CIF-FSA for Low-resource Automatic Speech Recognition in Jiao-Liao Mandarin

Xuchen Li, Ming Tan, Wei Liu, Kang Xie, Xuanda Chen, Jie Liu and Meixia Qu

Abstract—**Jiao-Liao Mandarin speech recognition constitutes a critical domain in Chinese dialect recognition research, holding considerable commercial value and profound cultural significance for preserving and transmitting this regional linguistic heritage. Nevertheless, the scarcity of annotated textual corpora and high-quality audio materials has substantially impeded progress in this research domain. To address this issue, this paper introduces the development of JLMS30, a 30-hour Jiao-Liao Mandarin speech recognition dataset created for research purposes. Utilizing this self-built dataset, we propose an optimized feature extraction method and a novel semantic modeling training strategy for Continuous Integrate-and-Fire (CIF) based speech recognition, called CIF-FSA. Firstly, the proposed CIF-FSA integrates a lightweight module (FG-Conv) to enhance discriminability in speech signal features. Subsequently, we propose a cross-modal knowledge distillation mechanism termed Semantic Acoustic Contrastive Distillation (SACD), which effectively transfers linguistic knowledge from a pre-trained language model into the CIF framework, improving its semantic modeling capabilities. Our model achieves a 15.5% relative reduction in Character Error Rate (CER) compared to prior approaches and reduces model parameters by 6%, significantly enhancing overall speech recognition accuracy. These findings highlight our dataset's and model design's efficacy in advancing speech recognition technology for low-resource languages.**

Index Terms—**Speech recognition，Jiao-Liao Mandarin，CIF，Knowledge distillation**

Manuscript received November 27, 2024; revised March 31, 2025.

This work was supported by the Key Lab of Information Network Security, Ministry of Public Security, and the Shenzhen Fundamental Research Program under Grant JCYJ20230807094104009.

Xuchen Li is a postgraduate student at Shenzhen Research Institute of Shandong University, Shandong University, Shenzhen 518057 China (e-mail: 202237557@mail.sdu.edu.cn).

Ming Tan is an Assistant Engineer of Office of Asset and Laboratory Management, Shandong University, Weihai 264209 China (e-mail: tom826@163.com).

Wei Liu is a lecturer at the College of Business Administration, Shandong University of Finance and Economics, Jinan 250014 China (e-mail: vivian.liu@sdufe.edu.cn).

Kang Xie is an engineer of Key Lab of Information Network Security, Ministry of Public Security, Shanghai 200031 China (e-mail: xiekang@stars.org.cn).

Xuanda Chen is a postgraduate student at Shenzhen Research Institute of Shandong University, Shandong University, Shenzhen 518057 China (e-mail: 202337570@mail.sdu.edu.cn).

Jie Liu is an engineer of Shenzhen Research Institute of Shandong University, Shandong University, Shenzhen 518057 China (e-mail: liujiesdwh@163.com).

Meixia Qu is an associate professor at the School of Mechanical, Electrical Information Engineering, Shandong University, Weihai 264209 China (Co- corresponding author e-mail: mxqu@sdu.edu.cn).

I. Introduction

END-TO-END Automatic Speech Recognition (ASR) technologies have demonstrated remarkable efficacy for widely spoken languages, such as English and Chinese, primarily attributable to the abundance of extensive speech data resources [1]. Conversely, approximately 6,000 languages globally are categorized as low-resource, needing more transcribed speech data [2], [3]. Transcribing speech using human-annotated data is often labor-intensive and costly [4], presenting significant challenges in developing high-performance ASR systems for low-resource languages [5].

Jiao-Liao Mandarin is a significant dialect in China, carrying profound historical and cultural significance. Spoken by a substantial population across multiple provinces, Jiao-Liao Mandarin is shaped by the diverse cultures and customs of various ethnic groups, endowing it with rich local characteristics and social practices. These factors confer considerable research value within the field of dialectology. The application of speech recognition technology to transcribe Jiao-Liao Mandarin into written form facilitates communication among residents on social networks. It aids linguists in examining the dialect's phonetic, lexical, and grammatical features. This initiative possesses substantial commercial potential and is essential for understanding the evolution of dialects and their interconnections with Standard Mandarin.

In this context, the present study establishes a novel Jiao-Liao speech corpus. It employs state-of-the-art deep learning techniques in ASR to explore end-to-end recognition of Jiao-Liao Mandarin. All models were trained from the ground up using this newly created dataset, and through enhancements to the base model, we achieved a significant improvement in recognition accuracy. This research addresses the challenge of dialect preservation in Jiao-Liao Mandarin speech recognition by developing an advanced deep neural network model. This study, therefore, aimed to achieve the following objectives:

(i) Establish a multi-domain Jiao-Liao Mandarin speech dataset for dialectological research.

(ii) Design an effective Jiao-Liao Mandarin speech recognition system by integrating a lightweight convolution module that enhances the discriminability of speech signal features.

(iii) Propose a new knowledge distillation strategy that improves the model's semantic modeling capabilities.

(iv) Play a crucial role in protecting and transmitting this dialect.

II. Related Works

A. Speech Recognition

With the advancement of deep learning, end-to-end models have exhibited robust recognition performance and system stability, establishing them as a prominent area of research for scholars both domestically and internationally [6] [7] [8]. Gulati et al. [9] introduced the Conformer model, which integrates a convolution module with an attention mechanism to extract local and global contextual information from audio, achieving state-of-the-art accuracy. Dong et al. [10] proposed the CIF mechanism to enhance the model's capability for acoustic boundary detection by integrating encoder outputs through cumulative weight calculations and weighted sums of states. Li [11] suggested employing Grouped Attention as a substitute for the traditional attention mechanism in the Conformer model, thereby reducing computational complexity and resource requirements while accelerating inference speed. Peng [12] recommended utilizing an attention mechanism alongside convolutional gating within a multi-layer perceptron module to extract and fuse global and local features independently. Kim [13] enhanced the Branchformer model by merging components and incorporating additional pointwise modules. Mai et al. [14] proposed an innovative attention mechanism termed HyperMixer, which extends the capabilities of the Conformer model while outperforming it in terms of inference speed, memory usage, parameter count, and available training data. Parcollet et al. [15] introduced a novel linear self-attention mechanism known as SummaryMixing to improve training efficiency, inference speed, and overall model performance. Numerous research initiatives are currently dedicated to developing more efficient linear attention mechanisms to minimize resource consumption while maximizing model performance. However, advancements in the convolution module still need to be improved. Traditional convolution modules rely on pointwise convolution to achieve linear combinations and fusion among neurons at specific channel positions. This methodology constrains their ability to effectively capture intricate feature relationships across spatial locations, impeding their ability to learn more descriptive cues. Inspired by [16], we propose a lightweight FG-Conv module. In this module, we employ pointwise group convolution to conduct independent convolution operations within each group, enhancing the model's capability to capture feature relationships across spatial dimensions and improving functional expressiveness. Furthermore, by implementing a channel shuffle strategy, we systematically or randomly rearrange the neurons across different channels or spatial locations to evaluate their significance and optimize the relationship between features and weights. These enhancements significantly improve the model's comprehension of complex speech signals and further elevate its performance in speech recognition tasks.

B. Jiao-Liao Mandarin Speech Recognition

Jiao-Liao Mandarin, a significant component of Chinese dialects, holds immense value for linguistic research and speech technology. Currently, the only available speech resource for scholarly investigation is the Jiao-Liao section of the KeSpeech dataset [17], which primarily emphasizes themes such as news, technology, and sports.

In the study of Jiao-Liao Mandarin, Shao et al. [18] proposed the Decoupling and Interacting Multi-task Network, which enhances recognition performance in speech and accent recognition tasks through joint training. This approach facilitates complementary information interactions at various granularities, significantly improving performance for both functions. Mu [19] introduced a unified generative error correction model for speech and accent recognition, termed MMGER, which employs multi-modal and multi-granularity calibration techniques. Tang et al. [20] implemented Pinyin regularization for prompts to fine-tune large language models, enhancing the error correction capabilities of automatic speech recognition systems. Chen et al. [21] developed the Layer-adapted Module model that extracts fine-grained prosodic information from different layers of the acoustic encoder and promotes frame-by-frame correction of ASR results via a cross-attention module. Gu et al. [22] proposed a personality-aware training framework to adapt pre-trained ASR models to target speakers, addressing mismatches between training and testing conditions in end-to-end automatic speech recognition.

Despite the technological advancements achieved through these methods, research explicitly focused on Jiao-Liao Mandarin remains relatively limited. Furthermore, existing datasets exhibit significant deficiencies in representing the unique vocabulary and regional idioms of Jiao-Liao Mandarin, which constrains the comprehensive development of speech recognition technology for this dialect. This paper establishes a Jiao-Liao Mandarin speech recognition dataset named JLMS30, encompassing common themes from daily life along with a rich array of unique vocabulary and regional idioms, thereby providing essential data support for advancing speech recognition technology in Jiao-Liao Mandarin.

C. Pre-trained Language Models for Automatic Speech Recognition

The emergence of pre-trained language models has created new opportunities for advancing speech recognition technology. Analyses of enhancing speech recognition using these models can be categorized into three distinct types: restorer-based, model-based, and knowledge distillation [23].

In contrast to the first two methods, knowledge distillation-based approaches focus on optimizing the speech recognition model. Lu et al. [24] proposed a cross-modal knowledge transfer learning framework that aligns hierarchical acoustic features with linguistic features, enabling the acoustic encoder to acquire rich linguistic knowledge. Futami et al. [25] aligned a Connectionist Temporal Classification model's frame-level predictions with BERT's word-level predictions and performed knowledge distillation. Kubo [24] investigated knowledge distillation methods involving attention-based decoders and Transducer-based decoders in conjunction with pre-trained language models. Han [26] introduced a hierarchical knowledge distillation (HKD) method for the CIF model, applying cross-modal distillation using contrastive loss at the acoustic level and regression loss at the linguistic level to

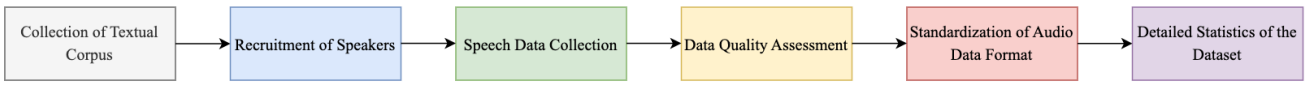

Fig. 1. Flowchart of the Dataset Construction Process.

extract knowledge from pre-trained language models into the ASR model.

While current knowledge distillation methods have made notable advancements in speech recognition, these studies focus excessively on transferring local textual information to acoustic representations, thereby overlooking the importance of overall semantic context. This paper introduces a sentence-level semantic contrast distillation strategy aimed at enhancing the learning of sentence coherence and improving the model's ability to manage long sentences and complex syntactic structures.

III. JLMS30 Datasets

This paper presents JLMS30, a Jiao-Liao Mandarin speech recognition dataset of 27,222 samples lasting 30 hours. This dataset is an extension and optimization of JLMS25 [27], further enhancing its scale and quality. As illustrated in Fig. 1, the process of constructing the dataset is outlined below:

(1) Textual Corpus Collection: To enrich the dataset with diverse and comprehensive textual content, three methods were utilized for its collection:

•Text from various online platforms, including social media, forums, and blogs, was gathered to ensure a rich and diverse dataset. Texts were carefully reviewed during the data cleaning process to remove any inappropriate content, such as politically sensitive issues, privacy violations, pornography, or violence. Additionally, overly long texts were shortened, unique tags, punctuation marks, and emojis were eliminated, and Arabic numerals were converted into their corresponding Chinese character forms.

•A portion of text from the Mandarin speech recognition dataset Aishell-1 [28] was incorporated, covering various domains such as finance, science and technology, and sports.

•Everyday expressions and folk proverbs collected from the Jiao-Liao Mandarin region were included. These texts highlight the area's unique linguistic characteristics.

(2) Recruitment of Speakers: To ensure both the linguistic authenticity and regional representativeness of the speaker samples, we meticulously selected speakers based on the following criteria:

•Speakers have resided in the Jiao-Liao Mandarin-speaking region for an extended period without any long-term (over one year) residence outside the area.

•Speakers demonstrate strong proficiency in the dialect and use Jiao-Liao Mandarin as their primary mode of communication in daily life.

(3) Speech Data Collection: We adopt a standardized process for speech data collection, which consists of the following steps: First, the researchers provide speakers with pre-prepared text materials in advance, ensuring they are well-acquainted with the content and can read it fluently. All recordings are conducted in a quiet environment to guarantee the quality of the speech samples. To enhance the diversity and representativeness of the speech data and ensure high-quality speech samples from a professional recording environment while also capturing more representative, everyday speech data, a hybrid collection approach combining both offline and online methods is employed:

•Offline Collection: In a quiet environment, speakers read the provided text aloud, sentence by sentence, using a high-fidelity AT2020 microphone. Afterward, the audio is edited using Adobe Audition software.

•Online Collection: Speakers use Android or iOS mobile devices to record their speech using the system's built-in recording applications. Before recording, all speakers receive standardized pronunciation guidance to ensure pronunciation and speech rate consistency.

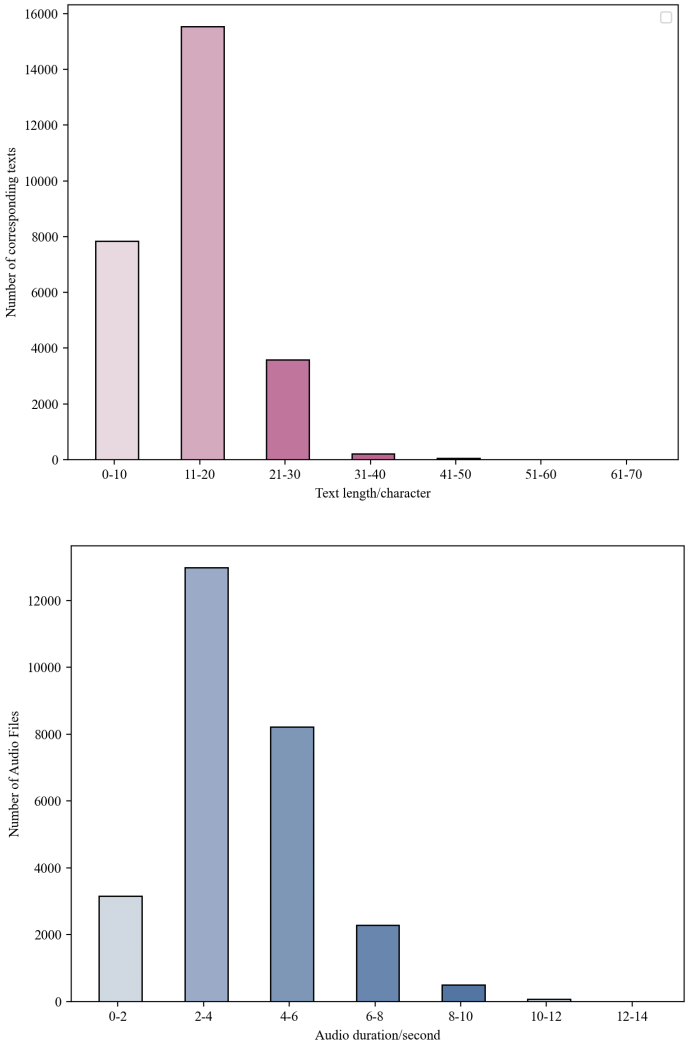

Fig. 2. The distribution text length and audio duration.

(4) Data Quality Assessment: We implement a rigorous quality control process to ensure the reliability of our speech data:

•Utilize professional audio analysis software to evaluate the SNR of the collected speech data, eliminating low-quality samples with excessive environmental noise.

•Assess the alignment between the spoken content and the provided text, removing samples with inconsistencies such as pronunciation errors, omissions, or additions.

(5) Standardization of Audio Data Format: All audio files were converted to mono WAV format, with a sampling rate of 16 kHz and a bit depth of 16.

(6) Detailed Statistics of the Dataset: We have conducted a detailed analysis of the distribution of text length and audio duration, the distribution of text domains, and the age and gender information of the speakers.

•The distribution of text length and audio duration is illustrated in Fig. 2. The text lengths range from 1 to over 50 characters, with the majority falling between 1 and 30 characters. As the number of characters increases, the frequency of texts decreases sharply. The audio samples range in duration from 1 to 12 seconds, with a significant proportion lasting between 2 and 6 seconds. This reflects the typical characteristics of everyday spoken communication.

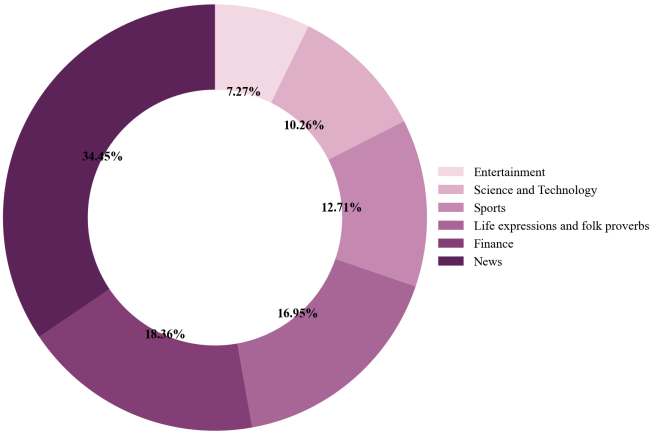

Fig. 3. The distribution of text domains.

•The distribution of text domains within the dataset is shown in Fig. 3. The corpus covers several common domains, with news-related texts accounting for the largest proportion at 34.45%. Texts related to daily expressions and folk proverbs make up 18.36%, ranking second and highlighting the richness of everyday spoken communication. Other domains, such as entertainment (16.95%), technology (12.71%), sports (10.26%), and finance (7.27%), also contribute significant proportions, together forming a diverse and representative distribution of text domains. This balanced distribution supports the development of a speech recognition model with broad adaptability, improving its generalization across various application scenarios.

• Regarding the age and gender information of the speakers, we recruited 57 local speakers from the Jiao-Liao Mandarin region, all of whom had no history of long-term residence outside the area, for audio collection. Fig. 4 provides a detailed overview of the gender and age distribution of the speakers. Overall, the proportion of female speakers is slightly higher than that of male speakers, with the largest age group comprising speakers aged 21 to 30.

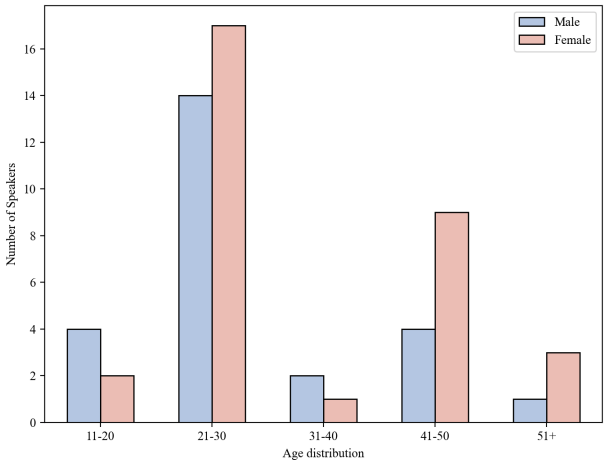

Fig. 4. Speaker age and gender information.

IV. Methods

This section delineates the selected baseline model for speech recognition, along with the newly proposed module and knowledge distillation strategy.

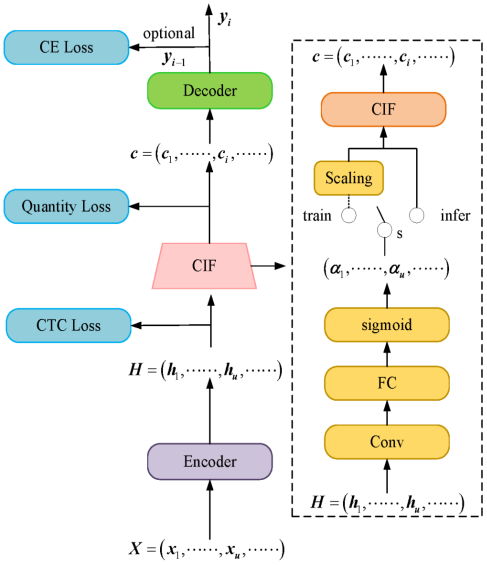

Fig. 5. The CIF-based ASR model.

A. CIF

The CIF [29], [30] functions as middleware that connects the acoustic encoder and decoder based on the principle of integrated distribution and has been extensively studied and applied. Due to the significantly higher sampling rate of acoustic features than textual features [6], directly transferring semantic knowledge to an end-to-end model presents challenges. Unlike other end-to-end speech recognition architectures, the CIF model can convert frame-level acoustic features into character-level features aligned with text. This capability markedly enhances the cross-modal learning process. Consequently, this paper focuses on CIF as the primary research object, with its model and module structure illustrated in Fig. 5.

B. FG-Conv module

Traditional convolution modules rely on pointwise convolutions to perform linear combinations and fusions

exclusively among neurons at channel positions. This approach limits their capacity to effectively capture complex feature relationships across spatial locations and constrains their ability to learn more descriptive cues. Inspired by the successful research [16], we propose a novel lightweight module termed FG-Conv, as illustrated in Fig. 6. In this module, we employ pointwise group convolution to execute independent convolution operations within each group, thereby enhancing the ability to capture feature relationships across spatial locations and improving the expressive power of the features. Furthermore, by implementing a channel shuffle strategy, we can systematically or randomly rearrange neurons at various channel or spatial positions to assess their importance and optimize features with complete weights.

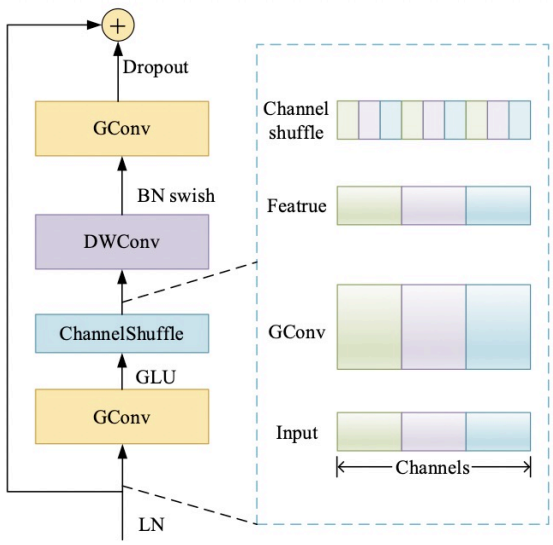

Fig. 6. The structure of FG-Conv module.

In speech recognition, input spectral features are often treated as images, making the task of speech signal recognition analogous to image classification. Building on this concept, we replaced the convolution module in the encoder with the FG-Conv module to enhance the distinguishability of speech signal features. This enhancement improves the representation of speech signals and ultimately increases the accuracy of predicted text.

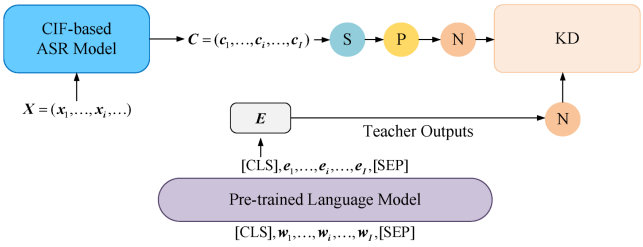

Fig. 7. Semantic acoustic knowledge distillation. P denotes projection, and N denotes L2 normalization.

C. Semantic acoustic contrastive distillation

Pre-trained language models demonstrate robust generalization capabilities, producing outputs rich in semantic knowledge and achieving impressive results across various natural language processing tasks. Han et al. [23] implemented a knowledge distillation strategy during training, leveraging pre-trained language models to extract semantic features for each character within the text data. The model effectively learns the relationship between acoustics and text by aligning character-level acoustic features with their corresponding semantic features, enhancing its understanding of speech transcription tasks. However, an excessive emphasis on fine-grained features during the semantic alignment from acoustics to text may compromise overall coherence, potentially resulting in the loss of critical semantic information. To bridge the semantic gap more effectively between acoustics and text, this paper introduces semantic acoustic contrastive distillation, which aims to leverage comprehensive semantic features of text to assist the model in learning acoustic features from a sentence-level perspective. The semantic acoustic knowledge distillation process is illustrated in Fig. 7.

During the training phase, character-level acoustic features $C=\left(c_1,\cdots\cdots,c_i,\cdots\cdots,c_I\right)$ are first processed $\sum_{i=1}^{I}c_i$ to obtain a representation C, which is then projected to match the dimensionality of the sentence-level semantic features E = [CLS]. These features are subsequently normalized to yield $\overline{C}$. The loss function for semantic acoustic contrastive distillation is calculated using the following formulas:

$$\mathrm{L}_{KD}^{\mathrm{cont}}=-\frac{1}{N}\sum_{n=1}^{N}log\frac{s\left(\overline{\mathbf{C}},\mathbf{E}\right)}{\sum_{k=1}^{K}s\left(\overline{\mathbf{C}},\overline{\mathbf{E}}_{n,k}\right)+s\left(\overline{\mathbf{C}},\mathbf{E}\right)} \tag{1}$$

where $s\left(x,y\right)$ is defined as $\exp\left(\left\langle x,y\right\rangle/\tau\right)$. Here, $\left\langle x,y\right\rangle$ denotes the inner product of vectors x and y. N indicates the batch size for the n-th audio sample. τ and K denote the temperature parameter and the number of negative samples for contrastive loss, respectively. $\overline{\mathbf{E}}_{n,k}$ represents the k-th negative teacher token representation sampled from all teacher token representations in the current batch (excluding positive samples).

In addition to examining the contrastive distillation strategy, this paper also explores the application of a mean squared error loss function for sentence-level semantic knowledge distillation for comparative analysis. The loss function is computed using the following formulas:

$$\mathrm{L}_{KD}^{mse}=\alpha_{mse}\cdot\frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}\left(\hat{C}_d^n-E_d^n\right)^2 \tag{2}$$

where D represents the dimensionality of the semantic features, and the coefficient α_{mse} represents the scaling weight associated with the loss function.

V. Experiments And Results

We developed a series of end-to-end Jiao-Liao Mandarin speech recognition models and performed a comparative performance analysis. Initially, we presented the results obtained from the baseline Jiao-Liao Mandarin speech recognition model. Subsequently, we introduced an enhanced model based on CIF-FSA and detailed its findings. The experimental results indicate that our proposed approach significantly enhances model performance and improves speech recognition accuracy.

A. Experimental Setup

The experiments utilized the JLMS30 dataset developed in this study. The dataset was partitioned into training, dev, and test sets according to a time ratio of 8:1:1. Table I provides detailed information regarding the division of the dataset.

TABLE I
INFORMATION OF JLMS30 DATASET.

Dataset	Speakers	Male	Female	#Sentences	Duration(h)
Train	47	30	17	21,364	23.7
Dev	5	2	3	3,200	3.4
Test	5	1	4	2,658	2.9

In Jiao-Liao Mandarin speech recognition, we use CER to evaluate model accuracy, with a lower CER indicating better performance. The CER is computed using the following formulas:

$$\mathrm{CER} = \frac{(S + D + I)}{N} \tag{3}$$

where N represents the length of the original string, S represents the number of substituted characters, D represents the number of deleted characters, and I represents the number of inserted characters.

We employ 80-dimensional log Mel-filter bank features (Fbank) as input, utilizing a window size of 25 ms and a shift of 10 ms. Before training, two data augmentation techniques, speed perturbation [31] and SpecAugment [32], are applied to enhance the dataset. Global CMVN [33] is subsequently employed to normalize the features. The dataset yields a vocabulary comprising 3,380 characters along with four unique tokens: <PAD>, <EOS>, <BOS>, and <UNK>, which collectively form the unit set. Bert-base-Chinese 2 is the pre-trained model for the experiments, and the Transformer LM is used for re-scoring. All experiments are conducted on an NVIDIA Tesla 4080 (16GB) GPU. The hyperparameters of the CIF model are detailed in Table II.

TABLE II
THE CIF-BASED ASR MODEL SETTINGS.

Hyperparameter	Value
Encoder Layers	15
Decoder Layers	2
Encoder Embed Dim	256
Decoder Embed Dim	256
Heads	4
FFN Dim	2048
Dropout	0.1
Max Token	1500
Train Shuffle	False

B. Results

We first compare the CIF-based ASR model with previous studies. As shown in Table III, all models demonstrate robust performance, confirming the dataset's reliability and validity. Our proposed CIF-FSA model achieves optimal performance with and without the LM. U2++ [34], with its bidirectional attention decoder, improves the Branchformer model by leveraging context from both directions. The Sum. Mix. [15] approach enhances the Branchformer model by replacing the original attention mechanism with a linear one. The CIF+HKD model, which integrates hierarchical knowledge distillation, ranks just below CIF-FSA, significantly improving recognition. On the Aishell-1 dataset, CIF+HKD achieves optimal performance without additional data. However, on the JLMS30 dataset, our CIF-FSA model outperforms CIF+HKD by 2.2% in CER on the test set, highlighting the superiority of our approach. We experiment with three settings: SACD, FG, and FSA which combine SACD and FG, with results showing performance improvements. Substituting the convolution module with the FG-Conv module reduces model parameters by 6% while improving performance. On the JLMS30 dataset, the CER decreases by 5.4%, indicating enhanced feature representation. The semantic acoustic contrastive distillation strategy transfers semantic knowledge from sentence-level features of the pre-trained LM to the CIF-based ASR model, reducing the CER by 9.6% compared to the baseline. Our method achieves a 15.5% reduction in CER, demonstrating its effectiveness and superiority.

TABLE III
MAIN RESULTS ON JLMS30 (CER %).

Model	#Params	w/o LM	w/ LM
		dev/test	dev/test
Non-autoregressive			
Transformer	30M	26.6/26.5	26.3/26.2
Conformer	46M	23.2/24.4	23.0/24.2
Eff Conformer	46M	23.5/24.1	23.3/23.9
Branchformer	41M	26.8/25.3	26.6/25.1
U2++ Branchformer	48M	25.4/24.2	25.2/24.0
Branchformer+Sum. Mix.	48M	25.7/25.9	25.5/25.7
E-Branchformer	46M	23.2/24.0	23.0/23.8
Autoregressive			
CIF	47M	25.8/25.8	25.6/25.6
CIF+HKD	47M	22.7/22.3	22.5/22.1
Autoregressive (Proposed)			
CIF-SACD	47M	23.5/23.3	23.3/23.1
CIF-FG	44M	25.3/24.4	25.1/24.2
CIF-FSA	44M	**22.6/21.8**	**22.4/21.6**

We compare loss functions by exploring methods for transferring semantic knowledge from pre-trained language models to speech recognition systems through character-level and sentence-level knowledge distillation strategies. In this context, AD and SD denote character-level and sentence-level knowledge distillation, respectively, with KD Loss encompassing mean squared error (MSE), cosine (COS), and contrastive distillation (CONT) losses.

TABLE IV
COMPARISON BETWEEN CONTRASTIVE LOSS AND OTHER DISTILLATION LOSSES (CER %). AD REPRESENTS CHARACTER-LEVEL KNOWLEDGE DISTILLATION, AND SD REPRESENTS SENTENCE-LEVEL KNOWLEDGE DISTILLATION. MSE, COS, AND CONT REPRESENT MEAN SQUARE ERROR LOSS, COSINE EMBEDDING LOSS, AND CONTRASTIVE LOSS, RESPECTIVELY.

Model	AD	SD	KD Loss	w/o LM	w/ LM
				dev/test	dev/test
CIF	×	×	×	25.8/25.8	25.6/25.6
	√	×	MSE	23.9/23.4	23.7/23.2
	√	×	COS	23.7/23.5	23.5/23.3
	√	×	CONT	23.6/23.4	23.4/23.2
	×	√	MSE	24.7/25.3	24.5/25.1
	×	√	CONT	23.5/23.3	23.3/23.1
CIF-FG	×	×	×	25.3/24.4	25.1/24.2
	√	×	MSE	23.4/22.2	23.2/22.0
	√	×	COS	23.0/22.9	22.8/22.7
	√	×	CONT	23.3/23.0	23.1/22.8
	×	√	MSE	23.0/22.2	22.8/22.0
	×	√	CONT	**22.6/21.8**	**22.4/21.6**

TABLE V

The Comparison of Some Cases on JLSD30.

Ground Truth Transcription	Baseline Predicted Transcription	CIF-SACD Predicted Transcription
考 虑 到 目 前 的 实 际 情 况	考 虑 到 目 前 的 事 迹 情 况	考 虑 到 目 前 的 实 际 情 况
传 闻 铁 路 部 融 资 两 千 亿	传 闻 铁 路 部 融 资 两 千 一	传 闻 铁 路 部 融 资 两 千 亿
是 名 副 其 实 的 骑 游 天 下	是 名 符 其 实 的 肌 肉 天 下	是 名 副 其 实 的 骑 游 天 下
提 高 全 社 会 福 利 水 平	提 高 全 社 会 富 力 水 平	提 高 全 社 会 福 利 水 平

Table IV shows that character-level and sentence-level knowledge distillation strategies effectively enhance the model's language modeling capabilities and improve performance. Notably, the contrastive loss function outperforms the others, likely due to its fundamental principle of minimizing the distance between similar samples while maximizing the distance between dissimilar ones; this encourages the model to learn latent semantic information more comprehensively rather than focusing solely on individual sample features. Furthermore, the sentence-level knowledge distillation strategy performs better than its character-level counterpart, particularly when utilizing the FG-Conv module. This suggests that the sentence-level knowledge distillation strategy captures more crucial semantic information, enhancing the model's more profound understanding of overall sentence semantics.

To further explore how the semantic acoustic contrastive distillation strategy improves model performance, this chapter presents a case analysis using representative samples from the JLSD30 test set. As shown in Table V, the experimental comparison includes three sets of data: the first column displays the accurate speech transcription, the second column presents the baseline model's predictions, and the third column shows the predictions obtained using the CIF-SACD method. In the visual presentation, misrecognized words from the baseline model are highlighted in red, while the corrected predictions from the CIF-SACD method are highlighted in blue. A comparative analysis reveals that the SACD strategy significantly enhances the semantic representation capability of the baseline model, especially in disambiguating homophones. This finding confirms the critical role of SACD strategy in improving the model's semantic understanding.

We explore the performance of the FG-Conv module under varying the number of groups within the channel shuffle strategy through hyperparameter experiments. In the experiment, a group number of 1 signifies the absence of channel shuffle strategy.

TABLE VI
EFFECTS OF THE NUMBER OF GROUPS (CER %).

Model	G=1	G=2	G=4	G=8	G=16
CIF-FG	25.3	26.1	25.1	24.4	25.4
CIF-FSA	21.9	23.1	22.0	**21.8**	**21.8**

As illustrated in Table VI, when the channel shuffle strategy is not employed, localized processing facilitates the recognition of subtle differences between spatial locations, enhancing feature expressiveness. However, each group performs convolution operations exclusively on its corresponding input group, resulting in relatively independent extracted features; consequently, the model's performance on the test set is only 1.9% lower than that of the baseline. When the channel shuffle strategy is implemented, particularly with the group number set to 8, the model demonstrates optimal performance on the test set, achieving a relative reduction of 3.5% in CER compared to the model without channel shuffle. This enhancement arises from the exchange of intergroup information following the grouped convolution layers, which mitigates high correlation among neighboring channel features and produces richer output features that more effectively represent acoustic characteristics. However, as the number of groups increases, model performance deteriorates, likely due to excessive grouping, which introduces disorder among acoustic features and hampers the model's ability to capture valuable information.

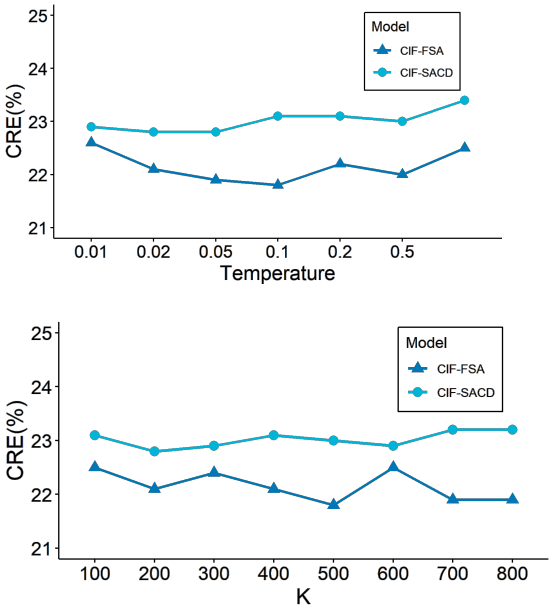

Fig. 8. Effects of the temperature and the number of negative samples.

We conduct extensive comparative experiments to assess the effect of temperature and the number of negative samples (K) on the semantic acoustic contrastive distillation strategy. The examined temperature values range from {0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1}, while K varies from {100, 200, 300, 400, 500, 600, 700, 800}.

The experimental results in Fig. 8 indicate that combining the CIF model with the semantic acoustic contrastive distillation strategy achieves optimal performance at a moderate temperature. Specifically, as the temperature increases, the CER shows a trend of first decreasing and then increasing. The CER exhibits minimal fluctuations with changes in K, indicating that K has a relatively limited impact on the model performance for this low-resource dataset. Furthermore, the incorporation of the FG-Conv module

generally leads to a reduction in CER, thereby confirming its effectiveness.

We explore the generality of the proposed method when using different pre-trained language models for semantic acoustic knowledge distillation. This set of experiments continues to use semantic acoustic knowledge distillation hyperparameters from the previous best results.

TABLE VII
EFFECTS OF DIFFERENT PLMS ON THE ASR PERFORMANCE.

Model	w/o LM	w/ LM
	dev/test	dev/test
CIF	25.8/25.8	25.6/25.6
+ bert-base-chinese	23.5/23.3	23.3/23.1
+ chinese-bert-wwm [35]	22.6/22.2	**22.4/22.1**
+ chinese-bert-wwm-ext [35]	22.8/22.5	22.6/22.3
+ chinese-roberta-wwm-ext [35]	23.3/23.2	23.1/23.0
+ chinese-lert-base [36]	22.8/22.3	22.6/22.1
+ albert-base-chinese	22.9/22.7	22.6/22.5
+ distilbert-base-zh-cased [37]	23.2/23.0	23.0/22.8
+ t5-base-chinese-cluecorpussmall [38]	23.1/22.7	22.9/22.6
+ chinese-macbert-base [35]	23.4/22.9	23.3/22.7

As shown in Table VII, our proposed semantic acoustic knowledge distillation method consistently achieves improvements when utilizing different pre-trained language models as the knowledge source for cross-modal distillation, demonstrating the approach's effectiveness and generalizability.

VI. Conclusion

We have developed the JLMS30 dataset, which comprises 30 hours of diverse Jiao-Liao Mandarin speech recognition data. This dataset encompasses a variety of general themes, including finance, science and technology, sports, entertainment, news, rich local characteristics, and folk proverbs. We designed a lightweight FG-Conv module to enhance feature expressiveness by capturing relationships across channels and spatial locations. Additionally, the channel shuffle strategy can systematically or randomly rearrange neurons across various channels or spatial positions to assess their importance and optimize features and their corresponding weights. Furthermore, we propose a semantic acoustic contrastive distillation strategy that aligns sentence-level semantic information from a pre-trained language model with high-level acoustic features. This approach narrows the semantic gap between acoustic and textual representations, thereby enhancing the model's language modeling capability and aiding in alleviating issues related to homophony.

In future work, we will continue to expand the data resources for Jiao-Liao Mandarin and explore additional speech processing tasks, including but not limited to speech recognition, speaker recognition, voice conversion, and speech synthesis. Simultaneously, we will investigate novel approaches to integrate the semantic knowledge of pre-trained language models into speech recognition systems using cross-modal knowledge distillation methods, thereby further enhancing the overall performance of these models.

Acknowledgment

The author would like to express sincere gratitude to Dr. Han Minglun for his invaluable guidance and assistance throughout this study.

References

[1] L. Dong, D. Qin, F. Bai, F. Song, Y. Liu, C. Xu, and Z. Ou, "Low-resourced speech recognition for iu mien language via weakly-supervised phoneme-based multilingual pre-training", arXiv preprint arXiv:2407.13292, 2024.

[2] L. Lonergan, M. Qian, N. N. Chiar´ain, C. Gobl, and A. N. Chasaide, "Towards dialect-inclusive recognition in a low-resource language: are balanced corpora the answer?", arXiv preprint arXiv:2307.07295, 2023.

[3] S. Kong, C. Li, C. Fang, and P. Yang, "Building a speech dataset and recognition model for the minority tu language", Applied Sciences, vol. 14, no. 15, pp. 6795, 2024.

[4] Y. Yang, H. Xu, H. Huang, E. S. Chng, and S. Li, "Speech-text based multi-modal training with bidirectional attention for improved speech recognition", ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.

[5] K. Bhogale, A. Raman, T. Javed, S. Doddapaneni, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages", ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.

[6] Z. Hu, Shanshan Tang, Y. Luo, F. Jian, and X. Si, "3DACRNN Model Based on Residual Network for Speech Emotion Classification", Engineering Letters, vol. 29, no. 2, pp. 400-407, 2021.

[7] Z. Hu, L. Wang, Y. Luo, Y. Xia, and H. Xiao, "Speech Emotion Recognition Model Based on Attention CNN Bi-GRU Fusing Visual Information", Engineering Letters, vol. 30, no. 2, pp. 427-434, 2022.

[8] K. Zheng, Z. Xia, Y. Zhang, X. Xu, and Y. Fu. "Speech Emotion Recognition based on Multi-Level Residual Convolutional Neural Networks", Engineering Letters, vol. 28, no. 2, pp. 559-565, 2020.

[9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition", arXiv preprint arXiv:2005.08100, 2020.

[10] L. Dong and B. Xu, "Cif: Continuous integrate-and-fire for end-to-end speech recognition", ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6079–6083.

[11] S. Li, M. Xu, and X.-L. Zhang, "Efficient conformer-based speech recognition with linear attention", in 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2021, pp. 448–453.

[12] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding", International Conference on Machine Learning. PMLR, 2022, pp. 17627–17643.

[13] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition", 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023, pp. 84–91.

[14] F. Mai, J. Zuluaga-Gomez, T. Parcollet, and P. Motlicek, "Hyperconformer: Multi-head hypermixer for efficient speech recognition", arXiv preprint arXiv:2305.18281, 2023.

[15] T. Parcollet, R. van Dalen, S. Zhang, and S. Bhattacharya, "Summarymixing: A linear-complexity alternative to self-attention for speech recognition and understanding", arXiv preprint arXiv:2307.07421.

[16] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.

[17] Tang Z., Wang D., Xu Y., Sun J., Lei X., Zhao S., Wen C., Tan X., Xie C., Zhou S. and Yan R., "Kespeech: An open source speech dataset of mandarin and its eight subdialects", Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[18] Q. Shao, P. Guo, J. Yan, P. Hu, and L. Xie, "Decoupling and interacting multi-task learning network for joint speech and accent recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 459–470, 2023.

[19] B. Mu, Y. Li, Q. Shao, K. Wei, X. Wan, N. Zheng, H. Zhou, and L. Xie, "Mmger: Multi-modal and multi-granularity generative error correction with llm for joint accent and speech recognition", arXiv preprint arXiv:2405.03152, 2024.
[20] Z. Tang, D. Wang, S. Huang, and S. Shang, "Pinyin regularization in error correction for chinese speech recognition with large language models", arXiv preprint arXiv:2407.01909, 2024.
[21] J. Chen, J. Fang, Y. Zheng, Y. Wang, and H. Fei, "Qifusion-net: Layer-adapted stream/non-stream model for end-to-end multi-accent speech recognition", arXiv preprint arXiv:2407.03026, 2024.
[22] Y. Gu, Z. Du, S. Zhang, Q. Chen, and J. Han, "Personality-aware training based speaker adaptation for end-to-end speech recognition", in Interspeech, vol. 2023, 2023, pp. 1249–1253.
[23] M. Han, F. Chen, J. Shi, S. Xu, and B. Xu, "Knowledge transfer from pre-trained language models to cif-based speech recognizers via hierarchical distillation", arXiv preprint arXiv:2301.13003, 2023.
[24] X. Lu, P. Shen, Y. Tsao, and H. Kawai, "Hierarchical cross-modality knowledge transfer with sinkhorn attention for ctc-based asr", ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 13116–13120.
[25] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Distilling the knowledge of bert for sequence-to-sequence asr", arXiv preprint arXiv:2008.03822, 2020.
[26] Y. Kubo, S. Karita, and M. Bacchiani, "Knowledge transfer from large-scale pretrained language models to end-to-end speech recognizers", ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 8512–8516.
[27] Li X., Wang Y., Liu X., Su K., Li Z., Wang Y., Jiang B., Xie K. and Liu J., "JLMS25 and Jiao-Liao Mandarin Speech Recognition Based on Multi-Dialect Knowledge Transfer.", Applied Sciences, vol. 15, no. 3, 2025.
[28] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline", in 2017 20th Conference of the oriental chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). IEEE, 2017, pp. 1–5.
[29] M. Han, L. Dong, S. Zhou, and B. Xu, "Cif-based collaborative decoding for end-to-end contextual speech recognition", ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6528–6532.
[30] M. Han, L. Dong, Z. Liang, M. Cai, S. Zhou, Z. Ma, and B. Xu, "Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection", ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 8532–8536.
[31] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition.", in Interspeech, vol. 2015, 2015, pp. 3586.
[32] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition", arXiv preprint arXiv:1904.08779, 2019.
[33] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition, Speech Communication", vol. 25, no. 1-3, pp. 133–147, 1998.
[34] D. Wu, B. Zhang, C. Yang, Z. Peng, W. Xia, X. Chen, and X. Lei, "U2++: Unified two-pass bidirectional end-to-end model for speech recognition", arXiv preprint arXiv:2106.05642, 2021.
[35] Cui Y, Che W, Liu T, Qin B, Wang S, Hu G, "Revisiting pre-trained models for Chinese natural language processing", arXiv preprint arXiv:2004.13922, 2020.
[36] Cui Y, Che W, Wang S, Liu T, "Lert: A linguistically-motivated pre-trained language model", arXiv preprint arXiv:2211.05344, 2022.
[37] Abdaoui A, Pradel C, Sigel G, "Load what you need: Smaller versions of multilingual bert", arXiv preprint arXiv:2010.05609, 2020.
[38] Zhao Z, Chen H, Zhang J, Zhao X, Liu T, Lu W, Chen X, Deng H, Ju Q, Du X, "Uer: An open-source toolkit for pre-training models", arXiv preprint arXiv:1909.05658, 2019.