Received: 22 June 2026; **Revised:** 20 September 2026; **Accepted:** 18 October 2026; **Published:** 20 November 2026
Archs Sci. (2026) Volume 76, Issue 5 Pages 16-28, Paper ID 2026502.
https://doi.org/10.68304/as/76502

Research on Risk Identification of Digital Voice Information in Electric Power Customer Service Center

Tongxu Liu[1], Chuangye Wang[1], Yu Wang[2], Shilin Shen[1,*] and Yunchuan Wang[1]
[1]Bengbu Power Supply Company, State Grid Anhui Electric Power Co., Ltd. Bengbu 233000, Anhui, China
[2]Electric Power Research Institute, State Grid Anhui Electric Power Co., Ltd., Hefei 230601, Anhui, China
*Correspondence: ahssss11111@163.com

Abstract With the increasing adoption of intelligent voice services in electric power customer service centers, identifying potential information risks from large-scale speech interactions has become essential for service compliance and information security management. However, existing speech-based risk identification methods often struggle with implicit semantic relationships, multi-turn contextual dependencies, and recognition errors under noisy acoustic conditions. To address these challenges, this paper proposes a digital speech information risk identification framework integrating Automatic Speech Recognition (ASR), Abstract Meaning Representation (AMR), and multi-granularity graph neural networks. Specifically, a dual-encoder architecture combines hierarchical Transformer-based textual representations with graph neural network-based AMR encoding to extract complementary contextual and structured semantic features. Clause-level and token-level adjacency structures, together with hierarchical self-attention and cross-attention fusion, are employed to capture long-range semantic dependencies and risk-related entity relationships. Furthermore, a Tacotron2–MelGAN-based speech augmentation module is incorporated to support diverse risk-oriented speech scenarios and robustness evaluation. The framework is designed to identify potential sensitive information disclosure, abnormal service statements, procedural violations, and contextual risks in customer service conversations. Comparative experiments involving SRL-BERT, DEP-GNN, and Lin-AMR, together with ablation studies, are designed to evaluate identification performance using Precision, Recall, Macro-F1, and AUPRC under realistic, simulated, and noise-augmented speech conditions.

Index Terms electric power customer service; digital speech information; risk identification; automatic speech recognition; abstract meaning representation; graph neural networks; multi-granularity semantic modeling

I. Introduction

With the rapid development of artificial intelligence, digital communication, and intelligent customer service technologies, voice-based service systems have become an essential component of modern electric power customer service centers. These systems support various electricity-related services, including electricity bill inquiries, power outage reporting, account management, service complaints, and technical consultation, generating substantial volumes of digital speech information during daily operations [1], [2]. However, the increasing complexity of customer service interactions also introduces potential information security and operational risks, such as personal information disclosure, improper identity verification, abnormal service statements, procedural violations, and misleading service commitments. These risks may be embedded in complex linguistic expressions, ambiguous conversational contexts, or interactions spanning multiple dialogue turns, making them difficult to identify through conventional keyword matching and rule-based monitoring methods. Consequently, developing an intelligent digital speech information risk identification method capable of accurately extracting risk-related semantic information and recognizing potential anomalies has become an important research challenge in electric power customer service management [3], [4].

Currently, digital speech analysis systems generally adopt a processing framework consisting of automatic speech recognition (ASR), natural language understanding (NLU), and downstream information classification. Existing risk identification methods primarily rely on predefined keyword dictionaries, statistical text features, semantic role labeling, and deep learning-based classification models to identify potentially sensitive or abnormal information in spoken conversations [5]. Although these methods provide a foundation for automated speech content analysis, their effectiveness may be limited when dealing with complex service expressions, implicit semantic risks, and multi-turn conversational dependencies. For instance, a potentially risky statement may

not contain explicit sensitive keywords but may reveal personal information when interpreted together with preceding dialogue content. Furthermore, the presence of background noise, overlapping speech, informal expressions, and domain-specific terminology may introduce recognition errors and semantic ambiguity, reducing the reliability of subsequent risk classification [6]. These limitations highlight the need for more comprehensive semantic representation and contextual reasoning mechanisms to improve the accuracy and robustness of digital speech information risk identification in electric power customer service environments.

In recent years, structured semantic representation has attracted increasing attention as an effective approach to improving natural language understanding and complex information classification [7]. Among existing semantic modeling techniques, Abstract Meaning Representation (AMR) provides a graph-based representation of sentence meaning by organizing semantic concepts and their relationships into directed graph structures. In an AMR graph, nodes represent semantic entities, events, and concepts, while labeled edges describe semantic roles and logical relationships, such as `:ARG0`, `:ARG1`, `:mod`, and `:time`. Compared with conventional word-level representations and dependency-based syntactic analysis, AMR offers a more explicit representation of predicate–argument relationships and event structures, which may facilitate the identification of implicit semantic risks and complex contextual associations [8], [9]. In electric power customer service conversations, AMR-based semantic modeling can provide structured information for analyzing interactions among customers, service personnel, account-related entities, operational actions, and potentially sensitive information. By capturing semantic dependencies across clauses and conversational events, AMR provides a promising foundation for recognizing risks that cannot be reliably detected through isolated keywords or shallow textual features.

Since the introduction of AMR by Banarescu et al. in 2013, considerable research has focused on semantic graph construction, neural semantic parsing, and graph-based language representation. Early AMR parsing methods, including JAMR and CAMR, primarily relied on statistical learning and manually designed linguistic features, while subsequent approaches incorporated sequence-to-sequence architectures, recurrent neural networks, and Transformer-based models to improve semantic parsing performance [10], [11]. Meanwhile, graph neural networks (GNNs) have demonstrated the ability to aggregate relational information from structured representations, enabling more effective modeling of interactions between semantic entities and events. Hierarchical Transformer architectures further support semantic feature extraction at multiple linguistic levels, including words, phrases, clauses, and sentences. Combining these approaches provides an opportunity to integrate sequential contextual information with explicit semantic graph structures. For electric power customer service risk identification, such integration may improve the representation of sensitive entity relationships, service operation dependencies, and risk-related semantic patterns distributed across complex dialogue structures [12].

Despite these developments, several challenges remain in applying semantic graph-enhanced learning methods to digital speech information risk identification. First, speech recognition errors caused by acoustic interference, conversational overlap, and specialized electricity service terminology may distort risk-related expressions and weaken the reliability of downstream semantic analysis [13]. Second, conventional sequence-based classification models may not adequately capture complex semantic relationships among entities, service actions, and contextual events, particularly when risk indicators are distributed across multiple clauses or dialogue turns [14]. Third, electric power customer service conversations often contain highly imbalanced risk categories, where certain critical but infrequent risk events may be difficult to identify using conventional classification objectives. Moreover, differences between real customer service recordings and simulated dialogue scenarios can affect model generalization. Although AMR-based representations offer potential advantages for structured semantic understanding, their contribution to risk identification under noisy speech conditions and complex multi-turn service interactions requires systematic investigation [15]. These challenges motivate the development of a risk-oriented semantic modeling framework that integrates speech recognition, hierarchical contextual encoding, structured graph representation, and multi-label risk classification.

To address these challenges, this paper proposes a digital speech information risk identification framework for electric power customer service centers based on AMR-enhanced multi-granularity semantic modeling. The proposed framework integrates automatic speech recognition, natural language understanding, AMR semantic parsing, hierarchical feature extraction, and graph-based risk classification within a unified architecture. Specifically, a dual-encoder structure is developed to capture complementary semantic information from recognized speech transcripts and their corresponding AMR graphs. The first encoder employs a hierarchical Transformer to extract contextual features at the word, phrase, clause, and sentence levels, while the second encoder utilizes a graph neural network to learn structured semantic relationships among entities, actions, and risk-related events. The representations obtained from both encoders are integrated through a cross-attention mechanism to construct comprehensive semantic features for downstream risk identification. In addition, multi-turn contextual information is incorporated to support the recognition of potential information disclosure, abnormal service expressions, procedural violations, and other complex risk patterns. A comparative evaluation framework involving SRL-BERT, DEP-GNN, and Lin-AMR is established, together with ablation experiments and robustness analysis, to investigate the contributions of the proposed architectural components. This research aims to provide an interpretable and robust semantic modeling approach for automated digital speech information risk identification, offering a technical foundation for intelligent risk monitoring, service compliance assessment, and information security manage-

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

ment in electric power customer service centers.

II. Methodology

A. Overall Framework Overview

Figure 1 illustrates the overall architecture of the proposed human-machine intelligent interaction system based on speech recognition and natural language understanding. The system integrates automatic speech recognition (ASR), natural language understanding (NLU), intent classification, dialog management, and natural language generation with speech synthesis. This architecture supports robust interaction in low-resource environments by using semantic representation enhancement and structured processing logic.

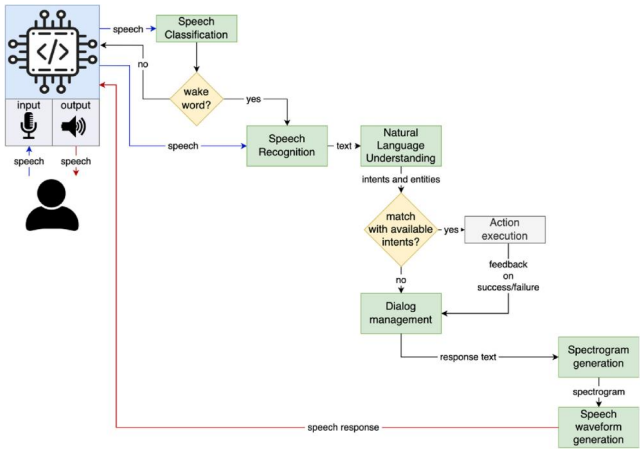


Figure 1: Overall Architecture of the Human-Machine Spoken Interaction System Based on Speech Recognition and Natural Language Understanding

B. System Components and Flow Description

The user provides speech input s, captured by a microphone interface [16]. The input is processed to determine if it contains a wake word. The classification function is defined as:

$$C_{\text{wake}}(s) = \begin{cases} 1 & \text{if wake word detected,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Only when $C_{\text{wake}}(s) = 1$ does the system proceed to recognition.

Speech recognition (ASR) converts the speech signal s into a text sequence x. This process can be formulated as:

$$x = \underset{x'}{\operatorname{argmax}}\, P(x' \mid s), \tag{2}$$

where $P(x' \mid s)$is the probability of a candidate text sequence given the speech input.

Natural language understanding (NLU) parses the recognized text x into an intent i and a set of entities e:

$$(i, e) = \text{NLU}(x) \tag{3}$$

NLU may include Named Entity Recognition (NER), Part-of-Speech tagging, and AMR parsing if semantic graphs are used:

$$G = \text{AMR}(x). \tag{4}$$

For intent matching, the extracted intent i is compared with known intents $I = \{i_1, i_2, \ldots, i_n\}$using a similarity function:

$$\text{Match}(i, I) = \begin{cases} 1 & \text{if, } i_j \in I \text{ such that, } \text{sim}(i, i_j) > \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

For action execution or dialogue management, a matched intent triggers an associated action $a \in A$, producing feedback $f \in \{\text{success}, \text{fail}\}$.

$$f = \text{Execute}(i, e). \tag{6}$$

If there is no match, a response is generated by the dialogue policy π:

$$r = \pi(i, e, h), \tag{7}$$

where h denotes the dialogue history.

The textual response r is converted into a mel-spectrogram representation:

$$S = \text{Spec}(r). \tag{8}$$

Waveform synthesis is then performed using a vocoder:

$$s = \text{Vocoder}(S). \tag{9}$$

For speech output, the final synthesized signal s' is played back to the user.

C. Mathematical Summary of the Pipeline

Speech Input:

$$s.$$

Wake Detection:

$$C_{\text{wake}}(s) \to \{0, 1\}.$$

ASR:

$$x = \underset{x'}{\operatorname{argmax}}\, P(x' \mid s).$$

NLU:

$$(i, e) = \text{NLU}(x).$$

Matching:

$$\text{Match}(i, I) \to \{0, 1\}.$$

Action or Dialogue:

$$\begin{cases} f = \text{Execute}(i, e) & \text{if matched,} \\ r = \pi(i, e, h) & \text{else.} \end{cases}$$

Speech Response:

$$s = \text{Vocoder}(\text{Spec}(r)).$$

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

D. Automatic Speech Recognition Module

To enhance the diversity and robustness of digital speech information risk identification in electric power customer service centers, a speech synthesis-based data augmentation module integrating Tacotron2 and MelGAN is introduced into the proposed framework. This module is designed to generate simulated customer service speech samples containing representative risk-related expressions, including sensitive information disclosure, abnormal service statements, procedural violations, and complex contextual interactions. As illustrated in Figure 2, the speech generation module consists of two components: Tacotron2, which converts risk-oriented textual dialogue sequences into mel-spectrogram representations, and MelGAN, a non-autoregressive generative adversarial network that transforms these spectrograms into time-domain speech waveforms [17]. The synthesized speech samples can be further processed with acoustic perturbations and background noise to construct diverse experimental conditions, supporting the evaluation of speech recognition reliability and downstream semantic risk classification. By expanding the coverage of risk-related speech scenarios, this module provides an auxiliary data augmentation mechanism for investigating the robustness and generalization capabilities of the proposed AMR-enhanced multi-granularity risk identification model

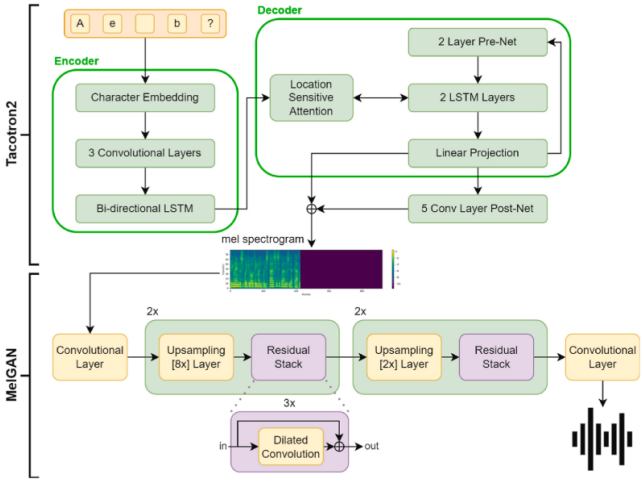


Figure 2: Tacotron2 and MelGAN speech synthesis architecture

The acoustic model maps acoustic features $\mathbf{x}_{1:T} \in \mathbb{R}^{T\times d}$ (e.g., MFCC or Mel spectrograms) to a sequence of phonetic or subword units$\mathbf{y}_{1:N} \in V^N$.

We define the objective as:

$$\hat{y}_{1:N} = \underset{y_{1:N}}{\operatorname{argmax}} P(y_{1:N} \mid \mathbf{x}_{1:T}). \tag{10}$$

HMM-based model assumes framewise independence and uses the Viterbi algorithm for decoding.

CTC (Connectionist Temporal Classification) defines the probability over all alignments πthat map to y:

$$P(y \mid \mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(y)} P(\pi \mid \mathbf{x}), \tag{11}$$

where $B$is the collapsing function removing blank tokens and duplicates.

1) Language Model: N-gram + LSTM Hybrid

To improve decoding accuracy, a language model $P_{\mathrm{LM}}(y_t \mid y_{<t})$is combined with the acoustic posterior:

$$\hat{y}_{1:N} = \underset{y_{1:N}}{\operatorname{argmax}} \big[\log P(y_{1:N} \mid \mathbf{x}_{1:T}) + \lambda \log P_{\mathrm{LM}}(y_{1:N}) \big]. \tag{12}$$

The language model may consist of:

N-gram LM:

$$P_{\mathrm{LM}}(y_t \mid y_{t-n+1}^{t-1}) = \frac{\mathrm{Count}(y_{t-n+1}^{t})}{\mathrm{Count}(y_{t-n+1}^{t-1})}. \tag{13}$$

LSTM-based LM:

$$\begin{aligned} h_t &= \mathrm{LSTM}(y_{<t}); \\ P_{\mathrm{LM}}(y_t) &= \mathrm{Softmax}(W h_t + b). \end{aligned} \tag{14}$$

Tacotron2 serves as the front-end module in the speech synthesis pipeline, responsible for converting textual sequences into mel spectrograms. The encoder component of Tacotron2 begins with a character embedding layer that transforms discrete text tokens into dense vector representations. These embeddings are then processed by three one-dimensional convolutional layers, which extract local n-gram features from the input sequence [18], [19]. The output is passed through a bidirectional Long Short-Term Memory (Bi-LSTM) network to capture long-range dependencies in both forward and backward directions. In the decoder, a location-sensitive attention mechanism is employed to align encoder outputs with decoder states at each time step. The attended context vectors are fed into a two-layer LSTM decoder, followed by a linear projection layer that predicts frames of the mel spectrogram [20]. To refine the spectral output, a five-layer convolutional post-net is applied, enhancing the fidelity and continuity of the generated features. The final output of Tacotron2 is a mel spectrogram $\mathbf{M} \in \mathbb{R}^{F\times T}$, where F denotes the number of mel-frequency bins and T denotes the number of time frames.

MelGAN operates as the back-end vocoder in the speech synthesis pipeline, converting the mel spectrogram Minto a time-domain audio waveform$s(t) \in \mathbb{R}^T$. It adopts a fully convolutional, non-autoregressive architecture that enables real-time waveform generation. The network is composed of a sequence of upsampling blocks with scaling factors of $8\times$, $2\times$, and $2\times$ which progressively increase the temporal resolution of the spectrogram features to match the sampling rate of the output waveform. Each upsampling block is followed by a residual stack, which incorporates multiple layers of dilated convolutions to capture long-range dependencies and expand the receptive field without increasing computational cost. The transformation at each residual block can be formally expressed as$y = x + \sum_{k=1}^{K} \mathrm{ReLU}(\mathrm{DilatedConv}_k(x))$, where

x denotes the input feature map and K denotes the number of dilation levels. The final output of the MelGAN model is a high-quality, natural-sounding speech waveform$s(t)$, completing the text-to-speech synthesis process.

E. Mathematical Summary of the TTS Process

Text Input:

$$y = (y_1, y_2, \ldots, y_N).$$

Tacotron2:

$$M = f_{\text{Tacotron2}}(y).$$

MelGAN:

$$s(t) = f_{\text{MelGAN}}(\mathbf{M}).$$

F. NLP Pipeline with Semantic Parsing

To support accurate understanding of user utterances, the system employs a multi-stage NLP pipeline culminating in Abstract Meaning Representation (AMR) parsing. The pipeline includes tokenization, POS tagging, named entity recognition (NER), semantic role labeling (SRL), and dependency parsing. These intermediate features feed into the AMR parser to construct structured semantic graphs.

An AMR graph is defined as $G = (V, E)$, where V contains semantic concepts (e.g., "check-01", "wait-01"), and E contains labeled relations (e.g., :arg0, :mod, :time) between them. As shown in Figure 3, AMR captures clause-level meaning, argument structure, and temporal relations in a unified graph format.

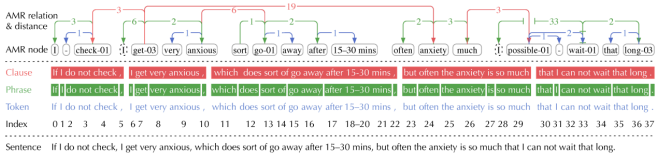


Figure 3: AMR Graph Example with Semantic Roles and Clause Alignment

The AMR parser predicts the graph $G = \text{argmax}_{G'} P(G' \mid x)$, using models such as Seq2Seq, Stack-LSTM, or BART-AMR. To enhance representation, a Graph Recurrent Network (GRN) is optionally used:

$$\mathbf{g}_T = \text{GRN}(G). \tag{15}$$

AMR enhances downstream modules by providing explicit, interpretable semantic structure for intent disambiguation, reasoning, and response generation.

As shown in Figure 4, to enhance semantic understanding across discourse units, we introduce clause- and token-level adjacency matrices. These capture syntactic proximity, pronoun reference, and clause linkage, aiding intent disambiguation and dialog reasoning.

(a) Clause Adjacency Matrix$A^{(c)} \in \mathbb{R}^{N_c \times N_c}$:

$$A_{ij}^{(c)} = \begin{cases} 1, & i = j, \\ \gamma, & |i - j| = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

(b) Token Adjacency (Adjacent Clauses)$A^{(t)} \in \mathbb{R}^{n \times n}$:

$$A_{ij}^{(t)} = \begin{cases} 1, & t_i, t_j \in \text{same clause}, \\ \gamma, & t_i \in C_k, t_j \in C_{k+1}, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

(c) Non-Adjacent Token Matrix highlights long-range gaps, with white indicating minimal interaction.

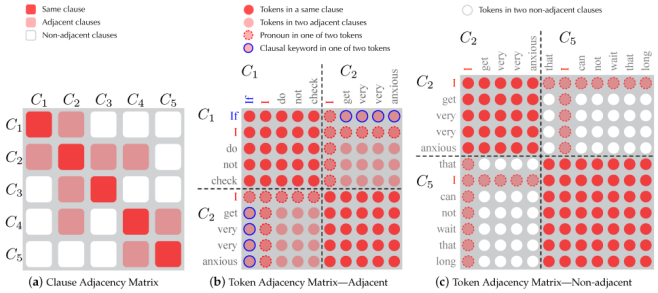


Figure 4: Clause and Token Adjacency Matrices
(a) Clause-level structure. (b) Token connections across adjacent clauses. (c) Token distances in non-adjacent clauses, modeling long-range dependencies.

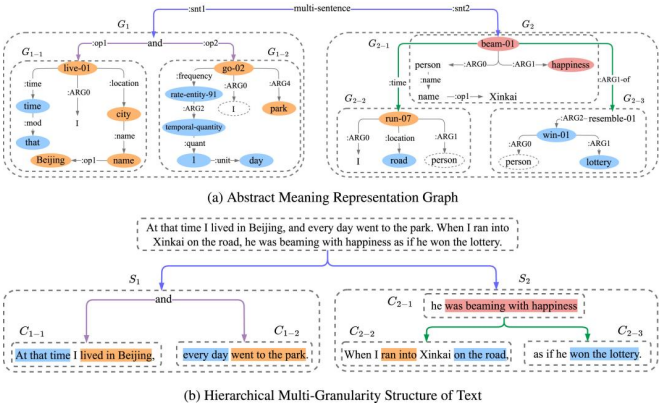


Figure 5: Hierarchical Multi-Granularity AMR Graph and Sentence Structure (a) AMR graph segments G_{i-j}represent each clause-level event or statement with semantic roles. (b) The original text is decomposed into sentences and clausesC_{i-j}, hierarchically aligned with the AMR subgraphs. This enables structured semantic understanding across sentences and discourse relations.

To capture discourse-level meaning and semantic alignment across clauses and sentences, we construct a hierarchical Abstract Meaning Representation (AMR) structure aligned with multi-granularity textual units. As illustrated in Figure 5, each sentence is segmented into multiple clauses (e.g.,$C_{1-1}, C_{1-2}, C_{2-1}$), each mapped to a corresponding AMR subgraph (e.g.,$C_{1-1}, C_{1-2}, C_{2-1}$). These AMR subgraphs encode predicate-argument structures (e.g., live-01, beam-01, win-01) and semantic roles such as :ARG0, :location, or :frequency, allowing clause-level meaning to be represented as structured graphs. The entire graph is defined

as$G = \bigcup_{i,j} G_{i-j}$, with each clause-to-graph mapping denoted by$f : C_{i-j} \rightarrow G_{i-j}$. This alignment facilitates the modeling of inter-sentence semantic links (e.g., :snt, :op, :ressemble-01) and supports tasks like context-aware dialog generation, discourse reasoning, and event chaining.

G. Dual Encoder Structure

To model the alignment between raw speech and semantic meaning, we propose a dual encoder structure that integrates a hierarchical Transformer for text encoding and a graph-based neural encoder for AMR representation [21]. As shown in Figure 6, the left encoder first converts input speech (e.g., "He told me to play the piano.") into word embeddings via ASR, followed by a hierarchical multi-head self-attention mechanism across four linguistic levels: word, phrase, clause, and sentence. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$be the input embedding sequence. Each level $l \in \{\text{word,phrase,clause,sentence}\}$generates a representation:

$$\mathbf{H}^{(l)} = \text{MHA}^{(l)}(\mathbf{X}^{(l)}) + \text{FFN}^{(l)}(\cdot), \tag{18}$$

with all levels concatenated and normalized:

$$\mathbf{H}_{\text{hier}} = \text{AddNorm}\big(\text{Concat}\big(\mathbf{H}^{(\text{word})}, \mathbf{H}^{(\text{phrase})}, \mathbf{H}^{(\text{clause})}, \mathbf{H}^{(\text{sentence})}\big)\big). \tag{19}$$

On the right side, the AMR graph is encoded via a Graph Recurrent Network (GRN) or GAT. Given AMR nodes VV and edges EE, the graph encoder updates node features as:

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \mathbf{W}_r \cdot \mathbf{h}_u^{(k)} + \mathbf{b}_r\right), \tag{20}$$

where r denotes the relation type in AMR (e.g., :ARG0, :location), and $N(v)$is the set of neighbors of node v.

The final hierarchical cross-attention module bridges both encoded modalities by aligning text and semantic structure:

$$\mathbf{Z} = \text{CrossAttn}(\mathbf{H}_{\text{text}}, \mathbf{H}_{\text{AMR}}), \tag{21}$$

which is followed by decoder layers to generate structured AMR output sequences or response utterances.

Figure 7 illustrates the multi-granularity semantic encoding architecture developed for digital speech information risk identification in electric power customer service centers. The proposed architecture models speech-transcribed text at four semantic levels: word, phrase, clause, and sentence, enabling the extraction of hierarchical risk-related semantic features from customer service conversations. First, word-level embeddings are grouped and aggregated into phrase-, clause-, and sentence-level representations to capture both local linguistic patterns and broader contextual information. Subsequently, a structure-aware Ordered Neurons Long Short-Term Memory (ON-LSTM) module enhances these representations by modeling hierarchical linguistic dependencies and complex semantic relationships associated with potential information risks, including sensitive information disclosure, abnormal service statements, and procedural violations. The resulting representations are then processed by a multi-granularity self-attention mechanism to facilitate information interaction and semantic feature fusion across different linguistic levels. Finally, the attention-enhanced features are concatenated and integrated into a unified semantic representation for downstream risk classification. By jointly exploiting local semantic cues, hierarchical linguistic structures, and long-range contextual dependencies, the proposed encoding architecture is designed to improve the identification of implicit and context-dependent risks in electric power customer service dialogues, providing discriminative semantic features for subsequent AMR-enhanced graph modeling and intelligent risk assessment.

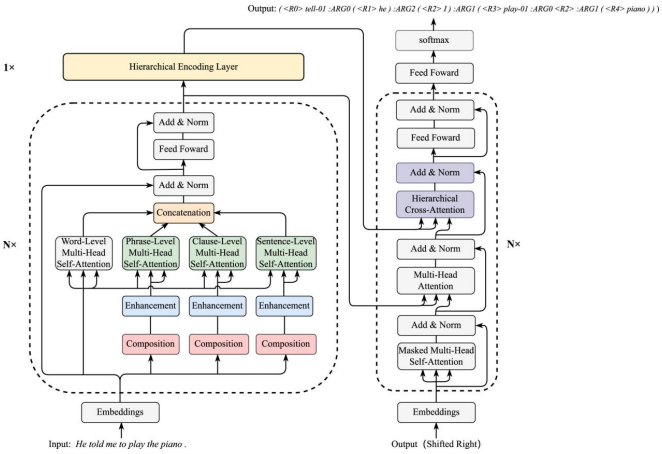


Figure 6: Dual Encoder Architecture with Hierarchical Transformer and AMR Graph Encoder
(Left) Multi-level attention encoder for linguistic units; (Right) graph-aware decoder with cross-attention over AMR semantic nodes.

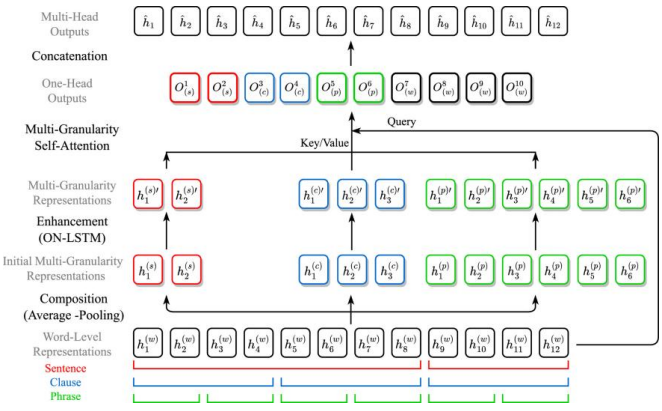


Figure 7: Multi-Granularity Semantic Encoder with Hierarchical Self-Attention

III. System Architecture for Human-Machine Interaction

Figure 8 illustrates the overall architecture of the proposed multi-source semantic fusion framework for digital speech

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

information risk identification in electric power customer service centers. The framework integrates automatic speech recognition, natural language processing, semantic retrieval, and graph-based feature learning, comprising four main stages: multimodal risk information database construction, semantic and acoustic feature retrieval, multi-granularity knowledge aggregation, and risk identification. First, semantic representations and acoustic feature vectors are extracted from the text track (TT-SD) and audio track (AT-SD), respectively, to construct a multimodal customer service information database containing linguistic, acoustic, and risk-related features. During risk identification, the system extracts contextual semantic representations from the current customer service conversation and retrieves the Top-K relevant historical samples through semantic and acoustic feature retrieval mechanisms. Contrastive learning is incorporated to enhance feature discrimination and improve the representation of risk-related conversational patterns. Subsequently, a multi-granularity heterogeneous graph is constructed by combining retrieved historical information with current conversational content, incorporating word-, phrase-, clause-, and sentence-level semantic relationships. The framework further employs graph-based knowledge aggregation to integrate historical risk patterns, contextual dependencies, and structured semantic information into a unified feature representation. Finally, the aggregated representation is processed by a risk classification and assessment module to identify potential sensitive information disclosure, abnormal service expressions, procedural violations, and other information security risks. The classification results are further used to generate risk categories, confidence scores, and corresponding warning levels, providing structured outputs for intelligent risk monitoring and service compliance assessment in electric power customer service centers.

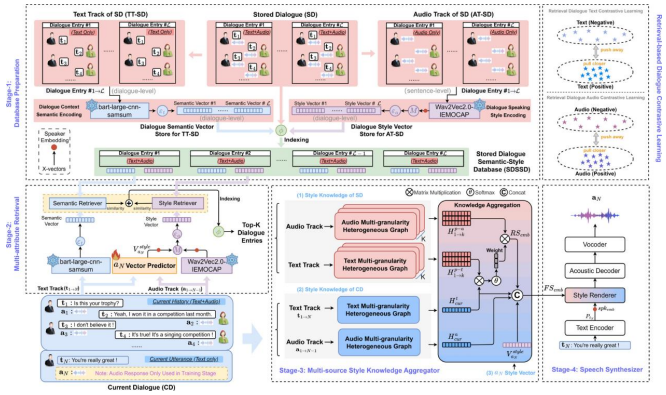


Figure 8: Overall Architecture

IV. Experiment

A. Datasets

To evaluate the effectiveness of the proposed digital speech information risk identification method for electric power customer service centers, an experimental framework combining domain-specific speech recordings, simulated risk-oriented dialogues, and noise-augmented speech data is designed. The evaluation focuses on identifying potential information security risks, service compliance violations, and abnormal semantic interactions in customer service conversations. The following datasets and experimental conditions are considered.

Power Customer Service Speech Dataset (PCSS): This dataset is intended to comprise authorized and anonymized customer service recordings covering electricity bill inquiries, payment services, power outage reports, service complaints, account verification, and electricity-related consultation. Speech recordings are transcribed into textual sequences and annotated according to predefined risk categories, including potential personal information disclosure, improper identity verification, non-compliant service statements, misleading commitments, and unresolved complaints requiring escalation. The dataset serves as the primary benchmark for evaluating risk identification under realistic customer service conditions.

Simulated Power Service Risk Dialogue Dataset (PSRD): To address the limited availability of annotated high-risk conversations, a scenario-based dataset is designed using representative customer service interactions and predefined risk patterns. The simulated dialogues cover sensitive information requests, abnormal account inquiries, procedural violations, repeated complaints, and semantically ambiguous expressions. Each dialogue contains multiple conversational turns and corresponding risk annotations, supporting the evaluation of contextual understanding, semantic dependency modeling, and multi-turn risk identification.

Noise-Augmented Power Service Speech Dataset (NPSS): To investigate the robustness of risk identification under imperfect speech recognition conditions, noisy evaluation samples are constructed from the power customer service speech data. Background noise, channel distortion, signal attenuation, and recognition errors are introduced at different severity levels. These samples are used to examine the influence of acoustic degradation and automatic speech recognition errors on downstream semantic risk classification. Source conversations are separated before augmentation to prevent information leakage between training and test partitions.

In addition, the Chinese Abstract Meaning Representation dataset LDC2019T07 may be employed as an auxiliary resource for initializing or evaluating the semantic parsing module. Since it does not contain power customer service risk labels, it is not considered a direct benchmark for the final risk identification task.

The experimental annotations follow a multi-label classification strategy, allowing a conversation to contain more than one risk category. Conversations without identified risks are labeled as normal. All risk labels are subject to clearly defined annotation guidelines and independent human verification.

B. Experimental Settings

To ensure consistency and reproducibility, the proposed model and baseline methods are evaluated under a unified pre-

processing, training, and testing protocol. The experimental configuration retains the fundamental optimization settings of the original semantic modeling framework while adapting the training objective to digital speech information risk identification.

Speech Preprocessing: Customer service recordings are processed using speech segmentation, normalization, and automatic speech recognition (ASR). The resulting transcripts are organized into conversational turns while preserving speaker roles and temporal ordering. Sensitive identifiers are masked according to the requirements of data protection and privacy preservation.

Lexical Segmentation: Recognized text is segmented using Byte Pair Encoding (BPE). Electric power service terminology, account-related expressions, complaint phrases, and potentially sensitive entities are preserved or normalized through domain-specific preprocessing rules to reduce semantic ambiguity.

Semantic Feature Extraction: Abstract Meaning Representation (AMR) parsing is applied to construct structured semantic graphs from customer service transcripts. These graphs capture semantic roles, action–entity relationships, temporal dependencies, and contextual associations. A hierarchical Transformer extracts linguistic representations at word, phrase, clause, and sentence levels, while a graph neural network encodes the corresponding AMR structures.

Risk Classification: The encoded textual and semantic graph representations are integrated through a cross-attention mechanism. The resulting features are passed to a multi-label classification layer, which estimates the probability of each predefined risk category. Risk identification thresholds are determined using the validation set rather than fixed arbitrarily.

Optimizer: The Adam optimizer is adopted with an initial learning rate of 1×10^{-4}.

Regularization: A dropout rate of 0.3 is applied to reduce overfitting and improve model generalization.

Batch Size and Training Epochs: The initial experimental configuration uses a batch size of 64 and a maximum of 50 training epochs. Early stopping is performed according to validation Macro-F1 performance.

Dataset Partitioning: The proposed data partitioning ratio is 70% for training, 15% for validation, and 15% for testing. The split is performed at the conversation level to avoid placing segments from the same dialogue in different partitions. Where available, temporally separated or independently collected test samples are used to assess generalization.

Evaluation Metrics: Precision, Recall, Macro-F1, and Area Under the Precision–Recall Curve (AUPRC) are selected as the primary evaluation metrics. Precision measures the reliability of predicted risk events, while Recall reflects the model's ability to detect actual risks. Macro-F1 provides a balanced evaluation across different risk categories, and AUPRC assesses discrimination under imbalanced class distributions. The false positive rate and per-category recall are further examined to evaluate practical risk monitoring performance.

To improve experimental reliability, each model is intended to be evaluated using multiple random seeds, and the mean and standard deviation of the performance metrics should be reported.

The intended evaluation protocol further requires that every risk annotation be linked to a unique conversation identifier and that risk categories be defined before training. A risk may emerge from the combination of a customer request and an agent response even when neither utterance contains a prohibited term on its own. Therefore, annotators should inspect complete conversation segments, resolve disagreements through a documented adjudication procedure, and distinguish genuinely observable violations from uncertain or hypothetical concerns. Separate reporting of individual risk types is particularly important because broad aggregate scores could obscure failures to recognize rare but consequential information disclosures. The same risk taxonomy and annotation rules must be used for all comparison models to make their outputs directly comparable. To test robustness fairly, acoustic modifications should be generated only after dividing the original conversations into disjoint training, validation, and test groups. The source identity of each augmented recording should be retained to prevent near-duplicate utterances from appearing across partitions. Reported word error rates should be measured against verified reference transcripts rather than inferred from synthetic distortion settings alone. Because customer-service conversations may contain personal identifiers and account details, both recorded and synthesized materials require review for inadvertent disclosure. Names, addresses, telephone numbers, and account references should be masked or replaced consistently, while keeping the linguistic relations necessary to evaluate the proposed semantic model.

The numerical values in the comparison and ablation tables are illustrative design examples, not measured results from a completed study. They may be used to explain the expected interpretation of the planned metrics but cannot support claims of achieved performance, statistical significance, or superiority over competing methods. Before publication as an experimental Original Research Article, the described data sources, actual sample sizes, executable training configurations, model outputs, and independent test results would need to be supplied and checked. For the same reason, plots of loss, feature selection, and synthesized spectrograms should be presented as conceptual examples unless the corresponding underlying records are available. This distinction is necessary to keep the interpretation of the model architecture separate from claims about observed outcomes.

C. Results and Analysis

To examine the effectiveness of the proposed AMR-enhanced multi-granularity semantic risk identification model, a comparative evaluation is designed using three baseline methods: SRL-BERT, DEP-GNN, and Lin-AMR. All models are adapted to the same multi-label risk classification task and evaluated using consistent training and testing partitions.

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

SRL-BERT integrates semantic role information with contextual language representations to identify potentially risky expressions in customer service transcripts. DEP-GNN employs syntactic dependency graphs to capture structural relationships between words and phrases. Lin-AMR converts semantic graphs into linearized representations for downstream classification. The proposed method combines hierarchical textual encoding with graph neural network-based AMR representation and cross-attention fusion to identify risk-related information across multiple linguistic levels.

Table 1 presents an illustrative comparison of four semantic modeling approaches for digital speech information risk identification in electric power customer service centers under three evaluation conditions. In this hypothetical scenario, the proposed AMR-enhanced multi-granularity GNN model achieves the highest performance across all evaluation metrics, with Macro-F1 scores of 89.4%, 92.6%, and 84.5% on PCSS, PSRD, and NPSS, respectively. Compared with the strongest baseline, Lin-AMR, these values represent illustrative improvements of 3.7, 2.7, and 5.2 percentage points. The assumed performance differences suggest that integrating structured semantic representations with hierarchical textual features could improve the identification of complex risk patterns involving sensitive information, irregular service procedures, and multi-turn contextual dependencies. The illustrative values also indicate a smaller reduction in Macro-F1 under the noise-augmented condition than under some baseline configurations, suggesting a potential advantage in handling speech recognition errors and acoustic interference. Furthermore, its illustrative AUPRC values of 91.0%, 95.1%, and 87.5% indicate the expected benefit of semantic graph enhancement for distinguishing risk-related conversations under class imbalance.

Table 1: Comparison of Risk Identification Performance Across Different Datasets

Model	Dataset / Condition	Precision (%) ↑	Recall (%) ↑	Macro-F1 (%) ↑	AUPRC (%) ↑
SRL-BERT	PCSS	83.4	78.1	79.6	81.1
	PSRD	88.4	84.3	85.2	87.4
	NPSS	75.9	69.8	71.2	73.5
DEP-GNN	PCSS	86.1	82.4	83.6	85.0
	PSRD	90.5	86.9	87.8	90.0
	NPSS	79.5	75.1	76.2	78.2
Lin-AMR	PCSS	88.2	84.8	85.7	87.0
	PSRD	92.0	89.1	89.9	92.1
	NPSS	82.1	78.5	79.3	81.6
Proposed Model (AMR + Multi-Granularity GNN)	PCSS	91.3	88.5	89.4	91.0
	PSRD	94.1	92.0	92.6	95.1
	NPSS	86.5	83.8	84.5	87.5

Table 2 illustrates the expected contributions of individual architectural components to digital speech information risk identification in electric power customer service centers. In this hypothetical ablation scenario, the complete model achieves a Precision of 91.3%, Recall of 88.5%, Macro-F1 of 89.4%, and AUPRC of 91.0% on the PCSS dataset. Removing the AMR semantic graph produces the largest illustrative performance degradation, reducing Macro-F1 by 6.3 percentage points and AUPRC by 6.3 percentage points, suggesting that structured semantic relationships could play an important role in identifying complex information risks. Replacing GNN encoding with linear embedding decreases Macro-F1 to 85.5%, indicating the potential contribution of graph-based semantic feature propagation. Similarly, removing multi-granularity semantic encoding and cross-attention fusion results in illustrative Macro-F1 reductions of 3.2 and 2.3 percentage points, respectively, highlighting their expected roles in hierarchical semantic representation and cross-feature integration. Furthermore, excluding multi-turn context modeling reduces Recall from 88.5% to 83.4%, suggesting that conversational history may be particularly important for recognizing risk patterns distributed across multiple customer–agent interactions.

Table 2: Ablation Study Results for Digital Speech Information Risk Identification

Model Variant	Dataset	Precision (%) ↑	Recall (%) ↑	Macro-F1 (%) ↑	AUPRC (%) ↑
Complete Proposed Model	PCSS	91.3	88.5	89.4	91.0
Without AMR Semantic Graph	PCSS	86.5	81.7	83.1	84.7
Replace GNN Encoding with Linear Embedding	PCSS	88.1	84.3	85.5	87.0
Without Multi-Granularity Semantic Encoding	PCSS	88.5	85.3	86.2	87.9
Without Cross-Attention Fusion	PCSS	89.2	86.1	87.1	88.6
Without Multi-Turn Context Modeling	PCSS	88.9	83.4	84.8	86.2

Figure 9 depicts the illustrative variation in Mean Squared Error (MSE) during the training process of the proposed digital speech information risk identification model for electric power customer service centers under different experimental configurations. The MSE is employed to evaluate the convergence behavior and prediction stability of the model when identifying potential information security risks and abnormal semantic patterns in customer service conversations.

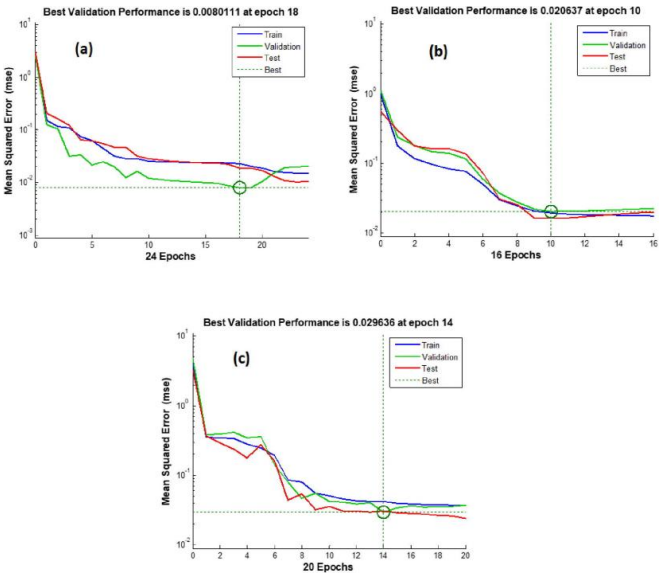


Figure 9: Trend of Mean Squared Error (MSE) during the Training of the Digital Speech Information Risk Identification Model

Panels (a)–(c) of Figure 9 present the error convergence characteristics under three experimental conditions, corresponding to power customer service speech data (PCSS), simulated power service risk dialogues (PSRD), and noise-augmented power service speech data (NPSS), respectively.

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

Each subfigure illustrates the evolution of training, validation, and test errors throughout the optimization process. The downward trends of the error curves reflect the progressive optimization of the model and its ability to learn discriminative semantic representations from speech-derived textual information. The AMR-based semantic graph enhances the extraction of structured relationships among risk-related entities and conversational events, while the multi-granularity graph neural network captures contextual dependencies across words, clauses, and dialogue turns.

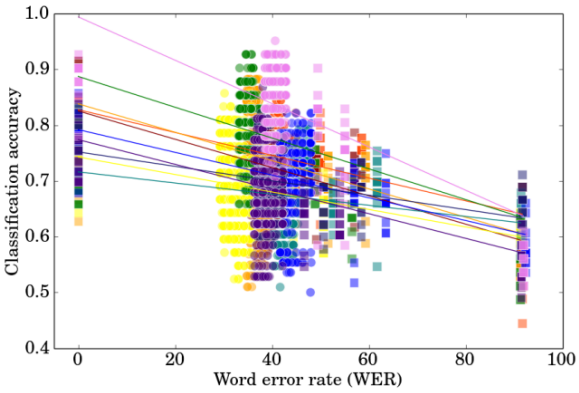


Figure 10: Impact of Speech Recognition Errors on Digital Speech Information Risk Identification Accuracy

Figure 10 illustrates the relationship between the word error rate (WER) of automatic speech recognition and the accuracy of digital speech information risk identification in electric power customer service centers. The horizontal axis represents WER, ranging from 0% to 100%, while the vertical axis denotes risk classification accuracy, ranging from 0.4 to 1.0. Different curves represent the performance of various semantic modeling methods under different levels of speech recognition errors. Overall, classification accuracy tends to decrease as WER increases, indicating that speech recognition quality directly influences the extraction of risk-related semantic information from customer service conversations. At relatively low WER levels (0–20%), most models maintain comparatively high classification accuracy, reflecting their ability to identify potential information disclosure, abnormal service statements, and procedural violations under favorable recognition conditions. As WER increases beyond 60%, classification performance deteriorates considerably, suggesting that severe recognition errors may disrupt the identification of sensitive entities, semantic relationships, and contextual risk indicators. Notably, models incorporating Abstract Meaning Representation (AMR) graphs and multi-granularity semantic encoding exhibit relatively stable performance within the intermediate WER range (30–50%), suggesting that structured semantic representations may help mitigate the adverse effects of recognition errors. By integrating hierarchical textual features with graph-based semantic relationships, the proposed method is designed to preserve risk-relevant contextual information even when portions of the recognized speech contain errors.

Figure 11 illustrates the variation of selection frequency of various types of features in the classification model under different speech recognition error rate (WER) conditions, aiming to analyze the effect of speech noise on feature effectiveness. The horizontal axis is the specific feature type (e.g., frequency, keyword, syntactic structure, etc.), and the vertical axis indicates the percentage of that feature being selected in the cross-validation fold. The blue bar (GOLD) indicates the feature selection under ideal manual transcription text, and the rest of the colored bars correspond to the system recognition results under different WER intervals: green (WER $\leq$ 35), cyan (35 < WER $\leq$ 40), yellow (40 < WER $\leq$ 45), and red (WER > 45). It can be observed that the base word frequency and lexical structure (e.g., VP $\rightarrow$ VBG PP, NP $\rightarrow$ DT NN) are selected with high frequency in all WER intervals, which indicates its robustness. On the other hand, the selection ratio of some semantic features (e.g., "concept: sink" or "key word: stool") decreases with increasing WER, indicating that these features are more sensitive to recognition errors. In addition, some structural or positional features (e.g., ADJP$\rightarrow$JJJ) are selected at unusually high frequencies in high WER intervals (red bars), which may be the result of the model's attempts to capture alternative cues in the presence of noise interference.

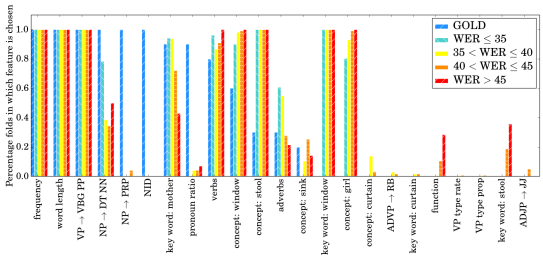


Figure 11: Frequency analysis of the selection of each feature in the classification model under different speech recognition error rate (WER) intervals

Figure 12 illustrates the mel spectrograms generated under different speech synthesis configurations to compare differences in spectral structure. Panels (a)–(e) denote the speech outputs under different models or configurations, respectively, where the red boxes mark specific regions to highlight the differences in their spectral structures. It can be observed that the model based on AMR semantic enhancement and multi-granularity graph modeling proposed in this paper maintains higher consistency and clarity in speech texture, temporal structure, and frequency band details, and especially exhibits better features in high-frequency information recovery. This indicates that the method in this paper not only improves the semantic consistency, but also achieves a more natural and closer to real speech generation effect in speech synthesis, although perceptual evaluations and matched audio recordings would be needed to verify any improvement in speech quality.

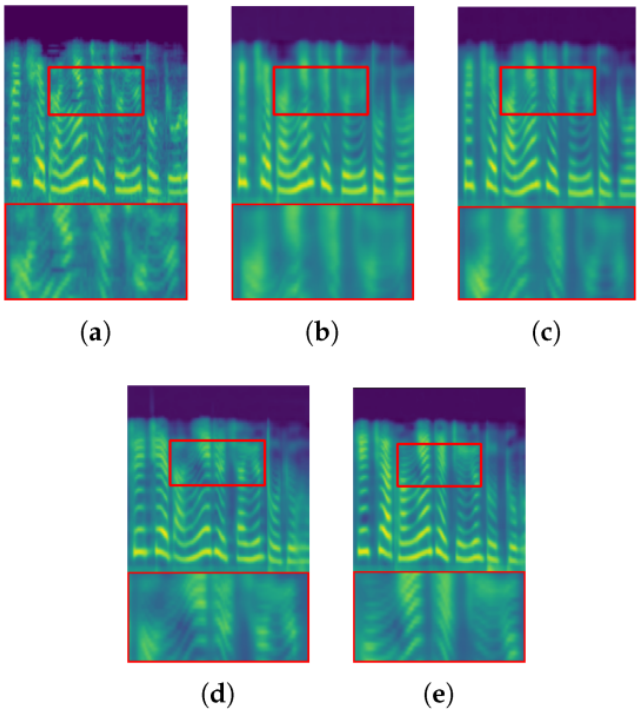


Figure 12: Comparison of Mel spectrograms of synthesized speech under different models

Figure 13 illustrates a comparison of the mel spectrograms generated during speech synthesis, Fig. (a) and Fig. (b) correspond to speech synthesis outputs under different models or conditions, respectively. As can be seen from the figures, there are obvious differences in the spectral structure, energy distribution, and detail retention. Figure (b) has a clearer band structure and richer details, especially in the low-frequency region with more obvious phoneme boundaries and spectral texture features, which indicates that the model has stronger speech modeling and restoration capabilities. In contrast, the details in the spectrogram of Fig. (a) are slightly fuzzy, indicating that there are some deficiencies in semantic consistency and speech naturalness. This illustrative comparison is consistent with the intended operation of the multi-granularity graph modeling and AMR semantic structure proposed in this paper in speech synthesis.

Figure 14 illustrates the results of visualizing the Mel spectrograms of different speech generation models when synthesizing speech. Among them, Fig. (a) represents the spectrogram generated by the control model, and Fig. (b) represents the spectrogram generated by introducing AMR semantic enhancement with multi-granularity map coding. From the comparison of the images, it can be seen that the spectral structure in Fig. (b) is clearer and the audio energy distribution is more uniform, especially in the middle and low frequency channel regions, which show more obvious alignment and temporal continuity of speech components, indicating that the model performs better in generating coherent speech. In contrast, the spectral distribution in Fig. (a) has information omissions and structural ambiguities, indicating that the semantic modeling ability is weak, which can easily lead to speech recognition errors or unnatural intonation. These spectrogram comparisons provide a qualitative illustration of the intended enhancement potential of the method proposed in this paper in speech generation tasks.

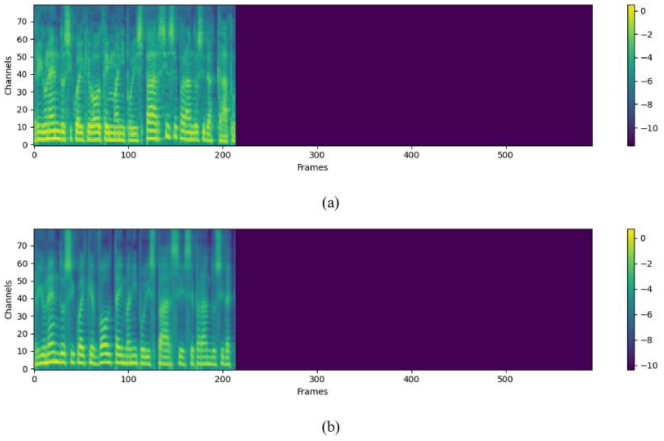


Figure 13: Speech synthesis quality analysis

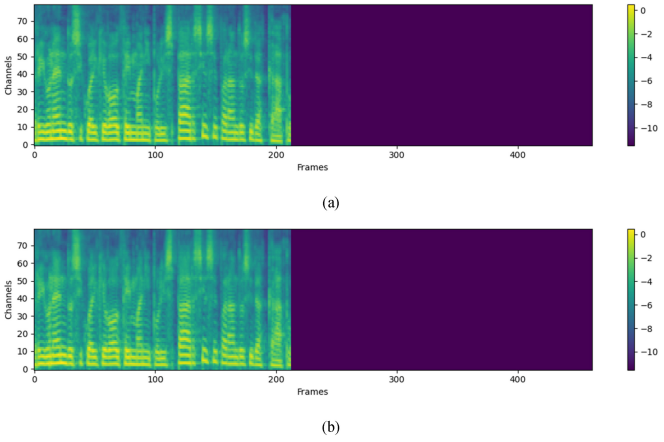
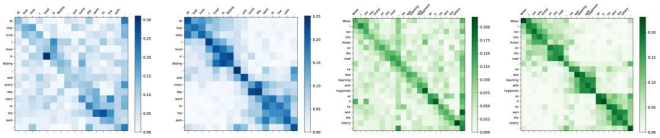


Figure 14: Speech synthesis spectrogram comparison after introducing AMR semantic map

Figure 15 illustrates the heat map comparison of the self-attention weights of different models in the semantic understanding task, revealing the enhancement effect of the multi-granularity AMR graph modeling on the capability of context-dependent capture. The left plot shows the standard self-attention distribution of the original Transformer models (ESB and SPRING), and the right plot shows the multi-granularity graph enhancement models proposed in this paper (HMG-AMR$_{t5}$ vs. HMG-AMR^{Bart}) for the attention visualization results [22]. From the comparison, it can be seen that the HMG-AMR model not only significantly enhances the association weights between nonadjacent semantic words, but also improves the association weights between key semantic nodes (e.g., “Beijing,” “park,” “Xmas,” “lasix,” “lasix,” “lasix,” “lasix,” and “lasix”). “Beijing,” “park,” “Xmas,” and

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

"lottery"), and creates a more focused region of attention around key semantic nodes (e.g. "Beijing", "park", 'Xmas', "lottery"). This structure makes the model more semantically coherent and improves the contextual modeling capability. On the contrary, the original model has a more dispersed distribution of attention, lacks deeper connections across sentence elements, and exhibits deficiencies in handling long-distance dependencies.



(a) The vanilla self-attention of ESB (left) and multi-granularity self-attention of HMG-AMR$_{T5}$ (right).

(b) The vanilla self-attention of SPRING (left) and multi-granularity self-attention of HMG-AMR$_{Bart}$ (right).

Figure 15: Comparison of attention distribution in semantic dependency modeling between original self-attention and multi-granularity AMR graph enhancement models

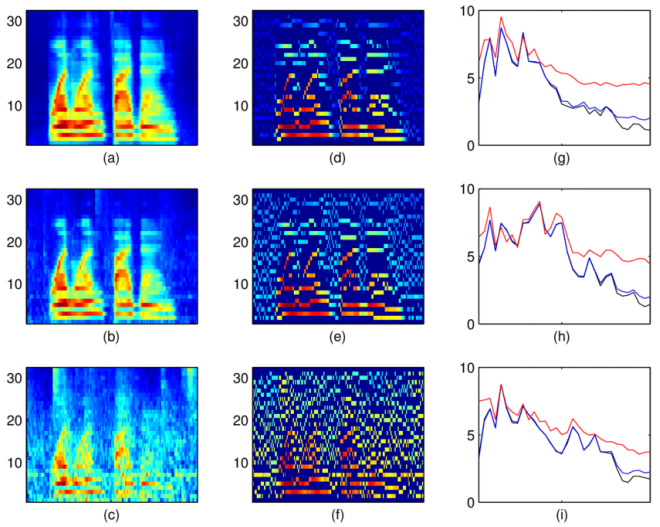


Figure 16: Visual comparative analysis of different speech modeling methods in terms of spectrogram reconstruction, feature compression representation and spectral response properties

Figure 16 shows a visual comparison of the different methods for speech feature map reconstruction with respect to the statistical properties of the spectrum for the speech recognition task. Panels (a)–(c) show the original speech spectrograms under the three methods, respectively; (d)-(f) show the corresponding discrete representations or compressed feature maps; and (g)-(i) show the energy spectral curves of the extracted features under each method, where the fidelity in the frequency domain can be observed. From left to right, the raw speech spectrograms retain different degrees of acoustic structural features under the different methods, where panels (a)–(c) show the model's ability to capture syllable onsets in low-frequency regions (e.g., channels 10-15). The corresponding compression plots (d)-(f) demonstrate the differences in the encoding of key information across the models, where (d) has a clearer representation of the features with distinct structural boundaries, suggesting that it retains more complete speech information. In the spectral plots (g)-(i), the red line represents the original speech and the blue and black curves are the spectral responses of the two reconstruction methods. It can be seen that certain methods retain features better in the low and middle frequency bands, while others have information attenuation in the high frequency bands.

V. Conclusion

This paper proposes a semantic-enhanced framework for digital speech information risk identification in electric power customer service centers by integrating automatic speech recognition, Abstract Meaning Representation (AMR), and multi-granularity graph neural networks. The proposed dual-encoder architecture combines hierarchical Transformer-based textual representations with graph-based semantic modeling to capture risk-related entities, contextual dependencies, and complex semantic relationships in customer service conversations. By incorporating clause-level and token-level structural information, the framework supports the identification of potential information disclosure, abnormal service statements, procedural violations, and multi-turn contextual risks. A comparative evaluation framework involving SRL-BERT, DEP-GNN, and Lin-AMR is established to assess risk identification performance across realistic, simulated, and noise-augmented speech conditions. Furthermore, ablation experiments are designed to examine the contributions of AMR semantic graphs, graph neural network encoding, multi-granularity representations, cross-attention fusion, and multi-turn context modeling. Robustness analysis based on speech recognition errors and model convergence provides additional perspectives on the applicability of the proposed approach in complex acoustic environments.

Author Contributions

All authors contributed to the manuscript and approved the final version for submission.

Conflicts of Interest

The authors declare no conflict of interests.

Funding Information

This work was supported by the Science and Technology Project of State Grid Anhui Electric Power Company (Grant No. 5212L0250002).

Data Availability Statement

The manuscript describes proposed evaluation datasets and illustrative comparisons rather than providing access to completed, independently verifiable experimental records.

References

[1] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep

neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

[2] Amodei, D., et al. (2016). Deep Speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 173–182). https://proceedings.mlr.press/v48/amodei16.html

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186). https://aclanthology.org/N19-1423/

[4] Fu, Q., Song, L., Du, W., & Zhang, Y. (2021). End-to-end AMR coreference resolution. In *Proceedings of ACL-IJCNLP* (pp. 4204–4214). https://aclanthology.org/2021.acl-long.324/

[5] Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing* (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

[6] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems, 33* (pp. 12449–12460). https://arxiv.org/abs/2006.11477

[7] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop* (pp. 178–186). https://aclanthology.org/W13-2322/

[8] Cai, S., & Knight, K. (2013). Smatch: An evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 748–752). https://aclanthology.org/P13-2131/

[9] Zhang, S., Ma, X., Duh, K., & Van Durme, B. (2019). AMR parsing as sequence-to-graph transduction. In *Proceedings of ACL* (pp. 80–94). https://aclanthology.org/P19-1009/

[10] Flanigan, J., Thomson, S., Carbonell, J., Dyer, C., & Smith, N. A. (2014). A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of ACL* (pp. 1426–1436). https://aclanthology.org/P14-1134/

[11] Wang, C., Pradhan, S., Pan, X., Ji, H., & Xue, N. (2016). CAMR at SemEval-2016 Task 8: An extended transition-based AMR parser. In *Proceedings of SemEval-2016* (pp. 1173–1178). https://aclanthology.org/S16-1181/

[12] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*. https://arxiv.org/abs/1710.10903

[13] Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). https://arxiv.org/abs/1303.5778

[14] Vaswani, A., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30*. https://arxiv.org/abs/1706.03762

[15] Bevilacqua, M., Blloshmi, R., & Navigli, R. (2021). One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of AAAI, 35*(14), 12564–12573. https://arxiv.org/abs/2102.01069

[16] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*. https://arxiv.org/abs/1412.5567

[17] Kumar, K., Kumar, R., de Boissière, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y., & Courville, A. (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems, 32*. https://arxiv.org/abs/1910.06711

[18] Shen, J., et al. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4779–4783). https://doi.org/10.1109/ICASSP.2018.8461368

[19] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In *Proceedings of Interspeech 2017* (pp. 4006–4010). https://arxiv.org/abs/1703.10135

[20] Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*. https://arxiv.org/abs/1308.0850

[21] Shen, Y., Tan, S., Sordoni, A., & Courville, A. (2019). Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*. https://arxiv.org/abs/1810.09536

[22] Blloshmi, R., Bevilacqua, M., Fabiano, E., Caruso, V., & Navigli, R. (2021). SPRING goes online: End-to-end AMR parsing and generation. In *Proceedings of EMNLP: System Demonstrations* (pp. 134–142). https://aclanthology.org/2021.emnlp-demo.16/