

ACRE: An Acoustic-Cue and Contextual-Knowledge Reasoning Enhancement Framework for Speech Relation Extraction

Wenxuan Mu¹ Jinzhong Ning^{1,*} Di Zhao² Yijia Zhang^{1,*}

¹School of Information Science and Technology, Dalian Maritime University, China

²School of Computer Science and Engineering, Dalian Minzu University, China

{mw, ningjinzong, zhangyijia}@dlmu.edu.cn, zhaodi@dlmu.edu.cn

Abstract

Speech relation extraction (SpeechRE) aims to extract structured relational triples directly from speech. However, existing studies still face three major challenges: the lack of Chinese SpeechRE benchmarks, insufficient modeling of speech-specific cues, and evaluation bias introduced by character-level exact matching in Chinese scenarios. To address these challenges, we first construct CS RTE, a real-world Chinese SpeechRE benchmark built from human speech, providing unified data and evaluation settings. We then propose ACRE, a reasoning enhancement framework for SpeechRE, which integrates acoustic-cue and contextual-knowledge into a structured acoustic-semantic reasoning chain, enabling the model to go beyond literal semantic content (what-to-say) and capture speech-specific features (how-to-say) for relation reasoning. We further combine supervised fine-tuning with reinforcement learning and design composite rewards to improve the quality of the model’s reasoning process. In addition, we propose pronunciation-aware evaluation (PAE), which considers pronunciation and tone consistency during entity matching to reduce the evaluation bias. Experiments on both Chinese and English SpeechRE benchmarks demonstrate the effectiveness of ACRE and the usefulness of PAE. Our source code and datasets are available at github.com/MWXGOD/CS RTE.

1 Introduction

Speech relation extraction (SpeechRE) aims to extract relational triples, e.g., <head entity, relation, tail entity>, directly from speech (Wu et al., 2022; Zhang et al., 2024). It plays an important role in many downstream applications, such as spoken language understanding (Arora et al., 2024), speech information retrieval (Lee et al., 2015), and spoken question answering (Nachmani et al., 2024).

*Corresponding author.

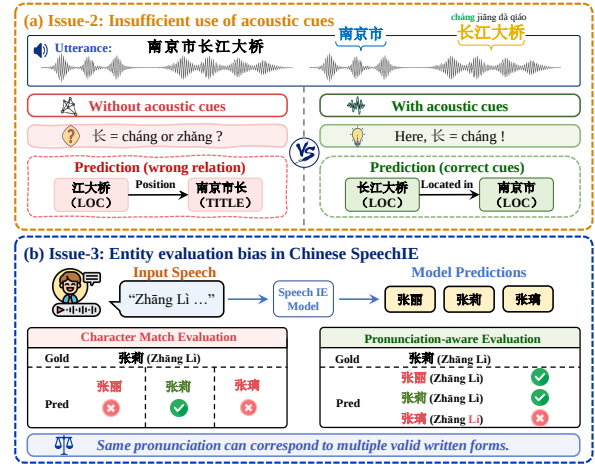


Figure 1: Two challenges in Chinese SpeechRE: (a) Issue-2, insufficient use of acoustic cues for relation understanding; (b) Issue-3, entity evaluation bias caused by Chinese homophones under character-level exact matching.

Compared with text relation extraction (TextRE), SpeechRE is more challenging because speech may contain background noise and accents, and lacks explicit semantic units such as words.

Early studies mainly focused on TextRE (Zhao et al., 2024). However, with the development of information technology, the amount of speech data has grown rapidly and has become widely available in daily life, such as in news broadcasts and short videos. Such data contains rich and valuable structured knowledge (Zhang et al., 2024; Ning et al., 2026). As a result, SpeechRE has attracted increasing attention in recent years. Existing studies on SpeechRE mainly focus on the cross-modal alignment problem. Early methods, such as LNA-ED (Wu et al., 2022) and MCAM (Zhang et al., 2024), were mainly conducted with synthetic speech data, while RPG-MoGe (Ning et al., 2026) further extends SpeechRE to real human speech, making the task closer to real-world application scenarios.

Despite these advances, existing works still face

the following limitations. **Issue-1: Lack of Chinese SpeechRE benchmarks.** Existing SpeechRE studies mainly focus on English datasets, while Chinese SpeechRE benchmarks remain under-explored. This limits the training, evaluation, and application of models in Chinese SpeechRE scenarios. **Issue-2: Insufficient use of acoustic cues.** Existing SpeechRE methods mainly focus on semantic alignment between speech and text at the literal meaning level (*what-to-say*), but make limited use of acoustic cues such as prosody and pronunciation (*how-to-say*). These cues can provide additional evidence for entity recognition and relation prediction, as shown in Figure 1 (a). Ignoring such acoustic information may weaken the model’s ability to capture relation expressions, thereby affecting the accuracy of SpeechRE.

Beyond the above issues, another important challenge remains to be addressed: **Issue-3: Entity Evaluation Bias in Chinese SpeechIE.** Existing speech information extraction (SpeechIE) studies usually follow the character-level exact matching used in text information extraction (TextIE) to determine whether an entity is correctly predicted. However, this metric may lead to unfair evaluation, especially in Chinese scenarios. Due to the large number of homophones in Chinese, especially in entities such as PER and LOC, many entities may not have a unique written form based only on speech and context. For example, as shown in Figure 1 (b), the pronunciation “zhāng lì” may be written as either “张丽” (zhāng lì) or “张莉” (zhāng lì). Existing evaluation metrics regard a prediction as wrong once its written form does not match the annotation, even if its pronunciation is correct. This may lead to evaluation bias and underestimate the model’s actual ability in SpeechIE.

To address the above challenges, this paper investigates Chinese SpeechRE from three aspects: benchmark construction, model design, and evaluation. For benchmark construction, to address **Issue-1**, we introduce Chinese Speech Relation Triplet Extraction (**CSRTE**), a benchmark for Chinese SpeechRE, designed to support model training, validation, and evaluation in Chinese spoken scenarios. CSRTE is constructed from real human speech data and provides a unified evaluation framework for Chinese SpeechRE.

For model design, we propose **ACRE**, an acoustic-cue and contextual-knowledge reasoning enhancement framework for SpeechRE. Specifically, to address **Issue-2**, ACRE explicitly models

pronunciation information and entity-related contextual knowledge, and organizes them into a structured acoustic-semantic chain-of-thought (SAS-CoT). In addition, we combine supervised fine-tuning (SFT) with reinforcement learning (RL), enabling progressive optimization of both reasoning quality and triple generation performance.

For evaluation, to address **Issue-3**, we propose a pronunciation-aware evaluation (**PAE**) strategy that considers pronunciation consistency during entity matching, allowing pronunciation-equivalent entities to be treated as valid matches in SpeechRE.

The main contributions of this paper are summarized as follows:

- We construct CSRTE, a Chinese real-world SpeechRE benchmark that provides a unified data resource and evaluation settings for Chinese SpeechRE research.
- We propose ACRE, an acoustic-semantic reasoning enhanced framework that introduces SAS-CoT for SpeechRE.
- We propose PAE, which incorporates pronunciation and tone consistency into entity matching to alleviate the bias of character-level exact matching in Chinese SpeechIE.
- Extensive experiments on both Chinese and English benchmarks demonstrate the effectiveness of ACRE.

2 Related Work

2.1 Speech Relation Extraction

As a core task of SpeechIE (Ning et al., 2026, 2025a,b, 2024), SpeechRE aims to extract relational triples directly from continuous speech signals. This task was first introduced by Wu et al. (2022), who constructed a synthetic speech benchmark from existing English TextRE datasets using Text To Speech (TTS) techniques (Jia et al., 2026) and developed a corresponding end-to-end modeling framework. Building on this, Zhang et al. (2024) further investigated cross-modal semantic alignment between speech and text, proposing to align speech and textual representations at multiple levels, including tokens, entities, and sentences, thereby improving model performance. Different from these studies based on synthetic speech,

Datasets	#Relations	#Instances			#Triplets			Avg. audio length
		Train	Dev	Test	Train	Dev	Test	
🔊 CoNLL04-SpeechRE	5	922	231	288	1,283	343	422	11.3s
👤 CommonVoice-SpeechRE	45	14,557	2,495	2,494	15,948	2,696	2,728	11.6s
👤 CSRTE-SpeechRE (Ours)	49	5,000	2,000	2,855	6,244	2,465	3,345	6.14s

Table 1: Statistics of different SpeechRE datasets. Datasets marked with 🔊 are constructed from synthesized speech, while datasets marked with 👤 are constructed from human-recorded speech.

Ning et al. (2026) extended SpeechRE to real-world speech scenarios by constructing a large-scale human-recorded SpeechRE dataset and adopting a generative end-to-end framework that jointly optimizes cross-modal representation learning and triple generation.

2.2 Speech Foundation Models

Early studies on Speech Foundation Models mainly focused on self-supervised representation learning, with representative models such as wav2vec (Schneider et al., 2019), wav2vec 2.0 (Baevski et al., 2020), and HuBERT (Hsu et al., 2021). To support speech-to-text tasks, including automatic speech recognition (ASR) and speech translation, researchers further explored encoder-decoder based generative speech models. SpeechT5 (Ao et al., 2022) proposed a unified framework for ASR, speech generation, and speech translation, while Whisper (Radford et al., 2023), trained with large-scale weak supervision, substantially improved multilingual speech recognition and robustness.

With the rapid development of large language models (LLMs), researchers have started to combine speech encoders with language models to build SpeechLLMs. Pengi (Deshmukh et al., 2023) made an early attempt to transfer language-model capabilities to general audio understanding tasks. SpeechGPT (Zhang et al., 2023) subsequently supported spoken dialogue and speech generation through cross-modal instruction tuning. Qwen-Audio (Chu et al., 2023), SALMONN (Tang et al., 2024), and Qwen2-Audio (Chu et al., 2024) further improved unified understanding of diverse audio inputs. More recently, MiMo-Audio (Xiaomi, 2025) demonstrated the few-shot generalization ability enabled by large-scale speech-language pretraining. Overall, SpeechLLMs have evolved into unified frameworks with stronger speech understanding and reasoning capabilities.

3 New Dataset

We present Chinese Speech Relation Triplet Extraction (CSRTE), a new Chinese SpeechRE dataset constructed from subsets of Common Voice 17 (Ardila et al., 2020) and AISHELL-1 (Bu et al., 2017). Common Voice 17 is a large-scale multilingual speech corpus containing 20,408 hours of validated recordings across 124 languages, contributed by volunteers worldwide. It is released under the CC0 license, which allows unrestricted use, modification, and redistribution. AISHELL-1 is a large-scale open-source Mandarin speech corpus released by Beijing Shell Shell Technology Co., Ltd. It contains approximately 178 hours of high-quality Mandarin speech from 400 speakers across different accent regions in China, and is distributed under the Apache 2.0 license.

Most utterances in Common Voice 17 and AISHELL-1 are negative instances without valid entity mentions or relation triplets. To identify potential positive samples, we first normalized the transcriptions and removed empty, malformed, duplicated, or noisy sentences. We then used a pre-trained BERT-based NER tagger fine-tuned on Chinese OntoNotes annotations to detect entity and value mentions, and retained utterances containing at least two predicted mentions as candidates. The tagger was used only for candidate filtering, while all final labels were manually annotated.

For the annotation schema, we developed a tailored annotation guideline for Chinese SpeechRE, where the entity and relation definitions were designed with reference to OntoNotes (Weischedel et al., 2022) and ACE 2004 (Doddington et al., 2004)/2005 (Walker et al., 2006). A team of 10 graduate students manually annotated approximately 10,000 transcriptions using Label Studio¹ (Tkachenko et al., 2020-2025). All annotators were native Chinese speakers with backgrounds in information extraction. Each sentence was indepen-

¹labelstud.io

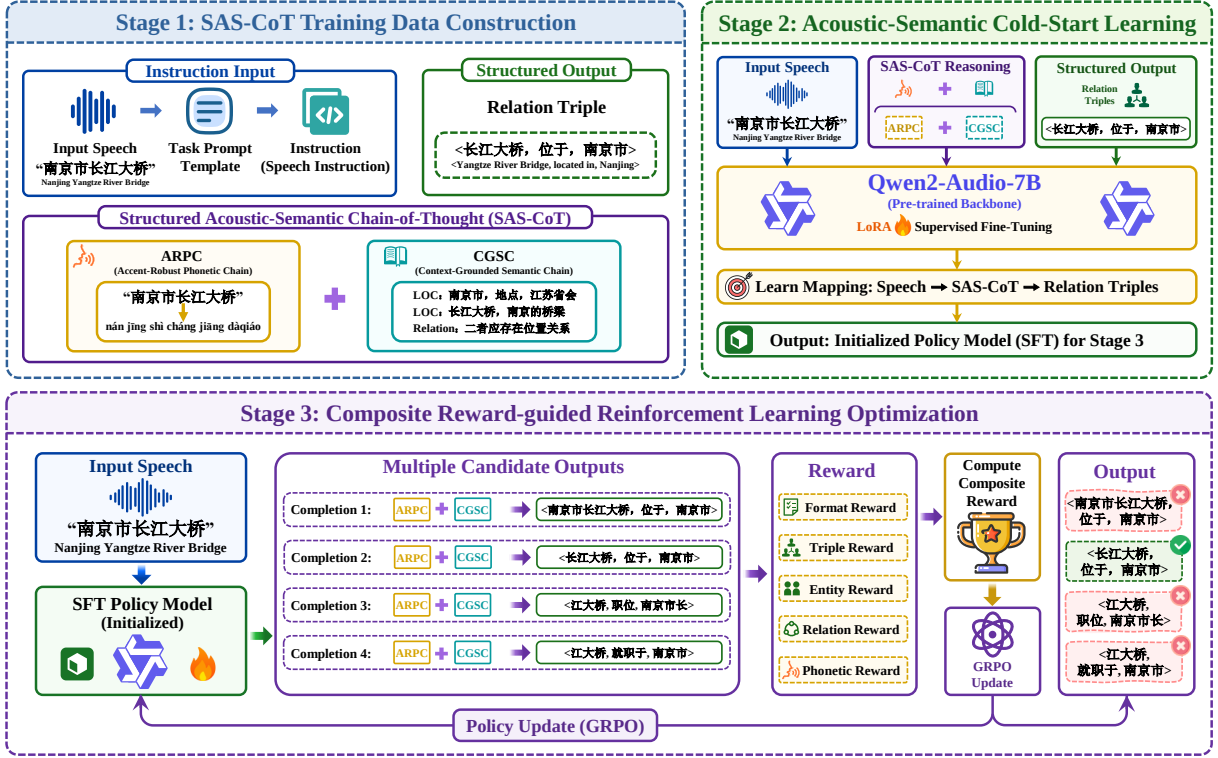


Figure 2: Overall architecture.

dently annotated by three annotators, and a senior reviewer resolved disagreements to produce the final labels. Annotation was conducted in batches of no more than 1000 sentences, with 10% of each batch randomly sampled for quality verification. If sentence-level accuracy fell below 95%, the entire batch was re-annotated. We further measured inter-annotator agreement on triples using Fleiss’ Kappa, obtaining a score of 0.847, which indicates substantial agreement. Table 1 summarizes the statistics of CS RTE and existing SpeechRE datasets.

4 Method

In this section, we formally define the SpeechRE task and present the detailed implementation of our proposed ACRE framework. It consists of three core modules: SAS-CoT Training Data Construction, Acoustic-Semantic Cold-Start Learning, and Reinforcement Learning Optimization. The architecture is illustrated in Figure 2.

4.1 Task Formulation

We study the task of SpeechRE. Given a raw speech signal A , the model is required to directly extract a set of relation triplets from the speech input, denoted as $Y = \{(h_i, r_i, t_i)\}_{i=1}^m$, where h_i and t_i denote the head and tail entities, respectively, and

r_i denotes the relation type between them.

4.2 SAS-CoT Training Data Construction

To enhance the SpeechRE model’s ability to explicitly reason over speech cues, we construct reasoning-aware training instances consisting of speech inputs, task instructions, structured reasoning chains, and target relation triplets.

Specifically, given a speech input A_j , a task instruction Q_j , a structured reasoning chain C_j , and the target relation triplets Y_j , we construct the input-output pair of the j -th training instance as:

$$\begin{aligned} \text{Input} &= \text{Template}(A_j, Q_j), \\ \text{Output} &= \text{Concatenate}(C_j, Y_j). \end{aligned} \quad (1)$$

For the structured reasoning chain C , we construct SAS-CoT as an intermediate reasoning representation. It consists of two components: the Accent-Robust Phonetic Chain (ARPC) for pronunciation-level cues and the Context-Grounded Semantic Chain (CGSC) for contextual entity-relation semantics.

Accent-Robust Phonetic Chain. Given the gold transcript x_j , we obtain the pronunciation chain through a pronunciation conversion function g_{pron} :

$$c_j^{\text{pron}} = g_{\text{pron}}(x_j). \quad (2)$$

For Chinese samples, g_{pron} converts Chinese characters into tone-marked Pinyin; for English samples, it converts words into IPA-style phonetic representations. Meanwhile, we preserve punctuation and the major boundary structure of the original sentence.

Context-Grounded Semantic Chain. When speech input contains ambiguity, the model can benefit from additional information to support relation reasoning. We therefore construct contextual reasoning knowledge about entities and relations to assist the model in relation prediction.

Specifically, given the gold transcript x_j and the entity list E_j , we construct the semantic chain through a function g_{sem} :

$$c_j^{\text{sem}} = g_{\text{sem}}(x_j, E_j). \quad (3)$$

In practice, g_{sem} is implemented using the GPT-4o (OpenAI, 2024), which generates concise entity background descriptions and preliminary relation reasoning.

The final SAS-CoT representation is obtained by combining the two components:

$$C_j = (c_j^{\text{pron}}, c_j^{\text{sem}}), \quad (4)$$

where c_j^{pron} and c_j^{sem} denote the outputs of ARPC and CGSC, respectively.

It is worth noting that gold transcripts and entity annotations are used only for SAS-CoT construction during training. During inference, the model uses only raw speech and the instruction, without gold annotations or external knowledge generators.

4.3 Acoustic-Semantic Cold-Start Learning

After constructing SAS-CoT, we use Qwen2-Audio-7B as the backbone and perform SFT with Low-Rank Adaptation (LoRA) (Hu et al., 2022). This stage teaches the model to generate the reasoning process and relation triplets from the speech input and task instruction. The target output is Z_j :

$$Z_j = [C_j; Y_j], \quad (5)$$

where C_j denotes SAS-CoT and Y_j denotes the final relation triplets. The SFT objective is to minimize the autoregressive negative log-likelihood:

$$\mathcal{L}_{\text{SFT}} = - \sum_{t=1}^T \log p_{\theta}(z_t | z_{<t}, A_j, Q_j), \quad (6)$$

where z_t denotes the t -th token in the target sequence Z_j , $z_{<t}$ denotes the preceding tokens, and T denotes the length of the target sequence. In this way, the SFT stage provides a stable output format and initial reasoning ability for the RL stage.

4.4 Reinforcement Learning Optimization

Although cold-start learning enables the model to acquire the basic generation format of SAS-CoT and the correspondence between SAS-CoT and relation triplets, it still optimizes reference-based next-token prediction rather than the SpeechRE task objective. To better align generation with the task objective, we introduce a RL optimization strategy based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024) after SFT, shifting the model from token-level imitation learning to reasoning-aware relative optimization.

This stage takes the SFT model as the initial policy model and keeps the same input-output format as the SFT stage. Given the speech input A_j and the task instruction Q_j , the model is required to generate a complete output containing both the SAS-CoT reasoning process and the final relation triplets:

$$O_j = [\hat{C}_j; \hat{Y}_j], \quad (7)$$

where $\hat{C}_j = (\hat{c}_j^{\text{pron}}, \hat{c}_j^{\text{sem}})$ denotes the generated SAS-CoT, and $\hat{Y}_j$ denotes the generated set of relation triplets.

To jointly optimize the final relation extraction results and the quality of SAS-CoT reasoning, we design a composite reward function. The overall reward is defined as:

$$R = \lambda_{\text{fmt}} R_{\text{fmt}} + \lambda_{\text{tri}} R_{\text{tri}} + \lambda_{\text{ent}} R_{\text{ent}} + \lambda_{\text{rel}} R_{\text{rel}} + \lambda_{\text{pron}} R_{\text{pron}} \quad (8)$$

where R_{fmt} denotes the format reward. R_{tri} , R_{ent} , and R_{rel} denote the triplet-level, entity-level, and relation-type-level F1 rewards, respectively. R_{pron} denote the pronunciation reasoning reward.

Specifically, we uniformly compute the pronunciation reasoning rewards using normalized edit-distance similarity:

$$R_x = 1 - \frac{\text{EditDistance}(\hat{c}^{\text{pron}}, c^{\text{pron}})}{\max(|\hat{c}^{\text{pron}}|, |c^{\text{pron}}|)}, \quad (9)$$

where $\hat{c}^{\text{pron}}$ and c^{pron} denote the model-generated reasoning chain and the reference reasoning chain, respectively.

After obtaining the above composite rewards, we optimize the policy model with GRPO. For each input sample (A_j, Q_j) , the current policy model samples a group of candidate outputs $\{O_j^i\}_{i=1}^G$, and computes the reward of each candidate output:

$$r_j^i = R(O_j^i, C_j, Y_j). \quad (10)$$

Then, GRPO normalizes the rewards within the same group to obtain the relative advantage:

$$\hat{A}_j^i = \frac{r_j^i - \mu_j}{\sigma_j + \delta}, \quad (11)$$

where μ_j and σ_j denote the mean and standard deviation of the rewards of the G candidate outputs for the same input, respectively, and δ a small constant for numerical stability.

Based on this relative advantage, the GRPO optimization objective can be simplified as:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{O_j^i \sim \pi_\theta} \left[\hat{A}_j^i \log \pi_\theta(O_j^i | A_j, Q_j) \right] + \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \quad (12)$$

where π_θ denotes the policy model, and π_{ref} denotes the reference policy. β controls the strength of the KL regularization term.

5 Experiments

5.1 Datasets

We conducted experiments on three SpeechRE datasets: CoNLL04, CommonVoice-SpeechRE, and our proposed CSRTE. CoNLL04 is English synthetic speech datasets constructed by converting text-based RE datasets into speech using TTS systems. CommonVoice-SpeechRE is an English SpeechRE dataset built from real human-recorded speech, providing a more realistic evaluation setting. In this paper, we further introduce CSRTE, a Chinese real-world SpeechRE benchmark built from human speech, which is designed to evaluate SpeechRE models under Chinese spoken scenarios. Detailed statistics are provided in Table 1.

5.2 Evaluation Metrics

Following previous SpeechRE studies, we use micro-F1 scores to evaluate model performance. Under strict evaluation, an entity is correct only if its surface form exactly matches the gold entity, and a relation triplet is correct only if its head entity, relation type, and tail entity all match the gold

annotation. We also report PAE, which converts predicted and gold entities into canonical pronunciation sequences and treats an entity as matched only when the pronunciation sequence, including tone for Chinese, is identical. Entity matching remains one-to-one, and relation-type matching remains unchanged, so pronunciation-equivalent entities can be treated as valid matches without relaxing the relation label constraint. PAE does not replace strict exact matching, but provides a complementary view for speech-based entity evaluation.

5.3 Experimental Settings

Our model was implemented with PyTorch and the MS-Swift framework, using Qwen2-Audio-7B² as the backbone. We adopted LoRA for parameter-efficient tuning with rank 8 and α 32. In the supervised fine-tuning stage, we used a learning rate of 2e-4 and trained for 10 epochs. The best checkpoint was selected according to the triple F1 on the development set. We then further optimized the model with GRPO using a learning rate of 5e-6 and KL coefficient $\beta = 0.04$. During inference, we used greedy decoding with temperature 0.0 and top-k 1.

5.4 Baselines

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

TextRE Models. TextRE models take textual transcripts as input and jointly extract entities and relations. Following Ning et al. (2026), we evaluate several TextRE baselines: TP-Linker (Wang et al., 2020), Spert (Eberts and Ulges, 2020), REBEL (Cabot and Navigli, 2021), BiRTE (Ren et al., 2022), and OD-RTE (Ning et al., 2023). In addition, we evaluate several LLM-based baselines, including GPT-4³ and Qwen2.5-7B⁴.

Pipeline SpeechRE Models. Pipeline SpeechRE models first transcribe speech into text with ASR and then apply TextRE models for relation triplet extraction. Following Ning et al. (2026), we use a wav2vec-large ASR model for transcription and adopt the same TextRE models described above.

End-to-End SpeechRE Models. End-to-end (End2End) SpeechRE models directly extract relation triplets from input speech without relying on intermediate text transcription. We evaluate

²huggingface.co/Qwen/Qwen2-Audio-7B

³openai.com/api/docs/models/gpt-4

⁴huggingface.co/Qwen/Qwen2.5-7B

Task	Model	🔥 CoNLL04-SpeechRE			👤 Common Voice-SpeechRE			👤 CSRTE		
		Entity	Relation	Triplet	Entity	Relation	Triplet	Entity	Relation	Triplet
TextRE	GPT-4 (ICL)	61.36	62.67	28.83	57.33	38.32	15.35	68.90	48.26	30.11
	Qwen2.5-7B (SFT)	83.71	82.40	58.49	71.02	78.37	50.81	82.48	88.04	65.12
	TP-Linker	78.63	83.49	58.56	64.61	69.31	46.61	75.58	79.07	60.54
	Spert	76.38	81.83	63.45	66.34	70.82	47.26	77.61	81.18	61.92
	REBEL	85.36	89.86	71.46	71.32	74.32	49.81	82.94	84.11	64.47
	BiRTE	79.34	87.34	64.61	67.61	72.35	47.10	78.94	82.49	61.36
	OD-RTE	81.81	82.35	60.57	70.61	69.67	47.34	81.64	80.06	61.58
SpeechRE (Pipeline)	GPT-4 _{pipe} (ICL)	24.41	70.31	7.13	23.61	44.35	4.94	25.18	55.02	11.62
	Qwen2.5-7B _{pipe} (SFT)	33.70	76.78	12.92	33.01	66.93	7.94	28.10	75.12	15.46
	TP-Linker _{pipe}	35.21	78.21	9.76	31.06	64.13	7.61	26.48	71.53	14.24
	Spert _{pipe}	30.43	75.95	11.88	32.61	64.48	7.54	26.72	72.41	14.16
	REBEL _{pipe}	37.06	83.35	14.01	31.54	66.10	7.92	27.50	74.82	15.12
	BiRTE _{pipe}	31.94	76.61	12.34	32.56	65.12	7.64	26.10	73.02	14.22
	OD-RTE _{pipe}	34.36	79.23	9.83	31.67	65.15	7.56	25.36	72.88	14.08
SpeechRE (End2End)	GPT-4o-Audio (ICL)	31.21	59.57	5.64	29.33	31.70	3.12	35.12	40.26	8.72
	Qwen2-Audio (ICL)	36.74	16.31	2.31	31.16	14.92	0.85	42.41	25.10	6.72
	Qwen2-Audio (SFT)	48.12	81.85	34.73	46.80	60.19	24.69	50.87	75.18	29.27
	LNA-ED	19.13	56.32	11.12	27.61	38.51	6.01	39.14	49.06	16.37
	MCAM	40.66	77.61	22.71	45.34	50.34	15.71	46.92	61.02	20.88
	RPG-MoGe	50.21	79.64	24.67	47.20	53.48	18.29	51.83	68.30	25.94
	ACRE _{Qwen2.5-Omni-3B}	56.28	82.90	37.41	49.52	63.18	27.12	50.92	73.06	31.02
	ACRE _{Qwen2.5-Omni-7B}	57.84	85.73	40.31	54.35	67.41	29.02	53.74	77.08	33.31
	ACRE _{Qwen2-Audio-7B}	58.09	87.14	42.07	53.66	69.58	30.41	53.96	78.25	34.90

Table 2: Performance comparison of different models on SpeechRE datasets under character-level exact-match evaluation. Entity F1 evaluates entity extraction, Relation F1 evaluates relation-type prediction without exact entity-pair matching, and Triplet F1 requires the complete relational triplet to match the gold annotation. ICL denotes in-context learning. Pipeline methods first transcribe speech with ASR and then perform RE, whereas End2End methods directly extract relational triplets from speech (without an intermediate transcription process).

general-purpose speech LLMs, including GPT-4o-audio⁵, Qwen2-Audio⁶. We also compare with representative SpeechRE methods, including LNA-ED(Wu et al., 2022), MCAM(Zhang et al., 2024), and RPG-MoGe(Ning et al., 2026). Finally, we report the results of our proposed ACRE framework.

5.5 Main Results

To evaluate the effectiveness of ACRE, we compare it with a broad range of strong baselines, including TextRE models, pipeline SpeechRE systems, and End2End SpeechRE models. As shown in Table 2, ACRE achieves the best performance among End2End SpeechRE methods on both Chinese and English datasets, and also consistently outperforms pipeline SpeechRE baselines. We summarize the main observations as follows.

(1) Compared with conventional end-to-end SpeechRE models, ACRE achieves consistent improvements across all datasets. This demonstrates the effectiveness of our acoustic-semantic reasoning framework. By explicitly incorporating pronun-

ciation cues and contextual knowledge into structured reasoning, ACRE provides a more effective solution for SpeechRE.

(2) Compared with LLM-based End2End baselines under both ICL and SFT settings, ACRE achieves better results. This shows that the improvement does not simply come from using a large backbone. In addition, SFT substantially improves relation prediction and triplet generation over ICL, indicating that task-specific supervision is crucial for structured SpeechRE.

(3) Among different backbones, the Qwen2.5-Omni-7B version of ACRE outperforms its 3B counterpart, showing the benefit of larger model capacity. The Qwen2-Audio-7B version achieves the best results, possibly because Qwen2-Audio is more specialized for audio-language understanding, whereas Omni models need to align additional modalities and may incur an extra “alignment tax” on speech-centric tasks.

(4) ACRE also consistently surpasses pipeline SpeechRE methods. The gap mainly comes from entity extraction rather than relation classification. This indicates that ASR errors mainly hurt entity

⁵openai.com/api/docs/models/gpt-4o-audio-preview

⁶huggingface.co/Qwen/Qwen2-Audio-7B

Task	Model	Metric	CoNLL04		CSRTE	
			Entity	Triplet	Entity	Triplet
SpeechRE (Pipeline)	REBEL _{pipe}	Strict PAE	37.06	14.01	27.50	15.12
			41.77	16.85	33.71	20.39
SpeechRE (End2End)	ACRE	Strict PAE	58.09	42.07	53.96	34.90
			63.91	45.61	62.45	42.26

Table 3: Comparison between character-level exact-match evaluation and PAE on CoNLL04 and CSRTE.

Model	CoNLL04			CSRTE		
	Entity	Relation	Triplet	Entity	Relation	Triplet
ACRE	58.09	87.14	42.07	53.96	78.25	34.90
w/o ARPC	55.21	85.63	39.74	51.08	75.92	31.76
w/o CGSC	56.42	82.87	37.61	52.18	76.86	31.54
w/o SAS-CoT	51.83	80.32	34.96	48.34	70.26	28.91
w/o GRPO	52.74	82.76	36.28	51.71	75.84	30.18

Table 4: Ablation study of ACRE on CoNLL04 and CSRTE. “w/o” denotes removing the corresponding component from the full model.

prediction, which further limits triplet extraction, especially in Chinese SpeechRE.

(5) Compared with TextRE models, ACRE is still less effective at extracting entities and complete triplets, but achieves more competitive results on relation prediction. This is expected because TextRE models take transcripts as input, where entity forms and boundaries are explicitly available. Relation types, however, must be inferred from input in both settings, which explains why ACRE achieves relation F1 comparable to TextRE models.

5.6 Results from the PAE Perspective

Table 3 compares character-level exact-match evaluation with the proposed PAE metric. Models obtain higher entity and RTE F1 scores under PAE, suggesting that exact matching may underestimate SpeechRE performance. PAE does not change model predictions, but evaluates them from a pronunciation-aware perspective.

The improvement is more evident on CSRTE than on CoNLL04. This is because Chinese contains many homophones, where the same pronunciation may correspond to different written forms, making some predictions rejected by strict matching still reasonable from a pronunciation perspective. The smaller gain on CoNLL04 suggests that exact matching can also introduce bias in English SpeechRE. Thus, PAE provides a complementary view, especially for Chinese SpeechRE.

5.7 Ablation Study

We conduct ablation experiments on CoNLL04 and CSRTE to analyze the contribution of each compo-

nent in ACRE. The results are shown in Table 4.

(1) Removing SAS-CoT causes the largest performance drop on CSRTE, indicating that the structured acoustic-semantic reasoning process plays a central role in SpeechRE.

(2) Removing GRPO also leads to clear degradation on CSRTE. This suggests that reward-guided optimization further aligns the generated reasoning process with the final relation extraction objective.

(3) For individual SAS-CoT components, removing ARPC mainly hurts entity extraction, suggesting that pronunciation cues help identify entities from speech. Removing CGSC causes a notable drop in Triplet F1, showing that contextual semantic knowledge provides useful evidence for the SpeechRE task.

6 Conclusion

In this paper, we address three key challenges in SpeechRE: the lack of Chinese benchmarks, insufficient use of speech-specific cues, and evaluation bias from exact matching. We construct CSRTE, a real-world Chinese SpeechRE benchmark, and propose ACRE, which integrates pronunciation and contextual semantics into structured acoustic-semantic reasoning. We further introduce PAE for pronunciation-aware entity matching. Experiments on Chinese and English datasets demonstrate the effectiveness of ACRE.

Acknowledgments

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (NSFC) [Grant No. 62506058] and the Young Teachers Science and Technology Innovation Project of the Fundamental Research Funds for the Central Universities [Grant No. 3132026362].

Limitations

The limitations of our work mainly include the following two aspects: 1) Although CSRTE is constructed from real human speech, it does not fully cover more complex speech scenarios, such as multi-speaker conversations, strong dialectal variations, code-switching, and extremely noisy environments. 2) ACRE mainly explores pronunciation for SpeechRE, while other speech-specific cues, such as pause boundaries, emotion, speaking style, and fine-grained prosodic variations, are not fully exploited, which may further improve the overall performance of our model.

References

- Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738. Association for Computational Linguistics.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.
- Siddhant Arora, Ankita Pasad, Chung-Ming Chien, Jionghao Han, Roshan Sharma, Jee-weon Jung, Hira Dharmyal, William Chen, Suwon Shon, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2024. On the evaluation of speech foundation models for spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11923–11938, Bangkok, Thailand. Association for Computational Linguistics.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. 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)*, pages 1–5. IEEE.
- Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 2370–2381.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *ArXiv*, abs/2407.10759.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.
- Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.
- George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*. European Language Resources Association (ELRA).
- Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, pages 2006–2013. IOS Press.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Zhenqi Jia, Yuan Zhao, Rui Liu, and Haizhou Li. 2026. Auemochat: Authentic emotion understanding and rendering for conversational speech synthesis. *arXiv preprint arXiv:2607.15755*.
- Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan. 2015. Spoken content retrieval: beyond cascading speech recognition with text retrieval. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(9):1389–1420.
- Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michele Tadmor Ramanovich. 2024. Spoken question answering and speech continuation using spectrogram-powered llm. In *International Conference on Learning Representations*, volume 2024, pages 51883–51898.
- Jinzhong Ning, Wenxuan Mu, Songtao Li, Yijia Zhang, Ling Luo, Yuanyuan Sun, Mingyu Lu, and Hongfei Lin. 2025a. Cner-omni: A unified dynamic modality learning framework for chinese named entity recognition across text and speech. *Neural Networks*, page 108410.
- Jinzhong Ning, Yuanyuan Sun, Bo Xu, Zhihao Yang, Ling Luo, and Hongfei Lin. 2024. Breaking the boundaries: A unified framework for chinese named entity recognition across text and speech. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1250–1260.
- Jinzhong Ning, Yuanyuan Sun, Zhihao Yang, Zhijun Wang, Ling Luo, Hongfei Lin, and Yijia Zhang. 2025b. Genen-mner: Enhancing nested chinese ner with multimodal fusion and alignment via speech-to-text generation. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1628–1640.

- Jinzhong Ning, Paerhati Tulajiang, Yingying Le, Yijia Zhang, Yuanyuan Sun, Hongfei Lin, and Haifeng Liu. 2026. Rpg-moge: Relation prompt-guided multi-order generative ensemble framework for speech relation extraction. *IEEE Transactions on Audio, Speech and Language Processing*.
- Jinzhong Ning, Zhihao Yang, Yuanyuan Sun, Zhizheng Wang, and Hongfei Lin. 2023. Od-rte: A one-stage object detection framework for relational triple extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11120–11135.
- OpenAI. 2024. Gpt-4o system card. *CoRR*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Feiliang Ren, Longhui Zhang, Xiaofeng Zhao, Shujuan Yin, Shilei Liu, and Bochao Li. 2022. A simple but effective bidirectional framework for relational triple extraction. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 824–832.
- Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. Interspeech 2019*, pages 3465–3469.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pages 16607–16629.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2025. [Label Studio: Data labeling software](https://github.com/HumanSignal/label-studio). Open source software available from <https://github.com/HumanSignal/label-studio>.
- Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus.
- 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 the 28th international conference on computational linguistics*, pages 1572–1582.
- Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2022. *OntoNotes Release 5.0*. Borealis.
- 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 the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10751–10762.
- LLM-Core Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners. *CoRR*.
- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.
- Liang Zhang, Zhen Yang, Biao Fu, Ziyao Lu, Liangying Shao, Shiyu Liu, Fandong Meng, Jie Zhou, Xiaoli Wang, and Jinsong Su. 2024. Multi-level cross-modal alignment for speech relation extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11975–11986.
- Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.*, 56(11).