1. Introduction

- **Task:** query-by-example spoken term detection
  - Given a spoken query, detecting whether an audio segment contains the spoken query
  - Matching of signals directly on the acoustic level without transcribing them into text.
- **Our Model:** end-to-end attention-based multi-hop
  - Supervised Learning: true labels to learn, classification problem
  - Unsupervised Learning: generating labels by teacher approach (e.g., Dynamic Time Warping)

<table>
<thead>
<tr>
<th>Supervised</th>
<th>Unsupervised</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="example.png" alt="Supervised" /></td>
<td><img src="example.png" alt="Unsupervised" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><img src="example.png" alt="Attention analysis" /></th>
<th>Distance: difference between the position with the highest attention weight and the end of the query</th>
</tr>
</thead>
</table>

2. Proposed Approach

- **Query Representation (A)**
  - Input MFCC feature sequence: \(X_1, X_2, \ldots, X_T\)
  - Using LSTM encode to \(V_Q\)
- **Audio Segment Representation (B)**
  - Input MFCC feature sequence: \(X_t', X_{t+1}', \ldots, X_{T}\)
  - Using the same LSTM encode each frame \(S_1, S_2, \ldots, S_T\)
  - **Attention Mechanism:**
    \[
    a_t^i = S_t \odot V_Q \quad \text{or} \quad \text{Cosine similarity}
    \]
  - **Audio segment vector:**
    \[
    V_S = \sum_{t=1}^{T} a_t^i S_t
    \]

- **Hopping**
  - Using attention mechanism repeatedly to extract more relative information from audio segment.
- **Keyword Detection**
  - Given \(V_Q\) and \(V_S^n\) to determine result by neural network.

3. Experiments

- **Data set:** LibriSpeech
- **Training set:**
  - Query set: 500
  - Query and Audio segment pair: 70,000
- **Testing set 1:**
  - Query acoustic feature from training set
  - Query set: 30
  - Query and Audio segment pair: 1,500
- **Testing set 2:**
  - Query Acoustic feature is different.
  - Query set: 30
  - Query and Audio segment pair: 1,500
- **Testing set 3:**
  - Query keyword didn’t present in training set.
  - Query set: 100
  - Query and Audio segment pair: 10,000

**Comparison Table**

<table>
<thead>
<tr>
<th>Method</th>
<th>Test set 1</th>
<th>Test set 2</th>
<th>Test set 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a): DTW</td>
<td>0.6173</td>
<td>0.5778</td>
<td>0.5678</td>
</tr>
<tr>
<td>(b): 1-hop</td>
<td>0.6523</td>
<td>0.6246</td>
<td>0.5754</td>
</tr>
<tr>
<td>(c): 2-hop</td>
<td>0.6472</td>
<td>0.6430</td>
<td>0.5842</td>
</tr>
<tr>
<td>(d): 3-hop</td>
<td>0.6676</td>
<td>0.6404</td>
<td>0.5837</td>
</tr>
<tr>
<td>(e): 4-hop</td>
<td>0.6417</td>
<td>0.6476</td>
<td>0.5792</td>
</tr>
<tr>
<td>(f): (a) + (d)</td>
<td>0.6789</td>
<td>0.6430</td>
<td>0.5830</td>
</tr>
</tbody>
</table>

| Unsupervised | | | |
| (e): 1-hop | 0.6128 | 0.5893 | 0.5548 |
| (g): 3-hop | 0.6141 | 0.5964 | 0.5702 |