RVQ-SNER: End-to-End Chinese Speech Named Entity Recognition via Quantized Acoustic Bottlenecks and Deep Acoustic-Semantic Fusion

Yaoqiang Zhou and Xudong Luo(✉)

School of Computer Science and Engineering,
Ministry-of-Education Key Lab of Education Blockchain and Intelligent Technology,
Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China
luoxd@mailbox.gxnu.edu.cn

Abstract. Conventional Speech Named Entity Recognition (SNER) typically relies on cascaded ASR (Automatic Speech Recognition)+NER (Named Entity Recognit ion) pipelines, which are hindered by error propagation and the underutilisation of acoustic cues. We propose an end-to-end Chinese SNER framework using Residual Vector Quantisation (RVQ) and deep acoustic-semantic fusion. The model extracts speech representations via a frozen Wav2Vec2-XLSR encoder, employing an RVQ-based bottleneck to reconstruct continuous quantized features that regularize the acoustic space and preserve semantic content. A Transformer decoder, trained with a joint CTC-attention objective, performs transcription while a gated deep-fusion mechanism integrates an external GPT model for linguistic consistency. For NER, a bidirectional multimodal fusion module aligns acoustic and semantic features before a GlobalPointer head performs span-level prediction. Experiments on AISHELL-NER and CNERTA yield F1-scores of 90.91% and 81.44%, respectively, outperforming pipeline, multimodal, and E2E baselines. These results demonstrate that deep bidirectional interaction between quantized acoustic streams and semantic contexts is essential for mitigating ASR error propagation and achieving robust Chinese SNER.

Keywords: Natural language processing, End to end named entity recognition, Multimodal, Cross-modal attention, Automatic speech recognition.

1 Introduction

Named Entity Recognition (NER) [1-4] is a core Natural Language Processing (NLP) task that identifies and classifies entity mentions, such as persons and locations, from unstructured data [5]. As a vital information extraction component, NER supports downstream applications including information retrieval [6-8] and knowledge graph construction [9]. With the proliferation of voice-interactive technologies [10], Speech Named Entity Recognition (SNER) has become a key challenge in spoken language understanding [11-15]. Unlike text-based NER, SNER identifies entities directly from

audio. In Chinese, this task is complicated by accent variations, speaking-rate fluctuations, and ambiguous entity boundaries [16], demanding robust acoustic and semantic modelling.

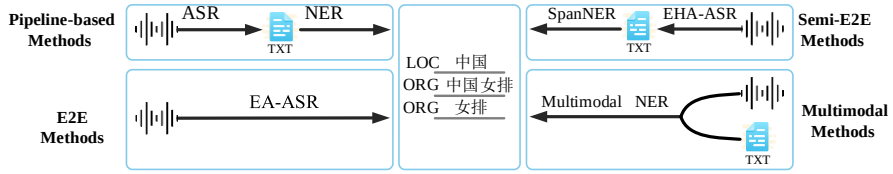


Fig. 1. Architectural comparison of three approaches for Speech NER.

As shown in Fig. 1, Existing SNER methods generally follow four paradigms pipeline-based, end-to-end (E2E), semi-E2E, and multimodal approaches. Pipeline methods are modular but suffer from Automatic Speech Recognition (ASR) error propagation [16], while E2E models mitigate this by jointly modelling acoustic and semantic information but face complex optimisation challenges. Multimodal methods attempt to incorporate prosodic cues, yet many current fusion strategies remain too shallow for effective interaction [17]. Overall, error accumulation and insufficient acoustic-semantic alignment remain primary obstacles.

To address these, in this paper we propose RVQ-SNER, an E2E Chinese SNER framework using Residual Vector Quantisation (RVQ) and deep acoustic-semantic fusion. Utilising an encoder-quantiser-decoder-fusion architecture, the model employs a pretrained Wav2Vec2 encoder and an RVQ module to suppress irrelevant acoustic variation. Quantized features are processed by a Transformer-based ASR decoder, followed by a multimodal fusion module and a GlobalPointer head [18] for span-level prediction.

Our main contributions are: 1) a quantisation-bottleneck-based E2E architecture that regularises the latent space using RVQ for improved robustness; 2) a deep acoustic-semantic fusion mechanism to explicitly align features for complex and nested entity recognition; 3) a multi-task training strategy that jointly optimises ASR and NER to enhance alignment quality; and 4) superior empirical performance on AISHELL-NER and CNERTA, outperforming various pipeline and E2E baselines.

2 Related Work

2.1 Speech Recognition Modelling

Early ASR systems relied on statistical models such as Hidden Markov Models (HMMs) [19] combined with GMMs or CRFs [20]. These approaches were later replaced by hybrid HMM-DNN architectures [21], where neural networks improved acoustic feature extraction. The field then moved toward E2E methods, including Connectionist Temporal Classification (CTC) [22], RNN-T [23], and Attention-based Encoder-Decoder (AED) models [24]. By mapping acoustic features directly to text, these methods simplified multi-stage pipelines, while joint CTC-attention frameworks [25]

further improved alignment and training stability. More recently, self-supervised learning has significantly advanced ASR significantly. Models such as wav2vec 2.0 and HuBERT, pretrained on large amounts of unlabeled data, learn rich acoustic representations that enhance accuracy and reduce reliance on labeled data [26], providing a strong foundation for downstream spoken language understanding tasks such as SNER.

2.2 Research on Chinese SNER

Chinese SNER has evolved from traditional transcription-based pipelines toward integrated E2E and multimodal frameworks. Compared with text-based NER, it must address speech variability, ASR errors, and ambiguous entity boundaries caused by the absence of explicit word delimiters. Early multimodal works, such as CNERTA and M3T [27], showed that acoustic cues (e.g., pauses and stress) can aid boundary detection. Subsequent studies on AISHELL-NER compared two-stage pipelines with E2E alternatives like entity-aware ASR. Although pipelines provide engineering stability, they suffer from cascading errors; recent efforts mitigate this by introducing Pinyin embeddings and adaptive homophone noise injection to improve robustness to transcription inaccuracies [28].

The field is now increasingly shifting toward generative E2E modeling. Recent approaches reformulate SNER as a prompt-based multi-task problem using Whisper encoders, or explore cross-modal alignment by integrating Whisper [29] with LLMs such as ChatGLM3 [30]. Despite these advances, current methods remain limited by residual ASR error propagation in pipeline-style designs and insufficient acoustic-semantic interaction in multimodal architectures.

2.3 Applications of Quantisation in Speech Processing

Vector Quantisation (VQ), introduced by VQ-VAE [31], converts continuous latent features into discrete codebook tokens, while RVQ [32,33] extends this idea with multi-stage residual encoding to achieve higher fidelity and more expressive discrete sequences. These quantisation techniques have mainly been used in speech and audio generation systems—such as LauraGPT [34], VALL-E [35], and UniAudio [36] — where discrete codecs balance compression and detail, and have also shown effectiveness in tasks like target-speaker [37] extraction and music generation [38]. In contrast, their application to speech understanding tasks, including SNER, remains limited. Recent work like LongCat-Audio-Codec [39] indicates that separating semantic and acoustic tokenisation can benefit ASR, suggesting that quantized representations may bridge acoustic modelling and semantic reasoning. Motivated by this, we explore whether RVQ can function as an acoustic bottleneck for Chinese SNER, enabling robust acoustic-semantic interaction within E2E framework.

3 Model

This section details our model.

3.1 Overall Structure

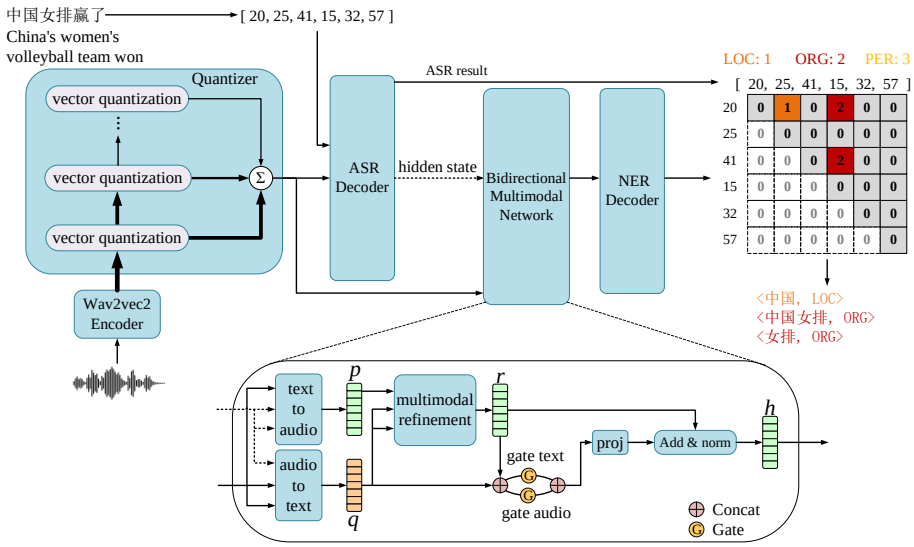


Fig. 2. Overall architecture of the proposed model.

As illustrated in Fig. 2, the model consists of five components: a Wav2Vec2 encoder, an RVQ bottleneck, an ASR decoder, a bidirectional multimodal network and a NER decoder. The pipeline first extracts high-dimensional speech representations from raw waveforms. An RVQ module then serves as an information bottleneck to suppress irrelevant acoustic variation. Unlike models that pass discrete indices, our framework uses reconstructed continuous feature vectors aggregated across multiple layers. Then, the ASR decoder processes these features to predict text tokens and produce high-level semantic hidden states. A multimodal fusion module then integrates these semantic states with the quantized acoustic features. Finally, a GlobalPointer head predicts a category-specific entity label matrix for span-level recognition. During inference, entities are recovered by mapping the matrix's boundary coordinates onto the generated text.

3.2 Acoustic Encoder

We employ the pretrained wav2vec2-large-xlsr-53-chinese-zh-cn[1] model to extract acoustic representations from raw speech. Given an input raw waveform $X = \{x_1, \ldots, x_T\}$, the encoder is kept frozen during training and inference to preserve pretrained knowledge and avoid overfitting, serving as a deterministic feature extractor:

$$H_{enc} = \text{Wav2Vec2}(X) \in \mathbb{R}^{T \times D}, \tag{1}$$

[1] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-Chinese-zh-cn

where T and D denote the temporal length and hidden dimension, respectively. These features H_{enc} form the input for the subsequent RVQ module.

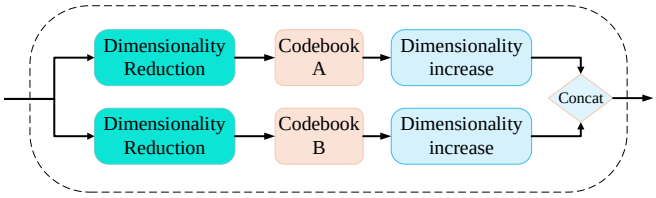


Fig. 3. Architecture of a single quantiser.

3.3 RVQ Module

To mitigate overfitting caused by high-dimensional acoustic features and to impose structured constraints on the latent space, we introduce a grouped residual vector quantization (GRVQ) module between the acoustic encoder and the ASR decoder. Importantly, although discrete codebooks are used internally, the quantization process operates entirely on continuous representations and outputs continuous features, ensuring compatibility with downstream neural components.

We initialize the residual with the encoder output:

$$\mathbf{r}_0 = \mathbf{H}_{enc}. \tag{2}$$

The RVQ module consists of L cascaded quantization stages. At each stage l, the model quantizes the residual from the previous stage. A single vector quantizer is illustrated in Fig. 3. At the l-th stage, the residual feature $\mathbf{r}_{l-1}$ is first projected into two continuous low-dimensional subspaces:

$$\mathbf{z}_{l,a} = f_{l,a}(\mathbf{r}_{l-1}), \quad \mathbf{z}_{l,b} = f_{l,b}(\mathbf{r}_{l-1}), \tag{3}$$

where $f_{l,a}$ and $f_{l,a}$ denote learnable linear projections.

Each sub-vector is then quantized using a cosine-similarity-based codebook. For a given input z, we first perform ℓ_2 normalization:

$$\tilde{\mathbf{z}} = \frac{\mathbf{z}}{||\boldsymbol{z}||_2}, \quad \tilde{\mathbf{e}}_k = \frac{\mathbf{e}_k}{||\mathbf{e}_k||_2}, \tag{4}$$

and select the nearest codeword:

$$k^* = arg\min_{\mathrm{k}} ||\tilde{\mathbf{z}} - \tilde{\mathbf{e}}||_2^2. \tag{5}$$

The quantized outputs are obtained via embedding lookup:

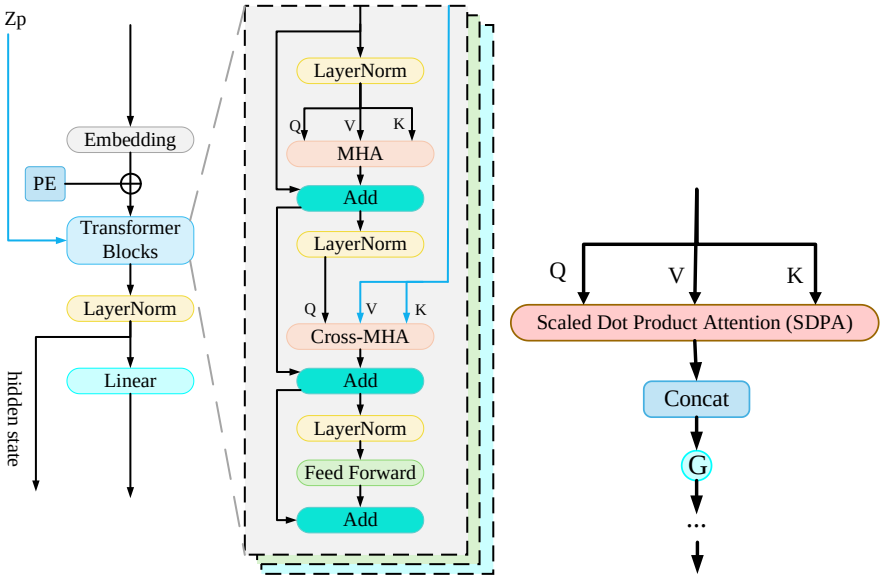
$$\mathbf{q}_{l,a} = \mathbf{e}_{k_a^*}, \quad \mathbf{q}_{l,b} = \mathbf{e}_{k_b^*}. \tag{6}$$

To enable E2E training, we apply a straight-through estimator (STE):

$$\hat{\mathbf{z}}_{l,a} = \hat{\mathbf{z}}_{l,a} + \mathrm{sg}\big(\mathbf{q}_{l,a} - \mathbf{z}_{l,a}\big), \tag{7}$$

$$\hat{\mathbf{z}}_{l,b} = \hat{\mathbf{z}}_{l,b} + \mathrm{sg}\big(\mathbf{q}_{l,b} - \mathbf{z}_{l,b}\big), \tag{8}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Notably, the representations remain continuous after this operation.



(a) Decoder architecture (b) Gated attention mechanism

Fig. 4. Overview of the decoder and gated attention mechanism. The proposed decoder architecture is optimised with a joint CTC-attention objective and enhanced by an external GPT language model, which is omitted from this simplified diagram for clarity.

The two quantized sub-representations are then concatenated and projected back to the original space:

$$\mathbf{q}_l = g_l([\tilde{\mathbf{z}}_{l,a}; \tilde{\mathbf{z}}_{l,b}]). \tag{9}$$

The residual is updated recursively:

$$\mathbf{r}_l = \mathbf{r}_{l-q} - \mathbf{q}_l. \tag{10}$$

This hierarchical process allows early stages to capture coarse structures, while later stages refine finer residual details.

After L stages, the final representation is reconstructed as:

$$\mathbf{Z}_q = \sum_{l=1}^{L} \mathbf{q}_l. \tag{11}$$

Importantly, $\mathbf{Z}_q \in \mathbb{R}^{T \times D}$ remains a continuous feature representation rather than a sequence of discrete indices.

To stabilize codebook learning, we adopt exponential moving average (EMA) updates:

$$n_k \leftarrow \lambda n_k + (1-\lambda)\sum_i \mathbb{I}[k_i = k], \tag{12}$$

$$\mathbf{m}_k \leftarrow \lambda \mathbf{m}_k + (1-\lambda)\sum_{i:k_i=k} \mathbf{z}_i, \tag{13}$$

$$\mathbf{e}_k = \frac{\mathbf{m}_k}{n_k+\epsilon} \tag{14}$$

In addition, rarely activated codewords are periodically reinitialized to prevent codebook collapse and maintain high utilization.

The ASR decoder employs a joint CTC-attention architecture [25] to ensure alignment stability. It consists of a Transformer-based attention branch for autoregressive dependency modelling and an auxiliary CTC branch to impose monotonic constraints. The joint training objective is:

$$\mathcal{L}_{\text{ASR}} = \alpha\mathcal{L}_{\text{ctc}} + (1-\alpha)\mathcal{L}_{\text{att}}. \tag{15}$$

As shown in Fig. 4(a), the decoder maps quantized acoustic memory $\mathbf{Z}_q$ into the textual space. We incorporate several architectural refinements to enhance feature selection and semantic modelling: Following [40], we introduce a fine-grained gating mechanism into the self-attention and cross-attention modules (see Fig. 4(b)). After scaled dot-product computation, an element-wise gating operation modulates the output along the feature dimension. This enables input-dependent control of information flow, improving robustness under complex acoustic conditions.

To leverage linguistic priors, we integrate a pretrained GPT model via a gated deep fusion strategy. At each step, a gating vector gt is computed to adaptively combine the decoder state h_t^{dec} with the language model output:

$$h_t^{\text{LM}} = \left[h_t^{\text{dec}}; g_t \odot h_t^{\text{LM}}\right] + b_f, \tag{16}$$

where

$$g_t = \sigma(W_g\left[h_t^{\text{dec}}; h_t^{\text{LM}}\right] + g_g), \tag{17}$$

$$h_t^{\text{dec}} = W_f. \tag{18}$$

This allows the model to rectify ASR errors using contextual knowledge from large-scale text corpora.

We replace standard Layer Normalisation with RMSNorm [41] for improved numerical stability and training efficiency. Furthermore, the feed-forward network (FFN) adopts the SwiGLU structure [42]:

$$\text{FFN}_{\text{SwiGLU}}(x) = (\text{SiLU}(xW_1) \odot xW_3)W_2. \tag{19}$$

SwiGLU enhances non-linear feature interaction, leading to better representational capacity and faster convergence.

3.4 Bidirectional Multimodal Synergy and Feature Interaction

Unlike language model fusion within the ASR decoder, we design a bidirectional multi-source NER fusion module specifically for downstream entity recognition, as illustrated in Fig. 2. The module integrates acoustic features from the quantizer with textual hidden states from the decoder to mitigate the impact of ASR errors. Although the overall framework is trained E2E, the NER component still depends heavily on decoder-side representations, which may be degraded by homophone substitutions, deletions, and other recognition errors. To address this limitation, the proposed module performs explicit acoustic-textual interaction to enhance semantic robustness. Inspired by the bidirectional multimodal synergy and feature interaction framework in [43], we further introduce two refinements. First, instead of fusing text-to-audio representations with the original textual features, we adopt a second-stage cross-modal attention mechanism, where text-to-audio representations are treated as queries and audio-to-text representations act as keys and values, enabling richer bidirectional interaction between modalities. Second, a dual-gating mechanism is incorporated to adaptively regulate modality-specific contributions, improving feature selectivity and enhancing robustness under noisy ASR conditions.

Formally, let the acoustic representations be denoted by $\mathbf{H}^a \in \mathbb{R}^{T_a \times D}$ and the textual hidden states by $\mathbf{H}^t \in \mathbb{R}^{T_t \times D}$. A bidirectional cross-modal encoder is then constructed to model the interaction between the two modalities in three stages. First, in the text-to-audio stage, textual representations attend to acoustic features:

$$\mathbf{H}^{t\to a} = \mathrm{CrossAttn}(\mathbf{H}^t, \mathbf{H}^a, \mathbf{H}^a). \tag{20}$$

Secondly, in the audio-to-text stage, acoustic representations attend to textual context:

$$\mathbf{H}^{a\to t} = \mathrm{CrossAttn}(\mathbf{H}^a, \mathbf{H}^t, \mathbf{H}^t). \tag{21}$$

Finally, the cross-modally enhanced representations are further integrated:

$$\mathbf{H}^{\mathrm{fusion}} = \mathrm{CrossAttn}(\mathbf{H}^{t\to a}, \mathbf{H}^{a\to t}, \mathbf{H}^{a\to t}). \tag{22}$$

To adaptively combine information from different modalities, we introduce a gating mechanism. The gating coefficients are computed through non-linear projections over concatenated features:

$$\mathrm{g}_t = \sigma(\mathbf{W}_t[\mathbf{H}^{\mathrm{fusion}}; \mathbf{H}^{t\to a}]), \tag{23}$$

$$\mathrm{g}_a = \sigma(\mathbf{W}_a[\mathbf{H}^{\mathrm{fusion}}; \mathbf{H}^{t\to a}]). \tag{24}$$

Finally, the fused representation is obtained as:

$$\mathbf{H}^* = \mathrm{LayerNorm}\left(\mathbf{H}^{\mathrm{fusion}} + \mathrm{Proj}\left(\mathrm{g}_t \odot \mathbf{H}^{\mathrm{fusion}}; \mathrm{g}_a \odot \mathbf{H}^{\mathrm{fusion}}\right)\right). \tag{25}$$

The effectiveness of this fusion mechanism depends strongly on the quality of the textual hidden states generated by the decoder. However, because the decoder is auto-regressive, a discrepancy arises between training and inference. During training, the decoder typically operates under teacher forcing, where hidden states are conditioned

on ground-truth token sequences. During inference, by contrast, the model must generate tokens based on its own previous predictions, which can lead to error accumulation and degraded textual representations, ultimately weakening the fused features used for NER.

To mitigate this issue, we introduce a dual-path curriculum learning strategy during training. Specifically, two parallel branches are constructed:

- Branch A uses ground-truth-conditioned representations, whilst Branch B relies on self-decoded representations. In the early stages of training, the model depends primarily on Branch A to stabilise optimisation.
- As training progresses, the contribution of Branch B is gradually increased, allowing the model to adapt to noisier and more realistic inputs.

Let e denote the current training epoch and E the curriculum schedule length.

- During the curriculum phase $e < E$, the weight of Branch A is updated using a linear annealing schedule:

$$\omega_a(e) = \lambda_{\text{start}} - (\lambda_{\text{start}} - \lambda_{\text{end}}) \cdot \frac{e}{E}, \tag{26}$$

and the weight of Branch B is defined as

$$\omega_b(e) = 1 - \omega_a(e). \tag{27}$$

- When $e \geq E$, the weights remain fixed:

$$\omega_a(e) = \lambda_{\text{end}},\ \omega_b(e) = 1 - \lambda_{\text{end}}. \tag{28}$$

This strategy progressively increases the contribution of the self-decoding branch, enabling the model to become more robust to noisy textual representations in later training stages and thereby improving resistance to ASR errors during inference.

3.5 NER Decoder

For entity prediction, we adopt the GlobalPointer head proposed by Su et al. [18]. Unlike conventional sequence labelling methods, which perform token-level tagging and often struggle with nested or overlapping entities, GlobalPointer formulates NER as a two-dimensional span prediction task by directly scoring all possible start-end position pairs. This design allows the model to identify entity boundaries in parallel and naturally supports complex entity structures without introducing the structural constraints typically associated with sequential labelling approaches.

4 Experiments

4.1 Datasets

We evaluate our framework on two public benchmark datasets derived from the AISHELL-1 Mandarin corpus [44]:

- AISHELL-NER [45]: It comprises 16 kHz studio-recorded speech across five domains, including finance, technology, sports, entertainment, and news. Three entity categories—Person (PER), Location (LOC), and Organisation (ORG)—are annotated using the BIO scheme, with additional special symbols (e.g., [], (), <>) for entity-aware ASR.
- CNERTA [27]: It adopts a linearised nested BIO labeling strategy to model overlapping entities. Approximately 30% of its entities are nested, posing a more challenging setting for accurate boundary recognition.

Both datasets provide paired speech-text annotations for E2E evaluation. Table 1 summarises the statistics of both datasets.

Table 1. Summary statistics of the Chinese speech named entity recognition datasets.

Dataset	Number of Entity Types	Sentences			Entity Instances		
		Train	Dev	Tese	PER	ORG	LOC
CNERTA	3	34,102	4,440	4,445	8,034	12,047	16,876
AISHELL-NER	3	120,098	14,326	7,176	18,642	25,351	24,611

4.2 Baselines

We compare the proposed model with several state-of-the-art Chinese SNER models as follows:

- LLMs: GLM-4.6[2] and Google AI Studio[3]. For GLM-4.6, which lacks direct speech support, we employ a pipeline using glm-asr for transcription followed by entity extraction.
- Pipeline methods: EA-ASR+BERT-PT [45], which employs a Conformer-based entity-aware ASR refined by a pretrained NER model; and the framework by Zhang et al. [28], which integrates character, glyph, and Pinyin embeddings with Adaptive Homophone Noise (AHN) to improve robustness against transcription errors.
- Semi-E2E and E2E methods: EHA-ASR [46], which predicts entity head tokens to constrain span classification; and fully E2E baselines, including IMAGE [47], a unified framework mapping waveforms directly to labelled text, and the Whisper–ChatGLM3 approach [30].

[2] https://open.bigmodel.cn/
[3] https://aistudio.google.com/

— Encoder-Decoder variants: an E2E model using a Wav2Vec2 encoder and a length adapter (1D-CNN and Transformer), which we test with both T5[4] and Qwen2.5[5] decoders, trained to generate text with explicit entity tags.

4.3 Evaluation Metrics

We assess the proposed model and the baseline models using distinct metrics in the tasks of ASR and NER: ASR Metric: Since Mandarin speech recognition is performed at the character level, we adopt the Character Error Rate (CER) to evaluate ASR performance. CER measures the normalized edit distance between the predicted transcription and the ground truth and is defined as:

$$\mathrm{CER} = \frac{S+D+I}{N} \times 100\%, \tag{29}$$

where S, D, and I denote the number of substitutions, deletions, and insertions required to transform the hypothesis into the reference, respectively, and N represents the total number of characters in the ground-truth transcription. A lower CER indicates better acoustic modeling and transcription accuracy.

NER Metrics: For the core NER task, we adopt the standard F1-score. An entity prediction is regarded as correct only when both its boundary offsets and semantic category exactly match the ground truth (Exact Match). Given the inherent class imbalance in NER datasets—where non-entity tokens dominate and certain entity types are relatively rare—we report the micro-averaged F1-score, which aggregates prediction statistics across all entity categories to provide a holistic evaluation of model performance. The metrics are defined as follows:

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{30}$$

4.4 Experimental Setup

We freeze the convolutional layers of the Wav2Vec2 encoder and fine-tune only its Transformer layers. The decoder consists of six Transformer layers featuring 16 attention heads, a 4096-dimensional SwiGLU feed-forward network, and a dropout rate of 0.1. Parameters are initialised via Xavier uniform initialisation, and beam search is employed during decoding.

The bidirectional multimodal network uses a dual-gate mechanism with GELU activation and a sigmoid function to dynamically weigh information sources. This module maintains the encoder's hidden dimensionality with 8 attention heads, while the NER decoder follows the GlobalPointer configuration [18].

Training employs AdamW with differential learning rates: 5×10^{-5} for the Wav2Vec2 encoder, 8×10^{-4} for the decoder and quantiser, and 2×10^{-4} for task-

[4] https://huggingface.co/Langboat/mengzi-t5-base
[5] https://qwenlm.github.io/blog/qwen2.5/

specific components. To ensure convergence, we use ReduceLROnPlateau and StepLR schedulers to decay the learning rate by a factor of 0.5 upon performance plateaus. Hardware and software details are provided in Table 2.

Table 2. Hardware and software configuration.

Hardware	Software
Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz	Ubuntu 18.04.5 LTS (Bionic Beaver)
32GB RAM	CUDA 10.1
Tesla V100S (32G	PyTorch 1.12.1+cu102
	Python 3.8.20

4.5 Comparison Experimental Results and Analysis

Table 3. Hardware and software configuration.

Paradigm	Method	CNERTA (nested)				AISHELL-NER			
		CER	P	R	F1	CER	P	R	F1
LLM API	GLM-4.6	–	46.25	49.62	47.88	–	58.71	64.82	61.61
	Google AI Studio	–	47.21	19.91	28.01	–	28.73	60.50	38.96
Pipeline	Conf. EA-ASR + BERT-PT [45]	–	–	–	–	4.79	75.54	74.27	74.90
	Conf.-ASR + AHN [28]	–	–	–	–	4.75	75.97	74.64	75.30
	Conf.-ASR + ChineseBERT	–	–	–	–	–	75.54	74.76	74.65
	MMS-ASR[6] + Bert-L	–	–	–	69.76	–	–	–	74.84
Semi-E2E	Semi-E2E [46]	8.30	73.79	50.73	60.12	–	–	–	–
Multimodal	BERT-MuIT-CRF	–	74.95	81.08	77.89	–	75.65	79.95	77.74
	BERT-UMT-CRF	–	75.96	80.18	78.01	–	75.24	80.92	77.98
	BERT-MMI-CRF [27]	–	80.95	74.98	77.85	–	–	–	–
	BiLSTM-GenEn-MNER [48]	–	75.25	66.40	70.55	–	89.61	89.53	**89.57**
	BERT-M3T	–	83.46	75.81	79.45	–	–	–	–
	CDP-MCNER [49]	–	80.20	81.93	**81.05**	–	77.92	80.58	79.23
	MacBERT-M3T [27]	–	83.99	77.46	80.59	–	–	–	–
	MacBERT-CMA-CRF	–	81.20	76.80	78.94	–	–	–	–
	MacBERT-USAF [50]	–	75.59	80.92	–	–	–	–	–
E2E	IMAGE [47]	–	–	–	70.36	–	–	–	75.76
	Whisper+WS [29]	–	–	–	–	**3.67**	–	–	80.30
	Whisper-ChatGLM3 [30]	–	–	–	–	3.88	–	–	80.50
	MMSpeech-ETG	–	–	–	69.82	–	–	–	75.42
	Wav2Vec2+LA+T5	–	–	–	–	–	75.15	75.95	75.55
	Wav2Vec2+LA+Qwen	–	–	–	62.05	–	–	–	77.56
	Ours (Full Model)	**2.52**	84.56	78.55	**81.44**	**2.47**	91.52	90.31	**90.91**

[6] https://modelscope.cn/models/iic/ofa_mmspeech_asr_aishell1_large_zh

As shown in Table 3, the proposed full model achieves the best overall performance on both datasets. On CNERTA, it attains an F1-score of 81.44% with a CER of 2.52%; on AISHELL-NER, it achieves an F1-score of 90.91% with a CER of 2.47%. Compared with the second-best results, the proposed method improves F1 by 0.39 and 1.34 percentage points on CNERTA and AISHELL-NER, respectively, whilst also yielding lower CER.

Compared with large language model APIs such as GLM-4.6 and Google AI Studio, the proposed model achieves substantially higher F1-scores, indicating that task-specific E2E optimisation is more effective for Chinese speech NER. Compared with pipeline approaches (e.g., Conf. EA-ASR + BERT-PT), the proposed framework mitigates error propagation, yielding consistent gains in both CER and NER performance.

The advantage of the proposed method is particularly evident on CNERTA, which contains many nested entities. Our model surpasses the strongest multimodal baseline, CDP-MCNER, by 0.39 percentage points in F1-score. This improvement is attributed to the GlobalPointer decoding head, which models complex span boundaries, and the deep acoustic-semantic fusion mechanism.

Compared with multimodal baselines such as MacBERT-M3T, the gains are mainly reflected in Recall, indicating improved entity coverage in nested scenarios. This likely stems from GlobalPointer's global-span modelling capability and the boundary cues provided by cross-modal interaction.

Across modelling paradigms, pipeline methods maintain stable ASR performance but suffer from limited NER due to error propagation. E2E methods generally outperform pipelines, yet exhibit variability, suggesting that naïve E2E modelling is insufficient for effective cross-modal alignment. Multimodal methods improve performance on CNERTA but yield only limited gains on AISHELL-NER, suggesting task sensitivity. In contrast, the proposed method achieves consistently strong results across both datasets, demonstrating better robustness and generalisation.

4.6 Ablation Studies

To assess the contribution of each major component, we conduct ablation studies on both CNERTA and AISHELL-NER. The results are reported in Table 4.

Table 4. Ablation experimental results.

Method	CNERTA (nested)				AISHELL-NER			
	CER	P	R	F1	CER	P	R	F1
w/o Joint CTC	5.35	80.12	75.44	77.71	5.15	85.42	83.21	84.30
w/o RVQ Bottleneck	4.56	82.31	76.88	79.51	4.32	88.15	87.42	87.78
w/o GPT Fusion	6.65	81.22	75.98	78.52	6.24	87.56	86.11	86.83
w/o Multimodal Fusion	3.11	78.22	74.15	76.12	2.85	82.11	80.45	81.27

Removing the multimodal fusion module results in the largest performance drop, with F1 Scores decreasing by 5.32 and 9.64 percentage points on CNERTA and AISHELL-NER, respectively. This confirms that explicit interaction between acoustic

features and decoder-side semantic representations is crucial for robust entity recognition, particularly for capturing boundary cues in Chinese speech. Removing GPT fusion also causes clear degradation, reflected in both higher CER and lower F1-score. This suggests that external linguistic priors meaningfully improve transcription quality and, consequently, downstream NER performance. Similarly, removing the RVQ bottleneck reduces performance on both datasets, indicating that quantized acoustic representations help filter task-irrelevant variation and regularise the latent space.

Finally, removing the joint CTC objective weakens alignment stability and leads to a marked decline in both ASR and NER performance. Overall, the ablation results show that each component contributes meaningfully, and that the full model benefits from the complementary effects of quantisation, linguistic fusion, multimodal interaction, and joint alignment supervision.

5 Conclusion

This paper proposes an E2E Chinese SNER framework that combines residual vector quantisation, joint CTC-attention decoding, GPT-enhanced semantic modelling, and deep acoustic-semantic fusion to improve robust entity extraction from speech, especially for complex and nested entities. Experiments on AISHELL-NER and CNERTA show strong performance, with F1-scores of 90.91% and 81.44%, respectively, outperforming representative pipeline, multimodal, and E2E baselines, while ablation studies confirm the effectiveness of the key components. Despite these gains, the method still incurs additional computational cost, and its robustness under low-resource conditions remains to be further examined. Future work will therefore focus on improving efficiency, strengthening cross-modal alignment, and extending the framework to multilingual and cross-domain scenarios. to improve generalisation.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (No. 61762016), Guangxi Key Lab of Multi-source Information Mining and Security (24-A-01-01), Ministry-of-Education Key Lab of Education Blockchain and Intelligent Technology (EBME24-05).

References

1. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Diffusionner: Boundary diffusion for named entity recognition. arXiv preprint arXiv:2305.13298 (2023)
2. Yang, B., Luo, X.: Recent progress on named entity recognition based on pretrained language models. In: 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 799–804. IEEE (2023)
3. Zhang, X., Luo, X.: A machine-reading-comprehension method for named entity recognition in legal documents. In: Tanveer, M., Agarwal, S., Ozawa, S., Ekbal, A., Jatowt, A. (eds) Neural Information Processing: ICONIP 2022. Communications in Computer and Information Science, vol 1793, pp. 224–236. Springer (2022)

4. Zhang, X., Luo, X., Wu, J.: A RoBERTa-GlobalPointer-based method for named entity recognition of legal documents. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2023)
5. Porjazoski, D., Leinonen, J., Kurimo, M.: Attention-based end-to-end named entity recognition from speech. In: Text, Speech, and Dialogue: TSD 2021. Lecture Notes in Computer Science, vol. 12848, pp. 469–480. Springer (2021)
6. Deng, Z., Gong, S., Luo, X., Liu, J.: A DeBERTa-GPLinker-based model for relation extraction from medical texts. In: 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 853–860. IEEE (2024)
7. Ding, W., Luo, X.: MedREX: A dual-biaffine attention and feature-enhanced joint model for Chinese medical entity-relation extraction. In: 2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 1448–1454. IEEE (2025)
8. Gong, S., Luo, X.: DGGCCM: A hybrid neural model for legal event detection. Artificial Intelligence and Law, 33.4: 1109-1149 (2025)
9. De, A., Guo, C.: An adaptive vector quantization approach for image segmentation based on SOM network. Neurocomputing 149, 48–58 (2015)
10. Ahmed S., Shumailov, I., Papernot, N., Fawaz, K.: Towards more robust keyword spotting for voice assistants. In: 31st USENIX security symposium (USENIX Security 22). pp. 2655–2672 (2022)
11. Caubrière, A., Rosset, S., Estève, Y., Laurent, A., Morin, E.: Where are we in named entity recognition from speech? In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4514–4520 (2020)
12. Huang, Q., Luo, X.: Automatic Help System with Voice Input for Passengers of Drop Taxi. ICEB 2019 Proceedings (Newcastle Upon Tyne, UK). 67. (2019)
13. Javed, M., Luo, X.: Sos intelligent emergency rescue system: Tap once to trigger voice input. In: Proceedings of the 2020 4th International Conference on Computer Science and Artificial Intelligence. pp. 187–193 (2020)
14. Huang, Q., Luo, X.: NotMeToo: An autonomous agent of preventing sexual harassment and crime. Available at SSRN 3980437 (2021)
15. Yan, M., Luo, X.: BERT-based detection of sexual harassment in dialogues. In: Proceedings of the 2021 5th International Conference on Computer Science and Artificial Intelligence. pp. 359–364 (2021)
16. Ma, R., Peng, M., Zhang, Q., Huang, X.: Simplify the usage of lexicon in Chinese NER. arXiv preprint arXiv:1908.05969 (2019)
17. Wang, S., Tan, H., Li, H.: A review of Chinese named entity recognition. In: Proceedings of the 4th International Conference on Artificial Intelligence and Intelligent Information Processing. pp. 290–300 (2025)
18. Su, J., Murtadha, A., Pan, S., Hou, J., Sun, J., Huang, W., Wen, B., Liu, Y.: GlobalPointer: Novel efficient span-based approach for named entity recognition. arXiv preprint arXiv:2208.03054 (2022)
19. Zhou, G., Su, J.: Named entity recognition using an HMM-based chunk tagger. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 473–480 (2002)
20. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
21. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29(6), 82–97 (2012)

22. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376 (2006)
23. Li, B., Chang, S.y., Sainath, T.N., Pang, R., He, Y., Strohman, T., Wu, Y.: Towards fast and accurate streaming end-to-end ASR. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6069–6073. IEEE (2020)
24. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 4960–4964. IEEE (2016)
25. Watanabe, S., Hori, T., Kim, S., Hershey, J.R., Hayashi, T.: Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE Journal of Selected Topics in Signal Processing 11(8), 1240–1253 (2017)
26. Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., et al.: Scaling speech technology to 1,000+languages. Journal of Machine Learning Research 25(97), 1–52 (2024)
27. Sui, D., Tian, Z., Chen, Y., Liu, K., Zhao, J.: A large-scale Chinese multimodal NER dataset with speech clues. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2807–2818 (2021)
28. Zhang, M., Qiao, X., Zhao, Y., Su, C., Li, Y., Zhu, M., Zhu, J., Li, Y., Zhao, X., Liu, Y., et al.: Incorporating pinyin into pipeline named entity recognition from Chinese speech. In: 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 947–953. IEEE (2023)
29. Yang, H., Zhang, M., Tao, S., Ma, M., Qin, Y.: Chinese ASR and NER improvement based on whisper fine-tuning. In: 2023 25th International Conference on Advanced Communication Technology (ICACT). pp. 213–217. IEEE (2023)
30. Li, Y., Yu, J., Zhang, M., Ren, M., Zhao, Y., Zhao, X., Tao, S., Su, J., Yang, H.: Using large language model for end-to-end Chinese ASR and NER. arXiv preprint arXiv:2401.11382 (2024)
31. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in Neural Information Processing Systems 30 (2017)
32. Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., Tagliasacchi, M.: Soundstream: An end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, 495–507 (2021)
33. Défossez, A., Copet, J., Synnaeve, G., Adi, Y.: High fidelity neural audio compression. arXiv preprint arXiv:2210.13438 (2022)
34. Wang, J., Du, Z., Chen, Q., Chu, Y., Gao, Z., Li, Z., Hu, K., Zhou, X., Xu, J., Ma, Z., et al.: LauraGPT: Listen, attend, understand, and regenerate audio with GPT. arXiv preprint arXiv:2310.04673 (2023)
35. Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al.: Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111 (2023)
36. Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Chang, X., Shi, J., Zhao, S., Bian, J., Wu, X., et al.: UNIAUDIO: An audio foundation model toward universal audio generation. arXiv preprint arXiv:2310.00704 (2023)
37. Tang, B., Zeng, B., Li, M.: LauraTSE: Target speaker extraction using auto-regressive decoder-only language models. arXiv preprint arXiv:2504.07402 (2025)

38. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., Défossez, A.: Simple and controllable music generation. Advances in Neural Information Processing Systems 36, 47704–47720 (2023)
39. Zhao, X., Xiang, H., Ye, S., Li, S., Tian, Z., Chen, G., Ding, K., Wan, G.: LongCat-AudioCodec: An audio tokenizer and detokenizer solution designed for speech large language models. arXiv preprint arXiv:2510.15227 (2025)
40. Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., et al.: Gated attention for large language models: Nonlinearity, sparsity, and attention-sink-free. arXiv preprint arXiv:2505.06708 (2025)
41. Zhang, B., Sennrich, R.: Root mean square layer normalization. Advances in Neural Information Processing Systems 32 (2019)
42. Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020)
43. Ghosh, S., Tyagi, U., Ramaneswaran, S., Srivastava, H., Manocha, D.: MMER: Multimodal multi-task learning for speech emotion recognition. arXiv preprint arXiv:2203.16794 (2022)
44. Bu, H., Du, J., Na, X., Wu, B., Zheng, H.: AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline. In: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). pp. 1–5. IEEE (2017)
45. Chen, B., Xu, G., Wang, X., Xie, P., Zhang, M., Huang, F.: AISHELL-NER: Named entity recognition from Chinese speech. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8352–8356. IEEE (2022)
46. Zhang, M., Qiao, X., Zhao, Y., Su, C., Li, Y., Li, Y., Piao, M., Peng, S., Tao, S., Yang, H.: Semi-end-to-end nested named entity recognition from speech. In: National Conference on Man-Machine Speech Communication. pp. 9–22. Springer (2023)
47. Ning, J., Sun, Y., Xu, B., Yang, Z., Luo, L., Lin, H.: Breaking the boundaries: A unified framework for Chinese named entity recognition across text and speech. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 1250–1260 (2024)
48. Ning, J., Sun, Y., Yang, Z., Wang, Z., Luo, L., Lin, H., Zhang, Y.: GenEn-MNER: Enhancing nested Chinese NER with multimodal fusion and alignment via speech-to-text generation. IEEE Transactions on Audio, Speech and Language Processing (2025)
49. Liu, R., Guo, X., Zhu, H., Wang, L.: A text-speech multimodal Chinese named entity recognition model for crop diseases and pests. Scientific Reports 15(1), 5429 (2025)
50. Liu, Y., Huang, S., Li, R., Yan, N., Du, Z.: USAF: Multimodal Chinese named entity recognition using synthesized acoustic features. Information Processing & Management 60(3), 103290 (2023)