RPG-MoGe: Relation Prompt-Guided Multi-Order Generative Ensemble Framework for Speech Relation Extraction

Anonymous ACL submission

Abstract

Speech Relation Extraction (SpeechRE) aims to extract relation triplets directly from speech data. However, existing datasets suffer from limited quantity and diversity of real-human speech in their training sets, while current models are constrained by fixed single-order generation templates and a lack of high-level semantic alignment, significantly hindering their performance. To address these challenges, we introduce **CommonVoice-SpeechRE**, a large-scale dataset comprising nearly 20,000 real-human speech samples from diverse speakers, establishing a new benchmark for SpeechRE research. Furthermore, we propose the **R**elation **P**rompt-**G**uided **M**ulti-**O**rder **G**enerative **E**nsemble (**RPG-MoGe**), a novel framework that features: (1) a multi-order triplet generation ensemble strategy, leveraging data diversity through diverse element orders during both training and inference, and (2) CNN-based latent relation prediction heads that generate explicit relation prompts to guide cross-modal alignment and accurate triplet generation. Extensive experiments demonstrate the superiority of our framework, outperforming state-of-the-art baselines. Our work not only provides a valuable dataset resource for the community but also offers an effective methodology to advance SpeechRE in real-world applications.

1 Introduction

Relation Extraction (RE), a fundamental task in information extraction, aims to extract structured knowledge in the form of relational triples (head entity, relation, tail entity) from unstructured data. RE plays a pivotal role in downstream applications such as knowledge graph construction and search engine optimization (Nasar et al., 2021). Despite its importance, most existing research focuses on **TextRE**, which extracts relational triples solely from plain text (Eberts and Ulges, 2020; Wang et al., 2020; Cabot and Navigli, 2021).

Dataset	CoNLL04	ReTACRED	Ours
#Rel.	5	40	45
#Train Sam.	922	33,477	14,557
#Dev Sam.	231	9,350	2,495
#Test Sam.	288	5,805	2,494
#Speaker	4	8	∼20,000

Table 1: Comparison of Key Statistics between existing datasets and the dataset proposed in this paper ("#Rel": Number of Relations; "Sam.": Samples; : Indicates samples with real-human speech; : Indicates samples with TTS synthetic speech)

However, with the exponential growth of speech data from sources such as news broadcasts, online meetings, and social media, there is a pressing need to extend RE to the speech domain. Speech data contains rich structured knowledge that can enhance knowledge graphs and support speech-related applications. This has led to the emergence of **Speech Relation Extraction (SpeechRE)**, a task that directly extracts relational triples from audio recordings.

Overall, SpeechRE is a relatively new research topic and remains underexplored. However, two notable works, LNA-ED (Wu et al., 2022) and MCAM (Zhang et al., 2024), have already made significant contributions. Wu et al. (2022) introduced the SpeechRE task by applying text-to-speech (TTS) to TextRE datasets, creating two synthetic speech benchmarks. They also provided the first SpeechRE baseline, LNA-ED, which uses a CNN-based length adapter to bridge a speech encoder and text decoder. Building on this, Zhang et al. (2024) developed two real-human-speech SpeechRE datasets and proposed MCAM, a more powerful model that employs a Multi-Level Cross-Modal Alignment Adapter to align tokens, entities, and sentences across speech and text.

Despite these advancements, existing approaches suffer from several limitations: (1) **Issue-1**: In their datasets, real-human speech data mainly

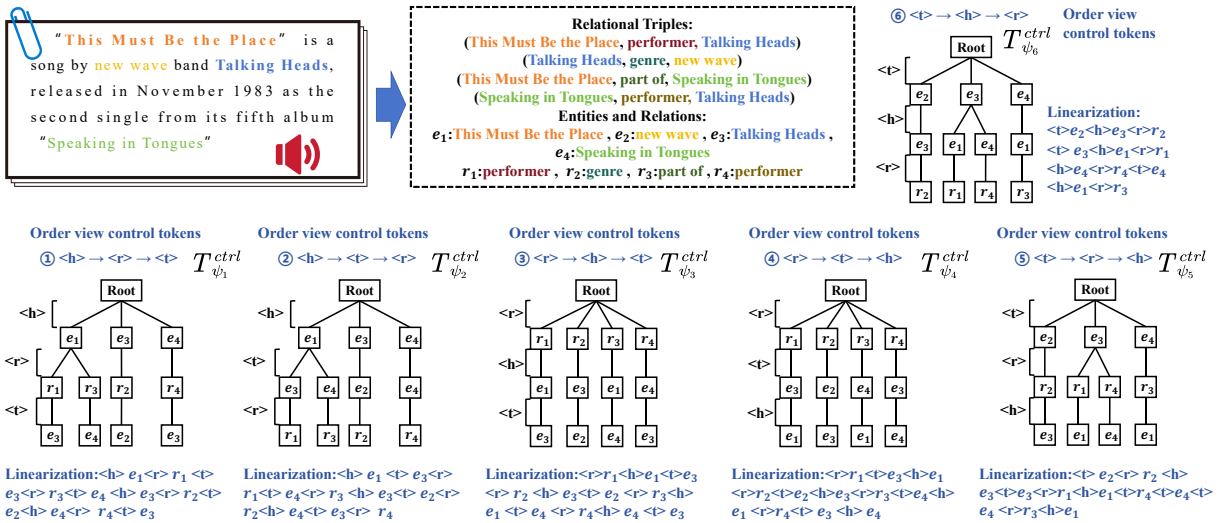


Figure 1: Explanation of the multi-view relation tree and its linearization process. Here, "<h>", "<r>", and "<t>" are special tokens representing the head entity, relation type, and tail entity of the relational triple respectively.

covers the test set, leaving limited training examples with few speakers (see Table 1). This may reduce the model's performance and generalization in real-world scenarios. (2) **Issue-2**: Current methods generate relational triples in a fixed order, ignoring the inherent diversity in the order of triple elements within the data. This restricts the model's ability to fully exploit the data. (3) **Issue-3**: Existing approaches primarily rely on semantic similarity for cross-modal alignment, overlooking high-level structured semantic cues such as entity relations.

To address these challenges, we propose a comprehensive solution that encompasses both data and model innovations.

For the data limitation (**Issue-1**), we introduce **CommonVoice-SpeechRE**, a newly annotated dataset comprising nearly 20,000 real speech recordings from diverse speakers. This dataset significantly expands the variety of speaker profiles and scenarios available for training (see Table 1).

For the model, we propose the **R**elation **P**rompt-**G**uided **M**ulti-**O**rder **G**enerative **E**nsemble (**RPG-MoGe**) framework. Specifically: (1) To mitigate the inherent limitation of fixed order in triplet generation templates (**Issue-2**), we introduce an innovative multi-view relation tree structure (depicted in Figure 1) to comprehensively capture the diverse ordering patterns of triplet elements. By linearizing these trees as generation targets, our model implements a multi-order triplet generation ensemble strategy during both training and inference phases, thereby fully exploiting the data's inherent diversity potential. (2) To alleviate the **Issue-3**, we design a CNN-based latent relation prediction head that identifies latent relations in the speech signal. These relations are used to construct explicit relation prompts, guiding the text decoder to generate relational triples and align speech and text modalities more effectively.

Our contributions can be summarized as follows:

- We present CommonVoice-SpeechRE, a large-scale, diverse real-human-speech dataset that sets a new benchmark for SpeechRE research.
- We propose RPG-MoGe, a novel framework that integrates multi-order triple generation and explicit relation prompts to fully exploit data diversity and high-level semantic cues.
- Extensive experiments on multiple SpeechRE benchmarks show that our approach outperforms state-of-the-art baselines, validating the effectiveness of our dataset and model design.

2 Related Work

2.1 Speech Relation Extraction

Speech Relation Extraction (SpeechRE) is a critical yet underexplored task in Information Extraction (IE) and Spoken Language Understanding (SLU) (Shon et al., 2022). While Speech Named Entity Recognition (Speech NER), an important subtask in both SLU and IE, has seen significant progress (Yadav et al., 2020; Ghannay et al., 2018; Chen et al., 2022), SpeechRE remains nascent, with limited advancements in datasets and models. Two

key contributions have shaped this field. Wu et al. (2022) introduced the SpeechRE task by converting TextRE datasets into synthetic speech using a text-to-speech (TTS) system, creating two benchmark datasets. They also proposed the LNA-ED model, which connects a speech encoder and text decoder via a CNN-based length adapter. Later, Zhang et al. (2024) advanced the field by constructing a dataset with real human speech and introducing the MCAM model, which employs a Multi-Level Cross-Modal Alignment Adapter to align speech and text across tokens, entities and sentences.

2.2 Multi-view Prompt Text Generation

Recent work in aspect-based sentiment analysis has shown that leveraging element order diversity in triples (Gou et al., 2023) or quadruples (Bai et al., 2024) during training and inference can enhance model performance and generalization. Inspired by this, we are the first to explore the impact of element order diversity in relational triplets on model performance in SpeechRE, a cross-modal text generation task involving both speech and text. This approach distinguishes our work from prior research and opens new avenues for improving SpeechRE through structured data diversity.

3 The New Dataset

We present CommonVoice-SpeechRE, a novel dataset derived from the English subset of the Common Voice 17.0 corpus (Ardila et al., 2020). Common Voice 17.0 is a large-scale, multilingual speech dataset comprising 20,408 validated hours of recordings across 124 languages, contributed by volunteers globally. Released under the CC-0 license, it permits unrestricted use, modification, and redistribution, making it an ideal foundation for secondary annotation tasks such as Speech Relation Extraction (SpeechRE).

Most samples in Common Voice 17.0 are negative examples lacking entities or relations. To identify potential positive samples, we employed a pre-trained BERT NER tagger[1] to analyze transcriptions and filter relevant data. We adopted entity and relation type definitions from the ACE04 and ACE05 datasets, crafting a tailored annotation guide. A team of 10 graduate students (all CET-6 certified) manually labeled approximately 20,000 transcriptions using Label Studio[2]. The annotation process involved dividing the data into batches of no more than 1,000 sentences, with 10% randomly selected for verification. Experienced annotators ensured sentence-level accuracy exceeded 95%; otherwise, the batch was re-annotated.

Detailed statistics of the CommonVoice-SpeechRE dataset are provided in the appendix due to page constraints. Sample data and annotation guidelines can be found in the supplementary material.

4 Methodology

In this section, we formally define the Speech Relation Extraction (SpeechRE) task and present the detailed implementation of our proposed RPG-MoGe framework.

4.1 Task Definition

Given a speech signal $\boldsymbol{S}$, the SpeechRE task aims to directly extract a set of relational triples $\Gamma = \{(h_i, r_i, t_i) \mid h_i, t_i \in \mathrm{E}, r_i \in \mathrm{R}\}$ from the speech signal, where E denotes the set of entities in the speech transcript, and R represents the set of predefined relations.

4.2 Details of the RPG-MoGe Framework

The ERP-MoGe framework consists of three core modules: a Speech Encoder, a Latent Relation Prediction Head, and a Text Decoder. The detailed structure is illustrated in Figure 2.

4.2.1 Speech Encoder

Given an input raw speech signal $\boldsymbol{S}$, we first convert it into log-mel spectrogram features $\boldsymbol{X}$. Subsequently, the features $\boldsymbol{X}$ are fed into the Whisper speech encoder (Radford et al., 2023) to extract high-level speech features $\boldsymbol{H}$ of the speech:

$$\boldsymbol{H} = \mathrm{WhisperEncoder}(\boldsymbol{X}) \in \mathbb{R}^{L_H \times d_h} \quad (1)$$

where $\mathrm{WhisperEncoder}(\cdot)$ represents the encoding operation of the Whisper encoder model, L_H and d_h are sequence length and dimension of speech features $\boldsymbol{H}$.

4.2.2 Latet Relation Prediction Head

The Latent Relation Prediction Head (LRPH) is designed to leverage semantic entity-relation cues by predicting latent relations in the speech signal. It consists of the following steps:

1.**CNN Layers**: We pass $\boldsymbol{H}$ through four CNN layers with ReLU activation to capture local patterns:

$$\boldsymbol{H}_{\mathrm{cnn}} = \mathrm{Conv}_4(\boldsymbol{H}) \quad (2)$$

[1] https://huggingface.co/flair/ner-english-ontonotes
[2] https://labelstud.io

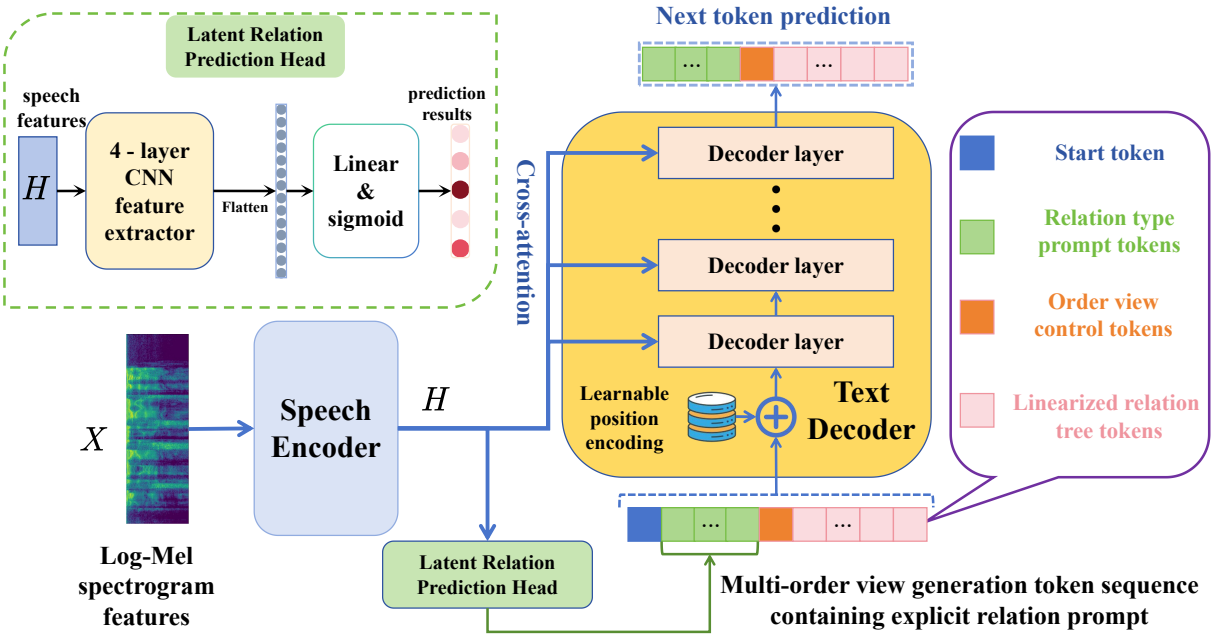


Figure 2: The overall architecture of RPG-MoGe.

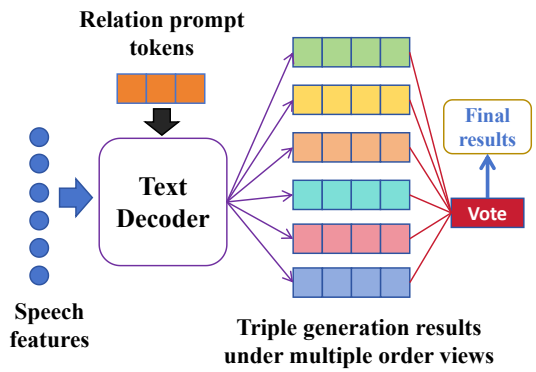


Figure 3: Implementation details for the Inference Phase in RPG-MoGe.

2.**Flattening and Linear Transformation**: The CNN output is flattened and fed into a linear layer to compute relation prediction scores:

$$\boldsymbol{H}_{\text{flat}} = \text{Flatten}(\boldsymbol{H}_{\text{cnn}}) \tag{3}$$

$$\boldsymbol{score}^{(R)} = \sigma\left(\boldsymbol{W}_{\text{lrp}}\boldsymbol{H}_{\text{flat}} + \boldsymbol{b}_{\text{lrp}}\right) \tag{4}$$

where σ is the sigmoid function, $\boldsymbol{W}_{\text{lrp}} \in \mathbb{R}^{|\text{R}| \times d_h}$ and $\boldsymbol{b}_{\text{lrp}} \in \mathbb{R}^{|\text{R}|}$ are learnable parameters, and $\boldsymbol{score}^{(R)} \in \mathbb{R}^{|\text{R}|}$ represents the scores for all predefined relation types.

3.**Loss Function**: We employ the Binary Cross Entropy (BCE) loss for training the LRPH module:

$$\mathcal{L}_{\text{lrp}} = -\frac{1}{|\text{R}|}\sum_{i=1}^{|\text{R}|}\Big[\text{y}_i^{(R)}\log(\text{score}_i^{(R)}) + (1-\text{y}_i^{(R)})\log(1-\text{score}_i^{(R)})\Big] \tag{5}$$

where $\text{y}^{(R)}$ denotes the ground-truth relation labels. Since each sample may contain multiple relations, this prediction task is a multi-label classification problem. In $\text{y}^{(R)}$, each element $\text{y}_i^{(R)}$ can be either 0 or 1, indicating the absence or presence of the i-th relation type, respectively. This approach enables the model to predict multiple relations simultaneously for each given input.

4.2.3 Multi-view Relation Tree and Linearization

To model the diversity introduced by permutations of triplet element orders, we propose the Multi-view Relation Tree structure. As depicted in Figure 1, each tree consists of four layers, with each layer (excluding the first) corresponding to an element of the triple. For a given sample, we can generate $P(3,3) = 6$ distinct relation trees by permuting the order of triplet elements.

Formally, for a speech signal $\boldsymbol{S}$ with a set of relation triplets $\mathcal{T}$, we apply the Treeify$(\cdot,\cdot)$ function to construct a relation tree $\mathcal{G}_{\psi_i}$ from a specific order perspective ψ_i:

$$\mathcal{G}_{\psi_i} = \text{Treeify}(\mathcal{T}, \psi_i) \tag{6}$$

where $\psi_i \in \Psi$ represents an order perspective, and Ψ encompasses all six possible order perspectives.

The relation tree $\mathcal{G}_{\psi_i}$ is then linearized into a token sequence using the SeqLin$(\cdot)$ operation:

$$T_{lin}^{\psi_i} = \text{SeqLin}(\mathcal{G}_{\psi_i}) \tag{7}$$

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
ReTACRED-SpeechRE	40	33,477	9,350	5,805	58,465	19,584	13,418	12.9s
CommonVoice-SpeechRE	45	14,557	2,495	2,494	15,948	2,696	2,728	11.6s

Table 2: Dataset statistics. : TTS-synthesized speech; : real human speech. ReTACRED-SpeechRE enumerates all entity pairs as triplets, including "no_relation" type, while the other two datasets only contain positive triplets.

4.2.4 Text Decoder

The Text Decoder uses relation prompts and multi-order triplet generation to decode relational triplets. We utilize the pre-trained Whisper decoder (Radford et al., 2023) for this purpose. The input token sequence to the decoder consists of three parts:

1.**Relation type prompt tokens**: $\boldsymbol{T}_{rel} = [t_1^{rel}, \ldots, t_n^{rel}]$, where t_i^{rel} are special tokens representing the predicted relation types generated by the Latent Relation Prediction Head. These tokens guide the decoder by incorporating latent relational cues from speech.

2.**Order view control tokens**: $\boldsymbol{T}_{\psi_i}^{ctrl} = \text{permute}\left([\langle h \rangle, \langle r \rangle, \langle t \rangle], \psi_i\right)$, which specify the order of special tokens $\langle h \rangle, \langle r \rangle, \langle t \rangle$ for a given perspective ψ_i, as illustrated in Figure 1.

3. **Linearized relation tree tokens**: $\boldsymbol{T}_{\psi_i}^{lin}$, which represent the linearized token sequence of the relation tree. This component encodes the hierarchical structure of the relation tree into a sequential format suitable for the decoder.

These components are concatenated into the decoder input sequence $\boldsymbol{T}_{dec} = [\boldsymbol{T}_{rel}, \boldsymbol{T}_{\psi_i}^{ctrl}, \boldsymbol{T}_{\psi_i}^{lin}]$. At the i-th decoding step, the probability distribution $\boldsymbol{p}_{t_i^{dec}}$ of the output token t_i^{dec} is computed as:

$$\boldsymbol{h}_{t_i^{dec}} = \text{WhisperDecoder}\left(\boldsymbol{H}, \boldsymbol{T}_{dec}^{<i}\right) \quad (8)$$

$$\boldsymbol{p}_{t_i^{dec}} = \text{Softmax}\left(\boldsymbol{W}_{lm}\boldsymbol{h}_{t_i^{dec}} + \boldsymbol{b}_{lm}\right) \quad (9)$$

where $\boldsymbol{h}_{t_i^{dec}}$ is the hidden state, and $\boldsymbol{W}_{lm}$, $\boldsymbol{b}_{lm}$ are learnable parameters.

The decoder is trained using the Cross-Entropy Loss:

$$\mathcal{L}_{dec} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|V|}\boldsymbol{y}_{t_i^{dec}}[j]\log(\boldsymbol{p}_{t_i^{dec}}[j]) \quad (10)$$

where N is the sequence length, $|V|$ is the vocabulary size, and $\boldsymbol{y}_{t_i^{dec}}$ is the token label at the i-th decoding step.

4.2.5 Training and Inference Strategies

During training, each sample is expanded into multiple generation targets corresponding to all possible order views for participation in training. The total loss combines the $\mathcal{L}_{\text{lrp}}$ and $\mathcal{L}_{dec}$:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{lrp}} + \mathcal{L}_{dec} \quad (11)$$

During inference, as illustrated in Figure 3, the text decoder takes the speech features $\boldsymbol{H}$ and relation prompt tokens $\boldsymbol{T}_{rel}$ as initial inputs. By varying the order view control tokens, the decoder autoregressively generates triplets under all order views. A triplet is included in the final results if it appears in more than λ_{vote} order views.

5 Experiments

5.1 Datasets & Evaluation Metrics

We conducted experiments on three datasets: CoNLL04-SpeechRE, ReTACRED-SpeechRE and the CommonVoice-SpeechRE dataset proposed in this paper. The CommonVoice-SpeechRE dataset includes diverse real human speech in its training, development, and test sets. For CoNLL04-SpeechRE and ReTACRED-SpeechRE, since the real human speech test set and partial real human speech training set proposed by Zhang et al. (2024) have not yet been released, we used the fully TTS-generated speech version released by Wu et al. (2022). Detailed statistics of the datasets are provided in Table 2. For evaluation metrics, following previous work (Wu et al., 2022; Zhang et al., 2024), we used the micro-F1 score to assess the performance of models in entity recognition, relation prediction, and relation triplet extraction. For an entity, relation or triple to be considered correct, it must exactly match its counterpart in the ground truth tags.

5.2 Experimental Settings

Our model was implemented using PyTorch-Lightning[3] and PyTorch (Paszke et al., 2019), with

[3] https://github.com/Lightning-AI/pytorch-lightning

	Model	External Resources	CoNLL04-SpeechRE			ReTACRED-SpeechRE			CommonVoice-SpeechRE		
			Entity	Relation	Triplet	Entity	Relation	Triplet	Entity	Relation	Triplet
TextRE	GPT-3.5(LLM)	-	58.74	49.45	22.27	40.46	17.63	3.22	53.74	28.41	10.73
	GPT-4(LLM)	-	61.36	62.67	28.83	47.4	39.12	9.12	57.33	38.32	15.35
	TP-Linker	-	78.63	83.49	58.56	50.46	51.83	20.39	64.61	69.31	46.61
	Spert	-	76.38	81.83	63.45	60.26	63.48	21.46	66.34	70.82	47.26
	REBEL	-	85.36	89.86	71.46	60.09	65.15	25.15	71.32	74.32	49.81
SpeechRE (Pipline)	GPT-3.5$_{pipe}$(LLM)	-	28.21	69.61	6.31	16.61	43.84	1.32	21.30	46.81	3.34
	GPT-4$_{pipe}$(LLM)	-	29.41	70.31	7.13	19.76	46.31	4.23	23.61	44.35	4.94
	TP-Linker$_{pipe}$	-	35.21	78.21	9.76	30.27	50.01	6.59	31.06	64.13	7.61
	Spert$_{pipe}$	-	30.43	75.95	11.88	34.36	57.17	6.89	32.61	64.48	7.54
	REBEL$_{pipe}$	-	37.06	83.35	14.01	32.07	51.97	6.49	31.54	66.10	7.92
SpeechRE (End2End)	GPT-4o-audio(LLM)	-	31.21	59.57	5.64	13.21	41.61	1.14	29.33	31.70	3.12
	Qwen2-audio(LLM)	-	36.74	16.31	2.31	10.50	23.61	0.31	31.16	14.92	0.85
	LNA-ED(520M)$_{ori}$	PL-FT	18.87	55.66	10.41	17.21	43.37	3.20	26.34	37.31	5.37
	LNA-ED(770M)$_{whi}$	-	19.13	56.32	11.12	18.26	43.15	3.67	27.61	38.51	6.01
	MCAM(520M)$_{ori}$	ASR-PTC	40.13	77.89	22.07	35.34	58.96	8.07	43.94	48.37	14.96
	MCAM(770M)$_{whi}$	-	40.66	77.61	22.71	35.61	**59.13**	8.21	45.34	50.34	15.71
	RPG-MoGe(250M)$_{whi}$	-	43.16	76.91	22.17	36.00	57.46	8.09	45.59	49.60	15.32
	RPG-MoGe(770M)$_{whi}$	-	**50.21**	**79.64**	**24.67**	**36.76**	58.38	**9.18**	**47.20**	**53.48**	**18.29**

Table 3: F1-score (%) comparison: RPG-MoGe versus baselines. Subscript $pipe$ denotes ASR+TextRE pipeline methods; 'PL-FT' indicates fine-tuning with pseudo-labeled data; 'ASR-PTC' refers to pre-training with ASR data. Subscript ori represents the original LNA-ED(Wu et al., 2022)/MCAM(Zhang et al., 2024) backbone: 24-layer Wave2vec encoder + 12-layer BART-large decoder (520M). Subscript whi denotes Whisper (Radford et al., 2023) backbones: 24-layer encoder/decoder (770M) or 12-layer encoder/decoder (250M).

OpenAI's Whisper[4] (Radford et al., 2023) as the backbone, specifically the whisper-small[5] (244M) and whisper-medium[6] (769M) versions. We optimized the model parameters using the Adam optimizer with a learning rate of 1e-5, a batch size of 12. Training epochs were set to 50 for CoNLL04-SpeechRE, 20 for ReTACRED-SpeechRE, and 10 for CommonVoice-SpeechRE. For the relation prediction head, we employed a four-layer CNN with 2D convolutions (kernel size = 3) and progressively increasing channel dimensions (16, 32, 64, 128). During inference, the voting threshold λ_{vote} for all order views was set to 2. All hyperparameters were tuned on the development set, and the best-performing checkpoint was selected for test set evaluation. Training was conducted on a single NVIDIA A40 GPU, while inference was performed on a single NVIDIA GeForce RTX 4090 GPU.

5.3 Baselines

To comprehensively evaluate the performance of our proposed model, we compare it with three categories of competitive baselines: (1) **TextRE Models.** These models are designed to jointly extract entities and relations from input text. For a fair comparison, following prior works (Wu et al., 2022; Zhang et al., 2024), we adopt three state-of-the-art TextRE models: TP-Linker (Wang et al., 2020), Spert (Eberts and Ulges, 2020), and REBEL (Cabot and Navigli, 2021). Additionally, to explore the potential of large language models (LLMs) in relation extraction, we include GPT-3.5[7] and GPT-4[8] as baselines, leveraging their in-context learning capabilities for TextRE tasks. (2) **Pipeline SpeechRE Models.** These models follow a two-stage pipeline: first, an Automatic Speech Recognition (ASR) module transcribes the input speech into text; second, a TextRE module extracts relation triplets from the transcribed text. To ensure a fair comparison, we follow the setup of prior works (Wu et al., 2022; Zhang et al., 2024) and employ the pre-trained wav2vec-large model as the ASR module. For the TextRE module, we use the same five TextRE models mentioned above, resulting in five pipeline models: TP-Linker$_{pipe}$, Spert$_{pipe}$, REBEL$_{pipe}$, GPT-3.5$_{pipe}$, and GPT-4$_{pipe}$. (3) **End-to-End SpeechRE Models.** These models are designed to directly extract relation triplets from input speech, without the intermediate step of text transcription. Our proposed RPG-MoGe also falls into this category. As baselines, we include two existing end-to-end SpeechRE models: LNA-ED (Wu et al., 2022) and MCAM (Zhang et al., 2024). Additionally, to explore the capabilities of recent advancements in speech-based LLMs, we introduce two in-context learning baselines: GPT-

[4]Whisper has become a standard backbone in speech processing, similar to BERT and BART in NLP.

[5]https://huggingface.co/openai/whisper-small.en

[6]https://huggingface.co/openai/whisper-medium.en

[7]gpt-3.5-turbo-0125

[8]gpt-4-turbo-2024-04-09

Model	CommonVoice-SpeechRE		
	Entity	Relation	Triplet
MCAM(520M)	43.94	48.37	14.96
RPG-MoGe(250M)	45.59	49.60	15.32
w/o RPG	44.61	48.06	14.89
w/o LRPH&RPG	43.86	48.36	14.61
w/o Moge(infer)	43.64	46.40	12.46
w/o Moge(infer&train)	40.68	45.94	11.15

Table 4: An ablation study of the RPG-MoGe(250M). 'RPG' denotes Relation Prompt Guide in text decoder; 'LRPH' refers to CNN-based Latent Relation Prediction Head; 'w/o MoGe (infer)' uses multi-order triplet generation in training and single-order in inference. 'w/o Moge(infer&train)' means single-order triplet generation in both training and inference.

Method	CommonVoice-SpeechRE		
	Entity	Relation	Triplet
Order#1	43.59	47.01	12.46
Order#2	42.76	44.32	11.73
Order#3	45.07	47.18	13.31
Order#4	43.80	44.50	12.01
Order#5	42.91	45.83	12.08
Order#6	43.70	47.45	12.56
Average	43.64	46.05	12.36
Ensemble	45.59(**+1.95**)	49.60(**+3.55**)	15.32(**+2.96**)

Table 5: Performance comparison of RPG-MoGe before and after voting ensemble of individual order view predictions.

4o-audio[9] and Qwen2-audio[10] (Chu et al., 2024).

5.4 Results and Analysis

5.4.1 Main Results

We conducted a comprehensive performance comparison between our proposed RPG-MoGe model and several strong baselines, including TextRE, SpeechRE (Pipeline), and SpeechRE (End2End). The experimental results, presented in Table 3, reveal the following key observations:

(1) RPG-MoGe outperforms all SpeechRE (End2End) baselines, achieving state-of-the-art performance in entity, relation, and triplet F1 scores across all datasets. Notably, RPG-MoGe with a 250M parameter Whisper backbone surpasses the SOTA baseline MCAM using a 520M backbone and matches MCAM's performance with a 770M backbone. This demonstrates RPG-MoGe's ability to leverage the diversity of relation triplet element orders and effectively utilize high-level semantic cues through its potential relation prediction head and explicit relation prompts.

(2) RPG-MoGe consistently outperforms all SpeechRE models in triplet extraction, highlighting the limitations of the pipeline approach, where cascading ASR with TextRE introduces significant errors. The end-to-end approach effectively mitigates error accumulation, improving entity, relation, and triplet extraction accuracy.

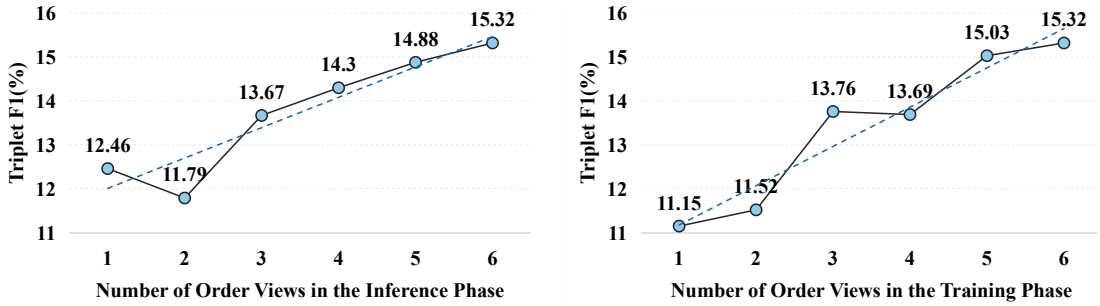
(3) Large language models without fine-tuning (e.g., GPT-3.5, GPT-4, GPT-4o-audio, Qwen2-audio) perform significantly worse on the datasets compared to fine-tuned smaller models, emphasizing the continued importance of developing fine-tuned models in TextRE and SpeechRE domains.

(4) Replacing the non-aligned Wave2vec and BART encoders in LNA-ED and MCAM with the pre-trained and aligned Whisper encoder and decoder eliminates the need for extensive external corpus alignment and improves performance. This also ensures a fairer comparison with RPG-MoGe, which utilizes Whisper as its backbone.

5.4.2 Ablation Study

To assess the contribution of different modules in RPG-MoGe, we conducted ablation experiments on the fully human-annotated CommonVoice-SpeechRE dataset (see Table 4). The results reveal two key insights: (1) Removing the Relation Prompt Guide (RPG) generated by the Latent Relation Prediction Head (LRPH) leads to performance degradation, demonstrating that LRPH enhances the model's ability to capture high-level semantic information (e.g., entity relations) from speech signals. Additionally, the RPG, derived from LRPH predictions, acts as a prompt for the text decoder, guiding the model to focus on potential relation types and enhancing speech-text alignment through high-level semantic information. (2) The strategy of integrating multi-order triplet generation significantly boosts model performance in both training and inference stages. As shown in the results, RPG-MoGe with a 250M-parameter backbone outperforms the state-of-the-art method MCAM, which uses a 520M-parameter backbone. This suggests that the effective information in speech data has not yet been fully exploited. For example, the diversity introduced by the order of relational triple elements, as utilized in this work, represents a potential avenue for further improving model performance.

[9] gpt-4o-audio-preview-2024-12-17
[10] Qwen2-Audio-7B-Instruct



(a) (b)

Figure 4: The impact of number of order views on RPG-MoGe performance.

5.4.3 Discussion on Multi-Order Triplet Generation Ensemble

To quantitatively assess the effectiveness of the Multi-Order Triplet Generation Ensemble strategy in RPG-MoGe, we conducted experiments to analyze the impact of the number of generation order views during training and inference (see Figure 4) and compared performance before and after ensemble across different order views (see Table 5). The results highlight two key findings: during training, increasing the number of generation order views significantly improves model performance, as it exposes the model to more diverse data, enhancing its ability to capture underlying patterns; during inference, ensembling predictions from multiple order views reduces individual model biases and errors, as predictions from different orders complement and correct each other, leading to higher overall accuracy.

5.4.4 Human vs. TTS Speech Data Analysis

To investigate the impact of using TTS-generated speech data for training and evaluating SpeechRE models, we synthesized TTS-based training and test sets for the CommonVoice-SpeechRE dataset using the same TTS tools applied to CoNLL04-SpeechRE and ReTACRED-SpeechRE. We conducted experiments comparing the performance of RPG-MoGe when trained and tested on human speech versus TTS-generated speech (see Table 6). The results reveal two key findings: (1) Models trained on TTS data exhibit significantly lower performance compared to those trained on human speech, indicating that human speech data better reflects real-world scenarios and yields more robust models; (2) Models trained on TTS data perform notably worse on human speech test sets than on TTS test sets, suggesting that TTS-generated data fails to accurately replicate real-world speech environments, leading to biased performance evaluation. These findings underscore the value of the fully human-annotated CommonVoice-SpeechRE dataset proposed in this work, which not only provides diverse, real-world training data for robust model development but also establishes a more reliable benchmark for evaluating SpeechRE performance in authentic settings.

Train \ Test		Speech (TTS)	Speech (Human)
Speech (TTS)	Entity	50.25	40.59
	Relation	53.88	42.13
	Triplet	18.78	12.17
Speech (Human)	Entity	41.91	45.59
	Relation	48.04	49.60
	Triplet	14.13	15.32

Table 6: Impact of TTS and Human Speech on Model Training and Performance Evaluation

Due to page constraints, the computational efficiency analysis of RPG-MoGe is provided in the appendix.

6 Conclusions

In this work, we address the limitations of existing datasets and models in Speech Relation Extraction (SpeechRE) by introducing CommonVoice-SpeechRE, a large-scale dataset with diverse real-human speech samples, and proposing RPG-MoGe, a novel framework that leverages a multi-order triplet generation ensemble strategy and CNN-based latent relation prediction heads to enhance triple generation and cross-modal alignment. Extensive experiments demonstrate the superiority of our approach, outperforming state-of-the-art baselines and setting a new benchmark for SpeechRE research. Our contributions provide both a valuable resource and an effective methodology, advancing the field toward real-world applications.

Limitations

While our work introduces significant advancements in Speech Relation Extraction (SpeechRE), it is not without limitations. The CommonVoice-SpeechRE dataset, despite its diversity, primarily focuses on English speech, which may restrict its generalizability to other languages. Additionally, the RPG-MoGe framework, though effective, relies on computationally intensive components such as multi-order ensemble strategies and cross-modal alignment, which could pose challenges for deployment in resource-constrained environments. Future work could explore multilingual extensions of the dataset and more efficient model architectures to further enhance the practicality and scalability of SpeechRE systems.

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Yinhao Bai, Yalan Xie, Xiaoyi Liu, Yuhua Zhao, Zhixin Han, Mengting Hu, Hang Gao, and Renhong Cheng. 2024. Bvsp: Broad-view soft prompting for few-shot aspect sentiment quad prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 8465–8482.

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.

Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang. 2022. Aishell-ner: Named entity recognition from chinese speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8352–8356. IEEE.

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 preprint arXiv:2407.10759*.

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.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1):1–39.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

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.

Tongtong Wu, Guitao Wang, Jinming Zhao, Zhaoran Liu, Guilin Qi, Yuan Fang Li, and Gholamreza Haffari. 2022. Towards relation extraction from speech. In *Empirical Methods in Natural Language Processing 2022*, pages 10751–10762. Association for Computing Machinery (ACM).

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4268–4272. ISCA.

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.

Method	Entity	Relation	Triplet	Speed(inference)
LNA-ED(770M)$_{whi}$	27.61	38.51	6.01	4.61sample/s
MCAM(770M)$_{whi}$	45.34	50.34	15.71	4.55sample/s
RPG-MoGe(250M)$_{whi}$	45.59	49.60	15.32	**4.68sample/s**
RPG-MoGe(770M)$_{whi}$	**47.20**	**53.48**	**18.29**	0.78sample/s

Table 7: Comparison of computational efficiency between RPG-MoGe and baseline methods. All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU with a batch size of 1.

Gender	Count
Male/Masculine	7598
Female/Feminine	5992
Unknown	5993

Table 8: Gender distribution of speech collectors in the dataset.

Age Group	Count
Unknown	5553
Twenties	5441
Thirties	2689
Forties	2052
Fifties	1263
Teens	1238
Sixties	1199
Seventies	118
Eighties	23
Nineties	7

Table 9: Age distribution of speech collectors in the dataset.

A Analysis on Model Efficiency

The computational efficiency of RPG-MoGe is evaluated against several baseline methods, including LNA-ED and MCAM, across key metrics such as entity, relation, and triplet F1 score (%), as well as inference speed. As shown in Table 7, RPG-MoGe demonstrates superior performance in terms of prediction accuracy. Specifically, the 770M parameter version of RPG-MoGe achieves the highest scores across all metrics, significantly outperforming both LNA-ED and MCAM. Notably, the 250M parameter version of RPG-MoGe also delivers competitive performance compared to the baseline methods.

However, during the inference phase, RPG-MoGe requires generating relational triplets from six different order views and ensembling them, which increases model complexity. This is reflected in the significantly lower inference speed of 0.78 samples per second for the 770M version. In contrast, RPG-MoGe (250M), due to its smaller parameter size, maintains the highest inference speed of 4.68 samples per second among all models. This trade-off between accuracy and efficiency underscores the importance of selecting the appropriate model size based on specific application requirements.

In summary, RPG-MoGe strikes a balance between computational efficiency and prediction accuracy. The 250M version offers a practical solution for scenarios prioritizing speed, while the 770M version delivers state-of-the-art performance for tasks where accuracy is paramount.

B Detailed Statistical Analysis of the Dataset

In this section, we present a comprehensive statistical analysis of the dataset, focusing on four key aspects: the distribution of relation triplet types, the gender distribution of speech collectors, the age distribution of speech collectors, and the accent distribution of speech collectors. These analyses provide valuable insights into the composition and diversity of the dataset.

B.1 Distribution of Relation Triplet Types

The distribution of relation triplet types in the dataset is presented in Table 10, which provides a detailed breakdown of the frequency and percentage of each triplet type. The dataset encompasses a wide variety of triplet types, reflecting the diversity and complexity of the relationships captured.

The most frequent triplet type is *Person:holds_title:Title*, which accounts for 3,333 instances (15.54% of the dataset). This is followed by *Person:affiliated_with:Organization* with 2,065 instances (9.63%) and *Location:located_in:Location* with 1,738 instances (8.11%). These triplet types dominate the dataset, indicating a strong focus on personal titles, organizational affiliations, and geographical relationships.

Other notable triplet types include *Organization:Establishes:Title* (1,233 instances, 5.75%), *Organization:located_in:Location* (1,163 instances, 5.42%), and *Person:creates:Work_of_art* (1,112 instances, 5.19%). These relationships highlight the dataset's coverage of organizational structures, geographical contexts, and creative works.

Less frequent triplet types, such as *Person:opposes:Regulation* (28 instances, 0.13%) and *Organization:violates_opposes:Regulation* (12 instances, 0.06%), represent more specialized or niche relationships. While these triplet types are underrepresented, they contribute to the dataset's richness and applicability to a broader range of tasks.

Overall, the distribution of relation triplet types demonstrates the dataset's comprehensive coverage of diverse relationships, ranging from common personal and organizational associations to more specialized interactions. This diversity is essential for training models capable of handling a wide array of real-world scenarios.

B.2 Gender Distribution of Speech Collectors

As shown in Table 8, the gender distribution of speech collectors is well-balanced, with 7,598 male/masculine speakers and 5,992 female/feminine speakers. This balanced representation ensures that the dataset is suitable for tasks requiring gender-neutral or gender-specific analysis, contributing to its overall robustness and fairness.

B.3 Age Distribution of Speech Collectors

The age distribution of speech collectors in the dataset, as shown in Table 9, exhibits a broad and diverse representation across various age groups, ranging from teens to individuals in their nineties. This diversity ensures the inclusion of a wide spectrum of vocal characteristics and speech patterns, enhancing the dataset's ability to support robust and generalizable SpeechRE models. The significant representation of younger and middle-aged speakers, complemented by meaningful contributions from older age groups, underscores the dataset's comprehensive coverage of speaker demographics, making it well-suited for real-world applications.

B.4 Diversity of Accents Among Speech Collectors

The distribution of accents among speech collectors, as shown in Table 11, demonstrates significant diversity. Among the labeled accents, "United States English" (4,967 instances) and "England English" (1,533) are the most prevalent, followed by "India and South Asia English" (1,453), "Canadian English" (1,032), and "German English, Non-native speaker" (896). Accents from regions such as Australia, Southern Africa, Scotland, and Northern Ireland are also well-represented, while accents from Singapore, Malaysia, and Hong Kong appear less frequently. Notably, some collectors exhibit multiple accent backgrounds, such as "United States English and England English" (40 instances) and "United States English and Transatlantic English" (31 instances). Overall, the dataset encompasses a wide range of English accents, reflecting the global diversity of English usage.

ID	Head_Entity	Relation	Tail_Entity	Quantity	Percentage
1	Person	holds_title	Title	3333	15.54%
2	Person	affiliated_with	Organization	2065	9.63%
3	Location	located_in	Location	1738	8.11%
4	Organization	Establishes	Title	1233	5.75%
5	Organization	located_in	Location	1163	5.42%
6	Person	creates	Work_of_art	1112	5.19%
7	Person	participates_in	Event	938	4.37%
8	Person	business/work	Person	778	3.63%
9	Person	family	Person	702	3.27%
10	Person	visits	Location	667	3.11%
11	Event	Occurs_at	Location	615	2.87%
12	Person	belongs_to_NORP	Organization	612	2.85%
13	Organization	Engagement	Event	548	2.56%
14	Person	resides_at	Location	548	2.56%
15	Person	performs	Work_of_art	539	2.51%
16	Organization	develops/produces/sells	Production	483	2.25%
17	Person	born_in	Location	404	1.88%
18	Person	is_character_in	Work_of_art	386	1.80%
19	Person	works_at	Location	385	1.80%
20	Person	owns/uses	Production	266	1.24%
21	Organization	is_subordinate_to	Organization	262	1.22%
22	Organization	publishes	Work_of_art	256	1.19%
23	Location	Adjacent	Location	232	1.08%
24	Person	died_at	Location	215	1.00%
25	Person	creates	Production	213	0.99%
26	Person	founds	Organization	205	0.96%
27	Event	Occurs_on	Date	185	0.86%
28	Person	competes_with	Person	142	0.66%
29	Person	coreference	Person	138	0.64%
30	Organization	manages/uses	Production	132	0.62%
31	Person	harms	Person	130	0.61%
32	Person	provides_services_to	Organization	117	0.55%
33	Organization	Organization	Event	106	0.49%
34	Person	leaves	Organization	91	0.42%
35	Person	leads_NORP	Organization	77	0.36%
36	Person	supports	Regulation	73	0.34%
37	Production	Originated_in	Location	57	0.27%
38	Person	critiques	Work_of_art	55	0.26%
39	Organization	collaborates_with	Organization	54	0.25%
40	Person	proposes	Regulation	47	0.22%
41	Organization	Complies_With	Regulation	43	0.20%
42	Organization	Legislative_Actions	Regulation	39	0.18%
43	Person	opposes	Regulation	28	0.13%
44	Person	Endorses	Production	18	0.08%
45	Organization	violates_opposes	Regulation	12	0.06%

Table 10: Distribution of relation triplet types: Frequency and percentage of each triplet type in the dataset.

Accent	Count
Unknown	7690
United States English	4967
England English	1533
India and South Asia	1453
Canadian English	1032
German English, Non native speaker	896
Australian English	627
Southern African	240
Scottish English	229
Northern Irish	143
Irish English	131
New Zealand English	112
Filipino	88
Liverpool English, Lancashire English	61
Singaporean English	48
England English, New Zealand English	47
United States English, England English	40
United States English, Transatlantic English	31
Hong Kong English	30
United States English, Midwestern, Low, Demure	21
Malaysian English	20
Welsh English	19
South African accent, Southern African	18
United States English, Scandinavian	14
Nepali	13
Academic southern English, England English	12
Southern United States, United States English	9
Northern Irish, Culchie	8
Southern United States, New Orleans dialect	8
United States English, Midwestern, Minnesotan	6
New Zealand English, England English	6
Polish	6
Southern Californian, United States English	5
United States English, Canadian English	5
West Indies and Bermuda	5
German	5
United States English, Midwestern	5

Table 11: Distribution of accents among speech collectors in the dataset.