Self-supervised Contrastive Cross-Modality Representation Learning for Spoken Question Answering

Chenyu You[†*] **Nuo Chen**[‡*] **Yuexian Zou**[‡§]
[†] Department of Electrical Engineering, Yale University, New Haven, CT, USA
[‡]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[§]Peng Cheng Laboratory, Shenzhen, China
chenyu.you@yale.edu
{nuochen,zouyx}@pku.edu.cn

Abstract

Spoken question answering (SQA) requires fine-grained understanding of both spoken documents and questions for the optimal answer prediction. In this paper, we propose novel training schemes for spoken question answering with a self-supervised training stage and a contrastive representation learning stage. In the self-supervised stage, we propose three auxiliary self-supervised tasks, including *utterance restoration*, *utterance insertion*, and *question discrimination*, and jointly train the model to capture consistency and coherence among speech documents without any additional data or annotations. We then propose to learn noise-invariant utterance representations in a contrastive objective by adopting multiple augmentation strategies, including *span deletion* and *span substitution*. Besides, we design a Temporal-Alignment attention to semantically align the speech-text clues in the learned common space and benefit the SQA tasks. By this means, the training schemes can more effectively guide the generation model to predict more proper answers. Experimental results show that our model achieves state-of-the-art results on three SQA benchmarks.

1 Introduction

Building an intelligent spoken question answering (SQA) system has attracted considerable attention from both academia and industry. In recent years, many significant improvements have achieved in speech processing and natural language processing (NLP) communities, such as multi-modal speech emotion recognition (Beard et al., 2018; Sahu et al., 2019; Priyasad et al., 2020; Siriwardhana et al., 2020), spoken language understanding (Mesnil et al., 2014; Chen et al., 2016, 2018; Haghani et al., 2018), and spoken question answering (Li et al., 2018; You et al., 2020a, 2021a,b). Among these topics, SQA is an especially challenging task, as it requires the machines to fully understand the semantic meaning in both speech and text data, and then provide the correct answer given a question and corresponding speech documents.

Automatic speech recognition (ASR) and text question answering (TQA) are two key components to build such a SQA system. The former module is used for transforming the speech sequences into text form, and the latter module trained on noisy ASR transcriptions utilizes NLP techniques to give a concrete answer. However, utilizing existing state-of-the-art SQA systems to retrieval answers still remain formidable challenges, such as ASR recognition errors. This is mainly because ASR systems usually fail to recognize the speech, leading to word errors (*e.g.*, "Barcelona" to "bars alone").

To address these issues, most existing SQA methods are either text-based (Li et al., 2018; Lee et al., 2018, 2019; Chuang et al., 2020) or fusion-based (You et al., 2021a, 2020a, 2021b). One line of research examines internal vector representations both in speech and text domains (Li et al., 2018; Lee et al., 2019), often using sub-word units for language modeling. Another line of work (You et al., 2021a,b) investigates the transfer learning problem about how to leverage a large amount of speech and text data to improve the performance of SQA. However, some critical challenges remain, such as robustness, generalization, and data efficiency.

Different from previous methods (Su and Fung, 2020; Li et al., 2018; Lee et al., 2019; You et al., 2021b), we move beyond leveraging dual nature of TQA and ASR to mitigate recognition errors. In this paper, we focus not only on extracting the cross-modality information for joint spoken and textual understanding, but also on the training procedure that may take the most advantage of the given dataset. Inspired by the recent advance in contrastive learning (Chen et al., 2020b; Khosla et al., 2020) and recent breakthrough (Devlin et al.,

[*]Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 28–39
November 7–11, 2021.

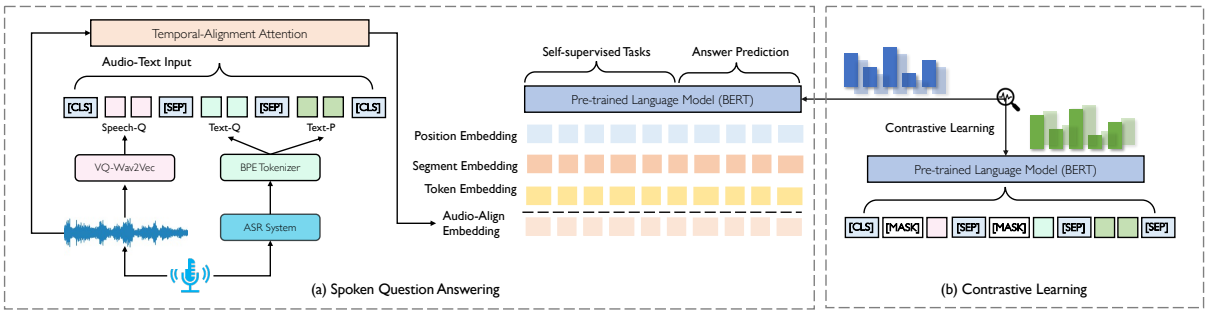

Figure 1: Overall architecture of our model: (a) For a spoken QA part, we use VQ-Wav2Vec and Tokenizer to transfer speech signals and text to discrete tokens. A Temporal-Alignment Attention mechanism is introduced to match each text embedding with the corresponding speech features. Then, we use BERT to learn sequential information of utterances with the proposed self-supervised tasks. We generate the final answer distribution on both domains. At inference time, we use the BERT only. (b) We incorporate contrastive learning strategies to train our SQA model in an auxiliary manner to improve the model performance.

2018; Liu et al., 2019; Rahman et al., 2020; Chen et al., 2020a) in the context of NLP, we propose a novel training framework for Spoken QA that integrates these two perspectives to improve spoken question answering performance. Our training framework contains two steps: (a) self-supervised training stage, and (b) contrastive training stage. During the self-supervised training stage, instead of building the complex spoken question answering model, we propose to learn a spoken question answering system based on pre-trained language models (PLMs) with several auxiliary self-supervised tasks. In particular, we introduce three self-supervised tasks, including *utterance restoration*, *utterance insertion*, and *question discrimination*, and jointly train the model with these auxiliary tasks in a multi-task setting. On the one hand, these auxiliary tasks enable the model to capture sequential order within the given passage. On the other hand, they effectively learn cross-modality knowledge without any additional dataset or annotations to generate better representations for answer prediction.

During the fine-tuning stage, along with the main QA loss, we incorporate the contrastive learning strategy to our framework in an auxiliary manner for the SQA tasks. Specifically, we use multiple augmentation strategies, including *span deletion* and *span substitution*, to develop the capability of learning noise-invariant utterance representations. In addition, we propose a novel attention mechanism, termed Temporal-Alignment Attention, to effectively learn cross-modal alignment between speech and text embedding spaces. By this mean, our proposed attention mechanism can encourage the training process to pay more attention to semantic relevance, consistency and coherency between speech and text in their contexts to provide better cross-modality representations for answer prediction. The overview of our framework is shown in Figure 1. We evaluate the proposed approach on the widely-used spoken question answering benchmark datasets - Spoken-SQuAD (Li et al., 2018), Spoken-CoQA (You et al., 2020a), and 2018 Formosa Grand Challenge (FGC). Experimental results show our proposed approach outperforms other state-of-the-art models when self-supervised training is preceded. Moreover, evaluation results indicate our learning schema can also consistently bring further improvements to the performance of existing methods with contrastive learning.

2 Related Work

Spoken Question Answering. Spoken question answering (Li et al., 2018; Lee et al., 2018, 2019; Su and Fung, 2020; Huang et al., 2021; You et al., 2021a,b, 2020a, 2021c; Chen et al., 2021) is a task of generating meaningful and concrete answers in response to a series of questions from spoken documents. Typical spoken QA systems focus on integrating ASR and TQA in one pipeline. ASRs are designed to transcribe audio recordings into written transcripts. However, current ASRs are not capable of processing every spoken document. Generated ASR transcripts may contain highly noisy data, which severely influences the performance of QA systems on speech documents. A number of works have explored the shortcomings of this issue. Li et al. (2018) and Lee et al. (2018) introduced sub-word unit strategy to alleviate the effects of speech recognition errors in SQA. SpeechBERT (Chuang et al., 2020) utilized the pre-trained BERT-

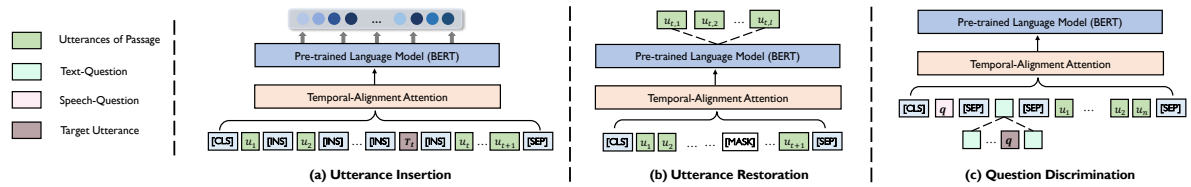

Figure 2: Auxiliary tasks in Self-supervised training.

based language model to effectively learn audio-text features. The model improved the performance of ASR by SpeechBERT. However, these works mainly focus on improving performance by exploiting internal information without considering learning the explicit mapping between human-made transcripts and corresponding ASR transcriptions, which is crucial to building Spoken QA systems. Lee et al. (2019) adopted an adversarial learning strategy to alleviate this gap to achieve remarkable performance improvements. In contrast to previous works in SQA, which only consider speech representations or confine to certain subtasks (*e.g.*, spoken multi-choice question answering and spoken conversational question answering), we not only model the interactions between speech and text data, but also focus on capturing semantic similarity. In parallel, our proposed method is a unified framework, which can be easily applied to a variety of downstream speech processing tasks.

Self-supervised Learning. Self-supervised learning (SSL) has become a promising solution for performance improvements by leveraging large amounts of unlabeled audio data. Substantial efforts have recently been dedicated to developing powerful SSL-based approaches in the machine learning community. (Oord et al., 2018; You et al., 2018; Schneider et al., 2019; Baevski et al., 2019; You et al., 2019b,a; Chung et al., 2019; Pascual et al., 2019; Liu et al., 2020; Chung et al., 2021; You et al., 2020b, 2021d,e). Oord et al. (2018) designed a Contrastive Predictive Coding (CPC) framework to learn compact latent representations to provide future predictions over future observations by combining autoregressive modeling and noise-contrastive estimation in an unsupervised manner. Later on, Schneider et al. (2019) further applied the learned generic speech representations to improve supervised ASR systems. Chung et al. (2019) and Liu et al. (2020) have taken advantage of state-of-the-art self-supervised pre-trained language models in the NLP community. These methods mainly focus on learning from audio data only, yet hardly exploit meaningful and relevant representations across both speech and text domains. Most recently, Khurana et al. (2020) investigated how to leverage speech-translation retrieval tasks into self-supervised learning. In this study, we explore an effective way to utilize cross-modality information via the self-supervised training scheme for SQA tasks without additional large-scale unlabeled datasets. In contrast, our proposed method yields such remarkable accuracy without using any extra data or annotations.

Contrastive Representation Learning. In parallel to self-supervised learning, an emerging subfield has explored the prospect of contrastive representation learning in the machine learning community (Kharitonov et al., 2021; Manocha et al., 2021; Oord et al., 2018; He et al., 2020; Chen et al., 2020b; Hjelm et al., 2018; Tian et al., 2019; Henaff, 2020; Wu et al., 2018; Khurana et al., 2020). This is often best understood as follows: pull together the *positive* and an anchor in embedding space, and push apart the anchor from many *negatives*. Thus, the choice of *negatives* can significantly determine the quality of the learned latent representations. Since contrastive learning is a framework to learn representations by comparing the similarity between different views of the data. In computer vision, Chen et al. (2020c) has demonstrated that the enlarged negative pool significantly enhances unsupervised representation learning. However, there are few attempts on contrastive learning to address downstream language processing tasks. Recently, few prior work (Kharitonov et al., 2021) incorporated CPC with time-domain data augmentation strategies into contrastive learning framework for speech recognition tasks. In contrast, we focus on learning interactions between speech and text modalities for spoken question answering tasks, and also introduce a set of auxiliary tasks on top of the former self-supervised training scheme to improve representation learning.

3 Methods

In this section, we first formalize the spoken question answering tasks. Furthermore, we introduce the key components of our method with self-supervised contrastive representation learning. Next, we describe the design of our proposed Temporal-Alignment Attention mechanism. Lastly, we discuss how to incorporate contrastive loss into our self-supervised training schema.

3.1 Task Formulation

Suppose that there is a dataset $\mathcal{D} \in \{Q_i, P_i, A_i\}_i^N$, where Q_i denotes a question, P_i denotes a passage with a answer A_i. In this study, similar to the SQA setting in (Lee et al., 2018; Kuo et al., 2020), we focus on extraction-based SQA, which can be applied to other types of language tasks. We use Spoken-SQuAD, Spoken-CoQA, and FGC to validate the robustness and generalization of our proposed approach. In Spoken-SQuAD, Q_i and A_i are both single sentences in text form, and P_i consists of multiple sentences in spoken form. In FGC, Q_i, A_i, and P_i are all in spoken form. Different from Spoken-SQuAD and FGC, Spoken-CoQA is in a multi-turn conversational SQA setting, which is more challenging than a single-turn setting. Moreover, it adopts Q_i in spoken form. The task is to learn a SQA model $\mathrm{G}(\cdot,\cdot)$ from $\mathcal{D}$ so that $\mathrm{G}(Q_i, P_i)$ can provide a most proper answer A_i to the given question Q_i.

3.2 Spoken question answering with PLMs.

Recent PLMs, such as BERT (Devlin et al., 2018) and ALBERT (Lan et al., 2020), learn meaningful language representations from large amounts of unstructured corpora, and have achieved superior performances on a wide range of downstream tasks in the domain of NLP. Following previous work (Lee et al., 2019), we consider building the SQA system with PLMs. We adopt BERT as the base model for a fair comparison. Similar to Lee et al. (2018), we concat ASR token sequences of a passage and a question as input to our SQA system. Specifically, given a passage $P=\{p_1, p_2, ..., p_n\}$ and a question $Q = \{q_1, q_2, ..., q_m\}$, we first concatenate all utterance sequences, which can be formulated as $\mathbf{X} = \{[\mathrm{CLS}], q_1, q_2, ..., q_m, [\mathrm{SEP}], p_1, p_2, ..., p_n, [\mathrm{SEP}]\}$. "[CLS]" and "[SEP]" denote begin token and separator token of each concatenated token sequence, respectively. We then utilize the pre-trained BERT to extract the hidden state features from the processed token sequences. Finally, we feed these representations to the following module, including a feed-forward network followed by a softmax layer, to obtain the probability distribution for each answer candidate given a textual passage-question pair. We use the cross-entropy loss as the question answering loss.

3.3 Self-supervised Training

Heading for a SQA model that can effectively make use of cross-modality knowledge with a limited number of training data and produce better contextual representations for answer prediction. To this end, we design three auxiliary self-supervised tasks, including *utterance restoration*, *utterance insertion*, and *question discrimination*. The objective of these auxiliary tasks is to capture the semantic relevance, coherence, and consistency between speech and text domains. Figure 2 illustrates three auxiliary self-supervised tasks. These tasks are jointly trained with the SQA model in a multi-task manner. More training examples of self-supervised training can be found in Table 1 and Appendix Table 4.

Utterance Insertion. PLMs often suffer from the limitations in capturing latent semantic and logical relationships in discourse-level, which refers to the problem that Next Sentence Prediction (NSP), the standard training objective of PLM-based approaches, negatively impact semantic topic shift without modeling coherence. One key reason is that NSP fails to capture sufficient semantic coherence with a incomprehensible passage (Lan et al., 2020), which leads performance degradation. Thus, learning the natural sequential relationship between consecutive utterances within a passage can significantly help the model understand the meaning of the passage.

In order to solve the above-mentioned problem, we design a more general self-supervised task with the spoken question answering context termed *utterance insertion*. In this way, it can enable the model to fully leverage the sequential relationship within a passage to improve the performance in calculating the semantic relevance between consecutive utterances. specifically, we first extract k consecutive utterances from one passage. Then we insert an utterance, which is randomly selected from another topic unrelated passage. Hence, suppose $k+1$ utterances consist of k utterances from the original passage and one from different corpus,

the goal is to predict the position of inserted utterance given the $k+1$ utterances. A special token $[\text{UI}]$ is introduced to be positioned before each utterance. The input can be formulated as follows:

$$\mathbf{X}_{\text{UI}} = [[\text{CLS}]\,[\text{UI}]_1\,u_1[\text{UI}]_2 \ldots [\text{UI}]_t\,u_{\text{INS}} \\ [\text{UI}]_{t+1}\,u_t \ldots [\text{UI}]_{k+1}\,u_k\,[\text{SEP}]], \quad (1)$$

where u_{INS} is the inserted utterance.

Utterance Restoration. One of the major tasks to train PLMs is mask token prediction (MTP), which requires the model to estimate the position of the masked utterance during the training stage. Although, recent work (Liu et al., 2019; Lan et al., 2020; Devlin et al., 2018; Joshi et al., 2020) found that utilizing this auxiliary task can improve model performance, it only focuses on learning syntactic and semantic representations of the word in token-level. However, spoken question answering is a more challenging task, which requires the deeper understanding of each utterance within a passage. To explicitly model the utterance-level interaction between utterances within a passage, we propose an utterance-level masking strategy termed *utterance restoration* to predict the utterance, which causes inconsistency. Specifically, suppose that a context is $c = \{u_1, u_2, ..., u_k\}$ including k consecutive utterances, we first randomly pick an utterance $u_t, t \in [0, k]$, and then replace all tokens in the u_t by using a special token $[\text{MASK}]$. Similarly, a special token $[\text{UR}]$ is introduced to be positioned before each utterance. To adapt the task in BERT, we formulate input of BERT encoder as follows:

$$\mathbf{X}_{\text{UR}} = [[\text{CLS}]\,[\text{UR}]_1\,u_1 \ldots [\text{UR}]_t\,u_{\text{MASK}} \\ [\text{UR}]_{t+1}\,u_{t+1} \ldots [\text{UR}]_k\,u_k[\text{SEP}]], \quad (2)$$

where u_{MASK} consists of only $[\text{MASK}]$ tokens, which has the same length with u_t.

Audio-Text Input. Inspired by recent success in video question answering (Kim et al., 2020), we leverage the cross-modality sequence modeling to generate audio-text sequence as input for *question discrimination* task. In this process, we utilize the BPE tokenizer to convert the ASR documents to a sequence of *Text-Question* and *Text-Passage* tokens, similar to PLMs (See Section 3.2). We utilize pre-trained VQ-Wav2Vec (Baevski et al., 2019) trained on Librispeech-960 (Panayotov et al., 2015) to encode speech signals to a sequence of input tokens for *Speech-Question*, since it outperforms the conventional RNN/CNN on sequence modeling.

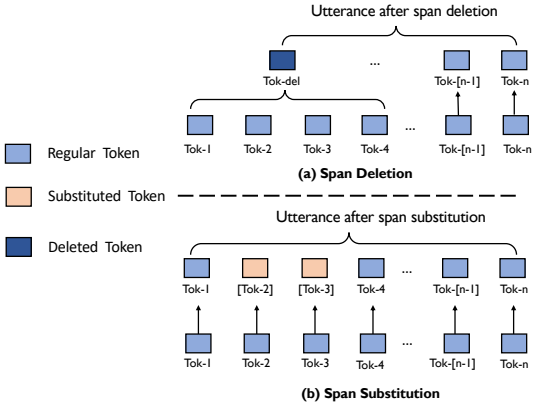

Figure 3: Auxiliary tasks in Contrastive Learning.

Question Discrimination. Recent work (Kuo et al., 2020) has shown that learning cross-modality representation is essential for SQA tasks. Hence we design *question discrimination* to consider building semantic alignments between speech and text by incorporating cross-modality knowledge into our model. Unlike the original goal of SQA (*i.e.*, finding the answer using a question and contextualized contexts in Section 3.2), we instead train the model to predict the proper text question using audio-text contexts. Specifically, we first randomly select $k-1$ questions in textual form from other passages, and then incorporate them into the corresponding question Q_t. We can reformulate the question as $\hat{Q} = \{Q_t^1, .., Q_t^{k-1}, Q_t\}$. The goal of this task is to find the correct *Text-Question* given a *Speech-Question* and *Text-Passage* contexts.

$$\bar{Q} = \text{argmax}\,\mathbb{P}(Q_i|Q_s, P), Q_i \in \hat{Q}, \quad (3)$$

where Q_s denotes the appropriate question in spoken form.

3.4 Temporal-Alignment Attention

Our proposed Temporal-Alignment Attention strategy is in the spirit of selectively leveraging cross-modality knowledge for SQA. Given an ASR token $\mathcal{U}^i$ and its corresponding acoustic-level MFCC features F^i, the goal is to enhance the SQA model by learning semantically meaningful alignment between speech and text domain [1]. To align speech and text embeddings, we use a simple fully-connected feed-forward layer. The speech embedding features $\hat{r}^i$ is processed by self-attention to obtain speech-aligned features. Formally, the proposed attention module is defined as follows:

[1] $\mathcal{U}^i$ can be any token in P_i and Q_i. Similar to (Kuo et al., 2020), for each acoustic frame, we use 40 MFCCs obtained from 40 FBANKs with 3 pitch features as input for ASR module and for our model.

Original text Input
[CLS] *Text-Question* [SEP] *Passage* [SEP]
Conceptual Audio-Text Input
[CLS] *Speech-Question* [SEP] *Text-Question option* [SEP] *Passage* [SEP]
Question-Discrimination Input
[CLS] *How does scholars divide the library?* [SEP] *What is the library for ?* [SEP] *The Vatican* ***at the stella clyde prairie****, more commonly called the Vatican Library or simply the* ***fact****, is the library of the Holy See, located in Vatican City...* [SEP]
Span Deletion Input
[CLS] *How does scholars divide the library?* [SEP] *What is* [DEL] *for ?* [SEP] *The Vatican* ***at the stella clyde prairie****,* [DEL] *commonly called the Vatican Library or simply the* ***fact****,* [DEL] *library of the Holy See, located in Vatican City...* [SEP]
Span Substitution Input
[CLS] *How does scholars divide the library?* [SEP] *What is the library for ?* [SEP] *The Vatican* ***at the stella clyde prairie****, more commonly named the Vatican Library or simply the* ***fact****, is the library of the Holy See, lied in Vatican City...* [SEP]

Table 1: Examples of audio-text input of our model. Original text input is used in a traditional BERT-liked model, *question discrimination* input is used in our self-supervised learning stage, and *span deletion* and *span substitution* inputs are used in a contrastive learning stage. Note that, for the readability, we do not use sub-word tokens in these examples. **Bold** denotes words in which the ASR error occurs. Blue and [DEL] represent the words in which the contrastive learning strategy is used.

$$
\begin{aligned}
r^i &= \sum_{i=1}^{|\mathcal{U}_i|} [\,\text{softmax}(W^i F^i) * F^i\,]_j, \\
\hat{r}^i &= \text{FNN}(r^i), \\
\{\mathbf{u}^i\} &= \text{Attention}(\hat{r}^i, \hat{r}^i, \hat{r}^i),
\end{aligned} \tag{4}
$$

where W^i is parameters. $*$ denotes element-wise multiplication. $[\cdot]_j$ is j-th column of a matrix. $\hat{r}^i$ and Attention are acoustic-level embedding and self-attention, respectively. Note that we set $\mathbf{u}^i$ of each special token (*e.g.*, [CLS]) to 0.

3.5 Contrastive Learning

Recent work (Wu et al., 2020) suggests two main arguments: (1) some deletion of unnecessary words in an utterance may not affect the original semantic meaning; (2) suppose that some necessary words (*e.g.*, not) are mistakenly deleted at times, it will provide extremely different semantic meaning. However, injecting some noises (*e.g.*, properly deleting some words) can improve the robustness of the model. Thus, in order to learn effective noise-invariant representation in sentence-level, we train our SQA model with a contrastive objective for performance improvement, in which we augment the training data with two sentence-level augmentation strategies, *span deletion* and *span substitution*[2]. The augmented input examples are shown in Figure 3. More training examples of contrastive learning can be found in Table 1.

- **Span Deletion**: we add one special token [DEL] to replace the deleted consecutive words of the utterance (*e.g.*, we randomly delete 5 spans, where each is of 5% length of the textual input sequences).
- **Span Substitution**: We randomly sample some words, and then replace them with synonyms to produce the augmented version (*e.g.*, we randomly select 30% spans of the utterances, and replace them with tokens which share similar semantic meanings).

In this stage, we first extract the [CLS] token representation $H \in R^{k \times d}$ from the last layer of the PLM, where $d = 768$ is the dimension of each word vector[3]. We create augmentations of original utterances with two sentence-level auxiliary tasks on top of the *Question Discrimination*, and then encode the augmented data using the same PLM, used in SQA section (See Figure 1 (a)), to construct the encoded representation $H_{anchor} \in R^{1 \times d}$. Our contrastive learning scheme consists of the following components: (1) we consider the representation corresponding to the correct Q_t as a *positive*, and others as many *negative*; (2) we use dot-production operation to compute the similarity scores between the joint speech-text representations and the anchor representation; (3) we apply a softmax function to the measured similarity scores. We leverage speech and text data for contrastive training, where the contrastive loss is as follows:

$$
\mathcal{L}_{con} = -\sum_{i}^{k} y_i \log(\text{softmax}(H \times H_{anchor}^T)) \tag{5}
$$

Multi-Task Learning Setup. We optimize our model with two main stages: (1) self-supervised training; (2) contrasitive learning. In the self-supervised training stage, we train our SQA model with three auxiliary tasks to obtain a better local optimum. We use binary cross-entropy loss in all proposed auxiliary tasks. The loss is computed by summing SQA answer prediction loss and all three

[2]The two augmentation strategies can happen in any position of the input.

[3]Similar to (Kuo et al., 2020), we use the [CLS] token to represent the sentence representation.

auxiliary SSL task losses with same ratio. In contrastive learning trainig stage, the loss is defined as a linear combination of SQA answer prediction loss and contrastive loss with the same ratio.

4 Experiments

In this section, we conduct experiments to compare our proposed method with various baselines and state-of-the-art approaches.

4.1 Datasets

We evaluate our approach on three benchmark datasets: Spoken-SQuAD (Li et al., 2018), Spoken-CoQA (You et al., 2020a), and FGC[4].

Spoken-SQuAD. Spoken-SQuAD (Li et al., 2018)[5] is a large listening comprehension corpus, where the training set and testing set consist of 37k and 5.4k question-answer pairs, respectively. The word error rate (WER) is around 22.77% in the training set, and around 22.73% on the testing set. The documents are in the form of speech, and the questions and answers are in the form of text, respectively. The manual transcripts of Spoken-SQuAD are collected from SQuAD benchmark dataset (Rajpurkar et al., 2016).

Spoken-CoQA. Spoken-CoQA (You et al., 2020a) is a large spoken conversational question answering (SCQA) corpus, where the training set and testing set consist of 40k and 3.8k question-answer pairs from 7 multiple domains, respectively. The WER is around 18.7%. The questions and passages are both in the form of text and speech, and answers are in the form of text, respectively. The goal is to generate a time span in the spoken multi-turn dialogues, and then answer questions based on the given passage and conversations.

FGC. FGC is a Chinese spoken multi-choice question answering (MCQA) corpus across a variety of domains. The number of question-answer pairs in the training set and testing set is 40k and 3.8k, respectively. Each PQC pair is composed of 1 passage, 1 question, and 4 corresponding answers, where only one answer is correct. All passages, questions, and multiple choices are in spoken form. Following the widely used setting in (Kuo et al., 2020), we apply the Kaldi toolkit to construct the ASR module. The WER is around 20.4%.

4.2 Implementation and Evaluation Setup

We utilize Pytorch to implement our model. We adopt BERT-base as our backbone encoder, which consists of 12 transformer layers. We set the maximum sequence length of input and the hidden vector dimension to 512 and 768, respectively. k in Section 3 is set to 9. We train our model on 2x 2080Ti for 2-3 days with a batch size of 4 per GPU using the Adam optimizer with an initial learning rate of 3×10^{-5}. For Spoken-CoQA, in order to utilize conversation history, we add the current question with previous 2 rounds of questions and ground-truth answers. When trained on FGC, we follow the standard multi-choice setting (Kuo et al., 2020), which takes questions, each candidate answers, and passages as inputs. We evaluate our model using the Exact Match (EM) and F1 to measure the performance of SQA models on Spoken-CoQA and Spoken-SQuAD, following previous work (Li et al., 2018; Kuo et al., 2020; Su and Fung, 2020). For FGC, we choose accuracy to evaluate the model performance on response quality.

4.3 Results

We report quantitative results on Spoken-SQuAD, Spoken-CoQA, and FGC datasets in Table 2. In our experiments, we set three aspects to study the effectiveness of key components of our method: (1) only using self-supervised learning strategies; (2) only using contrastive learning strategies; (3) we train the model with Temporal-Alignment Attention. Based on these initial aspects, we explore how effective each key component is for SQA.

We first evaluate if the model with three auxiliary tasks can generate a proper answer and how much improvement it can achieve over all evaluated models. For all datasets, our model significantly outperforms all evaluated methods on most of the metrics. Specifically, we observe that sequentially incorporating three proposed strategies brings superior performance improvements in terms of F1 and EM scores. Table 2 compares the importance of different auxiliary SSL tasks, which shows that QD $>$ UI $>$ UR in terms of response quality. This suggests that the auxiliary tasks can effectively aid the learning of the SQA model to learn more sequential information and cross-modality representations for the answer prediction.

[4]https://fgc.stpi.narl.org.tw/activity/techai2018

[5]In original Spoken-SQuAD dataset, questions are in text form. In this work, we utilize Google TTS to translate them into spoken form.

Method	Spoken-SQuAD	Spoken-CoQA					FGC
	Overall	Child.	Liter.	Mid-High.	News	Wiki	Overall	Acc
FlowQA (Huang et al., 2018)	56.7/70.8	22.6/35.8	22.4/35.2	22.0/34.2	21.4/33.6	22.0/34.7	22.1/34.7	-
BERT (Devlin et al., 2018)	58.6/71.1	41.7/55.6	40.1/54.6	39.8/52.7	40.1/53.8	40.6/53.8	40.6/54.1	77.0
BERT + SLU (Serdyuk et al., 2018)	59.3/71.7	42.0/55.7	41.4/54.6	40.0/53.1	40.5/54.0	41.1/54.6	41.0/54.4	77.6
Su and Fung (2020)	59.8/72.6	42.1/56.0	42.0/56.3	40.0/53.1	40.4/54.0	40.2/54.0	40.9/54.7	77.8
BERT+ TS-Attention (Kuo et al., 2020)	59.7/72.4	42.6/56.6	42.7/56.7	40.3/53.9	41.0/55.0	40.6/54.8	41.6/55.4	78.2
Only using Self-supervised Learning								
BERT + UR	59.4/71.7	42.6/55.8	41.9/55.6	40.6/53.8	40.9/54.0	40.7/54.3	41.5/54.7	77.5
BERT + UI	59.5/71.9	42.7/55.8	42.3/55.7	41.1/53.9	41.0/54.2	41.2/54.6	41.5/54.8	77.6
BERT + QD	59.9/72.4	43.0/56.2	42.2/55.7	41.2/54.3	41.5/54.4	41.6/54.8	41.9/55.1	78.0
BERT + UR + UI	59.8/72.6	43.1/56.3	42.3/55.7	41.5/54.5	41.4/54.6	41.5/54.9	41.9/55.2	78.1
BERT + UR + QD	60.2/72.6	43.4/56.7	42.6/55.9	41.8/54.7	41.5/54.9	42.0/55.4	42.5/55.5	78.4
BERT + UI + QD	60.5/73.0	43.5/56.8	42.5/56.1	41.6/55.0	41.2/54.8	42.0/55.6	42.4/55.6	78.5
BERT + UR + UI + QD	61.0/73.6	43.9/57.4	42.8/56.7	42.1/55.3	41.9/55.3	42.0/56.0	42.7/56.1	78.8
Only using Contrastive Learning								
BERT + SD	59.2/71.5	42.8/55.5	42.0/55.3	40.5/53.4	40.8/54.0	41.2/54.3	41.5/54.5	77.3
BERT + SS	59.4/71.5	42.9/55.7	42.1/55.6	40.3/53.4	41.0/54.1	41.4/54.2	41.5/54.6	77.4
BERT + SD + SS	59.6/71.8	43.3/56.1	42.4/55.6	41.2/54.2	41.4/54.5	41.2/54.5	41.9/54.9	77.9
BERT + T-A Attention	60.3/73.2	43.0/57.3	42.5/56.1	40.9/55.0	41.9/55.1	41.7/55.5	42.0/55.8	78.7
Ours	**62.5/75.5**	**46.5/59.5**	**46.1/59.1**	**44.3/57.3**	**44.9/57.6**	**45.2/58.0**	**45.4/58.3**	**81.3**

Table 2: The comparison between our method and other method on the SQA performance. UR, UI, and QD denote *utterance resorting*, *utterance insertion*, and *question discrimination*, respectively. SD and SS are *span deletion* and *span substitution*. T-A Attention denotes Temporal-Align Attention.

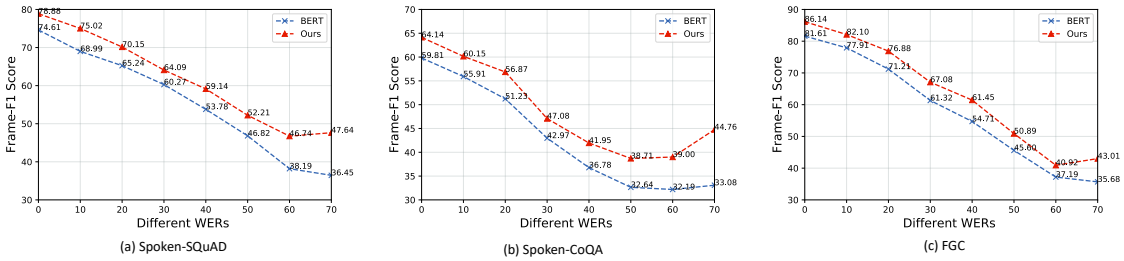

Figure 4: Performances of different WERs.

We then compare our method with other methods in terms of contrastive loss on three datasets. In Table 2, we utilize the proposed contrastive learning with the speech-text input as the auxiliary task, which consistently brings additional performance improvements on all datasets. When further explore the effectiveness of two augmentation strategies, we see that the model achieves comparable performances using SD or SS, and combining both of them enhances the capacity of the model to tackle many unseen sentence pairs. This indicates the importance of noise-invariant representations in boosting performance.

To validate the effectiveness of the proposed T-A Attention, we compare the models with T-A Attention and without it. The model with T-A Attention consistently shows remarkable performance improvements by 60.3%/73.2% (vs. 58.6%/71.1%) and 42.0%/55.6% (vs. 40.6%/54.1%) in terms of EM/F1 scores on Spoken-SQuAD and Spoken-CoQA, and 78.7% (vs. 77.0%) in terms of standard accuracy on FGC. Table 2 shows that our model achieves best results by 62.5%/75.5% (vs. 58.6%/71.1%), 45.4%/58.3% (vs. 40.6%/54.1%), and 81.3% (vs. 77.0%) across three datasets. This suggests that, by taking advantage of the proposed training scheme and T-A Attention, our model provides a more fine-grained understanding of spoken content to benefit the SQA answer prediction.

5 Ablation Study

Effects of Word Error Rates. To study how word error rates (WERs) will influence the model performance, we experiment with BERT, which is our baseline model, under different WERs. We randomly split three datasets into small-scale subsets of roughly equal training data size under different WERs for the ablation study. Then we compute Frame-level F1 score (Chuang et al., 2020) to evaluate the robustness of our proposed method with different WERs in Figure 4. We find that our model consistently achieves better results compared to the evaluated baseline. In addition, we find that higher WER leads to a consistent drop in all three spoken

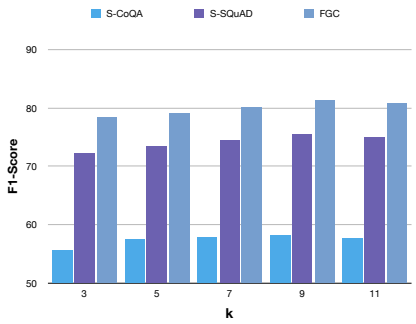

Figure 5: Effects of k.

Algorithm	S-SQuAD	S-CoQA	FGC
	F1	F1	Acc. (%)
BERT	71.1	54.1	77.0
w/ Co-Att (Lu et al., 2019)	72.8	55.0	77.9
w/ ICCN (Sun et al., 2020)	71.7	54.7	77.7
w/ S-Fusion (Siriwardhana et al., 2020)	68.1	51.8	75.1
w/ ST-Attention	**73.2**	**55.8**	**78.7**

Table 3: Effect of different attention mechanism.

question answering tasks. This suggests low WER brings these gains in all SQA settings.

Effects of Hyperparameter Selection. Self-supervised training enables the SQA model to capture sequential dependency between utterances along with semantic matching and maintain dialog coherence within a context. We explore the effects of different k, which determines the length of utterances in these auxiliary tasks. Figure 5 compares the performance of model with different k. We find that increasing the value of k clearly improves model performance, but it will not further increase after k = 9. We hypothesize that it gives rise to two potential reasons: (1) if the utterance length is too small within the context, the model cannot capture enough contextual information; (2) if the utterance length is too large, which introduces additional noise, it will not benefit the model performance. In our final models, we use k = 9 for self-supervised training.

Effects of T-A Attention. We further evaluate the effectiveness of various attention mechanisms in Table 3. We define BERT as the base model. We observe that the model with the proposed T-A attention strategy achieves state-of-the-art performance on three datasets. It clearly demonstrates T-A attention can effectively reduce the discrepancy between text and speech domains.

6 Conclusions

Spoken question answering requires fine-grained understanding of both speech and text data. To this end, we propose a novel training scheme for spoken question answering. By carefully designing several auxiliary tasks, we incorporate the self-supervised contrastive learning framework to capture consistency and coherence within speech documents and text corpus without any additional data. We further propose a novel Temporal-Alignment strategy to align audio features and textual concepts by performing mutual attention over two modalities. Our model achieves state-of-the-art performance on three SQA benchmark datasets. For future work, we will develop more effective auxiliary tasks to enhance the quality of answer prediction.

References

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. In *ICLR*.

Rory Beard, Ritwik Das, Raymond WM Ng, PG Keerthana Gopalakrishnan, Luka Eerens, Pawel Swietojanski, and Ondrej Miksik. 2018. Multi-modal sequence fusion via recursive attention for emotion recognition. In *COLING*, pages 251–259.

Nuo Chen, Fenglin Liu, Chenyu You, Peilin Zhou, and Yuexian Zou. 2020a. Adaptive bi-directional attention: Exploring multi-granularity representations for machine reading comprehension. In *ICASSP*.

Nuo Chen, Chenyu You, and Yuexian Zou. 2021. Self-supervised dialogue learning for spoken conversational question answering. In *INTERSPEECH*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020c. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore. 2018. Spoken language understanding without speech recognition. In *ICASSP*. IEEE.

Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. 2016. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Interspeech*, pages 3245–3249.

Yung-Sung Chuang, Chi-Liang Liu, and Hung-Yi Lee. 2020. SpeechBERT: Cross-modal pre-trained language model for end-to-end spoken question answering. In *INTERSPEECH*.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An unsupervised autoregressive model

for speech representation learning. *arXiv preprint arXiv:1904.03240*.

Yu-An Chung, Chenguang Zhu, and Michael Zeng. 2021. SPLAT: Speech-language joint pre-training for spoken language understanding. *arXiv preprint arXiv:2010.02295*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. 2018. From audio to semantics: Approaches to end-to-end spoken language understanding. In *SLT*. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *ICML*, pages 4182–4192. PMLR.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. FlowQA: Grasping flow in history for conversational machine comprehension. *arXiv preprint arXiv:1810.06683*.

Zhiqi Huang, Fenglin Liu, Xian Wu, Shen Ge, Helin Wang, Wei Fan, and Yuexian Zou. 2021. Audio-oriented multimodal machine comprehension via dynamic inter-and intra-modality attention. In *AAAI*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *TACL*.

Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. 2021. Data augmenting contrastive learning of speech representations in the time domain. In *SLT*, pages 215–222. IEEE.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *NeurIPS*.

Sameer Khurana, Antoine Laurent, and James Glass. 2020. Cstnet: Contrastive speech translation network for self-supervised speech representation learning. *arXiv preprint arXiv:2006.02814*.

Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. 2020. Self-supervised pre-training and contrastive representation learning for multiple-choice video QA. *arXiv preprint arXiv:2009.08043*.

Chia-Chih Kuo, Shang-Bao Luo, and Kuan-Yu Chen. 2020. An audio-enriched bert-based framework for spoken multiple-choice question answering. In *INTERSPEECH*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *ICLR*.

Chia-Hsuan Lee, Yun-Nung Chen, and Hung-Yi Lee. 2019. Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation. In *ICASSP*, pages 7300–7304. IEEE.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. ODSQA: Open-domain spoken question answering dataset. In *SLT*, pages 949–956. IEEE.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*.

Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP*, pages 6419–6423. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23.

Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. 2021. CDPAM: Contrastive learning for perceptual audio similarity. *arXiv preprint arXiv:2102.05109*.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2014. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, pages 5206–5210. IEEE.

Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, and Yoshua Bengio. 2019. Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv preprint arXiv:1904.03416*.

Darshana Priyasad, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2020. Attention driven fusion for multi-modal emotion recognition. In *ICASSP*, pages 3227–3231. IEEE.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Saurabh Sahu, Vikramjit Mitra, Nadee Seneviratne, and Carol Y Espy-Wilson. 2019. Multi-modal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription. In *Interspeech*, pages 3302–3306.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *ICASSP*, pages 5754–5758. IEEE.

Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. 2020. Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition. *arXiv preprint arXiv:2008.06682*.

Dan Su and Pascale Fung. 2020. Improving spoken question answering using contextualized word representation. In *ICASSP*, pages 8004–8008. IEEE.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *AAAI*, volume 34, pages 8992–8999.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Chenyu You, Nuo Chen, Fenglin Liu, Dongchao Yang, and Yuexian Zou. 2020a. Towards data distillation for end-to-end spoken conversational question answering. *arXiv preprint arXiv:2010.08923*.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021a. Contextualized attention-based knowledge transfer for spoken conversational question answering. In *INTERSPEECH*.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021b. Knowledge distillation for improved accuracy in spoken question answering. In *ICASSP*.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021c. MRD-Net: Multi-Modal Residual Knowledge Distillation for Spoken Question Answering. In *IJCAI*.

Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. 2019a. CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE Transactions on Medical Imaging*.

Chenyu You, Junlin Yang, Julius Chapiro, and James S. Duncan. 2020b. Unsupervised wasserstein distance guided domain adaptation for 3D multi-domain liver segmentation. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*.

Chenyu You, Linfeng Yang, Yi Zhang, and Ge Wang. 2019b. Low-Dose CT via Deep CNN with Skip Connection and Network in Network. In *Developments in X-Ray Tomography XII*. International Society for Optics and Photonics.

Chenyu You, Qingsong Yang, Lars Gjesteby, Guang Li, Shenghong Ju, Zhuiyang Zhang, Zhen Zhao, Yi Zhang, Wenxiang Cong, Ge Wang, et al. 2018. Structurally-sensitive multi-scale deep neural network for low-dose CT denoising. *IEEE Access*.

Chenyu You, Ruihan Zhao, Lawrence Staib, and James S Duncan. 2021d. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. *arXiv preprint arXiv:2105.07059*.

Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S. Duncan. 2021e. SimCVD: Simple Contrastive Voxel-Wise Representation Distillation for Semi-Supervised Medical Image Segmentation. *arXiv preprint arXiv:2108.06227*.

Appendix

A More Examples

Table 4 show examples used in the self-supervised training stage.

ASR Passage Title	Vatican-Library
ASR Question	How does scholars divide the library?
Original ASR Content	The Vatican **at the stella clyde prairie**, more commonly called the Vatican Library or simply the **fact**, is the library of the Holy See, located in Vatican City. Formally established in 1475, although it is much older, it is one of the oldest libraries in the world and contains one of the most significant collections of historical **tax**. It has 75,000 **courtesies** from throughout history, as well as 1.1 million printed books, which include some 8,500 **king abdullah**. The Vatican Library is a research library for history, **lot**, philosophy, science and theology. The Vatican Library is open to anyone who can document their qualifications **in** research needs. Photocopies for private study of pages from books published between 1801 and 1990 can be requested in person or by mail. In March 2014, **team** the Vatican Library began an initial four-year project of digitising its collection of manuscripts, to be made available online. The Vatican Secret Archives were separated from the library at the beginning of the 17th century; they contain another 150,000 items. Scholars have traditionally divided the history of the library into five periods, **pre ladder and ladder and having yon prevent a cannon vatican**. The **pre latter in period**, comprising the initial days of the library, dated from the earliest days of the Church. Only a handful of volumes survive from this period, **the summer** very significant.
Utterance Insertion	The Vatican **at the stella clyde prairie**, more commonly called the Vatican Library or simply the **fact**, is the library of the Holy See, located in Vatican City. Formally established in 1475, although it is much older, it is one of the oldest libraries in the world and contains one of the most significant collections of historical **tax**. It has 75,000 **courtesies** from throughout history, as well as 1.1 million printed books, which include some 8,500 **king abdullah**. The Vatican Library is a research library for history, **lot**, philosophy, science and theology. The Vatican Library is open to anyone who can document their qualifications **in** research needs. Photocopies for private study of pages from books published between 1801 and 1990 can be requested in person or by mail. The highly prized memorabilia which included item spanning the many stages of jackson's courier came for more than thirty fans associates and family members who contacted julian factions to sell their gifts and mementos of the singer. In March 2014, **team** the Vatican Library began an initial four-year project of digitising its collection of manuscripts, to be made available online. The Vatican Secret Archives were separated from the library at the beginning of the 17th century; they contain another 150,000 items. Scholars have traditionally divided the history of the library into five periods, **pre ladder and ladder and having yon prevent a cannon vatican**. The **pre latter in period**, comprising the initial days of the library, dated from the earliest days of the Church. Only a handful of volumes survive from this period, **the summer** very significant.
Utterance Restoration	The Vatican **at the stella clyde prairie**, more commonly called the Vatican Library or simply the **fact**, is the library of the Holy See, located in Vatican City. Formally established in 1475, although it is much older, it is one of the oldest libraries in the world and contains one of the most significant collections of historical **tax**. It has 75,000 **courtesies** from throughout history, as well as 1.1 million printed books, which include some 8,500 **king abdullah**. [MASK], [MASK], [MASK], . . . , [MASK]. Photocopies for private study of pages from books published between 1801 and 1990 can be requested in person or by mail. In March 2014, **team** the Vatican Library began an initial four-year project of digitising its collection of manuscripts, to be made available online. The Vatican Secret Archives were separated from the library at the beginning of the 17th century; they contain another 150,000 items. Scholars have traditionally divided the history of the library into five periods, **pre ladder and ladder and having yon prevent a cannon vatican**. The **pre latter in period**, comprising the initial days of the library, dated from the earliest days of the Church. Only a handful of volumes survive from this period, **the summer** very significant.

Table 4: Example of *Utterance Insertion* and *Utterance Restoration*. **Bold** denotes the words in which the ASR error occurs. Blue and [MASK] are the words in which the self-supervised learning strategies are used.