TTIC 31110
Speech Technologies

May 26, 2020

Announcements

- Project proposal
  - Thanks for some great proposals, and great feedback to each other!
  - Comments and grades (minus the "feedback grade") are available
  - Please take comments into consideration for the remainder of the project
  - Alert: disruption in computing resources is possible in the next 2 weeks – I will keep you posted
- HW4: Please submit PDF file separately

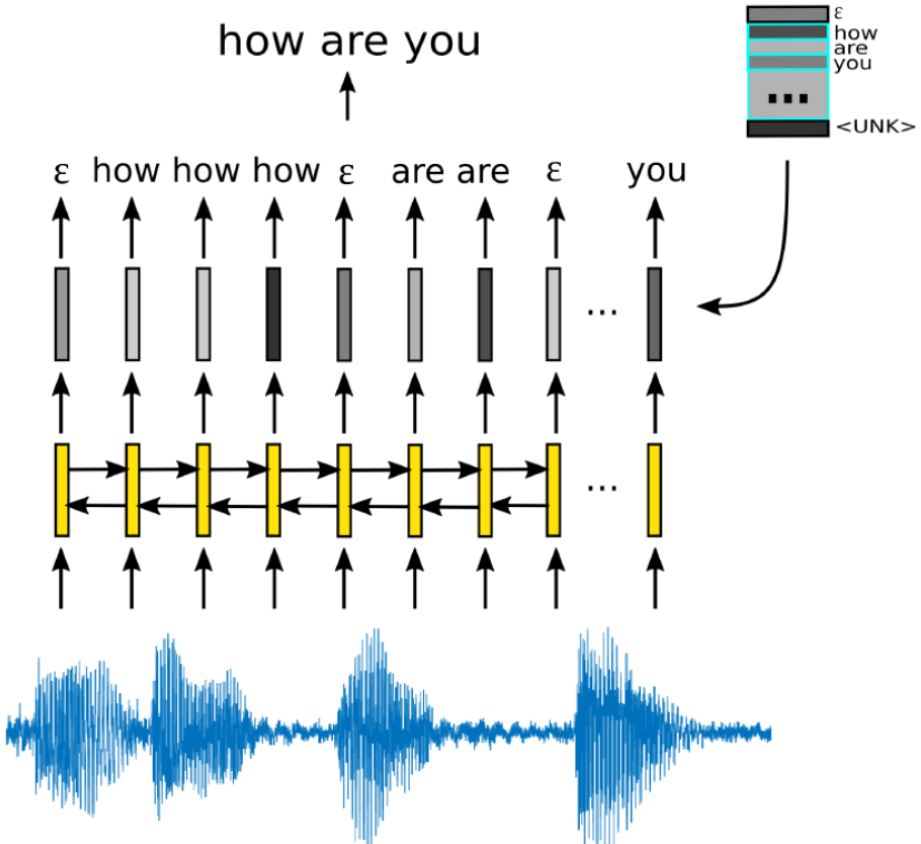
Outline

Today

Learning pre-trained representations for speech

Reminder: A typical neural speech recognizer

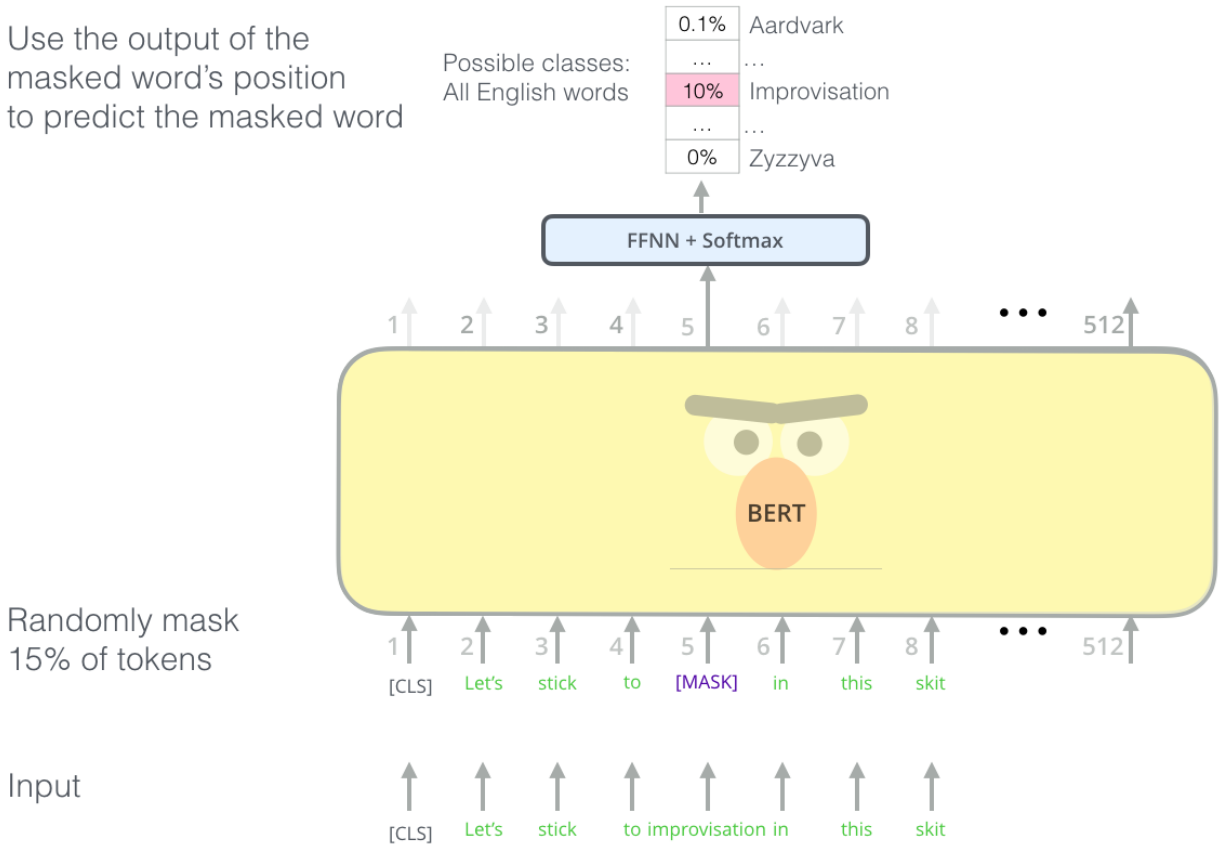
- Multiple layers of bidirectional recurrent neural network + linear layer + softmax
- Final layer weights represent a label embedding matrix
- Output labels can be whole words, but more commonly sub-words (characters, phones, etc.)
- Pre-trained representations initialize the first few layers



Background: Pre-trained text representations

Since 2018...

- Deep neural representations of text learned on unlabeled data

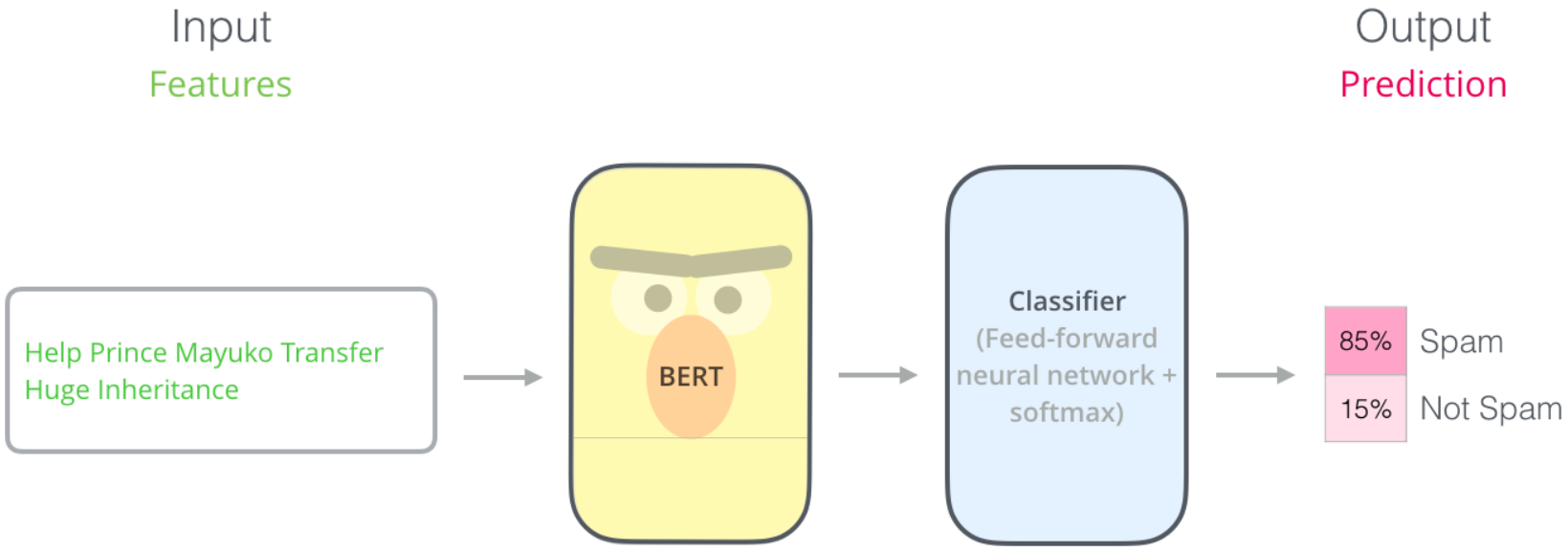


[jalammar.github.io/illustrated-bert]

Pre-trained text representations

Since 2018...

- Deep neural representations of text learned on unlabeled data
- Then combined with a simple classifier trained on much less task-specific labeled data



[jalammar.github.io/illustrated-bert]

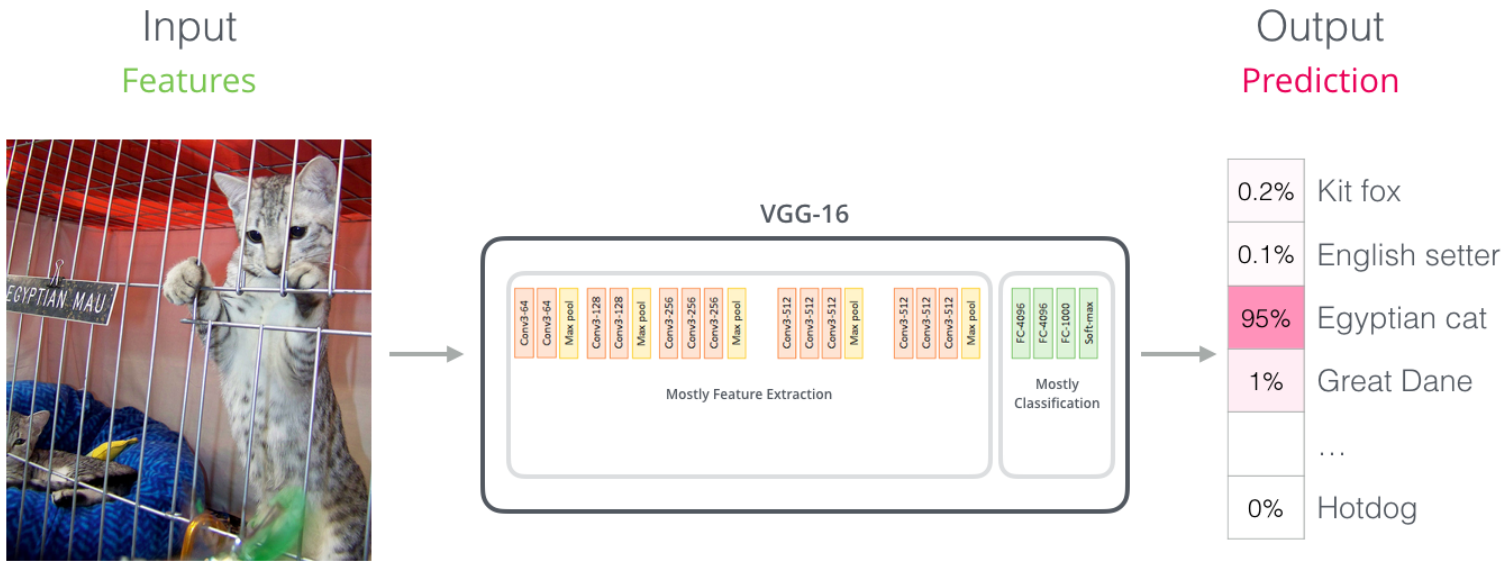
Pre-trained text representations

Since 2018...

- Deep neural representations of text learned on unlabeled data
- Then combined with a simple classifier trained on much less task-specific labeled data
- This works remarkably well!
- We have lots of unlabeled text, and lots of tasks with little labeled data!

Pre-trained representations for other types of data?

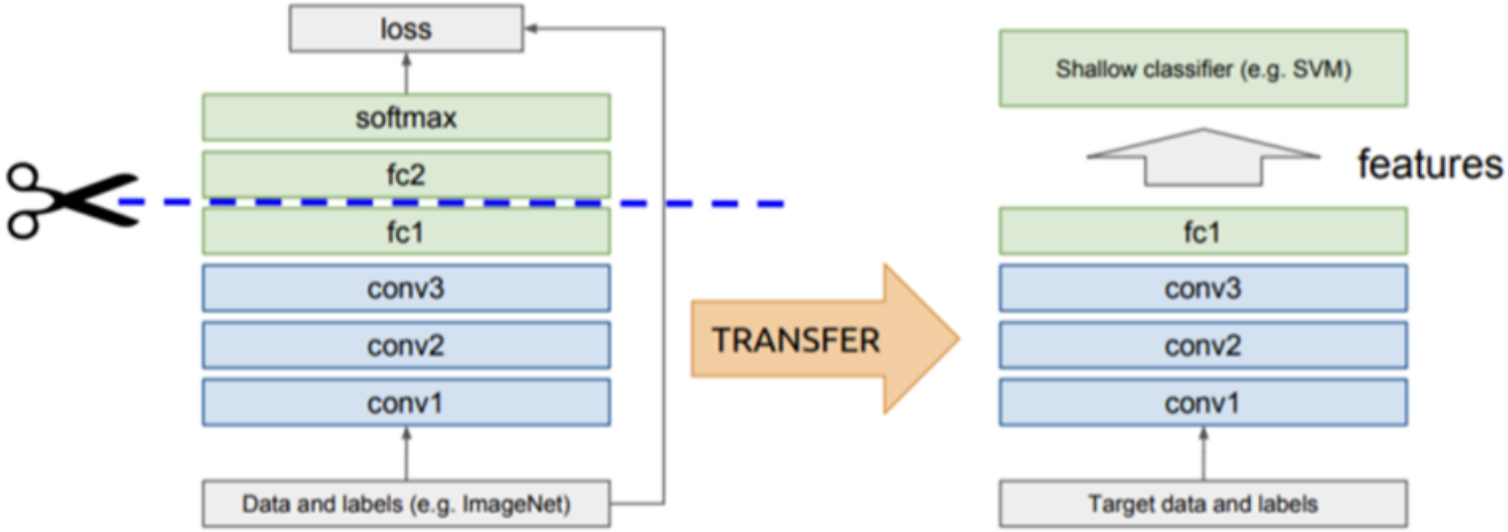
Computer vision: **supervised** pre-training



[jalammar.github.io/illustrated-bert]

Pre-trained representations for other types of data?

Computer vision: **supervised** pre-training



[towardsdatascience.com]

Pre-trained representations for speech?

No **standard** pre-training approach so far

- Neither unsupervised...
- ...nor supervised

But some promising approaches

Speech vs. text: Reminder from Lecture 2!

- Converting speech to text can be very hard (at least with limited supervision)
- Speech also provides some extra information (prosody, contextual information) (we'll ignore that for today, though)
- Human speech perception factors out irrelevant information
- Can we learn pre-trained representations that keep just the relevant information?

Aside: Pre-trained speech representations circa 2012

Pre-training in the original neural network-based speech recognition work (see the survey paper in the readings)

- Based on restricted Boltzmann machines
- Pre-training on same set as ASR training
- Goal is to initialize the network well to ease the burden on the optimizer

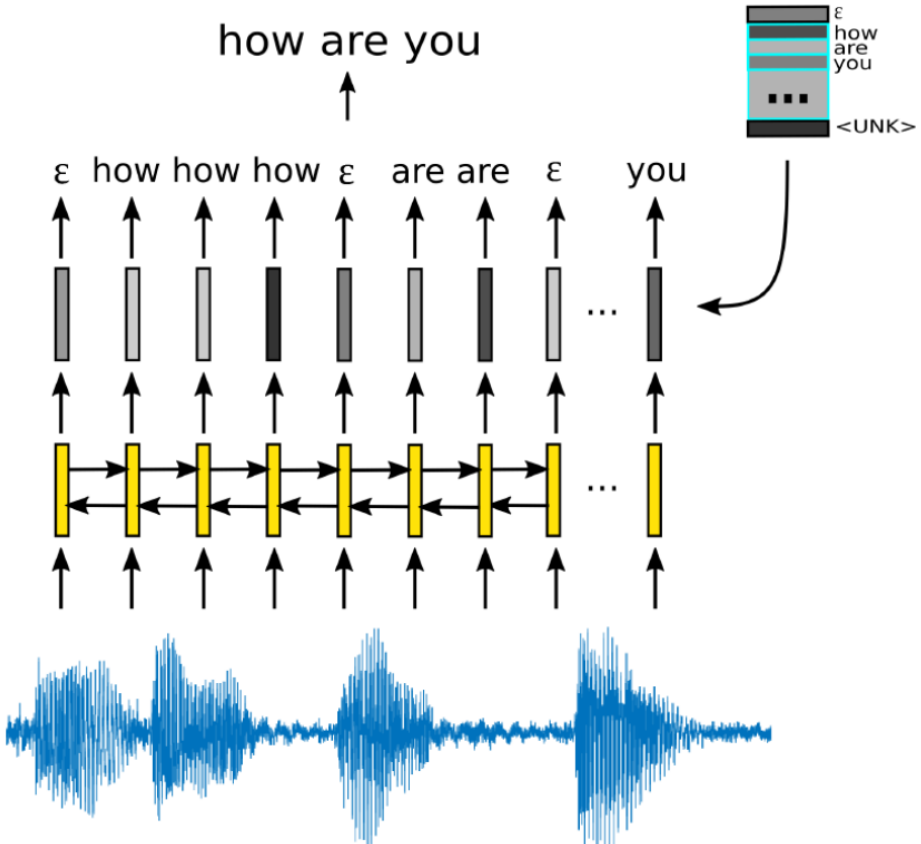
Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine* **29**(6):82–97, November 2012.

This is very different from current research

- Goal now is to use large amount of external data
- Viewed as learning good features, not aiding the optimization (though that is still a factor)

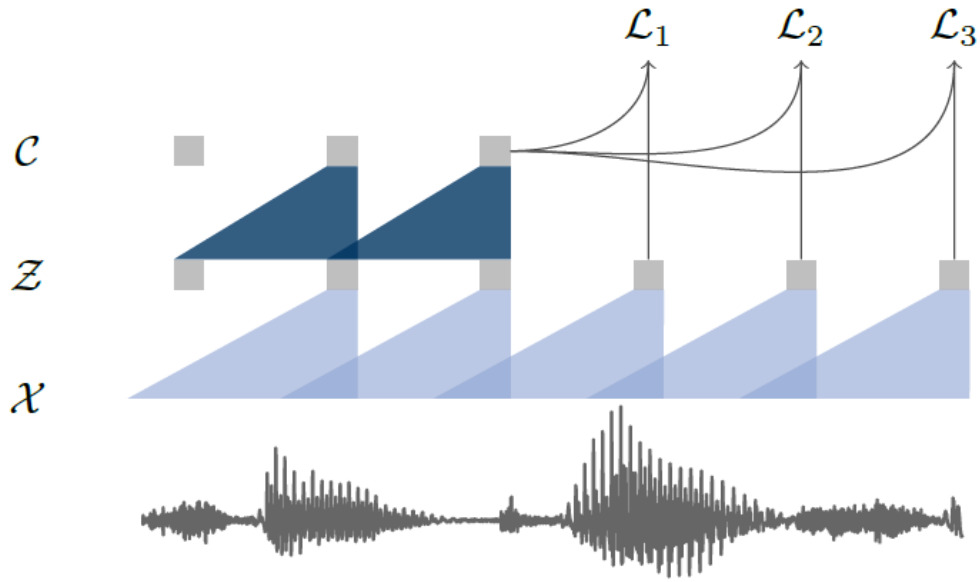
Reminder: A typical neural speech recognizer

- Multiple layers of bidirectional recurrent neural network + linear layer + softmax
- Final layer weights represent a label embedding matrix
- Output labels can be whole words, but more commonly sub-words (characters, phones, etc.)
- Pre-trained representations initialize the first few layers



First successful frame-level unsupervised pre-training approach: wav2vec

Based on contrastive predictive coding (CPC)



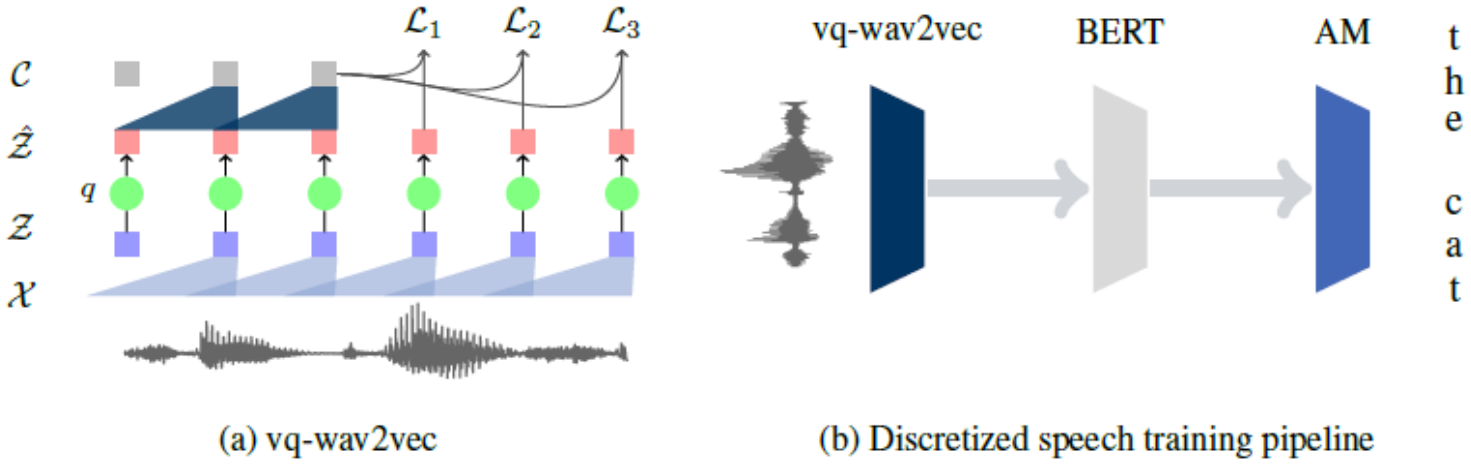
$$\mathcal{L}_k = -\sum_{i=1}^{T-k} \left(\log \sigma(\mathbf{z}_{i+k}^\top h_k(\mathbf{c}_i)) + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p_n} [\log \sigma(-\tilde{\mathbf{z}}^\top h_k(\mathbf{c}_i))] \right)$$

A. van den Oord, Yazhe Li, Oriol Vinyals, "Representation Learning with Contrastive Predictive Coding," arXiv:1807.03748.

S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," Interspeech 2019.

First successful frame-level unsupervised pre-training approach: wav2vec

- wav2vec applies CPC to the raw waveform (20-30% improvement in WER)
- vq-wav2vec: wav2vec + discretization + BERT (30-35%)
- One drawback: Atypical unidirectional architecture

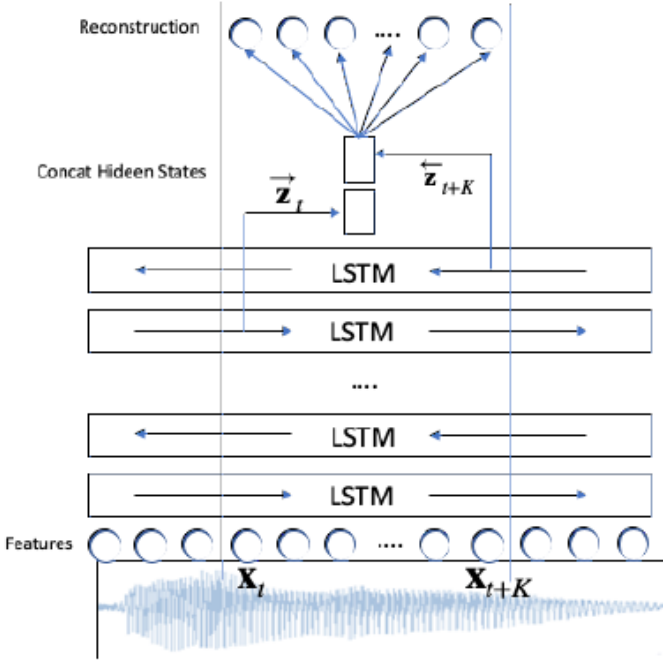


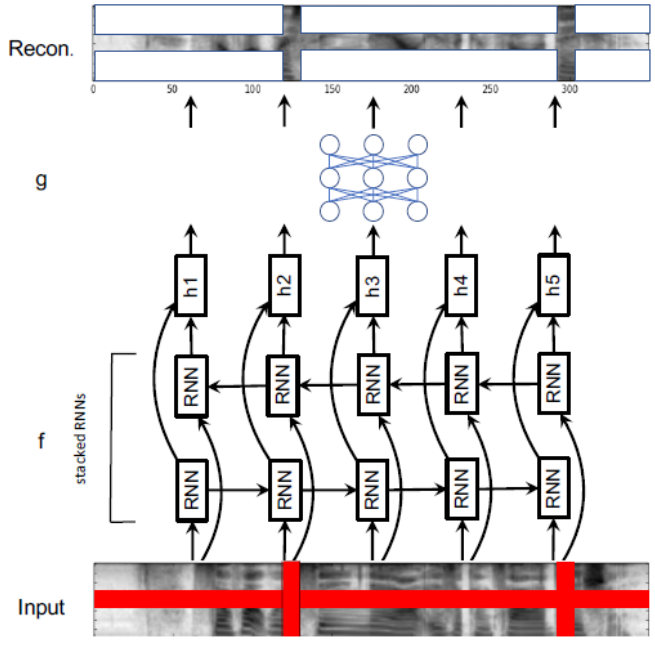
(a) vq-wav2vec (b) Discretized speech training pipeline

A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," ICLR 2020.

Unsupervised pre-trained acoustic representations for ASR

Bidirectional masked reconstruction-based (15-42% relative improvement)

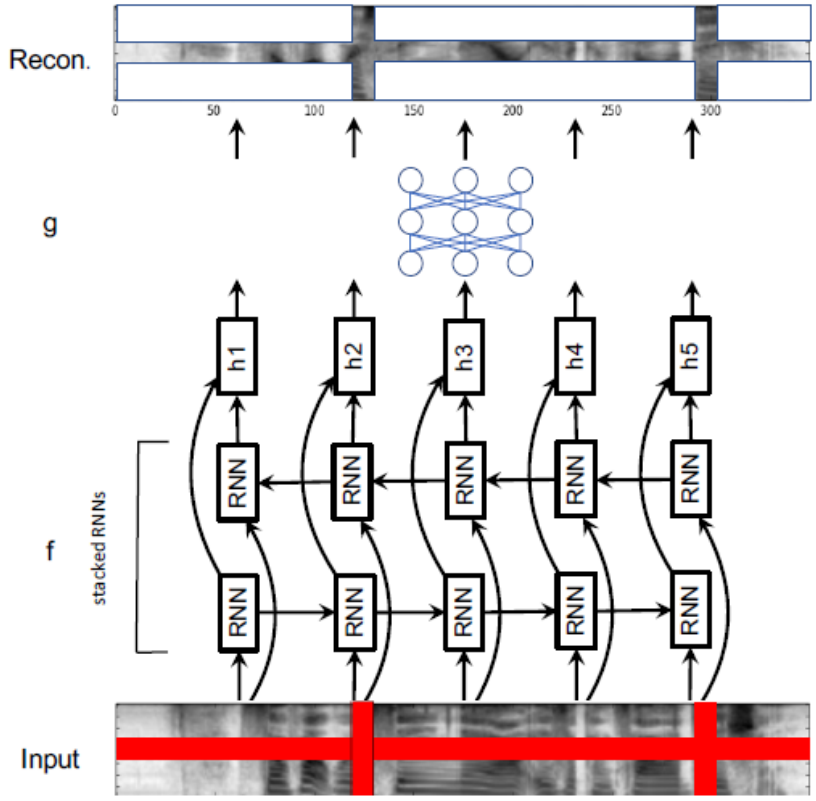




W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," ICASSP 2020.

S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," arXiv:1912.01679.

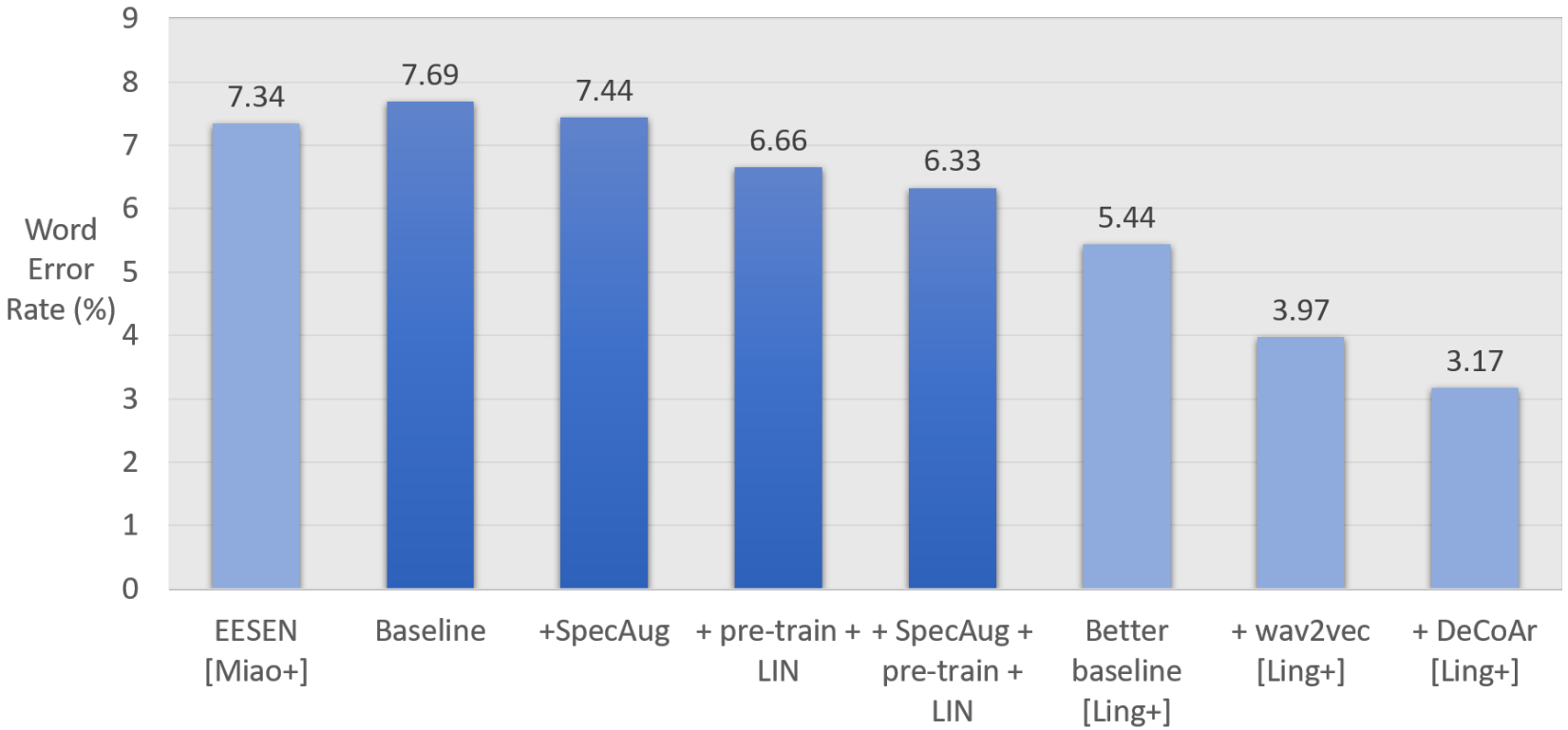
Learning bidirectional encoders via masked reconstruction



$$\mathcal{L}(X, M; f, g) = ||(1 - M) \odot [X - g(f(M \odot X))]\,||^2_{\text{Fro}}$$

Unsupervised pre-trained acoustic representations for ASR: Some results

- Pre-training on LibriSpeech, fine-tuning on WSJ
- Handle domain mismatch via a linear input layer (LIN)

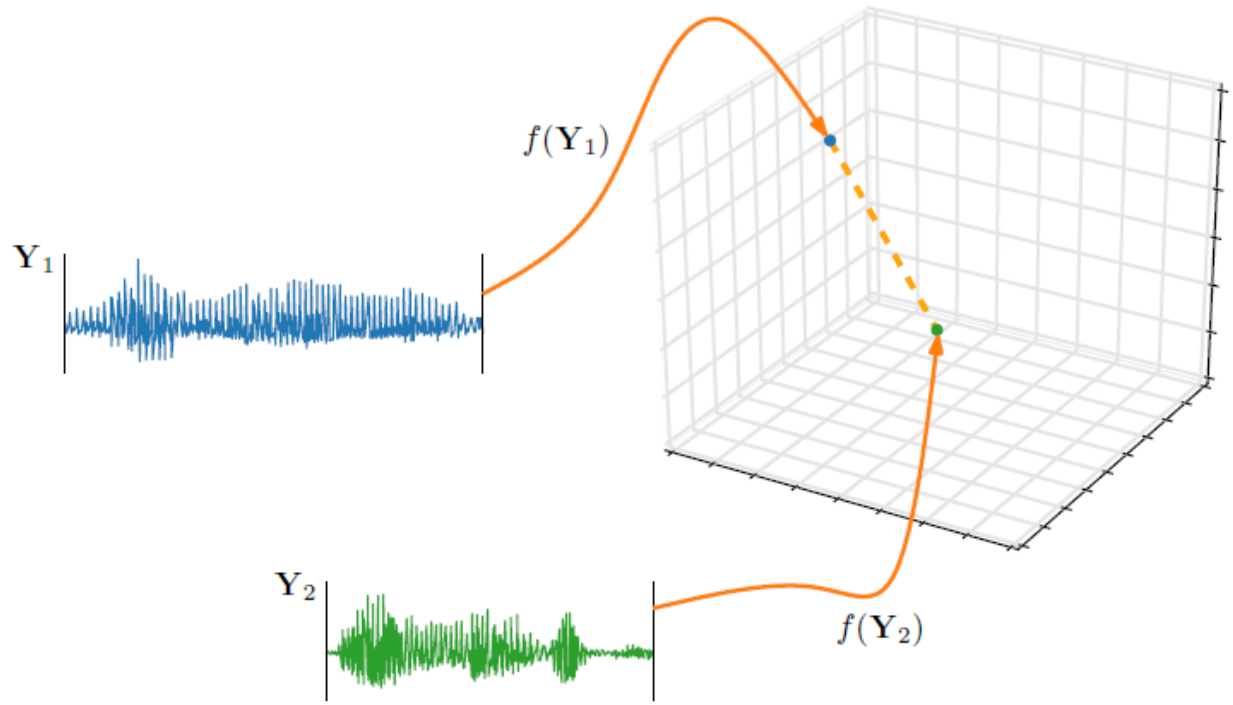


S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," ICASSP 2020.

Acoustic word embeddings

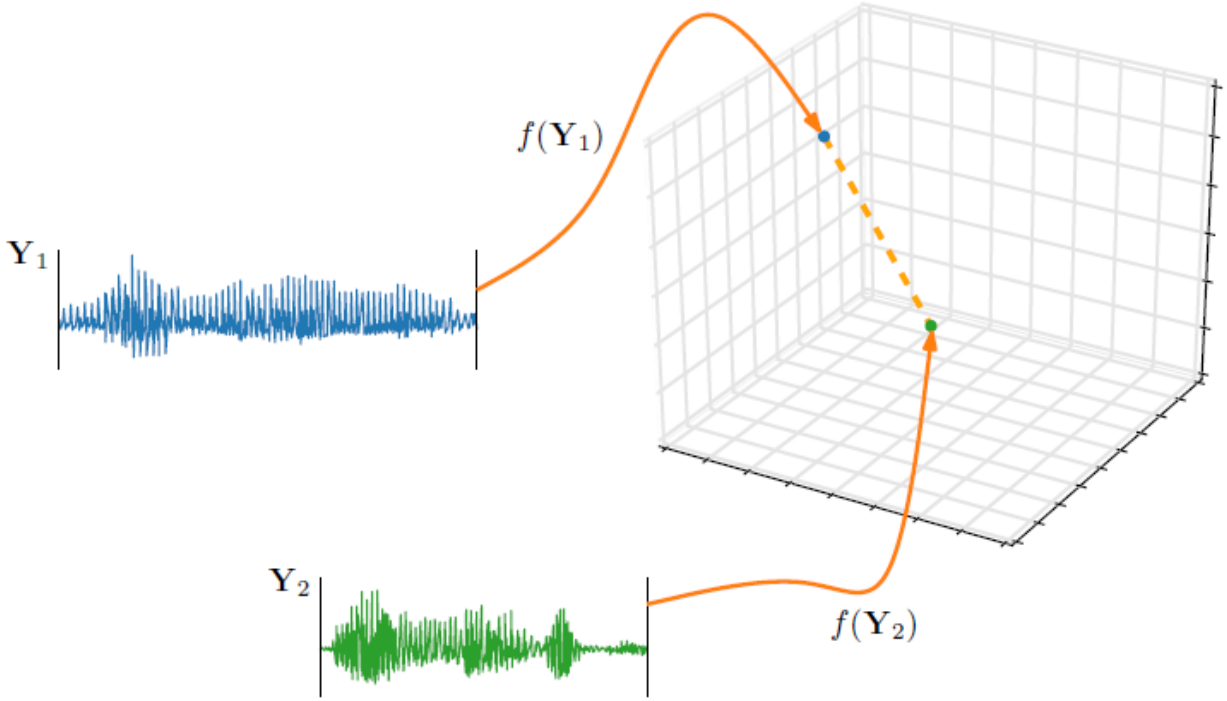
What if we want to represent **entire spoken words**?

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



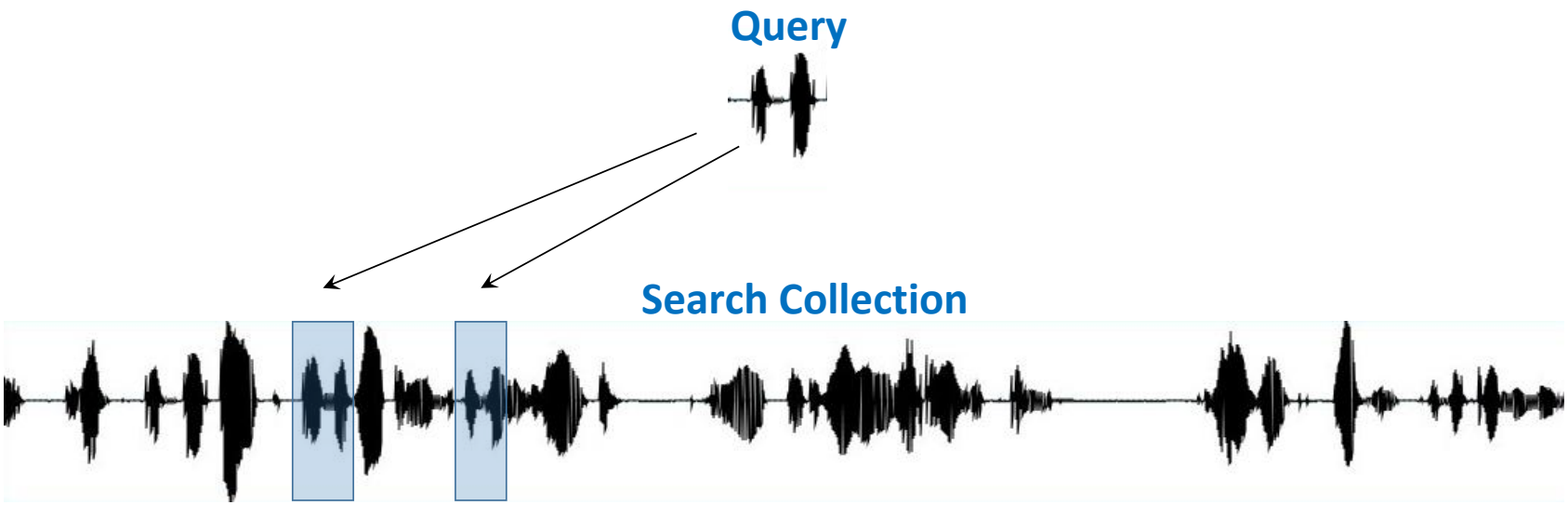
What makes a good acoustic word embedding?

- **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?



Applications of spoken word embeddings

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

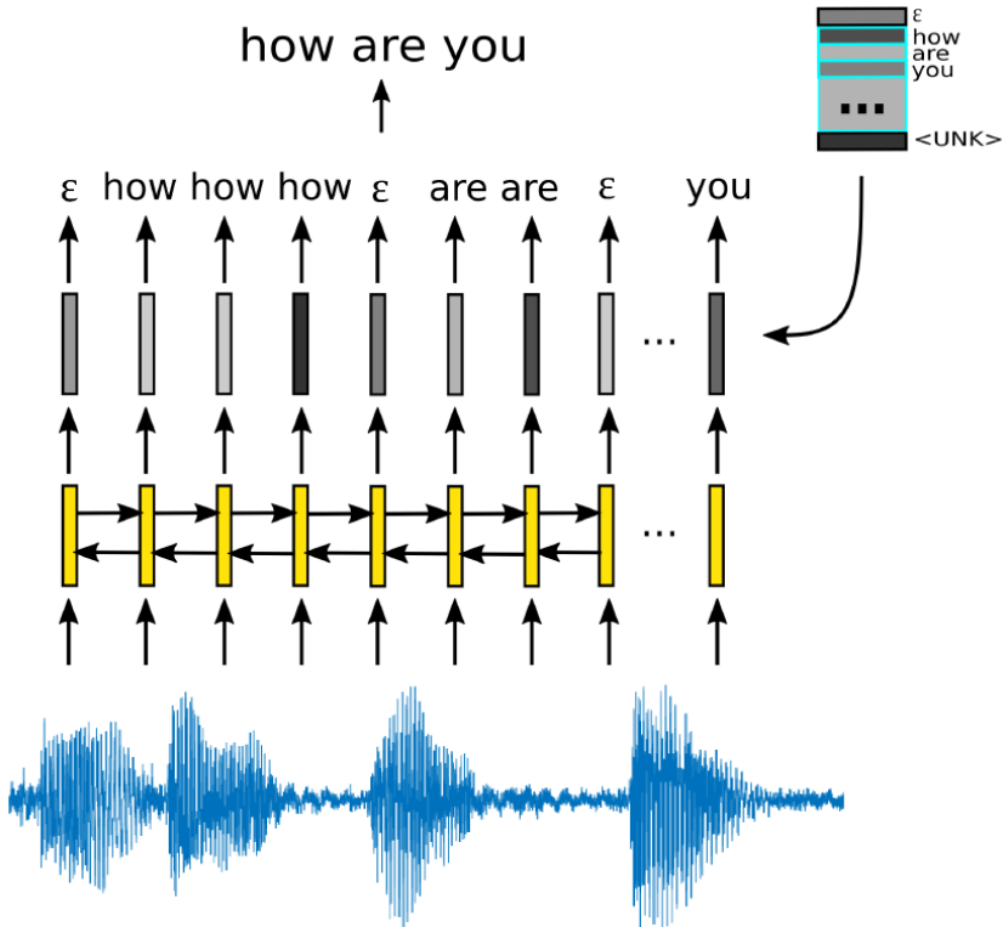


Useful for:

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

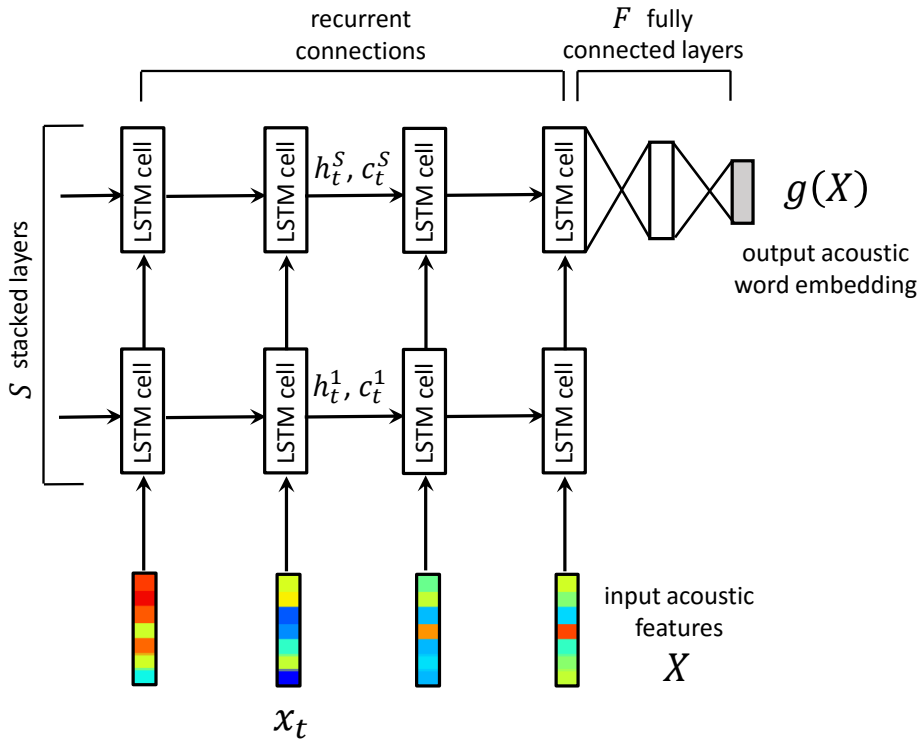
Applications of spoken word embeddings

Whole-word automatic speech recognition



Acoustic word embeddings: Neural models

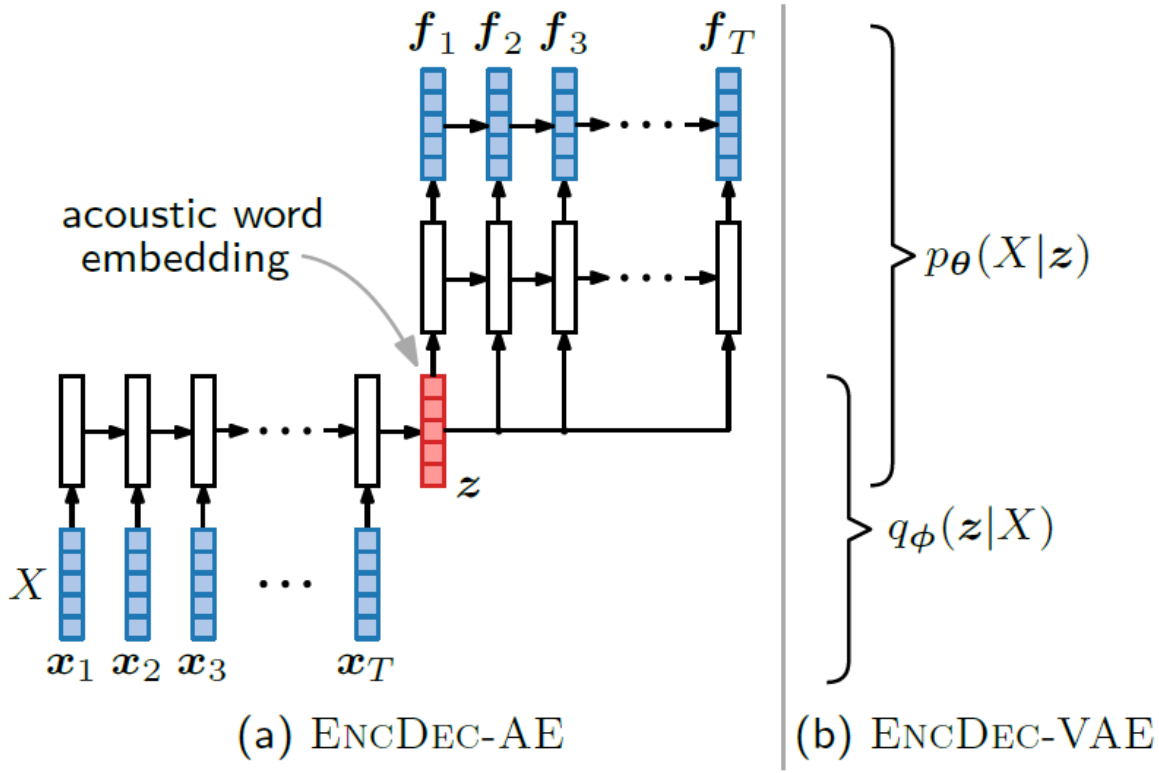
- **Input:** Raw speech features
- **RNN Model:** n_{rec} recurrent + n_{full} fully connected layers
- **Embedding** is activation vector of final layer



S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," SLT 2016

Acoustic word embeddings: Unsupervised learning

- **Training data:** Unlabeled speech waveforms, possibly with automatically discovered word-like segments



H. Kamper, "Truly unsupervised acoustic word embeddings," ICASSP 2019

Acoustic word embeddings: Supervised learning

- **Training data:** Pairs of same-word speech waveforms
- **Contrastive (triplet) loss:**
  - Bring together same-word pairs, while separating different ones by some margin

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

where $\mathbf{x}^-$ = random (or hard) negative example, m = margin

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