Multi-Level Cross-Modal Alignment for Speech Relation Extraction

Liang Zhang[1,2*] **Zhen Yang**[3*] **Biao Fu**[1] **Ziyao Lu**[3] **Liangying Shao**[1] **Shiyu Liu**[1]
Fandong Meng[3] **Jie Zhou**[3] **Xiaoli Wang**[1] and **Jinsong Su**[1,2†]
[1]School of Informatics, Xiamen University, China
[2]Xiamen Key Laboratory of Intelligent Storage and Computing,
School of Informatics, Xiamen University
[3]Pattern Recognition Center, WeChat AI, Tencent Inc, China
lzhang@stu.xmu.edu.cn, yz_tencent_person@163.com, jssu@xmu.edu.cn

Abstract

Speech Relation Extraction (SpeechRE) aims to extract relation triplets from speech data. However, existing studies usually use synthetic speech to train and evaluate SpeechRE models, hindering the further development of SpeechRE due to the disparity between synthetic and real speech. Meanwhile, the modality gap issue, unexplored in SpeechRE, limits the performance of existing models. In this paper, we construct two real SpeechRE datasets to facilitate subsequent researches and propose a Multi-level Cross-modal Alignment Model (MCAM) for SpeechRE. Our model consists of three components: 1) a speech encoder, extracting speech features from the input speech; 2) an alignment adapter, mapping these speech features into a suitable semantic space for the text decoder; and 3) a text decoder, autoregressively generating relation triplets based on the speech features. During training, we first additionally introduce a text encoder to serve as a semantic bridge between the speech encoder and the text decoder, and then train the alignment adapter to align the output features of speech and text encoders at multiple levels. In this way, we can effectively train the alignment adapter to bridge the modality gap between the speech encoder and the text decoder. Experimental results and in-depth analysis on our datasets strongly demonstrate the efficacy of our method. Our source code is available at https://github.com/DeepLearnXMU/SpeechRE-MCAM.

1 Introduction

Relation Extraction (RE) aims to extract structured knowledge from unstructured data in the form of relation triplet. Since such structured knowledge can benefit various downstream applications, many efforts have been devoted to this task. However, most studies (Wang et al., 2020; Eberts and Ulges, 2020; Cabot et al., 2021) in this regard focus on extracting relation triplets from plain text (*i.e.* TextRE), which severely limits the application scope of RE. In addition to text, a large amount of speech is continuously produced in our daily lives, including news reports, meetings, and dialogues, *etc.* Similar to text, these speeches often contain rich and valuable structured knowledge that can not only enrich existing knowledge graphs but also benefit various speech-related tasks. Therefore, how to effectively extract relation triples from speech is an crucial research topic, yet it remains under-explored.

Wu et al. (2022) are the first to explore the Speech Relation Extraction (SpeechRE) task. They first construct two benchmark datasets for this task by converting the input text from TextRE datasets into speech using a text-to-speech (TTS) system. However, this synthetic speech ususlly fails to accurately evaluate the model's performance in real-world scenarios. Meanwhile, due to the limited performance of the TTS system, the synthesized speech usually contains much noise, especially for longer input text. It leads to the SpeechRE model, trained on such synthetic speech, often demonstrating a subpar performance on real speech.

Meanwhile, Wu et al. (2022) propose a baseline model for SpeechRE, where a CNN-based length adapter is used to connect a speech encoder with a text decoder (See Figure 1(a)). Since the speech encoder and text decoder are pre-trained on corpora of different modalities, a significant modality gap exists between them, thus limiting the performance of this model. Moreover, SpeechRE models are usually required to comprehensively understand the input speech at multiple levels to effectively extract relation triplets from it. Particularly, the SpeechRE model first recognizes entities in the input speech based on its token/entity-level information, and then predict the relations between these entities according to its overall semantics at the sentenc

This work is done when Liang Zhang was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
*Equal contribution
†Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11975–11986
November 12-16, 2024

level. Thus, it is essential to bridge the modality gap between the speech encoder and text decoder at multiple levels for better SpeechRE model.

Although the modality gap issue has been widely studied in other speech-related tasks (Xu et al., 2021; Han et al., 2021; Ye et al., 2022), it has not yet been explored in the context of SpeechRE. In this paper, we first conduct a preliminary study to investigate the efficacy of advanced cross-modal alignment methods in other tasks for SpeechRE. We observe that token-level and sentence-level alignment methods can enhance the entity recognition and relation extraction performance of the SpeechRE model, respectively. Meanwhile, we notice some obvious problems with these methods: 1) the CTC-based token-level alignment method tends to overfit high-frequency tokens, leading to the collapse of speech features. 2) the compression-then-alignment-based sentence-level alignment method loses considerable fine-grained information in the input speech, significantly limiting the model's entity recognition capability.

Based on the above analysis, we propose a Multi-level Cross-modal Alignment Model (MCAM) for SpeechRE, which consists of a speech encoder, an alignment adapter, and a text decoder. We first use the speech encoder to extract speech features from the input speech. Then, we employ the alignment adapter to map the speech features into a suitable semantic space for the text decoder. Finally, the text decoder is utilized to autoregressively generate relation triplets based on these speech features.

The alignment adapter is designed to bridge the modality gap between our speech encoder and text decoder from multiple levels. To do this, during training, we introduce an additional text encoder to extract the text features from the input text, which will be removed during inference. Then, we train the alignment adapter to align the feature sequences produced by the speech and text encoders at three levels: 1) **Token-level Alignment**. We first concatenate the text features in the current batch to create a token feature matrix, and then use it to calculate alignment scores from speech features to token features. Lastly, we compute a CTC loss based on these scores and achieve the token-level alignment by minimizing the loss. Since a token has distinct features in various contexts, our method can effectively avoid the overfitting issue for high-frequency tokens. 2) **Entity-level Alignment**. Here, we construct a mixed feature sequence by replacing the text features of entities in the text

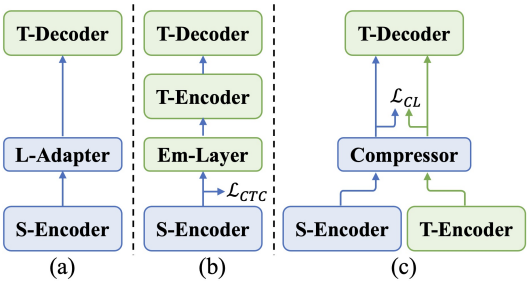

Figure 1: Illustration of the baseline (Wu et al., 2022) and its variants, where S = **S**peech, T = **T**ext, L= **L**ength, and Em = **Em**bedding.

feature sequence with their corresponding speech features. Then, we minimize the KL divergence between the output distributions generated by the text decoder based on the original text feature sequence and the mixed one. 3) **Sentence-level Alignment**. We first separately compress the speech and text feature sequences into R global feature vectors using R relation-specific learnable vectors, serving as their soft prompts. Then, we use a contrastive loss to align the final representations of soft prompts from different modalities. Through the soft prompt strategy, we can prevent the loss of fine-grained information during the compression process.

To evaluate the performance of SpeechRE models in real-world scenarios and facilitate future studies, we annotate two real SpeechRE datasets: CoNLL04 and ReTACRED. Experimental results on the two datasets show that our model consistently outperforms all baselines. Extensive ablation studies further demonstrate the effectiveness of various components in our model. Notably, we conduct extensive analysis experiments on our dataset with the aim of inspiring subsequent research.

2 Preliminary Study

In this section, we investigate the effectiveness of advanced cross-modal alignment methods from other speech-related tasks for SpeechRE. Here, we consider two alignment methods: CTC-based token-level alignment (See Figure 1(b)) (Xu et al., 2021; Zhang et al., 2023d) and compression-then-alignment-based sentence-level alignment (See Figure 1(c)) (Han et al., 2021; Wang et al., 2022). For the former, a CTC loss is employed to monotonically align speech features with the embedding vectors of their corresponding tokens. Meanwhile, these speech features are projected into the embedding space of the text encoder and serve as its input. For the latter, an attention-based semantic compressor is utilized to compress the feature se-

Model	CONLL04		
	ER	RP	RTE
LNA-ED (Wu et al., 2022)	18.87	55.66	10.41
+ *Token-level alignment*	21.53	55.90	11.38
+ *Sentence-level alignment*	12.75	58.44	8.08

Table 1: Performance of the baseline (LNA-ED) and its variations on the CONLL04 test set.

Datasets		CoNLL04	ReTACRED
#Relation		5	40
#Instance	train	922	33,477
	dev	231	9,350
	test	288	5,805
#Triplet	train	1,283	58,465
	dev	343	19,584
	test	422	13,418
#Speech AvgLen.		17.5	20.1
#Speaker		4	8

Table 2: Dataset statistics, where Speech AvgLen. = Speech Average Length (in seconds).

quences generated by the speech and text encoders into K global feature vectors, separately. Then, a contrastive loss is used to align the global feature vectors of different modalities and input them into the text decoder to generate relation triples.

In Table 1, we present the performance of these methods on the CoNLL04 test set, revealing several intriguing phenomena: (1) The token-level alignment enhances the model's entity extraction capability, but has a limited effect on its relation extraction performance. As shown in Figure 2, we analyse tokens generated by CTC greedy decoding (the nearest tokens to speech features). We find that the CTC loss tends to overfit high-frequency tokens, leading to the collapse of speech features and limiting the efficacy of this token-level alignment. (2) The sentence-level alignment effectively improves the model's performance in relation extraction, while significantly degrading its entity extraction capability. The primary reason is a significant loss of fine-grained information during compression, which is crucial for entity recognition.

These results suggest that naively applying cross-modal alignment methods from other tasks to SpeechRE is suboptimal. Therefore, we customize a more effective cross-modal alignment method for SpeechRE, which avoids the defects of the above methods and inherits their advantages.

3 SpeechRE Datasets

In this section, we provide a detailed description to the construction process of our SpeechRE datasets.

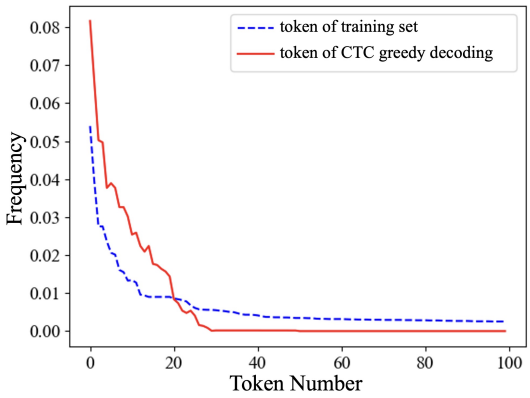

Figure 2: Token frequencies in the CONLL04 training set and those generated by CTC greedy decoding on the CONLL04 test set. Here, we sort tokens based on their frequencies in the CONLL04 training set and only report the top 100 most frequent tokens.

Given a TextRE instance (x, y), where x and y respectively denote the input text and output relation triplets, we create a SpeechRE instance (s, x, y) by converting x into its corresponding speech s in a human-read manner. Specifically, we invite 12 native English speakers to transform the input texts in the CoNLL04 (Roth and Yih, 2004) and ReTACRED (Stoica et al., 2021) datasets into fluent and clear speeches, thus obtaining two benchmark datasets for SpeechRE. We totally annotate 7,246 instances over one month and each speaker is responsible for annotating 242 instances per day (about 2 hours) to prevent fatigue. The statistics of these two datasets are presented in Table 2.

To ensure the quality of speeches, we employ the Whisper (Radford et al., 2023) model to transcribe each speech into text, and then compare the resulting text with its true input text. Meanwhile, we hire a professional English speaker to sample and check the annotated instances for further quality control. Notably, for the ReTACRED dataset, we only annotate its test set while retaining the training and validation sets synthesized by Wu et al. (2022). Using this dataset, we can effectively evaluate the model's robustness to the noise in synthetic data.

4 Our Model

In this section, we give a detailed description to the proposed MCAM. In terms of architecture, our model (MCAM) includes three modules: a speech encoder, an alignment adapter, and a text decoder. We first use the speech encoder to extract speech features from the input speech. Then, we employ the alignment adapter to map these speech features into an suitable semantic space, rendering

them more compatible with the text decoder. Finally, the text decoder autoregressively generates linearized relation triplets based on the speech features. Notably, the above model inference process corresponds to the blue arrows in Figure 3.

During training, we introduce a text encoder to serve as a semantic bridge between the speech encoder and text decoder, which is removed during model inference. Then, the alignment adapter is trained to align the output features of the speech and text encoders at three levels. In this way, the trained alignment adapter can effectively bridge the modality gap between the speech encoder and text decoder. In the following sections, we first individually detail each component involved in our model training (See Sections 3.1−3.3). Additionally, we introduce our model training in Section 3.4.

4.1 Speech Encoder & Text Encoder

Following Wu et al. (2022), we use the pre-trained wav2vec2.0 (Baevski et al., 2020) as our speech encoder to extract speech features $\mathbf{H}_s \in \mathbb{R}^{l_s \times d}$ from the input speech s, where l_s and d denote the number and dimension of speech features, respectively. Meanwhile, we adopt the BART encoder (Lewis et al., 2019) as our text encoder to generate text features $\mathbf{H}_t \in \mathbb{R}^{l_t \times d}$ for the input text x, where l_t represents the number of tokens in the input text.

4.2 Alignment Adapter

Our alignment adapter is designed to map speech features produced by the speech encoder to a suitable semantic space for the text decoder, effectively alleviating the modality gap issue between them. To do this, we train the alignment adapter to align the feature sequences ($\mathbf{H}_s$ and $\mathbf{H}_t$) generated by the speech and text encoders at three levels:

Token-Level Alignment. Here, we aim to monotonically align the speech and text features at the token level. Considering that the length l_s of the speech feature $\mathbf{H}_s$ is often relatively long, we employ two 1D convolution layers with a stride of 2 to shrink its length by a factor of 4, thus obtaining the new speech features: $\mathbf{H}_s = \mathbf{CNN}(\mathbf{H}_s)$. Meanwhile, we concatenate the text features in the current batch b to construct a token feature matrix $\mathbf{W} = [\mathbf{H}_t^1; \cdots; \mathbf{H}_t^{|b|}]$, where each element corresponds to a token feature. Next, we compute the alignment scores $\mathbf{A}$ from speech features to token features: $\mathbf{A} = \mathbf{H}_s \mathbf{W}^\top$, and calculate a CTC loss $\mathcal{L}_{\text{CTC}}$ based on $\mathbf{A}$. Finally, we achieve the token-level alignment by minimizing $\mathcal{L}_{\text{CTC}}$. In contrast, previous CTC-based token-level alignment methods (Xu et al., 2021; Zhang et al., 2023d) often use the static token embedding matrix as $\mathbf{W}$ to compute alignment scores $\mathbf{A}$, making them prone to overfitting high-frequency tokens. Since a token has distinct features in different contexts, our method exploits such dynamic feature to effectively avoid above overfitting issue. Meanwhile, our method inherits the advantages of contrastive learning by using our token feature matrix $\mathbf{W}$.

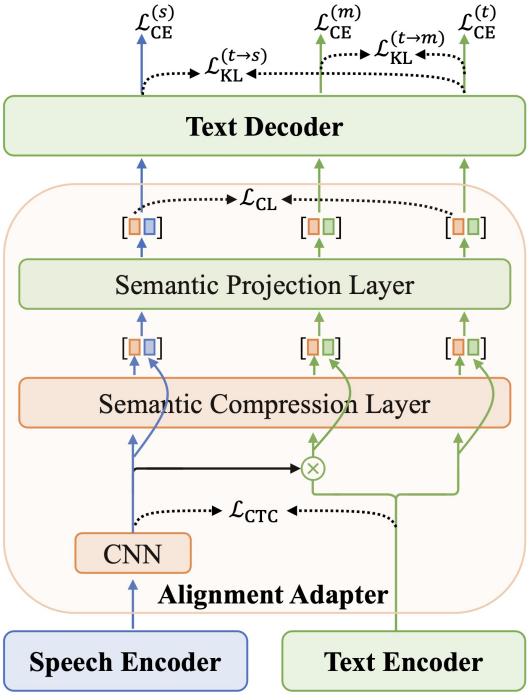

Figure 3: The overall architecture of our model. Here the blue arrows show the inference process of our model.

Entity-Level Alignment. It is designed to enhance cross-modal alignment for tokens within entities, thus improving the accuracy of our model in generating entity. To effectively achieve this, we first need to obtain the speech features of entity tokens, which poses a challenge. Previous studies usually use external tools (Fang et al., 2022) or Optimal Transport (OT) (Zhou et al., 2023) to determine the speech features corresponding to each token from the speech feature sequence ($\mathbf{H}_s$). However, these methods often suffer from issues such as error propagation, high complexity, and instability. Here, we propose a simple and effective method to attain the above goal. Considering the monotonicity and locality of the alignment from the text sequence to the speech sequence, we use a window-based attention mechanism to obtain the

speech feature $\mathbf{h}_i^{(s)}$ of each entity token:

$$\mathbf{h}_i^{(s)} = \text{softmax}\left(\mathbf{H}_t[i]\mathbf{H}_s^{\mathsf{T}}[s:e]\right)\mathbf{H}_s[s:e], \quad (1)$$

where i is the index of the current entity token in the input text ($\mathbf{H}_t$), $[s{=}i{*}(\frac{l_s}{l_t}){-}w : e{=}i{*}(\frac{l_s}{l_t}){+}w]$ refers to the local window corresponding to the entity token in the input speech ($\mathbf{H}_s$), the window size w is set as $2{*}\frac{l_s}{l_t}$ on both the left and right sides.

Then, we replace the text features of entity tokens in $\mathbf{H}_t$ with their speech features to obtain a mixed feature sequence $\mathbf{H}_m$. Lastly, we calculate a KL divergence loss $\mathcal{L}_{\text{KL}}^{(t\to m)}$ between the output distributions generated by the text decoder for $\mathbf{H}_t$ and $\mathbf{H}_m$ (See Equation 5). By minmizing $\mathcal{L}_{\text{KL}}^{(t\to m)}$, we can accomplish the entity-level alignment.

Sentence-Level Alignment. To this end, we first introduce a semantic compression layer that consists of a single attention layer and R relation-specific learnable query vectors $\mathbf{Q}{\in}\mathbb{R}^{R\times d}$. Its objective is to compress the speech features $\mathbf{H}_s$ and text features $\mathbf{H}_t$ into R global feature vectors:

$$\mathbf{G}_{t/s} = \text{Attention}\left(\mathbf{Q}, \mathbf{H}_{t/s}, \mathbf{H}_{t/s}\right) \in \mathbb{R}^{R\times d}. \quad (2)$$

However, this compression process often accompanies a significant loss of fine-grained information, as demonstrated in our preliminary study. To avoid this issue, we regard $\mathbf{G}_s$ and $\mathbf{G}_t$ as soft prompts of $\mathbf{H}_s$ and $\mathbf{H}_t$, denoted as $\widehat{\mathbf{H}}_s{=}[\mathbf{G}_s;\mathbf{H}_s]$ and $\widehat{\mathbf{H}}_t{=}[\mathbf{G}_t;\mathbf{H}_t]$. Following this, we use a semantic projection layer to map $\widehat{\mathbf{H}}_t$ and $\widehat{\mathbf{H}}_s$ into a shared semantic space, generating new speech and text features $\widetilde{\mathbf{H}}_s$ and $\widetilde{\mathbf{H}}_t$. Lastly, we perform the sentence-level alignment from R different (relation-specific) perspectives by minimizing the contrastive loss $\mathcal{L}_{CL}$ between the above soft prompts:

$$\mathcal{L}_{CL} = -\sum_{i=1}^{R}\log\frac{e^{\cos(\widetilde{\mathbf{H}}_t[i],\widetilde{\mathbf{H}}_s[i])/\tau}}{\sum_{j=1}^{R}e^{\cos(\widetilde{\mathbf{H}}_t[i],\widetilde{\mathbf{H}}_s[j])/\tau}}, \quad (3)$$

where the first R elements of $\widetilde{\mathbf{H}}_t$ and $\widetilde{\mathbf{H}}_s$ refer to the final features of their respective soft prompts, and temperature τ is set as 0.1 empirically. Intuitively, the relation binary classification loss can be applied to these soft prompts to further improve the model performance. However, this improvement is marginal in our experiments, so we did not do this.

Notably, we utilize the top N layers of the BART encoder as our semantic projection layer, and its remaining layers as our text encoder. In this way, we can further alleviate the above modality gap issue, while preventing the introduction of new parameters that could disrupt the compatibility between the text encoder and decoder.

4.3 Text Decoder

Following prior studies (Cabot et al., 2021; Wu et al., 2022), we treat SpeechRE as a sequence generation task and utilize the BART decoder as our text decoder. The text decoder focuses on autoregressively generating linearized relation triplets based on the speech or text features.

Back to Figure 3, through the alignment adapter, we obtain three feature sequences: $\widetilde{\mathbf{H}}_s$, $\widetilde{\mathbf{H}}_t$, and $\widetilde{\mathbf{H}}_m$. Subsequently, we feed them into the text decoder to derive their respective output distributions: $p(y|\widetilde{\mathbf{H}}_s)$, $p(y|\widetilde{\mathbf{H}}_t)$ and $p(y|\widetilde{\mathbf{H}}_m)$. Finally, we employ cross-entropy as our task loss:

$$\begin{aligned}\mathcal{L}_{\text{CE}} &= \mathcal{L}_{\text{CE}}^{(t)} + \mathcal{L}_{\text{CE}}^{(m)} + \mathcal{L}_{\text{CE}}^{(s)},\\ \mathcal{L}_{\text{CE}}^{(t/m/s)} &= -\sum_{i=1}^{|y|}\log p(y_i|y_{<i}, \widetilde{\mathbf{H}}_{t/m/s}).\end{aligned} \quad (4)$$

Moreover, we introduce two KL divergences loss for our model training:

$$\begin{aligned}\mathcal{L}_{\text{KL}} &= \mathcal{L}_{\text{KL}}^{(t\to m)} + \mathcal{L}_{\text{KL}}^{(t\to s)},\\ \mathcal{L}_{\text{KL}}^{(t\to m/s)} &= \text{KL}(p(y|\widetilde{\mathbf{H}}_{m/s}) \mid \text{sg}(p(y|\widetilde{\mathbf{H}}_t))),\end{aligned} \quad (5)$$

where $\text{sg}(\cdot)$ is the stop-gradient operator, $\mathcal{L}_{\text{KL}}^{(t\to m)}$ is used to achieve our entity-level alignment, and $\mathcal{L}_{\text{KL}}^{(t\to s)}$ represents the knowledge distillation loss from TexRE ($p(y|\widetilde{\mathbf{H}}_t)$) to SpeechRE ($p(y|\widetilde{\mathbf{H}}_s)$).

4.4 Model Training

To effectively train our model, we use two hyper-parameters (α and β) to balance the above losses, deriving the final training objective of our model:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{KL}} + \alpha\mathcal{L}_{\text{CL}} + \beta\mathcal{L}_{\text{CTC}}. \quad (6)$$

Since our model is insensitive to the weights of $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{KL}}$, we simply set their weights as 1.

To alleviate the data scarcity issue in SpeechRE, previous study (Wu et al., 2022) resorts to pseudo-labeling methods, where TextRE models are used to generate relation triples for the transcript text in ASR dataset, thus creating SpeechRE instances. However, the issue of error propagation limits the effectiveness of such methods. Hence, we directly utilize the ASR task to pre-train our model on the

ASR dataset, while incorporating our multi-level alignment method into this process. To do this, we convert ASR instances (s, x) into SpeechRE instances $(s, x, y{=}x)$ by treating its transcript text x as both the input and output text. Considering the lack of entity annotations in the transcript text x, we implement the entity-level alignment on a randomly-selected 30% of the tokens within x. Through the pre-training stage, we effectively improve the accuracy of our model in recognizing entities due to the correlation between the ASR and entity recognition task. Then, we fine-tune our model on our SpeechRE datasets.

To efficiently train our model without compromising performance, we follow Li et al. (2021) to freeze the FFN layers in the speech encoder, text encoder and text decoder during the above pre-training and fine-tuning stages. Meanwhile, we freeze the entire text encoder and decoder during the pre-training stage.

5 Experiment

5.1 Datasets & Evaluation Metrics

We conduct experiments on our two SpeechRE datasets (CoNLL04 and ReTACRED) and a Mixed-CoNLL04 dataset whose training and alidation sets are synthetic and test set is real (as in ReTACRED). To pre-train our model, we use the English portion of the MuST-C v2.0 (Di Gangi et al., 2019) En-Zh corpus as our pre-training ASR dataset. We adopt the micro-F1 score as a metric to assess the performance of models in entity recognition, relation prediction, and relation triplet extraction. For any entity, relation, or triplet to be considered correct, they must exactly match their counterpart in tags.

5.2 Settings

We implement our model based on Fairseq (Ott et al., 2019) and PyTorch (Paszke et al., 2019). To optimize our model, we use the Adam (Kingma and Ba, 2015) optimizer with parameters (0.99, 0.98), while setting the clip norm to 10. We pre-train our model for 32K update steps on the ASR dataset and fine-tune it for 16K update steps on SpeechRE datasets. Meanwhile, we set the early stopping to 20 update steps during the fine-tuning stage and apply a learning rate of 1e-4 for both training stages, monitored by a tri-stage scheduler. We select the best checkpoint on the validation set for testing. All experiments are conducted on 4 NVIDIA Tesla-V100 GPUs.

$\alpha\backslash\beta$	0.1	0.2	0.3	0.4	0.5
0.2	21.67	22.06	20.95	20.75	20.21
0.4	22.01	22.57	21.24	21.05	20.51
0.6	22.35	22.81	21.79	21.42	20.70
0.8	22.67	**23.20**	22.11	21.74	21.03
1.0	22.43	22.02	21.76	21.55	21.19

Table 3: The performance of our model with different values of α and β in relation triplet extraction on the CONLL04 validation set.

5.3 Baselines

We compare our model with the following three categories of competitive baselines:

TextRE Models. These models aim to jointly extract entities and relations from the input text. Following Wu et al. (2022), we consider TP-Linker (Wang et al., 2020) , Spert (Eberts and Ulges, 2020), and REBEL (Cabot et al., 2021) for comparison.

Pipeline SpeechRE Models. These models first use an ASR module to transcribe the input speech into text, and then feed the resulting text into a TextRE module for extracting relation triplets. Wu et al. (2022) employ the pre-trained wav2vec-large as the ASR module and the above three TextRE models as the TextRE module to construct three pipeline models for SpeechRE: TP-Linker$_{pipe}$, Spert$_{pipe}$, REBEL$_{pipe}$. Moreover, we further fine-tune the ASR module of these models on SpeechRE datasets to improve their performance. These pipeline models possess the same speech encoder and text decoder as our model, ensuring a fairer comparison.

End2End SpeechRE Models. These models are designed to directly extract relation triplets from the input speech, and our model also falls into this category. In this regard, the sole existing work is LNA-ED (Wu et al., 2022), which also serves as our base model. It uses a simple CNN-based length adapter to connect a pre-trained speech encoder (wav2vec2.0) with a pre-trained text decoder (BART decoder). Moreover, we compare our model with some representative cross-modal alignment models in other speech-related tasks: SATE (Xu et al., 2021), Chimera (Han et al., 2021), MSP-ST (Zhang et al., 2023d), and CMOT (Zhou et al., 2023), all of which usually align the speech and text features at the token or sentence level.

5.4 Hyper-parameter Settings

Effect of Hyper-parameters α and β The hyper-parameters α and β in Equation 6 play a crucial

	Model	CONLL04			ReTACRED			Mixed-CoNLL04		
		ER	RP	RTE	ER	RP	RTE	ER	RP	RTE
TextRE	TP-Linker (Wang et al., 2020)	78.63	83.49	58.56	50.46	51.83	20.39	78.63	83.49	58.56
	Spert (Eberts and Ulges, 2020)	76.38	81.83	63.45	60.26	63.48	21.46	76.38	81.83	63.45
	REBEL (Cabot et al., 2021)	85.36	89.86	71.46	60.09	65.15	25.15	85.36	89.86	71.46
SpeechRE (Pipline)	TP-Linker$_{pipe}$ (Wu et al., 2022)	35.21	78.21	9.76	30.27	50.01	6.59	33.63	76.57	8.21
	Spert$_{pipe}$ (Wu et al., 2022)	30.43	75.95	11.88	34.36	57.17	6.89	29.32	73.61	10.73
	REBEL$_{pipe}$ (Wu et al., 2022)	37.06	83.35	14.01	32.07	51.97	6.49	36.42	81.64	12.08
SpeechRE (End2End)	LNA-ED (Wu et al., 2022)	18.87	55.66	10.41	17.21	43.37	3.20	13.11	52.55	6.08
	SATE (Xu et al., 2021)	21.53	55.90	11.38	16.01	46.97	3.02	14.87	53.06	6.74
	Chimera (Han et al., 2021)	12.75	58.44	8.08	16.01	46.97	3.02	8.77	56.91	4.96
	MSP-ST (Zhang et al., 2023d)	26.60	70.33	13.15	19.03	48.97	4.07	19.31	67.70	8.32
	CMOT (Zhou et al., 2023)	28.24	70.95	14.02	20.77	49.41	4.65	20.49	67.84	9.05
	MCAM (ours)	**40.13**	**77.89**	**22.07**	**35.34**	**58.96**	**8.07**	**31.16**	**73.55**	**16.25**

Table 4: The model performance on the test sets of CONLL04, ReTACRED, and Mixed-CoNLL04. Here, ER = Entity Recognition, RP = Relation Prediction, and RTE = Relation Triplet Extraction

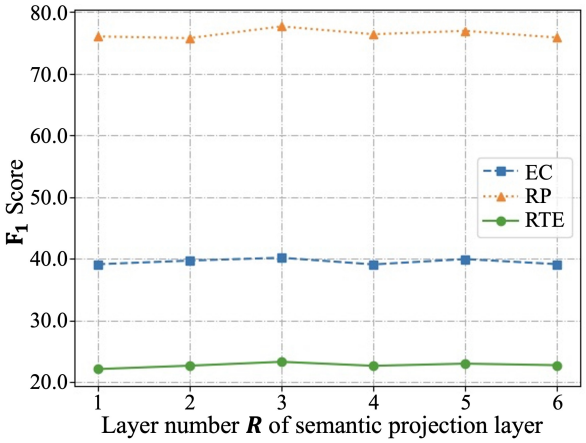

Figure 4: The performance of our model with different layer number N of semantic projection layer on the CONLL04 validation set.

role in balancing our multiple loss terms during model training. To investigate their influence on our model, we conduct experiments with different values of α and β on the CONLL04 validation set. As illustrated in Table 3, our model achieves the best performance in relation triplet extraction when α and β are set to 0.8 and 0.2, respectively. Meanwhile, we also observe that the performance of our model is not significantly affected by the values of α and β. Hence, we adopt α=0.8 and β=0.2 for all experiments.

Effect of Hyper-parameter N Our semantic projection layer aims to map the speech and text features into a shared semantic space, thereby further mitigating the modality gap between the speech encoder and text decoder. Meanwhile, all of its parameters are updated during our entire training process. Here, we will explore the impact of the layer number N of the projection layer on the overall performance (RTE) of our model. As illustrated in Figure 4, our model reaches its best performance when N is set to 3. Meanwhile, we notice that our model exhibits inferior performance when a larger value is assigned to N. The main reason for this is that the limited SpeechRE data poses a challenge in effectively training the model with a large number of parameters. Furthermore, our model is insensitive to N and consistently outperforms all pipeline and end2end SpeechRE models at all values of N.

5.5 Main Results

The experimental results on the three datasets we consider are shown in Table 4. After carefully analyzing these results, we draw several conclusions:

First, our model consistently outperforms all end-to-end SpeechRE models across all metrics (See **Bottom rows**). This indicates that our MCAM can bridge the modality gap between the speech encoder and text decoder more effectively than the cross-modal alignment methods in other speech-related tasks. Particularly, compared to our base model LNA-ED, our model achieves the improvements of **21.26/22.23/11.66** points, **18.13/15.59/4.87** points, and **18.05/21.00/10.17** points on the above three datasets, respectively. These results further demonstrate the effectiveness of our model.

Second, we note that end-to-end SpeechRE models exhibit better performance on CONLL04 than on Mixed-CoNLL04. It implies that SpeechRE models trained on synthetic data often underperform in real-world scenarios. However, our model continues to outperform all end-to-end models on both Mixed-CoNLL04 and ReTACRED, which demonstrates the robustness of our model

to noise in synthetic speech. Since annotating real SpeechRE data is costly, it is worthwhile to further explore the utilization of synthetic data for training a robust SpeechRE model in real-world scenarios.

Third, our model exceeds all pipeline SpeechRE models in relation triplet extraction on all datasets. This be attributed to the error propagation issue, which makes it more challenging for pipeline models to generate complete relation triplets accurately. Although incorporating stronger (larger) ASR modules into the pipeline model may alleviate this issue, it can also lead to larger inference latency. Hence, we leave the development of a strong pipeline model with low inference latency for future study.

Lastly, we observe that end-to-end SpeechRE models usually perform better in relation prediction than in entity recognition. Thus, enhancing the entity recognition ability of the SpeechRE model is crucial for further improving its overall performance. As in previous studies, we also discover that there is still significant room for improvement in the performance of the SpeechRE model compared to the TextRE models.

5.6 Ablation Studies

We further conduct extensive ablation studies by removing different components from our model to comprehend their different impacts. We compare our model with the following variants in Table 5.

(1) *w/o. Token-level alignment*. In this variant, we neglect the token-level alignment and remove the loss $\mathcal{L}_{\text{CTC}}$ from entire training objective. As shown in **Line 1**, this change leads to a significant performance drop across all metrics. Meanwhile, we note that the token-level alignment primarily enhances the overall performance (RTE) of the model by improving its entity extraction capability (ER).

(2) *w/o. Entity-level alignment*. When we discard the entity-level alignment and remove its corresponding loss $\mathcal{L}_{\text{KL}}^{(t\rightarrow m)}$ from our training objectives, the entity extraction performance of our model suffers a significant decline (See **Line 2**). It suggests that our entity-level alignment can indeed improve the accuracy of our model in generating entity.

(3) *w/o. Sentence-level alignment*. This variant removes the two components responsible for the sentence-level alignment from our model: the semantic compression layer and the contrastive loss $\mathcal{L}_{\text{CL}}$. As reported in **Line 3**, the model performance exhibits a more significant decrease of 3.44 points in relation prediction, compared to that in entity recognition and relation triplet extraction. This confirms that the sentence-level alignment mainly contributes to enhancing the relation extraction capability of our model.

Model	CONLL04		
	ER	RP	RTE
Ours	**40.13**	**77.89**	**22.07**
1 *w/o.* Token-level alignment	36.61	76.57	19.47
2 *w/o.* Entity-level alignment	36.09	76.83	19.14
3 *w/o.* Sentence-level alignment	38.75	74.94	19.56
4 *w/o.* Semantic projection layer	38.01	76.08	19.77
5 *w/o.* Pre-training	33.62	74.68	17.96
6 *w/o.* Alignment in pre-training	35.75	76.16	19.07

Table 5: Ablation results on the CONLL04 test set.

(4) *w/o. Semantic Projection Layer*. Our semantic projection layer aims to project the speech and text features into a shared semantic space, thereby further mitigating the modality gap issue between the speech encoder and text decoder. To prove its effectiveness, we remove it from our model, resulting in a decline in our model performance across all metrics (See **Line 4**). This indicates that semantic projection layer is necessary for our model.

(5) *w/o. Pre-training*. Here, we directly finetune our model on SpeechRE datasets without pretraining on ASR data. As shown in **Line 5**, this variant is significantly inferior to our model, suggesting that our pre-training method is more suitable for SpeechRE than prior pseudo-labeling methods.

(6) *w/o. Alignment in pre-training*. In this case, we pre-train our model on the ASR dataset without using our cross-modal alignments, which also leads to a significant performance decline across all metrics (See **Line 6**). It indicates that our cross-modal alignments play a crucial role in the effectiveness of our pre-training method.

Please see **Appendix A** for more further analyses.

6 Related Work

Relation extraction (RE) is a fundamental task in information extraction, which aims to extract structured knowledge from unstructured data in the form of relation triplet (Song et al., 2019; Han et al., 2020; Wu et al., 2022; Zhang et al., 2022, 2023b,a,c; Yue et al., 2024). In this regard, dominant studies mainly focus on extracting relation triplets from plain text (Cabot et al., 2021; Lu et al., 2022; Wang et al., 2023). However, in addition to text, plenty of speech data is continuously produced in our daily lives. These speech data often contain rich and valuable structured knowledge. Therefore, it is meaningful to extract relation

triplets from speech data. To do this, Wu et al. (2022) are the first to explore this task. Nonetheless, they mainly use synthetic data to train and evaluate SpeechRE models, hindering the further development of SpeechRE due to the disparity between synthetic and real speeches. Moreover, their proposed baseline model fails to effectively align the two modalities of speech and text, resulting in poor performance.

In other speech-related tasks, such as Speech Translation (ST) and Automatic Speech Recognition (ASR), researchers have proposed many modality alignment methods (Tang et al., 2021; Han et al., 2021; Xu et al., 2021; Wang et al., 2022; Zhang et al., 2023d; Zhao et al., 2024). For example, Tang et al. (2021) propose an attention-based regularization to pull the representations from different modalities closer. Han et al. (2021) introduce a shared semantic projection module to map speech and text features into a common semantic space and align them via contrastive learning. Wang et al. (2022) mix up the feature sequences of different modalities, and then take both the unimodal speech sequence and multimodal mixed sequence as inputs to the translation model in parallel, where their output predictions are regularized with a self-learning framework. Zhang et al. (2023d) first employ a CTC loss to align speech features and token embeddings, and use contrastive loss to align speech and text features at the sentence level. Although these methods have achieved improvements in their respective tasks, they often only consider the alignment between speech and text features at the token or sentence level.

In this work, we first create two real SpeechRE datasets to facilitate future studies. Consi the SpeechRE model is usually required to comprehensively understand the input speech from multiple levels, we propose a Multi-level Cross-modal Alignment Model (MCAM) for SpeechRE. It aligns the speech and text features at the token, entity, and sentence levels for better SpeechRE.

7 Conclusion and Future Work

In this paper, we first annotate two real SpeechRE datasets to facilitate future research in the field of SpeechRE. Then, we propose a new SpeechRE model MCAM, which consists of a speech encoder, an alignment adapter, and a text decoder. The alignment adapter aims to project the speech features, extracted by the speech encoder from the input speech, into a suitable semantic space for the text decoder, enabling it to effectively generate relation triplets based on these speech features. By doing so, we successfully bridge the modality gap between the speech encoder and text decoder. To efficiently train our adapter, we introduce an additional text encoder and train the adapter to align the output features of the speech and text encoders at three levels. Experiments on our two SpeechRE datasets demonstrate the effectiveness of our model.

In future, we plan to apply our model to other speech-related tasks, such as ST and ASR, so as to verify its generality. Moreover, exploring methods to train a zero-shot SpeechRE model using only vast ASR and TextRE data is an interesting future research direction.

Acknowledgments

The project was supported by National Natural Science Foundation of China (No. 62276219), and the Public Technology Service Platform Project of Xiamen (No. 3502Z20231043). We also thank the reviewers for their insightful comments.

Limitations

The limitations of our method mainly include following two aspects: 1) Our model does not fully exploit the intrinsic information contained in input speech, such as emotion inflections and pauses, which could be beneficial to our task. 2) During the pre-training stage, we do not use the existing large amount of TextRE data, which may effectively enhance the overall performance of our model.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of NeurIPS*.

Pere-Lluís Huguet Cabot, Roberto Navigli, Pere-Lluís Huguet Cabot, and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Proceedings of Findings of EMNLP*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *Proceedings of NAACL*.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *Proceedings of ECAI*.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of ACL*.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Proceedings of ACL*.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of AACL*.

Xuming Hu, Zhijiang Guo, Zhiyang Teng, Irwin King, and Philip S Yu. 2023. Multimodal relation extraction with cross-modal retrieval and synthesis. In *Proceedings of ACL*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation with efficient finetuning of pretrained models. In *Proceedings of ACL*.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of ACL*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of NIPS*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of NAACL*.

Linfeng Song, Yue Zhang, Daniel Gildea, Mo Yu, Zhiguo Wang, and Jinsong Su. 2019. Leveraging dependency forest for neural medical relation extraction. In *Proceedings of EMNLP*.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of AAAI*.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of AAAI*.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of ACL*.

Chen Wang, Yuchen Liu, Boxing Chen, Jiajun Zhang, Wei Luo, Zhongqiang Huang, and Chengqing Zong. 2022. Discrete cross-modal alignment enables zero-shot speech translation. In *Proceedings of ACL*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of COLING*.

Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. 2023. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of ACL*.

Tongtong Wu, Guitao Wang, Jinming Zhao, Zhaoran Liu, Guilin Qi, Yuan-Fang Li, and Gholamreza Haffari. 2022. Towards relation extraction from speech. In *Proceedings of EMNLP*.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Qi Ju, Tong Xiao, Jingbo Zhu, et al. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. In *Proceedings of ACL*.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of NAACL*.

Hao Yue, Shaopeng Lai, Chengyi Yang, Liang Zhang, Junfeng Yao, and Jinsong Su. 2024. Towards better graph-based cross-document relation extraction via non-bridge entity enhancement and prediction debiasing. In *Proceedings of Findings of ACL*.

Liang Zhang, Zijun Min, Jinsong Su, Pei Yu, Ante Wang, and Yidong Chen. 2023a. Exploring effective inter-encoder semantic interaction for document-level relation extraction. In *Proceedings of IJCAI*.

Liang Zhang, Jinsong Su, Yidong Chen, Zhongjian Miao, Zijun Min, Qingguo Hu, and Xiaodong Shi. 2022. Towards better document-level relation extraction via iterative inference. In *Proceedings of EMNLP*.

Liang Zhang, Jinsong Su, Zijun Min, Zhongjian Miao, Qingguo Hu, Biao Fu, Xiaodong Shi, and Yidong Chen. 2023b. Exploring self-distillation based relational reasoning training for document-level relation extraction. In *Proceedings of AAAI*.

Liang Zhang, Chulun Zhou, Fandong Meng, Jinsong Su, Yidong Chen, and Jie Zhou. 2023c. Hypernetwork-based decoupling to improve model generalization for few-shot relation extraction. In *Proceedings of EMNLP*.

Yuhao Zhang, Chen Xu, Bojie Hu, Chunliang Zhang, Tong Xiao, and Jingbo Zhu. 2023d. Improving end-to-end speech translation by leveraging auxiliary speech and text data. In *Proceedings of AAAI*.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. 2024. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of AAAI*.

Changmeng Zheng, Junhao Feng, Yi Cai, Xiaoyong Wei, and Qing Li. 2023. Rethinking multimodal entity and relation extraction from a translation point of view. In *Proceedings of ACL*.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. Cmot: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of ACL*.

Model	CONLL04		
	ER	RP	RTE
1 REBEL	85.36	89.86	71.46
2 MCAM (ours)	40.13	77.89	22.07
Speaker-Independent SpeechRE			
3 MCAM (ours)	32.05	73.77	16.54
4 *w/o. All alignments*	28.27	70.90	13.16
Speech-Enhanced TextRE			
5 MCAM (ours)	88.57	90.16	74.12
6 *w/o. All alignments*	86.90	89.45	72.08
Zero-Shot SpeechRE			
7 MCAM (ours)	17.69	51.40	6.21
8 *w/o. All alignments*	13.79	48.12	2.65

Table 6: Ablation results on the CONLL04 test set.

Appendix

A Further Analysis

In this section, we focus on exploring some interesting research directions in the SpeechRE domain, while validating the effectiveness of our model in these directions. We present the experimental results in Table 6.

Speaker-Independent SpeechRE. The robustness of SpeechRE models to the speaker of the input speech directly affects their practicality in real-world scenarios. Therefore, we re-partitioned the CONLL04 dataset to ensure that the speakers in the training set, validation set, and test set are entirely distinct. Subsequently, we evaluate our model on this new dataset (See **Line 3-4**). Although the speaker difference degrade the performance of our model, our cross-modal alignment method can effectively alleviate this degradation and improve the model's robustness to speakers. The probable reason is that our alignment method can assist the model to learn speaker-independent speech features.

Speech-Enhanced TextRE. Many studies (Sun et al., 2021; Zheng et al., 2023; Wu et al., 2023; Hu et al., 2023) have explored the use of images to enhance TexRE. Similar to image, speech also contains rich multimodal information that can be used to enhance TextRE. Here, we provide features generated by both the text and speech encoder to the text decoder during the fine-tuning stage. As shown in **Line 5**, speech features can indeed improve the performance of TextRE. Meanwhile, removing our cross-modal alignment method leads to a decline in the model performance across all metrics (See **Line 6**). This suggests that our method can help the TextRE model learn beneficial multimodal information for relation extraction from speech.

Zero-Shot SpeechRE. Given that SpeechRE data is scarce and expensive to obtain, it holds practical significance to train an end-to-end SpeechRE model using only ASR and TextRE data. In this case, we first pre-train our model on the ASR dataset using the ASR task along with our cross-modal alignment method. Then, we solely fine-tune the text decoder on the TextRE dataset and freeze the other components in our model. As shown in **Line 7**, the SpeechRE model trained under this setting achieves a lower performance. The primary reasons for this is the modality gap between the speech encoder and text decoder, as well as the length differences between the speech and text feature sequences. Through the ablation study in **Line 8**, we find that our cross-modal alignment methods can alleviate such modality gap.