Assignment 3

Aryan Kumar M23CSA510

April 12, 2025

GitHub Repository: `https://github.com/Aryank47/speech-understanding-assignment3`

Contents

1 Title of the Paper

DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering

2 Summary

The paper has proposed a framework called Discrete Spoken Unit Adaptive Learning (DUAL), which is the first “known” framework for textless spoken question answering that does not rely on ASR transcripts. It uses a self-supervised speech pre-trained model (HuBERT) to extract frame-level speech representations and applies k-means clustering (k=64, 128, 512) to generate discrete units. These units are then fed into a pre-trained language model (Longformer) adapted to handle longer sequences, enabling the direct prediction of answer boundaries in time. The study introduces a new benchmark dataset (NMSQA), enabling more realistic evaluations, and demonstrates that DUAL achieves competitive results compared to conventional cascaded ASR–TQA methods, especially when high ASR error rates are present.

3 What is Spoken Question Answering (SQA)?

Spoken Question Answering (SQA) is a task where the system is required to locate or extract answers directly from spoken content (i.e., audio recordings) in response to a spoken or written question. Instead of relying solely on text documents, SQA deals with audio documents where the retrieval is based on the audio signal itself. Applications of SQA include:

- **Virtual Assistants:** Devices like smart speakers or phone-based assistants (e.g., Siri, Google Assistant) use SQA to interact naturally with users.
- **Interactive Voice Response (IVR) Systems:** In customer service settings, SQA can help extract answers from recorded customer queries or support calls.
- **Accessibility Tools:** Systems designed for individuals with visual impairments may leverage SQA to read and understand spoken information.
- **Multimedia Information Retrieval:** SQA can be used in scenarios where audio or video content needs to be indexed and queried based on its speech content, such as lecture recordings or conference presentations.

4 Architecture

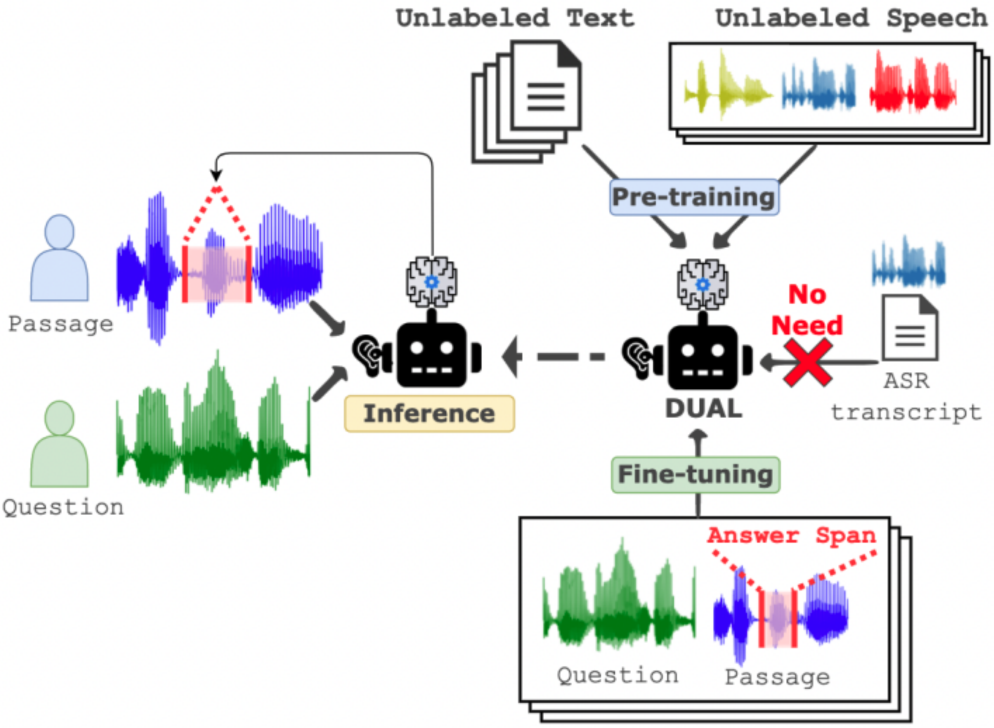

Figure 1: DUAL framework for textless (ASR transcript-free) SQA

4.1 How does the DUAL architecture work? What is the flow of the DUAL framework?

DUAL (Discrete Spoken Unit Adaptive Learning) introduces a novel textless approach for Spoken Question Answering (SQA) that bypasses the ASR stage. It works in the following way:

- **Speech Content Encoder (SCE):** The model first uses a self-supervised speech pre-trained network (HuBERT) to extract high-dimensional, frame-level features from the raw audio of both the spoken question and the spoken passage. Since the extracted features are continuous and can be excessively long, k-means clustering is applied (with different cluster sizes, 64, 128, 512) to quantize the speech into discrete units. This effectively transforms the continuous audio into a sequence of discrete "tokens"
- **Pre-trained Language Model (PLM) Adaptation:** These discrete tokens are then used as inputs to a pre-trained language model (Longformer) which is adapted to handle long sequences. The PLM is fine-tuned so that it learns to predict the start and end indices corresponding to the answer span directly from the discrete representation.

Flow of the Framework:

- **Input:** Spoken question and spoken passage.
- **Feature Extraction:** HuBERT processes the audio to generate frame-level representations.
- **Quantization:** K-means clustering converts these features into discrete units, forming a condensed sequence.
- **QA Prediction:** The Longformer model, using its adapted input embeddings, processes the concatenated discrete tokens and predicts the answer span as indices.
- **Mapping Back:** The predicted indices are mapped back into the corresponding time intervals in the original audio, giving the final answer in spoken form.

4.1.1 Evaluation Metrics

The paper uses two primary metrics to evaluate the performance of the spoken question answering (SQA) task:

- **Frame-level F1 Score (FF1):**
  FF1 is computed at the frame level in the audio signal to assess the overlap between the predicted answer span and the ground truth span. It is defined as the harmonic mean of precision and recall, where precision and recall are calculated over individual frames. Mathematically,

$$\text{FF1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Audio Overlapping Score (AOS):**
  AOS quantifies the temporal overlap between the predicted answer interval and the ground truth interval. If T_{overlap} denotes the duration of overlap and T_{union} is the total duration covered by both intervals, then

$$\text{AOS} = \frac{T_{\text{overlap}}}{T_{\text{union}}}$$

  A higher AOS indicates a better match between the predicted span and the actual answer span.

Why These Metrics?

Since SQA involves extracting precise answer spans from continuous audio signals, it is essential to measure both the exact temporal boundaries (with FF1) and the overall duration overlap (with AOS) between the predicted and ground truth segments. Together, these metrics provide a robust evaluation of the system's performance in localizing spoken answers.

5 Strengths

- **Innovative Approach:** Introduces a textless SQA method that bypasses error-prone ASR transcripts, addressing challenges inherent to low-resource and non-written languages.
- **Robust Representations:** Utilizes state-of-the-art self-supervised learning (HuBERT) and quantization through k-means clustering to capture fine-grained acoustic information.
- **Integration of Long-Context Modeling:** The paper uses Longformer to handle longer sequences efficiently accommodate the extensive duration of spoken passages.
- **Comprehensive Experiments:** Detailed experiments, including ablation studies and evaluations across synthesized and human speech, strengthen the validation of the proposed method.
- **New Novel Benchmark Dataset:** The paper has also released a novel and open-source dataset for SQA benchmark corpus called NMSQA. It has data with more realistic scenarios.

6 Weaknesses

- **Cluster Sensitivity:** The performance is sensitive wrt cluster size 128 vs. 512, which is a potential limitation in the robustness of the discrete unit extraction. For ex: HuBERT-128 with Longformer model gave the following output on the test (human) 55.9 for FF1 and 49.1 for AOS, whereas HuBERT-512 with Longformer gave 17.3 and 12.5 respectively, which is lower than HuBERT-128.
- **Dependence on Pre-trained PLMs:** Relying on text pre-trained language model weights for adaptation could lead to suboptimal performance if the domain mismatch is significant.
- **Generalizability Concerns:** The experiments are mostly centered on a single dataset (NMSQA) and Wikipedia-based content; further validation on diverse languages and domains is lacking.
- **Potential Overhead in Fine-Tuning:** The transfer of pre-trained embeddings and the adaptation process could result in complex tuning requirements, which might limit scalability.
- **Limited Analysis on Error Sources:** Although the paper shows robust performance at high ASR error rates (up to 70%), there is not enough info about the specific failure modes or error types in the proposed framework.

7 Minor Questions and Weaknesses

- The paper could further explain how the discrete unit quantization parameters might be optimally chosen for varying speech characteristics.
- A clarification on how handling noisy inputs or accents in human speech data would strengthen the evaluation.
- Also, more experiments on real-world conversational data could provide further insights into practical deployment scenarios.

8 Suggestions

The authors can maybe explore an adaptive technique for selecting the optimal number of clusters based on the input speech characteristics, potentially using data-driven methods to reduce sensitivity in performance.

Incorporating domain adaptation strategies for the PLM might also help mitigate the potential mismatch between text pre-training and acoustic inputs.

Also, expanding the experiments to include different languages and noisy real-world scenarios could strengthen the generalizability claims.

More detailed error analysis would help identify specific limitations and guide improvements.

A detailed analysis of why the performance dropped drastically for K = 512 (17.3 of FF1) is needed.

9 Rating and Justification

I would rate this paper **7/10**.

The innovative textless approach and comprehensive experimental validation is great, but sensitivity to clustering and limited multi-language and noisy environment experiments suggest room for further refinement and robustness.