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 i.e HuBERT to extract frame-level speech representations and applies k-means clustering with (k=64, 128, 512) to generate discrete units. These units are then fed into a pre-trained language model i.e Longformer which is adapted to handle longer sequences, enabling the direct prediction of answer boundaries in time. The paper shows that DUAL produces comparable and in some cases better results when compared to traditional cascaded ASR–TQA approaches, particularly when large ASR error rates are present, as shown in the results. It also adds a new benchmark dataset (NMSQA), based on Wikipedia-based material which allows for more realistic comparisons.

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:** Virtual assistants, such as Google Assistant, Siri, and smart speakers..
- **IVR Systems:** SQA can assist in extracting responses from recorded customer enquiries or support calls for customer service applications.
- **Accessibility Tools:** SQA can be used by systems made for people with visual impairments to read and comprehend 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

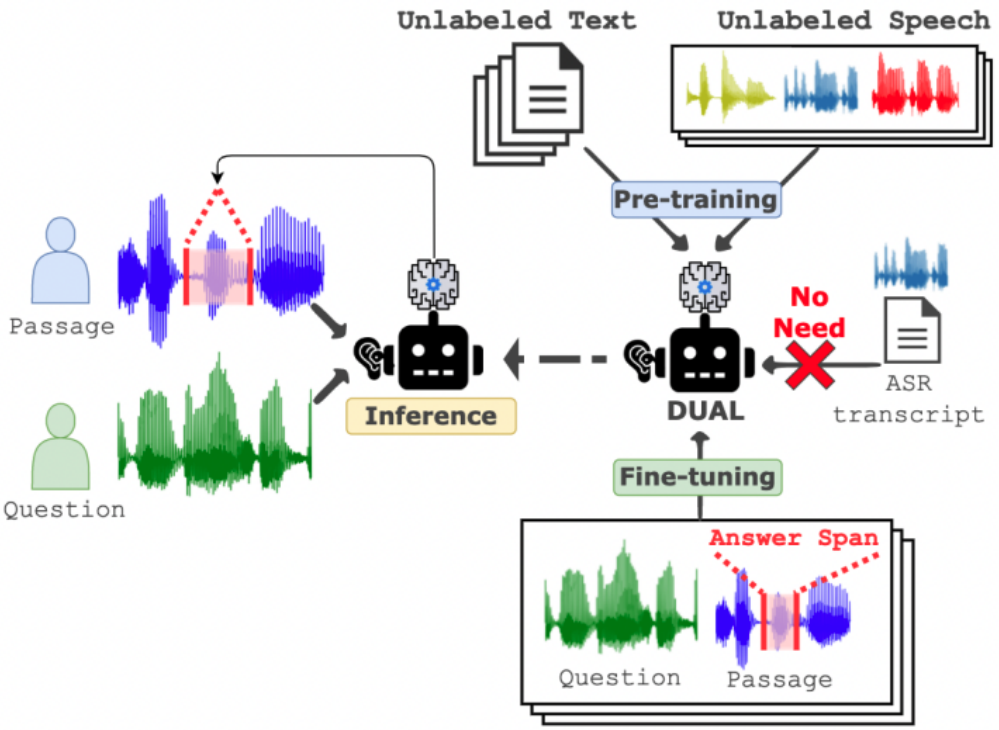

Figure 1: DUAL framework for textless (ASR transcript-free) SQA

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 initially extracts high-dimensional, frame-level characteristics from the raw audio of the spoken question and the spoken passage using a self-supervised speech pre-trained network (HuBERT). The speech is quantified into discrete pieces using k-means clustering, which has various cluster sizes of 64, 128, and 512. This is done since the extracted characteristics are continuous and may be unduly long. This successfully converts the continuous audio into a series of distinct "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 passage and inquiry
- **Feature Extraction:** HuBERT creates frame-level representations by processing the audio.
- **Quantization:** These features are broken down into discrete units using K-means clustering, creating a compacted sequence.
- **QA Prediction:** The Longformer model processes the concatenated discrete tokens using its modified input embeddings and forecasts the answer span (duration) as indices.
- **Mapping Back:** The final spoken response is obtained by mapping the anticipated indices back into the appropriate time periods in the original audio.

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 get the overlap between the predicted answer time and the ground truth time. It is defined as the harmonic mean of precision and recall, where precision and recall are calculated over individual frames.

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

- **Audio Overlapping Score (AOS):**
  AOS quantifies the temporal overlap between the predicted answer duration and the ground truth duration. 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 needs to measure both the exact time boundaries (with FF1) and the overall time overlap (with AOS) between the predicted and ground truth segments.

Table 2: *The performance of the proposed (DUAL) and baseline (cascade) approaches on the NMSQA dev and test sets. "Longformer†" indicates the Longformer model fine-tuned on clean text SQuAD-v1.1, while the normal "Longformer" was only pre-trained by unlabeled text data. The number after HuBERT (64, 128, 512) are numbers of clusters. "synth" and "human" represent the synthesized and human speech respectively.*

Input	Model	dev (synth)		test (human)	
		FF1	AOS	FF1	AOS
Baseline - Cascade (with ASR transcripts)					
SB	Longformer†	56.7	49.7	17.3	15.3
W2v2	Longformer†	65.7	58.3	64.2	57.4
Proposed - DUAL (without ASR transcripts)					
HuBERT-64	Longformer	47.8	42.2	39.0	33.0
HuBERT-128	Longformer	54.2	48.5	**55.9**	**49.1**
HuBERT-512	Longformer	**55.0**	**49.6**	17.3	12.5

Figure 2: Results

Table 3: *Ablation study on embedding assignment. All experiments used the HuBERT-128 setting. Performance was measured on the NMSQA dev set.*

Embedding Assignment	**FF1**	**AOS**
Most frequent	54.2	48.5
Least frequent	46.9	41.7
Random	51.7	46.2
Re-init	8.9	7.2
Scratch (baseline)	6.1	4.9

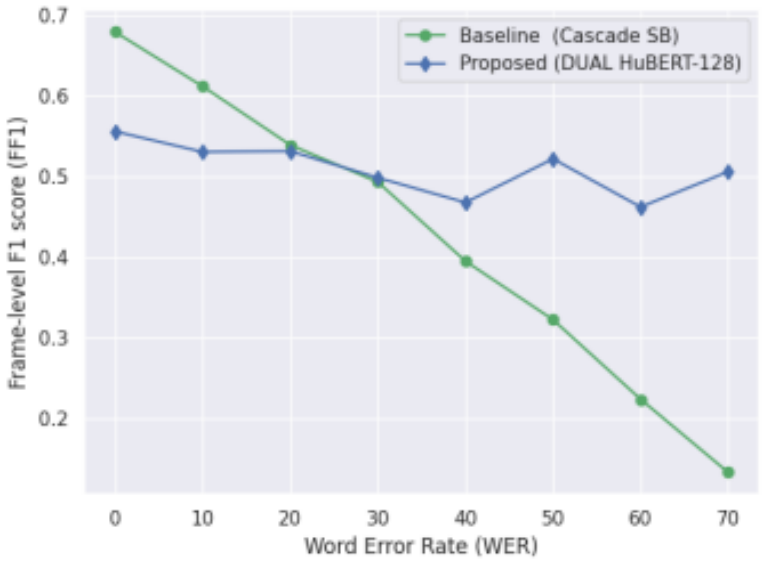

Figure 3: *Frame-level F1 (FF1) scores for DUAL and cascade approach (SB), evaluated on the small groups of full NMSQA dev set at different levels of ASR (SB) WER.*

Figure 3: Results

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:** If there is a large domain mismatch, depending on text pre-trained language model weights for adaptation may result in less than ideal performance.
- **Generalizability Concerns:** There is a lack of additional validation on a variety of languages and domains, and the studies are mostly focused on a single dataset (NMSQA) and Wikipedia-based material.
- **Potential Overhead in Fine-Tuning:** Complex tuning requirements resulting from the adaptation process and the transfer of pre-trained embeddings may restrict scalability.
- **Limited Analysis on Error Sources:** Despite the paper's strong performance at high ASR error rates (up to 70%), the suggested framework's precise failure modes and error kinds are not sufficiently described.

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.
- Additional research on conversational data from the actual world may also shed more light on practical deployment possibilities.

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 novel 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.