Audio Word2Vec: Unsupervised Learning of Audio Segment Representations using Sequence-to-sequence Autoencoder
Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, Lin-Shan Lee
College of Electrical Engineering and Computer Science, National Taiwan University, Taipei, Taiwan

1. Abstract
Audio Word2Vec:
- Offering any variable-length audio segment a fixed length vector representation (embedding) just as Word2Vec
- Use Seq2seq Autoencoder for training
- Unsupervised Learning (no label information is needed)
- Example Application: Query-by-Example Spoken Term Detection (STD)

2. Proposed Model: Seq2seq Autoencoder

3. Application: Query-by-Example STD

4. Experiment
LibriSpeech Corpus
1. Training set: a 5.4 hours audio archive (produced by 40 speakers)
2. Testing set: another 5.4 hours audio archive (produced by another 40 speakers)

Experimental Results

5. Analysis of the Embedding

<table>
<thead>
<tr>
<th>Phoneme Sequence Edit Distance</th>
<th># of pairs</th>
<th>Average cosine similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SA</td>
<td>DSA</td>
</tr>
<tr>
<td>0 (same)</td>
<td>95222</td>
<td>0.4847 0.5012</td>
</tr>
<tr>
<td>1</td>
<td>95355</td>
<td>0.4016 0.4237</td>
</tr>
<tr>
<td>2</td>
<td>987934</td>
<td>0.2674 0.2798</td>
</tr>
<tr>
<td>3</td>
<td>4651318</td>
<td>0.0835 0.0986</td>
</tr>
<tr>
<td>4</td>
<td>4791482</td>
<td>0.0255 0.0196</td>
</tr>
<tr>
<td>5 or more</td>
<td>3078684</td>
<td>0.0051 0.0013</td>
</tr>
</tbody>
</table>

6. Property: Audio Word2Vec
The vector difference of the embedding (projected to 2-dim) for the two words in a pair:
\[
\delta(\text{new}) - \delta(\text{few}) \approx \delta(\text{night}) - \delta(\text{fight})
\]
→ Phoneme replacement is realizable!

7. Conclusions
1. Seq2seq Autoencoder was proposed to learn fixed-length embedding of variable-length audio segment without supervision.
2. Outperformed DTW in Query-by-Example STD