Overview

Few-shot Spoken Language Understanding (SLU)
- Assumes limited labeled speech data access, alongside more readily obtainable text data.

Speech-text models for few-shot & zero-shot SLU
- Match the performance of previous models with 0-20% of speech data.

Analysis of hidden representations
- Explains the zero-shot text-to-speech transferability.
- Suggests fine-tuning with bottom layers frozen, which improves zero-shot performance.

Speech-Text Models

• Learn shared representations for speech & text.
• Improve ASR, speech translation, etc.

Conclusion

• Speech-text models exhibit zero-shot transferability from text to speech in SLU.
• Few-shot performance matches previous work trained with 5+ times more speech data.
• Bottom layers are task-agnostic and top layers are task-specific.
• Freezing bottom layers enhances zero-shot performance.

Analysis

Average Neuron-Wise Correlation (ANC)\(^{[4]}\)

\[
\frac{1}{d} \sum_{i=1}^{d} \text{corr}(X_i, Y_i)
\]

\(X, Y \in \mathbb{R}^d\): different views (e.g. text & speech) of the same data instance.

NER (sequence labeling)

• Bottom layers align speech & text into a shared space.
• Fine-tuning only influences top layers.
• Tasks affect top layers more than input modalities → top layers are task-specific.

Few-Shot SLU

Fine-Tuning

Speech-Text Model

Speech-Text Model

Speech-Text Model

Speech-Text Model

Testing

All-speech

Few-shot

Zero-shot

SLUE Benchmark\(^{[1]}\)

Sentiment Analysis (Classification)

<table>
<thead>
<tr>
<th>Labeled Data</th>
<th>Speech-Only</th>
<th>Speech-Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Text</td>
<td>HubERT</td>
<td>Speech-LM-P</td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 hr</td>
<td></td>
<td>36.9</td>
</tr>
<tr>
<td>12.8 hrs</td>
<td></td>
<td><strong>43.0</strong></td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td><strong>45.2</strong></td>
</tr>
<tr>
<td>10 mins full</td>
<td></td>
<td>45.2</td>
</tr>
<tr>
<td>1 hr full</td>
<td></td>
<td>46.4</td>
</tr>
</tbody>
</table>


