MRD-Net: Multi-Modal Residual Knowledge Distillation for Spoken Question Answering

Chenyu You\textsuperscript{1*}, Nuo Chen\textsuperscript{2*}, Yuexian Zou\textsuperscript{2,3†}

\textsuperscript{1}Department of Electrical Engineering, Yale University, USA
\textsuperscript{2}ADSPLAB, School of ECE, Peking University, Shenzhen, China
\textsuperscript{3}Peng Cheng Laboratory, Shenzhen, China
chenyu.you@yale.edu, \{nuochen,zouyx\}@pku.edu.cn

Abstract

Spoken question answering (SQA) has recently drawn considerable attention in the speech community. It requires systems to find correct answers from the given spoken passages simultaneously. The common SQA systems consist of the automatic speech recognition (ASR) module and text-based question answering module. However, previous methods suffer from severe performance degradation due to ASR errors. To alleviate this problem, this work proposes a novel multi-modal residual knowledge distillation method (MRD-Net), which further distills knowledge at the acoustic level from the audio-assistant (Audio-A). Specifically, we utilize the teacher (T) trained on manual transcriptions to guide the training of the student (S) on ASR transcriptions. We also show that introducing an Audio-A helps this procedure by learning residual errors between T and S. Moreover, we propose a simple yet effective attention mechanism to adaptively leverage audio-text features as the new deep attention knowledge to boost the network performance. Extensive experiments demonstrate that the proposed MRD-Net achieves superior results compared with state-of-the-art methods on three spoken question answering benchmark datasets.

1 Introduction

In recent years, spoken question answering (SQA) has received remarkable attention by researchers. The goal of SQA is to fully understand the spoken content of the passage and questions, and then provide an accurate language answer. Commonly, this challenging task includes two sub-tasks: automatic speech recognition (ASR) and text question answering (TQA). In general, the SQA system first utilizes the ASR module to transfer spoken content to sequential utterances in text form, and then employs the TQA module to tackle the auto-transcribed text documents. In this way, taking ASR transcriptions as input brings noisy signals (e.g., substitution error), which misleads the system to make wrong predictions. In other words, this leads to a suboptimal performance of spoken question answering.

Several studies have been proposed to alleviate automatic speech recognition errors [Li et al., 2018; Lee et al., 2018; Lee et al., 2019; You et al., 2020b; Kuo et al., 2020], Li et al. [2018] and Lee et al. [2018] utilized sub-word units to generate the auto-transcriptions to mitigate the impact of speech recognition errors. However, such methods bring the SQA system limited performance gains. More recently, Lee et al. [2019] learned the domain-invariant representations by projecting audio and text features into the common latent space for the development of SQA systems. However, these methods turn out to be susceptible to training instability. Subsequently, You et al. [2021] demonstrated that knowledge distillation [Hinton et al., 2015] (KD) proved to be a promising way to achieve superior performance, in which employs a soft-label distillation loss to transfer the knowledge of T trained on the clean text (manual transcriptions) to S trained on the ASR transcriptions. Nevertheless, they only focus on resolving the challenge in a textual way without considering the acoustic-level information to further improve the model performance.

The machine learning research community has devoted substantial efforts to leverage various types of knowledge from multiple domains to improve the model performance. Previous work [Su and Fung, 2020] has shown that spoken content can provide additional cues for remarkable performance improvements. Moreover, Siriwardhana et al. [2020] adopted the pre-trained “BERT-like” language model in the self-supervised manner to fuse both speech and text features to tackle the multi-modal speech emotion recognition task. Concurrent with our research, Kuo et al. [2020] built an audio-enriched BERT-based (aeBERT) framework for the spoken multiple-choice question answering task. However, most existing methods are not unified approach to address spoken language tasks, which is impractical in real-world applications.

In this paper, we propose a deep multi-modal residual knowledge distillation framework for SQA tasks, called MRD-Net. Specifically, multi-domain features are treated as knowledge in this work. In our teacher-student paradigm, we first train the teacher model (T) on text transcriptions, and then the student model (S) is trained on ASR transcriptions with the goal of achieving the comparable performance. To

\textsuperscript{*}Equal contribution.

\textsuperscript{†}Contact Author
further enhance the process of knowledge transfer, we propose an audio-assistant model (Audio-A) to learn the residual error between the hidden state features of $T$ and $S$ (See Figure 1). Furthermore, we introduce a novel attention (ST-Attention) mechanism by utilizing the scaled dot-product between multi-modal key-value and query vector pairs in the attention module as the audio-text knowledge to boost the performance (See Figure 2). With such knowledge transferring scheme, $S$ can achieve the ideal mimicking on $T$ efficiently. We validate the proposed method on three benchmark datasets, and experimentally demonstrate that our proposed method achieves superior performance compared with state-of-the-art methods.

The contributions of our work are summarized as follows:
- We propose a novel distillation model, MRD-Net, for the spoken question answering task, by introducing the multi-modal residual knowledge distillation (MRKD) strategy to alleviate the ASR errors. Moreover, we propose an Audio-A to learn residual error between the teacher and student models, which helps significantly improve model performance.
- We build a novel attention mechanism, termed as ST-Attention, to aid multi-modal knowledge transfer process with diverse-grained representations. Moreover, we further validate the proposed attention mechanism on other downstream multi-modal language processing tasks, and show its generalization ability.
- We evaluate the proposed MRD-Net on three commonly used speech question answering datasets, and prove that MRD-Net significantly outperforms state-of-the-art methods.

2 Related Work

2.1 Spoken Question Answering

Spoken question answering aims at enabling machine learning systems to automatically find the relevant answers from the given spoken documents. Typically, the spoken question answering system includes two main modules: automatic speech recognition and text question answering (TQA). Specifically, the ASR module first converts spoken content into noisy transcriptions, and then the noisy transcriptions are treated as input to TQA module to find the answers in given spoken transcriptions by leveraging robust information retrieval (IR) techniques. But existing work [Su and Fung, 2020] has proved that noisy ASR errors would significantly degrade the SQA performance. Much effort has been devoted to alleviate this issue. Utilizing sub-word unit in SQA [Li et al., 2018; Lee et al., 2018] achieved competitive performance in SQA tasks by mitigating the impact of speech recognition errors. The key idea of such methods is that ASR errors can be viewed as substitutions of word sequences for another ones (e.g., “feature” to “feather”). Therefore, the sub-word information can still be transferred correctly while the common error was missing or wrongly added one or two word transcriptions in ASR systems. A very recent work, called SpeechBERT [Chuang et al., 2020], introduced a pre-trained BERT-based language model for SQA tasks to jointly learn audio-text features for significant accuracy performance improvements.

2.2 Knowledge Distillation

In KD scheme [Hinton et al., 2015], the teacher model $T(\cdot)$ is to transfer richer knowledge to the student model $S(\cdot)$. In other words, the student network is trained with the purpose of fully reproducing the predictive behavior of the teacher network. Suppose $f_T$ and $f_S$ denote the behavior functions of $T$ and $S$, respectively, behavior functions aim to transform network inputs into informative feature representations, and distill the knowledge from $T$ to $S$. To further promote the distillation process, many attempts [Hahn and Choi, 2019; Gao et al., 2020] have been proposed. Gao et al. [2020] proposed residual knowledge distillation (RKD) strategy by adopting a residual learning strategy into the standard KD framework. Especially, RKD aims to supervise $S$ to acquire knowledge from $T$, and then employ an assistant model $A$ to learn the residual difference between $S$ and $T$, such that $A$ can further ease the process of knowledge transfer. Although our study shares a similar topic, our research method differs. Our experimental results show significant performance improvements by enabling the Audio-assistant model to capture the multi-modal residual knowledge to refine the learned feature representations. Most importantly, we focus on the question of utilizing multiple domain knowledge for network performance improvements.

3 Method

In this section, we detail a novel distillation method for the spoken question answering task. First, we present the design decisions for the proposed BERT-based network. We then present the proposed multi-modal residual knowledge distillation (MRKD) algorithm (See Figure 1). Next, we introduce the proposed ST-Attention mechanism (See Figure 2), and perform prediction-layer distillation to guide the learning framework throughout the training process. Finally, we describe how to incorporate the progressive learning in the training procedure in Section 3.5.

3.1 Problem Formulation

The backbone of the proposed MRD-Net is similar to BERT [Devlin et al., 2019]. We follow the spoken question answering convention in [Lee et al., 2018; Kuo et al., 2020] that set the ASR token sequences of passages and corresponding questions as input to the following module. Specifically, a special token [CLS] is added to the beginning of the input sequences, and a special boundary token [SEP] is applied to separate token sequences. Given a passage $P = \{p_1, p_2, ..., p_n\}$, and the corresponding question $Q = \{q_1, q_2, ..., q_m\}$, these two sequences are packed together into $\{[CLS], q_1, q_2, ..., q_m, [SEP], p_1, p_2, ..., p_n, [SEP]\}$. Then the pre-trained BERT is employed to extract hidden state features from the given token sequences. A task-specific layer finally predicts the possibility of each token. Although directly adopting the pre-trained BERT language model has achieved good performance, there exist noisy ASR errors in the transcriptions, which significantly lead to the degradation of SQA performance. Thus,
the key research problem becomes how to effectively alleviate the ASR errors and robust behavior functions. Different from previous natural language processing (NLP) models, we consider how to leverage spoken features to achieve better performance.

3.2 Multi-Modal Residual Knowledge Distillation

Knowledge Distillation. MRD-Net uses a pretrained BERT as the backbone of teacher and student models, consisting of 12 Transformer layers with the word embedding dimension of 768. Specifically, \( T \) is trained on manual transcriptions, and \( S \) on ASR transcriptions, including highly noisy ASR errors. Thus, there exists a substantial gap between the learning capacities of \( T \) and \( S \).

Audio-Assistant. To narrow down the performance gap between the representation abilities of \( T \) and \( S \), we use a BERT-like network trained on discretized speech tokens as the Audio-A. Concretely, a pre-trained VQ-Wav2Vec [Schneider et al., 2019] is trained on Librispeech-960 [Panayotov et al., 2015] to encode speech signals to a sequence of input tokens. Audio-A then utilizes contextual information from speech token sequences to aid the \( S \) in the mimicking process by learning the residual error between the feature maps of \( S \) and \( T \). It is worth noting that Audio-A shares the similar structure with \( S \) and \( T \) (See Section 4.2).

Intuitively, we perform MRKD to encourage \( S \) to fully acquire knowledge from \( T \) in the distillation process. More concretely, in our proposed MRKD setting, \( S \) targets to mimic the hidden state representations of \( T \) to achieve comparable performance. Since there exists a huge gap of the learning capacities between \( T \) and \( S \), we introduce Audio-A to facilitate the knowledge transfer process.

In MRKD, each model consists of the \( L \)-layer Transformer model, as shown in Figure 1. Let \( \mathbf{h}_T^A \), \( \mathbf{h}_T^T \), and \( \mathbf{h}_S^S \) denote the hidden state features of Audio-A, \( T \) and \( S \), respectively. Then, we train Audio-A by minimizing the following loss function:

\[
L_A = \frac{((\mathbf{h}_T^T - \mathbf{h}_S^S)(\mathbf{h}_T^A)^T)}{||\mathbf{h}_T^T - \mathbf{h}_S^S||_2 ||\mathbf{h}_T^A||_2},
\]

where the size of hidden state features \( \mathbf{h}_T^A \), \( \mathbf{h}_T^T \), and \( \mathbf{h}_S^S \) are \( l_1 \times d_1 \), \( l_2 \times d_2 \), and \( l_2 \times d_2 \), respectively. \( l_1 \) and \( l_2 \) refer to the length of input speech word tokens and text word tokens, and \( d_1 \) and \( d_2 \) are the word embedding dimensions, respectively. Then, we compute residual-error-involved feature representations by summing the feature maps with the residual error with \( \mathbf{h}_L = \mathbf{h}_T^T + \mathbf{h}_T^A \), will be used as the input of the following St-Attention.

3.3 ST-Attention

To achieve more accurate distillation process, we propose a novel ST-Attention mechanism by fusing audio features and textual features (See Figure 2). Let \( \mathbf{h}_L \) and \( s_L \) represent feature representations in both text and spoken forms. The calculation of ST-Attention mechanism depends on the three components of key, query and value matrices with Transformer layers. Different from self-attention mechanism, we first set key-value pairs from textual features and queries from acoustic information as input. The attention head computes attention maps \( a_L \) in the speech domain with additional text context information. Then, we choose \( a_L \) as key-value pairs with \( h_L \) as queries into the following module to generate audio-
Multi-Head Attention
Add & Norm
FFN
Add & Norm
QKV QKV
Add & Norm
FFN
Adapter
Adapter
FFN
Adapter
Figure 2: Overview of ST-Attention mechanism.

In addition to mimicking the behavior of the L-th layer, we also utilize the standard knowledge distillation [Hinton et al., 2015]. Specifically, we compute the soft cross-entropy loss to guide S to learn from the softened output of T. More concretely, \( w^T_L \) and \( w^S_L \) denote the logits vectors generated by T and S, and \( y \) represents the corresponding sequence of ground-truth. We perform prediction layer distillation, and the objective is defined as:

\[
L_S = \alpha L_{NLL}(p_T(w^S_L), p_T(w^T_L)) + (1 - \alpha) L_{CE}(w^S_L, y),
\]

(2)

where \( L_{NLL} \) and \( L_{CE} \) denote the negative log-likelihood loss and cross-entropy loss, respectively. \( p_T() \) refers to the softmax function with temperature \( \tau \). \( \alpha \) is a hyper-parameter.

In the training phase, to achieve better performance, the knowledge transfer process consists of two steps. S first distills knowledge from the softened output of T, and then Audio-A is introduced to help refine the feature representations by capturing the underlying knowledge to improve the performance significantly.

### 3.5 MRKD with Progressive Learning

Inspired by the recent success [Wang et al., 2018], we incorporate progressive learning into MRKD to enable more effective knowledge transfer. In this fashion, S is able to capture both high-level and low-level information from T, leading to significant performance gains. Concretely, we perform MRKD with each block to advance the process of the knowledge transfer. For example, in the first layer, \( S_1 \) attempts to mimic the feature map of \( T_1 \), and then Audio-A1 learns the residual error given the intermediate feature maps. After that, residual-error-involved features are utilized as the input for the following training. Following this learning strategy, both S is able to learn richer knowledge from T at multiple levels. In this way, S can be optimized at one time with

\[
L_A = \sum_{i=1}^{L} \frac{(h_T^i - h_S^i)(h_A^i)^T}{||h_T^i - h_S^i||_2 ||h_A^i||_2}.
\]

(3)

### 4 Experiments

#### 4.1 Datasets

In this section, we evaluate the effectiveness of the proposed method on a variety of speech question answering tasks.

**Spoken-SQuAD** Spoken-SQuAD [Li et al., 2018] is an English listening comprehension dataset, which contains 37k ASR transcripts question pairs in the training set and 5.4k in the testing set, respectively. In Spoken-SQuAD, the spoken documents are in spoken form, and the questions and answers are in the text form. Specially, the word error rate (WER) is 22.77% on the training set, and 22.73% on the testing set, respectively. The manual written documents of Spoken-SQuAD are from the original SQuAD dataset [Rajpurkar et al., 2016], which is one of the most popular machine reading comprehension benchmark datasets.

**FGC** 2018 Formosa Grand Challenge (FGC) dataset \(^1\) is a Mandarin Chinese spoken multi-choice question answering (MCQA) dataset, which includes 7k passage-question-choice (PQC) pairs as the training set and 1.5k as the development set, respectively. Each PQC pair consists of a passage, a question, and four corresponding answers, in that only one choice is the correct answer. In FGC dataset, all passages, questions, and multiple choices are in spoken form. Following the standard setting in [Kuo et al., 2020], we utilize the Kaldi toolkit to build up our ASR system where the WER is about 20.4%.

**Spoken-CoQA** Spoken-CoQA [You et al., 2020a] is an English spoken conversational question answering (SCQA) dataset, which consists of 40k and 3.8k question-answer pairs from 4k conversations in the training set and 380 conversations in test set from seven diverse domains, respectively. The WER is 18.7%. In Spoken-CoQA, questions and passages are both in text and spoken form, and answers are in the text form.

\(^1\)https://fgc.stpi.narl.org.tw/activity/techai2018

---

**Table 1: Performance of BERT model trained on different types of transcriptions on Spoken-CoQA.**

<table>
<thead>
<tr>
<th>Data Usage</th>
<th>Experimental Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Dev &amp; Test</td>
</tr>
<tr>
<td>Manual</td>
<td>Manual</td>
</tr>
<tr>
<td>ASR</td>
<td>ASR</td>
</tr>
<tr>
<td>ASR</td>
<td>Manual</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual Manual</td>
<td>67.73</td>
<td>77.71</td>
</tr>
<tr>
<td>Manual ASR</td>
<td>41.79</td>
<td>54.70</td>
</tr>
<tr>
<td>ASR ASR</td>
<td>40.58</td>
<td>54.12</td>
</tr>
<tr>
<td>ASR Manual</td>
<td>42.53</td>
<td>54.81</td>
</tr>
</tbody>
</table>
Wav2Vec as tokenizer. The maximum sequence lengths of \( T \) and \( S \) are 512, and the Audio-A is 1024. We utilize AdamW optimizer in training, and the learning rate is set to 8e-6. All models are trained using 4 as the batch size. The hyperparameter \( \tau \) and \( \alpha \) are set to 1 and 0.9, respectively. Specially, when training MRD-Net on Spoken-CoQA dataset, we utilize conversation history via adding the last question with \( \tau \) and \( \alpha \), respectively. For brevity, \( pl \) distillation denotes prediction layer distillation. \( MSE \) denotes using mean-squared-error loss function instead of Equation 1.

![Table 2: Comparison results between our proposed model with other methods. Progressive MRD-Net is MRD-Net with progressive learning.](image)

### 4.2 Experimental Settings

In this study, we use BPE as the tokenizer to generate token sequences as input for both \( T \) and \( S \), and Audio-A adopts VQ-Wav2Vec as tokenizer. The maximum sequence lengths of \( T \) and \( S \) are 512, and the Audio-A is 1024. We utilize AdamW optimizer in training, and the learning rate is set to 8e-6. All models are trained using 4 as the batch size. The hyperparameter \( \tau \) and \( \alpha \) are set to 1 and 0.9, respectively. Specially, when training MRD-Net on Spoken-CoQA dataset, we utilize conversation history via adding the last question with \( \tau \) and \( \alpha \), respectively. For brevity, \( pl \) distillation denotes prediction layer distillation. \( MSE \) denotes using mean-squared-error loss function instead of Equation 1.

### 4.3 Results

To demonstrate the performance degradation caused by ASR errors, we first evaluate the standard BERT on Spoken-CoQA dataset, which only set text document as inputs for answer predictions. In Table 1, we find that using standard BERT to train and to make inferences on the manual transcriptions achieves 67.73%/77.11% on EM/F1 scores, which outperforms that in other settings. This indicates that speech recognition errors lead to poor performances of the SQA systems, confirming the similar patterns in [Lee et al., 2019].

We also demonstrate the effectiveness of our proposed method. The results on Spoken-SQuAD, Spoken-CoQA, and FGC datasets are reported in Table 2. From Table 2, we can draw two following conclusions. i) The proposed method significantly outperforms all other evaluated methods. For example, on FGC dataset, aeBERT is the best language model among all the 7 evaluated methods. Compared with aeBERT, our proposed model achieves 81.91% (+ 3.71% improvement) in terms of accuracy. Moreover, MRD-Net is capable of achieving superior performance on Spoken-SQuAD and Spoken-CoQA datasets. The results suggest that our proposed method is capable of improving the performance in a variety of language tasks. ii) Here we introduce the progressive learning strategy to MRD-Net (Progressive MRD-Net). Table 2 shows that Progressive MRD-Net can perform better than that without using progressive learning, improving the results by 0.81%/0.96% and 0.98% on Spoken-SQuAD, 1.51%/1.21% and 0.66% on Spoken-CoQA and 0.82% on FGC, in terms of EM, F1, and AOS scores, respectively. This suggests that some missing low-level feature representations embedded in both text corpora and speech documents can help with the performance improvements.

### 5 Ablation Study

In this section, we conduct ablation studies to analyze the following contributions of: i) validating the effectiveness of different components, ii) examining different distillation objectives, iii) investigating how ST-Attention results in performance gain, iv) exploring how the hyperparameter selection affects network performance. Note that here we utilize the pre-trained BERT as the backbone for \( T \) and \( S \) in this study.

**Different Components of MRD-Net**. We report results in Table 3. We can clearly see that removing either perdition layer distillation or Audio-A leads to a significant performance drop over three benchmark datasets. This confirms the
We investigate the effectiveness of ST-Attention mechanism. As shown in Table 5, using ST-Attention achieves superior performance compared with other methods. Furthermore, we can see that using Adapters results in small performance degradation. This demonstrates that Adapters can yield performance gains by filtering out useless knowledge. To illustrate the generality and robustness of the proposed ST-Attention mechanism, we further validate ST-Attention on the multi-modal emotion recognition task. For a fair comparison, we adopt various multi-modal fusion mechanisms into the BERT-liked model [Sirivardhana et al., 2020]. Following the setting in [Sun et al., 2020], we choose four most commonly used emotion categories, including Happy, Sad, Angry, and Neutral on IEMOCAP dataset [Busso et al., 2008]. In parallel, we use Binary Accuracy and F1-Score for evaluation. As shown in Table 6, we observe that our method achieves better performance than other evaluated methods among all emotion categories.

Effects of Temperature $\tau$. We present the model performance with different temperature $\tau$ in Figure 3. $\tau$ is chosen from $\{1, 2, 4, 6, 8\}$. From Figure 3, MRD-Net achieves superior results when $\tau$ is set to 1.

6 Conclusion

We propose a novel speech question answering framework, called MRD-Net, which further distills knowledge from T to alleviate ASR errors. Specially, we introduce audio-A to aid knowledge transfer process for performance improvements. In addition, we introduce ST-Attention mechanism by incorporating audio-text representations to further distill multi-domain knowledge to bring further performance improvements. Extensive experiments demonstrate the effectiveness of the proposed method on a variety of tasks. In future work, we will explore new training strategies by only leveraging acoustic features to achieve comparable performance.
References


