Acoustic word embeddings

Karen Livescu
Toyota Technological Institute at Chicago
Joint work with...

Wanjia He
Katie Henry
Aren Jansen
Keith Levin
Herman Kamper
Shane Settle
Weiran Wang
(Textual) word embeddings

- Representation of (written) words as vectors in $\mathbb{R}^d$
- Semantically similar words should have similar vectors
- Examples: latent semantic analysis, word2vec, GloVe

[Figure credit: Yves Peirsman]
Spoken words have lots of variants (lots = a continuum)

Variations due to pronunciation variant, speaker, acoustic environment, mood, state of inebriation...
Acoustic word embeddings

- Representation of a spoken word as a vector in $\mathbb{R}^d$
- "Spoken word" = speech signal of arbitrary length corresponding to a word
- **Minimal goal:** Same-word signals should cluster together
- **Other desiderata?** Maybe.
Acoustic word embeddings: Applications

- Any task involving distances between speech segments as a subroutine
- E.g., query-by-example search:

Downstream applications:
- Open-vocabulary search in lectures, YouTube videos, ...
- Low-resource/unwritten/unknown language
Acoustic word embeddings: More applications

Any task involving \textbf{distances between speech segments} as a subroutine

- Query-by-example search
- Spoken term detection
- Spoken term discovery
- Whole-word speech recognition
Query-by-example search: Classic approach

Dynamic time warping (DTW):

- Slow
- Hard to improve via machine learning

[Figure credit: Proenca et al. 2015]
Query-by-example search with acoustic word embeddings

Everything can be learned from data!
Preliminaries: Acoustic features

waveform

spectrogram

MFCCs
Embedding approaches: Template-based

[ASRU 2013]

Represent a word segment $Y$ as a vector of distances to a set of other (template) word segments $\{R_1, \ldots, R_m\}$, $m \approx 10,000$:

$$f(Y) = [\text{DTW-dist}(Y, R_1) \ldots \text{DTW-dist}(Y, R_m)]$$

Then reduce dimensionality
Evaluation: Word discrimination task

- **Input**: Pair of acoustic signals
- **Output**: “Same word” or “different words”
- (Proxy task for query-by-example search)
- **Method**: Threshold the distance between the embeddings:
  \[ d_{\cos}(x_1, x_2) = 1 - \frac{x_1^T x_2}{\|x_1\| \|x_2\|} \]
- **Evaluation**: Average precision (AP)
Neural embeddings: CNN-based [ICASSP 2016]

Embedding is activation vector of top layer

\[ x_i = f(Y_i) \]

\[ \times n_{\text{full}} \]

\[ \times n_{\text{conv}} \]

\[ Y_i \]
Neural embeddings: RNN-based [SLT 2016]

Embedding is

- Final hidden state vector
- Concatenation of hidden states at the two ends
- Activation vector of fully connected layer following RNN
Training objectives

Word classifier log loss
- After adding a softmax layer

Contrastive loss (triplet loss, Siamese networks)
- Bring together same-word pairs, separate different ones

\[ l(x_1, x_2) = \max\{0, m + d_{\cos}(x_1, x_2) - d_{\cos}(x_1, x_3)\} \]
Word discrimination results

Average precision (AP)

- DTW
- DTW + learned frame features [Carlin et al. 11]
- Template-based [ASRU 2013]
- CNN classifier [ICASSP 2016]
- CNN Siamese [ICASSP 2016]
- RNN classifier [SLT 2016]
- RNN Siamese [SLT 2016]
Visualization: RNN embeddings

2-dimensional t-SNE embeddings [van der Maaten & Hinton 2008]
Search for spoken query in a 433-hour speech corpus

- **DTW baseline**: Optimized for speed, uses locality-sensitive hashing (LSH) to quickly pre-select likely frame matches [Jansen & van Durme 2012]

- **AWE-based search**: Uses LSH to find approximate nearest neighbor embeddings [Levin+ 15, Interspeech 2017]

<table>
<thead>
<tr>
<th>System</th>
<th>FOM (↑)</th>
<th>OTWV (↑)</th>
<th>P@10 (↑)</th>
<th>Time (s) (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTW</td>
<td>6.7</td>
<td>2.7</td>
<td>44.0</td>
<td>24.70</td>
</tr>
<tr>
<td>Template-based</td>
<td>24.5</td>
<td>14.4</td>
<td>34.5</td>
<td>0.08</td>
</tr>
<tr>
<td>Siamese RNN</td>
<td>43.3</td>
<td>22.4</td>
<td>60.2</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Visualization: Query-by-example task

[Interspeech 2017]

t-SNE visualization of hits/misses for several queries
Query-by-example: Speed-performance tradeoff [Interspeech 2017]
Relating spoken and written words

- We should be able to use the fact that the acoustics correspond to some written character sequence
- Also, some tasks involve “distances” between speech segments and written words
  - Spoken term detection
  - Automatic speech recognition

“Barack Obama” = ?
Joint learning of acoustic and text embeddings [ICLR 2017]

Character-based RNN text embedding:
Joint learning of acoustic and text embeddings [ICLR 2017]

Contrastive losses:

\[
\begin{align*}
    l_0(x, c) &= \max\{0, m + d_{\cos}(x, c) - d_{\cos}(x, c^-)\} \\
    l_1(x, c) &= \max\{0, m + d_{\cos}(x, c) - d_{\cos}(c^-, c)\} \\
    l_2(x, c) &= \max\{0, m + d_{\cos}(x, c) - d_{\cos}(x^-, c)\} \\
    l_3(x, c) &= \max\{0, m + d_{\cos}(x, c) - d_{\cos}(x, x^-)\}
\end{align*}
\]

Variants:

- A combination of the above losses
- Cost-sensitive margin that scales with orthographic distance
Word discrimination results

Average precision (AP)

- DTW
- DTW + learned frame features [Carlin et al. 11]
- Template-based [ASRU 2013]
- CNN classifier [ICASSP 2016]
- CNN Siamese [ICASSP 2016]
- RNN classifier [SLT 2016]
- RNN Siamese [SLT 2016]
- RNN multi-view [ICLR 2017]
Visualization: Character RNN embeddings
Learned jointly with acoustic RNN embeddings
Summary

- Acoustic word embeddings encapsulate a speech segment of arbitrary length with a fixed-dimensional vector
- Can be learned jointly with textual embeddings
- Outperform DTW on word discrimination, query-by-example search in both performance and speed

**Related work:**
- Word discrimination: [Chung+ 2016]
- Spoken term discovery: [Kamper+ 2016]
- Spoken term detection/search: [Chen+ 2015, Audhkasi+ 2017]

**Ongoing/future work:**
- Hierarchical embeddings: structure above/below the word
- Joint acoustic-semantic embeddings, NLP on speech (See next talk!)
- Applications to music, general audio
Thank you!

- **Collaborators:** Wanjia He, Katie Henry, Aren Jansen, Herman Kamper, Keith Levin, Shane Settle, Greg Shakhnarovich, Weiran Wang
- **Support:** NSF, Google
- **Code:** https://github.com/kamperh/, https://github.com/opheadacheh/Multi-view-neural-acoustic-words-embeddings
- **References:**
  - [ICASSP 2016] H. Kamper, W. Wang, and K. Livescu, Deep convolutional acoustic word embeddings using word-pair side information
  - [SLT 2016] S. Settle and K. Livescu, Discriminative acoustic word embeddings: Recurrent neural network-based approaches
  - [ICLR 2017] W. He, W. Wang, and K. Livescu, Multi-view recurrent neural acoustic word embeddings
  - [Interspeech 2017] S. Settle, K. Levin, H. Kamper, and K. Livescu, Query-by-example search with discriminative neural acoustic word embeddings