End-to-end Contrastive Language-Speech Pretraining Model For Long-form Spoken Question Answering

Jiliang Hu[2], Zuchao Li[1,*], Baoyuan Qi[4], Liu Guoming[4], Ping Wang[3]
[1]School of Artificial Intelligence, Wuhan University, Wuhan, China,
[2]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China,
[3]School of Information Management, Wuhan University, Wuhan, China,
[4]Xiaomi, Beijing, China.
{jilianghu, zcli-charlie}@whu.edu.cn, {qibaoyuan, liuguoming}@xiaomi.com, wangping@whu.edu.cn

Abstract

Significant progress has been made in spoken question answering (SQA) in recent years. However, many existing methods, including large audio language models, struggle with processing long audio. Follow the success of retrieval augmented generation, a speech-related retriever shows promising in help preprocessing long-form speech. But the performance of existing speech-related retrievers is lacking. To address this challenge, we propose CLSR, an end-to-end contrastive language-speech retriever that efficiently extracts question-relevant segments from long audio recordings for downstream SQA task. Unlike conventional speech-text contrastive models, CLSR incorporates an intermediate step that converts acoustic features into text-like representations prior to alignment, thereby more effectively bridging the gap between modalities. Experimental results across four cross-modal retrieval datasets demonstrate that CLSR surpasses both end-to-end speech related retrievers and pipeline approaches combining speech recognition with text retrieval, providing a robust foundation for advancing practical long-form SQA applications.

Code — https://github.com/193746/CLSR

Introduction

The question answering (QA) task requires the model to find the answer to a question from a given context. When the answer is a specific segment within the context, the task is classified as extractive QA; conversely, if the answer cannot be directly derived from the context and necessitates additional reasoning by the model, it is termed abstractive QA (Shih et al. 2023a). In the realm of spoken question answering, the context is presented in audio format (Li et al. 2018), and certain complex SQA tasks also require the questions to be delivered in audio format (Shon et al. 2022). Despite advancements in SQA methodologies (Lee, Chen, and Lee 2019; You et al. 2022), the majority of existing SQA models are limited to processing short audio segments (under one minute). However, many real-world dialogue scenarios, such as meetings, lectures, and online discussions, often involve voice recordings that exceed ten minutes, posing challenges for current SQA techniques.

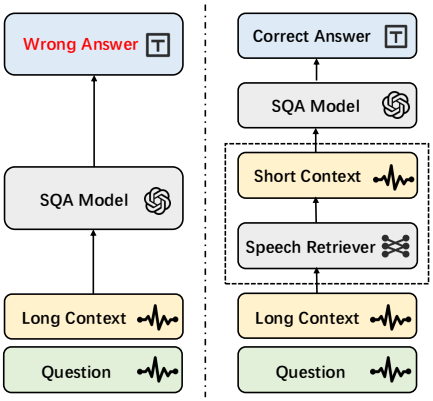

Figure 1: Using a speech retrieval model to simplify long audio context into several key audio segments can help improve the quality of subsequent LALM response.

The development of large language models (LLMs) is advancing rapidly. Notable examples, such as GPT (Brown 2020) and LLaMA (Touvron et al. 2023), among others (Li et al. 2023; Yao, Li, and Zhao 2024), have demonstrated significant success across various traditional natural language processing (NLP) tasks, including question answering. In the speech domain, numerous large audio language models (LALMs) have emerged, showcasing remarkable capabilities in speech comprehension (Chu et al. 2023; Radford et al. 2023). The retrieval-augmented generation (RAG) paradigm (Lewis et al. 2020) enhances LLMs' natural language understanding by integrating external knowledge (Gupta, Ranjan, and Singh 2024). Specifically, RAG employs a retriever to assess the similarity between user queries and segments within a knowledge database, selecting the top-k most relevant segments as supplementary context for the LLM. This approach enables the LLM to better comprehend queries and generate more accurate responses. Currently, RAG is commonly used for long-context reasoning tasks (Li et al. 2025; Guo et al. 2025). For example, in question-answering tasks that involve lengthy input contexts like full articles, RAG works by identifying and retrieving the most pertinent con-

*Corresponding author.

text segments. This process minimizes the inclusion of irrelevant information, which can otherwise introduce errors into the answer and reduce inference speed. Given the effectiveness of RAG in text-based long-context QA, a pertinent question arises for long-form SQA: Can RAG be similarly employed to extract problem-relevant segments from audio inputs to serve as context for subsequent LALM processing?

In this paper, we propose CLSR, an end-to-end (E2E) contrastive language-speech retriever designed to distill lengthy speech recordings into a selection of audio clips that are most pertinent to a given query. These audio clips are used for subsequent LALM inference. Unlike conventional E2E speech-text contrastive learning models, CLSR does not endeavor to align acoustic representations and text representations directly. Instead, it first converts acoustic representations into text-like representations, which are then aligned with actual text representations. The extraction of text-like representations primarily employs continuous integrate-and-fire (CIF) to map acoustic representations from temporal steps to token numbers, followed by the application of a vector quantizer (VQ) based adaptor to refine these acoustic representations into text-like forms. We conduct a comparative analysis of CLSR against standard E2E speech-text retrievers and pipeline retrievers, which integrate speech-to-text models with text contrastive learning models, across four datasets: Spoken-SQuAD, LibriSQA, SLUE-SQA-5, and DRCD. The experimental findings demonstrate that CLSR exhibits superior retrieval performance, suggesting that the use of text-like representations as an intermediary between acoustic and text representations enables CLSR to more effectively discern the similarities and distinctions between these two modalities, thereby enhancing the accuracy of pairing speech with text or speech with speech. The contributions of this paper are as follows:

(1) To our knowledge, this study represents the inaugural introduction of the RAG concept within the domain of SQA and its application to address challenges associated with lengthy speech inputs.

(2) The proposed model initially transforms acoustic representations into text-like representations, subsequently aligning these text-like representations with text representations, which effectively mitigates modal discrepancies and facilitates cross-modal alignment.

(3) The CLSR demonstrates superior performance compared to both E2E and cascade speech retrieval systems on four datasets: Spoken-SQuAD, LibriSQA, SLUE-SQA-5, and DRCD.

Related Work

Currently, there are many works related to SQA. Chuang et al. (2019) propose a pre-trained model called SpeechBERT for the E2E SQA task. Through the training stage called initial phonetic spatial joint embedding for audio words, it aligns the generated audio embeddings with the text embeddings produced by BERT (Devlin et al. 2019) within the same hidden space. Shih et al. (2023a) introduce GSQA, which enables the SQA system to perform abstractive reasoning. They first utilize HuBERT (Hsu et al. 2021) to convert the input speech into discrete units, and then employ a sequence-to-sequence SQA model finetuned from the text QA model LongT5 (Guo et al. 2022) to generate answers in the form of discrete units. Lin et al. (2024) focus on open-domain SQA in scenarios whose paired speech-text data is unavailable. They propose SpeechDPR, which utilizes a bi-encoder retriever framework and learns a sentence-level semantic representation space by extracting knowledge from a combined model of automatic speech recognition (ASR) and text retrieval. Johnson et al. (2024) introduce a retriever that employs deep Q-learning to bypass irrelevant audio segments in longer audio files, thereby enhancing the efficiency of SQA. The latter two articles are related to retriever, which is similar to our paper; however, they have limitations: the performance of the former is inferior to that of the pipeline model, and the latter poorly adapts to the high-dimensional, complex feature space of audio, easily falls into local optima during training due to unbalanced exploration and exploitation.

Since the inception of GPT, RAG has advanced rapidly (Zhao et al. 2025), while research on speech RAG has been comparatively limited. Yang et al. (2024) utilize RAG for spoken language understanding (SLU). They first employ a pre-trained ASR encoder to extract acoustic features. Next, they perform a similarity calculation to identify audio-text label pairs in the training set that are similar, subsequently incorporating this label information into the SLU decoder through a cross-attention mechanism. Wang et al. (2024) propose a joint speech and language model based on RAG, which enhances performance on the named entity recognition task. They compute the similarity between the input speech query embeddings and the entity embeddings in the database to extract the K entities most relevant to the query, using these entities as additional inputs to the model. Currently, there is no SRAG model designed for long-form SQA task.

Method

Preliminary

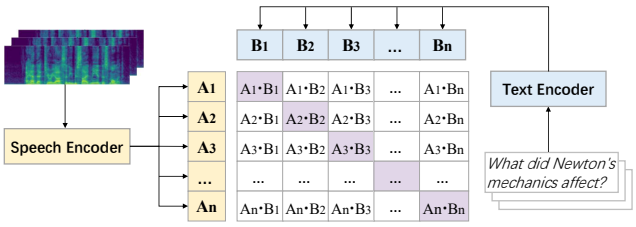

Figure 2: The architecture of typical E2E speech-text contrastive model.

Consider the SQA task, where questions are presented in text format and contexts are delivered in speech format. Let $X = \{x_1, x_2, \ldots, x_t\}$ denote the speech context, represented as a sequence of t frames. Let $Z = \{z_1, z_2, \ldots, z_m\}$ denote the question text, represented as a sequence of m tokens, where each token z_i belongs to a vocabulary V (i.e., $z_i \in V$). Figure 2 illustrates the architecture of a typical

E2E audio-text contrastive model, such as CLAP (Wu et al. 2023). This model employs a speech encoder $A(\cdot)$ and a text encoder $B(\cdot)$ to extract acoustic features and textual features, respectively. The similarity S between the two feature representations is then quantified using cosine similarity. The formula is as follows, where $||.||$ denotes the L2 norm.

$$S_{X,Z} = ||A(X)|| \cdot ||B(Z)||$$

Contrastive learning models of this type focus on deriving sentence-level features. Two common methods exist for this purpose. The first introduces a trainable `CLS` token during encoding; the representation corresponding to this token is then used as the global sentence feature. The second method computes the average of all token-level features (of sequence length n) to produce a single feature vector. Both methods are equally applicable to extracting features from entire audio segments.

During training, the model learns to minimize the negative log-likelihood (NLL) loss between the representations of paired questions and contexts. This NLL loss comprises two symmetric components: one for retrieving the context given the question, and the other for retrieving the question given the context. The specific formulation is as follows:

$$NLL_{A,B} = -\frac{1}{2}(\sum_{i=0}^{m} log\frac{e^{S_{X,z_i}}}{e^{S_{X,Z}}} + \sum_{i=0}^{t} log\frac{e^{S_{x_i,Z}}}{e^{S_{X,Z}}})$$

Overview

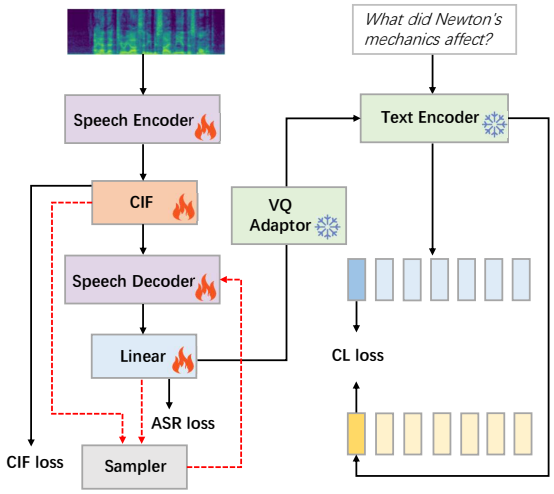

Figure 3: The architecture of proposed model, CLSR. The red line is only used during training.

Figure 3 shows the specific architecture of CLSR. The left half features a non-autoregressive attention encoder-decoder (AED) framework based on CIF (Dong and Xu 2020). It takes the speech context X as input and produces the corresponding token probability distribution D, $D = \{d_1, d_2, d_3, \ldots, d_n\}$. Both the speech encoder and decoder adopt the SAN-M (Gao et al. 2020) structure, which is a specialized Transformer (Vaswani et al. 2017) layer that integrates a self-attention mechanism with deep feed-forward sequential memory networks (DFSMN). Initially, the framework uses the speech encoder to extract acoustic features H^s.

$$H^s = SEncoder(X)$$

Next, it maps H^s from the time step to the number of tokens through the soft and monotonic alignment mechanism, CIF, obtaining an acoustic representation E^a, which is aligned with the token probability distribution.

$$E^a = CIF(H^s)$$

Then, as shown in the following formula, it predicts the corresponding token distribution through the speech decoder and a fully connected layer. In addition, following Gao et al. (2022), we use a sampler to optimize the training process of this framework. The sampler does not contain any learnable parameters and is designed to enhance the context modeling capability of the decoder by sampling text features into E^a. It will be elaborated on in subsequent chapter.

$$D = W \cdot Decoder(H^s, E^a) + b$$

Next, we utilize the VQ adaptor to map the token distribution to text-like embeddings.

$$E^{Y'} = VQAdaptor(D)$$

The right half of CLSR is a Transformer-based text encoder that receives either text embeddings or text-like embeddings as input and output corresponding text representations. We obtain the sentence-level representation by inserting the `CLS` token. When aligning the text question with the speech context, we input the text-like embeddings $E^{Y'}$ of the context and the text embeddings E^Z of the question into the text encoder to obtain their respective sentence-level representations. We then use cosine similarity to evaluate the similarity between them.

$$S_{Y',Z} = ||TEncoder(E^{Y'})|| \cdot ||TEncoder(E^Z)||$$

Continuous Integrate-and-Fire

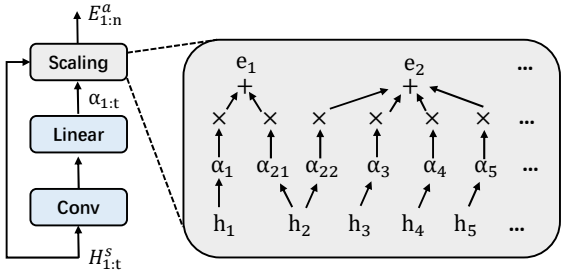

Figure 4: The explanation of CIF workflow. The gray box on the right shows an example of CIF, where $\alpha = \{0.8, 0.3, 0.4, 0.4, 0.1\}$ and the threshold β=1.

Figure 4 explains the workflow of the CIF. Through convolution operations and linear mapping, it calculates the

weight distribution α, $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_t \mid \alpha_i \in [0, 1]\}$. Each α_i shows the valid information contained in relevant h_i of the acoustic feature $H^s_{1:t}$.

$$\alpha_{1:t} = W \cdot conv(H^s_{1:t}) + b$$

Then, it gathers the weights and combines $H^s_{1:t}$ until the total weight hits a specified threshold β, signaling that an acoustic boundary has been attained. When reaching the threshold, if the current state of α overflows, it will be used for the next round of weight accumulation.

Sampler

To enhance the capacity of the selected non autoregressive AED framework to model token probability distributions, we introduce a training optimization module called the sampler. When the sampler is enabled, the training process of the framework consists of two rounds. In the first round, we do not utilize the sampler; instead, we directly employ the acoustic features E^a obtained from the CIF module to predict the probability distribution of tokens. By applying $argmax$, we can derive the transcription result Y^{asr}.

$$Y^{asr} = \arg\max_{y_i \in V}(W \cdot Decoder(H^s, E^a) + b)$$

Next, we proceed to the second round of training and initiate sampling. We first compare the ASR output Y^{asr} with the ground-truth context Y^{con} to identify tokens containing transcription errors and their respective positions. Then merge the correct embeddings of erroneous tokens from E^c (the embeddings of Y^{con}) into the acoustic features E^a through selective replacement, generating semantic features E^s. This process is formalized as:

$$E^s[:, i, :] = \begin{cases} E^c[:, i, :] & \text{if token } i \text{ is erroneous} \\ E^a[:, i, :] & \text{otherwise} \end{cases}$$

After generating E^s, we complete the sampling step. It is important to note that not every correct embedding from an erroneous token will be incorporated into E^a; this decision is governed by the mixing ratio λ ($\lambda \in (0, 1)$).

$$E^s = Sampler(E^a, E^c, \lceil \lambda \sum_{i=1}^{N} (y_i^{asr} \neq y_i^{con}) \rceil)$$

Afterwards, use E^s instead of E^a to calculate the probability distribution of the tokens.

$$D^{'} = W \cdot Decoder(H^s, E^s) + b$$

It is important to note that during the initial training phase, no gradient backpropagation is performed, and Y^{asr} is solely utilized to determine the sampling number for the sampler. $D^{'}$ obtained in the second phase is then used to calculate the ASR loss.

Regarding the real text embeddings E^c, Gao et al. (2022) uses the embedding layer of the speech decoder to derive them. However, in our proposed model, this layer is not trained, and its weights may not effectively represent the text embedding space. Consequently, we use the weights of the linear layer, which is used to generate the probability distribution of the tokens, to calculate E^c.

Adaptor

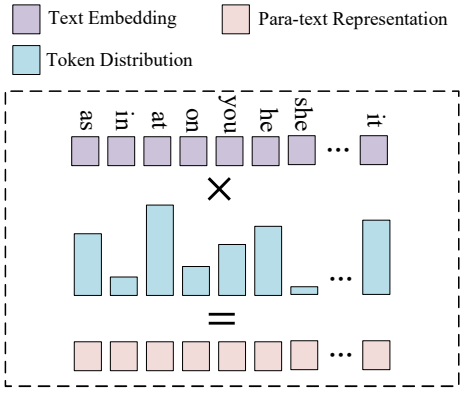

Figure 5: The mapping process of the adaptor.

After obtaining the probability distribution D of the tokens, we use an adaptor to map it to the text-like embedding $E^{Y'}$. The adaptation involves two steps: quantization and mapping. The quantization converts each token's probability distribution into the index of the highest-probability token in vocabulary V. Following Shih et al. (2023b), we first select the token index q_i with the highest probability from each distribution d_i:

$$q_i = \arg\max_{j \in [1, |V|]} d_{i,j}$$

Since q_i is non-differentiable, directly using it would break the computational graph. For gradient propagation, we use the temperature-scaled softmax distribution $\tilde{p}_i$, where $\gamma = 0.1$ is a hyper-parameter:

$$\tilde{p}_i = softmax(d_i / \gamma)$$

Through the straight-through gradient estimator (Bengio, Léonard, and Courville 2013), we combine the discrete token q_i (represented as a one-hot vector) with the continuous approximation $\tilde{p}_i$ while maintaining gradient flow. The formula is as follows, where $sg(x) = x$ and $\frac{d}{dx}\text{sg}(x) = 0$ is the stop-gradient operator:

$$q_i^{\text{st}} = q_i + \tilde{p}_i - \text{sg}(\tilde{p}_i)$$

The quantized token representation q_i^{st} are collected into a matrix Q^{st}. Next, we map Q^{st} to the embedding layer of the text encoder. The specific operation, illustrated in Figure 5, involves matrix multiplication with the embedding layer weights W^{te}:

$$E^{Y'} = Matmul(Q^{st}, W^{te})$$

Loss Function

The adopted framework calculates three loss functions during training: cross-entropy (CE), mean absolute error (MAE), and minimum word error rate (MWER) loss. CE and MWER are used to optimize the model's transcription

ability, while MAE facilitates convergence of the CIF. According to Gao et al. (2022), the loss function for the ASR part is:

$$\mathcal{L}_{ASR} = \gamma \mathcal{L}_{CE} + \mathcal{L}^{N}_{werr}(x, y^*)$$

$$\mathcal{L}^{N}_{werr}(x, y^*) = \sum_{y_i \in sample} p(y_i \mid x)[\mathcal{W}(y_i, y^*) - W]$$

We also use NLL loss to optimize the model's ability in aligning the question representation with the context representation. The overall loss function can be expressed as follows, where α and β are parameters that control the proportions of the CIF loss and the contrastive loss, with $\alpha \in (0, 1)$ and $\beta \in (0, 1)$.

$$\mathcal{L}_{total} = (1 - \alpha - \beta)\mathcal{L}_{ASR} + \alpha \mathcal{L}_{MAE} + \beta \mathcal{L}_{NLL}$$

Experiment

Configuration

Dataset	Language	Type		Size		
		Question	Context	Train	Val	Test
Spoken-SQuAD	English	Text	Speech	37,107	5,351	-
Spoken-SQuAD*	English	Text	Speech	29,227	3,884	-
LibriSQA	English	Text	Speech	104,014	2620	-
SLUE-SQA-5	English	Speech	Speech	46,186	1,939	2,382
DRCD*	Chinese	Speech	Speech	25,321	1,425	-

Table 1: Experimental datasets. Datasets marked with asterisks are filtered to ensure one-to-one correspondence between problems and contexts.

We conduct experiments on four datasets: Spoken-SQuAD (Li et al. 2018), LibriSQA (Zhao et al. 2024), SLUE-SQA-5 (Shon et al. 2022), and DRCD. Table 1 displays detailed information about these datasets. Li et al. (2018) use Google text-to-speech (TTS) system to generate the spoken versions of the articles in SQuAD (Rajpurkar 2016). Given that SQuAD is a many-to-one dataset, where multiple questions correspond to the same context, it is unsuitable for training text-speech retrievers. So, we filter the original Spoken-SQuAD dataset to ensure that each question corresponds one-to-one with its context; the filtered dataset is referred to as Spoken SQuAD*. LibriSQA is adapted from the ASR dataset LibriSpeech (Panayotov et al. 2015). The authors input the textual document of each speech segment from LibriSpeech into ChatGPT and request the generation of corresponding text question-answer pairs. We use the first part of LibriSQA, which presents questions without options, and the answers are complete sentences. SLUE-SQA-5 is adapted from five text QA datasets, and both the questions and contexts consist of authentic audio recordings. DRCD (Shao et al. 2018) is originally a Chinese QA dataset. Similar to SQuAD, it is also a many-to-one dataset. We first filter it into a one-to-one dataset and then use the TTS model (Li et al. 2020) to synthesize the speech versions of each question-context pair for its training set. Lee et al. (2018) offer spoken version of DRCD's development set, which we use for testing.

CLSR is built with Paraformer (Gao et al. 2022) (220M) and frozen BGE-base (Chen et al. 2024) (109M). Two kinds of baselines are compared: an E2E text-speech contrastive model like Figure 2 and a cascaded model that first transcribes speech with an ASR module followed by a text QA module. For the former, CLAP and SpeechDPR are selected; for the latter, we employ Whisper (Radford et al. 2023) (244M) for ASR and BGE-base for text QA. Evaluation uses word error rate (WER) for ASR performance and top-k retrieval recall for retrieval ability. Experiments are conducted on FunASR (Gao et al. 2023) and ModelScope. The loss weights α and β are set to $\frac{1}{3}$. Models are trained to convergence with the Adam optimizer at a learning rate of 5e-5.

Main Result

Table 2 shows the comparison results of CLSR and other models across four datasets. We additionally provide the results of using BGE for clean text question-context retrieval. In terms of E2E text-to-speech contrastive models, the results of CLSR are significantly better than those of CLAP and SpeechDPR. We found that CLAP cannot learn the relevance between text questions and speech contexts effectively on four datasets, suggesting that CLAP is not well-suited for text-to-speech content alignment. In fact, CLAP is more appropriate for sound and text alignment.

SpeechDPR is committed to using text-less data for training. Although they use ASR models and text QA models for knowledge distillation, the scarcity of data hampers its ability to achieve optimal performance. It is worth noting that we do not conduct large-scale pre-training prior to training CLSR. All leading contrastive learning models, such as BGE, have undergone extensive pre-training, which enhances their retrieval capabilities. Nevertheless, CLSR still achieves results that are second only to BGE in clean text retrieval and even surpasses BGE's performance on Spoken-SQuAD*, highlighting the advantages of CLSR's architecture.

Compared to conventional E2E contrastive models that directly perform text-to-speech or speech-to-speech alignment, CLSR utilizes text-like representations to alleviate the differences between speech and text modalities. It first maps speech representations into text-like representations and then aligns these text-like representations with actual text representations (or aligns text-like representations with other text-like representations) within the text modality. Leveraging the robust performance of text contrastive models, this approach enhances the alignment between speech and text (or between speech and speech), thereby facilitating more accurate pairing with the context most relevant to the question.

When conducting a comparative analysis of CLSR and Whisper+BGE, we find that their retrieval performances on three English datasets are quite similar; however, CLSR demonstrates certain advantages. In terms of transcription ability, CLSR significantly outperforms Whisper+BGE. This indicates that the joint training of CLSR effectively optimizes both the ASR module and the contrastive learning module. Given that Whisper's performance in Chinese speech recognition is not exceptional, we have opted not to

Dataset	Model	Paradigm	Type		ASR	Q-C Retrieval (↑)			C-Q Retrieval (↑)		
			Question	Context	WER (↓)	R@1	R@5	R@10	R@1	R@5	R@10
Spoken-SQuAD*	BGE	E2E	Text	Text	0	67.12	85.20	89.44	65.63	84.14	89.06
	CLAP	E2E	Text	Speech	-	2.93	9.92	14.84	3.20	10.15	15.23
	Whisper+BGE	Pipeline	Text	Transcript	19.39	69.93	86.61	90.53	**67.97**	**85.76**	89.65
	CLSR	E2E	Text	Speech	**15.14**	**70.03**	**86.90**	**90.68**	67.84	85.69	**90.17**
LibriSQA	BGE	E2E	Text	Text	0	86.91	94.31	95.92	86.87	94.73	96.60
	CLAP	E2E	Text	Speech	-	0.04	0.19	0.38	0.08	0.19	0.50
	Whisper+BGE	Pipeline	Text	Transcript	4.32	83.70	93.28	94.92	85.15	93.40	95.27
	CLSR	E2E	Text	Speech	**4.09**	**85.04**	**93.36**	**95.04**	**85.53**	**94.01**	**95.57**
SLUE-SQA-5	BGE	E2E	Text	Text	0	38.71	72.26	84.34	35.68	70.11	82.28
	CLAP	E2E	Text	Speech	-	11.17	28.67	38.16	11.21	28.59	38.12
	SpeechDPR	E2E	Speech	Speech	-	-	-	19.94*	-	-	-
	Whisper+BGE	Pipeline	Transcript	Transcript	36.41	29.98	60.41	72.71	29.85	60.75	**73.47**
	CLSR	E2E	Speech	Speech	**16.69**	**30.65**	**62.19**	**74.43**	**29.89**	**62.18**	73.05
DRCD*	BGE	E2E	Text	Text	0	90.67	97.12	98.74	89.26	97.75	98.39
	CLAP	E2E	Text	Speech	-	0.35	1.33	2.95	0.35	1.12	1.61
	CLSR	E2E	Speech	Speech	**5.56**	76.21	87.79	90.03	75.23	88.21	91.51

Table 2: Main results of the proposed model across four datasets. Results for BGE are included as a reference benchmark, showing theoretical limits under optimal ASR conditions (100% accuracy). The SpeechDPR's paper only provides the result of R@20. CLAP is composed of HTSAT (Chen et al. 2022) and RoBERTa (Liu 2019).

train Whisper on DRCD*.

Ablation Result

To demonstrate the effectiveness of the quantizer and sampler in CLSR, as well as the potential for multi-stage training to improve model performance. We conduct a series of ablation experiments on Spoken-SQuAD. The results are shown in Table 3. The first two rows of the results show the value of the quantizer. When the quantizer is not utilized, the model may achieve a lower WER; however, its comparative learning ability significantly diminishes. The top-10 retrieval recall rate of "CLSR w/o VQ" is only comparable to the top-1 retrieval recall rate of "CLSR w/ VQ". The results in the sixth and seventh rows show the effectiveness of the sampler. After introducing the sampler, CLSR not only improves retrieval ability, but also improves ASR performance.

Before joint training, we can pre-train the ASR module and the BGE module of CLSR separately. In the experiment, we use 460 hours of clean LibriSpeech data to pre-train Paraformer, and use Spoken-SQuAD's clean text question-context pairs to train BGE. Comparing the second and fourth rows of the experimental results, it is not difficult to find that pre-training the BGE is beneficial; incorporating the pre-trained BGE during joint training enhances various retrieval metrics of the CLSR. In addition, through the comparison between the fourth and sixth rows, it can be found that pre-training Paraformer can improve the model's transcription performance while also slightly improving its retrieval ability. It should be noted that in order to improve the training speed of the model, we froze BGE, which has strong retrieval performance, during joint training. Therefore, we can freeze the ASR module after joint training and train BGE for a few epochs separately, which is called post-train in the table. It is hoped that this approach can make BGE better adapt to the text-like representation provided by the ASR module. Unfortunately, post-train can only slightly improve the performance of the model, as evidenced by rows 2 and 3, 4 and 5, 7 and 8 in the table. In short, through ablation experiments, we have shown that both quantizers and samplers are inseparable for CLSR, and that pre-training the ASR module and BGE module of CLSR is of significant importance.

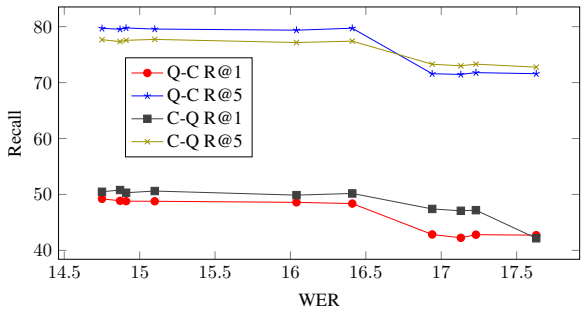

Figure 6: The correlation between the retrieval ability and speech recognition ability of CLSR.

To assess transcription error impact on CLSR's retrieval ability, we evaluate it on Spoken-SQuAD (results in Figure 6). Lower WER correlates with higher recall rates. Crucially, a WER of 16.75% marks a threshold: recall drops significantly above this value.

In order to further demonstrate the superiority of the proposed model over the traditional E2E speech-related contrastive model, which consists of two encoders, we construct a new baseline: ParaBGE, to compare the retrieval capability with CLSR. ParaBGE is composed of the speech encoder from Paraformer and the text encoder from BGE. The sizes of each module in both models are identical to those in CLSR. The experimental results are shown in Table 4. All retrieval metrics of CLSR far exceed ParaBGE, indi-

Pre-train		Joint-train		Post-train	ASR	Q-C Retrieval (↑)			C-Q Retrieval (↑)		
ASR	BGE	VQ	Sampler	BGE	WER (↓)	R@1	R@5	R@10	R@1	R@5	R@10
×	×	×	×	×	16.13	15.29	34.14	44.18	15.75	36.11	46.16
×	×	✓	×	×	17.00	42.52	71.46	78.36	46.86	72.66	79.95
×	×	✓	×	✓	17.00	45.11	75.31	82.90	48.05	75.82	83.18
×	✓	✓	×	×	17.00	48.10	78.28	84.98	49.45	76.79	83.42
×	✓	✓	×	✓	17.00	48.31	78.55	84.73	50.08	77.16	83.68
✓	✓	✓	×	×	16.18	49.00	79.20	85.69	50.31	77.48	84.21
✓	✓	✓	✓	×	15.01	49.65	79.61	85.91	50.59	77.71	84.38
✓	✓	✓	✓	✓	**15.01**	**49.82**	**79.63**	**85.83**	**50.63**	**77.69**	**84.56**

Table 3: Ablation results in Spoken-SQuAD.

Dataset	Model	Paradigm	ASR	Q-C Retrieval (↑)			C-Q Retrieval (↑)		
			WER (↓)	R@1	R@5	R@10	R@1	R@5	R@10
Spoken-SQuAD	ParaBGE	E2E	-	17.79	38.68	48.35	17.03	38.31	48.91
	CLSR	E2E	15.01	**49.82**	**79.63**	**85.83**	**50.63**	**77.69**	**84.56**
LibriSQA	ParaBGE	E2E	-	29.31	50.27	59.70	20.57	39.28	49.28
	CLSR	E2E	4.09	**85.04**	**93.36**	**95.04**	**85.53**	**94.01**	**95.57**
SLUE-SQA-5	ParaBGE	E2E	-	7.31	21.83	32.75	7.52	21.96	33.12
	CLSR	E2E	16.69	**30.65**	**62.19**	**74.43**	**29.89**	**62.18**	**73.05**

Table 4: Comparison results between traditional E2E contrastive model and CLSR.

cating that CLSR has a stronger question-context alignment ability. Although ParaBGE can optimize parameters towards the direction of aligning question and context representation during training, its performance is not ideal. As we mentioned earlier, such model heavily rely on pre-training with large-scale corpora. However, high-quality speech-text pairs are already very scarce, so for E2E speech related retrieval models, it is difficult to achieve excellent results. However, CLSR alleviates the modal differences between speech and text by using text-like representation as a bridge, shifting the alignment of speech to text alignment. With the powerful generalization ability of text contrastive learning models, it can achieve excellent retrieval capabilities comparable to cascade models and text contrastive models without the need for long-term, large-scale pre-training.

Long-form SQA Evaluation

w/ CLSR	EM (↑)	F1 (↑)	Cost time (s)	SpeedUP
×	18.00	23.55	7935.00	1.00X
✓	**27.60**	**35.10**	**783.00**	**10.13X**

Table 5: The effectiveness of CLSR when applied to long audio SQA.

To validate our model for real-world long audio SQA—enabling downstream LALMs to simplify inputs and enhance inference speed and accuracy—we conduct SQA tasks on a modified SLUE-SQA-5 dataset using CLSR and Qwen-Audio. To simulate long-context inference, we replace the contextual audio in 500 randomly selected test instances with full documents from the Spoken Wikipedia corpus (Köhn, Stegen, and Baumann 2016), each averaging 30 minutes in length.

Without CLSR, Qwen-Audio directly generates answers from the input speech document and text question. With CLSR, we first segment the speech document into 40-second intervals, assess each segment's similarity to the text question, and select the most relevant one as Qwen-Audio's contextual input. Prior to testing, both Qwen-Audio and CLSR were trained using the original training subset of SLUE-SQA-5. The results of the testing are presented in Table 5. Following the application of CLSR for long audio reduction, there is a notable enhancement in Qwen-Audio's exact match (EM) and macro-F1 scores for SQA, alongside a tenfold reduction in inference time. These findings underscore the significance of CLSR in the preprocessing of long audio, demonstrating its capacity to not only enhance the inference accuracy of downstream LALM but also to substantially decrease inference time.

Conclusion

In this paper, we introduce CLSR, an E2E contrastive language-speech retriever designed to distill lengthy speech recordings into a limited number of clips that are most pertinent to a given query. By employing a text-like representation as an intermediary state, CLSR exhibits strong capability of cross-modal question-context alignment. Experimental findings demonstrate that CLSR's retrieval performance significantly outstrips that of existing E2E speech-related retrievers and is competitive with both cascaded models and text-based retrievers.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62306216), the Fundamental Research Funds for the Central Universities (No.2042025kf0026), the Technology Innovation Program of Hubei Province (Grant No. 2024BAB043).

References

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Chen, K.; Du, X.; Zhu, B.; Ma, Z.; Berg-Kirkpatrick, T.; and Dubnov, S. 2022. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 646–650. IEEE.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Chuang, Y.-S.; Liu, C.-L.; Lee, H.-Y.; and Lee, L.-s. 2019. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. *arXiv preprint arXiv:1910.11559*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dong, L.; and Xu, B. 2020. Cif: Continuous integrate-and-fire for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6079–6083. IEEE.

Gao, Z.; Li, Z.; Wang, J.; Luo, H.; Shi, X.; Chen, M.; Li, Y.; Zuo, L.; Du, Z.; Xiao, Z.; et al. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*.

Gao, Z.; Zhang, S.; Lei, M.; and McLoughlin, I. 2020. San-m: Memory equipped self-attention for end-to-end speech recognition. *arXiv preprint arXiv:2006.01713*.

Gao, Z.; Zhang, S.; McLoughlin, I.; and Yan, Z. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*.

Guo, J.; Li, Z.; Wu, J.; Wang, Q.; Li, Y.; Zhang, L.; Zhao, H.; and Yang, Y. 2025. ToM: Leveraging Tree-oriented MapReduce for Long-Context Reasoning in Large Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 17804–17823.

Guo, M.; Ainslie, J.; Uthus, D. C.; Ontanon, S.; Ni, J.; Sung, Y.-H.; and Yang, Y. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 724–736.

Gupta, S.; Ranjan, R.; and Singh, S. N. 2024. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. *arXiv preprint arXiv:2410.12837*.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460.

Johnson, A.; Plantinga, P.; Sun, P.; Gadiyaram, S.; Girma, A.; and Emami, A. 2024. Efficient SQA from Long Audio Contexts: A Policy-driven Approach. In *Proc. Interspeech 2024*, 1350–1354.

Köhn, A.; Stegen, F.; and Baumann, T. 2016. Mining the spoken wikipedia for speech data and beyond. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4644–4647.

Lee, C.-H.; Chen, Y.-N.; and Lee, H.-Y. 2019. Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7300–7304. IEEE.

Lee, C.-H.; Wang, S.-M.; Chang, H.-C.; and Lee, H.-Y. 2018. ODSQA: Open-domain spoken question answering dataset. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 949–956. IEEE.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, C.-H.; Wu, S.-L.; Liu, C.-L.; and Lee, H.-y. 2018. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*.

Li, N.; Liu, Y.; Wu, Y.; Liu, S.; Zhao, S.; and Liu, M. 2020. Robutrans: A robust transformer-based text-to-speech model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8228–8235.

Li, Q.; Xiao, T.; Li, Z.; Wang, P.; Shen, M.; and Zhao, H. 2025. Dialogue-rag: Enhancing retrieval for llms via node-linking utterance rewriting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 24423–24438.

Li, Z.; Zhang, S.; Zhao, H.; Yang, Y.; and Yang, D. 2023. Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer. *arXiv preprint arXiv:2307.00360*.

Lin, C.-J.; Lin, G.-T.; Chuang, Y.-S.; Wu, W.-L.; Li, S.-W.; Mohamed, A.; Lee, H.-y.; and Lee, L.-S. 2024. SpeechDPR: End-To-End Spoken Passage Retrieval For Open-Domain

Spoken Question Answering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12476–12480. IEEE.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518. PMLR.

Rajpurkar, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Shao, C. C.; Liu, T.; Lai, Y.; Tseng, Y.; and Tsai, S. 2018. DRCD: A Chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.

Shih, M.-H.; Chung, H.-L.; Pai, Y.-C.; Hsu, M.-H.; Lin, G.-T.; Li, S.-W.; and Lee, H.-y. 2023a. GSQA: An End-to-End Model for Generative Spoken Question Answering. *arXiv preprint arXiv:2312.09781*.

Shih, Y.-J.; Wang, H.-F.; Chang, H.-J.; Berry, L.; Lee, H.-y.; and Harwath, D. 2023b. Speechclip: Integrating speech with pre-trained vision and language model. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, 715–722. IEEE.

Shon, S.; Arora, S.; Lin, C.-J.; Pasad, A.; Wu, F.; Sharma, R.; Wu, W.-L.; Lee, H.-Y.; Livescu, K.; and Watanabe, S. 2022. SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks. *arXiv preprint arXiv:2212.10525*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, M.; Shafran, I.; Soltau, H.; Han, W.; Cao, Y.; Yu, D.; and El Shafey, L. 2024. Retrieval Augmented End-to-End Spoken Dialog Models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12056–12060. IEEE.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Yang, H.; Zhang, M.; Wei, D.; and Guo, J. 2024. Srag: speech retrieval augmented generation for spoken language understanding. In *2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT)*, 370–374. IEEE.

Yao, Y.; Li, Z.; and Zhao, H. 2024. SirLLM: Streaming Infinite Retentive LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2611–2624.

You, C.; Chen, N.; Liu, F.; Ge, S.; Wu, X.; and Zou, Y. 2022. End-to-end spoken conversational question answering: Task, dataset and model. *arXiv preprint arXiv:2204.14272*.

Zhao, Y.; Li, Z.; Zhao, H.; Qi, B.; and Guoming, L. 2025. DAC: A Dynamic Attention-aware Approach for Task-Agnostic Prompt Compression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 19395–19407.

Zhao, Z.; Jiang, Y.; Liu, H.; Wang, Y.; and Wang, Y. 2024. LibriSQA: A Novel Dataset and Framework for Spoken Question Answering with Large Language Models. *IEEE Transactions on Artificial Intelligence*.