incur an extra cost during inference.

Multi-task learning. Another way to incorporate auxiliary abilities to the ASR model is multi-task learning. For example, distillation tasks [32] were added with ASR tasks to transfer the teacher model's (e.g. SSL model's) knowledge to the student model (e.g. SL model) [33–35]. The existing methods usually assume the teacher model is superior to the student model, while the performance gaps of the SL and SSL models might be small in resource-rich settings. The joint of SL and SSL training losses for ASR tasks were studied in [36–38]. However, as mentioned in [38], it might be difficult to balance the SSL and SL components systematically. Differently, our work focuses on weight-based fusion, which is more efficient to inherit the abilities of the individual models. Moreover, the multi-task learning method might also complement the moderate fine-tuning in the proposed OTF, which will be investigated in our future work.



(a) English ASR models (b) Chinese ASR models

Figure 2: *Loss landscapes of the linear interpolation along the three kind of models ($\alpha = 1$): 1) SL models, 2) SSL models, 3) fused models via our proposed OTF, with their initial models ($\alpha = 0$), respectively, where α is the interpolation coefficient.*

3. Our Proposed Approach

3.1. Advantages for fusing SL and SSL ASR models[1]

Complementarity in recognition accuracy. Although the gaps in WER/CER of the SL and SSL ASR models are small in resource-rich settings, we find that the recognition errors of the two models might be different for a certain utterance. The potential gain by the fusion of SL and SSL models can be estimated via picking the best hypothesis among the individual models for each test utterance to calculate the error rate performance. It can be regarded as an approximated upper bound of model fusion, as shown in Table 1-2. Numerically, compared with the best among SL and SSL models, the relative WER/CER reduction of the upper bound can lead up to 21% for the English and Chinese models. This implies a large potential for improvement for the model fusion as we expected.

Complementarity in optimization property. To find out the advantages of SL and SSL models with abundant speech labels, we analyze the optimization properties of the two models via the visualization of loss landscape [22]. Specifically, given the weights of an SL/SSL ASR model θ and its initialized weights θ_0 before fine-tuning, we plot the loss landscape $\mathcal{L}(\theta(\alpha))$ of the validation dataset, with α the coefficient for the weights' linear interpolation $\theta(\alpha) = (1-\alpha)\theta_0 + \alpha\theta$; and with $\mathcal{L}$ the ASR loss function used for fine-tuning. As shown in Fig. 2, the curves of the loss landscape show that each of SL and SSL has its advantage in the optimization process: 1) The flat minima points of SSL models present a better generalization than the SL models' sharp minima points [39]; and 2) The Chinese SL model achieves a better minima (and lower CERs on Chinese Cantonese test sets in Table 2) than the SSL model when large amounts of in-domain labeled speech is available. The results qualitatively indicate the SSL and SL models' complementarity in generalization and in-domain task learning, respectively.

[1]Details about the models are shown in Sec. 4.1.

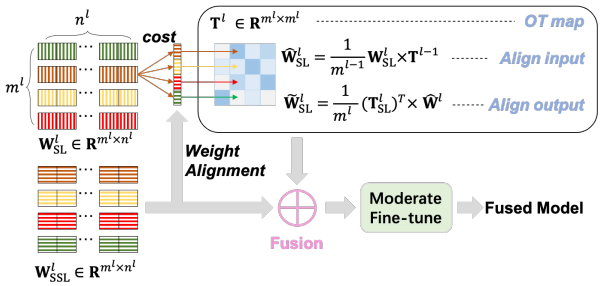


Figure 3: *Illustration of the layer-wise weight alignment and fusion via optimal transport. The l-th layer's weight matrix $\mathbf{W}^l_{\mathrm{SL}}$ of the SL model is aligned with weight matrix $\mathbf{W}^l_{\mathrm{SSL}}$ of the SSL model with the transport maps.*

3.2. Model fusion by optimal transport

Key intuition: Our work aims to fuse the SL and SSL ASR models into a single model with the same architecture while taking advantage of both models. To achieve this, we propose OTF, which matches similar weights in each layer for model aggregation (see Fig 3). Instead of searching over the space of permutation matrices, transport mapping matrices are calculated to softly align one model's weight to another model according to the cost (e.g. Euclidean distance [40]) across the two weights' row vectors [29]. Then each layer's input and output are aligned by the maps of the preceding and current layers, respectively.

Given an SL model $\mathcal{M}_{\mathrm{SL}}$ and an SSL model $\mathcal{M}_{\mathrm{SSL}}$ that are composed of L layers of weight $\mathbf{W}^l_{\mathrm{SL}} \in \mathbf{R}^{m^l \times n^l}$ and $\mathbf{W}^l_{\mathrm{SSL}} \in \mathbf{R}^{m^l \times n^l}$ with $l \in [1, 2, \cdots, L]$, n^l and m^l the dimensions[2] of input and output, respectively. Note that the input dimension of the current layer is equal to the output dimension of the preceding layer. Then the transport map of the l-th layer $\mathbf{T}^l \in \mathbf{R}^{m^l \times m^l}$ is estimated by minimizing the cost $\mathcal{L}_{\mathrm{t}}$ in transferring the weight distribution of $\mathbf{W}^l_{\mathrm{SL}}$ to $\mathbf{W}^l_{\mathrm{SSL}}$ [29],

$$\mathcal{L}_{\mathrm{t}} = \min_{\mathbf{T}^l} \left\langle \mathbf{T}^l, \mathbf{D}^l \right\rangle_F \tag{1}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product, $\mathbf{D}^l \in \mathbf{R}^{m^l \times m^l}$ is the Euclidean distance matrix of the rows in $\mathbf{W}^l_{\mathrm{SL}}$ to $\mathbf{W}^l_{\mathrm{SSL}}$. $\mathbf{T}^l \mathbf{1}_{m^l}$ and $(\mathbf{T}^l)^T \mathbf{1}_{m^l}$ are constrained to be equal to the m^l-dimensional uniform distributions. The optimization is to find a better way (i.e. $\mathbf{T}^l$) to move the l^{th} layer of the SL model's weight to the SSL model's by minimizing the defined cost – If a row of the SL model's weight matrix is similar to a row of the SSL model's weight, the transportation cost is defined as small. The Frobenius inner product is the sum of the cost to move the SL model's weight to the SSL model's. Since our method fuses the two models' weight directly, we do not need any data to perform the optimization. Note that the data would be alternatively used to align the output features of each layer [29].

Denoting the map of the preceding layer as $\mathbf{T}^{l-1}$, the input weights are aligned to obtain the intermediate matrix as [29]:

$$\widehat{\mathbf{W}}^l_{\mathrm{SL}} = \frac{1}{m^{l-1}} \mathbf{W}^l_{\mathrm{SL}} \times \mathbf{T}^{l-1} \tag{2}$$

Then, the output weights are aligned via pre-multiplying $\widehat{\mathbf{W}}^l_{\mathrm{SL}}$ by the transposition of $\mathbf{T}^l$, yields

$$\widetilde{\mathbf{W}}^l_{\mathrm{SL}} = \frac{1}{m^l} (\mathbf{T}^l)^T \times \widehat{\mathbf{W}}^l_{\mathrm{SL}} \tag{3}$$

Finally, we calculate the fused weight of the l-th layer by $\widehat{\mathbf{W}}^l_{\mathrm{F}} = \frac{1}{2}(\mathbf{W}^l_{\mathrm{SLL}} + \widetilde{\mathbf{W}}^l_{\mathrm{SL}})$, which is fine-tuned on the labeled dataset to improve the recognition performance. Since the fused model $\mathcal{M}_{\mathrm{F}}$ is averaged with the weight alignment, it can still partly recognize the speech and only moderate/limited fine-tuning steps are needed in this stage. The effectiveness of weight alignment and fine-tuning are studied in Sec. 4.3.

[2]Following [29], the weights of convolution and bias layers are reshaped to the 2-dimensional regular weights.

Table 1: *WER performance of English systems and relative WER improvement of OTF over SSL model (in brackets) (%).*

Methods	model size	training data (h)		test-clean		test-other	
		unlabeled	labeled	offline	online	offline	online
SL [41]	0.1B	NA	1k	3.6	4.7	9.7	12.4
SSL [42]	0.1B	1k	1k	3.3	4.2	7.7	10.3
Upper-bound	0.2B	NA	NA	**2.6**	**3.4**	**6.7**	**9.0**
SBF [23]	0.2B	NA	NA	3.3	4.1	7.9	10.3
EBF [24]	0.2B	NA	1k	3.3	4.3	7.7	10.2
OTF	0.1B	NA	1k	**3.0(+9.1)**	**4.0(+4.8)**	**7.4(+3.9)**	**10.1(+1.9)**

Table 2: *CER performance of Chinese systems and relative CER improvement of OTF over SSL model (in brackets) (%).*

Methods	model size	training data (h)		test-Mandarin		test-Cantonese	
		unlabeled	labeled	offline	online	offline	online
SL [41]	0.1B	NA	2.6k	12.8	15.7	12.6	13.8
SSL [42]	0.1B	150k	2.6k	12.2	14.5	12.9	14.0
Upper-bound	0.2B	NA	NA	**10.1**	**11.9**	**10.6**	**12.1**
SBF [23]	0.2B	NA	NA	11.6	13.5	12.1	13.8
EBF [24]	0.2B	NA	2.6k	11.6	13.5	12.2	13.8
OTF	0.1B	NA	2.6k	**11.1(+9.0)**	**12.7(+12.4)**	**11.8(+8.5)**	**13.1(+6.4)**

4. Experiments and Discussion

4.1. Experimental setup

Data preparation. Two datasets in different languages are used in our experiments: 1) English data: the public LibriSpeech [43] which contains 1k-hour speech-text pairs; and 2) Chinese data: our in-house 150k-hour Mandarin-Cantonese dataset and only 2.6k-hour data are labeled, with a ratio of Mandarin to Cantonese 10:1. For LibriSpeech, the WER performance of the ASR model is evaluated on the original test-clean (considered easier) and test-other (considered harder and noisier) datasets. For the other one, 5% of the labeled Mandarin and Cantonese samples are randomly selected as the test sets, named test-Mandarin and test-Cantonese. The 80-dimensional Mel-spectrograms of each utterance are pre-processed as the input to the model with 25ms window size and 10ms step size.

Models. Considering the preference of unified offline and online ASR in industrial applications [41,42,44], we use the State-Of-The-Art (SOTA) SSL-based unified framework – UFO2 [42] to test the performance of OTF in both offline and online modes. The model consists of 2 convolutional sub-sampling layers and 12 Conformer blocks with dimension 512 for the encoder. The number of model parameters is 0.1 billion. More details about the model architecture can be referred to [42]. Here, we focus on the fusion of SL and SSL models in two scenarios as below:

1) English ASR models

- $\mathcal{M}_{\mathrm{SL}}$: trained on the 1k-hour public LibriSpeech dataset.
- $\mathcal{M}_{\mathrm{SSL}}$: pre-trained and fine-tuned on the 1k-hour LibriSpeech dataset successively.

2) Chinese ASR models

- $\mathcal{M}_{\mathrm{SL}}$: trained on the 2.6k-hour labeled Chinese dataset.
- $\mathcal{M}_{\mathrm{SSL}}$: pre-trained on the 150k-hour unlabeled Chinese dataset and fine-tuned on the 2.6k-hour dataset successively.

Both the pre-training and fine-tuning are optimized with a mini-batch of 96 and using the same learning rate scheduler as [42]. Note that the SL and SSL models are fully fine-tuned with 200 epochs and 80 epochs for the English and Chinese tasks, respectively. The epoch number in the moderate fine-tuning of the proposed OTF is set to 10 for the two languages.

Table 3: *Ablation studies on weight alignment and Moderate Fine-Tuning (MFT) for English models in WER (%).*

Methods	FT data (h)	test clean		test other	
		offline	online	offline	online
SL	NA	3.6	4.7	9.7	12.4
+ MFT	1k	3.6	4.7	9.8	12.5
SSL	NA	3.3	4.2	7.7	10.3
+ MFT	1k	3.3	4.3	7.7	10.3
Direct Avg.	NA	100	100	100	100
+ *fully* FT	1k	3.5	4.6	9.6	12.1
Aligned Avg.	NA	52.7	63.0	78.4	82.9
+ MFT (Our OTF)	1k	**3.0**	**4.0**	**7.4**	**10.1**

Table 4: *Ablation studies on weight alignment and Moderate Fine-Tuning (MFT) for Chinese models in CER (%).*

Methods	FT data (h)	test Mandarin		test Cantonese	
		offline	online	offline	online
SL	NA	12.8	15.7	12.6	13.8
+ MFT	2.6k	12.7	15.6	12.8	13.9
SSL	NA	12.2	14.5	12.9	14.0
+ MFT	2.6k	12.1	14.4	12.9	14.1
Direct Avg.	NA	100	100	100	100
+ *fully* FT	2.6k	12.9	15.8	13.5	14.2
Aligned Avg.	NA	97.5	99.2	92.1	97.6
+ MFT (Our OTF)	2.6k	**11.1**	**12.7**	**11.8**	**13.1**

Decoding. Following the representative SSL works [6, 7, 31], we evaluate the performance of all models using the Connectionist Temporal Classification (CTC) beam search decoder [45] with 10 the decoding beam size. Note that no language model is applied in our experiments.

4.2. Main experiment

Baseline methods. To evaluate the performance of our method on the fusion of SL and SSL models, we implement the SOTA EBF method via combining the outputs of the two models' encoders [24], and the existing SBF method based on the average word confidence [23] as our baseline methods.

System performance. As shown in Table 1-2, we compare OTF with the performance of the SL and SSL models, and the baseline methods in offline and online modes. Compared with the SL and SSL models, OTF achieves a large improvement with lower WERs/CERs. Numerically, compared with the better of SL and SSL models, OTF yields relative WERs/CERs reductions up to 9.1% and 12.4% in English and Chinese, respectively. Although SBF and EBF are effective in model fusion, they suffer from a high inference cost with aggregating the two models' parameters. Besides, we find the baseline methods can cause a worse performance in some test sets. We infer 1) the confidence scoring of SBF might be not accurate enough for hypothesis selections; and 2) it would be challenging for EBF to align the features of two adequately-trained but diverse models. Compared with the baselines, our method consistently achieves the best performance in both the English and Chinese scenarios. Moreover, since each layer of the individual models' weights are fused into a single one, OTF is more efficient without increasing the model size than the baseline methods [23,24].

4.3. Ablation study and discussion

To further evaluate the effectiveness of OTF, we analyze the proposed method from the following three perspectives.

1) Ablation study. To analyze the effect of weight alignment and moderate fine-tuning in OTF, we compare our proposed method with: direct averaging of the SL and SSL models; SL models with moderate fine-tuning; and SSL models with moderate fine-tuning (see Table 3-4). **(a) Weight alignment.** Before fine-tuning, we find that direct averaging of the two models will cause the recognition results to be all empty with 100% error rate. Although the performance can be substantially improved after fully fine-tuning on labeled data, the performance is inferior to the better one of the individual models. We infer the direct averaging of the two different models would damage the well-trained models and lead to a suboptimal solution. Differently, our method averages the layer-wise weights via alignment, which can still partly recognize the speech even without fine-tuning. We also find that the averaging after alignment for the ASR tasks achieves a larger performance degradation compared with the image classification tasks [29]. We infer it is more difficult to align the weights of sequence-to-sequence ASR models than the instance classification models. However, the proposed OTF still achieves the best performance than other methods with limited fine-tuning steps. **(b) Moderate fine-tuning.** As for the moderate fine-tuning in OTF, we compared SL and SSL models with the same fine-tuning to test if the performance improvement of our method is caused by more training steps. Since the SL and SSL are adequately trained, the performance is almost the same before and after the moderate fine-tuning. The experimental results show that our method fuses the two models effectively, and can obtain a better initialization to be fine-tuned for performance enhancement.

2) Visualization of loss landscape. From the perspective of the optimization property, we plot the loss landscape of OTF, as shown in Fig. 2. The curves of loss landscape show that our method achieves the flattest and lowest minimum points than both the SL and SSL models. It implies that OTF fuses the two models and incorporates both of the individual advantages in generalization and in-domain optimization as we expected.

3) Performance on small labeled datasets. As so far, we have verified the effectiveness of OTF on abundant datasets, i.e. one thousand hours and more. However, we still wonder about the performance when only a small amount of labeled dataset is available. Here, we assume that only a 100-hour train-clean subset of Librispeech [43] and a 50-hour Cantonese dataset are labeled. As shown in Table 5-6, the relative improvement of OTF decreases when there is a large performance gap between the two models. We infer that if there is a large gap between the two constituent models, the better one will dominate the results of model fusion. Nevertheless, our method can still improve the performance of the fused models with small labeled datasets.

Table 5: *Performance on small English dataset in WER (%).*

Methods	training data (h)		test clean		test other	
	unlabeled	labeled	offline	online	offline	online
SL [41]	NA	100	8.3	9.7	21.9	25.3
SSL [42]	1k	100	5.5	6.9	12.8	**16.8**
OTF	NA	100	**5.4**	**6.5**	**12.7**	**16.8**

Table 6: *Performance on small Cantonese dataset in CER (%).*

Methods	training data (h)		test Cantonese	
	unlabeled	labeled	offline	online
SL [41]	NA	50	16.6	18.2
SSL [42]	150k	50	14.6	16.7
OTF	NA	50	**14.2**	**16.3**

5. Conclusions

We proposed a novel weight-based fusion method for SL and SSL ASR models via optimal transport, which improved the recognition performance without increasing the inference cost. Extensive experiments on English and Chinese dataset were conducted to verify the effectiveness of the proposed method. However, models with the same architecture are used to test the performance, the fusion of heterogeneous models will be further investigated in our future work.

6. References

[1] A. Babu, C. Wang, A. Tjandra *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022.

[2] A. Khurana, A. Laurent, and J. Glass, "Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0," in *Proc. ICASSP*, 2022.

[3] J. Zhao and W. Zhang, "Improving automatic speech recognition performance for low-resource languages with self-supervised models," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[4] J. Zuluaga-Gomez, A. Prasad, I. Nigmatulina *et al.*, "How does pre-trained wav2vec 2.0 perform on domain-shifted asr? an extensive benchmark on air traffic control communications," in *Proc. SLT*, 2022.

[5] B. Thomas, S. Kessler, and S. Karout, "Efficient adapter transfer of self-supervised speech models for automatic speech recognition," in *Proc. ICASSP*, 2022.

[6] A. Baevski, Y. Zhou, A. Mohamed *et al.*, "Wav2vec2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[7] W. Hsu, B. Bolte, Y. Tsai *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[8] S. Chen, C. Wang, Z. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[9] A. Mohamed, H. Lee, L. Borgholt *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[10] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *Proc. ICASSP*, 2020.

[11] S. Yang, P. Chi, Y. Chuang *et al.*, "Superb: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021.

[12] Y. Chung, Y. Zhang, W. Han *et al.*, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. ASRU*, 2021.

[13] Y. Zhang, D. Park, W. Han *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[14] C. Wang, Y. Wu, S. Liu *et al.*, "Unispeech at scale: An empirical study of pre-training method on large-scale speech recognition dataset," *arXiv preprint arXiv:2107.05233*, 2021.

[15] A. Radford, J. Kim, T. Xu *et al.*, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[16] J. Li, Y. Wu, Y. Gaur *et al.*, "On the comparison of popular end-to-end models for large scale speech recognition," in *Proc. Interspeech*, 2020.

[17] A. Gulati, J. Qin, C. Chiu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020.

[18] L. Fu, X. Li, R. Wang *et al.*, "Scala: Supervised contrastive learning for end-to-end speech recognition," in *Proc. Interspeech*, 2022.

[19] W. Hsu, A. Sriram, A. Baevski *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Proc. Interspeech*, 2021.

[20] C. Zan, L. Ding, L. Shen *et al.*, "On the complementarity between pre-training and random-initialization for resource-rich machine translation," in *Proc. COLING*, 2022.

[21] S. Bucci, A. D'Innocente, Y. Liao *et al.*, "Self-supervised learning across domains," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[22] H. Li, Z. Xu, G. Taylor *et al.*, "Visualizing the loss landscape of neural nets," in *Proc. NeurIPS*, 2018.

[23] V. Soto, O. Siohan, M. Elfeky *et al.*, "Selection and combination of hypotheses for dialectal speech recognition," in *Proc. ICASSP*, 2016.

[24] A. Arunkumar, V. Sukhadia, and S. Umesh, "Investigation of ensemble features of self-supervised pretrained models for automatic speech recognition," in *Proc. Interspeech*, 2022.

[25] C. Tang, Y. Wang, X. Chen *et al.*, "Exploring effective fusion algorithms for speech based self-supervised learning models," in *Proc. NCMMSC*, 2022.

[26] T. Wu, T. Hsu, C. Li *et al.*, "The efficacy of self-supervised speech models for audio representations," in *Proc. HEAR (NeurIPS 2021 Competition)*, 2022.

[27] Y. Sun and L. Fu, "Stacking ensemble learning for non-line-of-sight detection of global navigation satellite system," *IEEE Transactions on Instrumentation and Measurement*, 2022.

[28] M. Yurochkin, M. Agarwal, S. Ghosh *et al.*, "Bayesian nonparametric federated learning of neural networks," in *Proc. ICML*, 2019.

[29] S. Singh and M. Jaggi, "Model fusion via optimal transport," in *Proc. NeurIPS*, 2020.

[30] D. Qiu, Q. Li, Y. He *et al.*, "Learning word-level confidence for subword end-to-end asr," in *Proc. ICASSP*, 2021.

[31] A. Baevski, W. Hsu, Q. Xu *et al.*, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. ICML*, 2022.

[32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NeurIPS*, 2015.

[33] K. Huang, T. Feng, Y. Fu *et al.*, "Ensemble knowledge distillation of self-supervised speech models," in *Proc. ICASSP*, 2023.

[34] S. Cao, Y. Kang, Y. Fu *et al.*, "Improving streaming transformer based asr under a framework of self-supervised learning," in *Proc. Interspeech*, 2021.

[35] R. Takashima, L. Sheng, and H. Kawai, "Investigation of sequence-level knowledge distillation methods for ctc acoustic models," in *Proc. ICASSP*, 2019.

[36] C. Talnikar, T. Likhomanenko, R. Collobert *et al.*, "Joint masked cpc and ctc training for asr," in *Proc. ICASSP*, 2021.

[37] C. Wang, Y. Wu, Y. Qian *et al.*, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *Proc. ICML*, 2021.

[38] J. Bai, B. Li, Y. Zhang *et al.*, "Joint unsupervised and supervised training for multilingual asr," in *Proc. ICASSP*, 2022.

[39] N. Keskar, D. Mudigere, J. Nocedal *et al.*, "On large-batch training for deep learning: Generalization gap and sharp minima," in *Proc. ICLR*, 2017.

[40] H. Wang, M. Yurochkin, Y. Sun *et al.*, "Federated learning with matched averaging," in *Proc. ICLR*, 2020.

[41] Z. Yao, D. Wu, X. Wang *et al.*, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*, 2021.

[42] L. Fu, S. Li, Q. Li *et al.*, "Ufo2: A unified pre-training framework for online and offline speech recognition," in *Proc. ICASSP*, 2023.

[43] V. Panayotov, G. Chen, D. Povey *et al.*, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[44] J. Yu, W. Han, A. Gulati *et al.*, "Dual-mode asr: Unify and improve streaming asr with full-context modeling," in *Proc. ICLR*, 2021.

[45] A. Graves, S. Fernandez, F. Gomez *et al.*, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.