

A POTENTIAL ETHICAL CONSIDERATIONS

We conduct all procedures for data collection and annotation by ethical principles and with the informed consent of the participants.

- **Privacy Concerns.** In developing the SpeechEE dataset, we prioritized privacy and ethical data use by meticulously ensuring that the combined datasets did not contain any personally identifiable or sensitive information. Our dataset, derived from existing textual EE datasets, represents a diverse array of news types, and we have made certain that all data used are legally sourced and devoid of any individual-specific details. Additionally, to further protect privacy, the dataset is publicly available, allowing for transparency and accountability in its use. These precautions are designed to prevent any potential misuse of personal information while fostering a secure environment for academic and applied research.

- **Annotator and Compensation.** Recognizing the crucial role of human annotators in developing our dataset, we employed highly skilled senior postgraduate students trained specifically for these tasks. The time required to annotate each segment of the dataset is substantial, typically ranging from 3 to 5 minutes per segment due to the complexity and precision required. To fairly compensate for their effort and expertise, annotators are paid 1 yuan (approximately \$0.15 USD) for each segment they annotate. Moreover, the compensation scheme for linguistics and computer science experts who contribute to our project is carefully calibrated to reflect their time investment and expertise, ensuring equitable remuneration. This approach underlines our commitment to ethical practices in compensating all contributors fairly for their labor and intellectual input.

- **Consent and Transparency.** In the creation of datasets, it is paramount that all contributors—whether their participation involves providing speech samples directly or indirectly—are fully informed about how their data will be used and have explicitly consented to it. Transparency about the data collection process and the intended use of the data is crucial to maintaining ethical standards.

- **Bias and Fairness.** Given that the dataset comprises diverse scenarios, languages, and speaker styles, it is important to systematically analyze and address any potential biases that may be present. These biases could manifest in the form of underrepresentation of certain languages, dialects, or demographic groups. Efforts should be made to ensure the SpeechEE system does not perpetuate or amplify existing biases.

- **Impact on Employment.** The automation of EE from speech could potentially impact jobs that traditionally rely on human transcription and analysis, such as secretarial or journalistic roles. It is important to consider the broader socio-economic implications of this technology and engage with affected stakeholders to explore supportive measures, such as retraining programs.

- **Misuse Potential.** Lastly, the capability to automatically extract structured information from speech could be misused in scenarios like surveillance without consent, eavesdropping, or other forms of intrusion into personal or confidential communications. Strong guidelines and possibly regulatory frameworks should be proposed to prevent misuse of this technology, ensuring it is used ethically and responsibly.

B EXTENDED MODEL IMPLEMENTATION AND SETTINGS

Here, we provide additional technical details about the model as an expansion of the main paper.

B.1 More Model Details on Pipeline SpeechEE Method

ASR Model. The whisper model is chosen as our ASR module for its high capability in learning speech features and outstanding ASR performance. In comparison to alternative ASR tools such as Wav2Vec 2.0 [2] and HuBERT [16], the whisper model undergoes training on a substantially labeled audio corpus (approximately 680,000 hours), a volume approximately ten times larger than the pre-training data (60,000 hours) utilized for unsupervised tasks like masked prediction in Wav2Vec 2.0. Consequently, it is capable of directly acquiring the mapping from speech to text, thereby exhibiting superior performance in speech recognition compared to other ASR models. Moreover, the whisper model integrates data from diverse languages and domains, aligning with the multilingual and varied domain characteristics of the SpeechEE dataset we have constructed.

The whisper model consists of three parts: an acoustic feature extraction module, a transformer encoder, and a decoder.

- **Feature Extraction:** Given a speech, the acoustic feature extraction module aims to get a log Mel spectrogram representation with 80 channels by using a 25 ms window and a 10 ms step.
- **Encoder:** After 2 one-dimensional convolutional layers and GLUE activation function for length reduction, the audio representation is fed into 12 layers of transformer modules to encode the acoustic features and get the encoder output last hidden state.
- **Decoder:** The decoder uses the same number of transformer modules as the encoder. The last hidden state of the encoder is fed into the decoder through a cross-attention mechanism. Then the decoder autoregressively predicts textual tokens based on the hidden state and previously predicted tokens.

TextEE Model. Text2Event is a generative-based E2E EE method where the entire EE process is modeled uniformly in a sequence-to-structure architecture. All trigger words, arguments, and their type labels are generated uniformly as natural language words to extract events from text in a direct manner.

- **Why do we choose Text2Event for the SpeechEE task?** Text2Event method is chosen for the SpeechEE task because it is data-efficient which means it can be learned using only coarse parallel text-record annotations, i.e. <sentence, event records>, instead of fine-grained token-level annotations. That matches the SpeechEE task perfectly where fine-grained trigger and argument mention annotation is absent because the speech signal is boundless. Therefore, following Text2Event, we adopt the generated-based EE method as the TextEE module of pipeline SpeechEE.
- **Decoding Strategy.** Different from the greedy decoding strategy where the model tends to select the token with the highest

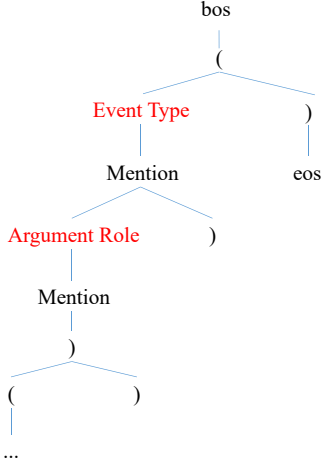


Figure 11: The trie of decoding an event structure.

predicted probability at each decoding step, a trie-based decoding strategy is used in the SpeechEE decoder module. This is because a greedy decoding algorithm can not guarantee to generate valid event structures. In other words, it may lead to invalid event types, mismatched argument types, and incomplete structures.

In addition, the greedy decoding algorithm ignores useful knowledge of event patterns that can effectively guide the decoding. In the SpeechEE model, a trie-based constraint decoding method dynamically selects and prunes a candidate vocabulary based on the currently generated state. The candidate vocabulary consists of event schema, mention to extract, and structure indicator. The event schema, including the event type and the argument role bonding to the event type, is injected into the decoder as external event knowledge to realize the controllable event record generation. As shown in Fig. 11, where “Event Type” and “Argument Role” are pre-defined and serve as constrained candidate vocabulary at the certain decoding step to guarantee the correctness of the event scheme.

B.2 Model Implementation Configurations

Evaluation Metrics. We evaluate the SpeechEE model by using four subtasks in EE.

- **Trigger Identification (TI):** A trigger is correctly identified if it matches a reference trigger.
- **Trigger Classification (TC):** A trigger is correctly classified if its event type and trigger mention match the reference.
- **Argument Identification (AI):** An argument is correctly identified if its event type and argument mention match a reference argument.
- **Argument Identification (AC):** An argument is correctly classified if its event type, argument role and argument mention all match a reference argument.

Model Configurations. We use the pre-trained whisper-large-v2 encoder, T5-large decoder, and Bart-large decoder for the E2E

Table 7: The main hyperparameters of the SpeechEE model.

Hyperparameters	Value
epoch	30
learning rate	5e-5
batch size	16
convolutional layer	2
kernel size	3
stride	2

SpeechEE model. For efficient training, we freeze the acoustic feature extraction module of whisper and train the self-attention encoder, Shrinking Unit module, cross-attention between encoder and decoder, and Entity Dictionary attention. We train all SpeechEE models for 30 epochs and optimize the models using AdamW [26]. We conduct experiments on NVIDIA A100 80GB. All our models are evaluated on the best-performing checkpoint on the validation set. The detailed hyperparameters setting is shown in Table 7.

C EXTENDED DATA SPECIFICATION

C.1 Data Source Description

The raw data is from well-known textual EE datasets, including five sentence-level datasets ACE05-EN⁺ [39], ACE05-ZH [39], PHEE [35], CASIE [32] and GENIA [19]; two document-level datasets RAMS [11] and WikiEvents [22]; one dialogue-level dataset Duconv [42] which is revised from Semantic Role Labeling task, which is pretty much the same task of EE.

- **ACE05-EN⁺**⁸ is a benchmark dataset for event extraction in the English language. It is created as part of the Automatic Content Extraction (ACE) program. The genres include newswire, broadcast news, broadcast conversation, weblogs, discussion forums, and conversational telephone speech. Particularly, for the ACE05-EN⁺ dataset, we follow the same split and preprocessing step with the previous work [24], which extended the original ACE05-EN data by considering multi-token event triggers and pronoun roles and marked it with a ‘+’.
- **ACE05-ZH**⁹ is a Chinese version of the ACE05 dataset. It is developed to facilitate research in event extraction from Chinese texts. Like ACE05-EN, it contains annotated data for training and testing, covering 33 event types. ACE05-ZH plays a crucial role in advancing event extraction research for the Chinese language domain.
- **PHEE**¹⁰ is a dataset for pharmacovigilance comprising over 5000 annotated events from medical case reports and biomedical literature. It is designed for biomedical event extraction tasks.
- **CASIE**¹¹ is an event extraction dataset focusing on the cybersecurity domain. The corpus contains 1000 annotations

⁸<https://blender.cs.illinois.edu/software/oneie/>

⁹<https://catalog.ldc.upenn.edu/LDC2006T06>

¹⁰<https://github.com/zhaoyuesun/phee>

¹¹<https://github.com/Ebiquity/CASIE>

and source files. It annotates event instances with rich annotation and defines five event types in the cybersecurity domain: Databreach, Phishing, Ransom, Discover, and Patch.

- **GENIA**¹² is the primary collection of biomedical literature compiled and annotated within the scope of the GENIA project. It contains 1,999 Medline abstracts, selected using a PubMed query for the three MeSH terms “human”, “blood cells”, and “transcription factors”. The corpus has been annotated with various levels of linguistic and semantic information. We follow the data processing approach as [40] and the GENIA dataset is only used for event detection tasks with five event types.
- **RAMS**¹³ is a widely used document-level event extraction dataset. It contains 9,124 annotated events from news based on an ontology of 139 event types and 65 roles. In the RAMS dataset, the document is relatively short, and each document is annotated with one event.
- **WikiEvents**¹⁴ is a document-level event extraction benchmark dataset that includes complete event and coreference annotation. The corpus is collected from English Wikipedia articles that describe real-world events and then follow the reference links to crawl related news articles. It contains 3,241 events in total, covering 50 event types and 59 argument roles.
- **Duconv**¹⁵ is a dialogue-level dataset which originates from the conversational semantic role labeling task. It contains 3,000 dialogue sessions focusing on movies and stars domain. We format this dataset for the EE task where the predicate works as the trigger in the EE task and the arguments of the predicate work as the argument mentions. Due to the predicate having no fine-grained classification, we consider Duconv as an EE dataset with only one event type and eight argument roles.

C.2 Specification on Data Construction

Based on the textual EE dataset, we employ a method of manual reading to record corresponding speech signals. Considering the high cost associated with manual recording, we conduct random sampling from each textual EE dataset according to the original scale. Specifically, we randomly sample 1,000 instances for ACE05-EN⁺, ACE05-ZH, GENIA, and RAMS datasets, 500 instances for PHEE and CASIE datasets, 140 dialogues for the Duconv dataset and 100 long documents for WikiEvents dataset.

Human-reading Speech Construction. To ensure the diversity of languages and speaker styles in the human-reading SpeechEE dataset, we have separately enlisted the assistance of 10 native speakers for both the Chinese and English languages. These speakers encompass a wide range of ages, genders, and tones, contributing to the diverse styles present in the human-reading SpeechEE dataset. During recording sessions, participants are instructed to maintain a distance of 25 cm between their phones and their mouths, ensuring clarity and accuracy in reading the EE text within a quiet room. Furthermore, to emulate real-life scenarios, we have also

recorded the speech that includes background noise. In these instances, speakers are prompted to record their speech amidst various environmental noise. The noisy background covers ten different scenarios, including street, airport, classroom, cafeteria, supermarket, stadium, meeting room, office, pet store, and rainy days outdoors. A detailed description of the environment settings is shown in Table 8. Following the recording process, two experts meticulously conduct cross-inspections to uphold the high quality of the human-reading speech data.

Automatic Speech Synthesis. First, we introduce the two TTS systems and explain the rationale behind our choice. Bark, a transformer-based TTS model developed by SunoAI, can generate highly realistic, multilingual speech, along with additional audio features such as music, background noise, and simple sound effects. Furthermore, the model is capable of simulating nonverbal expressions like laughter, sighing, and crying. Given its exceptional performance in generating English speech, we prefer Bark as the TTS model for sentence-level English SpeechEE texts because Bark can’t support synthesizing long texts into speech. Therefore, for longer English passages and Chinese text, we employ Edge-TTS as the TTS tool, mainly because of its ability to produce lengthy content and its relatively high quality in generating Chinese speech.

When building the automatic synthesis speech, we also take the speaker’s style diversity into full account. So, we use different voice presets in the TTS system to control the synthesized intonation. Specifically, for the Bark TTS system, we use 10 different English speaker voices; for the Edge-TTS system, we use 7 different English speaker voices and 6 Chinese speaker voices. Both TTS systems cover male and female voices to guarantee the diversity of genders and tones. In addition, we also manually add some background noise into the raw synthesis speech. The ambiances include the cafeteria, meeting room, street, classroom, and rainy days outdoors, which cover many real-world speech scenarios.

After using the TTS model to convert the text of the textual EE train set to audio, we reconstruct the dataset as speech-event records pairs for the SpeechEE task and filter the nonsense instances that only contain meaningfulness information for EE such as “##20010615”, “...um” because these meaningfulness instances may generate abnormal audios such as empty cases, which will cause error for the following speech processing. Finally, we construct above 200 hours of SpeechEE synthesis dataset illustrated in Table 2. The synthesis data will be used to help augment the training corpus for better modeling SpeechEE task.

Quality Control. As for quality controlling, we employ two indicators to show the data quality, the objective indicator, and the subjective indicator. For the objective indicator, we mainly focus on the accuracy of synthesis speech. Therefore, a SOTA ASR model is used to evaluate the word error rate (WER) of the synthesis data. The ASR model we choose is whisper-large-v2¹⁶. The average WER for the English corpus and average CER for the Chinese corpus are 11.4% and 7.2% respectively, which shows the synthesis speech can keep most of the semantic information. For the subjective indicator, we consider naturalness as the quality indicator of synthesis speech. So we launch a listening test to evaluate our SpeechEE dataset. Two

¹²<https://github.com/openbiocorpora/genia-event>

¹³<https://nlp.jhu.edu/rams/>

¹⁴<https://github.com/raspberryyice/gen-arg>

¹⁵https://github.com/syxu828/CSRL_dataset

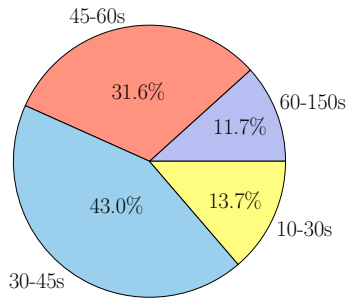
¹⁶<https://huggingface.co/openai/whisper-large-v2>

Table 8: Environment settings on the speech ambiances.

Environment	Description
Street	Busy city streets with pedestrians passing and car noise.
Airport	Crowded airports with broadcasting and crowd noise.
Classroom	Classrooms where students have lively discussions during the break of the class.
Cafeteria	The cafeteria with the clattering of dishes and utensils and the chatting noise of people.
Supermarket	Busy supermarkets with the shoppers talking noise and the sound of cash registers.
Stadium	School stadiums with the sound of sports and cheering.
Meeting room	Meeting rooms with multi-speakers talking and arguing.
Office	Working offices with the noise of printers, photocopiers, ringing telephones, and conversations of workers.
Pet store	Pet stores with the bark of animals and chatting noise of customers.
Rainy days outdoors	Rainy days with the noise of wind and rain outdoors.

Table 9: The age information of the recruited speakers.

Stage	Age periods
Kid	8-12
Teenager	13-19
Young adults	20-30
Middle aged	31-50
Seniors	51-70

**Figure 12: Speech duration statistics of RAMS dataset.**

experts are recruited and asked to rate the naturalness of the synthesis speech using the 10-scale mean opinion score. The Cohesion Kappa score is utilized to assess the level of agreement among experts, measuring their consistency. Cohen’s Kappa score, a statistical metric for inter-rater reliability, is employed to determine the extent of agreement among experts beyond what could be attributed to chance. Following the standard quality control process, we compute the Cohesion Kappa score, attaining a 0.81 for the synthesis SpeechEE dataset, indicative of the good quality of our datasets. Data with Kappa scores falling below a predefined threshold undergo thorough review and discussion to maintain the good quality of our dataset.

C.3 Specification on SpeechEE Duration

For sentence-level datasets, all synthesis datasets except ACE05-ZH are generated by the Bark TTS model, which is limited from 1 second to 15 seconds. For the RAMS dataset in which the speech duration differs obviously from 10 seconds to three minutes, we count the speech duration distributions and show them in Fig. 12. RAMS

dataset is mainly composed of speech within 60 seconds, which shows its documents are shorter but more numerous (9,124 documents). In contrast, the WikiEvents dataset consists of 246 long documents which average 4 minutes duration. These two document-level datasets are quite complementary when evaluating the SpeechEE model under different cases.

C.4 Specification on SpeechEE Data Insight

- **Two Construction Approaches:** Our SpeechEE dataset contains two construction approaches: manually crafted human-reading speech, and synthesized speech generated using high-performance TTS systems. On one hand, human-reading speech includes rich information from real-life scenarios such as emphasis, pauses, onomatopoeia, and emotional cues. On the other hand, synthesized speech serves as an efficient data augmentation method, aiding in the rapid construction of large-scale training corpora and addressing the high cost associated with manually constructing datasets.

- **Multiple Scenarios:** Our SpeechEE dataset covers three major common scenarios of existing EE: sentence-level, document-level, and dialogue-level, which can help to evaluate the performance comprehensively.

- **Multiple Languages:** Our SpeechEE dataset covers two languages, Chinese for ACE05-ZH and Duconv datasets and English for the other six datasets.

- **Diverse Domains:** Our SpeechEE dataset covers a wide range of topics and fields including news reports, medicine effects, genetic biology, cybersecurity, movies and stars, and other general domains. The diverse domains enable models to be trained and applied in a wider range of real-world scenarios.

- **Rich Tones and Genders:** Our SpeechEE dataset also considers the diversity of speaker styles. Specifically, we have 17 English speaker voices and 6 Chinese speaker voices in the synthesis speech by choosing different TTS voice presets. For human-reading speech, we recruit 10 different native speakers of English and Chinese respectively. These voices include men and women and differ in speech volume, speed, and intonation.

- **Multiple Ages:** Our SpeechEE dataset covers all kinds of age stages by considering different age periods of speakers. The detailed age information is shown in Table 9.

- **Different Ambiences:** Our SpeechEE dataset considers two background settings, including the quiet background and various noisy scenarios in the real world. The ambiances include 10 different background settings such as car noise in the street, crowd noise in the cafeteria and classroom, multi-speaker noise in the meeting room, and rainy day noise outdoors, which can be found in Table 8.

- **Large Scale and High Quality:** Our SpeechEE dataset not only contains the manually crafted human-reading speech, but we also augment it with synthesis speech by using TTS systems to enlarge its scale by a factor of 10. At the same time, strict human cross-inspection is conducted to ensure the high quality of the whole speech data.