Few-Shot Spoken Language Understanding Via Joint Speech-Text Models

Chung-Ming Chien\textsuperscript{1}, Mingjiamei Zhang\textsuperscript{2}, Ju-Chieh Chou\textsuperscript{1}, Karen Livescu\textsuperscript{1}
TTIC\textsuperscript{1}, The University of Chicago\textsuperscript{2}
Highlight

• We use speech-text models for few-shot & zero-shot Spoken Language Understanding (SLU).
Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Understanding (SLU).
  - Match the performance of previous models with 0-20% of speech data.
Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Understanding (SLU).
  - Match the performance of previous models with 0-20% of speech data.
- We analyze pre-trained & fine-tuned speech-text models.
Highlight

• We use speech-text models for few-shot & zero-shot Spoken Language Understanding (SLU).
  - Match the performance of previous models with 0-20% of speech data.
• We analyze pre-trained & fine-tuned speech-text models.
  - Explain the zero-shot text-to-speech transferability of speech-text models.
Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Understanding (SLU).
  - Match the performance of previous models with 0-20% of speech data.
- We analyze pre-trained & fine-tuned speech-text models.
  - Explain the zero-shot text-to-speech transferability of speech-text models.
  - Suggest fine-tuning with bottom layers frozen, which improves zero-shot performance.
Background: Speech-Text Pre-Trained Models
Background: Speech-Text Pre-Trained Models

- Models studied:
  - SpeechLM \cite{1}
  - SpeechUT \cite{2}

Background: Speech-Text Pre-Trained Models

- Models studied:
  - SpeechLM [1]
  - SpeechUT [2]

Background: Speech-Text Pre-Trained Models

- Models studied:
  - SpeechLM [1]
  - SpeechUT [2]

Background: Speech-Text Pre-Trained Models

- Models studied:
  - SpeechLM [1]
  - SpeechUT [2]

Background: Speech-Text Pre-Trained Models

- Models studied:
  - SpeechLM \[1\]
  - SpeechUT \[2\]

Background: Speech-Text Pre-Trained Models

- Models studied:
  - SpeechLM \[^1\]
  - SpeechUT \[^2\]


Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Understanding (SLU).
  - Match the performance of previous models with 0-20% of speech data.

- We analyze pre-trained & fine-tuned speech-text models.
  - Explain the zero-shot text-to-speech transferability of speech-text models.
  - Suggest fine-tuning with bottom layers frozen, which improves zero-shot performance.
Few-Shot & Zero-Shot Spoken Language Understanding
Few-Shot & Zero-Shot Spoken Language Understanding

- Assumptions
  - Limited labeled speech data
  - More text data
Few-Shot & Zero-Shot Spoken Language Understanding

• Assumptions
  - Limited labeled speech data
  - More text data
Few-Shot & Zero-Shot Spoken Language Understanding

- Assumptions
  - Limited labeled speech data
  - More text data
Experimental Setups

• SLU tasks: SLUE Benchmark [3]
  - Sentiment Analysis (SA)
    ▸ Classification: “positive,” “neutral,” or “negative” sentiments
  - Named Entity Recognition (NER)
    ▸ Sequence labeling

• Speech-text models fine-tuned with labeled text data + different amounts of labeled speech data

• Other details follow the default setup of the SLUE benchmark

Sentiment Analysis

- Zero-shot performance comparable to models using full speech data.

<table>
<thead>
<tr>
<th>Sentiment Analysis Accuracy (%)</th>
<th>Labeled Data</th>
<th>Prior work: Speech-Only</th>
<th>Speech-Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speech Text</td>
<td>HuBERT SpeechLM-P SpeechLM-H</td>
<td></td>
</tr>
<tr>
<td>Baselines</td>
<td>1 hr -</td>
<td>36.9 37.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.8 hrs -</td>
<td>43.0 45.6 45.3</td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>- full</td>
<td>45.2 45.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10 mins full</td>
<td>45.2 38.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 hr full</td>
<td>46.4 43.4</td>
<td></td>
</tr>
</tbody>
</table>
Sentiment Analysis

- Zero-shot performance comparable to models using full speech data.

<table>
<thead>
<tr>
<th>Sentiment Analysis Accuracy (%)</th>
<th>Labeled Data</th>
<th>Prior work: Speech-Only</th>
<th>Speech-Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speech</td>
<td>Text</td>
<td>HuBERT</td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 hr</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.8 hrs</td>
<td>-</td>
<td>43.0</td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>full</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 mins</td>
<td>full</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 hr</td>
<td>full</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sentiment Analysis

- Zero-shot performance comparable to models using full speech data.

<table>
<thead>
<tr>
<th>Sentiment Analysis Accuracy (%)</th>
<th>Labeled Data</th>
<th>Prior work: Speech-Only</th>
<th>Speech-Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speech</td>
<td>Text</td>
<td>HuBERT</td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 hr</td>
<td>-</td>
<td>-</td>
<td>36.9</td>
</tr>
<tr>
<td>12.8 hrs</td>
<td>-</td>
<td>-</td>
<td>43.0</td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>full</td>
<td></td>
<td>45.2</td>
</tr>
<tr>
<td>10 mins</td>
<td>full</td>
<td></td>
<td>45.2</td>
</tr>
<tr>
<td>1 hr</td>
<td>full</td>
<td></td>
<td>46.4</td>
</tr>
</tbody>
</table>
Named Entity Recognition

![Graph showing F1 scores with LM decoding vs. hours of speech data used. The graph compares different models: Speech only, Speech-text, HuBERT baseline, SpeechLM-P, SpeechLM-H, and SpeechUT.](image-url)
Named Entity Recognition

Graph showing F_1 scores with LM decoding versus hours of speech data used (log-scale). The graph includes lines for Speech only, Speech-text, and HuBERT baseline, with corresponding colors and markers.
Named Entity Recognition

- SpeechUT has great zero-shot performance.
Named Entity Recognition

- SpeechUT has great zero-shot performance.
Named Entity Recognition

- SpeechUT has great zero-shot performance.
- Speech+text fine-tuning is better than speech-only fine-tuning.
Named Entity Recognition

• SpeechUT has great zero-shot performance.

• Speech+text fine-tuning is better than speech-only fine-tuning.
Named Entity Recognition

- SpeechUT has great zero-shot performance.

- Speech+text fine-tuning is better than speech-only fine-tuning.
  - Outperforms HuBERT (speech-only) with 20% of speech data.
Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Understanding (SLU).
  - Match the performance of previous models with 0-20% of speech data.

- We analyze pre-trained & fine-tuned speech-text models.
  - Explain the zero-shot text-to-speech transferability of speech-text models.
  - Suggest fine-tuning with bottom layers frozen, which improves zero-shot performance.
Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) \[4\]
  \[
  \frac{1}{d} \sum_{i=1}^{d} \text{corr}(X_i, Y_i)
  \]
  - with \(X, Y \in \mathbb{R}^d\) representing different views (e.g. text & speech) of the same data instance.

Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC)\textsuperscript{[4]}
  \[
  \frac{1}{d} \sum_{i=1}^{d} \text{corr}(X_i, Y_i)
  \]
- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.

Analysis Method: Average Neuron-Wise Correlation

• Average Neuron-Wise Correlation (ANC)\(^4\)

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

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

Analysis Method: Average Neuron-Wise Correlation

• Average Neuron-Wise Correlation (ANC)\(^4\)

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

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

Analysis Method: Average Neuron-Wise Correlation

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

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

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

\[\text{[4]}\text{ M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.}\]
Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) \(^4\)
  \[
  \frac{1}{d} \sum_{i=1}^{d} \text{corr}(X_i, Y_i)
  \]

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

Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) \[^4\]
  \[
  \frac{1}{d} \sum_{i=1}^{d} \text{corr}(X_i, Y_i)
  \]
- with \(X, Y \in \mathbb{R}^d\) representing different views (e.g. text & speech) of the same data instance.

Analysis Method: Average Neuron-Wise Correlation

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

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

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

Analysis Method: Average Neuron-Wise Correlation

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

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

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

Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC)\footnote{M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.}

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

- with \(X, Y \in \mathbb{R}^d\) representing different views (e.g. text & speech) of the same data instance.
Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC)\[^{[4]}\]
  \[
  \frac{1}{d} \sum_{i=1}^{d} \text{corr}(X_i, Y_i)
  \]
  - with \(X, Y \in \mathbb{R}^d\) representing different views (e.g. text & speech) of the same data instance.

ANC scores between speech and text representations in pre-trained and fine-tuned models
Representation Alignment in Pre-Trained & Fine-Tuned Models

\( corr(X_i, Y_i) \): whether speech & text representations are aligned.

ANC scores between speech and text representations in pre-trained and fine-tuned models
Representation Alignment in Pre-Trained & Fine-Tuned Models

ANC scores between speech and text representations in pre-trained and fine-tuned models

![Graph showing ANC scores between speech and text representations in pre-trained and fine-tuned models. The graph displays different models and their performance across transformer layer indices.]

- Pre-trained model
- Fine-tuned on NER
- Fine-tuned on SA

Legend:
- SpeechLM-P
- SpeechLM-H
ANC scores between speech and text representations in pre-trained and fine-tuned models

![Graph showing ANC scores for different models and layers]
Speech-text models learn aligned speech & text representations in bottom layers.
• Speech-text models learn aligned speech & text representations in bottom layers.

ANC scores between speech and text representations in pre-trained and fine-tuned models
• Speech-text models learn aligned speech & text representations in bottom layers.

• Pre-trained & fine-tuned models are similar in bottom layers and differ more in top layers.
Representation Alignment in Pre-Trained & Fine-Tuned Models

- Speech-text models learn aligned speech & text representations in bottom layers.

- Pre-trained & fine-tuned models are similar in bottom layers and differ more in top layers.
  - Fine-tuning affects top layers more.
Top Layers Are Task Specific

ANC scores between speech representations in models with different fine-tuning setups

- FT speech SA vs. FT text SA
- FT speech SA vs. FT speech NER

Transformers layer index

ANC scores

Top Layers Are Task Specific

$corr(X_i, Y_i)$: how much pre-trained and fine-tuned models differ.

ANC scores between speech representations in models with different fine-tuning setups.
Top Layers Are Task Specific

ANC scores between speech representations in models with different fine-tuning setups

- FT speech SA vs. FT text SA
- FT speech SA vs. FT speech NER
Top Layers Are Task Specific

- We compare ANC scores between speech representations in models with different fine-tuning setups.
Top Layers Are Task Specific

- We compare:
  - Models fine-tuned on the same task with different input modalities.
Top Layers Are Task Specific

- We compare
  - Models fine-tuned on the same task with different input modalities.
  - Models fine-tuned on different tasks with the same input modality.
Top Layers Are Task Specific

- We compare
  - Models fine-tuned on the same task with different input modalities.
  - Models fine-tuned on different tasks with the same input modality.
- During fine-tuning, the task makes a larger difference than the input modality to top layers.

![ANC scores between speech representations in models with different fine-tuning setups](image)
Inspired By the Analysis...
Inspired By the Analysis...

- Bottom layers align speech & text representations.
Inspired By the Analysis...

- Bottom layers align speech & text representations.
  - Should not be affected by fine-tuning.
Inspired By the Analysis...

- Bottom layers align speech & text representations.
  - Should not be affected by fine-tuning.
- Top layers are task specific.
Inspired By the Analysis...

- Bottom layers align speech & text representations.
  - Should not be affected by fine-tuning.
- Top layers are task specific.
  - Should be fine-tuned.
Inspired By the Analysis...

- Bottom layers align speech & text representations.
  - Should not be affected by fine-tuning.
- Top layers are task specific.
  - Should be fine-tuned.
- How about fine-tuning only top layers and keeping bottom layers frozen?
Fine-Tuning with Bottom Layers Frozen

$F_1$ scores for NER with varying number of frozen layers during fine-tuning
Fine-Tuning with Bottom Layers Frozen

- All-speech & few-shot: slight performance reduction.

$F_1$ scores for NER with varying number of frozen layers during fine-tuning
Fine-Tuning with Bottom Layers Frozen

- All-speech & few-shot: slight performance reduction.
- Zero-shot: significant improvements in text-to-speech transferability.
Conclusion

• Speech-text models for few-shot SLU.
  - Speech-text models exhibit zero-shot transferability from text to speech.
  - Few-shot performance matches previous work trained with only 20% of speech data.

• Analysis of speech-text models.
  - Bottom layers are task-agnostic and top layers are task-specific.
  - Freezing bottom layers enhances zero-shot performance.