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

Shane Settle

Kartik Audhkhasi, Karen Livescu, Michael Picheny

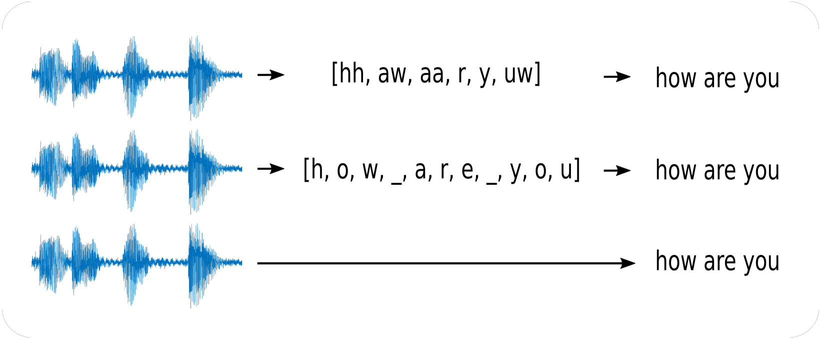
ICASSP 2019

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

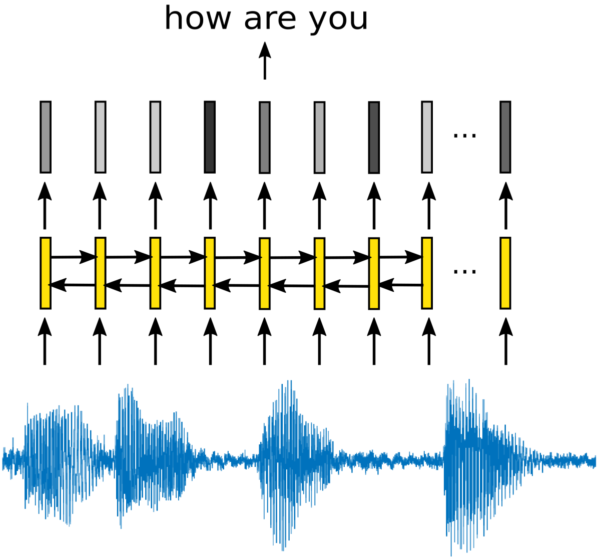
IBM Research AI

Models for Speech Recognition

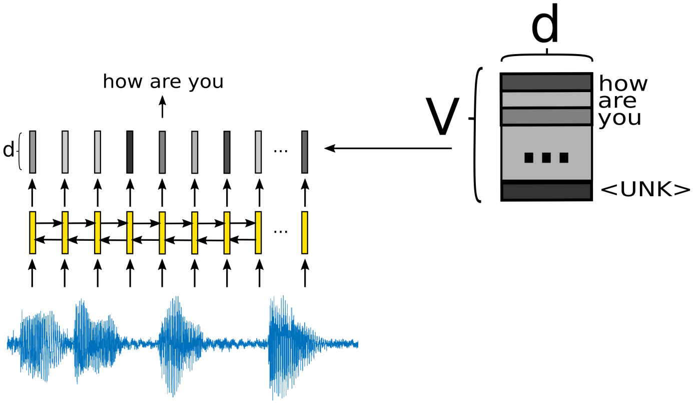
- Traditional models for speech recognition are sub-word based
- Acoustics-to-word (A2W) models directly map input acoustic features to words without the need for additional decoding [Soltau+, Audhkhasi+ 2017] [Audhkhasi+, Li+, Yu+ 2018]



Acoustics-to-Word (A2W) Models for Speech Recognition

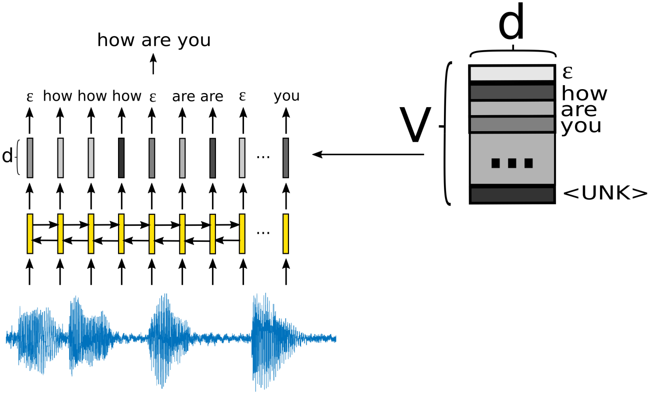


Acoustics-to-Word (A2W) Models for Speech Recognition



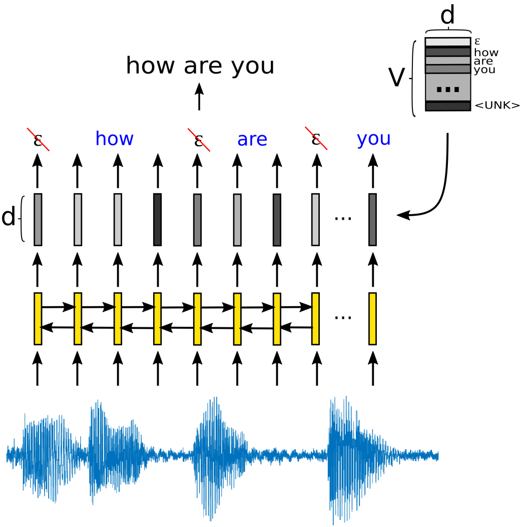
Acoustics-to-Word (A2W) Models for Speech Recognition

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



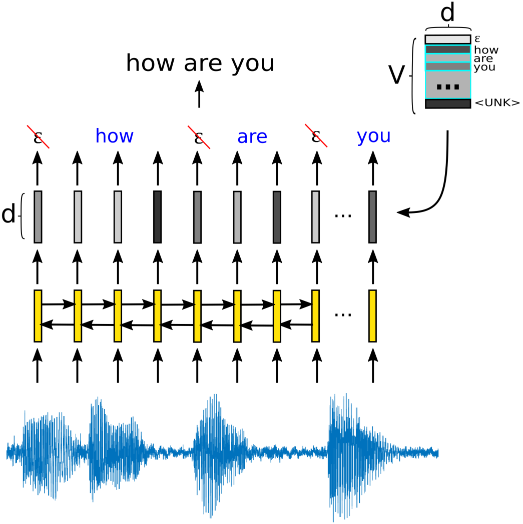
Acoustics-to-Word (A2W) Models for Speech Recognition

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



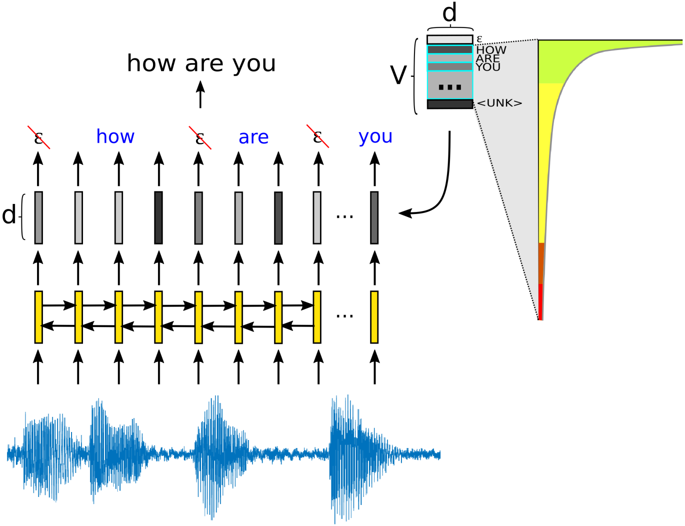
Acoustics-to-Word (A2W) Models for Speech Recognition

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

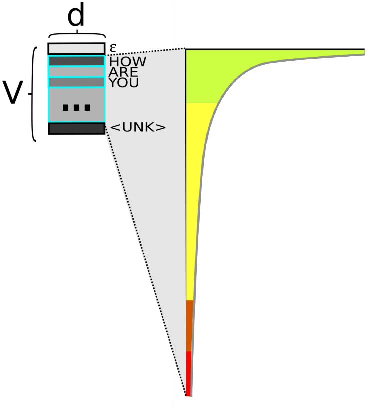
Connectionist Temporal Classification (CTC) [Graves+ 2006] resolves input/target length disparity to allow for frame-wise prediction.



Acoustically Grounded Word Embedding Motivation

While prior work [Soltau+ 2018] matches sub-word performance training on 125Khrs, a gap remains for smaller datasets:

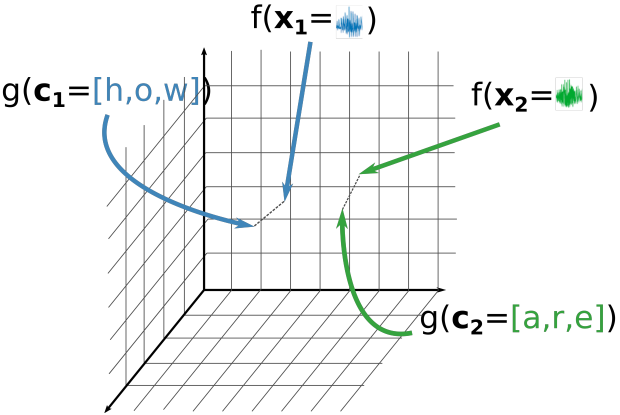
- difficulty learning rare/infrequent words
- out-of-vocabulary words



Idea: Use pre-trained acoustically grounded word embeddings to improve quality of the learned word embedding matrix

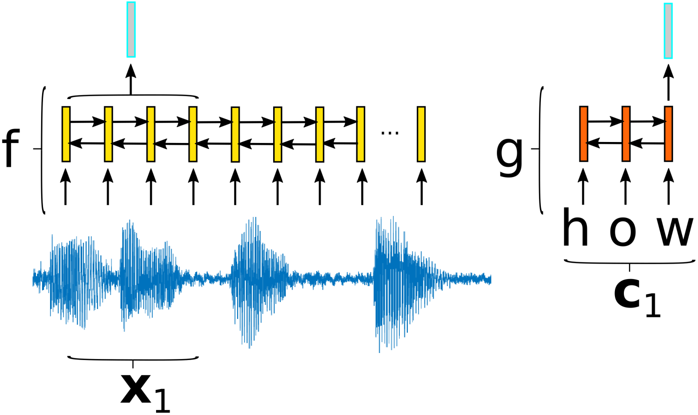
Acoustically Grounded Word Embeddings (AGWE)

Given (*acoustic*, *character*) word pairs $(\mathbf{x}, \mathbf{c})$, we train embedding functions $f(\cdot)$ and $g(\cdot)$ to learn mappings into a shared space:



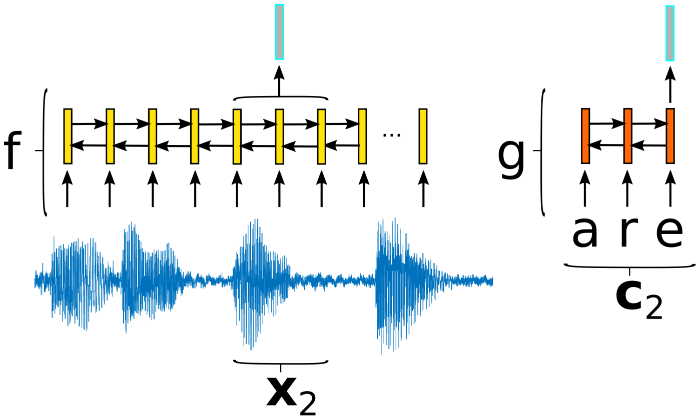
Acoustically Grounded Word Embeddings (AGWE)

Given (*acoustic*, *character*) word pairs $(\mathbf{x}, \mathbf{c})$, we train embedding functions $f(\cdot)$ and $g(\cdot)$ to learn mappings into a shared space:



Acoustically Grounded Word Embeddings (AGWE)

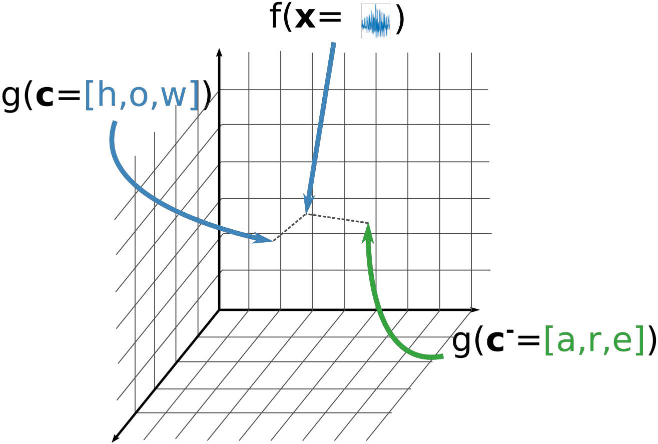
Given (*acoustic*, *character*) word pairs $(\mathbf{x}, \mathbf{c})$, we train embedding functions $f(\cdot)$ and $g(\cdot)$ to learn mappings into a shared space:



Acoustically Grounded Word Embeddings (AGWE)

Most offending different character sequence: [He+ 2017]

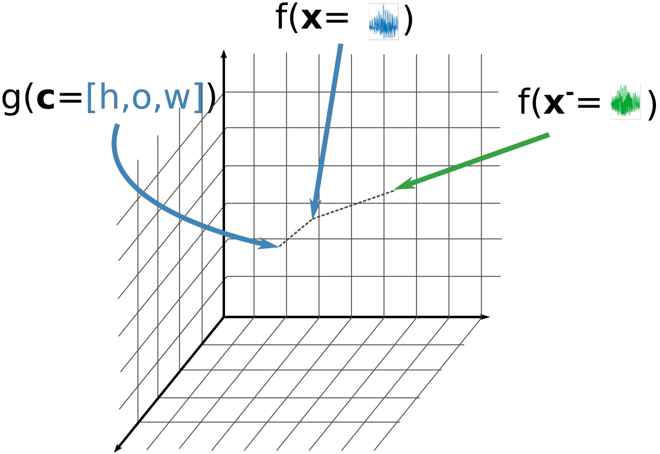
$$\max \left\{ 0, m + d_{\cos}(f(\mathbf{x}), g(\mathbf{c})) - \min_{\mathbf{c}^- \neq \mathtt{char}(\mathbf{x})} d_{\cos}(f(\mathbf{x}), g(\mathbf{c}^-)) \right\}$$



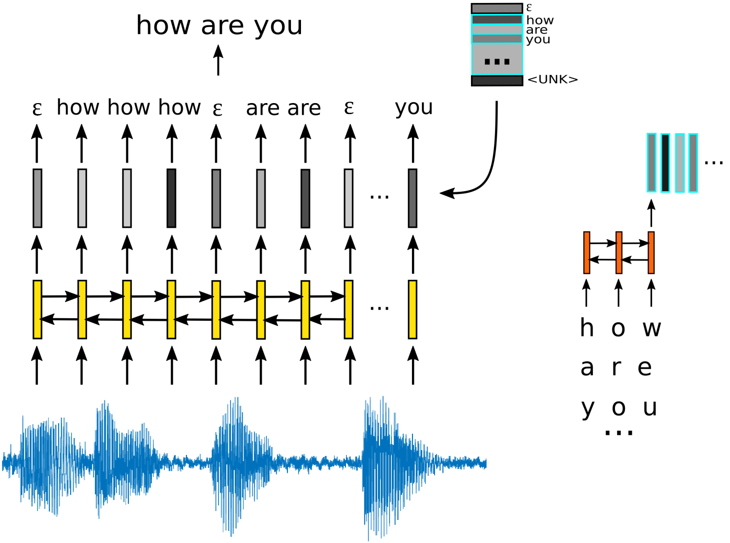
Acoustically Grounded Word Embeddings (AGWE)

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

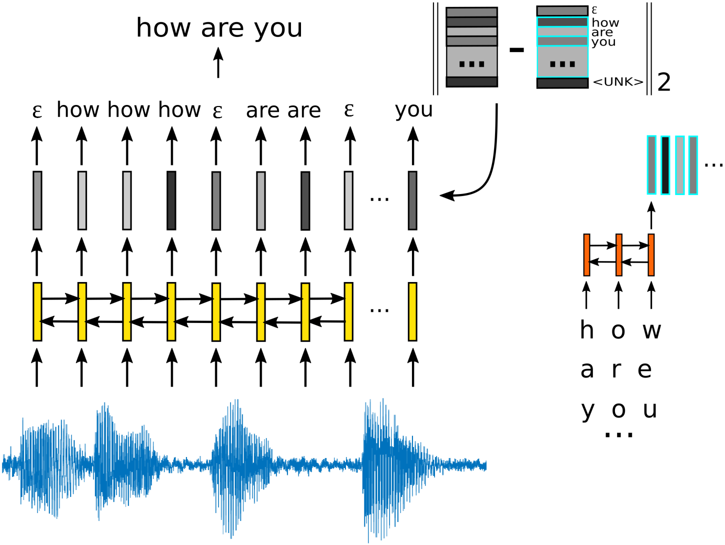
$$\max\left\{0, m + d_{\cos}(g(\mathbf{c}), f(\mathbf{x})) - \min_{\mathtt{char}(\mathbf{x}^-)\neq\mathbf{c}} d_{\cos}(g(\mathbf{c}), f(\mathbf{x}^-))\right\}$$



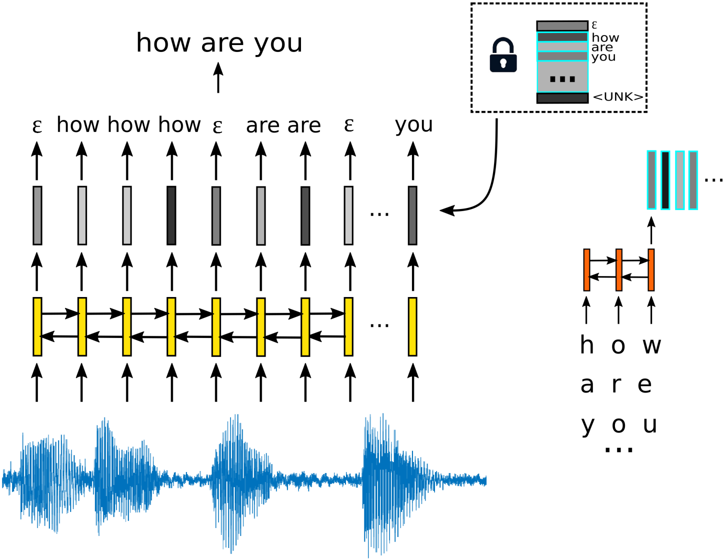
Acoustics-to-Word Recognition: AGWE Initialized



Acoustics-to-Word Recognition: AGWE Regularized



Acoustics-to-Word Recognition: AGWE Frozen



Experimental Setup

Data

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

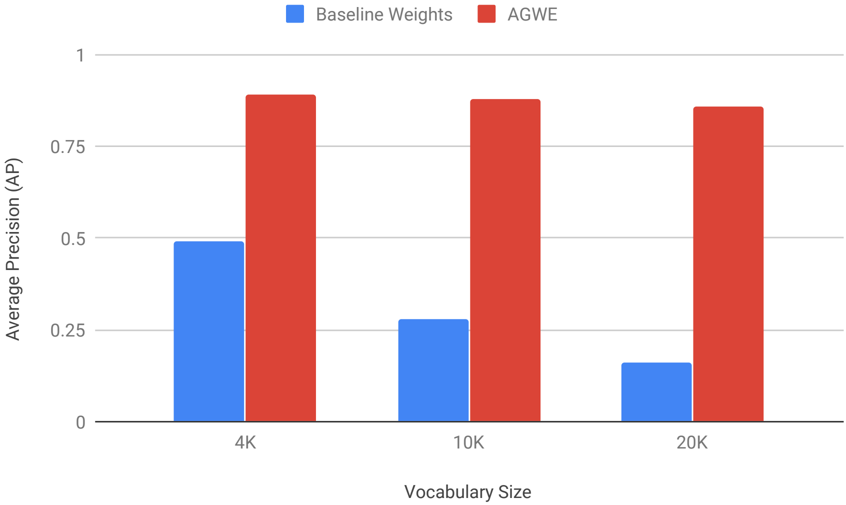
Acoustically grounded word embeddings (AGWE)

- Acoustic view: 6-BLSTM (512d) $\rightarrow$ 256d
- Character view: 64d char embed $\rightarrow$ 1-BLSTM (512d) $\rightarrow$ 256d
- Tuned on development set word discrimination performance

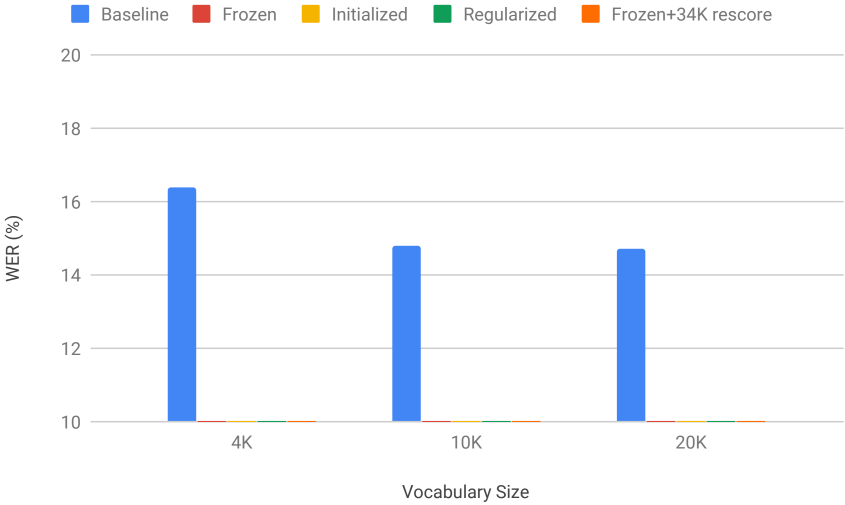
Acoustics-to-word (A2W) recognition

- Acoustic view $\rightarrow$ prediction layer over $|V|$ words
- Word error rate reported on Hub5-2000 Switchboard evaluation set

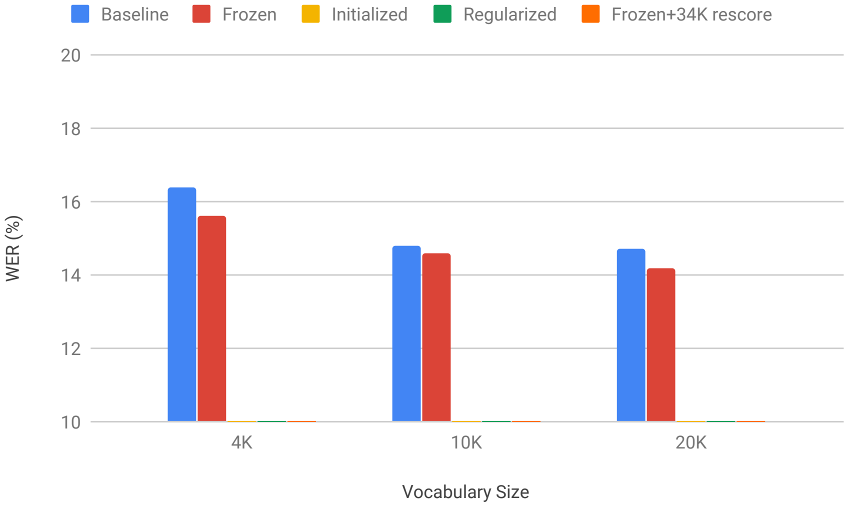
Word Discrimination Development Set Results



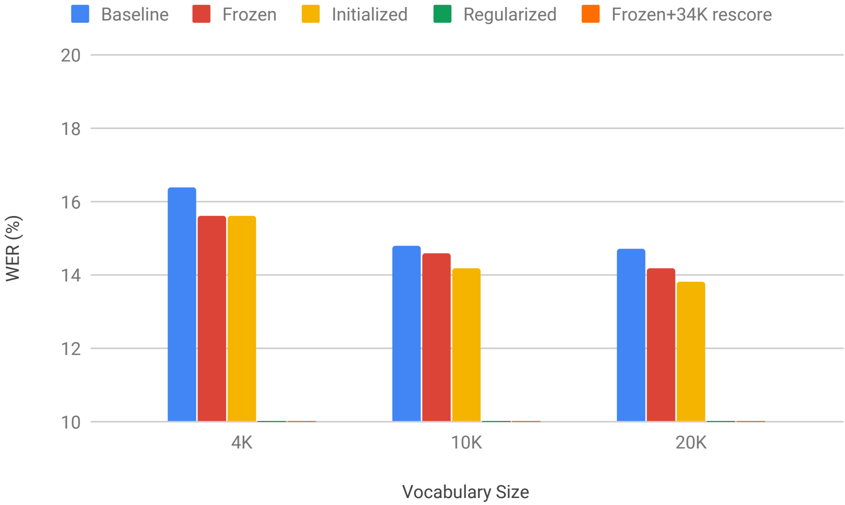
Acoustics-to-Word Recognition: Switchboard Results



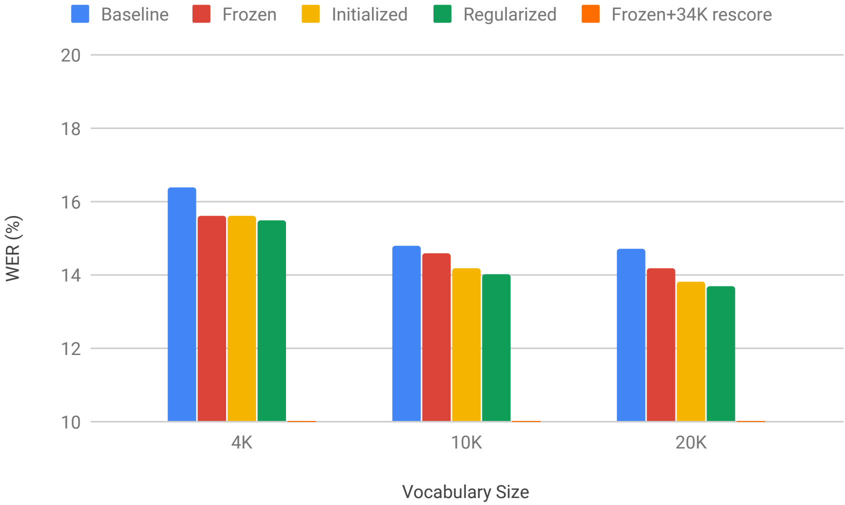
Acoustics-to-Word Recognition: Switchboard Results



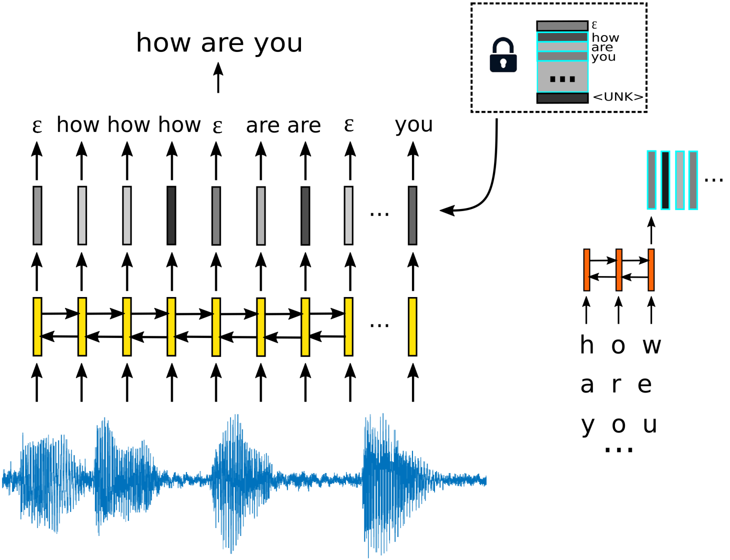
Acoustics-to-Word Recognition: Switchboard Results



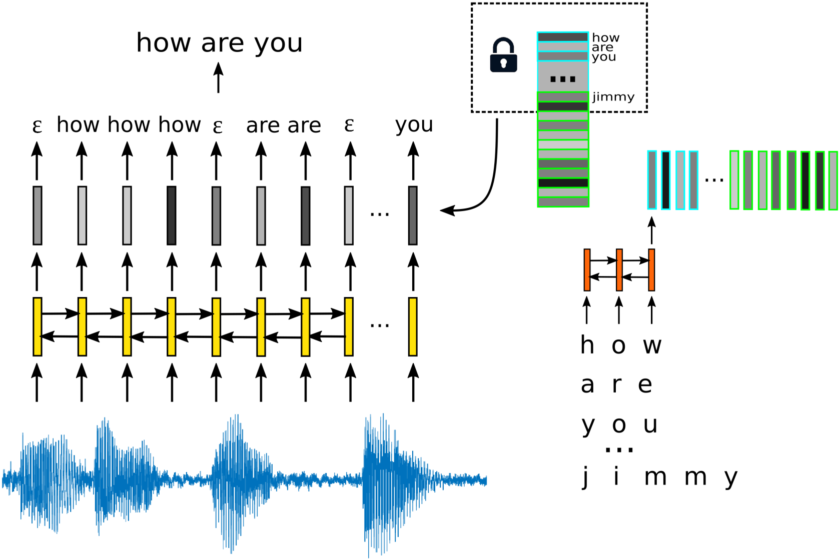
Acoustics-to-Word Recognition: Switchboard Results



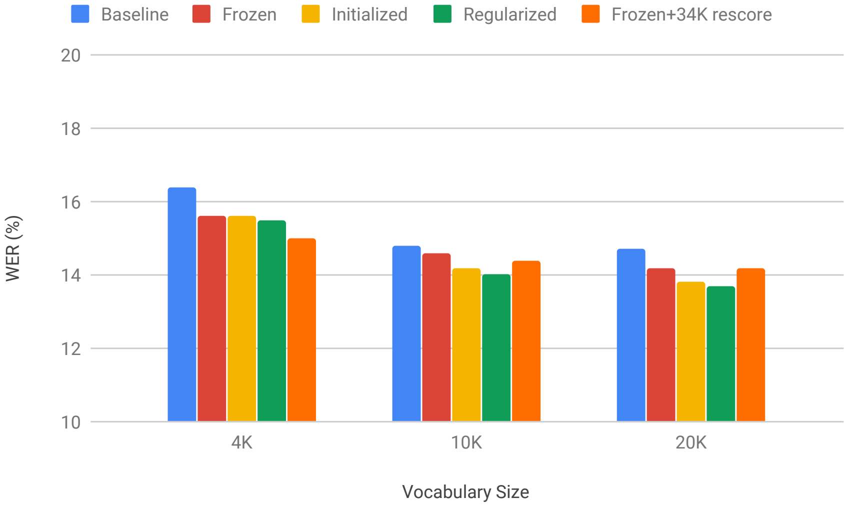
Acoustics-to-Word Recognition: Vocabulary Extension



Acoustics-to-Word Recognition: Vocabulary Extension



Acoustics-to-Word Recognition: Switchboard Results



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):

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):

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

REF: `but those LOANS ARE so much cheaper`
HYP (1st pass): `but those <UNK> so much cheaper`
HYP (rescoring):

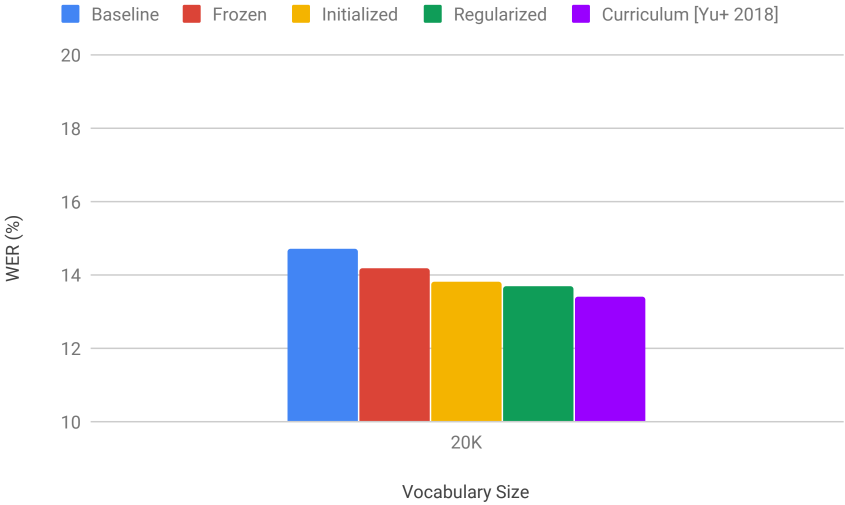
Acoustics-to-Word Recognition: Frozen+34K Rescores

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



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