
# Learning Word Embeddings from Speech

NIPS Workshop on Machine Learning for Audio Signal Processing

December 8, 2017

Yu-An Chung  James Glass

MIT Computer Science and Artificial Intelligence Laboratory

Cambridge, MA

# Outline

- **Motivation**
- Proposed Approach
- Experiment
- Conclusion

# Motivation

- GloVe and word2vec transform words into **fixed dimensional** vectors.
  - Obtained by unsupervised learning from co-occurrences information in the text
  - Contain semantic information about the word

- Humans learn to speak before they can read or write.
- Machines can learn semantic word embeddings from raw text.

Can machines learn semantic word embeddings
from ***speech*** as well?

## Text (written language)

Audio signal processing is currently undergoing a paradigm change, where data-driven machine learning is replacing hand-crafted feature design. This has led some to ask whether audio signal processing is still useful in the era of machine learning.

Input

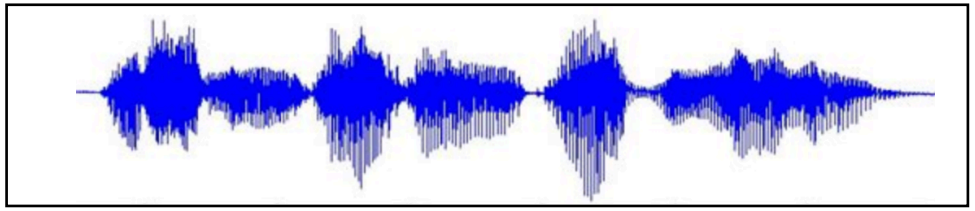
**Learning system**
such as GloVe and word2vec

Output

**Word embeddings**

audio
signal
processing
⋮
learning

## Speech (spoken language)



Input

**Learning system**
**our goal**

Output

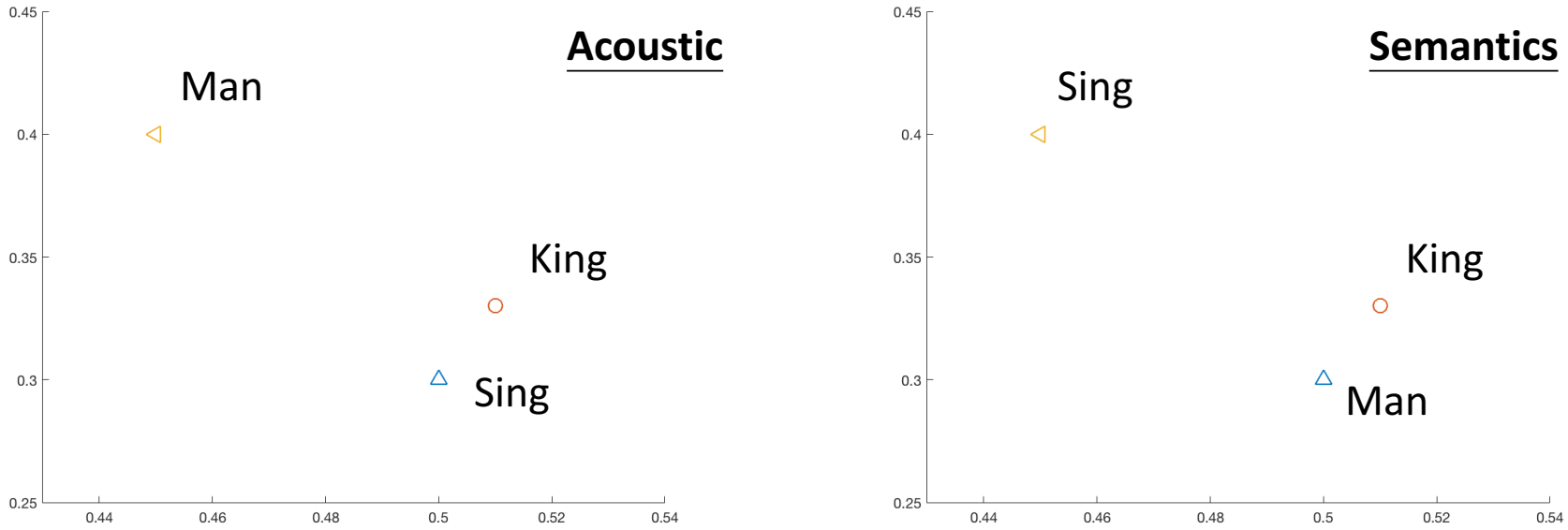
**Word embeddings**

audio
signal
processing
⋮
learning

# Acoustic Word Embeddings

**We aim to learn embeddings that capture semantic information rather than acoustic-phonetic structure!**

- Also learn fixed-length vector representations (embeddings) from speech
  - Audio segments that sound alike would have embeddings nearby in the space.
  - Capture ***phonetic*** structure




References:
[1] Multi-view recurrent neural acoustic word embeddings. He et al., ICLR 2017
[2] Discriminative acoustic word embeddings: Recurrent neural network-based approaches. Settle and Livescu, SLT 2016
[3] Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. Chung et al., Interspeech 2016
[4] Deep convolutional acoustic word embeddings using word-pair side information. Kamper et al., ICASSP 2016
[5] Word embeddings for speech recognition. Bengio and Heigold, Interspeech 2014

# Outline

- Motivation
- **Proposed Approach**
- Experiment
- Conclusion

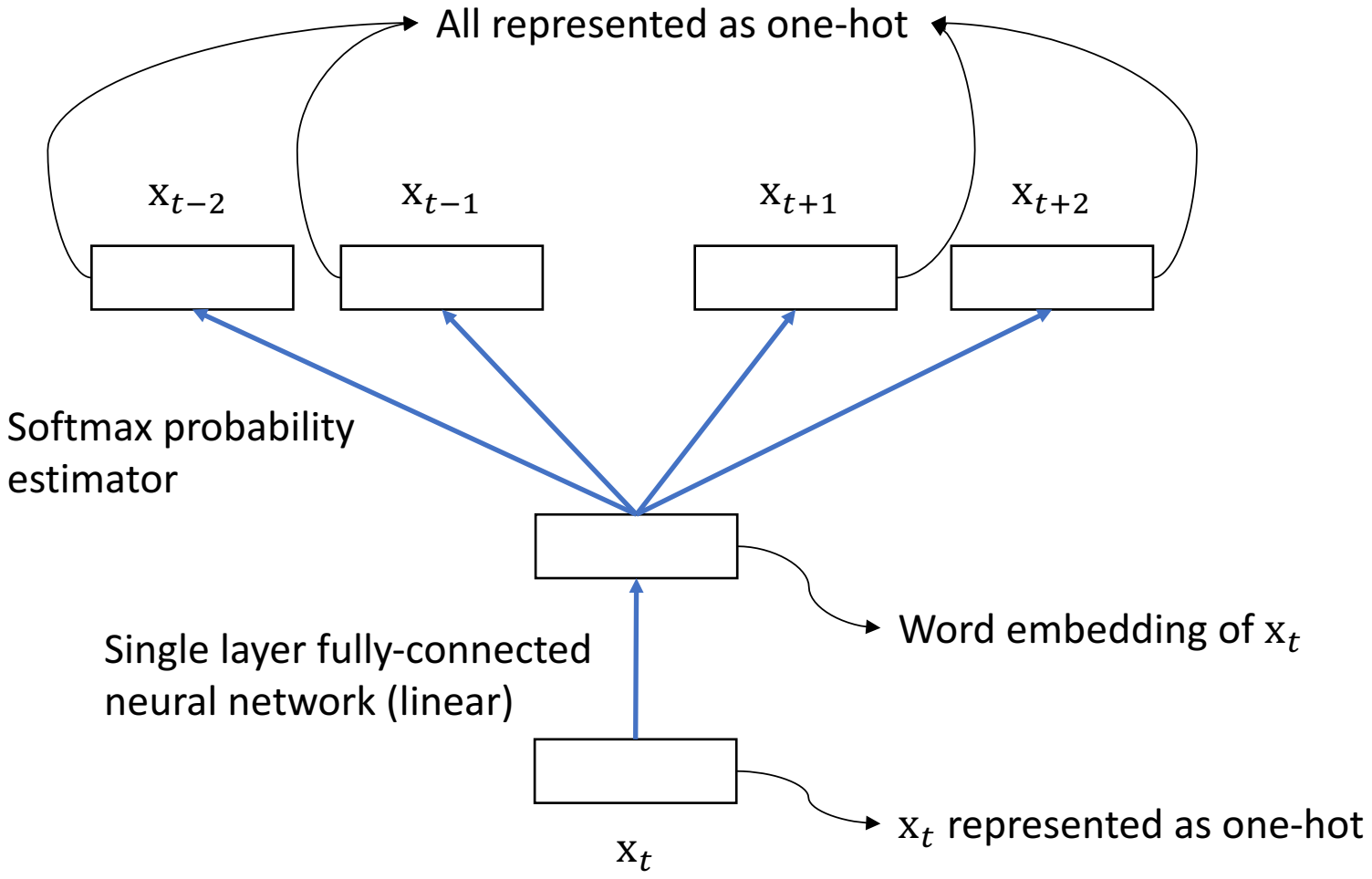
# Our approach is inspired by word2vec (skip-gram)

**Text**

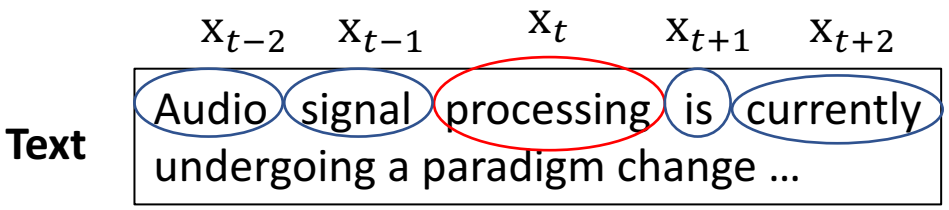
$x_{t-2}$ $x_{t-1}$ $x_t$ $x_{t+1}$ $x_{t+2}$

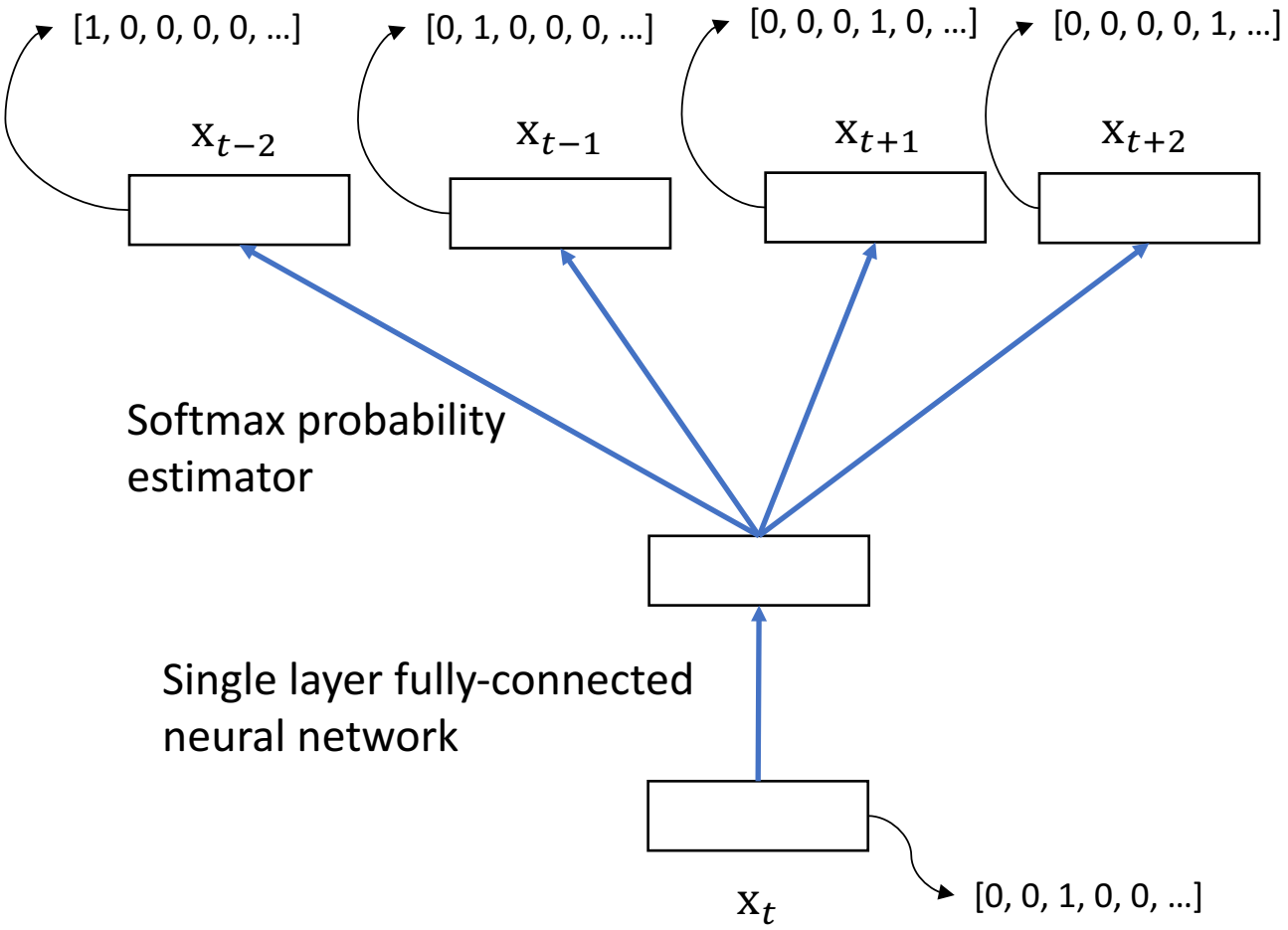
Audio signal processing is currently undergoing a paradigm change …

Window size = 2

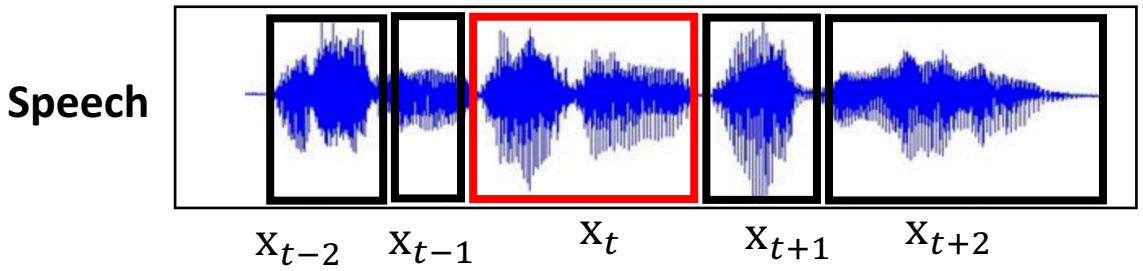


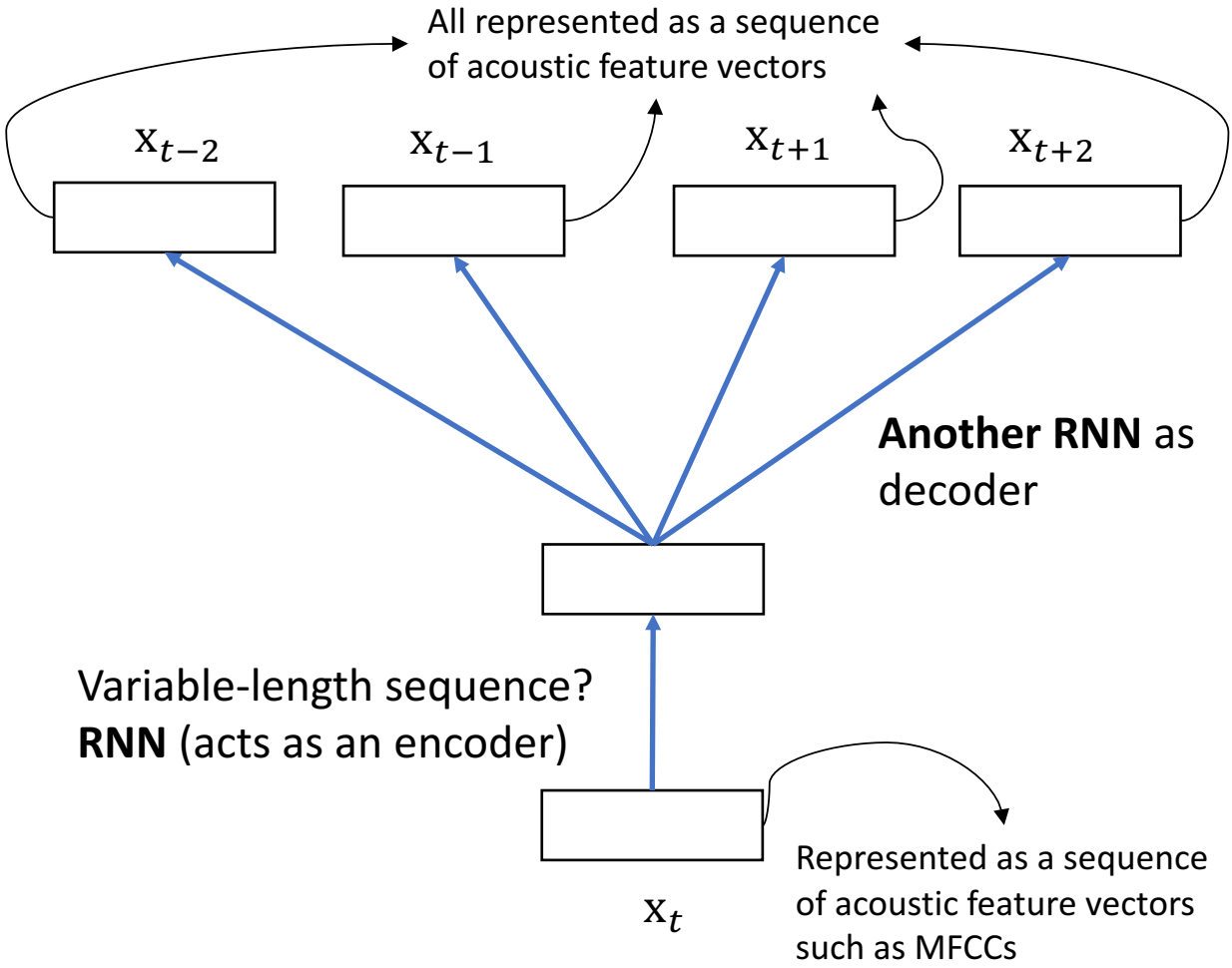
# Word2Vec (skip-gram) for text



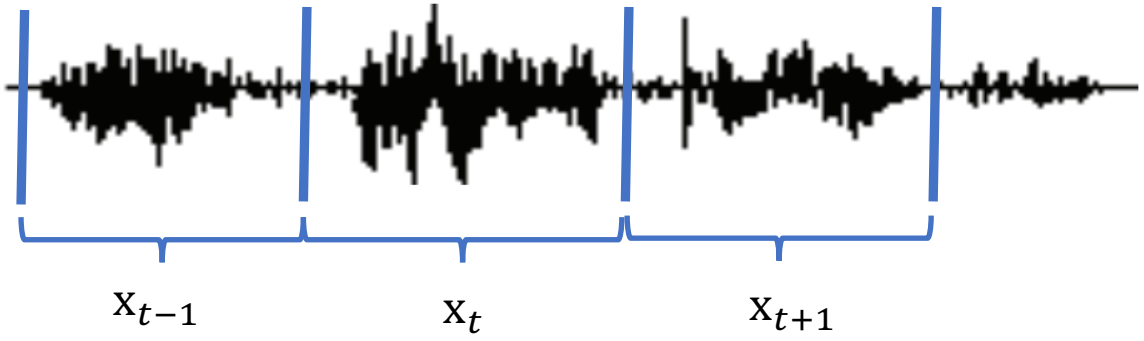




# Our approach



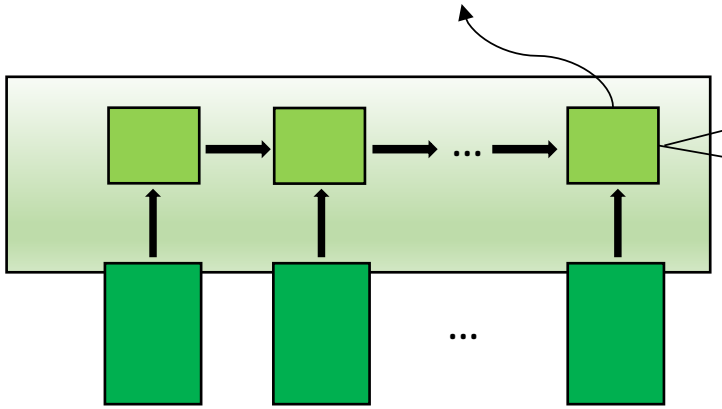
**Speech**




Here window size = 1

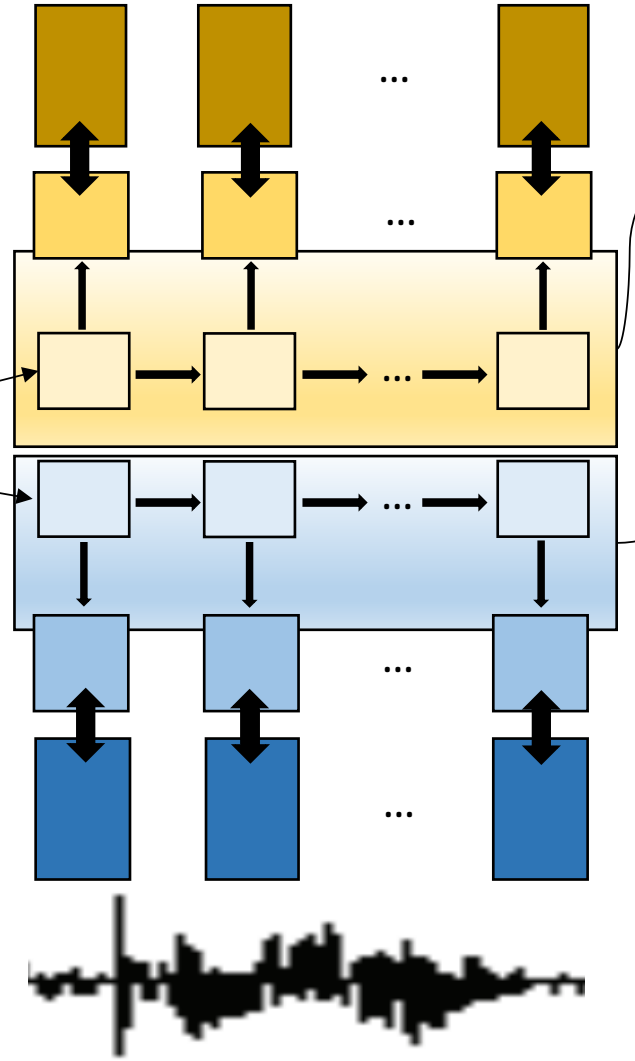
**Learned word embedding of $x_t$**

Encoder RNN



$x_t$ =

$x_{t-1}$ =




$x_{t+1}$ =

# Outline

- Motivation
- Proposed Approach
- **Experiment**
- Conclusion

# Corpus & Model Architecture

- LibriSpeech - a large corpus of read English speech (500 hours)
- Acoustic features consisted of 13-dim MFCCs produced every 10ms
- Corpus was segmented via forced alignment
  - Word boundaries were used for training our model

- Encoder RNN is a 3-layer LSTM with 300 hidden units (dim = 300)
- Decoder RNN is a single-layer LSTM with 300 hidden units

# Task: 13 word similarity benchmarks

- The 13 benchmarks contain different numbers of pairs of English words that have been assigned similarity ratings by humans.

- Each benchmark evaluate the word embeddings in terms of different aspects, e.g.,
  - RG-65 and MC-30 focus on nouns
  - YC-130 and SimVerb-3500 focus on verbs
  - Rare-Word focuses on rare words

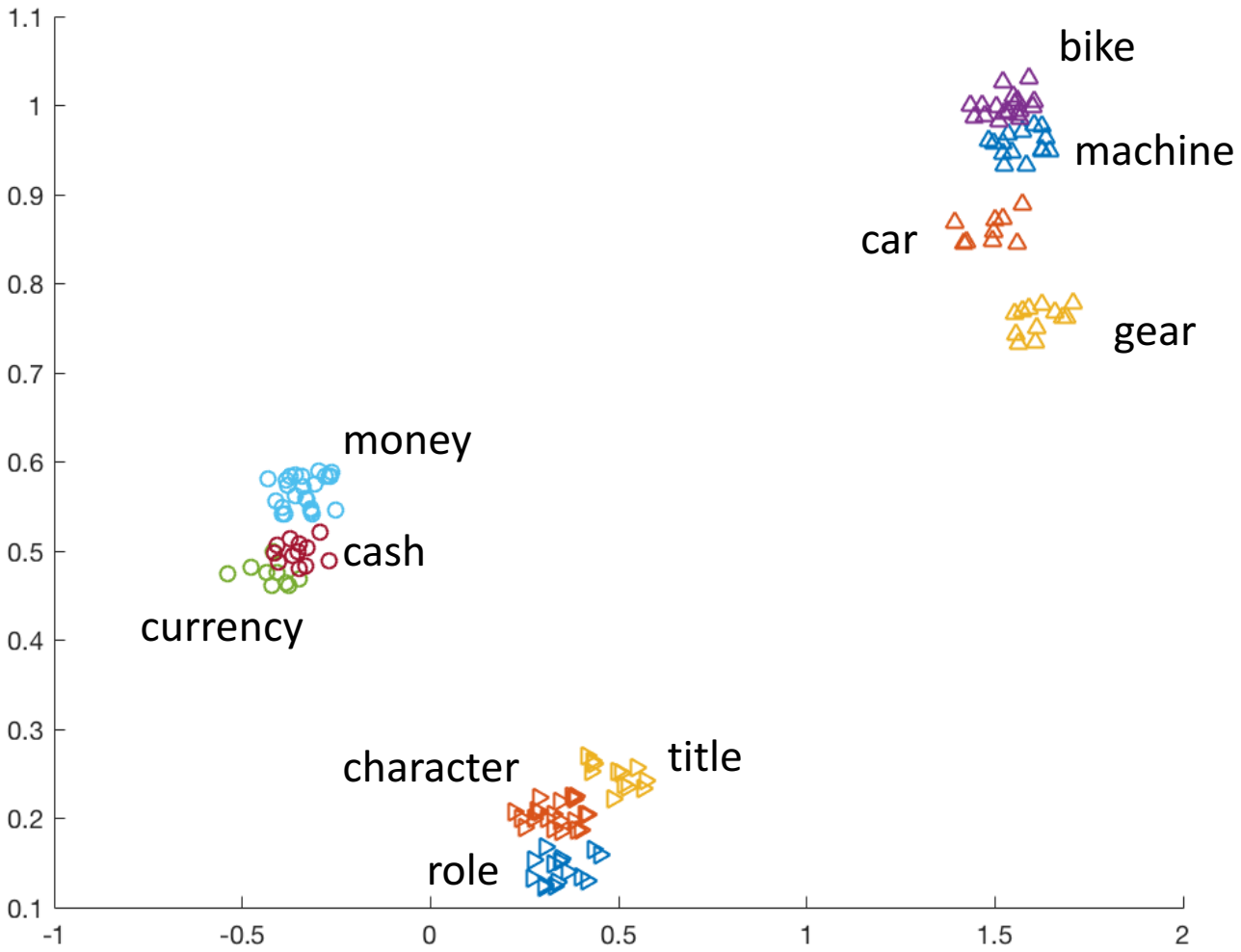
- Spearman's rank correlation coefficient $\rho$ between the rankings produced by the model against the human rankings (the higher the better)

- Embeddings representing the audio segments of the same word were averaged to obtain one single 300-dim vector/word

# Experimental Results

| No. | Dataset | #(word pairs) | **Our model** | | GloVe Wikipedia 2014 | |
|---|---|---|---|---|---|---|
| | | | #(not found) | $\rho$ | #(not found) | $\rho$ |
| 1 | WS-353 | 353 | 21 | 0.5324 | 0 | 0.6054 |
| 2 | WS-353-REL | 252 | 12 | 0.4959 | 0 | 0.5725 |
| 3 | WS-353-SIM | 203 | 7 | 0.5842 | 0 | 0.6638 |
| 4 | MC-30 | 30 | 0 | 0.6647 | 0 | 0.7026 |
| 5 | RG-65 | 65 | 0 | 0.7274 | 0 | 0.7662 |
| 6 | Rare-Word | 2034 | 783 | 0.3158 | 252 | 0.4118 |
| 7 | MEN | 3000 | 122 | 0.6877 | 0 | 0.7375 |
| 8 | MTurk-287 | 287 | 13 | 0.5647 | 0 | 0.6332 |
| 9 | MTurk-771 | 771 | 22 | 0.6010 | 0 | 0.6501 |
| 10 | YP-130 | 130 | 0 | 0.5173 | 0 | 0.5613 |
| 11 | SimLex-999 | 999 | 0 | 0.2985 | 0 | 0.3705 |
| 12 | Verb-143 | 144 | 0 | 0.2877 | 0 | 0.3051 |
| 13 | SimVerb-3500 | 3500 | 126 | 0.2023 | 2 | 0.2267 |

# t-SNE Visualization



# Impressive, but why still worse than GloVe?

1. Different speech and text training data (LibriSpeech vs. Wikipedia)

2. Inherent variability in speech production - unlike textual data, every instance of any spoken word ever uttered is different:
   - Different speakers
   - Different speaking styles
   - Environmental conditions
   - Just name a few of the major influences on a speech recording

# Outline

- Motivation
- Proposed Approach
- Experiment
- **Conclusion**

# Conclusion

- We proposed a model for learning semantic word embeddings from speech:
  - Mimics the architecture of the textual skip-gram word2vec
  - Uses two RNNs to handle variable-length input and output sequences
- Showed impressive results (not too worse than GloVe trained on Wikipedia) on word similarity tasks
- Verified that machines are capable of learning semantics word embeddings from speech!

# Future Works

1. Assuming perfect word boundaries is unrealistic - try train the model on very likely imperfect segments obtained by existing segmentation methods

2. Overcome speech recording issues - try remove the speaker information

3. Compare with word2vec/GloVe trained on LibriSpeech transcriptions

4. Evaluate the word embeddings on downstream applications - their effectiveness on real tasks is actually what we care

Thank you!