End-to-end Spoken Conversational Question Answering: Task, Dataset and Model

Chenyu You[♣*], **Nuo Chen**[♠*], **Fenglin Liu**[♠],
Shen Ge[♡], **Xian Wu**[♡], **Yuexian Zou**[♠]
[♣]Yale University [♡]Peking University [♠]Tencent
[♣]chenyu.you@yale.edu [♠]{nuochen,fenglinliu98,zouyx}@pku.edu.cn
[♡]{shenge,kevinxwu}@tencent.com

Abstract

In spoken question answering, the systems are designed to answer questions from contiguous text spans within the related speech transcripts. However, the most natural way that human seek or test their knowledge is via human conversations. Therefore, we propose a new **S**poken **C**onversational **Q**uestion **A**nswering task (SCQA), aiming at enabling the systems to model complex dialogue flows given the speech documents. In this task, our main objective is to build the system to deal with conversational questions based on the audio recordings, and to explore the plausibility of providing more cues from different modalities with systems in information gathering. To this end, instead of directly adopting automatically generated speech transcripts with highly noisy data, we propose a novel unified data distillation approach, DDNET, which effectively ingests cross-modal information to achieve fine-grained representations of the speech and language modalities. Moreover, we propose a simple and novel mechanism, termed Dual Attention, by encouraging better alignments between audio and text to ease the process of knowledge transfer. To evaluate the capacity of SCQA systems in a dialogue-style interaction, we assemble a **Spoken Co**nversational **Q**uestion **A**nswering (Spoken-CoQA) dataset with more than 40k question-answer pairs from 4k conversations. The performance of the existing state-of-the-art methods significantly degrade on our dataset, hence demonstrating the necessity of cross-modal information integration. Our experimental results demonstrate that our proposed method achieves superior performance in spoken conversational question answering tasks.

1 Introduction

Conversational question answering (CQA) has been studied extensively over the past few years within the natural language processing (NLP) communities (Zhu et al., 2018; Liu et al., 2019; Yang et al., 2019). Different from traditional question answering (QA) tasks, CQA aims to enable models to learn the representation of the context paragraph and multi-turn dialogues. Existing CQA methods (Huang et al., 2018a; Devlin et al., 2018; Xu et al., 2019; Gong et al., 2020) have achieved superior performances on several benchmark datasets, such as QuAC (Choi et al., 2018) and CoQA (Elgohary et al., 2018).

Current CQA research mainly focuses on leveraging written text sources in which the answer can be extracted from a large document collection. However, humans communicate with each other via spontaneous speech (*e.g.*, meetings, lectures, online conversations), which convey rich information. Consider our multimodal experience, fine-grained representations of both audio recordings and text documents are considered to be of paramount importance. Thus, we learn to draw useful relations between modalities (speech and language), which enables us to form fine-grained multimodal representations for end-to-end speech-and-language learning problems in many real-world applications, such as voice assistant and chat robot.

In this paper, we propose a novel and challenging spoken conversational question answering task - SCQA. An overview pipeline of this task is shown in Figure 1. Collecting such a SCQA dataset is a non-trivial task, as in contrast to current CQA tasks, we build our SCQA with two main goals as follows: (1) SCQA is a multi-turn conversational spoken question answering task, which is more challenging than *only* text-based task; (2) existing CQA methods rely on a single modality (text) as the context source. However, plainly leveraging uni-modality information is naturally undesirable for end-to-end speech-and-language learning problems since the useful connections between speech and text are elusive. Thus, employing data from

[*]Equal contribution.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1219 - 1232

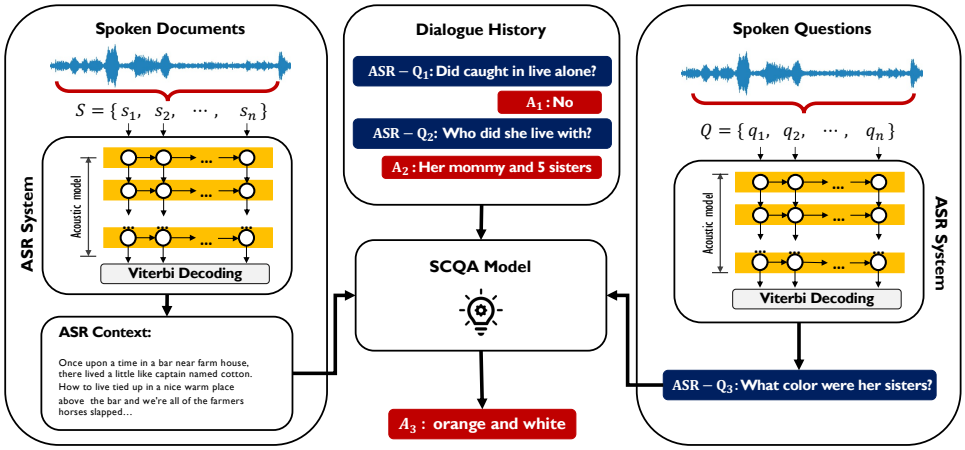

Figure 1: An illustration of flow diagram for spoken conversational question answering tasks with an example from our proposed Spoken-CoQA dataset.

Manual Transcript	ASR Transcript
Once upon a time, in a barn near a farm house, there lived a little white kitten named Cotton. Cotton lived high up in a nice warm place above the barn where all of the farmer's horses slept. But Cotton wasn't alone in her little home above the barn, oh no. She shared her hay bed with her mommy and 5 other sisters...	Once upon a time in a **bar** near farm house, there lived a little **like captain** named cotton. **How to live** tied up in a nice warm place above the **bar** and **we're** all of the farmers horses **slapped**. But **caught in** was not alone in her little home above the bar **in now**. She shared her hey bed with her mommy and 5 other sisters...
Q_1: Did Cotton live alone? A_1: no R_1: Cotton wasn't alone.	ASR-Q_1: Did **caught in** live alone? A_1: no R_1: **Caught in** wasn't alone.
Q_2: Who did she live with? A_2: with her mommy and 5 sisters R_2: with her mommy and 5 other sisters	ASR-Q_2: Who did she live with? A_2: with her mommy and 5 sisters R_2: with her mommy and 5 other sisters
Q_3: What color were her sisters? A_3: orange and white R_3: her sisters were all orange with beautiful white tiger stripes	ASR-Q_3: What color were her sisters? A_3: orange and white R_3: her sisters were all orange with beautiful white tiger stripes

Table 1: An example from Spoken-CoQA. We can observe large misalignment between the manual transcripts and the corresponding ASR transcripts. Note that the misalignment is in **bold** font and the example is the extreme case. For more dataset information, please see Section 5 and Appendix Section "More Information about Spoken-CoQA".

the context of another modality (speech) can allow us to form fine-grained multimodal representations for the downstream speech-and-language tasks; and (3) considering the speech features are based on regions and are not corresponding to the actual words, this indicates that the semantic inconsistencies between the two domains can be considered as the *semantic gap*, which requires to be resolved by the downstream systems themselves.

In order to provide a strong baseline for this challenging multi-modal spoken conversational question answering task, we first present a novel knowledge distillation (KD) method for the proposed SCQA task. Our intuition is that speech utterances and text contents share the dual nature property, and we can take advantage of this property to learn the correspondences between these two forms. Specifically, we enroll multi-modal knowledge into the *teacher* model, and then guide the *student* (*only* trained on noisy speech documents) to boost network performance. Moreover, considering that the semantics of the speech features and the textual representations are usually inconsistent, we introduce a novel mechanism, termed Dual Attention, to encourage fine-grained alignments between audio and text to close the cross-modal semantic gap between speech and language. One example of cross-modal gap is shown in Table 1. The experimental results show that our proposed DDNET achieves remarkable performance gains in the SCQA task. To the best of our knowledge, we are the first work in spoken conversational question answering task.

Our main contributions are as follows:

- We propose Spoken Conversational Question Answering task (SCQA), and comprise Spoken-CoQA dataset for machine comprehension of spoken question-answering style conversations. To the best of our knowledge, our Spoken-CoQA is the first spoken conversational question answering dataset.
- We develop a novel end-to-end method based

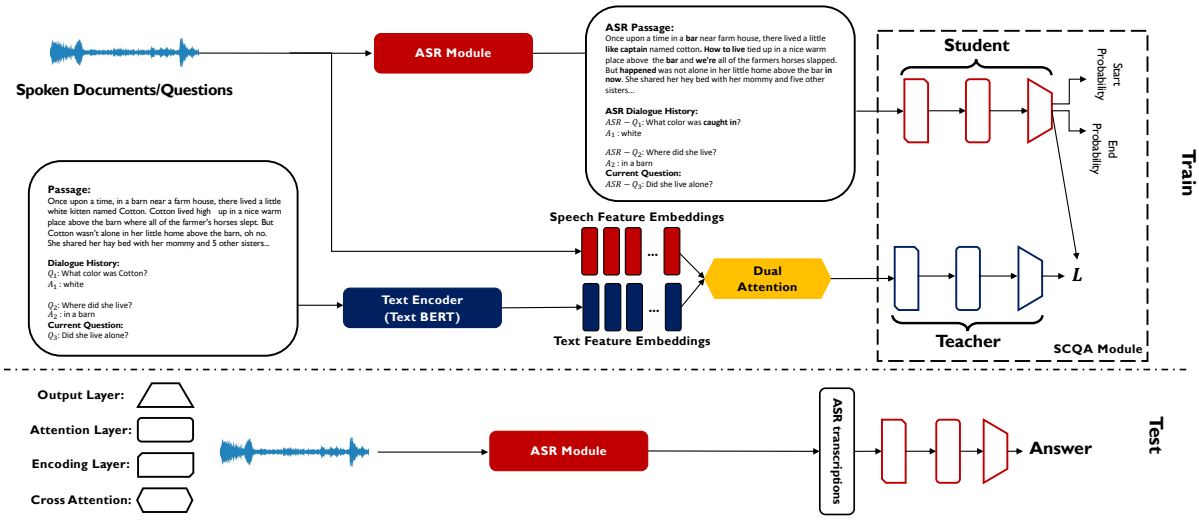

Figure 2: An illustration of the architecture of DDNET. In training stage, we adopt the teacher-student paradigm to enable the student model (only trained on speech documents) to achieve good performance. As for test, we only use student model for inference.

on data distillation to learn both from speech and language domain. Specifically, we use the model trained on clear texts as well as recordings to guide the model trained on noisy speech transcriptions. Moreover, we propose a novel Dual Attention mechanism to align the speech features and textual representations in each domain.

- We demonstrate that, by applying our proposed DDNET on several previous baselines, we can obtain considerable performance gains on our proposed Spoken-CoQA dataset.

2 Related Work

Text Question Answering. In recent years, the natural language processing research community has devoted substantial efforts to text question answering tasks (Huang et al., 2018a; Zhu et al., 2018; Xu et al., 2019; Zhang et al., 2020; Gong et al., 2020; Chen et al., 2020, 2021a). Within the growing body of work on machine reading comprehension, an important sub-task of text question answering, two signature attributes have emerged: the availability of large benchmark datasets (Choi et al., 2018; Elgohary et al., 2018; Reddy et al., 2019) and pre-trained language models (Devlin et al., 2018; Liu et al., 2019; Lan et al., 2020). However, these existing works typically focus on modeling the complicated context dependency in text form. In contrast, we focus on enabling the machine to build the capability of language recognition and dialogue modeling in both speech and text domains.

Spoken Question Answering. In parallel to the recent works in natural language processing (Huang et al., 2018a; Zhu et al., 2018), these trends have also been pronounced in the speech field (Chen et al., 2018; Haghani et al., 2018; Lugosch et al., 2019; Palogiannidi et al., 2020; You et al., 2021a,b,c,d, 2020a; Chen et al., 2021b; Xu et al., 2021; Su et al., 2020, 2021), where spoken question answering (SQA), an extended form of QA, has explored the prospect of machine comprehension in spoken form. Previous work on SQA typically includes two separate modules: automatic speech recognition (ASR) and text question answering. It involves transferring spoken content to ASR transcriptions, and then employs NLP techniques to handle speech tasks.

Existing methods (Tseng et al., 2016; Serdyuk et al., 2018; Su and Fung, 2020) focus on optimizing each module in a two-stage manner, where errors in the ASR module would result in severe performance loss. Lee et al. (2019) proved that utilizing clean texts can help model trained on the ASR transcriptions to boost the performance via domain adaptation. Chuang et al. (2019) cascaded the BERT-based models as a unified model, and then trained it in a joint manner of audio and text. However, the existing SQA methods aimed at solving a single question given the related passage, without building and maintaining the connections of different questions in the human conversations. In addition, we compare our Spoken-CoQA with existing SQA datasets (See Table 2). Unlike existing SQA datasets, Spoken-CoQA is a multi-turn con-

Dataset	Conversational	Spoken	Answer Type
TOEFL (Tseng et al., 2016)	×	√	Multi-choice
S-SQuAD (Li et al., 2018)	×	√	Spans
ODSQA (Lee et al., 2018)	×	√	Spans
S-CoQA	√	√	Free-form

Table 2: Comparison of Spoken-CoQA with existing spoken question answering datasets. S-SQuAD and S-CoQA denote Spoken-SQuAD and Spoken-CoQA, respectively.

versational SQA dataset, which is more challenging than single-turn benchmarks.

Knowledge Distillation. Hinton et al. (2015) introduced the idea of Knowledge Distillation (KD) in a *teacher-student* scenario. In other words, we can distill the knowledge from one model (massive or *teacher* model) to another (small or *student* model). Previous work has shown that KD can significantly boost prediction accuracy in natural language processing and speech processing (Kim and Rush, 2016; Hu et al., 2018; Huang et al., 2018b; Hahn and Choi, 2019; Liu et al., 2021b,a; Cheng et al., 2016b; Cheng and You, 2016; Cheng et al., 2016a; You et al., 2020b, 2021e, 2022b,a, 2018, 2019a,b; Lyu et al., 2018, 2019; Guha et al., 2020; Yang et al., 2020; Ma et al., 2021a,b), while adopting KD-based methods for SQA tasks has been less explored. In this work, our goal is to handle the SCQA tasks. More importantly, we focus the core nature property in speech and text: Can spoken conversational dialogues further assist the model to boost the performance? Finally, we incorporate the knowledge distillation framework to distill reliable dialogue flow from the spoken contexts, and utilize the learned predictions to guide the *student* model to train well on the noisy input data.

3 Task Definition

In this section, we propose the novel SCQA task and collect a Spoken-CoQA (S-CoQA) dataset, which uses the spoken form of multi-turn dialogues and spoken documents to answer questions in multi-turn conversations.

Given a spoken document D^s, we use D^t to denote the clean original text and D^a to denote the ASR transcribed document. We also have $Q^a_{1:L}$=$\{q^a_1, q^a_2, ..., q^a_L\}$, which is a collection of L-turn ASR transcribed spoken questions $Q^s_{1:L}$, as well as $A^t_{1:L}$= $\{a^t_1, a^t_2, ..., a^t_L\}$ which are the corresponding answers to the questions in clean texts. The objective of SCQA task is then to generate the answer a^t_L for question q^a_L, given document D^a, multi-turn history questions $Q^a_{1:L-1}$=$\{q^a_1, q^a_2, ..., q^a_{L-1}\}$, and reference answers $A^t_{1:L-1}$= $\{a^t_1, a^t_2, ..., a^t_{L-1}\}$. In other words, our task in the testing phase can be formulated as

$$\{D^s, Q^s_{1:L}\} \xrightarrow{\text{ASR}} \{q^a_L, D^a, Q^a_{1:L-1}, a^t_{1:L-1}\} \rightarrow a^t_L \quad (1)$$

Please note that in order to improve the performance, in the training phase, we make use of auxiliary information which are the clean texts of document D^t and dialogue questions Q^t=$\{q^t_1, q^t_2, \ldots, q^t_L\}$, to guide the training of student model. As a result, the training process could be formulated as below:

$$\left.\begin{array}{l}\text{student: } \{D^s, Q^s_{1:L}\} \xrightarrow{\text{ASR}} \{q^a_L, D^a, Q^a_{1:L-1}, a^t_{1:L-1}\} \\ \text{teacher: } \{D^t, Q^t_{1:L}\}\end{array}\right\} \rightarrow a^t_L$$

However, in the inference stage, these additional information of D^t and $Q^t_{1:L}$ are not needed.

4 DDNet

In this section, we propose DDNET to deal with the SCQA task, which is illustrated in Figure 2. We first describe the embedding generation process for both audio and text data. Next, we propose *Dual Attention* to fuse the speech and textual modalities. After that, we present the major components of the DDNET module. Finally we describe a simple yet effective distillation strategy in the proposed DDNET to learn enriched representations in the speech-text domain comprehensively.

4.1 Embedding

Given spoken words $S = \{s_1, s_2, ..., s_m\}$ and corresponding clean text words $T = \{t_1, t_2, ..., t_n\}$, we utilize Speech-BERT and Text-BERT to generate speech feature embedding $\mathbf{E}_s$=$\{\mathbf{E}_{s1}, \mathbf{E}_{s2}, ..., \mathbf{E}_{sm}\}$ and context word embedding $\mathbf{E}_t$=$\{\mathbf{E}_{t1}, \mathbf{E}_{t2}, ..., \mathbf{E}_{tn}\}$[1], respectively. Concretely, for speech input, we first use vq-wav2vec (Baevski et al., 2019) to transfer speech signals into a series of tokens, which is the standard tokenization procedure in speech related tasks. Next, use Speech-BERT (Chuang et al., 2019), a variant of BERT-based models retrained on our Spoken-CoQA dataset, to process the speech sequences for training. The text contents are embbed into a sequence of vectors via our text encoder - Text-BERT, with the same architecture of BERT-base (Devlin et al., 2018).

[1]In our implement, the padding strategy is used to keep m and n to be the same as the max sequence length.

4.2 Dual Attention

Dual Attention (DA) is proposed to optimize the alignment between speech and language domains by capturing useful information from the two domains. In particular, we first use cross attention to align speech and text representations in the initial stage. After that, we utilize contextualized attention to further align the cross-modal representations in the contextualized word-level. Finally, we employ the self-attention mechanism to form fine-grained audio-text representations.

Cross Attention. Inspired by ViLBERT (Lu et al., 2019), we apply the co-attention transformer layer, a variant of Self-Attention (Vaswani et al., 2017), as the Cross Attention module for the fusing of speech and text embeddings. The Cross Attention is implemented by the standard Attention module involving Multi-Head Attention (MHA) and Feed-Forward Network (FFN) (Vaswani et al., 2017) as below:

$$\begin{aligned} \text{Attention}(Q, K, V) &= \text{FFN}(\text{MHA}(Q, K, V)) \\ \text{CrossAttention}(F_1, F_2) &= \text{Attention}(F_1, F_2, F_2) \end{aligned} \tag{2}$$

where Q, K, V denote query, key, and value matrices, and F_1, F_2 denote features from difference modalities, respectively. The co-attention module then use the Cross Attention function to compute the cross attention-pooled features, by querying one modality using the query vector of another modality.

$$\begin{aligned} \hat{\mathbf{E}}_s^{\text{cross}} &= \text{CrossAttention}(\mathbf{E}_s, \mathbf{E}_t) \\ &= \text{Attention}(\mathbf{E}_s, \mathbf{E}_t, \mathbf{E}_t), \\ \hat{\mathbf{E}}_t^{\text{cross}} &= \text{CrossAttention}(\mathbf{E}_t, \mathbf{E}_s) \\ &= \text{Attention}(\mathbf{E}_t, \mathbf{E}_s, \mathbf{E}_s), \end{aligned} \tag{3}$$

where $\hat{\mathbf{E}}_s^{\text{cross}} \in \mathbb{R}^{n \times d}$, $\hat{\mathbf{E}}_t^{\text{cross}} \in \mathbb{R}^{n \times d}$ and d is the dimension of feature vectors.

Contextualized Attention (CA). After obtaining *speech-aware* representation $\hat{\mathbf{E}}_s^{\text{cross}}$ and *text-aware* representation $\hat{\mathbf{E}}_t^{\text{cross}}$, our next goal is to construct more robust contextualized cross-modal representations by integrating features from both modalities. The features with fused modalities are computed as follows:

$$\mathbf{H}_{\text{CA}} = \text{ReLU}(\hat{\mathbf{E}}_s^{\text{cross}} W_1^T)\text{ReLU}(\hat{\mathbf{E}}_x^{\text{cross}} W_1^T) W_2^T, \tag{4}$$

where $W_1, W_2 \in \mathbb{R}^{n \times d}$ are trainable weights.

Self-Attention. To build a robust SCQA system, special attention needs to be paid on the sequential order of the dialogue, since the changes in utterances order may cause severely low-quality and in-coherent corpora. As a result, to capture the long-range dependencies such as co-references for the downstream speech-and-language tasks, similar to (Li et al., 2016; Zhu et al., 2018), we introduce a self-attention layer to obtain the final Dual Attention (DA) representations.

$$\begin{aligned} \mathbf{E}_{\text{DA}} &= \text{SelfAttention}(\mathbf{H}_{\text{CA}}) \\ &= \text{Attention}(\mathbf{H}_{\text{CA}}, \mathbf{H}_{\text{CA}}, \mathbf{H}_{\text{CA}}). \end{aligned} \tag{5}$$

4.3 Key Components

The framework of our SCQA module is similar to recent works (Zhu et al., 2018; Huang et al., 2017), which is divided into three key components: Encoding Layer, Attention Layer and Output Layer.

Encoding Layer. Then documents and conversations (questions and answers) are first converted into the corresponding feature embeddings (i.e., character embeddings, word embeddings, and contextual embedding). The output contextual embeddings are then concatenated by the aligned cross-modal embedding $\mathbf{E}_{\text{DA}}$ to form the encoded input features:

$$\mathbf{E}_{enc} = [\mathbf{E}_t; \mathbf{E}_{\text{DA}}]. \tag{6}$$

Attention Layer. We compute the attention on the context representations of the documents and questions, and extensively exploit correlations between them. Note that we adopt the default attention layers in four baseline models.

Output Layer. After obtaining attention-pooled representations, the Output Layer computes the probability distributions of the start and end index within the entire documents and predicts an answer to current question:

$$\begin{aligned} \mathcal{L} = &-\log\mathcal{P}(\text{st} = a_{L,\text{st}}|\mathbf{X}) \\ &-\log\mathcal{P}(\text{ed} = a_{L,\text{ed}}|\mathbf{X}) \end{aligned} \tag{7}$$

where $\mathbf{X}$ denotes the input document D and Q^L, and "st", "ed" denote the start and end positions.

4.4 Knowledge Distillation

In previous speech-language models, the only guidance is the standard training objective to measure the difference between the prediction and the reference answer. However, for noisy ASR transcriptions, such criteria may not be suitable enough. To overcome such problem, we distill the knowledge

Domain	Passages	QA-Pairs	Passage Length	Avg.Turns
Children	357	3.5k	212	9.8
Literature	898	8.7k	275	9.7
Mid./High School	878	9.5k	308	10.8
News	967	9.5k	271	9.8
Wikipedia	864	8.9k	249	10.3
Overall	3964	40.1k	270	10.1

Table 3: Statistical analysis on Spoken-CoQA.

from our *teacher* model, and use them to guide the *student* model to learn contextual features in our SCQA task. Concretely, we set the model trained on the speech document and the clean text corpus as the *teacher* model and trained on the ASR transcripts as the *student* model, respectively. Thus, the *student* trained on low-quality data learns to absorb the knowledge that the *teacher* has discovered. Given the z_S and z_T as the prediction vectors by the *student* and *teacher* models, the objective is defined as:

$$\mathcal{L}_{\text{SCQA}}=\sum_{x\in\mathcal{X}}(\alpha\tau^2\text{KL}(p_\tau(z_S),p_\tau(z_T))+(1-\alpha)\mathcal{L}), \tag{8}$$

where $\text{KL}(\cdot)$ denotes the Kullback-Leibler divergence. $p_\tau(\cdot)$ is the softmax function with temperature τ, and α is a balancing factor.

5 Experiments and Results

In this section, we first describe the collection and filtering process of our proposed Spoken-CoQA dataset in detail. Next, we introduce several state-of-the-art language models as our baselines, and then evaluate the robustness of these models on our proposed Spoken-CoQA dataset. Finally, we provide a thorough analysis of different components of our method. Note that we use the default settings in all evaluated methods.

Data Collection. We detail the procedures to build Spoken-CoQA as follows. First, we select the conversational question-answering dataset CoQA (Reddy et al., 2019)[2] as our basis data since it is one of the largest public CQA datasets. CoQA contains around 8k stories (documents) and over 120k questions with answers. The average dialogue length of CoQA is about 15 turns, and the answers areis in free-form texts. In CoQA, the training set and the development set contain 7,199 and 500 conversations over the given stories, respectively. Therefore, we use the CoQA training set as our reference text of the training set and the CoQA development set as the test set in Spoken-CoQA.

Next, we employ the Google text-to-speech system to transform the questions and documents in CoQA into the spoken form, and adopt CMU Sphinx to transcribe the processed spoken contents into ASR transcriptions. In doing so, we collect more than 40G audio data, and the data duration is around 300 hours. The ASR transcription has a kappa score of 0.738 and Word Error Rates (WER) of 15.9%, which can be considered sufficiently good since it is below the accuracy threshold of 30% WER (Gaur et al., 2016). For the test set, we invite 5 human native English speakers to read the sentences of the documents and questions. The sentences of one single document are assigned to a single speaker to keep consistency, while the questions in one example may have different speakers. All speech files are sampled at 16kHz, following the common approach in the speech community. We provide an example of our Spoken-CoQA dataset in Table 1 and Fig. 5.

Data Filtering In our SCQA task, the model predicts the start and end positions of answers in the ASR transcriptions. As a result, during data construction, it is necessary for us to perform data filtering by eliminating question-answer pairs if the answer spans to questions do not exist in the noisy ASR transcriptions. We follow the conventional settings in (Lee et al., 2018)[3]. In our approach, an ASR question will be removed if the ground-truth answers do not exist in ASR passages. However, when coreference resolution and inference occurs, the contextual questions related to the previous ones are required to be discarded too. For the case of coreference resolution, we change the corresponding coreference. For the case of coreference inference, if the question has strong dependence on the previous one that has already been discarded, it will also be removed. After data filtering, we get a total number of our Spoken-CoQA dataset, we collect 4k conversations in the training set, and 380 conversations in the test set in our Spoken-CoQA dataset, respectively. Our dataset includes 5 domains, and we show the domain distributions in Table 3.

Baselines. For SCQA tasks, our DDNET is able to utilize a variety of backbone networks for SCQA

[2]Considering that the test set of CoQA (Reddy et al., 2019) idoes not publicly availablesh the test set, we follow the widely used setting in the spoken question answering task (Li et al., 2018), where we divide Spoken-CoQA dataset into train and test set.

[3]We compare different Speech APIs, *e.g.*, Google and CMU. Considering the quality of generated speech transcripts, we choose Google TTS for TTS and CMU Sphinx for ASR.

	CoQA				S-CoQA			
	CoQA dev		S-CoQA test		CoQA dev		S-CoQA test	
Methods	EM	F1	EM	F1	EM	F1	EM	F1
FlowQA (Huang et al., 2018a)	66.8	75.1	44.1	56.8	40.9	51.6	22.1	34.7
SDNet (Zhu et al., 2018)	68.1	76.9	39.5	51.2	40.1	52.5	41.5	53.1
BERT-base (Devlin et al., 2018)	67.7	77.7	41.8	54.7	42.3	55.8	40.6	54.1
ALBERT-base (Lan et al., 2020)	71.4	80.6	42.6	54.8	42.7	56.0	41.4	55.2
Average	68.5	77.6	42	54.4	41.5	54.0	36.4	49.3

Table 4: Comparison of four baselines (FlowQA, SDNet, BERT, ALBERT). Note that we denote Spoken-CoQA test set as S-CoQA test for brevity.

tasks. We choose several state-of-the-art language models (FlowQA (Huang et al., 2018a), SDNet (Zhu et al., 2018), BERT-base (Devlin et al., 2018), ALBERT (Lan et al., 2020)) as our backbone network baselines. We also compare our proposed DDNET with several state-of-the-art SQA methods (Lee et al., 2018; Serdyuk et al., 2018; Lee et al., 2019; Kuo et al., 2020). To use the *teacher-student* architecture in our models, we first train baselines on the CoQA training set as *teacher* and then evaluate the performances of testing baselines on CoQA dev set and Spoken-CoQA dev set. Finally, we train the baselines on the Spoken-CoQA training set as *student* and evaluate the baselines on the CoQA dev set and Spoken-CoQA test set. We provide quantitative results in Table 4.

Experiment Settings. We use the official BERT (Devlin et al., 2018) and ALBERT (Lan et al., 2020) as our textual embedding modules. We use BERT-base (Devlin et al., 2018) and ALBERT-base (Lan et al., 2020), which both include 12 transformer encoders, and the hidden size of each word vector is 768. BERT and ALBERT both utilize BPE as the tokenizer, but FlowQA and SDNet use SpaCy (Honnibal and Montani, 2017) for tokenization. Under the circumstance when tokens in spaCy (Honnibal and Montani, 2017) correspond to more than one BPE sub-tokens, we average the BERT embeddings of these BPE sub-tokens as the final embeddings for each token. For fair comparisons, we use standard implementations and hyper-parameters of four baselines for training. The balancing factor α is set to 0.9, and the temperature τ is set to 2. We train all models on 4 24GB RTX GPUs, with a batch size of 8 on each GPU. For evaluation, we use three metrics: Exact Match (EM), F_1 score and Audio Overlapping Score (AOS) (Li et al., 2018) to compare the model performance comprehensively. Please note that the metric numbers of baseline may be different from that in the CoQA leader board as we use our own implementations, Note that, we only utilize the *student* network for inference.

Results. We compare several *teacher-student* pairs on CoQA and Spoken-CoQA dataset and the quantitative results are shown in Table 4. We can observe that the average F1 scores is 77.6% when training on CoQA (text document) and testing on CoQA dev set. However, when training the models on Spoken-CoQA (ASR transcriptions) and testing on Spoken-CoQA test set, the average F1 scores drops significantly to 49.3%. For FlowQA, the performance even drops by 40.4 pts in terms of F1 score. This corroborates the importance of mitigating ASR errors.

Table 5 compares our approach DDNET to all the previous results. As shown in the table, our distillation models achieve strong performance, and incorporating DA mechanism further improves the results considerably. Our DDNET using BERT-base models as backbone achieves similar or better results compared to all the state-of-the-art methods, and we observe that using a larger encoder ALBERT-base will give further bring large gains on performance.

As seen from Table 6, we find that our best model ALBERT-base *only* trained with KD achieve an absolute EM/F1 improvement of +1.7pts/+1.7pts, +2.5pts/+2.5pts, on CoQA and S-CoQA, respectively. This shows that cross-modal information is useful for the model, hence demonstrating that such information is able to build more robust contextualized cross-modal representations for the network performance improvements. As shown in Table 6, we also observe that our approach ALBERT-base *only* trained with DA outperforms the original method by an absolute EM/F1 of +1.4pts/+1.2pts, +1.8pts/+2.0pts, on CoQA and S-CoQA, respectively. This indicates that the fine-grained alignment between audio and text learned through DA during training benefits the downstream speech-and-language tasks. Overall, our results suggest that such a network notably improves prediction performance for spoken conversational question answering tasks. Such significant improvements demonstrate the effectiveness of DDNET.

Methods	CoQA dev EM	CoQA dev F1	CoQA dev AOS	S-CoQA test EM	S-CoQA test F1	S-CoQA test AOS
FlowQA (Huang et al., 2018a)	40.9	51.6	30.6	22.1	34.7	16.7
FlowQA + sub-word unit (Li et al., 2018)	41.9	53.2	31.4	23.3	36.4	17.4
FlowQA+ SLU (Serdyuk et al., 2018)	41.2	52.0	30.6	22.4	35.0	17.1
FlowQA + back-translation (Lee et al., 2018)	40.5	52.1	30.8	22.9	35.8	17.3
FlowQA + domain adaptation (Lee et al., 2019)	41.7	53.0	31.8	23.4	36.1	17.7
FlowQA + **Dual Attention**	42.3	53.0	32.7	23.5	38.8	18.9
FlowQA + **Knowledge Distillation**	42.5	53.7	32.1	23.9	39.2	18.4
FlowQA + **Dual Attention+Knowledge Distillation**	**44.3**	**55.9**	**34.4**	**26.3**	**42.4**	**21.1**
SDNet (Zhu et al., 2018)	40.1	52.5	41.1	41.5	53.1	42.6
SDNet + sub-word unit (Li et al., 2018)	41.2	53.7	41.9	41.9	54.7	43.4
SDNet+ SLU (Serdyuk et al., 2018)	40.2	52.9	41.2	41.7	53.2	42.6
SDNet + back-translation (Lee et al., 2018)	40.5	53.1	41.5	42.4	54.0	42.9
SDNet + domain adaptation (Lee et al., 2019)	41.0	53.9	42.0	41.7	54.6	43.6
SDNet + **Dual Attention**	41.7	55.2	43.4	43.2	56.1	44.2
SDNet + **Knowledge Distillation**	41.7	55.6	43.6	43.6	56.7	44.3
SDNet + **Dual Attention+Knowledge Distillation**	**44.3**	**57.9**	**44.0**	**45.9**	**59.1**	**46.8**
BERT-base (Devlin et al., 2018)	42.3	55.8	50.1	40.6	54.1	48.0
BERT-base + sub-word unit (Li et al., 2018)	43.2	56.8	51.1	41.6	55.4	48.9
BERT-base+ SLU (Serdyuk et al., 2018)	42.5	56.1	50.3	41.0	54.6	48.1
BERT-base + back-translation (Lee et al., 2018)	42.9	56.5	50.5	41.5	55.2	48.6
BERT-base + domain adaptation (Lee et al., 2019)	43.1	57.0	51.0	41.7	55.7	49.0
aeBERT (Kuo et al., 2020)	43.0	56.9	51.5	41.8	55.6	50.3
BERT-base + **Dual Attention**	44.3	58.3	52.6	42.7	57.0	51.1
BERT-base + **Knowledge Distillation**	44.1	58.8	52.9	42.8	57.7	51.3
BERT-base + **Dual Attention+Knowledge Distillation**	**46.5**	**61.1**	**55.1**	**45.6**	**60.1**	**53.6**
ALBERT-base (Lan et al., 2020)	42.7	56.0	50.4	41.4	55.2	49.5
ALBERT-base + sub-word unit (Li et al., 2018)	43.7	57.2	51.2	42.6	56.8	50.3
ALBERT-base + SLU (Serdyuk et al., 2018)	42.8	56.3	50.5	41.7	55.7	49.7
ALBERT-base + back-translation (Lee et al., 2018)	43.5	57.1	50.9	42.4	56.4	50.0
ALBERT-base + domain adaptation (Lee et al., 2019)	43.5	57.0	51.5	42.7	56.7	50.7
ALBERT-base + **Dual Attention**	44.7	59.4	52.0	43.8	58.4	51.3
ALBERT-base + **Knowledge Distillation**	44.8	59.6	52.7	43.9	58.7	51.6
ALBERT-base + **Dual Attention+ Knowledge Distillation**	**47.3**	**61.9**	**55.5**	**46.1**	**61.3**	**53.6**

Table 5: Comparison of key components in DDNET. We denote the model trained on speech document and text corpus as the *teacher* model, and the one trained on the ASR transcripts as the *student* model.

Methods	SQuAD dev EM	SQuAD dev F1	S-SQuAD test EM	S-SQuAD test F1
FLowQA (Huang et al., 2018a)	51.9	65.7	49.1	63.9
FlowQA +**DA**	53.6	67.3	50.4	65.3
FLowQA+ **KD**	53.5	67.3	50.9	65.8
FLowQA+**DA**+ **KD**	**55.6**	**68.8**	**52.8**	**68.0**
SDNet (Zhu et al., 2018)	56.1	70.5	57.8	71.8
SDNet + **DA**	58.3	71.4	59.3	73.8
SDNet + **KD**	58.7	71.9	59.2	73.6
SDNet + **DA**+ **KD**	**60.1**	**73.7**	**60.9**	**75.7**
BERT-base (Devlin et al., 2018)	58.3	70.2	58.6	71.1
BERT-base + **DA**	59.9	72.8	61.0	74.1
BERT-base + **KD**	60.1	72.2	60.8	73.8
BERT-base + **DA**+ **KD**	**62.1**	**74.6**	**63.3**	**76.0**
ALBERT-base (Lan et al., 2020)	59.1	71.9	59.4	72.2
ALBERT-base + **DA**	60.5	73.1	61.2	74.2
ALBERT-base + **KD**	60.8	73.6	61.9	74.7
ALBERT-base + **DA**+ **KD**	**62.6**	**75.7**	**64.1**	**77.1**

Table 6: Comparison of our method. We set the model on text corpus as the *teacher* model, and the one on the ASR transcripts as the *student* model. DA and KD represent Dual Attention and knowledge distillation.

6 Ablation Study

We conduct ablation studies to show the effectiveness of several components in DDNet in this section and appendix.

Multi-Modality Fusion Mechanism. To study the effect of different modality fusion mechanisms, we introduce a novel fusion mechanism *Con Fusion*: first, we directly concatenate two output embedding from speech-BERT and text-BERT models, and then pass it to the encoding layer in the following SCQA module. In Table 8, we observe that Dual Attention mechanism outperform four baselines with *Con Fusion* in terms of EM and F1 scores. We further investigate the effect of unimodel input. Table 8 shows that *text-only* performs better than *speech-only*. One possible reason for this performance is that only using speech features can bring additional noise. Note that speech-only (text-only) means that we only feed the speech (text) embedding for speech-BERT (text-BERT) to the encoding layer in the SCQA module.

7 Conclusions

In this paper, we have presented SCQA, a new spoken conversational question answering task, for enabling human-machine communication. We make our effort to collect a challenging dataset - Spoken-CoQA, including multi-turn conversations and pas-

sages in both text and speech form. We show that the performance of existing state-of-the-art models significantly degrade on our collected dataset, hence demonstrating the necessity of exploiting cross-modal information in achieving strong results. We provide some initial solutions via knowledge distillation and the proposed dual attention mechanism, and have achieved some good results on Spoken-CoQA. Experimental results show that DDNET achieves substantial performance improvements in accuracy. In future, we will further investigate the different mechanisms of integrating speech and text content, and our method also opens up the possibility for downstream spoken language tasks.

References

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations (ICLR)*.

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

Nuo Chen, Linjun Shou, Min Gong, Jian Pei, and Daxin Jiang. 2021a. From good to best: Two-stage training for cross-lingual machine reading comprehension. *arXiv preprint arXiv:2112.04735*.

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

Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore. 2018. Spoken language understanding without speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Long Cheng and Chenyu You. 2016. Hybrid non-linear dimensionality reduction method framework based on random projections. In *International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. IEEE.

Long Cheng, Chenyu You, and Liguo Chen. 2016a. Identification of power line outages based on pmu measurements and sparse overcomplete representation. In *International Conference on Information Reuse and Integration (IRI)*. IEEE.

Long Cheng, Chenyu You, and Yani Guan. 2016b. Random projections for non-linear dimensionality reduction. *International Journal of Machine Learning and Computing*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yung-Sung Chuang, Chi-Liang Liu, and Hung-Yi Lee. 2019. SpeechBERT: Cross-modal pre-trained language model for end-to-end spoken question answering. *arXiv preprint arXiv:1910.11559*.

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*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. Dataset and baselines for sequential open-domain question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yashesh Gaur, Walter S Lasecki, Florian Metze, and Jeffrey P Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th Web for All Conference*.

Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. 2020. Recurrent chunking mechanisms for long-text machine reading comprehensio. *arXiv preprint arXiv:2005.08056*.

Indranil Guha, Syed Ahmed Nadeem, Chenyu You, Xiaoliu Zhang, Steven M Levy, Ge Wang, James C Torner, and Punam K Saha. 2020. Deep learning based high-resolution reconstruction of trabecular bone microstructures from low-resolution ct scans using gan-circle. In *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*.

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 *Spoken Language Technology Workshop (SLT)*. IEEE.

Sangchul Hahn and Heeyoul Choi. 2019. Self-knowledge distillation in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 423–430.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. 2018. Attention-guided answer distillation for machine reading comprehension. *arXiv preprint arXiv:1808.07644*.

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

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *arXiv preprint arXiv:1711.07341*.

Mingkun Huang, Yongbin You, Zhehuai Chen, Yanmin Qian, and Kai Yu. 2018b. Knowledge distillation for sequence model. *Interspeech*, pages 3703–3707.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1317–1327.

Chia-Chih Kuo, Shang-Bao Luo, and Kuan-Yu Chen. 2020. An audio-enriched bert-based framework for spoken multiple-choice question answering. *arXiv preprint arXiv:2005.12142*.

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 *International Conference on Learning Representations (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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. ODSQA: Open-domain spoken question answering dataset. In *Spoken Language Technology Workshop (SLT)*. 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*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Fenglin Liu, Xian Wu, Chenyu You, Shen Ge, Yuexian Zou, and Xu Sun. 2021a. Aligning source visual and target language domains for unpaired video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. 2021b. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems (NeurIPS)*.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*.

Qing Lyu, Chenyu You, Hongming Shan, and Ge Wang. 2018. Super-resolution mri through deep learning. *arXiv preprint arXiv:1810.06776*.

Qing Lyu, Chenyu You, Hongming Shan, Yi Zhang, and Ge Wang. 2019. Super-resolution mri and ct through gan-circle. In *Developments in X-ray tomography XII*. International Society for Optics and Photonics.

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. 2021a. Good students play big lottery better. *arXiv preprint arXiv:2101.03255*, 3.

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. 2021b. Undistillable: Making a nasty teacher that cannot teach students. *arXiv preprint arXiv:2105.07381*.

Elisavet Palogiannidi, Ioannis Gkinis, George Mastrapas, Petr Mizera, and Themos Stafylakis. 2020. End-to-end architectures for asr-free spoken language understanding. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Dan Su and Pascale Fung. 2020. Improving spoken question answering using contextualized word representation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8004–8008. IEEE.

Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems (NeurIPS)*.

Kun Su, Xiulong Liu, and Eli Shlizerman. 2021. How does it sound? *Advances in Neural Information Processing Systems (NeurIPS)*.

Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial TOEFL listening comprehension test by machine. *arXiv preprint arXiv:1608.06378*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems (NeurIPS)*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Review conversational reading comprehension. *arXiv preprint arXiv:1902.00821*.

Weiyuan Xu, Peilin Zhou, Chenyu You, and YueXian Zou. 2021. Semantic transportation prototypical network for few-shot intent detection. In *INTERSPEECH*.

Linfeng Yang, Rajarshi P Ghosh, J Matthew Franklin, Simon Chen, Chenyu You, Raja R Narayan, Marc L Melcher, and Jan T Liphardt. 2020. Nuset: A deep learning tool for reliably separating and analyzing crowded cells. *PLoS computational biology*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems (NeurIPS)*.

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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

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

Chenyu You, Nuo Chen, and Yuexian Zou. 2021d. Self-supervised contrastive cross-modality representation learning for spoken question answering. In *Findings of the Association for Computational Linguistics: EMNLP*.

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*. Springer International Publishing.

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, Fenglin Liu, Sandeep Chinchali, Ufuk Topcu, Lawrence Staib, and James S Duncan. 2022a. Class-aware generative adversarial transformers for medical image segmentation. *arXiv preprint arXiv:2201.10737*.

Chenyu You, Ruihan Zhao, Lawrence Staib, and James S Duncan. 2021e. 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. 2022b. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020. SG-Net: Syntax-guided machine reading comprehension. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. SDNet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

Appendix

A Temperature τ

To study the effect of temperature τ, we conduct the additional experiments of four baselines with the standard choice of the temperature $\tau \in \{1, 2, 4, 6, 8, 10\}$. All models are trained on Spoken-CoQA dataset, and validated on the Spoken-CoQA test set. We present the results in Figure 4. When τ is set to 2, four baselines all achieve their best performance in term of F1 and EM metrics.

B Effects of Different Word Error Rates

We study how the network performances change when trained with different word error rates (WER) in Figure 3. Specifically, we first split Spoken-SQuAD and Spoken-CoQA into smaller groups with different WERs. Then we utilize Frame-level F1 score (Chuang et al., 2019) to validate the effectiveness of our proposed method on Spoken-CoQA. In Figure 3, we find that all evaluated networks for two tasks are remarkably similar: all evaluated models suffer larger degradation in performance at higher WER, and adopting knowledge distillation strategy is capable of alleviating such issues. Such phenomenon further demonstrates the importance of *knowledge distillation* in the case of high WER.

C Results on Human Recorded Speech

The results using BERT-base as the baseline are shown in Table 7. We train the model in the Spoken-CoQA training dataset and evaluate the model in both machine synthesized and human recorded speech. As shown in Table 7, the average EM/F1/AOS scores using BERT fell from 40.6/54.1/48.0 to 39.4/53.0/46.8, respectively. In addition, the similar trends can be observed on our proposed method. We hypothesise that the human recorded speech introduces additional noise during training, which leads to the performance degradation.

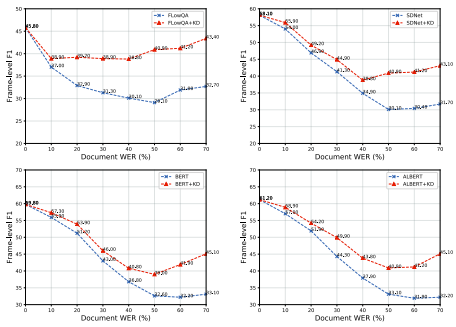

Figure 3: Comparison of different WER on Spoken-CoQA.

Table 7: Comparisons between human recorded speech and synthesized speech. We employ BERT as our base model.

Dataset	Method	EM	F1	AOS
Machine	BERT	40.6	54.1	48.0
	BERT+KD+DA	45.6	60.1	53.6
Human	BERT	39.4	53.0	46.8
	BERT+KD+DA	44.7	59.4	53.1

Table 8: Comparison of different fusion mechanisms in DDNET.

Models	CoQA dev EM	CoQA dev F1	S-CoQA test EM	S-CoQA test F1
FlowQA (Huang et al., 2018a)	40.9	51.6	22.1	34.7
+ *speech-only*	40.8	51.2	21.8	34.0
+ *text-only*	41.1	51.7	22.4	35.3
+ *Con Fusion*	41.0	52.0	22.1	35.2
+ **Dual Attention**	**42.3**	**53.0**	**23.5**	**38.8**
SDNet (Zhu et al., 2018)	40.1	52.5	41.5	53.1
+ *speech-only*	39.3	51.6	40.9	52.28
+ *text-only*	40.2	52.7	41.5	53.3
+ *Con Fusion*	40.3	52.6	41.5	53.2
+ **Dual Attention**	**41.7**	**55.2**	**43.2**	**56.1**
BERT-base (Devlin et al., 2018)	42.3	55.8	40.6	54.1
+ *speech-only*	41.9	55.8	40.2	54.1
+ *text-only*	42.4	56.0	40.9	54.3
+ *Con Fusion*	42.3	56.0	40.8	54.1
+ **Dual Attention**	**44.3**	**58.3**	**42.7**	**57.0**
ALBERT-base (Lan et al., 2020)	42.7	56.0	41.4	55.2
+ *speech-only*	41.8	55.9	41.1	54.8
+ *text-only*	42.9	56.3	41.4	55.7
+ *Con Fusion*	42.7	56.1	41.3	55.4
+ **Dual Attention**	**44.7**	**59.4**	**43.8**	**58.4**

D More Information about Spoken-CoQA

To perform qualitative analysis of speech features, we visualize the log-mel spectrogram features and the mel-frequency cepstral coefficients (MFCC) feature embedding learned by *DDNet* in Figure 5. We can observe how the spectrogram features respond to different sentence examples. In this example, we observe that given the text document (ASR-document), the conversation starts with the question Q_1 (ASR-Q_1), and then the system requires to answer Q_1 (ASR-Q_1) with A_1 based on a contiguous text span R_1. Compared to the existing benchmark datasets, ASR transcripts (both the document and questions) are much more difficult for the machine to comprehend questions, reason among the passages, and even predict the correct answer.

E More Comparisons on Spoken-SQuAD

To verify that our proposed DDNET is not biased towards specific settings, we conduct a series of experiments on Spoken-SQuAD (Li et al., 2018) by training the *teacher* model on textual documents, and the *student* model on the ASR transcripts. From the Table 5, compared with the performances on Spoken-CoQA, all baselines performances improve by a large margin, indicating our proposed dataset is a more challenging task for current models. We verify that, in the setting (KD+DA), the model consistently achieves significant performance boosts on all

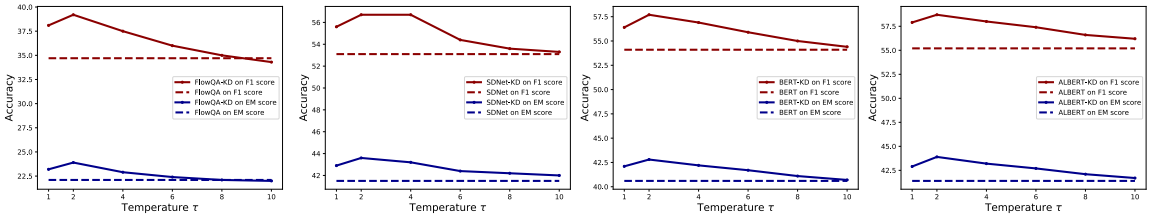

Figure 4: Ablation studies of temperature τ on DDNET performance (FlowQA, SDNet, BERT, ALBERT). Red and blue denote the results on Spoken-CoQA test set.

baselines. Specifically, for FlowQA, our method achieves 55.6%/68.8% (vs.51.9%/65.7%), and 52.8%/68.0% (vs.49.1%/63.9%) in terms of EM/F1 score over the text documents and ASR transcriptions, respectively. For SDNet, our method outperforms the baseline without distillation, achieving 60.1%/73.7% (vs.56.1%/70.5%) and 60.9%/75.7% (vs.57.8%/71.8%) in terms of EM/F1 score. As for two BERT-based models (BRET-large and ALBERT-large), our methods with KD consistently improve EM/F1 scores to 62.1%/74.6% (vs.58.3%/70.2%) and 63.3%/76.0% (vs.58.6%/71.1%); 62.6%/75.7% (vs.59.1%/71.9%) and 64.1%/77.1% (vs.59.4%/72.2%), respectively. These results confirm the importance of knowledge distillation strategy and dual attention mechanism.

F Broader Impact

In this section, we acknowledge that our work will not bring potential risks to society considering the data is from open source with no private or sensitive information. We also discuss some limitations of our work. First, we admit that using Google TTS for TTS and CMU Sphinx for ASR may affect the distribution of errors compared with the human recorded speech. Second, we currently cover only English language but it would be interesting to see that contributions for other languages would follow. Finally, as our collection comes with reliable data, it should trigger future analysis works on analyzing spoken conversational question answering biases.

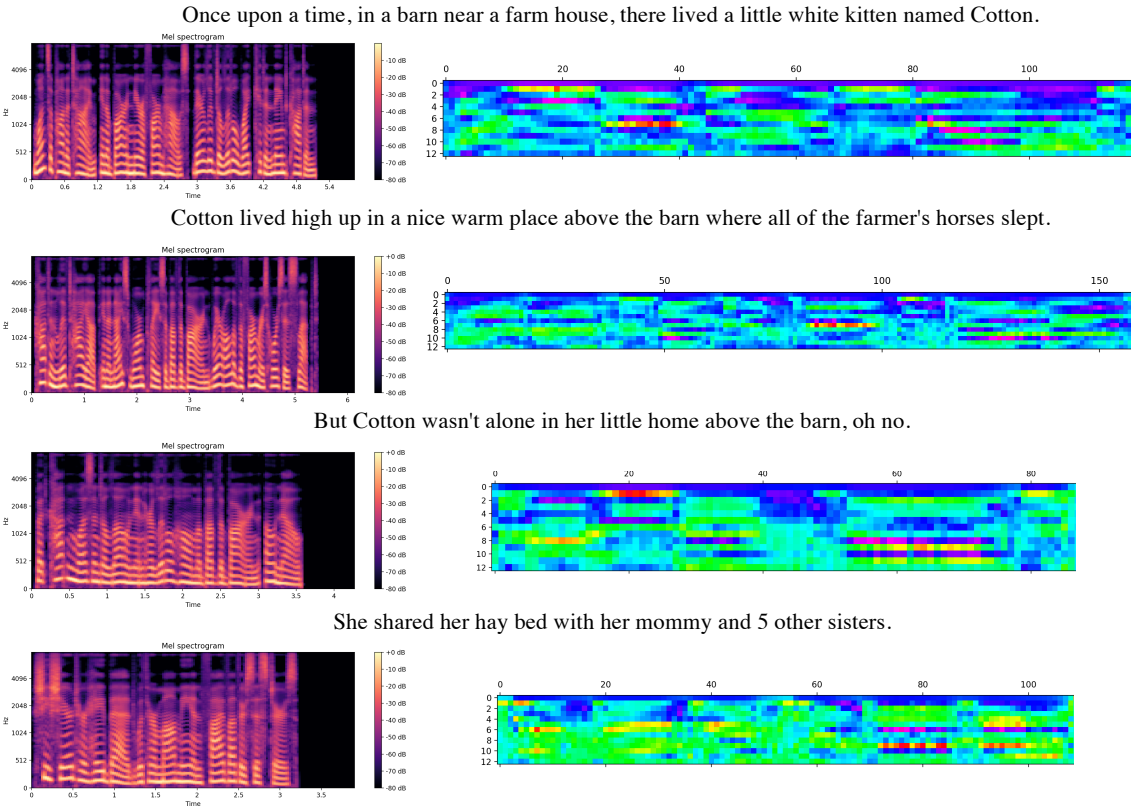

Figure 5: Examples of the log-mel spectrograms and the corresponding MFCC feature embedding. It can see that the log-mel spectrograms corresponds to different example sentences from the Spoken-CoQA dataset.