TTIC 31110
Speech Technologies

May 28, 2020

Announcements

- Project presentations
  - Schedule for interim presentations (June 2) and final presentations (Jun 11) has been posted
  - Many requests for switches to June 11... any volunteers to present June 2?

Outline

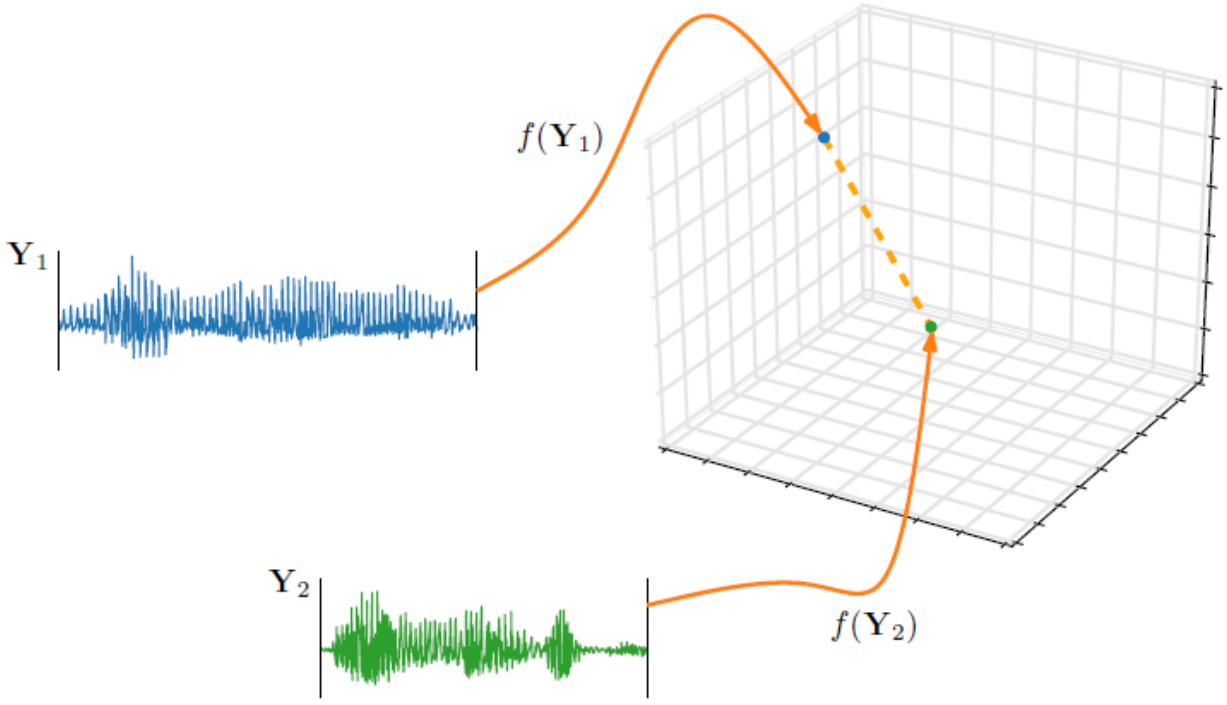
Learning pre-trained representations for speech

Higher-level “downstream” tasks

Acoustic word embeddings

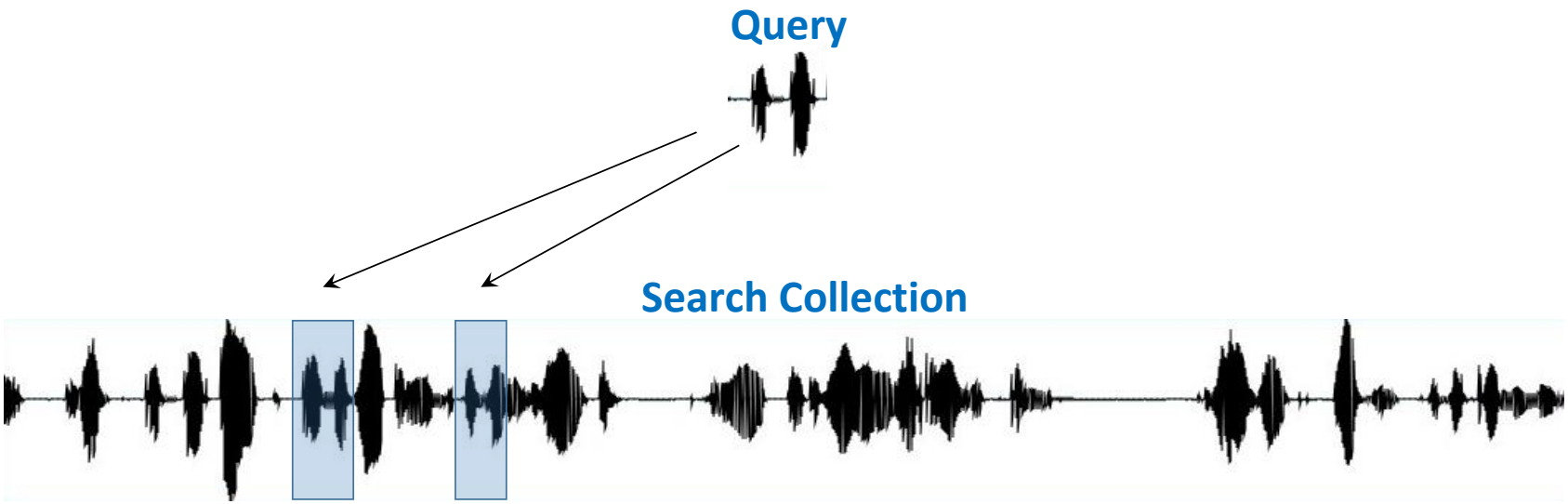
What if we want to represent **entire spoken words**?

- **Acoustic word embedding** = Function that maps from a spoken word to a vector
- **Spoken word** = speech signal of arbitrary duration corresponding to a word



Applications of spoken word embeddings

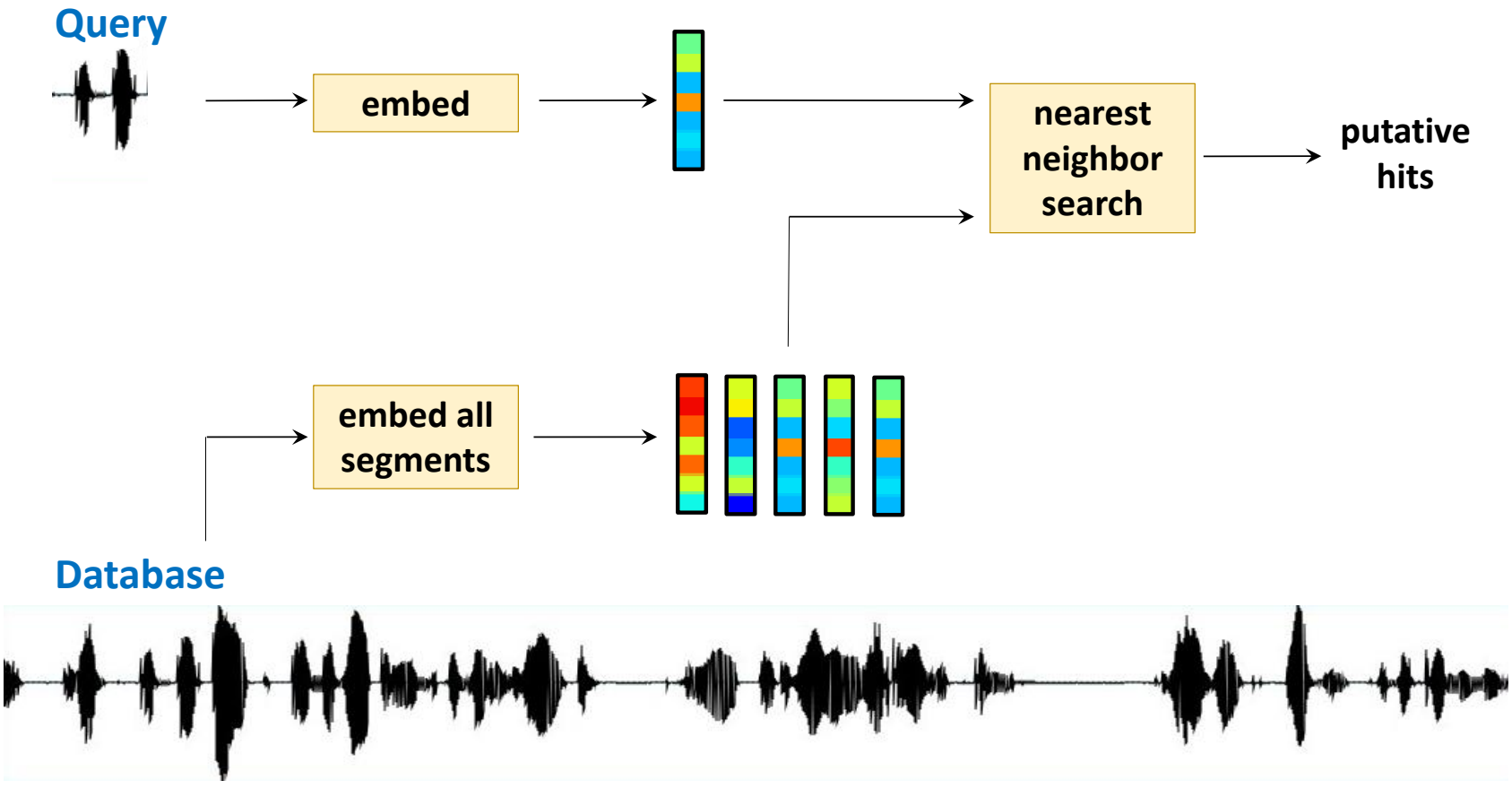
Query-by-example search: Given spoken query, find examples of it in a search collection



Useful for:

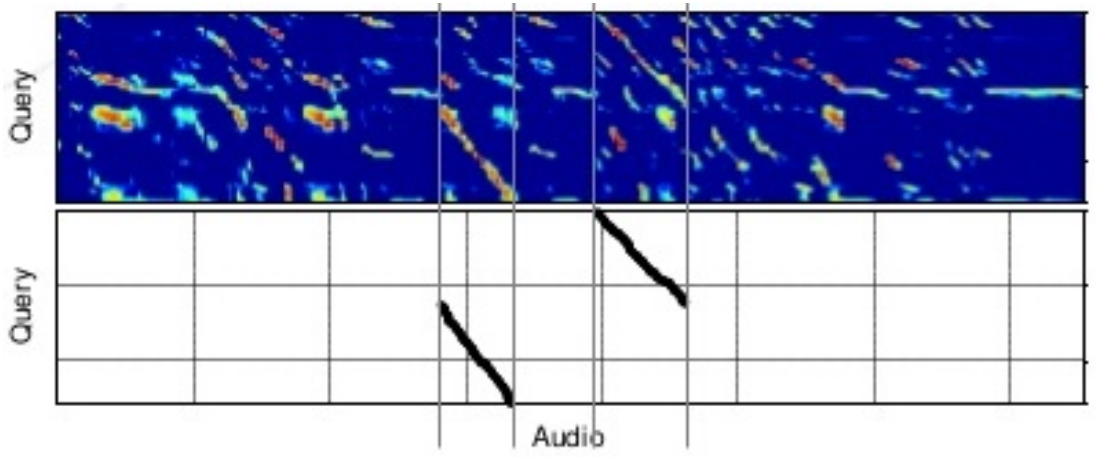
- Open-vocabulary search
- Search in multiple/low-resource/unwritten languages

Query-by-example search



[Figure credit: Herman Kamper]

Query-by-example: Classic approach



[Figure credit: Proenca et al. 2015]

Dynamic time warping (DTW)

- Slow
- Sensitive to nuisance variations: noise, speaker, ...
- Hard to learn and tune

Query-by-example results

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

System	**P@10 (↑)**	**Time (s) (↓)**
DTW baseline [Jansen & van Durme 2012]	44.0	24.70
Acoustic word embeddings [Interspeech 2017]	**60.2**	0.38

Embedding-based search is both more accurate and faster than DTW baseline

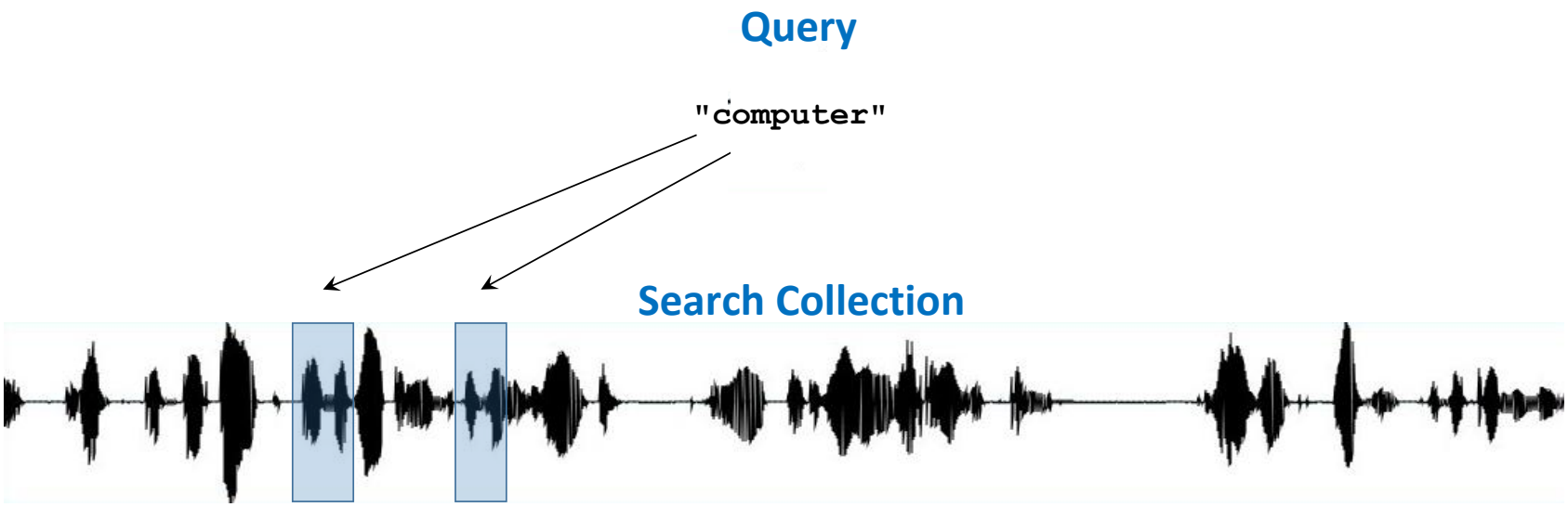
Settle, Levin, Kamper, and Livescu, "Query-by-Example Search with Discriminative Neural Acoustic Word Embeddings," Interspeech 2017

Combining acoustic and written word embeddings

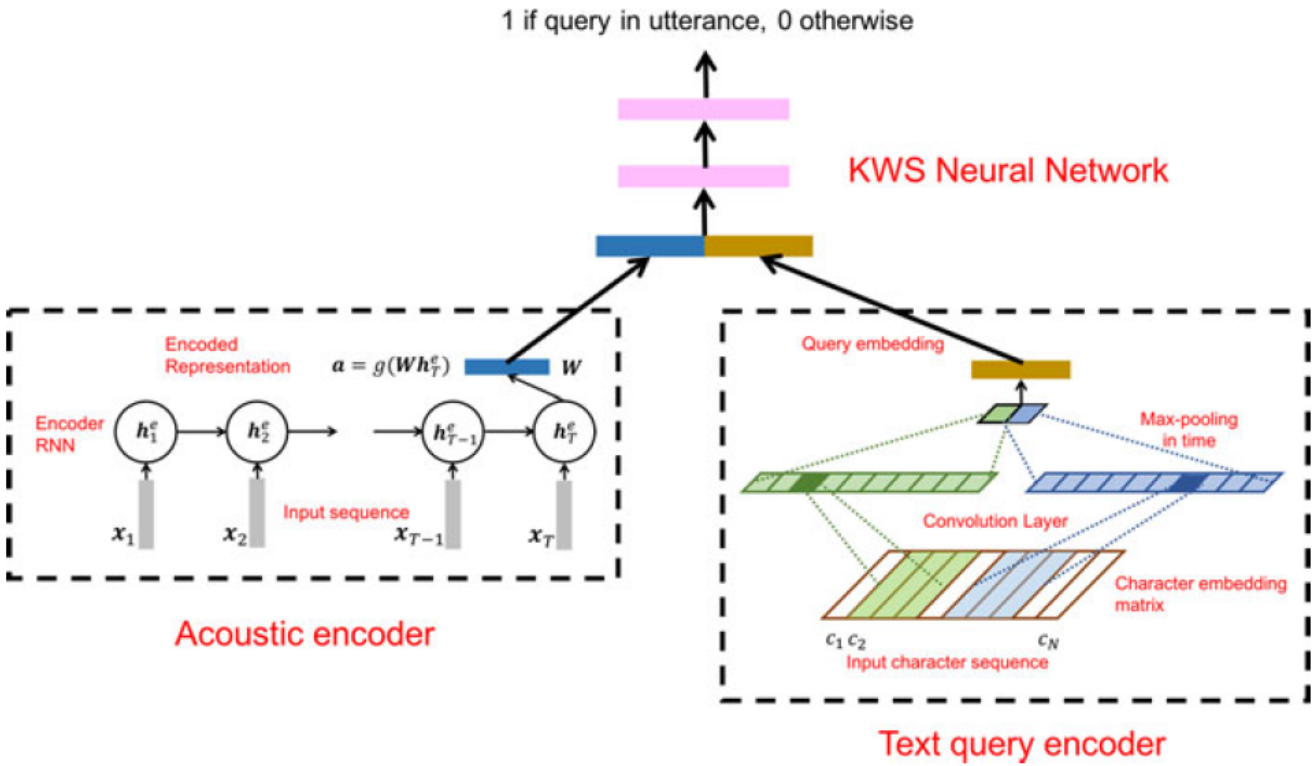
Some tasks involve comparing speech segments and written words

- Keyword search / spoken term detection
- Whole-word speech recognition

This suggests learning acoustic word embeddings as well as written word embeddings that represent the way a word sounds



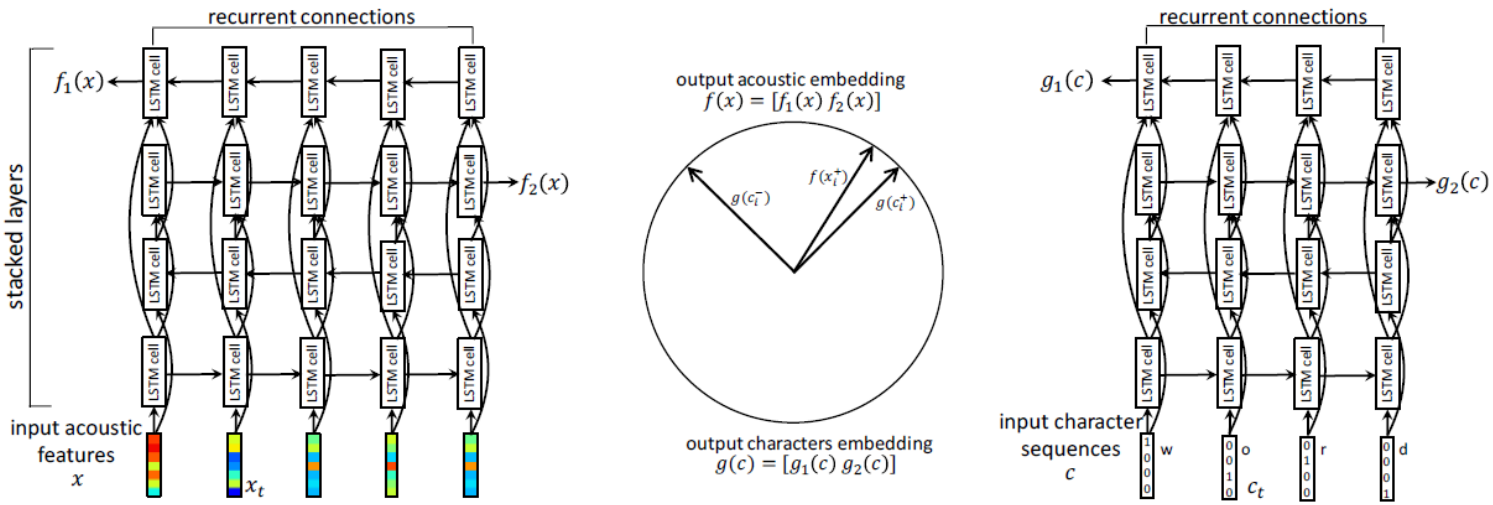
Combining acoustic and written word embeddings for spoken term detection



Audhkhasi, Rosenberg, Sethy, Ramabhadran, and Kingsbury, "End-to-end ASR-free keyword search from speech," *IEEE Journal of Selected Topics in Signal Processing* **11**(8):1351–1359, 2017.

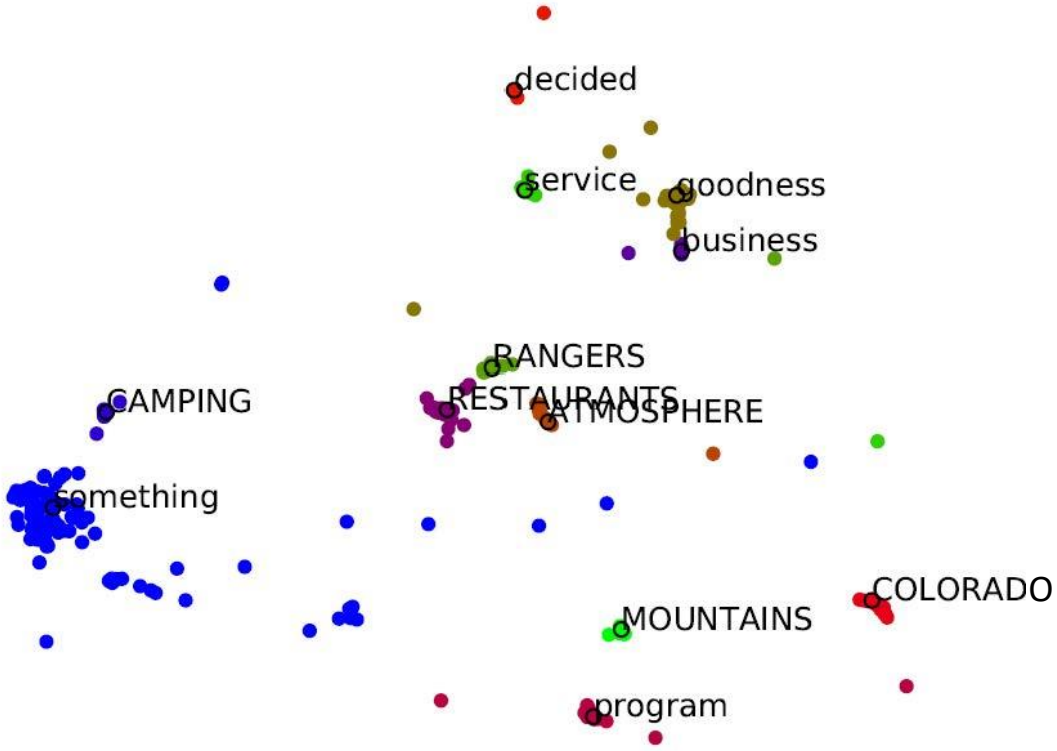
Joint learning of acoustic and written word embeddings

- **Approach:** Learn a pair of neural embedding functions
  - Acoustic word embedding (speech → vector)
  - Acoustically grounded word embedding (character sequence → vector)
- **Training data:** Pairs of matched (acoustic, written) words



He, Wang, and Livescu, "Multi-view recurrent neural acoustic word embeddings," ICLR 2017

Visualization of jointly learned embeddings

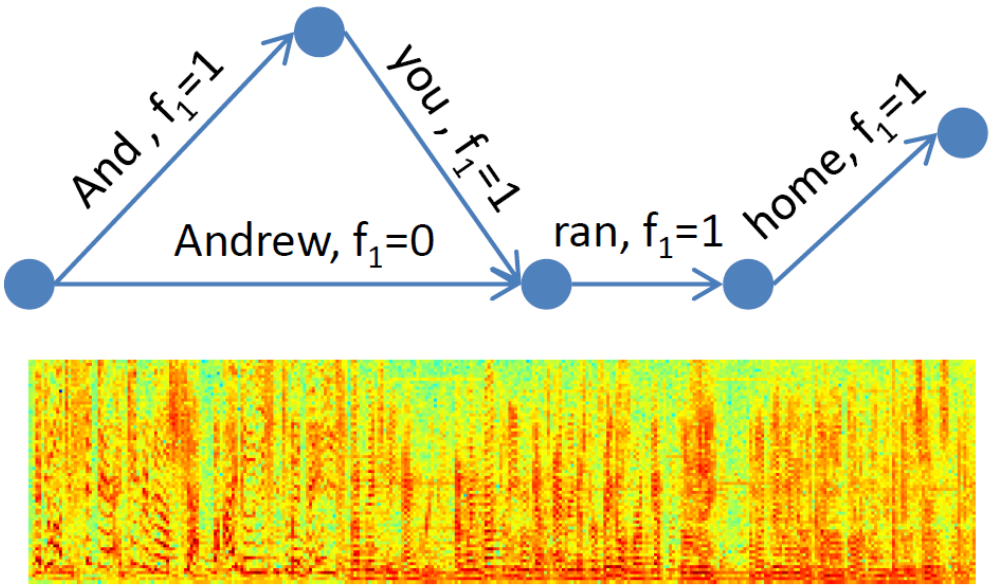


- Acoustic word embeddings cluster around corresponding written word embedding
- Previously unseen words are embedded well

Whole-word ASR with word embeddings

First attempt: Maas *et al.* 2012

- Learn acoustic word embeddings that approximate pre-trained written word embeddings
- Use learned embeddings to rescore output lattices from first-pass HMM-based speech recognizer

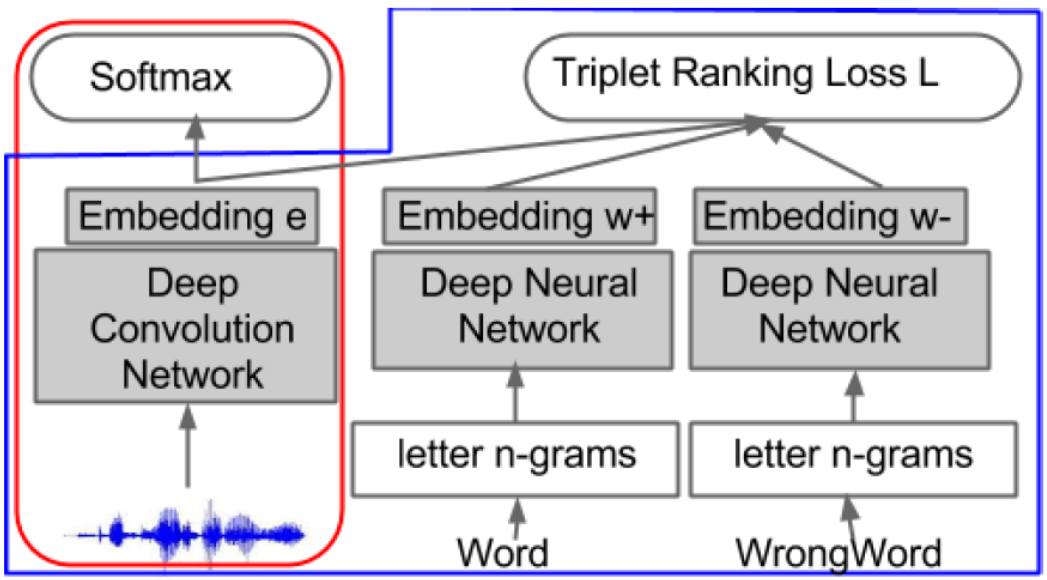


Maas, Miller, O'Neil, Ng, and Nguyen, "Word-level acoustic modeling with convolutional vector regression," ICML Workshop on Representation Learning, 2012

Whole-word ASR with word embeddings

Second attempt: Bengio and Heigold 2014

- Jointly learn acoustic and written word embeddings
- Use learned embeddings to rescore output lattices from first-pass HMM-based speech recognizer

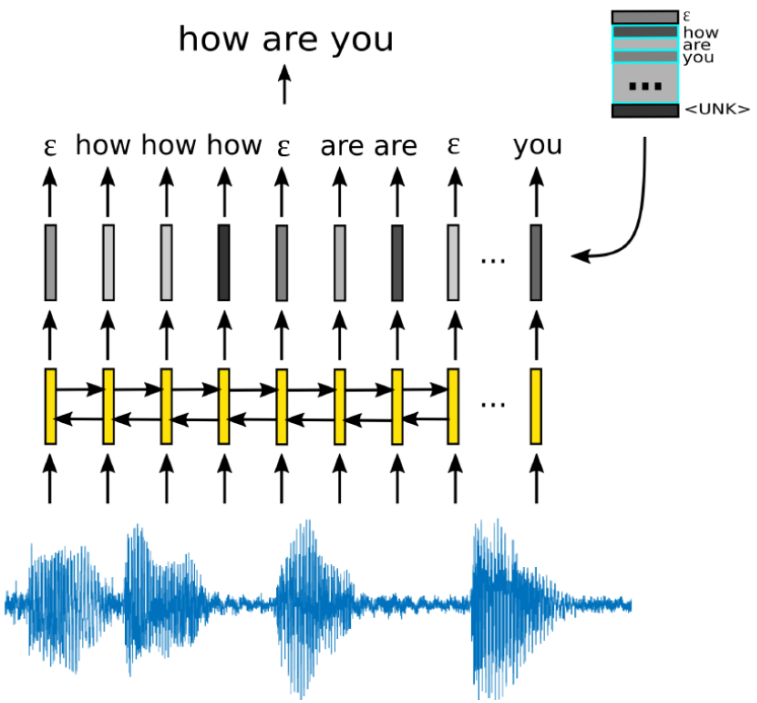


Bengio and Heigold, "Word embeddings for speech recognition," ICASSP 2014

Whole-word ASR with word embeddings

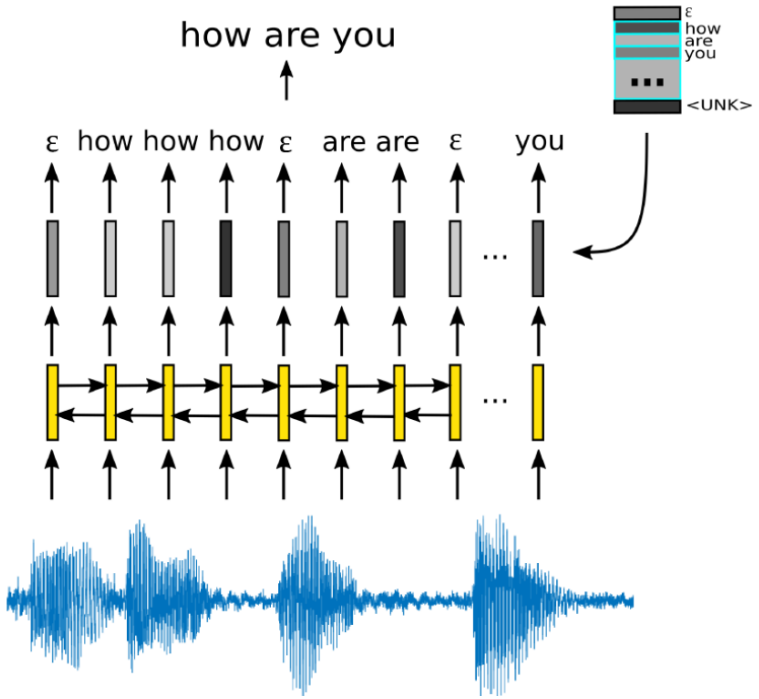
End-to-end whole-word neural ASR [Audhkhasi+ 2017, Soltau+ 2018, Li+ 2018]

- Output labels are whole words
- No need for phones, word pieces, lexicons, ...
- Note: The final layer weights represent a word embedding matrix



Whole-word ASR with word embeddings

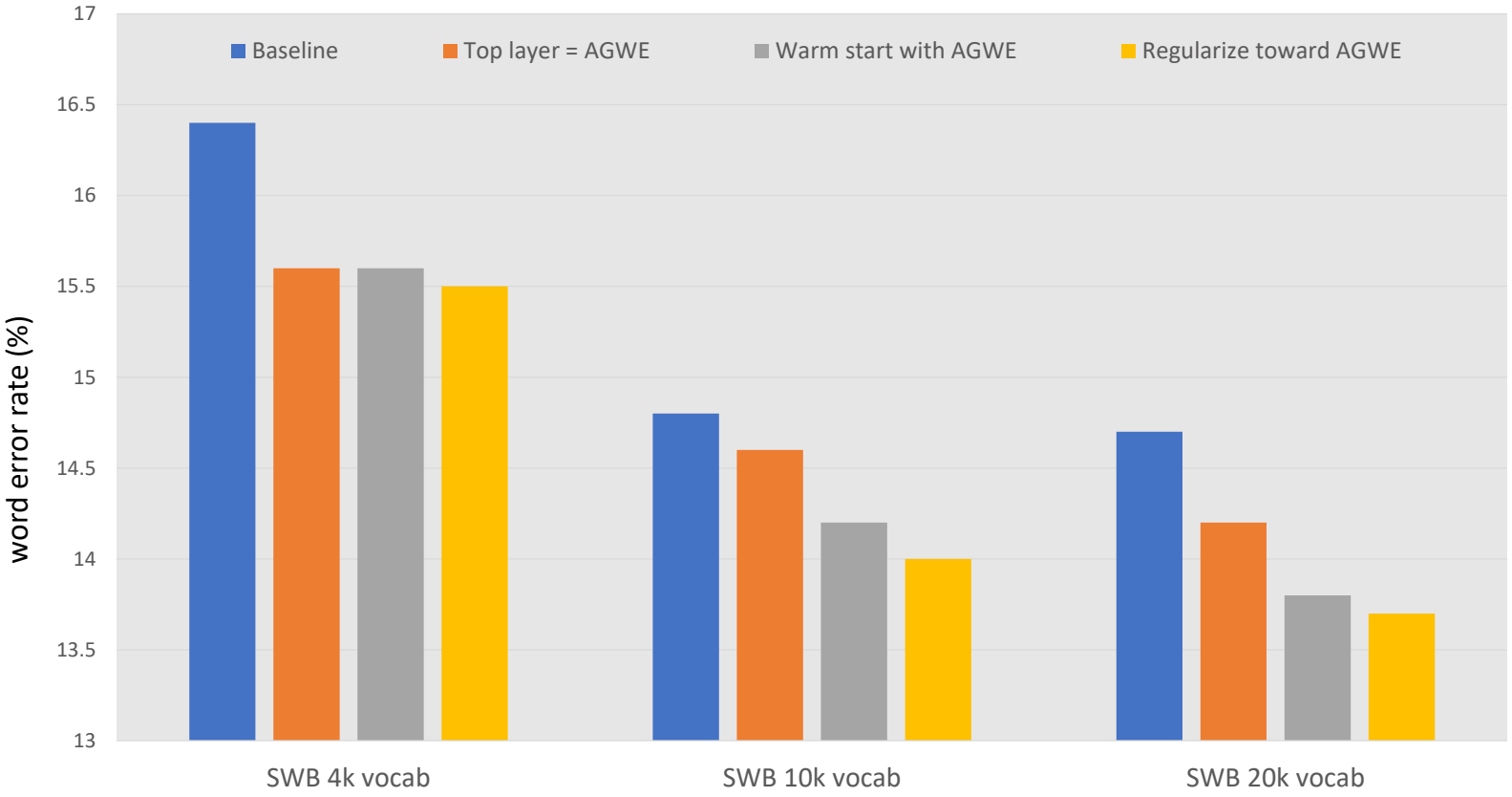
- **Problem:** Many words are rare $\longrightarrow$ poorly learned
- **Idea:** Use pre-trained acoustically grounded embeddings for the final weight layer
- Pre-trained embeddings are **parametric functions** $\longrightarrow$ can handle rare/unseen words



Settle, Audhkhasi, Livescu, Picheny, "Acoustically Grounded Word Embeddings

Whole-word ASR with word embeddings

Switchboard conversational telephone speech recognition results:



Bonus: on-the-fly vocabulary extension $\longrightarrow$ extra 0.6% improvement in 4K vocab case Settle, Audhkhasi, Livescu, Picheny, "Acoustically Grounded Word Embeddings for Improved Acoustics-to-word Speech Recognition," ICASSP 2019

Whole-word ASR with word embeddings

Learning the acoustically grounded word embeddings together with recognizer training

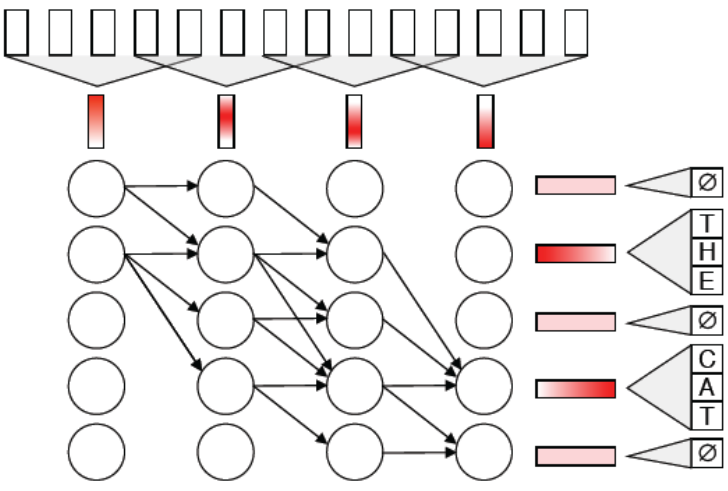


Figure 1: A CTC trained acoustic model combined with the character-based word model. The ∅ denotes BLANK.

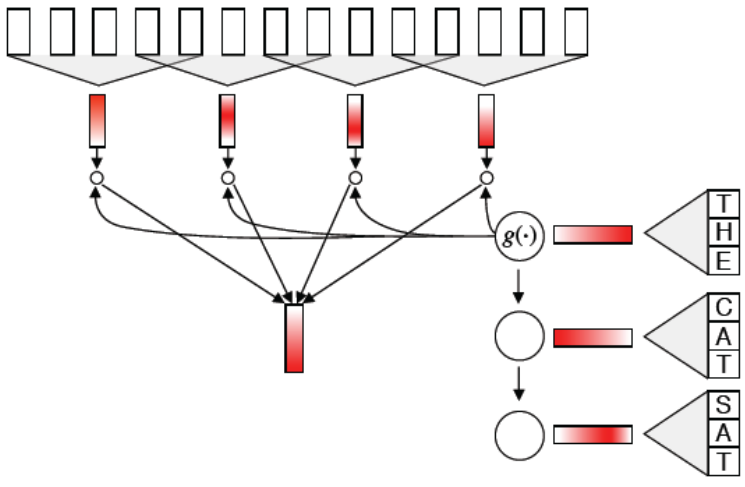


Figure 2: A seq2seq model combined with the character-based word model. The function $g(\cdot)$ denotes the decoder RNN.

Collobert, Hannun, and Synnaeve, "Word-level speech recognition with a dynamic lexicon," arXiv:1906.04323

Whole-word ASR with word embeddings

Learning the acoustically grounded word embeddings together with recognizer training

Table 1: Comparison of our word-level approaches against a word-piece baseline, with and without language model (LM) decoding. All models have an overall stride of 8 except rows with $s = 16$, which have a stride of 16.

Model	LM	$\|\mathcal{D}\|$ train	$\|\mathcal{D}\|$ test	Dev WER clean	Dev WER other	Test WER clean	Test WER other
word piece seq2seq [16]	None	10k	10k	5.04	14.45	5.36	15.64
word piece CTC	None	10k	10k	6.58	17.52	6.69	18.23
word piece CTC	4-gram	10k	10k	6.05	13.85	6.42	14.81
word-level seq2seq	None	89k	89k	5.99	16.73	6.22	17.32
word-level CTC	None	89k	89k	5.42	15.62	5.63	16.27
word-level CTC, $s = 16$	None	89k	89k	5.61	15.49	5.49	16.01
word-level CTC	4-gram	89k	89k	4.35	12.10	4.42	12.82
word-level CTC, $s = 16$	4-gram	89k	89k	4.43	12.00	4.51	13.00
word-level CTC	4-gram	89k	200k	**4.09**	**11.12**	**4.26**	**11.98**
word-level CTC, $s = 16$	4-gram	89k	200k	4.17	11.21	4.27	12.20

Collobert, Hannun, and Synnaeve, "Word-level speech recognition with a dynamic lexicon," arXiv:1906.04323

Pre-trained speech representations: Summary

Last lecture: Unsupervised pre-trained frame representations

- Some very recent progress in pre-training frame-level representations for speech recognition
- No "universal" pre-trained model yet
- Very little work on downstream tasks other than speech recognition

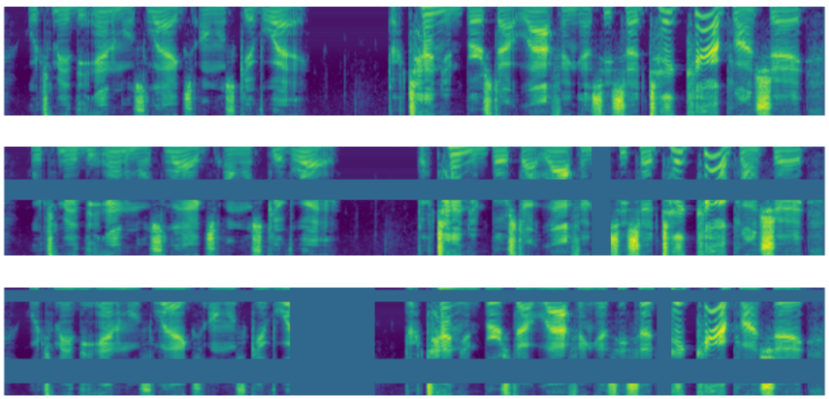
Pre-trained acoustic word embeddings

- Good results on speech search, whole-word speech recognition
- Unsupervised embeddings catching up to supervised ones

Some other good ideas

Data augmentation

- Speed perturbation, vocal tract length perturbation
- Simulated noise or room acoustics
- SpecAugment



D. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," Interspeech 2019.

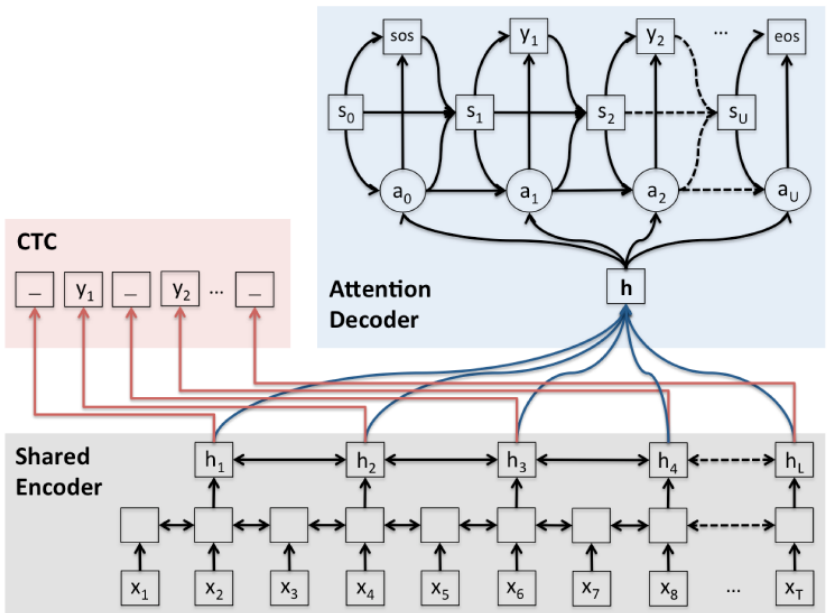
T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," Interspeech 2015.

N. Jaitly and G. Hinton, "Vocal Tract Length Perturbation (VTLP) improves speech recognition," in ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013.

Some other good ideas

Multi-task learning (MTL)

- Multiple decoders (e.g., CTC + attention-based)
- Task loss + unsupervised representation loss
- Final task loss + intermediate losses ("hierarchical MTL")



S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," ICASSP 2017.

G. Kurata and K. Audhkhasi, "Multi-task CTC Training with Auxiliary Feature Reconstruction for End-to-end Speech Recognition," Interspeech 2019.

Beyond speech recognition to higher-level tasks

Often speech recognition is not the end goal

- Search in spoken corpora (discussed earlier)
- Speech understanding tasks (e.g., getting Siri, Alexa, etc. to do something)
- Speech translation
- Speech summarization
- ...

Using/adapting pre-trained text representations for speech

Useful for

- Rescoring ASR outputs
- Downstream tasks that use the ASR outputs

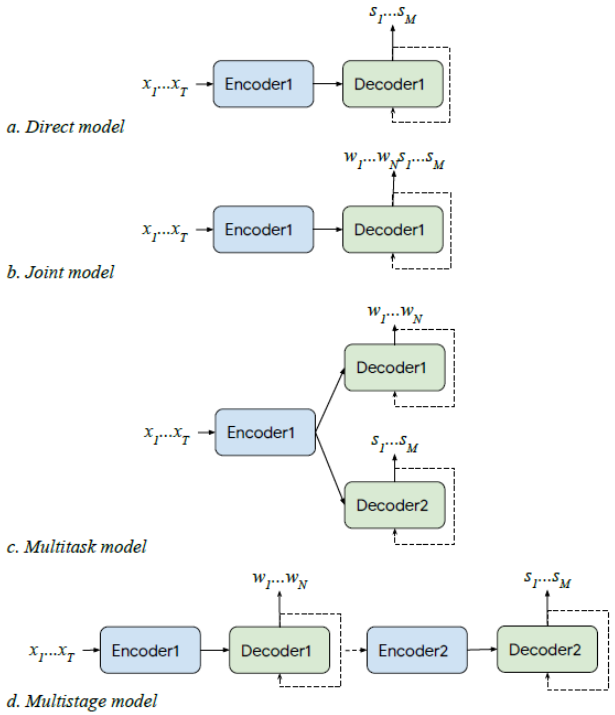
Huang and Chen, "Adapting pretrained transformer to lattices for spoken language understanding," ASRU 2019.

Huang and Peng, "An empirical study of efficient ASR rescoring with transformers," arXiv:1910.1145.

Sperber, Neubig, Pham, and Waibel, "Self-Attentional Models for Lattice Inputs," arXiv:1906.01617.

End-to-end spoken language understanding

A bit speculative at this point; results are not better than pipeline ASR + NLP approaches

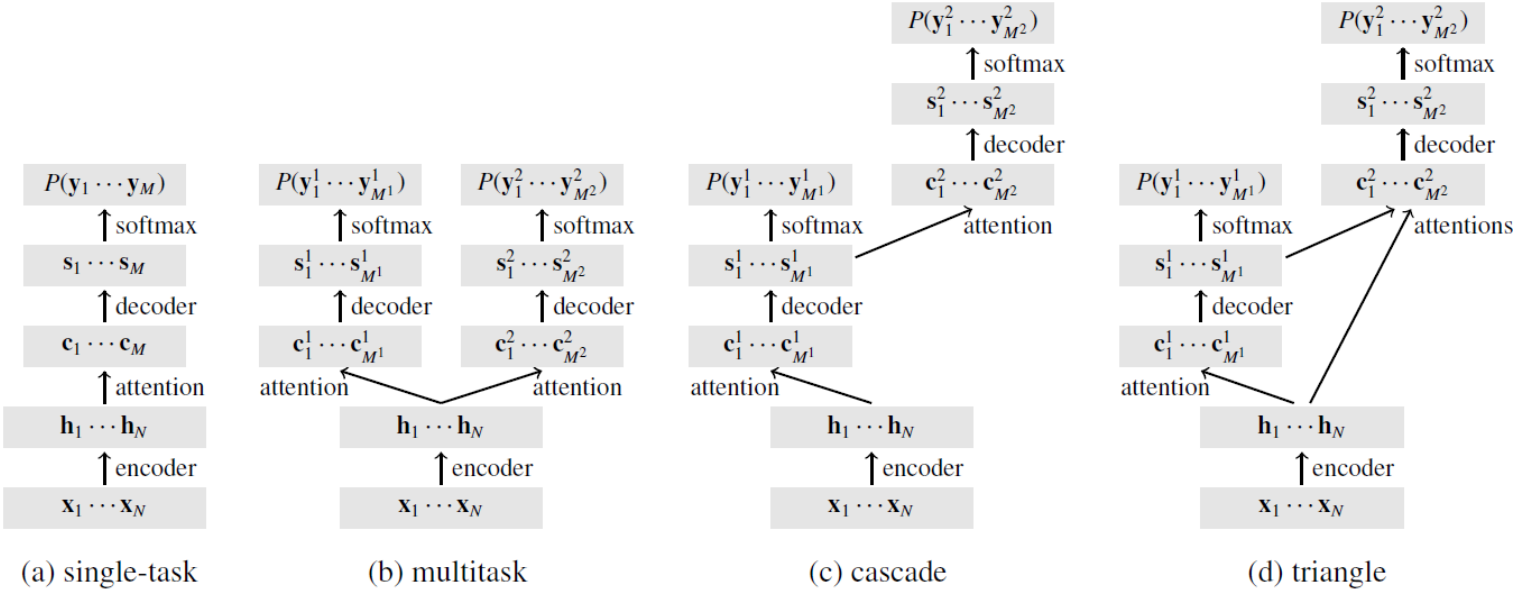


P. Haghani, A. Narayanan, M. Bacchiani, G. Chuang, N. Gaur, P. Moreno, R. Prabhavalkar, Z. Qu, A. Waters, "From audio to semantics: Approaches to end-to-end spoken language understanding," SLT 2018.

D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," arXiv:1802.08395.

Speech translation

End-to-end approach improve over pipelines

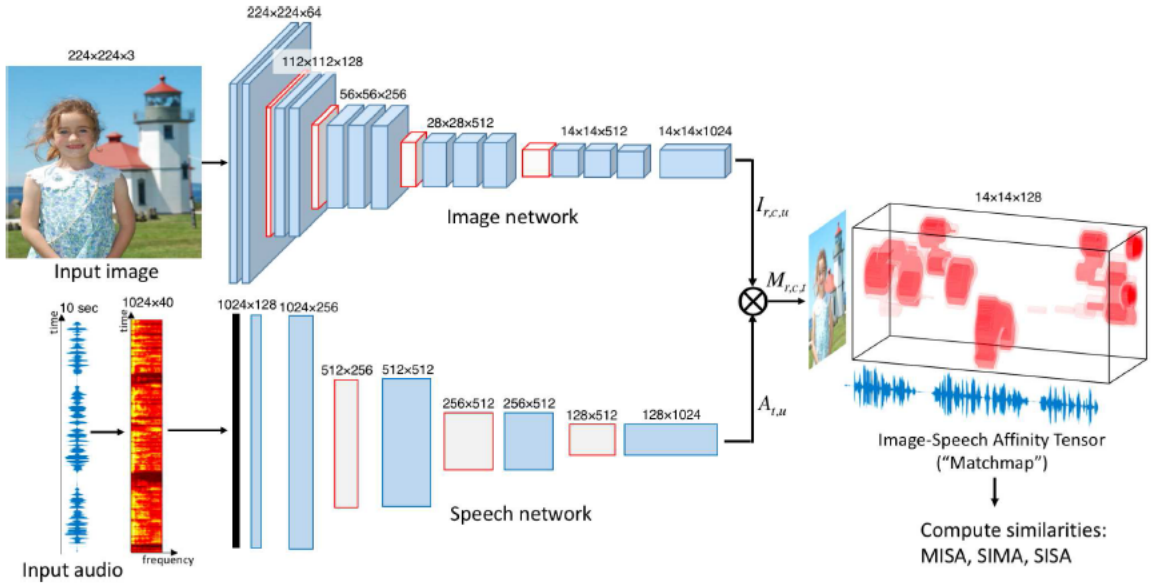


A. Anastasopoulos and D. Chiang, "Tied Multitask Learning for Neural Speech Translation," NAACL HLT 2018.

RJ Weiss, J Chorowski, N Jaitly, Y Wu, and Z Chen, "Sequence-to-Sequence Models Can Directly Translate Foreign Speech," Interspeech 2017.

Some other good ideas: visually grounded representations

Joint learning of semantic representations of speech and images



D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," ECCV 2018.

Some other good ideas: visually grounded representations

Visualization of learned semantic units

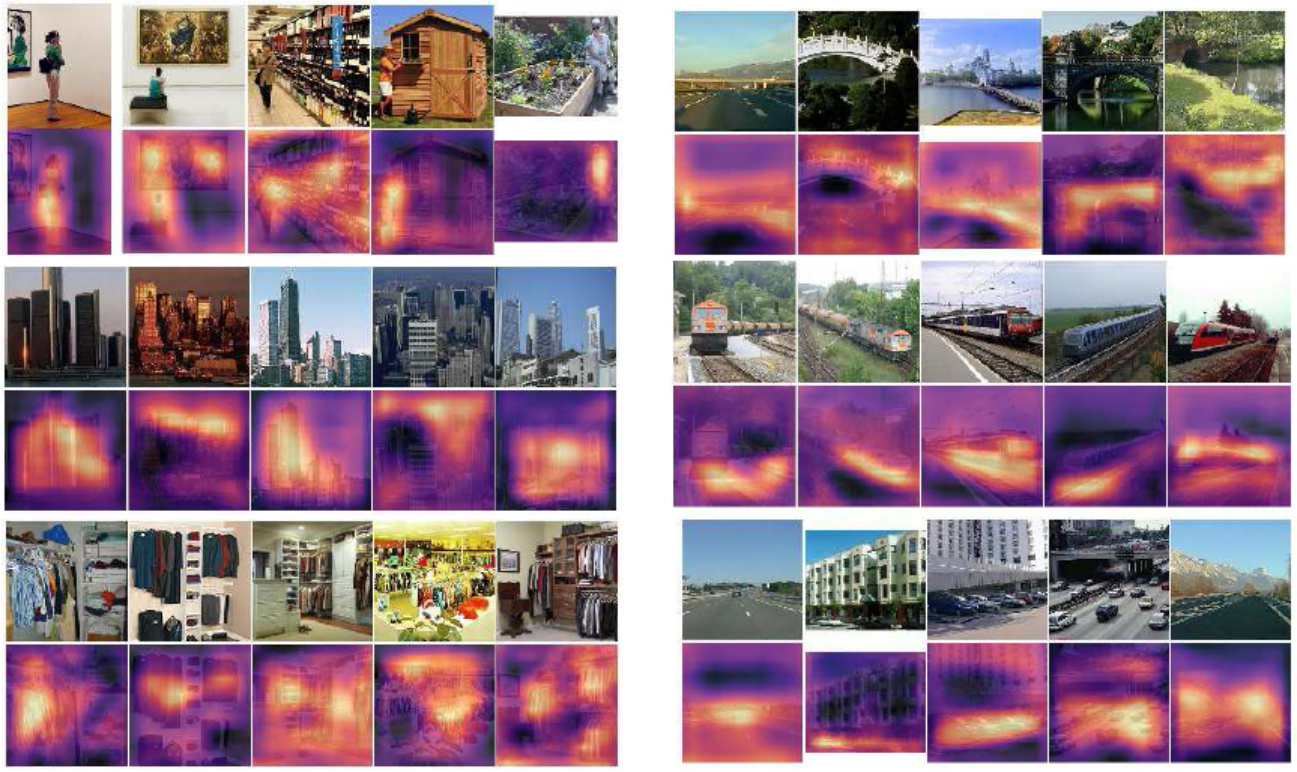
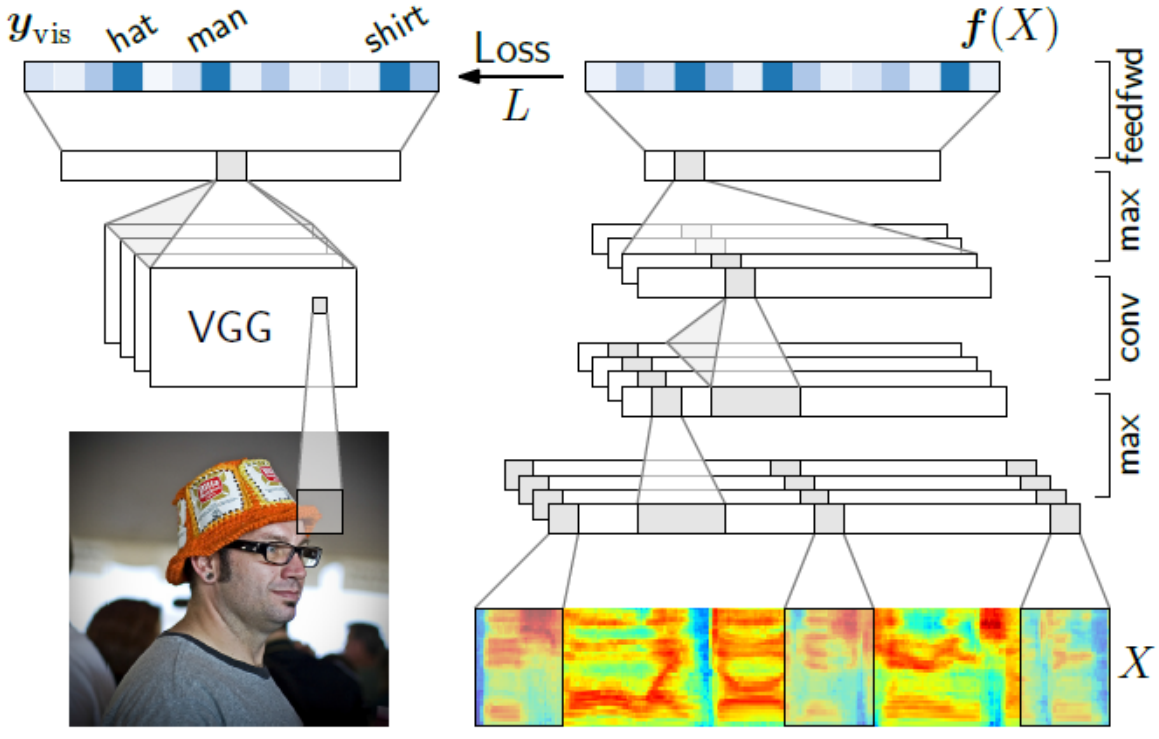


Fig. 4: Speech-prompted localization maps for several word/object pairs. From top to bottom and from left to right, the queries are instances of the spoken words "WOMAN," "BRIDGE,", "SKYLINE", "TRAIN", "CLOTHES" and "VEHICLES" extracted from each image's accompanying speech caption.

Some other good ideas: visually grounded representations

A slightly more supervised approach:



H. Kamper, S. Settle, H. Kamper, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," Interspeech 2017.

Some other good ideas: visually grounded representations

Some example keyword predictions

Table 2: *Example input utterances and BoW predictions of VisionSpeechCNN for $\alpha = 0.7$. Orange shows correct predictions.*

Transcription of input utterance	Predicted BoW labels
a little girl is climbing a ladder	child, girl, little, young
a rock climber standing in a crevasse	climbing, man, rock
man on bicycle is doing tricks in an old building	bicycle, bike, man, riding, wearing
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
a snowboarder jumping in the air with a person riding a ski lift in the background	air, man, person, snow, snowboarder

H. Kamper, S. Settle, H. Kamper, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech,"