Acoustically grounded word embeddings for improved acoustics-to-word speech recognition

Shane Settle
Kartik Audhkhasi, Karen Livescu, Michael Picheny

ICASSP 2019
Models for Speech Recognition

- Traditional models for speech recognition are sub-word based
- Acoustics-to-word (A2W) models directly map input acoustic features to words without the need for additional decoding [Soltau+, Audhkhasi+ 2017] [Audhkhasi+, Li+, Yu+ 2018]

![Waveform](image-url)
Acoustics-to-Word (A2W) Models for Speech Recognition

how are you
Acoustics-to-Word (A2W) Models for Speech Recognition
Acoustics-to-Word (A2W) Models for Speech Recognition

Connectionist Temporal Classification (CTC) [Graves+ 2006] resolves input/target length disparity to allow for frame-wise prediction.
Acoustics-to-Word (A2W) Models for Speech Recognition

Connectionist Temporal Classification (CTC) [Graves+ 2006] resolves input/target length disparity to allow for frame-wise prediction.
Acoustics-to-Word (A2W) Models for Speech Recognition

Connectionist Temporal Classification (CTC) [Graves+ 2006] resolves input/target length disparity to allow for frame-wise prediction.
Acoustics-to-Word (A2W) Models for Speech Recognition

Connectionist Temporal Classification (CTC) [Graves+ 2006] resolves input/target length disparity to allow for frame-wise prediction.

![Diagram showing how A2W models work](image)
While prior work [Soltau+ 2018] matches sub-word performance training on 125Khrs, a gap remains for smaller datasets:

- difficulty learning rare/infrequent words
- out-of-vocabulary words

**Idea:** Use pre-trained acoustically grounded word embeddings to improve quality of the learned word embedding matrix
Acoustically Grounded Word Embeddings (AGWE)

Given \((\text{acoustic, character})\) word pairs \((x, c)\), we train embedding functions \(f(\cdot)\) and \(g(\cdot)\) to learn mappings into a shared space:
Acoustically Grounded Word Embeddings (AGWE)

Given (acoustic, character) word pairs \((x, c)\), we train embedding functions \(f(\cdot)\) and \(g(\cdot)\) to learn mappings into a shared space:
Acoustically Grounded Word Embeddings (AGWE)

Given \((\text{acoustic, character})\) word pairs \((x, c)\), we train embedding functions \(f(\cdot)\) and \(g(\cdot)\) to learn mappings into a shared space:
Acoustically Grounded Word Embeddings (AGWE)

Most offending different character sequence: [He+ 2017]

\[
\max \left\{ 0, m + d_{\cos}(f(x), g(c)) - \min_{c^- \neq \text{char}(x)} d_{\cos}(f(x), g(c^-)) \right\}
\]
Acoustically Grounded Word Embeddings (AGWE)

Most offending different spoken word example: [He+ 2017]

\[
\max \left\{ 0, m + d_{\cos}(g(c), f(x)) - \min_{\text{char}(x^-) \neq c} d_{\cos}(g(c), f(x^-)) \right\}
\]
Acoustics-to-Word Recognition: AGWE Initialized
Acoustics-to-Word Recognition: AGWE Regularized
Acoustics-to-Word Recognition: AGWE Frozen

how are you

ε how how how ε are are ε you

ε how how how ε are are ε you

how are you

<UNK>

how are you

...
Experimental Setup

Data
- 300h Switchboard corpus; conversational telephone English
- Standard log-Mel spectral features

Acoustically grounded word embeddings (AGWE)
- Acoustic view: 6-BLSTM (512d) → 256d
- Character view: 64d char embed → 1-BLSTM (512d) → 256d
- Tuned on development set word discrimination performance

Acoustics-to-word (A2W) recognition
- Acoustic view → prediction layer over $|V|$ words
- Word error rate reported on Hub5-2000 Switchboard evaluation set
Word Discrimination Development Set Results

![Bar Chart]

**Average Precision (AP)**

- **Baseline Weights**
- **AGWE**

- Vocabulary Size:
  - 4K
  - 10K
  - 20K

- Baseline Weights:
  - 0.5
- AGWE:
  - 0.75
  - 1.0

---

**Vocabulary Size**

**Average Precision (AP)**
Acoustics-to-Word Recognition: Switchboard Results

Vocabulary Size

<table>
<thead>
<tr>
<th></th>
<th>4K</th>
<th>10K</th>
<th>20K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>16</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>Frozen</td>
<td>12</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Initialized</td>
<td>14</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Regularized</td>
<td>18</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>Frozen+34K rescore</td>
<td>20</td>
<td>19</td>
<td>18</td>
</tr>
</tbody>
</table>
Acoustics-to-Word Recognition: Switchboard Results

![Graph showing WER (%) for different vocabulary sizes and model configurations.]

- Vocabulary Size: 4K, 10K, 20K
- WER (%):
  - Baseline
  - Frozen
  - Initialized
  - Regularized
  - Frozen+34K rescore

The graph visualizes the WER (%) for different vocabulary sizes and model configurations.
Acoustics-to-Word Recognition: Switchboard Results

![Graph showing WER (%) for different vocabulary sizes and model configurations.]

- **Baseline**
- **Frozen**
- **Initialized**
- **Regularized**
- **Frozen+34K rescore**

Vocabulary Size:
- 4K
- 10K
- 20K
Acoustics-to-Word Recognition: Switchboard Results

The diagram shows the Word Error Rate (WER) for different vocabulary sizes (4K, 10K, 20K) across various models. The models compared are:
- Baseline
- Frozen
- Initialized
- Regularized
- Frozen+34K rescore

The baseline model generally performs best, with the lowest WER across all vocabulary sizes. The WER decreases as the vocabulary size increases for all models, indicating improved performance with larger vocabularies. The Frozen+34K rescore model shows notable improvement over the Baseline, especially at the 20K vocabulary size.
Acoustics-to-Word Recognition: Vocabulary Extension
Acoustics-to-Word Recognition: Vocabulary Extension
Acoustics-to-Word Recognition: Switchboard Results

<table>
<thead>
<tr>
<th>Vocabulary Size</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4K</td>
<td>10</td>
</tr>
<tr>
<td>10K</td>
<td>12</td>
</tr>
<tr>
<td>20K</td>
<td>14</td>
</tr>
</tbody>
</table>

- Baseline
- Frozen
- Initialized
- Regularized
- Frozen+34K rescore
Acoustics-to-Word Recognition: Frozen+34K Rescores

**REF:** some REMINDERS for me as we are talking
**HYP (1st pass):**
**HYP (rescoring):**

**REF:** fair and speedy TRIAL
**HYP (1st pass):**
**HYP (rescoring):**

**REF:** but those LOANS ARE so much cheaper
**HYP (1st pass):**
**HYP (rescoring):**
<table>
<thead>
<tr>
<th></th>
<th>REF:</th>
<th>HYP (1st pass):</th>
<th>HYP (rescoring):</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>some REMINDERS for me as we are talking</td>
<td>some &lt;UNK&gt; for me as we are talking</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>fair and speedy TRIAL</td>
<td>fair and speedy &lt;UNK&gt;</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>but those LOANS ARE so much cheaper</td>
<td>but those &lt;UNK&gt; so much cheaper</td>
<td></td>
</tr>
</tbody>
</table>
Acoustics-to-Word Recognition: Frozen+34K Rescores

REF: some REMINDERS for me as we are talking
HYP (1st pass): some <UNK> for me as we are talking
HYP (rescoring): some REMINDERS for me as we are talking

REF: fair and speedy TRIAL
HYP (1st pass): fair and speedy <UNK>
HYP (rescoring): fair and speedy TRIAL

REF: but those LOANS ARE so much cheaper
HYP (1st pass): but those <UNK> so much cheaper
HYP (rescoring): but those LOANER so much cheaper
Acoustics-to-Word Recognition: Switchboard Results

Vocabulary Size

WER (%)

Baseline  Frozen  Initialized  Regularized  Curriculum [Yu+ 2018]

20K

10

12

14

16

18

20
Conclusion

- Pre-trained acoustically grounded word embeddings (AGWEs) give consistent improvements in A2W recognition
- AGWEs allow straightforward test time vocabulary extension
- Ongoing work includes curriculum learning, joint training, and application to low resource languages