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

Chenyu You[1*], **Nuo Chen**[2*], **Yuexian Zou**[2,3†]
[1]Department of Electrical Engineering, Yale University, USA
[2]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[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

[*]Equal contribution.
[†]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,

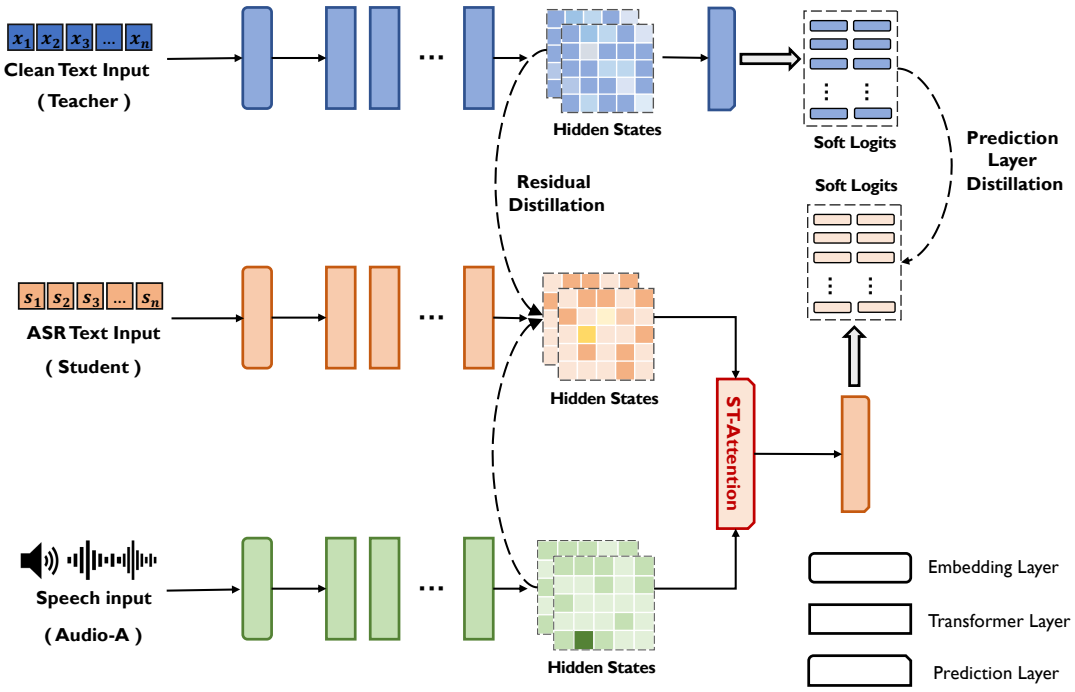

Figure 1: Overview of the proposed MRD-Net.

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 $\mathbf{h}_L^A$, $\mathbf{h}_L^T$, and $\mathbf{h}_L^S$ denote the hidden state features of Audio-*A*, *T* and *S*, respectively. Then, we train Audio-*A* by minimizing the following loss function:

$$\mathcal{L}_{\mathcal{A}} = \frac{(\mathbf{h}_L^T - \mathbf{h}_L^S)(\mathbf{h}_L^A)^T}{\left\|\mathbf{h}_L^T - \mathbf{h}_L^S\right\|_2 \left\|\mathbf{h}_L^A\right\|_2}, \tag{1}$$

where the size of hidden state features $\mathbf{h}_L^A$, $\mathbf{h}_L^T$, and $\mathbf{h}_L^S$ 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 $\mathbf{h}_L = \mathbf{h}_L^S + \mathbf{h}_L^A$, 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 $\mathbf{s}_L$ 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 heads compute attention maps $\mathbf{a}_L$ in the speech domain with additional text context information. Then, we choose $\mathbf{a}_L$ as key-value pairs with $\mathbf{h}_L$ as queries into the following module to generate audio-

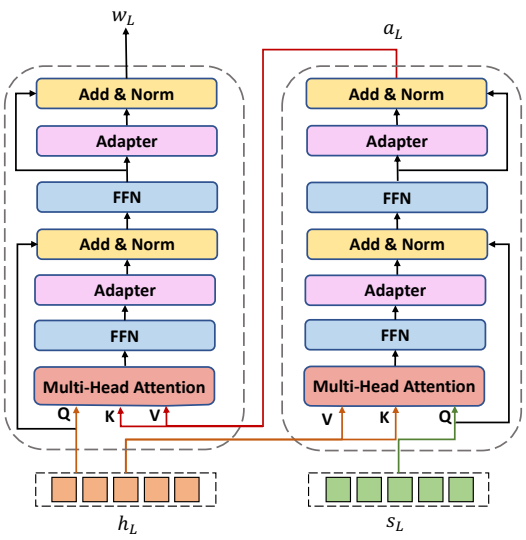

Figure 2: Overview of ST-Attention mechanism.

text feature representations $\mathbf{w}_L$. To ease the process of knowledge transfer, we adopt the Adapter [Houlsby *et al.*, 2019] in the ST-Attention mechanism with the parameter-efficient tuning to yield a compact and extensible model for every new NLP task. The feature map $\mathbf{w}_L$ is used for inference.

3.4 Prediction Layer Distillation

In addition to mimick 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, $\mathbf{w}_L^T$ and $\mathbf{w}_L^S = \mathbf{w}_L$ 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:

$$\mathcal{L}_{\mathcal{S}} = \alpha \mathcal{L}_{\mathcal{NLL}}(p_\tau(\mathbf{w}_L^S), p_\tau(\mathbf{w}_L^T)) + (1-\alpha)\mathcal{L}_{\mathcal{CE}}(\mathbf{w}_L^S, y), \tag{2}$$

where $\mathcal{L}_{\mathcal{NLL}}$ and $\mathcal{L}_{\mathcal{CE}}$ denote the negative log-likelihood loss and cross-entropy loss, respectively. $p_\tau(\cdot)$ refers to the softmax function with temperature τ. α is a hyper-parameter .

In the training phase, to achieve better performance, the knowledge transfer process consists of two steps. S first distill 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-A_1 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

$$\mathcal{L}_{\mathcal{A}} = \sum_{i=1}^{L} \frac{(\mathbf{h}_i^T - \mathbf{h}_i^S)(\mathbf{h}_i^A)^T}{\left\|\mathbf{h}_i^T - \mathbf{h}_i^S\right\|_2 \left\|\mathbf{h}_i^A\right\|_2}. \tag{3}$$

Data Usage		Experimental Results	
Training	Dev & Test	EM	F1
Manual	Manual	67.73	77.71
Manual	ASR	41.79	54.70
ASR	ASR	40.58	54.12
ASR	Manual	42.53	54.81

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

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-choices (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

Method	Spoken-SQuAD			Spoken-CoQA			FGC
	EM	F1	AOS	EM	F1	AOS	Accuracy (%)
SDNet [Zhu *et al.*, 2018]	57.81	71.84	64.72	41.51	53.12	42.57	76.70
HMM [Luo *et al.*, 2019]	54.43	65.11	46.88	37.71	50.09	34.75	72.00
Lee et al. [2019]	51.11	63.11	-	-	-	-	69.88
Speech-BERT [Chuang *et al.*, 2020]	51.09	63.09	59.61	40.41	51.77	43.11	68.72
vanilla BERT [Devlin *et al.*, 2019]	58.31	70.20	64.12	40.58	54.12	48.01	77.00
Su and Fung [2020]	59.71	70.94	65.01	42.11	55.64	48.17	77.78
aeBERT [Kuo *et al.*, 2020]	59.37	70.36	67.87	44.38	55.67	50.34	78.20
MRD-Net	62.31	74.95	70.81	49.24	60.78	54.02	81.91
Progressive MRD-Net	**63.12**	**75.89**	**71.79**	**50.75**	**61.99**	**54.68**	**82.73**

Table 2: Comparison results between our proposed model with other methods. Progressive MRD-Net is MRD-Net with progressive learning.

respectively. In the spoken conversational question answering task, the machine comprehension systems aim to fully understand the spoken multi-turn dialogues, and then answer questions among the passage and conversations.

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 τ and α 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 previous 2 rounds of questions and ground-truth answers. When training MRD-Net on FGC, we concatenate ASR token sequences of a passage, a question, and the corresponding answers in training. We train our student model using 2 NVIDIA 2080Ti GPU. We choose Exact Matched (EM) percentage and F1 score as the evaluation metrics for Spoken-CoQA and Spoken-SQuAD, and use accuracy to evaluate the model on FGC. Furthermore, we adopt Audio Overlapping Score (AOS) [Li *et al.*, 2018] to evaluate overlap of time span between predictions and ground-truth answers.

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%/77.71% 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.

Model	S-SQuAD	S-CoQA	FGC
	F1	F1	Accuracy (%)
MRD-Net	74.95	60.78	81.91
- w/o *pl* distillation	73.87	59.23	80.75
- w/o Audio-*A*	73.09	**58.12**	80.42
- w/o ST-Attention	**72.81**	58.61	**79.07**
- w/ *MSE*	74.32	60.13	81.20

Table 3: Ablation study of different components of MRD-Net on Spoken-SQuAD (S-SQuAD) and Spoken-CoQA (S-CoQA), respectively. For brevity, *pl* distillation denotes prediction layer distillation. *MSE* denotes using mean-squared-error loss function instead of Equation 1.

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

Algorithm	S-SQuAD	S-CoQA	FGC
	F1	F1	Accuracy (%)
KD (Text only)	72.23	57.68	78.81
RKD (Text only)	73.09	58.12	80.75
KD (Audio only)	64.51	52.31	70.01
"[CLS]"-based	69.32	53.10	76.43
Attention-based	74.11	59.11	80.78
MRD-Net	**74.95**	**60.78**	**81.91**

Table 4: Comparison results of different distillation objectives. *Text only* and *Audio only* denote *T* is trained on text or speech documents, respectively, and *S* is trained on ASR transcripts. '[CLS]"-based means using hidden features from the "[CLS]" token as distilled knowledge. Attention-based represents distilling knowledge from self-attention weights instead of using hidden state features in transformer block. To make better evaluation, all *T* and *S* model are fixed to the same structure.

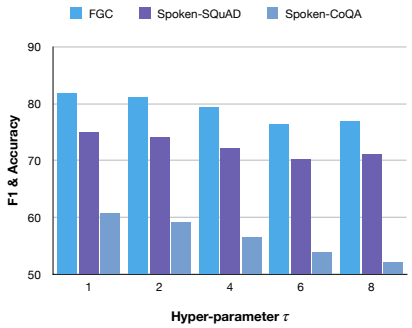

Figure 3: Performance of different temperature τ.

importance of the knowledge transfer process and audio features in this setting. The results also indicate that the acoustic-level information distilled from Audio-*A* can help *S* to deeply model the behavior of *T*. Specifically, we observe that without using ST-Attention leads to the performances drop by 2.14%, 2.17% and 2.84%, respectively. Moreover, it is worth to note that using mean-squared-error (MSE) loss instead of Eq.1 also suffers performance degradation.

Effects of Distillation Objectives. We investigate the effects of distillation objectives on the proposed method. The quantitative results are reported in Table 4. From Table 4, we can obtain the following conclusion. i) Compared with directly using knowledge distillation strategy on either text corpora or audio documents, the proposed model achieves better performance by using both acoustic and textual features. This indicates the importance of multi-domain knowledge for SQA tasks. ii) Directly utilizing the `[CLS]` token in BERT [Devlin *et al.*, 2019] to process passage for the following knowledge transfer process leads to the significant performance drop. This shows that, compared with the simple sentence-level classifier tasks, the SQA task is more challenging for machine systems to fully comprehend the passages and make correct predictions. iii) Our proposed MRKD approach outperforms the attention-based distillation approach [Jiao *et al.*, 2020]. Although knowledge embedded in attention weights contains more syntax and coreference information, it is hard to align similar abstract concepts

Algorithm	S-SQuAD	S-CoQA	FGC
	F1	F1	Accuracy (%)
Multi-T [2019]	71.93	56.69	78.76
Co-Att [2019]	72.82	58.07	79.63
ICCN [2020]	71.71	57.31	79.01
S-Fusion [2020]	68.16	51.79	75.13
ST-Attention	**74.95**	**60.78**	**81.91**
- w/o Adapters	74.41	60.15	81.13

Table 5: Comparison results of ST-Attention mechanism.

Algorithm	Happy		Sad		Angry		Neutral	
	Acc(h)	F1(h)	Acc(h)	F1(h)	Acc(h)	F1(h)	Acc(h)	F1(h)
Multi-T [2019]	84.4	81.9	77.7	74.1	73.9	70.2	62.5	59.7
ICCN [2020]	87.41	84.72	86.26	85.93	88.62	88.02	69.73	68.47
Co-Attention [2019]	88.64	87.61	89.01	88.3	93.04	93.21	80.31	79.09
Shallow-Fusion [2020]	89.71	88.34	89.48	89.2	93.82	93.9	80.93	81.01
St-Attention	**90.76**	**89.42**	**90.54**	**89.7**	**94.21**	**94.76**	**82.31**	**82.77**

Table 6: Comparison results of multi-modal emotion fusion mechanisms on ICMOCAP.

in speech representations. In contrast, our proposed method can effectively tackle the above issues.

Effects of ST-Attention. We also 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 [Siriwardhana *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 τ. We present the model performance with different temperature τ in Figure 3. τ is chosen from $\{1, 2, 4, 6, 8\}$. From Figure 3, MRD-Net achieves superior results when τ 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

[Busso *et al.*, 2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 2008.

[Chuang *et al.*, 2020] Yung-Sung Chuang, Chi-Liang Liu, and Hung-Yi Lee. SpeechBERT: Cross-modal pre-trained language model for end-to-end spoken question answering. In *INTERSPEECH*. 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*, 2019.

[Gao *et al.*, 2020] Mengya Gao, Yujun Shen, Quanquan Li, and Chen Change Loy. Residual knowledge distillation. *arXiv preprint arXiv:2002.09168*, 2020.

[Hahn and Choi, 2019] Sangchul Hahn and Heeyoul Choi. Self-knowledge distillation in natural language processing. In *RANLP*, 2019.

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

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019.

[Jiao *et al.*, 2020] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *EMNLP*, 2020.

[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*. ISCA, 2020.

[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*. IEEE, 2018.

[Lee *et al.*, 2019] Chia-Hsuan Lee, Yun-Nung Chen, and Hung-yi Lee. Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation. In *ICASSP*. IEEE, 2019.

[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*. ISCA, 2018.

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

[Luo *et al.*, 2019] Shang-Bao Luo, Hung-Shin Lee, Kuan-Yu Chen, and Hsin-Min Wang. Spoken multiple-choice question answering using multimodal convolutional neural networks. In *ASRU*. IEEE, 2019.

[Panayotov *et al.*, 2015] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP*. IEEE, 2015.

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

[Schneider *et al.*, 2019] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*. ISCA, 2019.

[Siriwardhana *et al.*, 2020] Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. Jointly Fine-Tuning "BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. In *INTERSPEECH*. ISCA, 2020.

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

[Sun *et al.*, 2020] Zhongkai Sun, Prathusha Kameswara Sarma, William A. Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *AAAI*. AAAI Press, 2020.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *ACL*, 2019.

[Wang *et al.*, 2018] Hui Wang, Hanbin Zhao, Xi Li, and Xu Tan. Progressive blockwise knowledge distillation for neural network acceleration. In *IJCAI*, 2018.

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

[You *et al.*, 2020b] Chenyu You, Nuo Chen, and Yuexian Zou. Contextualized attention-based knowledge transfer for spoken conversational question answering. *arXiv preprint arXiv:2010.11066*, 2020.

[You *et al.*, 2021] Chenyu You, Nuo Chen, and Yuexian Zou. Knowledge distillation for improved accuracy in spoken question answering. In *ICASSP*. IEEE, 2021.

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