The 24th INTERSPEECH Conference, August 20th to 24th, 2023, Dublin, Ireland

# Knowledge Transfer from Pre-trained Language Models to CIF-based Speech Recognizers via Hierarchical Distillation

*Minglun Han, Feilong Chen, Jing Shi, Shuang Xu, Bo Xu.*
*Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China.*
*E-mail: hanminglun.cn@gmail.com*

*Knowledge Transfer from Pre-trained Language Models to CIF-based Speech Recognizers via Hierarchical Distillation*

## Abstract

Large-scale pre-trained language models (PLMs) have shown great potential in natural language processing tasks. Leveraging the capabilities of PLMs to enhance automatic speech recognition (ASR) systems has also emerged as a promising research direction. However, previous works may be limited by the inflexible structures of PLMs and the insufficient utilization of PLMs. To alleviate these problems, we propose the hierarchical knowledge distillation (HKD) on the continuous integrate-and-fire (CIF) based speech recognition models. To transfer knowledge from PLMs to the ASR models, HKD employs cross-modal knowledge distillation with contrastive loss at the acoustic level and knowledge distillation with regression loss at the linguistic level. Compared with the original CIF-based model, our method achieves 15% and 9% relative error rate reduction on the AISHELL-1 and LibriSpeech, respectively.

## Introduction

End-to-end (E2E) models have recently made remarkable progress on speech recognition tasks. Compared with hybrid models, E2E models are optimized in a unified structure. However, the tight integration in this unified structure hinders the infusion of linguistic knowledge and limits the use of large-scale corpora.

Currently, there are two popular approaches widely used to leverage unpaired text for E2E ASR models: language model (LM) fusion and re-scoring. Apart from them, using large-scale pre-trained language models (PLMs) to improve language modeling of ASR models is also a practical approach to make use of unpaired text dataset. PLMs possess powerful language modeling abilities, and their outputs contain rich linguistic information that can improve ASR language modeling. Therefore, employing PLMs to improve speech recognition has gradually become an important research direction. Until now, the methods used to improve ASR with PLMs can be categorized into three classes.

### Taxonomy

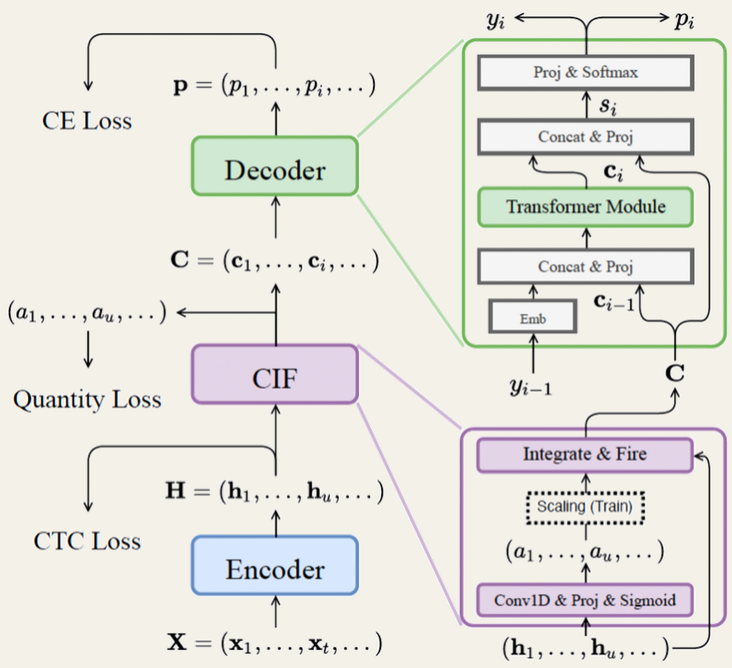
1. Re-scorer based method
2. Model-based method
3. Knowledge distillation (KD) based method

In this paper, we propose a knowledge transfer strategy called hierarchical knowledge distillation (HKD). HKD transfers linguistic knowledge from PLMs to different levels of the CIF-based ASR model, including the acoustic level. However, it is not easy to directly transfer linguistic knowledge to the acoustic level of E2E models. Unlike other end-to-end mechanisms, the continuous integrate-and-fire (CIF) mechanism, which generates token-level acoustic representations aligned with the text, provides a natural option for the KD at the acoustic level. Thus, we develop the HKD based on the CIF-based ASR model.

### Hierarchical Knowledge Distillation

- Acoustic level → Contrastive loss
- Linguistic level → Regression loss

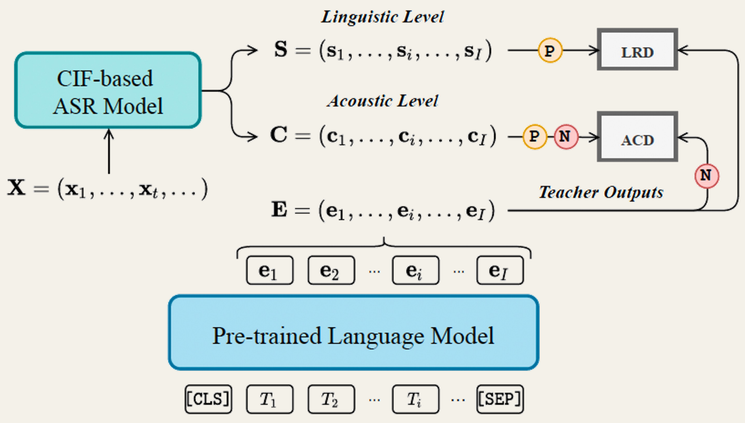
## Methods




**CIF-based model:** The CIF-based ASR model in this work consists of an acoustic encoder, a CIF module, and a decoder.

Acoustic Level: $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_i, \dots, \mathbf{c}_I)$

Linguistic Level: $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_i, \dots, \mathbf{s}_I)$




**Hierarchical knowledge distillation:** LRD denotes linguistic regression distillation, and ACD denotes acoustic contrastive distillation. `P` is projection, and `N` is L2 normalization. The total training loss can be written as $\mathcal{L}_{Total} = \mathcal{L}_{ASR} + \lambda_{AD}\mathcal{L}_{AD} + \lambda_{LD}\mathcal{L}_{LD}$

**Acoustic contrastive distillation:**

Two problems: (1) Modality gap (2) Structure gap. Contrastive loss forces the model to pull together the positive pairs and push apart the negative pairs.

$$\mathcal{L}_{AD}^{cont} = -\frac{1}{N}\sum_{n=1}^{N}\frac{1}{I^n}\sum_{i=1}^{I^n} log\frac{s(\bar{\mathbf{c}}_i, \bar{\mathbf{e}}_i)}{\sum_{k=1}^{K} s(\bar{\mathbf{c}}_i, \bar{\mathbf{e}}_{n,i,k}^{-}) + s(\bar{\mathbf{c}}_i, \bar{\mathbf{e}}_i)}$$

**Linguistic regression distillation:**

Using MSE loss to distill the knowledge from PLMs to the final linguistic representations.

$$\mathcal{L}_{LD}^{mse} = \alpha_{mse} \cdot \frac{1}{N}\sum_{n=1}^{N}\frac{1}{I^n}\sum_{i=1}^{I^n}\sum_{d=1}^{D}(\hat{s}_{i,d}^{n} - e_{i,d}^{n})^2$$

## Experimental Results

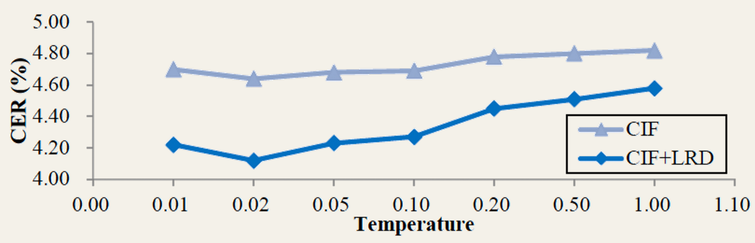
### Main Results on AISHELL-1

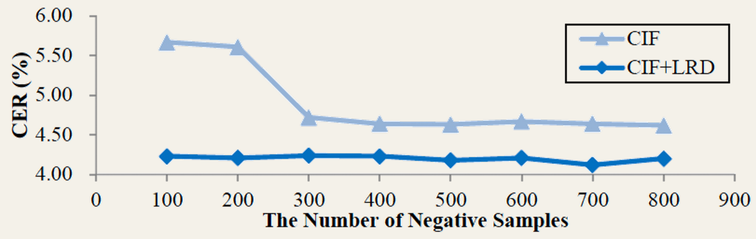
| Model | LM | # Param | dev (%) | test (%) |
|---|---|---|---|---|
| ESPnet Conformer [35] | ✗ | 46 M | 4.5 | 4.9 |
| ESPnet Conformer [35] | ✓ | 46 M | 4.4 | 4.7 |
| Branchformer [36] | ✗ | 45 M | 4.2 | 4.4 |
| WeNet [37] | ✓ | 46 M | - | 4.4 |
| Icefall | ✓ | - | - | 4.3 |
| Neural Transducer [38] | ✓ | 90 M | 3.8 | 4.1 |
| CIF | ✗ | 47 M | 4.5 | 4.9 |
| + ACD | ✗ | 47 M | 4.2 | 4.7 |
| + LRD | ✗ | 47 M | 4.0 | 4.5 |
| + HKD | ✗ | 47 M | 3.8 | 4.2 |
| CIF | ✓ | 47 M | 4.4 | 4.8 |
| + ACD | ✓ | 47 M | 4.2 | 4.6 |
| + LRD | ✓ | 47 M | 4.0 | 4.4 |
| + HKD | ✓ | 47 M | **3.8** | **4.1** |

### Comparison between Distillation Losses

| Model | LRD | AD | AD Loss | w/o LM dev / test | w/ LM dev / test |
|---|---|---|---|---|---|
| CIF | ✗ | ✗ | - | 4.5 / 4.9 | 4.4 / 4.8 |
| | ✗ | ✓ | MSE | 4.4 / 4.9 | 4.4 / 4.8 |
| | ✗ | ✓ | COS | 4.5 / 4.9 | 4.4 / 4.8 |
| | ✗ | ✓ | CONT | 4.2 / 4.7 | 4.2 / 4.6 |
| | ✓ | ✗ | - | 4.0 / 4.5 | 4.0 / 4.4 |
| | ✓ | ✓ | MSE | 4.0 / 4.5 | 4.0 / 4.5 |
| | ✓ | ✓ | COS | 4.1 / 4.5 | 4.0 / 4.4 |
| | ✓ | ✓ | CONT | **3.8 / 4.2** | **3.8 / 4.1** |

### Investigations on effects of Hyper-parameters




### Main Results on LibriSpeech

| Model | dev clean | dev other | test clean | test other |
|---|---|---|---|---|
| CIF | 3.0 | 7.3 | 3.3 | 7.7 |
| + ACD | 3.0 | 7.2 | 3.2 | 7.3 |
| + LRD | 2.8 | 6.9 | 3.1 | 7.1 |
| + HKD | **2.7** | **6.9** | **3.0** | **7.0** |

## Takeaways

1. With the help of pretrained language models, HKD effectively improves the ASR performance in both Chinese and English;
2. Contrastive loss shows superior performance in acoustic-level distillation in our experiments;
3. Acoustic-level distillation and linguistic-level distillation are complementary. Linguistic-level distillation helps stabilize the training of acoustic-level distillation.

## Conclusion

In this work, we introduce a hierarchical knowledge distillation strategy to transfer PLM knowledge to different levels of the CIF-based ASR model. Specifically, we use acoustic contrastive distillation at the acoustic level and linguistic regression distillation at the linguistic level. Compared to the CIF-based ASR baseline, our method brings 15% relative CER reduction on AISHELL-1 and 9% relative WER reduction on LibriSpeech. We will explore our methods with larger-scale language models in the future.

## Acknowledgements

1. The National Key Research and Development Program of China (2018AAA0100400)
2. The Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDA27030300)
3. The National Natural Science Foundation of China (62206294)

## Contact Information

*E-mail : hanminglun.cn@gmail.com*

Github

Google Scholar