Acoustic (and acoustically grounded) word embeddings

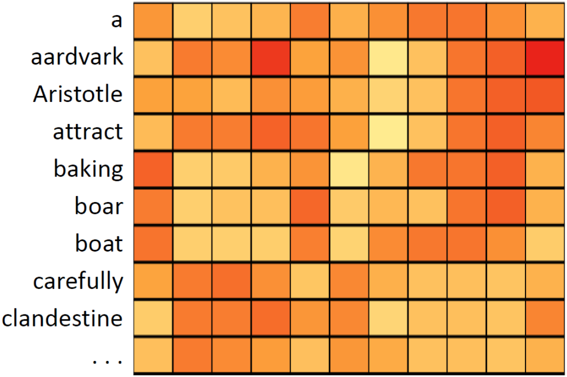
Karen Livescu

IEEE SLT 2018
Workshop on Spoken Language Technology

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

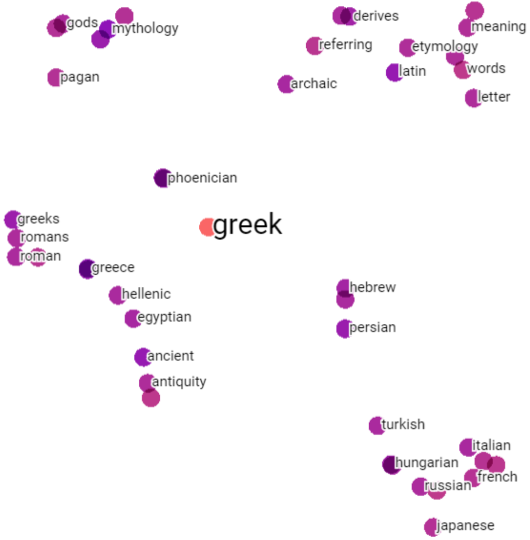
(Written) word embeddings

- Representation of written words as continuous-valued vectors
- Makes it easy to quantify word similarity
- Often used as pretrained parameters in neural models
- Examples: latent semantic analysis, word2vec, GloVe



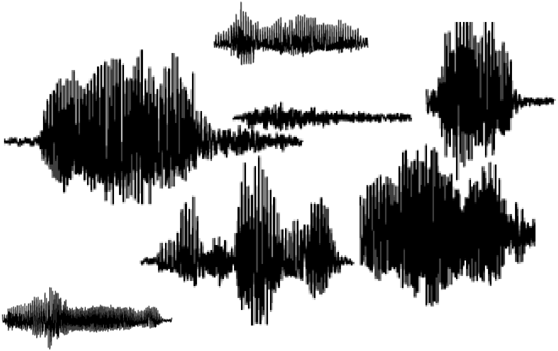
(Written) word embeddings

Usually, we want semantically similar words to have similar vectors



Should we embed spoken words as vectors?

- (-) Speech is already continuous-valued
- (-) Spoken words have lots (a continuum!) of variants
  - Speaking rate, pronunciation variant, speaker, acoustic environment, intonation, fatigue, inebriation...
- (-) So, can't write down a matrix of spoken word embeddings
- (+) But spoken words are hard to compare... vectors are much easier



Talk preview

- There is a growing body of work related to acoustic word embeddings and related ideas
- This talk: Exploration of 3 ideas
  - **Part I**: Acoustic word embeddings
  - **Part II**: Acoustically grounded word embeddings
  - **Part III**: Acoustic-semantic embeddings via visual grounding

Part I:
Acoustic word embeddings

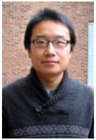
Katie Henry

Aren Jansen

Herman Kamper

Keith Levin

Shane Settle

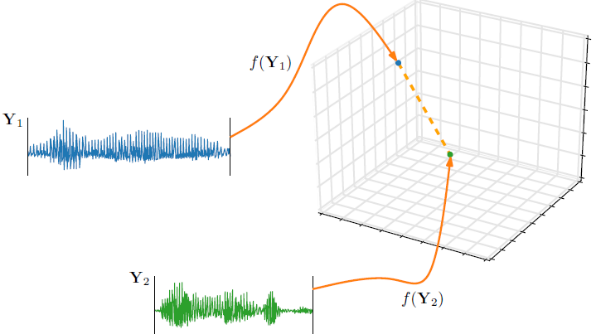


Weiran Wang

[ASRU 2013] Levin, Henry, Jansen, & Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," ASRU 2013
[SLT 2016] Settle & Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," SLT 2016
[Interspeech 2017] Settle, Kamper & Livescu, "Query-by-Example Search with Discriminative Neural Acoustic Word Embeddings," Interspeech 2017

Acoustic word embeddings

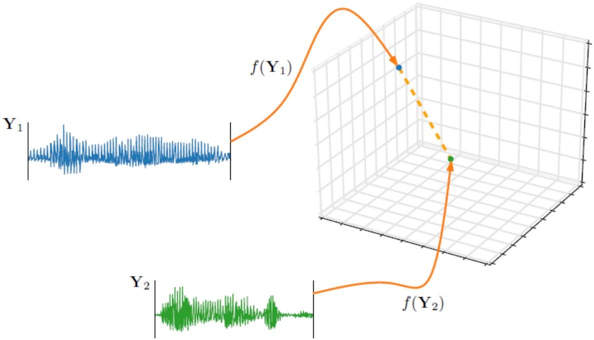
- Computed by a function that maps from a spoken word to a vector
- "Spoken word" = speech signal of arbitrary length corresponding to a word



[Figure credit: Herman Kamper]

What makes a good acoustic word embedding?

- **Same-word** signals should have similar vectors: factor out speaker, acoustic environment, ...
- **Phonetically similar** words should have similar vectors?
- **Semantically similar** words should have similar vectors?

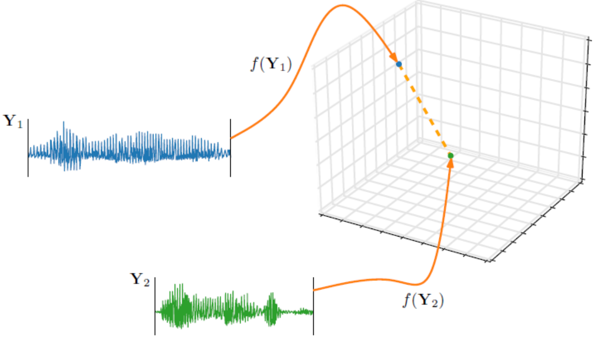


[Figure credit: Herman Kamper]

Applications of acoustic word embeddings

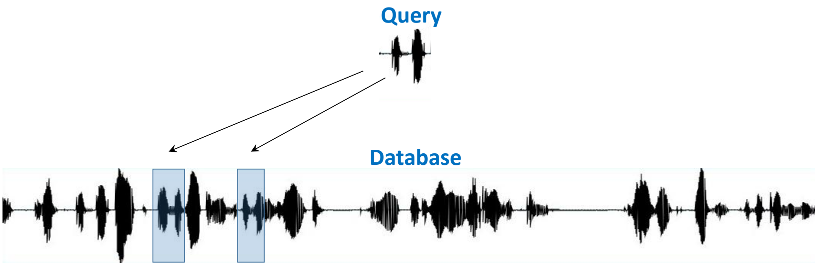
Any task involving similarity between speech segments

- Query-by-example search
- Whole-word speech recognition
- Spoken term discovery



[Figure credit: Herman Kamper]

Query-by-example search

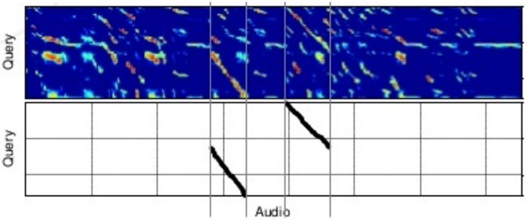


[Figure credit: Herman Kamper]

Applications:

- Open-vocabulary search
- Search in low-resource/unwritten/unknown language data
- Multilingual search

Query-by-example: Classic approach

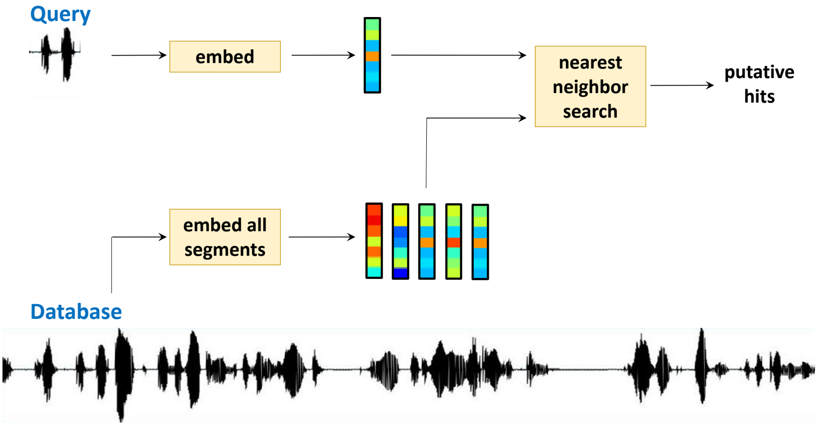


[Figure credit: Proenca et al. 2015]

Dynamic time warping (DTW)

- Slow
- Hard to tune (frame distance function, move costs)
- Sensitive to nuisance variations: noise, speaker, ...
- Hard to learn end-to-end

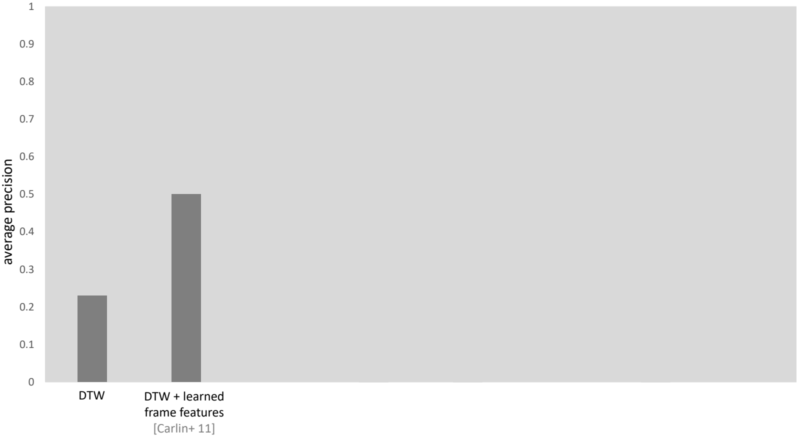
Query-by-example with acoustic word embeddings



An initial task: Word discrimination

Proxy task for query-by-example

- **Input:** Pair of acoustic signals
- **Output:** "Same word" or "different words"
- **Baseline approach:** Threshold the DTW distance
- **Evaluation:** Average precision (AP) over all thresholds
- **Test set:** $\sim 11k$ word segments ($\sim 60M$ pairs)



First embedding approach: Template-based [ASRU 2013]

- Embedding of word segment $\mathbf{X}$ is a vector of distances to a set of other (template) segments $\{\mathbf{R}_1, \ldots, \mathbf{R}_m\}, m \approx 10,000$:

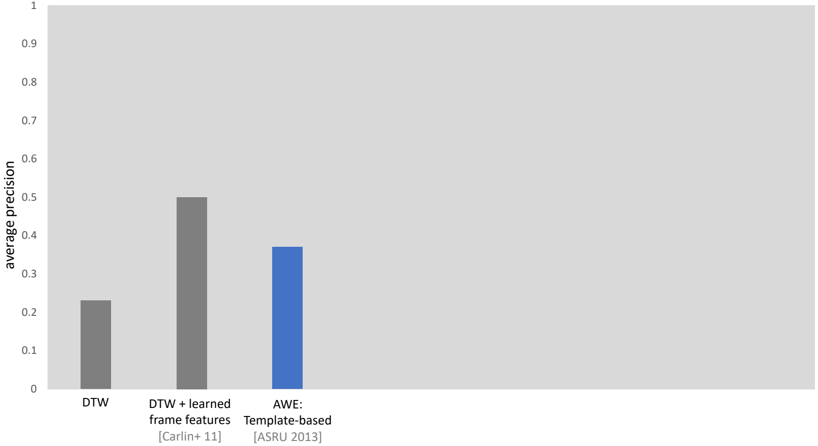
$$f(\mathbf{X}) = [d_{DTW}(\mathbf{X}, \mathbf{R}_1) \ \ldots \ d_{DTW}(\mathbf{X}, \mathbf{R}_m)]$$

- Then (optionally) reduce dimensionality

Word discrimination results

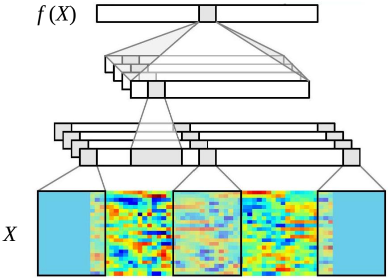
Embedding-based approach: Threshold the cosine distance between the embeddings $d_{\cos}(x_1, x_2) = 1 - \frac{x_1^T x_2}{||x_1|| ||x_2||}$

- Template-based embeddings outperform vanilla DTW
- DTW with learned distance function does better, but requires ~ 200 hours of labeled data



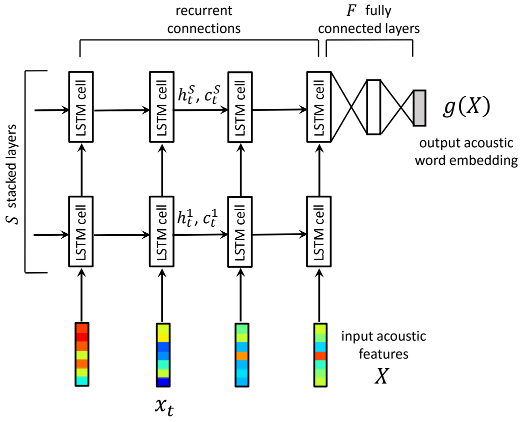
Neural embeddings: CNN-based [ICASSP 2016]

- **Input:** MFCCs, padded to fixed duration
- **Model:** n_{conv} convolutional + n_{full} fully connected layers
- **Embedding** is activation vector of top layer



Neural embeddings: RNN-based [SLT 2016]

- **Input:** MFCCs (without padding)
- **Model:** n_{rec} recurrent + n_{full} fully connected layers
- **Embedding** is activation vector of final fully connected layer



Training objectives

Word classifier log loss

- Add a softmax layer to predict word w
- $l(\mathbf{x}, w) = \log p(w|\mathbf{x})$

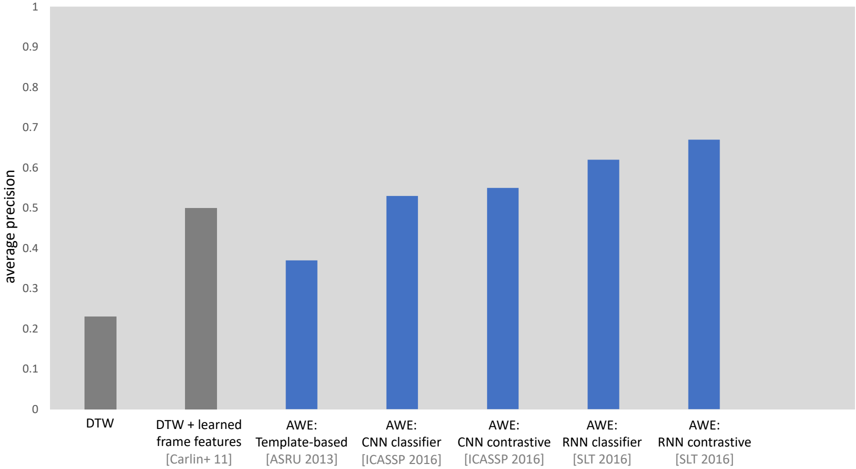
Contrastive (triplet) loss

- Bring together same-word pairs, separate different ones

$$l(\mathbf{x}_1, \mathbf{x}_2) = max\{0, m + d_{\cos}(\mathbf{x}_1, \mathbf{x}_2) - d_{\cos}(\mathbf{x}_1, \mathbf{x}^-)\}$$

  where $\mathbf{x}^- =$ random (or hard) negative example, $m =$ margin
- Weaker supervision (no word labels, only same-word pairs)

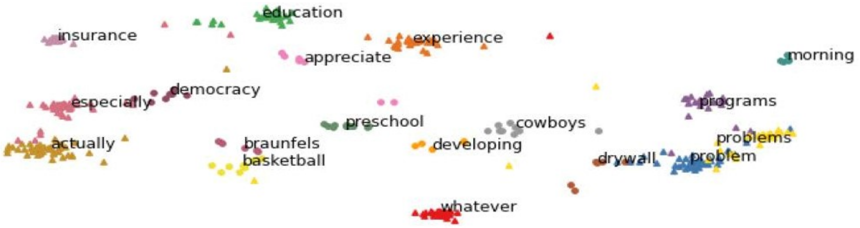
Word discrimination results



Visualization: RNN embeddings

2-dimensional t-SNE embeddings [van der Maaten & Hinton 2008]

△ = word types seen at training time
◯ = not seen at training time



Evaluation on query-by-example

Task: Search for matches to a spoken query in a 433-hour corpus

- **DTW baseline**: 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]

System	**P@10 (↑)**	**Time (s) (↓)**
DTW [Jansen & van Durme 2012]	44.0	24.70
Template-based [Levin+ 15]	34.5	**0.08**
RNN AWE (contrastive) [Interspeech 2017]	**60.2**	0.38

Related work

Autoencoder-based embeddings

- [Y.-A. Chung+ Interspeech 2016, Y.-H. Wang+ ICASSP 2018, C.-H. Shen+ ICASSP 2018]
- [Audhkhasi+ ICASSP 2017]

Unsupervised embeddings for spoken term discovery and unsupervised speech recognition

- [Kamper+ SLT 2014, Interspeech 2015, CSL 2017, arXiv 2018]

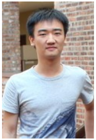
Acoustic word embeddings for segmental speech recognition

- [Maas+ ICML WRL 2012, Bengio & Heigold ICASSP 2014]

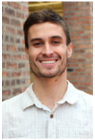
Future work: More comparisons among embedding approaches

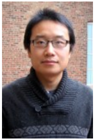
Part II:
Acoustically grounded word embeddings


Kartik Audhkhasi


Wanjia He


Michael Picheny


Shane Settle


Weiran Wang

[ICLR 2017] He, Wang, & Livescu, "Multi-view recurrent acoustic word embeddings," ICLR 2017

Joint learning of acoustic + written embeddings [ICLR 2017]

Motivation:

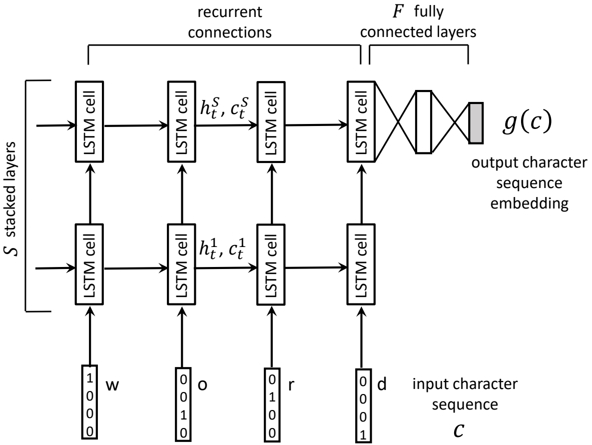
- Learn better acoustic embeddings by relating them to a written character sequence
- Some tasks involve "distances" between speech segments and written words
  - Spoken term detection ("Query-by-text")
  - Automatic speech recognition

Approach: Learn a pair of RNN-based embedding functions

- Acoustic word embedding (speech $\rightarrow$ vector)
- Acoustically grounded word embedding (character sequence $\rightarrow$ vector)

"Barack Obama" $\stackrel{?}{=}$ 

Character RNN-based acoustically grounded word embedding



Joint learning of acoustic and acoustically grounded word embeddings

Given a matched (acoustic, written) word pair $(\mathbf{x}, \mathbf{c})$

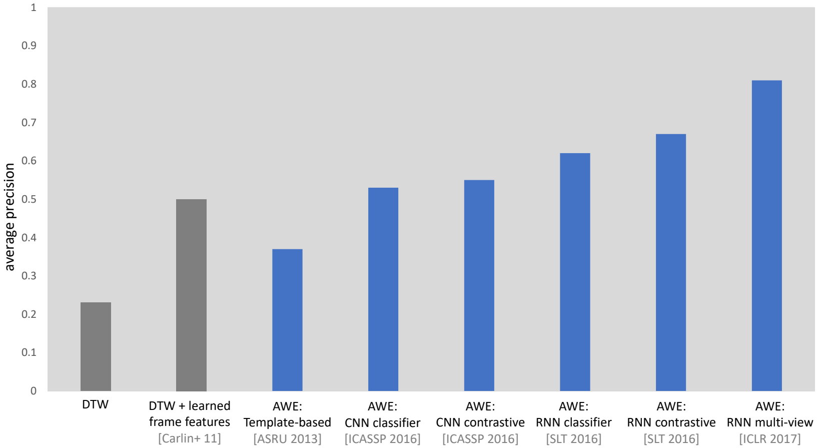
$$
\begin{aligned}
l_0(\mathbf{x}, \mathbf{c}) &= max\{0, m + d_{\cos}(\mathbf{x}, \mathbf{c}) - d_{\cos}(\mathbf{x}, \mathbf{c}^-)\} \\
l_1(\mathbf{x}, \mathbf{c}) &= max\{0, m + d_{\cos}(\mathbf{x}, \mathbf{c}) - d_{\cos}(\mathbf{c}^-, \mathbf{c})\} \\
l_2(\mathbf{x}, \mathbf{c}) &= max\{0, m + d_{\cos}(\mathbf{x}, \mathbf{c}) - d_{\cos}(\mathbf{x}^-, \mathbf{c})\} \\
l_3(\mathbf{x}, \mathbf{c}) &= max\{0, m + d_{\cos}(\mathbf{x}, \mathbf{c}) - d_{\cos}(\mathbf{x}, \mathbf{x}^-)\}
\end{aligned}
$$

Variants:

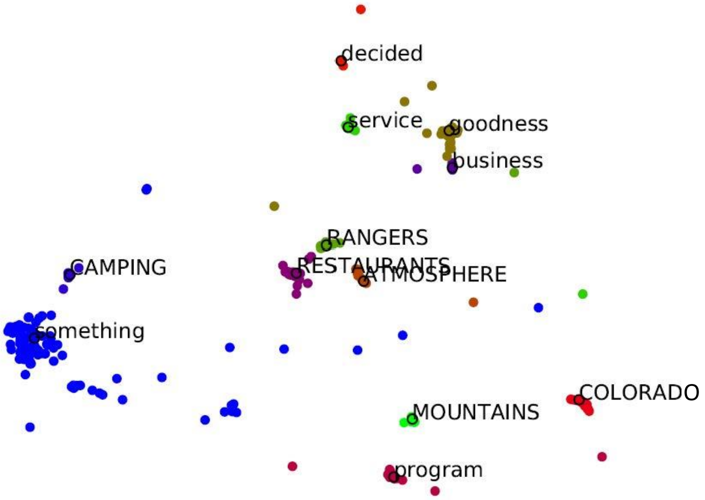
- Weighted combination of these losses
- Cost-sensitive margin that scales with orthographic distance

Word discrimination results

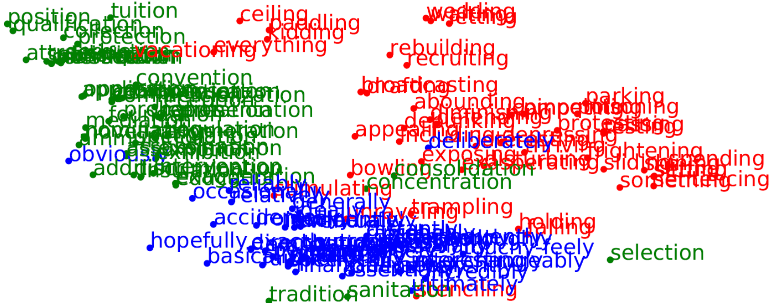
(Using just the acoustic word embeddings)



Visualization of acoustically grounded word embeddings

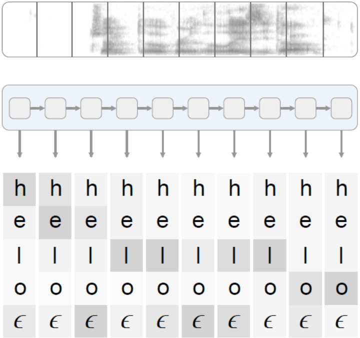


Visualization of acoustically grounded word embeddings



Acoustically grounded word embeddings for speech recognition

- Ongoing work with Shane Settle, Kartik Audhkhasi (IBM), Michael Picheny (IBM)
- Background: Connectionist temporal classification (CTC) [Graves+ 2006]



[Figure credit: https://distill.pub/2017/ctc/]

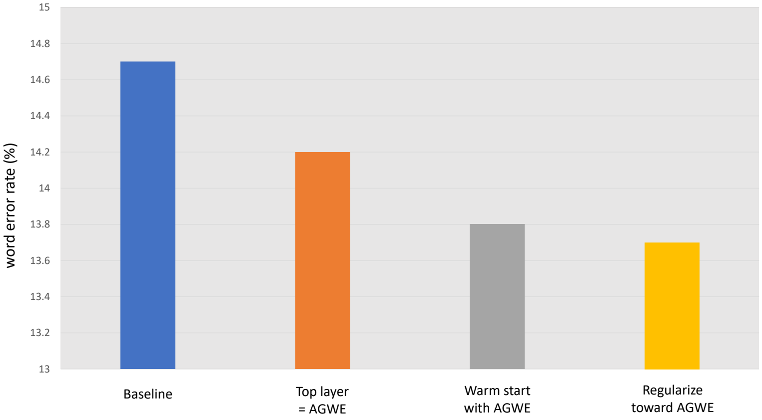
Background: Whole-word CTC

Several groups have started studying whole-word ASR

- Output labels are whole words (no typos to fix)
- Now the final layer weights represent a word embedding matrix
- Many rare words $\Longrightarrow$ many rows are learned very poorly
- **Idea:** Use pre-trained acoustically grounded word embeddings

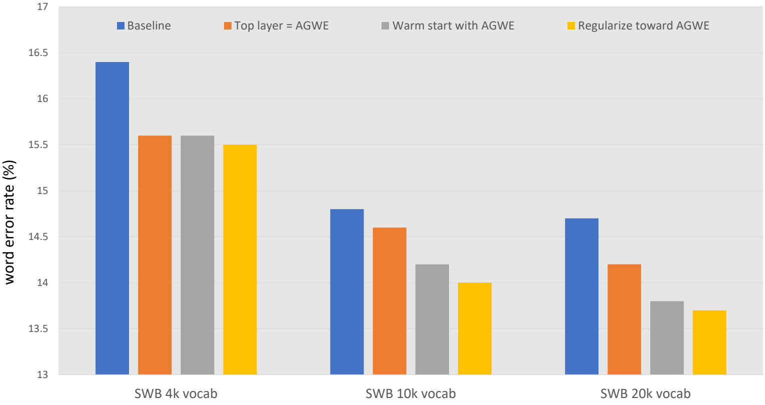
Improving ASR with acoustically grounded word embeddings

Switchboard conversational telephone speech recognition:



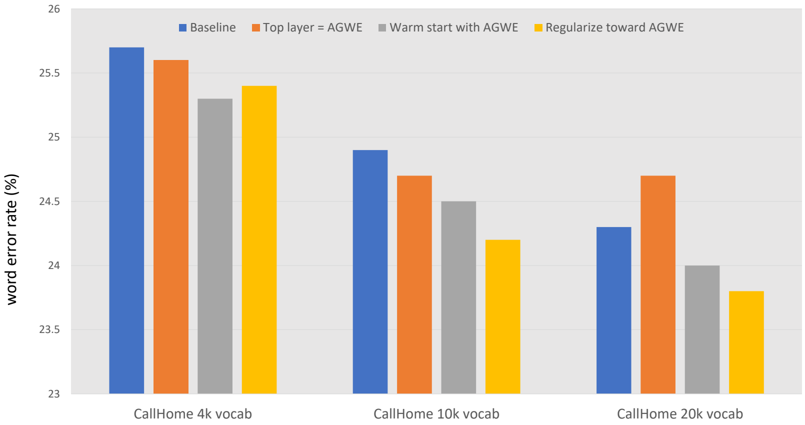
Improving ASR with acoustically grounded word embeddings

Switchboard conversational telephone speech recognition:



Improving ASR with acoustically grounded word embeddings

CallHome conversational telephone speech recognition (slight domain mismatch, and more speaker mismatch):



Related work

Character sequence autoencoders for spoken term detection

- [Audhkhasi+ ICASSP 2017]

Phonetically oriented word embeddings for ASR error detection

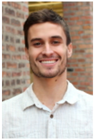
- [Ghannay+ ACL WEVSRNLP 2016, Interspeech 2016]

Jointly learned acoustic and acoustically grounded word embeddings for segmental speech recognition

- [Bengio & Heigold ICASSP 2014]

Part III: Acoustic-semantic embeddings via visual grounding

Herman Kamper



Shane Settle

Greg Shakhnarovich

[Interspeech 2017] Kamper, Settle, Livescu, and Shakhnarovich "Visually grounded learning of keyword prediction from untranscribed speech," Interspeech 2017.

[TASLP 2018] Kamper, Livescu, and Shakhnarovich "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE/ACL TASLP* 2018.

Acoustic-semantic embeddings

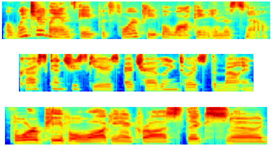
- Thus far: Embeddings that represent (mostly) acoustic-phonetic information
- What about acoustic embeddings that represent **meaning**?
- Useful for semantic search, speech understanding, ...
- One possibility: extend text embedding approaches to speech [Chung & Glass Interspeech 2018, Palaskar & Metze arXiv 2018, Y.-C. Chen+ SLT 2018]
- More challenging than text embedding learning
  - Less speech data available than text
  - Speech data is more computationally demanding (1 text "frame" $\approx$ 500 speech frames)
- Can we use some weaker semantic supervision to learn from less data?

Images as weak semantic labels for speech

We use images as **weak labels** to learn semantic embeddings

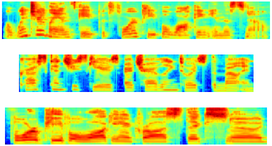
- Data set from [Harwath & Glass ASRU 2015]
- (Slightly different setting from before: We will learn **whole-utterance** embeddings)
- Won't compare directly to other acoustic-semantic approaches

Play



Images as weak semantic labels for speech

Play



What can we hope to learn from such data?

- Off-the-shelf image taggers work pretty well! Use one to get labels!
- Probably can't learn a complete speech recognizer this way
- But maybe learn to predict keywords?

Related work

Joint acoustic-visual embeddings

- [Harwath & Glass ASRU 2015, ACL 2017; Harwath+ NIPS 2016, ACL 2017; Leidal+ ASRU 2017; Harwath PhD Dissertation 2018]
- [Gelderloos & Chrupala COLING 2016; Chrupala+ ACL 2017]

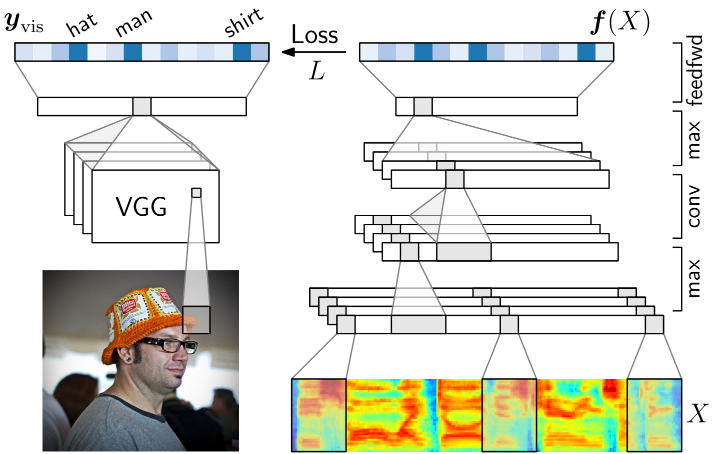
Linguistic unit discovery from multi-modal inputs in unwritten languages

- [Scharenborg+ ICASSP 2018]

Main difference from related work: We use visual taggers to produce **weak textual labels** to enable text-mediated tasks

Visually grounded keyword prediction

Idea: Use an image tagger to get soft textual labels [Kamper+ 17]

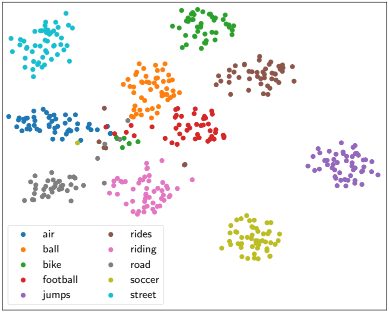


[Figure credit: Herman Kamper]

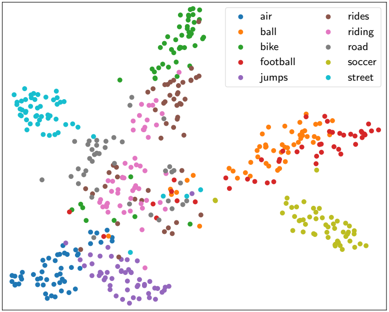
Keyword prediction examples

Input utterance	Predicted BoW labels
Play man on bicycle is doing tricks in an old building	**bicycle**, bike, **man**, riding, wearing
a little girl is climbing a ladder	child, **girl**, **little**, young
a rock climber standing in a crevasse	climbing, man, **rock**
a dog running in the grass around sheep	**dog**, field, **grass**, **running**
a man in a miami basketball uniform looking to the right	ball, **basketball**, **man**, player, **uniform**, wearing

Visually grounded embeddings are more semantic



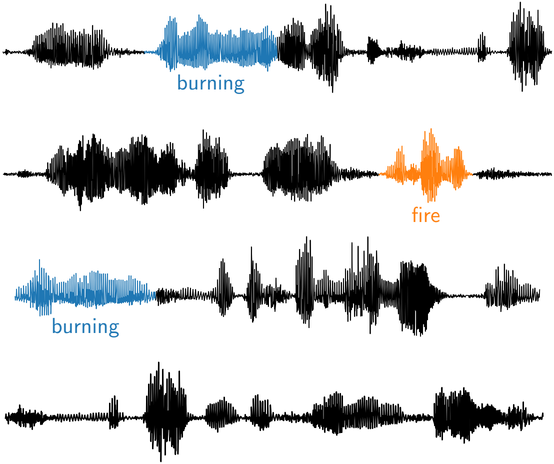
(a) Text-supervised



(b) Visually grounded

Task: Semantic speech retrieval

Written query:
burning



[Figure credit: Herman Kamper]

Semantic speech retrieval evaluation

Training

- **Data:** 8000 images with 5 spoken captions each ($\sim$37 hours of speech) [Harwath & Glass ASRU 2015]
- **Weak labels:** From image tagger trained on external data (Flickr30k + MSCOCO)

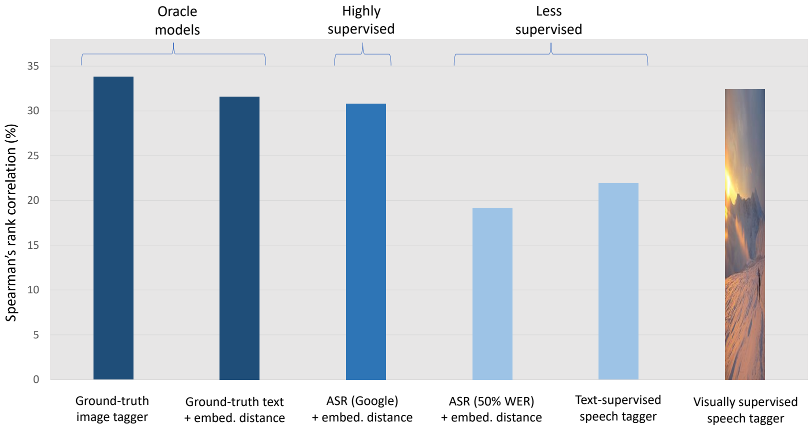
Testing

- **Prediction:** Output words w where $f_w(X) > \alpha$
- **Evaluation:** Use the predicted words for semantic speech search, and measure typical search performance metrics (P@10, P@N, EER, AP, Spearman's ρ)
- **Ground truth:** Human (MTurk) judgments

Semantic speech retrieval evaluation

In terms of correlation with human scores:

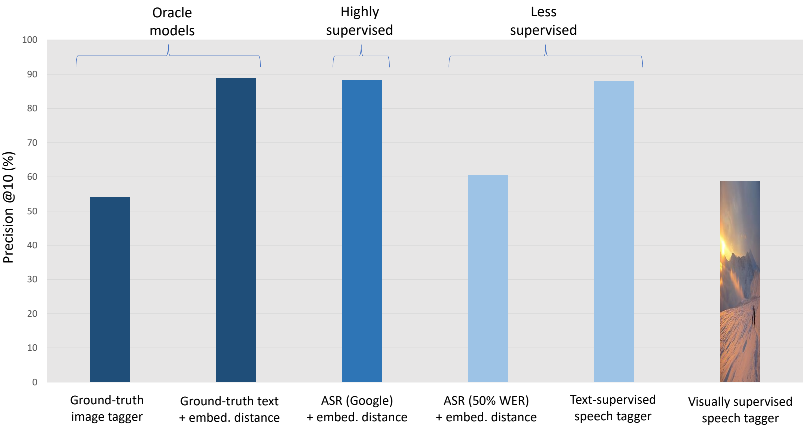
- Visually grounded model performs about as well as oracle models
- Much better than text-supervised model



Semantic speech retrieval evaluation

In terms of Precision @10:

- Visually grounded model performs about as well as 50% WER speech recognizer and ground-truth image tagger
- Main benefit of visually grounded model: Finding non-exact matches



Summary

3 ideas

- Acoustic word embeddings that respect phonetic similarity
- Acoustic word embeddings that respect semantic similarity
- Acoustically grounded (written) word embeddings that respect phonetic similarity

Ongoing/future work

- Joint acoustic-semantic embeddings for NLP on speech
- Hierarchical embeddings: structure above/below the word
- More thorough comparisons among approaches

Thanks!