SQuAD-SRC: A Dataset for Multi-Accent Spoken Reading Comprehension

Yixuan Tang , Anthony K.H. Tung
Department of Computer Science, National University of Singapore
{yixuan, atung}@comp.nus.edu.sg

Abstract

Spoken Reading Comprehension (SRC) is a challenging problem in spoken natural language retrieval, which automatically extracts the answer from the text-form contents according to the audio-form question. However, the existing spoken question answering approaches are mainly based on synthetically generated audio-form data, which may be ineffectively applied for multi-accent spoken question answering directly in many real-world applications. In this paper, we construct a large-scale multi-accent human spoken dataset SQuAD-SRC, in order to study the problem of multi-accent spoken reading comprehension. We choose 24 native English speakers from six different countries with various English accents and construct audio-form questions to the correspondent text-form contents by the chosen speakers. The dataset consists of 98,169 spoken question answering pairs and 20,963 passages from the popular machine reading comprehension dataset SQuAD. We present a statistical analysis of our SQuAD-SRC dataset and conduct extensive experiments on it by comparing cascaded SRC approaches and the enhanced end-to-end ones. Moreover, we explore various adaption strategies to improve the SRC performance, especially for multi-accent spoken questions.

1 Introduction

Recent years have witnessed rapid progress in models bridging semantic connections between speech and natural language modalities in tasks such as automatic speech recognition (ASR), spoken translation and spoken language understanding. Compared to these fields, the task of spoken question answering (SpokenQA) is less investigated [Li *et al.*, 2018; You *et al.*, 2022]. Currently, QA systems are widely adopted in real-world applications, allowing users to interact with them via speech interfaces [Raux *et al.*, 2005]. Taking smart voice assistants [Hoy, 2018] as an example, users ask questions in audio forms while the QA system can explore numerous knowledge resources stored in text format to retrieve the answers. The abundance of large-scale textual QA datasets [Nguyen *et al.*, 2016; Rajpurkar *et al.*, 2016; Trischler *et al.*, 2017; Yang *et al.*, 2018] and visual QA datasets [Goyal *et al.*, 2019; Hudson and Manning, 2019] have provoked the development of high-performance QA models. Therefore, the construction of a high-quality large-scale SpokenQA benchmark is in crucial need.

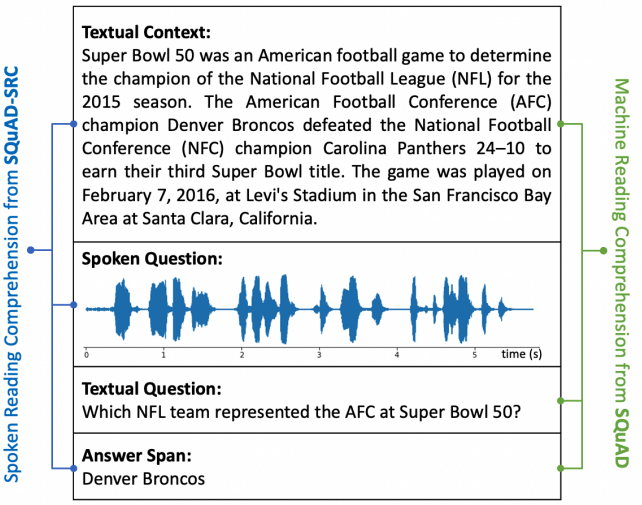


Figure 1: Comparison between textual reading comprehension and spoken reading comprehension with an example from SQuAD-SRC.

In general, SpokenQA aims to generate the answer to a given context-question pair with audio input. Several SpokenQA datasets have been proposed [Abdelnour *et al.*, 2018; Fayek and Johnson, 2020; Huang *et al.*, 2021; Lipping *et al.*, 2022; Li *et al.*, 2018; Lee *et al.*, 2018]. However, they suffer from at least one of the following defects. First, The audio utterances are synthetically generated via text-to-speech (TTS) engines [Li *et al.*, 2018; You *et al.*, 2022]. Due to the lack of variations, it cannot well reflect the challenges associated with complex human voices. Second, the scale of data is too small to train powerful large-scale neural networks [Huang *et al.*, 2021; Lipping *et al.*, 2022; Lee *et al.*, 2018].

To address the issues above, we introduce an open-source, large-scale, naturally recorded, multi-accent spoken reading comprehension (SRC) dataset named SQuAD-SRC. The SRC task proposed differs from the existing SpokenQA datasets

Dataset	language	Data source	Modality: q, c, a	Audio generation	# QA pairs
Spoken SQuAD [Li *et al.*, 2018]	English	SQuAD	text, audio, either	synthetic	42k
Spoken-CoQA [You *et al.*, 2022]	English	CoQA	audio, audio, text	synthetic	120k
DAQA [Fayek and Johnson, 2020]	English	N.A.	text, audio, text	synthetic	599k
CLEAR [Abdelnour *et al.*, 2018]	English	N.A.	text, audio, text	synthetic	2M
L-TOEFL & CET [Huang *et al.*, 2021]	English	TOEFL & CET	text, audio, text	natural	2k
Clotho-AQA [Lipping *et al.*, 2022]	English	Clotho	text, audio, text	natural	12k
ODSQA [Lee *et al.*, 2018]	Chinese	DRCD	audio, audio, text	natural	3k
SQuAD-SRC (ours)	English	SQuAD	audio, text, text	natural	98k

Table 1: Comparison of existing SpokenQA datasets with SQuAD-SRC. The Modality column refers to the modality of {question, context, answer} triplets provided in the datasets. For audio generation, synthetic means generated by algorithms, and natural means recorded by humans.

by asking the model to answer spoken questions based on the textual context given, which suits the need of real-world application scenarios [Hoy, 2018]. SQuAD-SRC is built upon the popular machine reading comprehension dataset, SQuAD [Rajpurkar *et al.*, 2016]. It contains questions on paragraphs from Wikipedia, where the answer to each question is a text span of the context. SQuAD-SRC provides spoken questions for the entire training and development set of SQuAD v1.1, allowing a direct performance comparison with existing MRC models. We hire qualified English speakers from six different countries to read the questions in a quiet environment. The speaker selection and audio collection process are carefully designed. In total, we collect 98,169 manually verified high-quality utterances. Figure 1 shows an example from SQuAD-SRC. The SQuAD-SRC is available at https://github.com/tangyixuan/SQuAD-SRC.

Our dataset poses new challenges for the SRC models. First, there is a large semantic gap between the spoken question and textual context. The two semantic latent spaces need to be aligned to allow meaningful information interaction between question and context. Meanwhile, the length of frame-level features for audio signals is much longer than corresponding tokenized transcripts. Moreover, human voices contains more variations than synthetically generated sounds, including accents, tones, speaking speed and background noises, which contribute to the SRC task difficulties.

For the SRC task, an intuitive strategy is to cascade the ASR component with a machine comprehension model. The combination of a large-scale pre-trained ASR model and a transformer-based language model fine-tuned on MRC task provides a strong baseline. Note that the input is unimodal for each step in the pipeline. The important information embedded in the acoustic features is not exploited by the downstream MRC model. For instance, speakers may put sound emphasis on the essential words in the spoken question, which can provide guidance for the model to attend. In contrast, while end-to-end models are able to extract information from raw audio efficiently, they need to bridge the large modality gap between speech and language.

To better understand the properties of SQuAD-SRC, we perform statistical analysis of the audio utterances collected and conduct extensive experiments on representative MRC models. The main contributions of this paper are as follows:

- Unlike the previous studies, we present the problem of SRC with spoken questions and textual context, following the setting widely adopted in real-world applications. We construct an open-source, large-scale, naturally recorded, multi-accent SRC dataset SQuAD-SRC.
- We conduct extensive experiments with representative MRC models on SQuAD-SRC. The performance drops significantly when evaluated on audio questions, indicating the challenges posed by our dataset. Furthermore, the performance degrades as word error rate increases for questions with different accents.
- We investigate several strategies to remedy the errors introduced by ASR. Training on spoken data helps to improve both the overall performance and the robustness over different accents. In addition, We incorporate cross-modal representations to encode audio signals in an end-to-end approach, which serves as a preliminary exploration to inspire future end-to-end models.

2 SQuAD-SRC Dataset

We construct SQuAD-SRC based on the popular machine comprehension dataset SQuAD v1.1 [Rajpurkar *et al.*, 2016]. We follow the same data partition of SQuAD and use the 87.6k examples from their entire training set for our training set, and 10.6k examples from their development set for our test set, since their test set is not made publicly available.

2.1 Data Collection and Verification

The main objective of our work is to collect a natural SRC dataset with diverse accents. We carefully designed the procedure for speaker selection, data collection and verification. Figure 2 illustrates the whole pipeline.

We hire annotators from six different countries, including the United States (US), the United Kingdom (UK), China (CN), India (IN), Japan (JP) and Thailand (TH), with two males and two females from each country. To ensure the variety of accents, annotators are required to be native speakers of the language of the country they come from. For annotators from non-native English-speaking countries (CN, IN, JP, TH), they are required to satisfy all three requirements as follows to ensure the speaking quality: (1) have passed at least one language qualification test for English, such as CET,

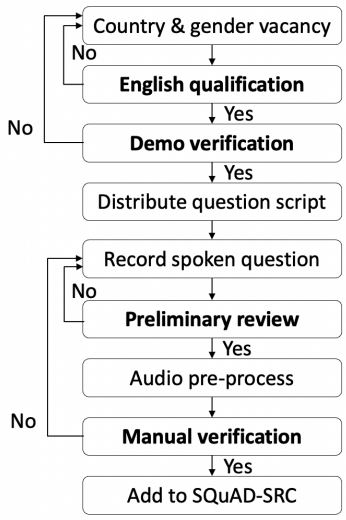


Figure 2: The overflow of data collection procedure.

TOEIC, TOEFL, (2) have been learning English for more than five years, (3) currently use oral English for education or work. The candidates satisfying all the qualifications are required to record a demo of five questions given for manual verification first. Only those with qualified demo audio proceed to record the official data. The candidate selection process is completed until all country and gender vacancies are occupied.

The questions in the training and development sets of SQuAD are randomly shuffled and divided into 24 portions, one for each annotator. Consequently, the training and test set of SQuAD-SRC follow similar accent distribution. The speakers are instructed to read the question fluently and record it in a quiet environment. The submission is done in batches of 50 utterances each. We conduct a round of preliminary review to check the general recording quality, including audio format, data completeness and background noise. All examples in an unqualified batch need to be re-recorded. We process the qualified audio utterances to remove the silence longer than 500ms at the beginning and the end of the audio. After that, all audio utterances are manually verified to match the question script given. Unqualified recordings are re-recorded until they pass the manual verification.

Following the common standard adopted by the speech community, we provide the audio for questions in raw waveform with a 16 kHz sampling rate and 16-bit encoding depth. For each utterance, we provide the question id from SQuAD, the anonymous speaker id (SPK001 to SPK024), and the gender and country information of the speaker. In total, we collect 129 hours of training data and 15 hours of test data.

2.2 Data Statistics

To further demonstrate the properties of SQuAD-SRC, we present detailed statistics about the audio duration, country and gender distributions and the word error rate (WER) of the spoken question recordings.

Country	US	UK	IN	TH	JP	CN
WER (%)	17.1	18.2	24.5	27.4	30.1	30.2

Table 2: Word error rate (WER) for questions recorded by annotators from different countries.

Audio Duration. Figure 3a shows the distribution percentage of audio question duration. The blue bins refer to the questions in the training set while the orange ones refer to those in the test set. Similar for both sets, the duration of audio recordings approximately follows a Gaussian distribution with a medium value close to 5 seconds. Over 99.9% of audio recordings are shorter than 15 seconds.

Country. Figure 3b shows the distribution of questions recorded by annotators from the different countries. SQuAD-SRC covers two native English-speaking countries, US and UK, and four non-native English-speaking countries, China, India, Japan and Thailand. The spoken questions are evenly distributed through the speakers from different countries for both the training and the test set.

Gender. Figure 3c shows the proportion of questions recorded by male speakers and female speakers. The gender distribution is fairly balanced for audio utterances provided in SQuAD-SRC.

Word Error Rate (WER). To evaluate the quality of the spoken questions collected, we run automatic speech recognition (ASR) on SQuAD-SRC using the wav2vec 2.0 large model [1] [Baevski *et al.*, 2020]. The WER on the whole dataset is 24.5. Table 2 presents WER for questions recorded by annotators from different countries in ascending order.

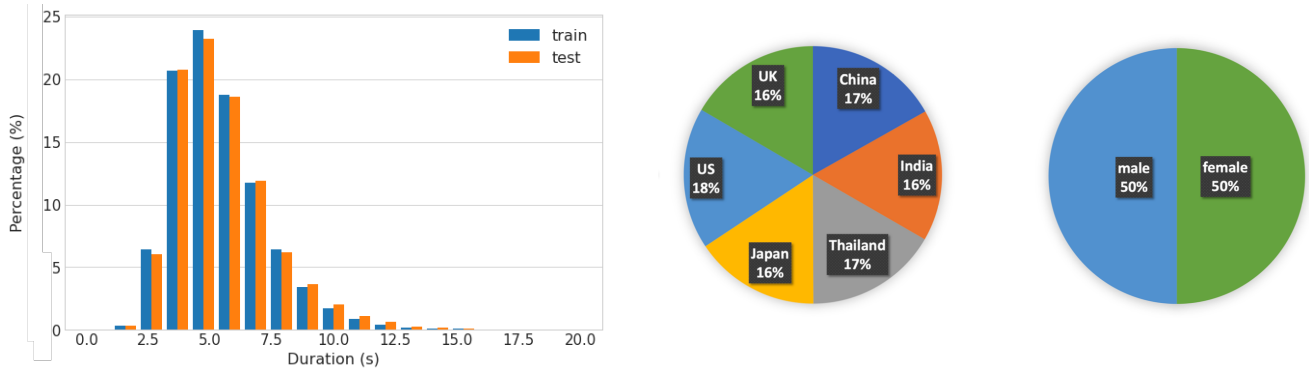
2.3 Advantages of SQuAD-SRC

As illustrated by Table 1, SQuAD-SRC advances existing SpokenQA datasets in the following aspects. First, compared to textual questions with spoken context, our setting of spoken questions with text context is more widely adopted in real-world applications, such as smart voice assistants. Second, to the best of our knowledge, SQuAD-SRC is the largest SpokenQA dataset naturally recorded by humans. The size of data is crucial to boost progress in deep learning fields. Meantime, we provide audios for the complete SQuAD training and development set, allowing SpokenQA models to directly compare performance with existing MRC models trained on the textual version of SQuAD. Third, all audios are recorded in a quiet and professional environment, resulting in high-quality utterances in SQuAD-SRC. Most importantly, the spoken questions are recorded by speakers from six different countries, which encourages accent diversity. It makes the dataset more challenging and helps to improve the robustness of the SRC model trained.

3 Method

We formulate the SRC problem as follows. Given a spoken question $\mathbf{q}^s$ and corresponding context $\mathbf{c}^t$, the goal is

[1] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self



(a) Duration distribution for spoken questions (b) Country distribution (c) Gender distribution

Figure 3: The statistics of our SQuAD-SRC dataset.

to find the answer $\mathbf{a}^t$, which is a text span of $\mathbf{c}^t$. We denote tokens for the textual context by $\mathbf{c}^t = \{c_1^t, c_2^t, ..., c_m^t\}$, where m is the number of tokens in $\mathbf{c}^t$. In other words, we aim to predict the start and end positions of the answer so that $\mathbf{a}^t = \{c_{y_1}^t, ..., c_{y_2}^t\}$, where y_1 and y_2 are integers and $1 \leq y_1 < y_2 \leq m$. For the rest of this paper, we use the superscript ^{t} to denote representations for input in text form, ^{s} for spoken form and ^{a} for ASR transcriptions.

3.1 Cascaded Model

As shown in Figure 4a, the cascaded model consists of an automatic speech recognition (ASR) component and a machine reading comprehension (MRC) model. Given a question $\mathbf{q}^s$ in the audio waveform, the ASR model first transcribes it into text script $\mathbf{q}^a = \{q_1^a, q_2^a, ..., q_l^a\}$, where l is the number of words in $\mathbf{q}^a$. Then the ASR transcription for the spoken question and its textual context $\mathbf{c}^t$ are fed into a MRC for answer span prediction.

We adopt state-of-the-art model wav2vec 2.0 [Baevski *et al.*, 2020] for ASR. For MRC, we select four models that achieve top performance on SQuAD, including two transformer-based models BERT [Devlin *et al.*, 2019] and ALBERT [Chi *et al.*, 2021], and two conventional MRC models, R-NET [Wang *et al.*, 2017] and BiDAF [Seo *et al.*, 2017]. Transformer-based models concatenate $\mathbf{q}^a$ and $\mathbf{c}^t$ into one sequence $\mathbf{t} = \{[\texttt{CLS}], q_1^a, ..., q_l^a, [\texttt{SEP}], c_1^t, ..., c_m^t, [\texttt{SEP}]\}$ and apply self-attention to allow information interaction among the words in $\mathbf{t}$. A linear layer with softmax is inserted on top to predict the start and end positions of the answer. In contrast, conventional MRC models encode the question $\mathbf{q}^a$ and context $\mathbf{c}^t$ into contextual representations $\mathbf{h}_q^a$ and $\mathbf{h}_c^t$ using separate bi-directional RNN layers. Then cross-attention is adopted on $\mathbf{h}_q^a$ and $\mathbf{h}_c^t$ to integrate information from the question and the context for answer prediction. The MRC models are fine-tuned on SQuAD official dataset and the ASR transcriptions individually for comparison.

3.2 End-to-End Model

Although large-scale pre-trained models are available for ASR and MRC, ASR error propagation and loss in useful acoustic features are inevitable for cascaded methods. Furthermore, both steps in the cascaded method need to handle pre-process and post-process separately, resulting in slow inference. To allow meaningful interaction between spoken questions and textual context, we explore an end-to-end model on SQuAD-SRC. Figure 4b shows the model architecture. It consists of a text encoder, an audio encoder, a bi-attention layer and a prediction layer. Our end-to-end SRC model takes in question and context with different modalities, while not requiring strict token-level speech-text alignment.

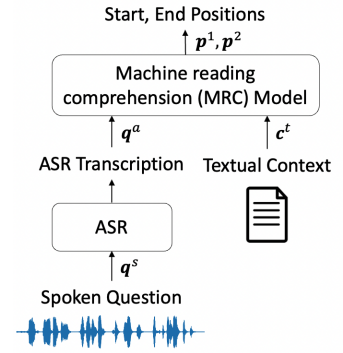
Text Encoder

We encode words in context $\mathbf{c}^t$ into embeddings $\mathbf{e}^t = \{\mathbf{e}_1^t, \mathbf{e}_2^t, ..., \mathbf{e}_m^t\}$ by concatenating Glove word embeddings [Pennington *et al.*, 2014] and character-level embeddings. Let d denote the hidden dimension, a bi-directional Long Short-Term Memory Network (bi-LSTM) [Graves *et al.*, 2005] layer is used to generate the contextual representations $\mathbf{v}^t \in \mathbb{R}^{m \times 2d}$ for the embedded context, i.e.

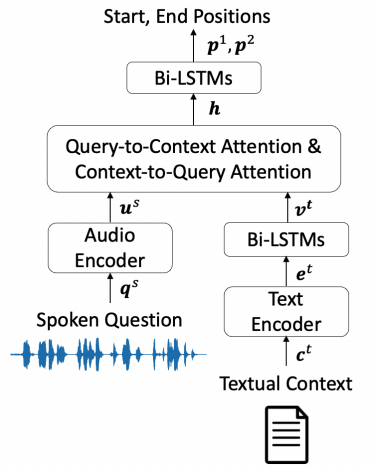
$$\mathbf{v}^t = bi{-}LSTM(\mathbf{e}^t) = \{\overrightarrow{LSTM}(\mathbf{e}_i^t) || \overleftarrow{LSTM}(\mathbf{e}_i^t)\}_{i=1}^m \quad (1)$$

Audio Encoder

A pre-trained audio representation model can extract informative acoustic features. However, it is impractical to directly interact such speech representations with the text representations due to the gap between the two semantic latent spaces. We adopt the unified cross-modal network SpeechT5 [Ao *et al.*, 2022] as our audio encoder. The feature extractor from wav2vec 2.0 is used as pre-net to process raw audio waves. The length of spoken question is downsampled by 320 times via seven convolutional layers. The encoder-decoder network of SpeechT5 is pre-trained with speech, text and speech-text joint tasks. A cross-modal vector quantification component is used to implicitly learn the alignment between acoustic and textual representations on unlabeled data. We use the audio pre-net and the pre-trained encoder of SpeechT5 to encode the spoken questions into hidden representations, i.e. $\mathbf{u}^s = g(f(\mathbf{q}^s)) = \{\mathbf{u}_1^s, \mathbf{u}_2^s, ..., \mathbf{u}_n^s\}$, where $f()$ refers to the pre-net and $g()$ refers to the encoder. We add a linear layer to transform the hidden dimension of $\mathbf{u}^s$ into $2d$ to match that of $\mathbf{v}^t$.



(a) Cascaded method



(b) End-to-end model

Figure 4: Model architecture for SRC.

Shared Semantic Interaction

We adapt BiDAF [Seo *et al.*, 2017] for shared semantic interaction in our end-to-end model. The question representation $\mathbf{u}^s$ and context representation $\mathbf{v}^t$ are fed into a bi-directional attention flow layer to fuse information between the acoustic features and textual features. First, we compute a similarity score matrix $S \in \mathbb{R}^{m \times n}$. The entry $S_{i,j} = \mathbf{w}_s[\mathbf{v}_i^t; \mathbf{u}_j^s; \mathbf{v}_i^t \odot \mathbf{u}_j^s]^\top \in \mathbb{R}$ measures the similarity between the ith context word and jth question frame, where $\mathbf{w}_s$ is a learnable weight matrix.

The context-to-query (C2Q) attention aggregates related question information for each token in context by $\mathbf{v}' = S_r\mathbf{u}^s \in \mathbb{R}^{m \times 2d}$, where S_r applies softmax on S by row. The query-to-context (Q2C) attention extracts context information most related to question by $\mathbf{u}' = S_c^\top \mathbf{v}^t \in \mathbb{R}^{1 \times 2d}$, where $S_c = softmax(max_{row}(S)) \in \mathbb{R}^{m \times 1}$ and max_{row} retrieves the maximum value for each row. We expand $\mathbf{v}'$ form $\mathbb{R}^{1 \times 2d}$ to $\mathbb{R}^{m \times 2d}$ by repeating the row m times.

To integrate all information, the query-aware context representation $\mathbf{h}$ is generated by $\mathbf{h} = [\mathbf{v}^t; \mathbf{v}'; \mathbf{v}^t \odot \mathbf{v}'; \mathbf{v}^t \odot \mathbf{u}']$. Finally, $\mathbf{h}$ is fed into two bi-LSTM layers and a feed-forward layer with softmax to predict the probability distribution of answer start and end positions. We fine-tune the proposed end-to-end model on SQuAD-SRC.

3.3 Training Objective

Similar to MRC, we minimize the cross-entropy loss between the ground truth start and end positions y_1, y_2, and the predicted distributions $\mathbf{p}^1$, $\mathbf{p}^2$ over all N examples from the training set:

$$L_{SRC} = -\frac{1}{N}\sum_{i=1}^{N} log(\mathbf{p}^1_{y_1}) + log(\mathbf{p}^2_{y_2}) \quad (2)$$

4 Experiments

In this section, we first introduce the evaluation metrics and the baseline MRC models used in the cascaded method. We analyze the impact of ASR error and different accents on the performances of cascaded models. Then we show the improvement made by training the model on ASR transcriptions of spoken questions. At last, we show the preliminary results achieved by our end-to-end model.

4.1 Evaluation Metrics

The answers of SQuAD-SRC are text spans, we follow the convention of machine comprehension and adopt the Exact Match (EM) and F1 scores between the ground-truth answers $\mathbf{a}^t$ and the predicted ones $\hat{\mathbf{a}}^t$ as the evaluation metrics. Punctuations and articles are removed in text spans before evaluation. The EM measures the percentage of examples in which $\hat{\mathbf{a}}^t$ and $\mathbf{a}^t$ are perfectly matched. Similar to human judgment, the F1 score measures the level of word overlap. For one example, it is calculated based on precision and recall with regard to the number of word tokens in $\mathbf{a}^t$ and $\hat{\mathbf{a}}^t$. The final score is averaged over all examples of SQuAD-SRC.

4.2 Baselines

We choose four representative MRC models below as baselines.

- **BERT** is a large-scale language model pre-trained with masked language modeling and next-sentence prediction tasks. When fine-tuned on MRC task, the question and context are concatenated into one sequence with a `[SEP]` token in between. The sequence is fed into self-attention layers and a linear layer is added on top to predict the start and end positions.
- **ALBERT** optimizes BERT by sharing parameter weights across self-attention layers and adding an inter-sentence coherence loss to the pre-training objective. The setting to fine-tune ALBERT on MRC remains the same as BERT.
- **R-NET** encodes the question and context using bi-GPU layers. The network contains a query-context matching and a context-context matching layer for information interaction. Then, a pointer network is adopted to locate the answer span.
- **BiDAF** encodes the context and question using Glove word embedding and character level embedding. It proposes a bidirectional attention flow mechanism to generate query-aware context representations. Then the context representations are fed into bi-LSTM layers for start and end position prediction.

Model	Text-dev		Audio-test		F1-drop
	EM	F1	EM	F1	
BERT	79.6	87.4	63.1	73.6	13.8
ALBERT	84.3	91.1	48.0	65.4	25.7
R-NET	70.4	79.3	54.2	64.6	14.7
BiDAF	64.0	75.0	49.2	61.3	13.7

Table 3: Experimental results on cascaded method with representative MRC models. The same ASR component wav2vec 2.0 is used for all models. The MRC models are trained on the SQuAD training set and evaluated on SQuAD dev set (Text-dev) and ASR transcriptions of SQuAD-SRC test set (Audio-test).

4.3 Results

Impact of ASR Errors and Accents

To validate the challenges brought by our dataset, we evaluate SQuAD-SRC against the cascaded method with representative MRC models trained on the official SQuAD training set. Table 3 presents the experimental results when they are tested against the official SQuAD development set (Text-dev) and SQuAD-SRC test set (Audio-test). Note the content of examples is the same for the two evaluation sets, while the only difference is the modality of questions. Compared to the performance on Text-dev, the performances of all models degrade significantly on Audio-test. The average drop in F1-scores is 17.0 for the four models, indicating the cascaded methods suffer from error accumulation problems.

Training on ASR Transcriptions

To remedy the drop in performance brought by spoken questions, we train the four MRC models in the cascaded method on ASR transcripts of SQuAD-SRC and compare the results with those trained on official SQuAD. As illustrated in Table 4, training with ASR transcripts achieves consistent performance gain across all models, resulting in an F1-raise of 10.2 on average. The difference in performance is most obvious for ALBERT.

To further investigate the impact of accents on SRC, we collect the performance of cascaded models on questions spoken by annotators from different countries. Table 4 shows the F1 scores with countries sorted in increasing WER order. The models achieve the best performance on questions spoken by annotators from US or UK, and their performance drops as the WER of questions increases. The last column F1-range refers to the performance variations on the questions spoken by annotators from different countries. It is calculated as the difference between the highest F1 score and the lowest F1 score among the six countries. Training on SQuAD-SRC improves the averaged F1-range from 11.4 to 6.2 on the four models, indicating that it helps to improve robustness over input with different accents.

End-to-End Model Performance

Table 5 demonstrates the performance results on our end-to-end model. To allow a direct comparison, we train a cascaded model by replacing the SpeechT5 audio encoder used in the end-to-end model with the SpeechT5 text encoder to encode question text. The remaining components of the model are unchanged. Given the modality gap between the question and the context, t is encouraging that the end-to-end model achieves performance comparable to the cascaded method. Our model can viewed as a starting point to inspire future end-to-end models on SRC tasks.

C: Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

Q_GT: What month, day and year did Super Bowl 50 take place?
Q_ASR: what month day and year did superbo fifty take place?
A: February 7, 2016
A_text: February 7, 2016 ✓
A_ASR: February 7, 2016 ✓
A_audio: February 7, 2016 ✓

Q_GT: Who won Super Bowl 50?
Q_ASR: who won super ball fifty?
A: Denver Broncos
A_text: The American Football Conference (AFC) champion Denver Broncos ✗
A_ASR: Denver Broncos ✓
A_audio: Carolina Panthers ✗

Q_GT: Super Bowl 50 determined the NFL champion for what season?
Q_ASR: superbo fifty determine the n f l champion forward season?
A: the 2015 season
A_text: Super Bowl 50 ✗
A_ASR: Super Bowl 50 ✗
A_audio: 2015 season ✓

Figure 5: Examples of answers predicted by different models. For each example, we present the context (C), the ground truth question (Q_GT), the ASR transcription of the spoken question (Q_ASR), the ground truth answer (A) and the predicted answers of three models built upon BiDAF. A_text, A_ASR and A_audio refer to the answers predicted by the cascaded models trained on SQuAD, ASR transcription of SQuAD-SRC and the end-to-end model trained on spoken questions of SQuAD-SRC correspondingly.

4.4 Qualitative Analysis

To further interpret the results, we present some correct and wrongly answered questions by different models for qualitative analysis. As demonstrated in Figure 5, it is easier for models to correctly locate the answer when the ASR is more similar to the ground truth text. Meanwhile, we observe that the end-to-end model sometimes can successfully locate the answer when the ASR transcriptions are too noisy for the cascaded methods.

5 Related Work

SpokenQA Benchmarks

Previous SpokenQA datasets mainly focus on posing textual questions on audio materials. CLEAR [Abdelnour *et al.*, 2018] and DAQA [Fayek and Johnson, 2020] construct audio scenes containing elementary sounds and generate textual QA pairs about the scenes via algorithms. Huang *et al.* [2021] collect 2,200 listening comprehension MCQs from human English tests to form the SpokenQA dataset. Clotho-AQA [Lipping *et al.*, 2022] provides 1,991 audio files, each with

Model	Text-dev			Country of speaker						
	EM	F1	F1-raise	US	UK	IN	TH	JP	CN	F1-range
Cascaded method with MRC models trained on SQuAD										
BERT	63.1	73.6	-	79.4	79.2	71.4	72.4	67.6	71.1	11.8
ALBERT	48.0	65.4	-	72.1	72.5	64.2	63.1	59.4	60.4	13.1
R-NET	54.2	64.6	-	68.5	70.1	63.3	62.5	60.1	62.5	10.0
BiDAF	49.2	61.3	-	64.9	67.4	59.6	59.7	56.9	58.6	10.5
Cascaded method with MRC models trained on SQuAD-SRC										
BERT	72.5	80.8	7.2	83.1	83.5	81.2	79.1	77.8	79.8	5.7
ALBERT	78.1	85.7	20.3	88.5	88.6	85.3	83.7	82.1	85.3	6.5
R-NET	62.7	71.6	7	73.8	75.1	71.9	69.1	68.8	70.8	6.3
BiDAF	56.2	67.4	6.1	69.0	70.9	68.0	64.7	64.6	66.9	6.3

Table 4: Comparison of experimental results on cascaded method with representative MRC models trained separately on the SQuAD training set and ASR transcriptions of SQuAD-SRC and evaluated on SQuAD-SRC test set. F1-raise refers to the raise in F1 score brought by comparing models trained on SQuAD-SRC with those trained on SQuAD. We also present the F1 scores for spoken questions recorded by annotators from different countries. The column F1-range refers to the performance difference between the countries with the highest and lowest F1 scores.

Model	Training data	EM	F1
cascaded	textual question	42.8	53.2
end-to-end	spoken question	42.6	52.2

Table 5: Comparison of results on the end-to-end model and cascaded model with the same architecture. The end-to-end model is trained with spoken questions in SQuAD-SRC while the cascaded model is trained with textual questions in SQuAD.

six cloud-sourced textual QA pairs. It can be formulated as a multi-class classification task thanks to single-word answers. Spoken SQuAD [Li *et al.*, 2018] is built upon the same QA dataset as ours. Instead of recording by humans, it generates spoken context using TTS. Due to the removal of QA pairs whose answers are not contained in the ASR transcriptions of the context, it provides an incomplete spoken version of the SQuAD dataset, including 37,111 examples (42%) for training and 5,351 (50%) examples for test.

For the SpokenQA datasets that provide audio for questions, Spoken-CoQA [You *et al.*, 2022] contains 40k QA pairs with spoken question and context generated by TTS from conversational QA dataset CoQA. ODSQA [Lee *et al.*, 2018] contains 3,654 QA pairs with spoken questions and context from a Chinese extraction-based MRC dataset.

Table 1 summarizes important properties of existing SpokenQA datasets. They either are limited in size or use synthetically generated audios, which cannot reflect the real-world need to handle complex human voices.

SpokenQA Models

A common solution for the SpokenQA task is to cascade an ASR model with a textual QA model. Previous works propose to improve the cascaded method from different perspectives, such as using multi-set pre-trainig [Su and Fung, 2020], self-supervised learning [Chen *et al.*, 2021], knowledge distillation [You *et al.*, 2021a; You *et al.*, 2022], and enhancing with aligned acoustic features [Kuo *et al.*, 2021; Kuo *et al.*, 2020; You *et al.*, 2021b]. Cascaded methods suffer from ASR error accumulation. Meantime, acoustic features useful for the downstream QA task is not fully utilized.

For end-to-end models, DIIA [Huang *et al.*, 2021] integrates information from ground truth textual and spoken context to predict answers. However, ground truth text is not provided in common SpokenQA setting. Lipping *et al.* [2022] directly concatenate the audio encoding for context and text encoding for the question to classify the answer. They show model that takes in only the textual question performs as good as the model that takes both inputs, indicating audio information is not utilized. SpeechBERT [Chuang *et al.*, 2020] tries to bridge the gap by jointly pre-training a BERT-based model on both audio and text data. DUAL [Lin *et al.*, 2022] proposes to perform the SpokenQA task with audio representations only. Both models achieve lower performance than the corresponding cascade models. It remains challenging for end-to-end models to fully utilize the audio information.

6 Conclusion and Future Work

In this paper, we reformulate the SpokenQA task with spoken questions and textual context. We introduce a large-scale, multi-accent SRC dataset recorded by humans. SQuAD-SRC challenges models' abilities to handle the speech-text semantic gap and robustness over speech variations. Extensive experiments show degraded performance on top MRC models, especially on questions with higher WER. We investigate strategies to address the issue. In the end-to-end method, acoustic features provide extra information for the downstream reading comprehension task. Meanwhile, it contains irrelevant features such as accents, genders, speaking speed and environment noises. How to disentangle semantic information embedded in spoken questions with irrelevant noises remains challenging and will be addressed in future work.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[Abdelnour *et al.*, 2018] Jérôme Abdelnour, Giampiero Salvi, and Jean Rouat. CLEAR: A dataset for compositional language and elementary acoustic reasoning. *CoRR*, abs/1811.10561, 2018.

[Ao *et al.*, 2022] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *ACL (1)*, pages 5723–5738. Association for Computational Linguistics, 2022.

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020.

[Chen *et al.*, 2021] Nuo Chen, Chenyu You, and Yuexian Zou. Self-supervised dialogue learning for spoken conversational question answering. In *Interspeech*, pages 231–235. ISCA, 2021.

[Chi *et al.*, 2021] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. Audio albert: A lite bert for self-supervised learning of audio representation. In *SLT*, pages 344–350. IEEE, 2021.

[Chuang *et al.*, 2020] Yung-Sung Chuang, Chi-Liang Liu, Hung-yi Lee, and Lin-Shan Lee. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. In *INTERSPEECH*, pages 4168–4172. ISCA, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics, 2019.

[Fayek and Johnson, 2020] Haytham M. Fayek and Justin Johnson. Temporal reasoning via audio question answering. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2283–2294, 2020.

[Goyal *et al.*, 2019] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127(4):398–414, 2019.

[Graves *et al.*, 2005] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional LSTM networks for improved phoneme classification and recognition. In *ICANN (2)*, volume 3697 of *Lecture Notes in Computer Science*, pages 799–804. Springer, 2005.

[Hoy, 2018] Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[Huang *et al.*, 2021] Zhiqi Huang, Fenglin Liu, Xian Wu, Shen Ge, Helin Wang, Wei Fan, and Yuexian Zou. Audio-oriented multimodal machine comprehension via dynamic inter- and intra-modality attention. In *AAAI*, pages 13098–13106, 2021.

[Hudson and Manning, 2019] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019.

[Kuo *et al.*, 2020] Chia-Chih Kuo, Shang-Bao Luo, and Kuan-Yu Chen. An audio-enriched bert-based framework for spoken multiple-choice question answering. In *INTERSPEECH*, pages 4173–4177. ISCA, 2020.

[Kuo *et al.*, 2021] Chia-Chih Kuo, Kuan-Yu Chen, and Shang-Bao Luo. Audio-aware spoken multiple-choice question answering with pre-trained language models. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3170–3179, 2021.

[Lee *et al.*, 2018] Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-yi Lee. ODSQA: open-domain spoken question answering dataset. In *SLT*, pages 949–956, 2018.

[Li *et al.*, 2018] Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. In *INTERSPEECH*, pages 3459–3463, 2018.

[Lin *et al.*, 2022] Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu-Wen Yang, Hsuan-Jui Chen, Shuyan Annie Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, and Lin-Shan Lee. DUAL: discrete spoken unit adaptive learning for textless spoken question answering. In *INTERSPEECH*, pages 5165–5169. ISCA, 2022.

[Lipping *et al.*, 2022] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering. In *EUSIPCO*, pages 1140–1144. IEEE, 2022.

[Nguyen *et al.*, 2016] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, volume 1773 of *CEUR Workshop Proceedings*, 2016.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL, 2014.

[Rajpurkar *et al.*, 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392, 2016.

[Raux *et al.*, 2005] Antoine Raux, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskénazi. Let's go public! taking a spoken dialog system to the real world. In *INTERSPEECH*, pages 885–888. ISCA, 2005.

[Seo *et al.*, 2017] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *ICLR (Poster)*. OpenReview.net, 2017.

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

[Trischler *et al.*, 2017] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Rep4NLP@ACL*, pages 191–200, 2017.

[Wang *et al.*, 2017] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *ACL*, pages 189–198. Association for Computational Linguistics, 2017.

[Yang *et al.*, 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pages 2369–2380, 2018.

[You *et al.*, 2021a] Chenyu You, Nuo Chen, and Yuexian Zou. Mrd-net: Multi-modal residual knowledge distillation for spoken question answering. In *IJCAI*, pages 3985–3991. ijcai.org, 2021.

[You *et al.*, 2021b] Chenyu You, Nuo Chen, and Yuexian Zou. Self-supervised contrastive cross-modality representation learning for spoken question answering. In *EMNLP (Findings)*, pages 28–39. Association for Computational Linguistics, 2021.

[You *et al.*, 2022] Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou. End-to-end spoken conversational question answering: Task, dataset and model. In *NAACL-HLT (Findings)*, pages 1219–1232, 2022.