1. Abstract
Speech2Vec:
- Representing any variable-length speech segment as fixed-dimensional vector (embedding) that captures the semantics of the speech segment
- Learning from speech instead of text
- Using RNN encoder-decoder architecture as backbone
- Training Methodology:
  - Skip-grams or CBOW (borrowed from Word2Vec)
  - Unsupervised learning (no label information is needed)

3. Word Similarity Experiments
- Used the 500 hours subset of LibriSpeech as the speech corpus to train Speech2Vec
  - Speech was pre-segmented into segments corresponding to words
  - Each segment was mapped to an embedding; for experimental purpose, embeddings of segments corresponding to the same word was averaged.
- Please refer to our paper for hyperparameters.
- Compared to word embeddings obtained from Word2Vec trained on the transcripts of the same subset.

Spearman’s rank correlation coefficient between the rankings produced by each model against the human rankings. The higher the better. Best performance is marked in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>Speech2Vec</th>
<th>Word2Vec</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cbow</td>
<td>skipgrams</td>
</tr>
<tr>
<td>Verb-143</td>
<td>0.223</td>
<td>0.315</td>
</tr>
<tr>
<td>SimLex-999</td>
<td>0.235</td>
<td>0.292</td>
</tr>
<tr>
<td>MC-30</td>
<td>0.716</td>
<td>0.846</td>
</tr>
<tr>
<td>WS-353</td>
<td>0.343</td>
<td>0.508</td>
</tr>
<tr>
<td>WS-353-SIM</td>
<td>0.484</td>
<td>0.663</td>
</tr>
<tr>
<td>WS-353-REL</td>
<td>0.192</td>
<td>0.346</td>
</tr>
<tr>
<td>RG-65</td>
<td>0.705</td>
<td>0.790</td>
</tr>
<tr>
<td>MEN</td>
<td>0.509</td>
<td>0.619</td>
</tr>
</tbody>
</table>

- For both Speech2Vec and Word2Vec, skip-grams outperforms its CBOW counterpart.
- Skip-grams Speech2Vec outperforms skip-grams Word2Vec.
- Speech2Vec is able to capture semantic information from speech corpus.
- Need to validate on downstream tasks.

2. Proposed Model: Speech2Vec

Input: a pre-segmented speech corpus consisting of N speech segments $x^{(1)}$, $x^{(2)}$, ..., $x^{(N)}$

Skip-grams:
- Encoder encodes the input speech segment $x^{(i)}$ into a fixed-length embedding $z^{(i)}$
- Decoder aims to predict the nearby speech segments $x^{(i-1)}$ and $x^{(i+1)}$ from $z^{(i)}$ (window size = 2 as example)

CBOW:
- Encoder encodes nearby speech segments $x^{(i-1)}$ and $x^{(i+1)}$ into embeddings and computes their sum $z^{(i)}$
- Decoder aims to predict $x^{(i)}$ from $z^{(i)}$

4. Follow-Up Work
Unsupervised cross-modal alignment of speech and text embedding spaces.
Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. 2018.

1. Learn speech embeddings from large speech corpora using Speech2Vec (unsupervised)
2. Learn word embeddings from large text corpora using Word2Vec (unsupervised)
3. Align (learn a mapping) the two embedding spaces using adversarial training technique (unsupervised)

→ Towards unsupervised automatic speech recognition & unsupervised speech-to-text translation