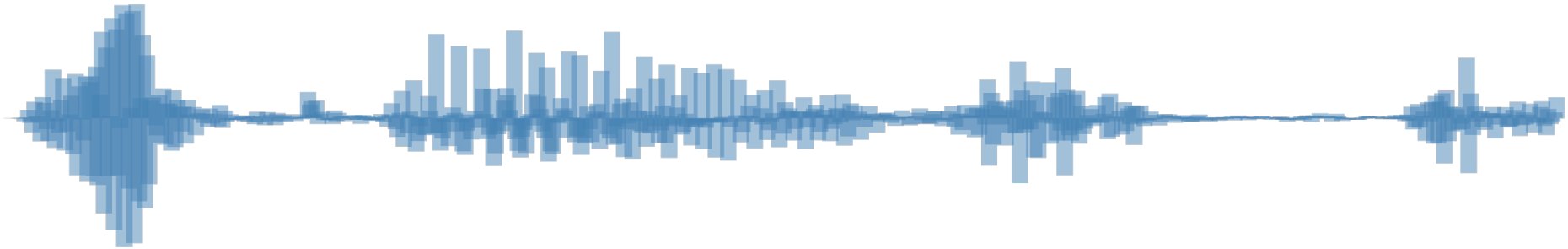
Models for End-to-end Speech Recognition

Awni Hannun
awni@cs.stanford.edu



What is End-to-End?

- First pass
- Minimal input preprocessing
- Minimal output processing
- External language model?

Connectionist Temporal Classification (CTC)*

- *Alignment Free*: Overcomes alignment issue by marginalizing over all allowed alignments between X and Y.

$$p(Y \mid X) = \sum_{A \in \mathcal{A}_{X,Y}} \prod_{t=1}^{T} p_t(a_t \mid X)$$

The CTC conditional **probability**

marginalizes over the set of valid alignments

computing the **probability** for a single alignment step-by-step.

*Graves, et al., 2006

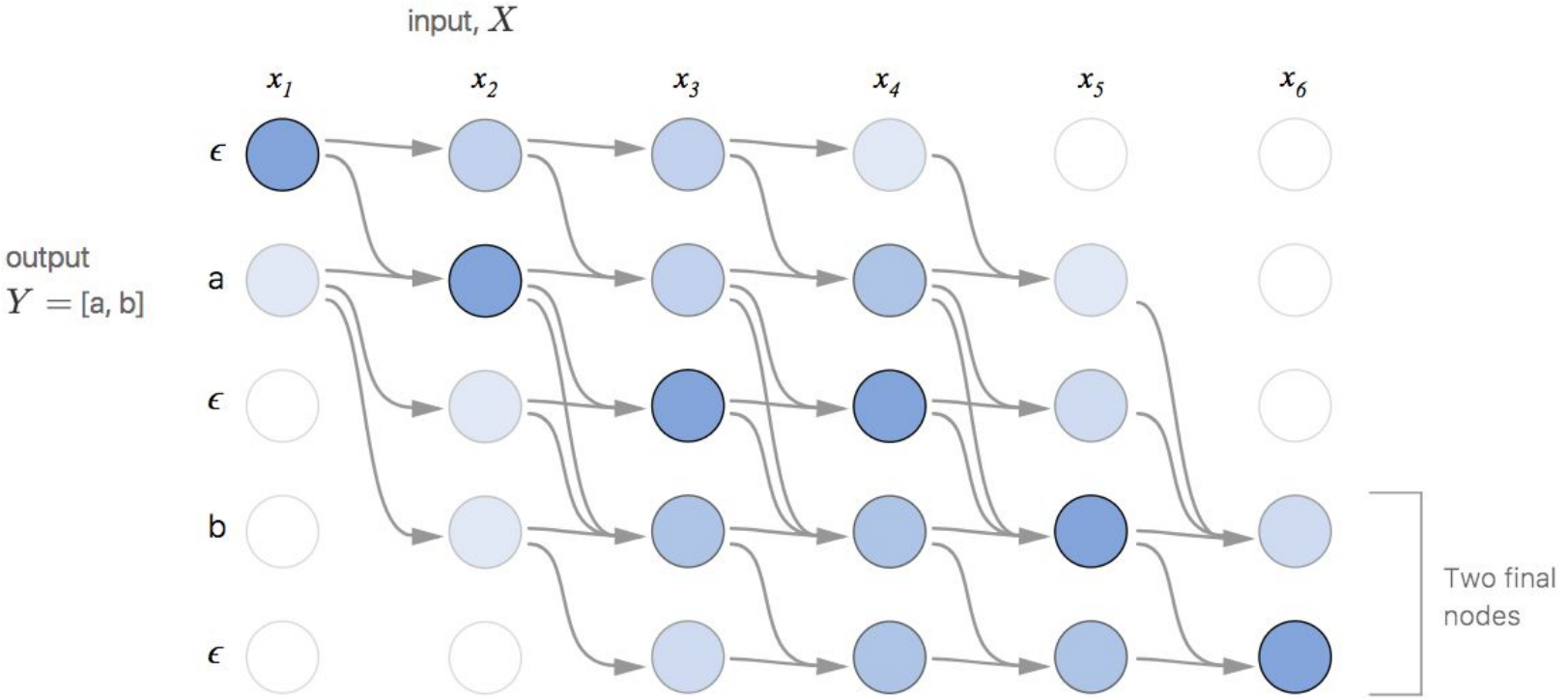
CTC Alignments

First, merge repeat characters.

Then, remove any ϵ tokens.

The remaining characters are the output.

CTC: Dynamic Programming



Node (s, t) in the diagram represents $\alpha_{s,t}$ – the CTC score of the subsequence $Z_{1:s}$ after t input steps.

Inference

- To find the best transcript, solve the following optimization problem:

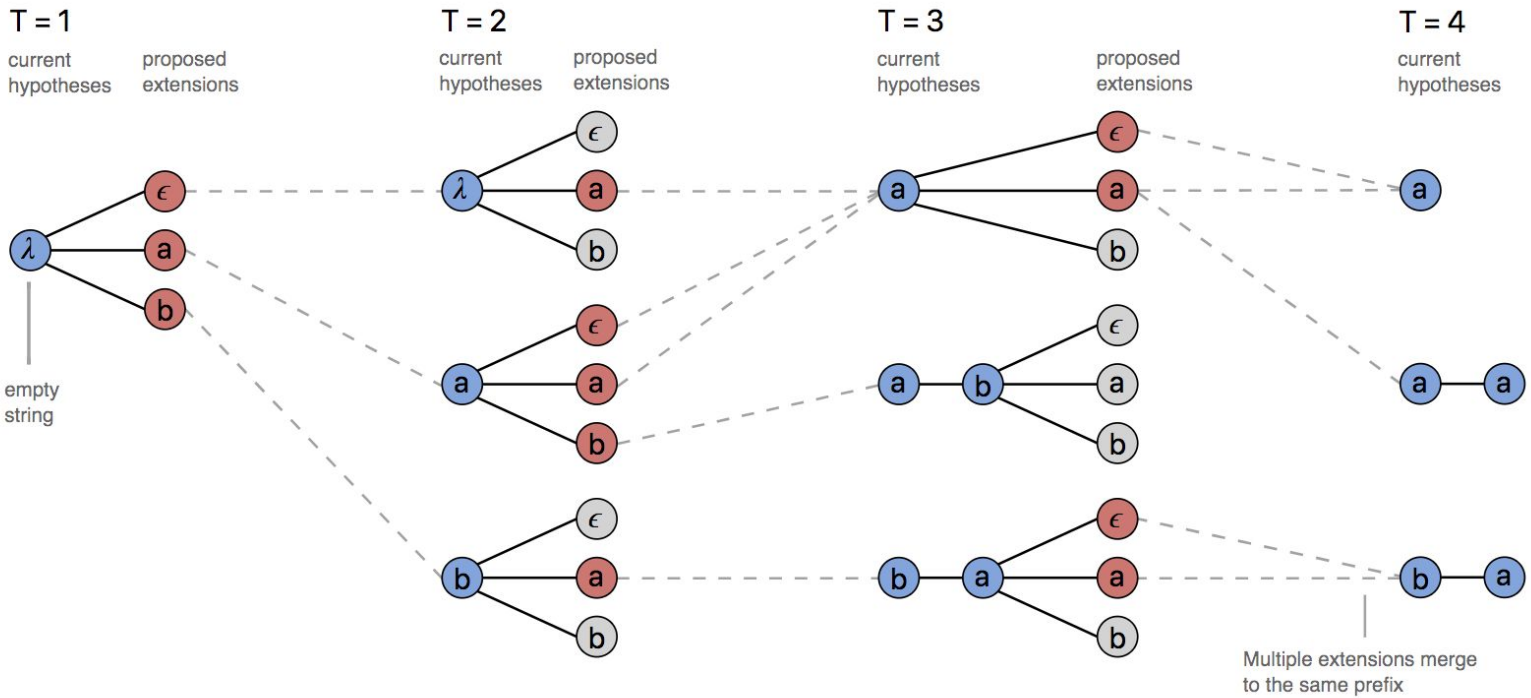
$$Y^* = \underset{Y}{\operatorname{argmax}} \; p(Y \mid X) \cdot p(Y)^{\alpha} \cdot L(Y)^{\beta}$$

The CTC conditional probability.

The language model probability.

The "word" insertion bonus.

Inference: CTC Beam Search



The CTC beam search algorithm with an output alphabet $\{\epsilon, a, b\}$ and a beam size of three.

Properties of CTC

- Outputs are conditionally independent given input

If we predict an 'A' as the first letter then the suffix 'AA' should get much more probability than 'riple A'. If we predict 't' first, the opposite should be true.

Example from: *Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition*, W. Chan, N. Jaitly, Q.V. Le, O. Vinyals, ICASSP, 2016

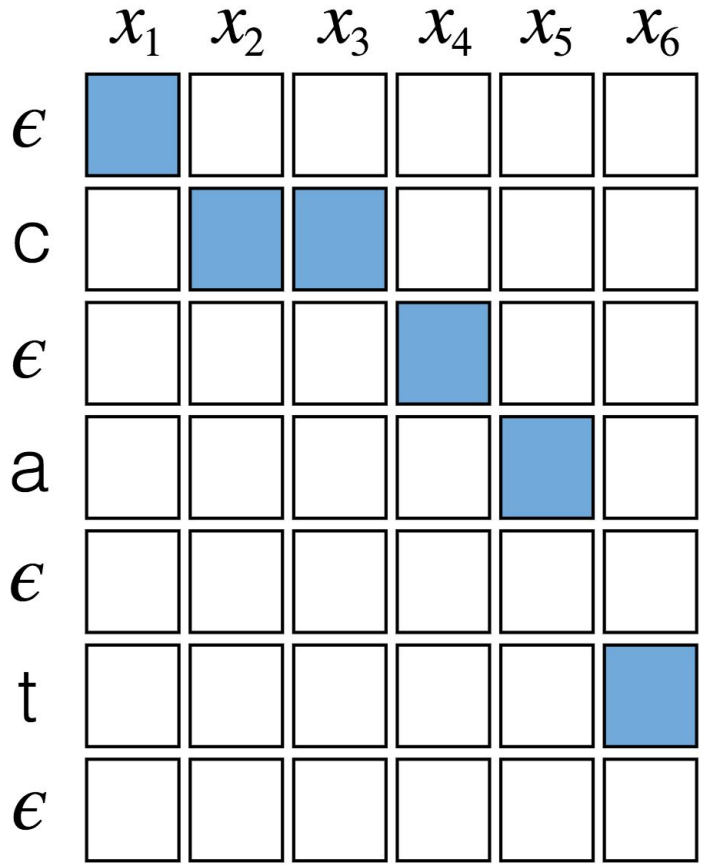
Properties of CTC

Many-to-one

- Many inputs can align to at most one output
- Implication: input must be longer than output

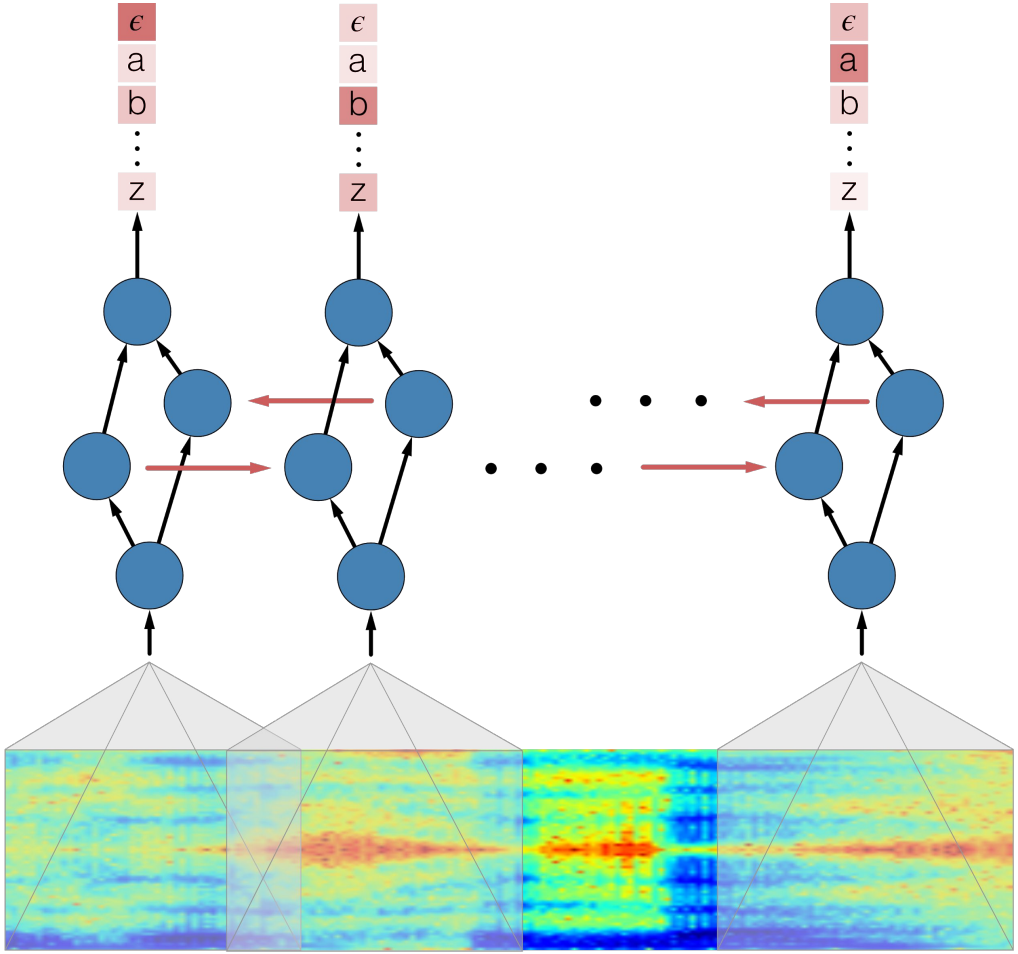
Monotonic

- Alignments are monotonic
- Can't move backwards in the output when moving forward in input



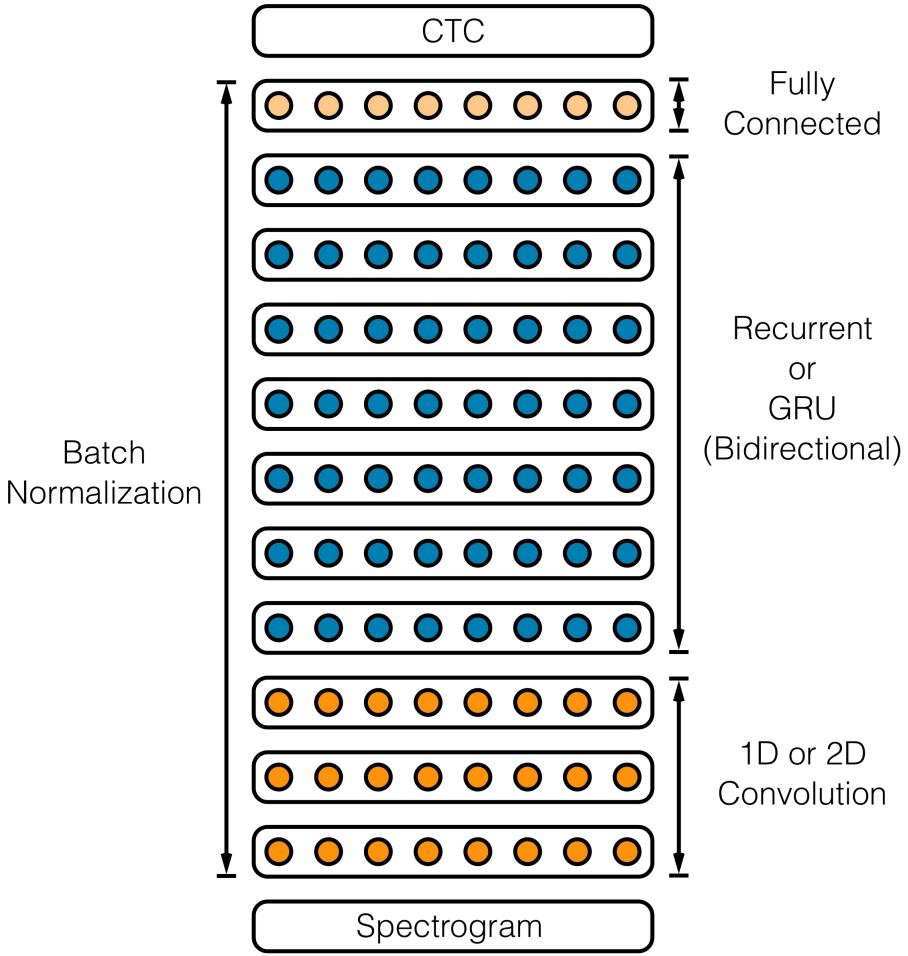
Alignment-free Speech Recognition

- Input spectrogram
- Output characters
- Train with CTC



Scaling up End-to-End Speech Recognition

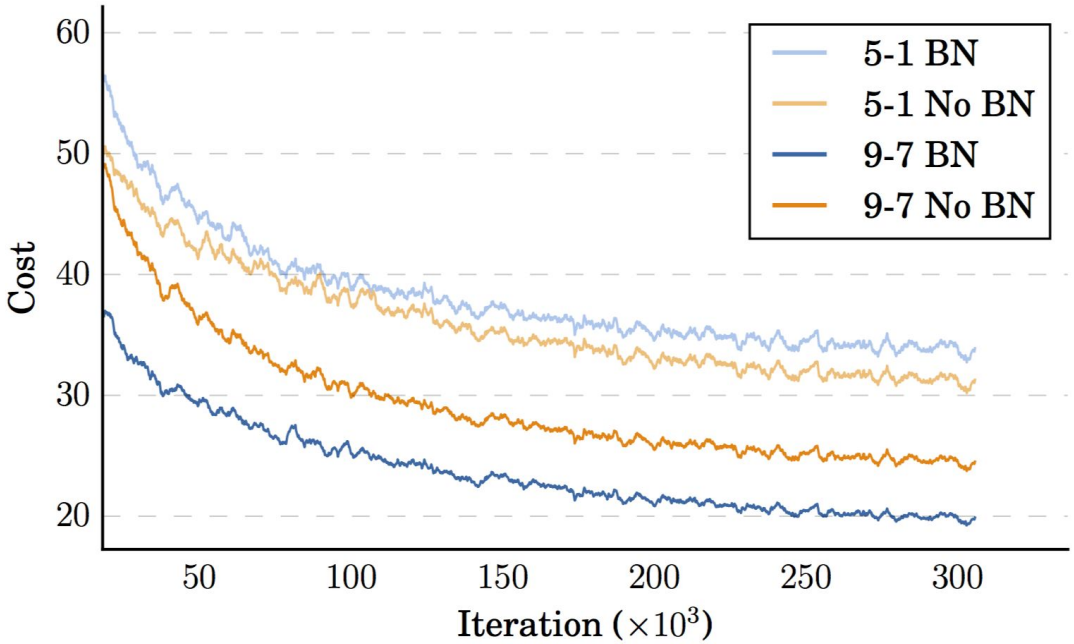
- Algorithmic improvements to improve convergence and generalization:
  - Convolutions in early layers.
  - Recurrent Batch Normalization
  - Length-based curriculum learning



Scaling up End-to-End Speech Recognition

Recurrent Batch Normalization

- Sequence BN:
  - Compute statistics over full sequence
  - Apply transform between RNN layers
- Improves convergence for deep models



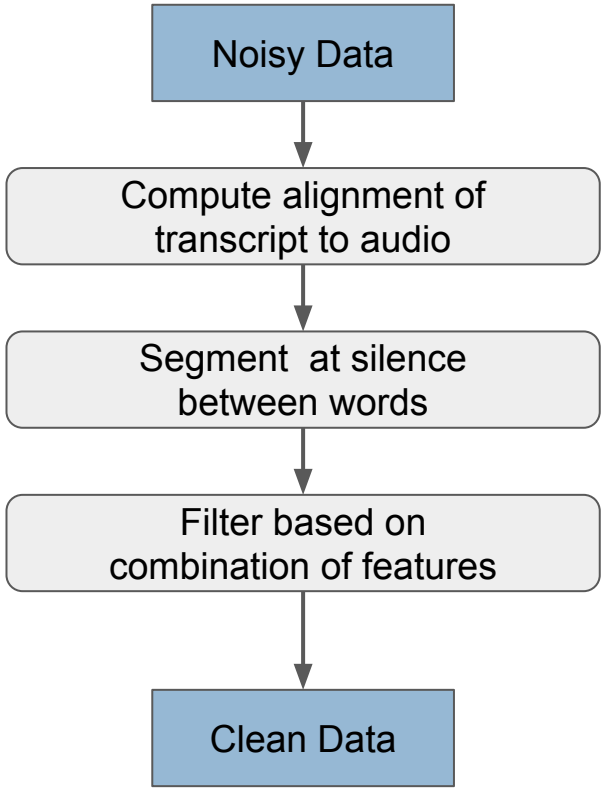
Scaling up End-to-End Speech Recognition

Data capture pipeline

Solve for CTC Viterbi alignment

Use stretches of ϵ between words to denote silence.

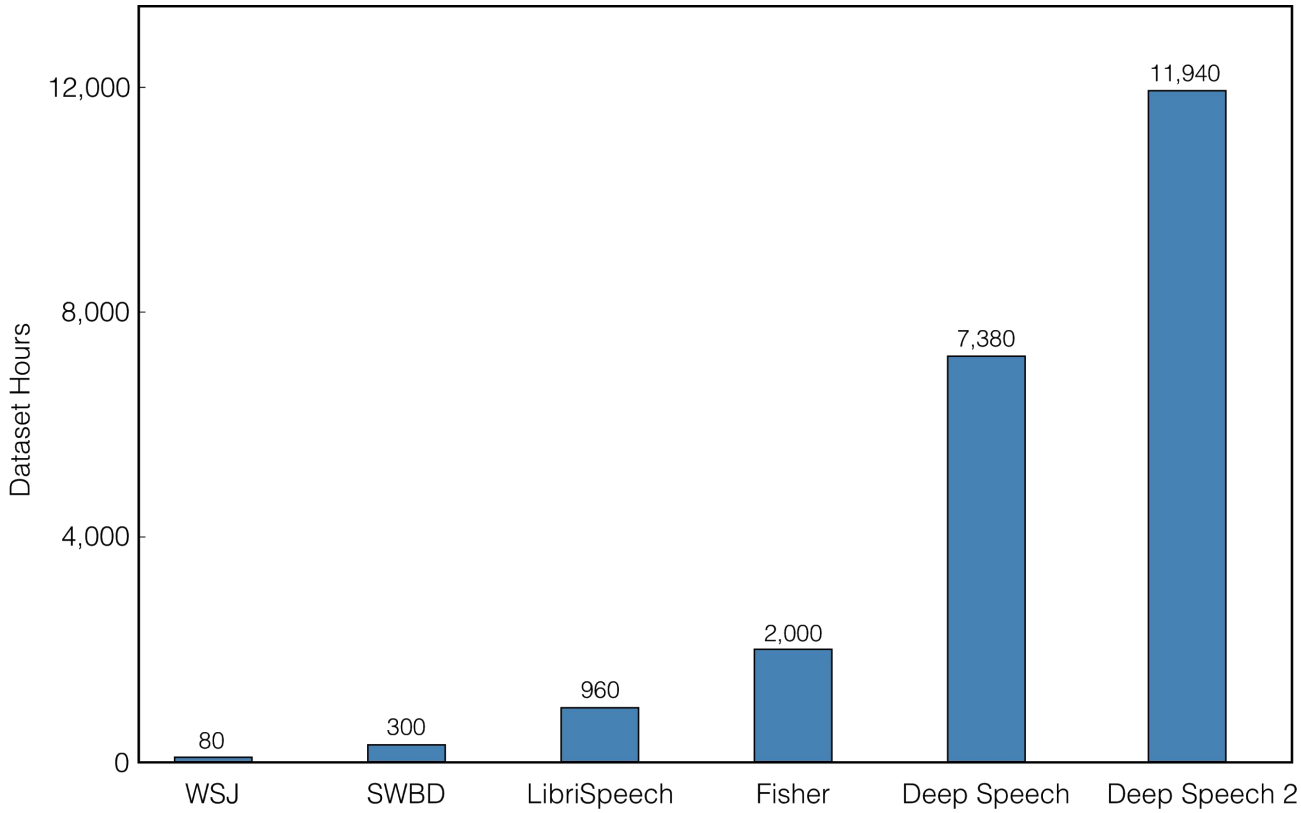
Reduces WER from 17% to 5% retaining 50% of the data.



Scaling up End-to-End Speech Recognition

English Dataset Sizes

- 12,000 hours
- 8 million utterances
- Average length is 7 seconds

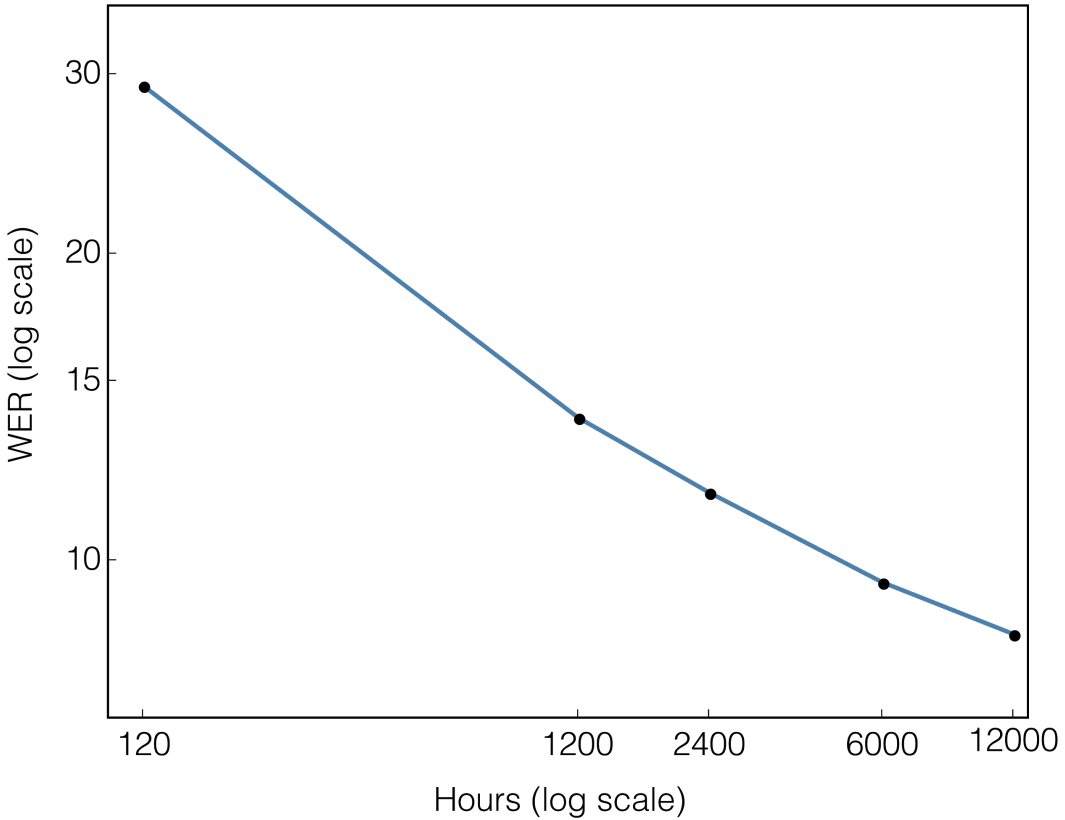


Scaling up End-to-End Speech Recognition

WER by dataset size (hours)

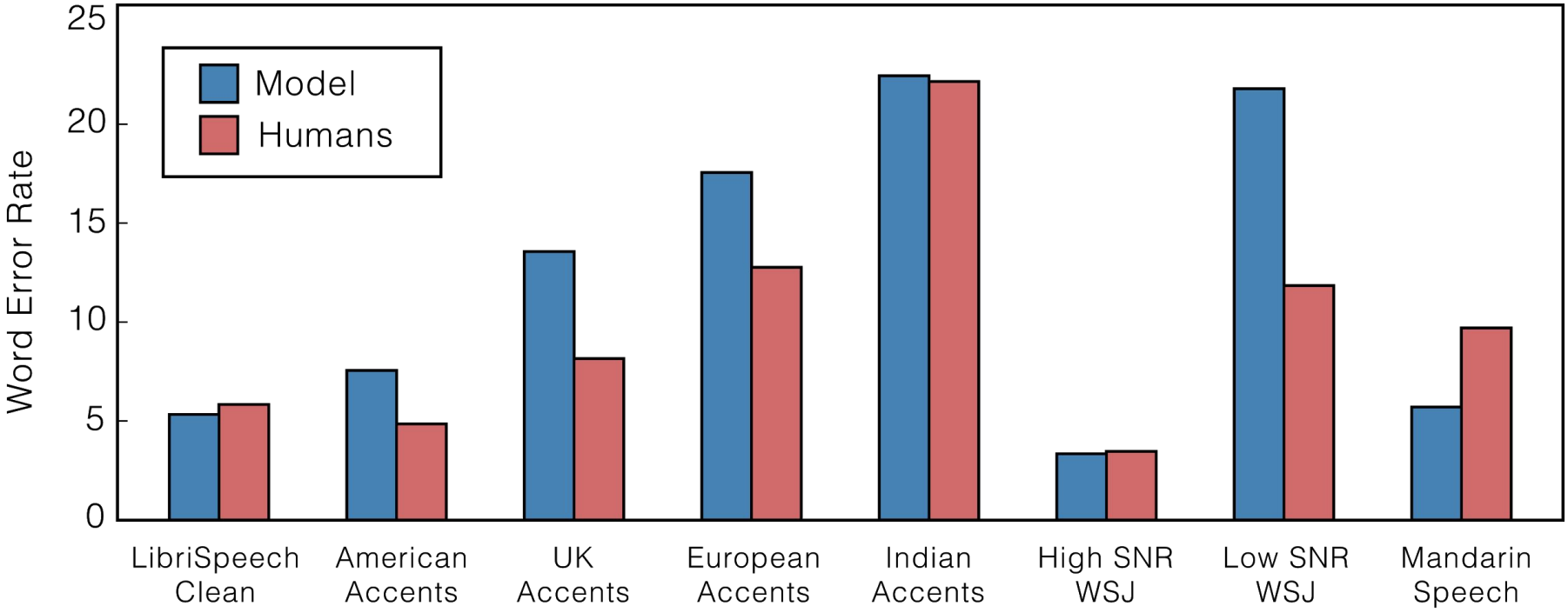
- Generalization improves as power law with dataset size

$$\mathcal{L}_{\mathrm{WER}} = |D|^{-\alpha}$$



Scaling up End-to-end Speech Recognition

Comparison to human transcribers

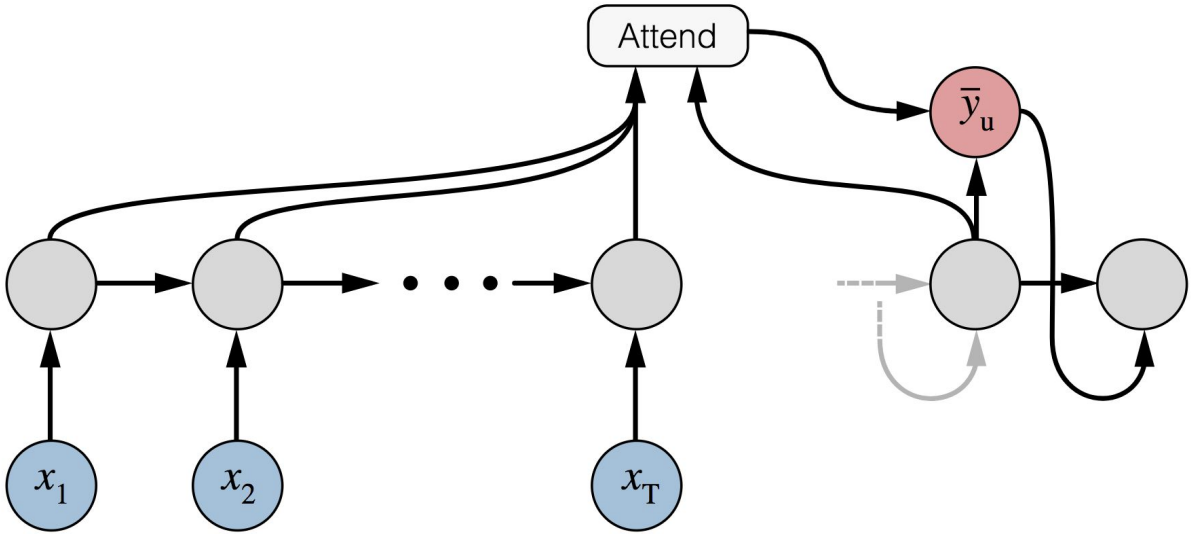


Alternative Models

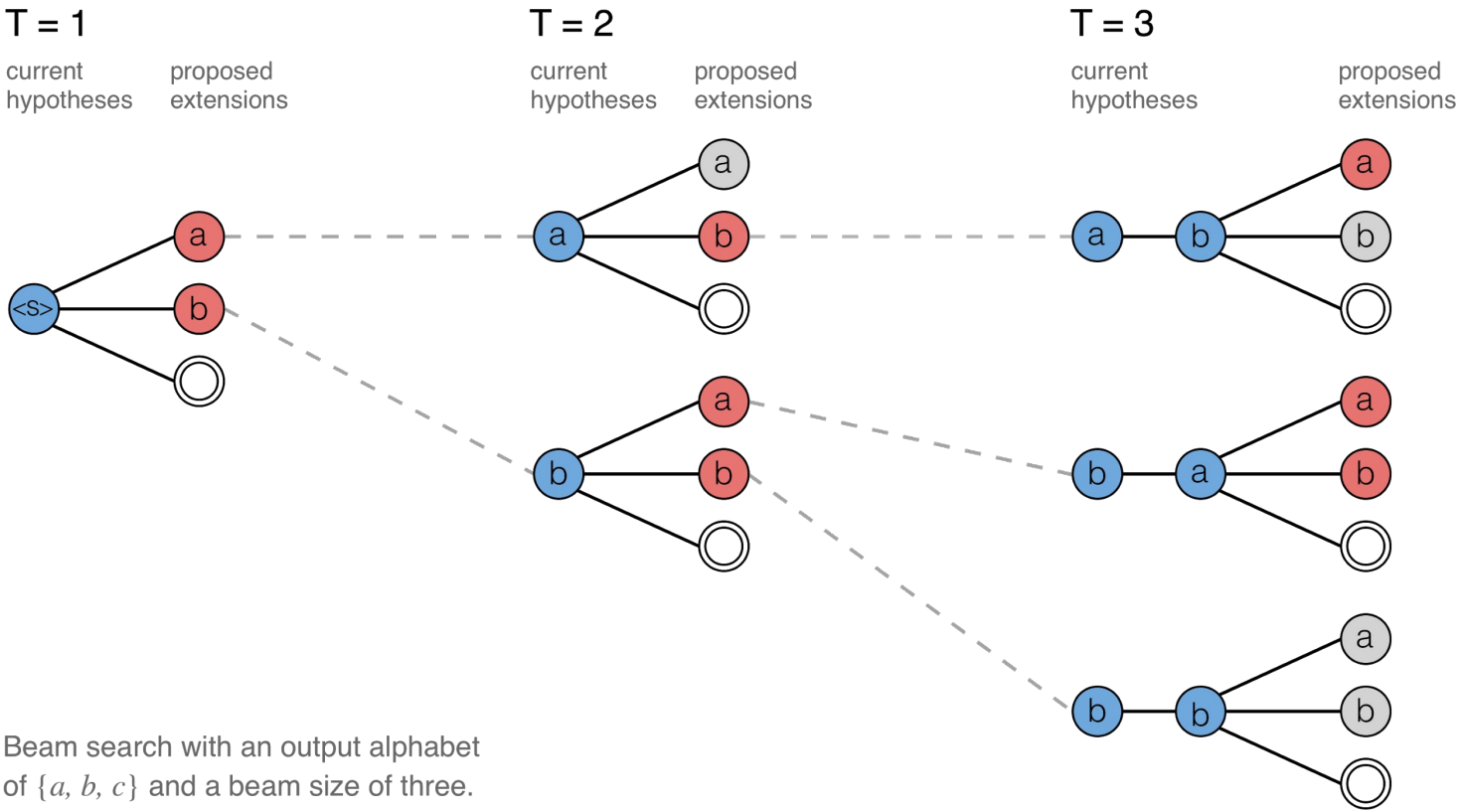
- What is the best sequence model for “end-to-end” speech?
- Alternatives include:
  - CTC
  - Seq2seq with attention
  - RNN transducer
  - Other variations
- Trade-offs not well understood

Sequence-to-sequence

- No conditional independence assumption
- No monotonic alignment assumption
- Output length can be longer than input length
    - Subsample input a lot more



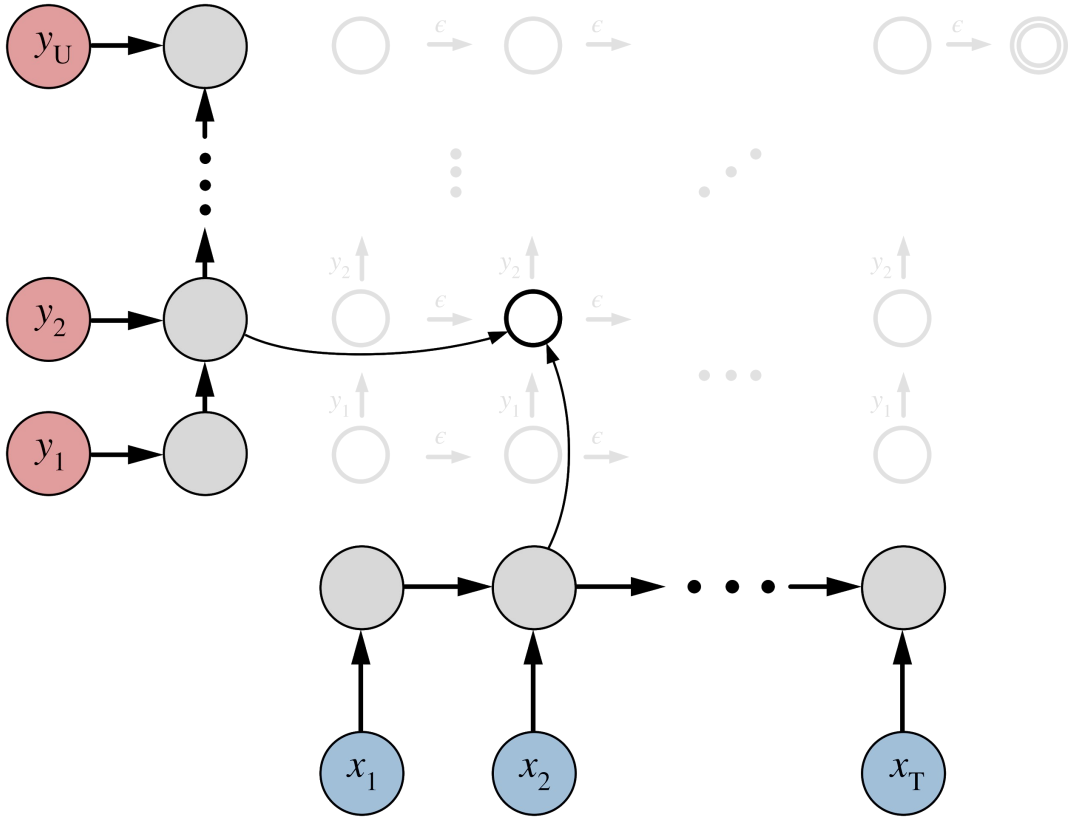
Inference: Sequence-to-sequence



Beam search with an output alphabet of $\{a, b, c\}$ and a beam size of three.

RNN Sequence Transducer

- Encoder and decoder
- Combine every input-output pair



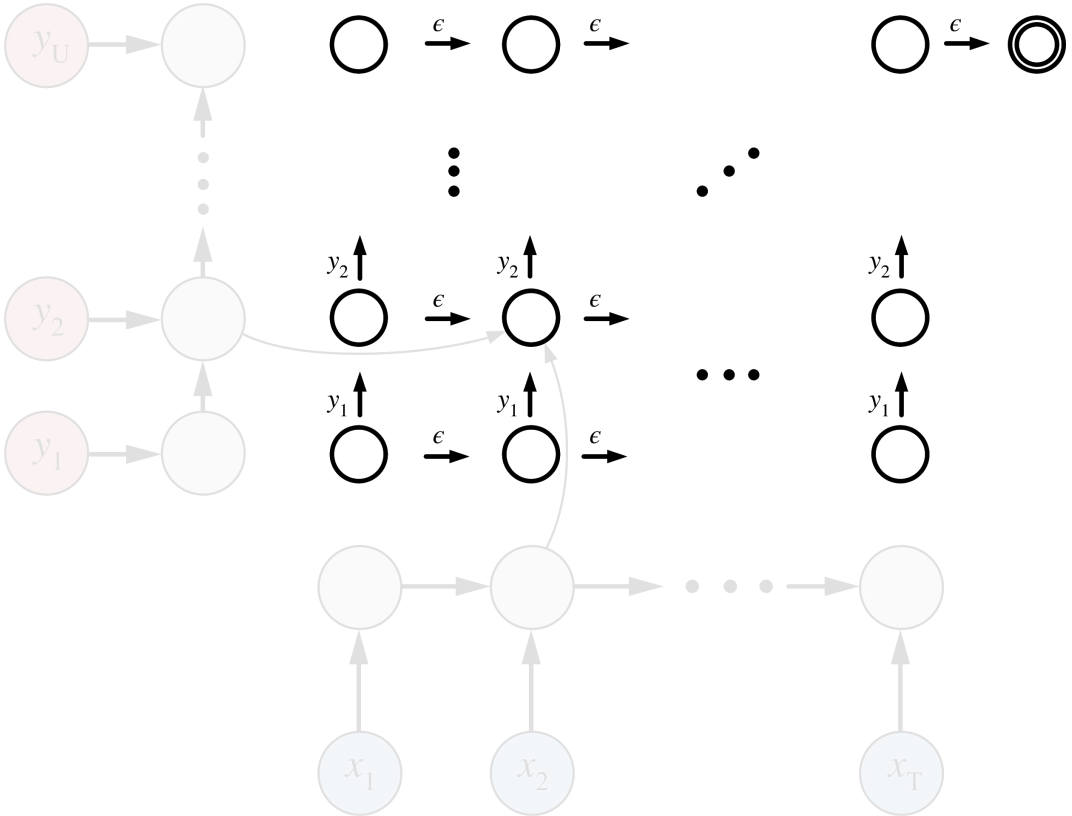
RNN Sequence Transducer

Like CTC

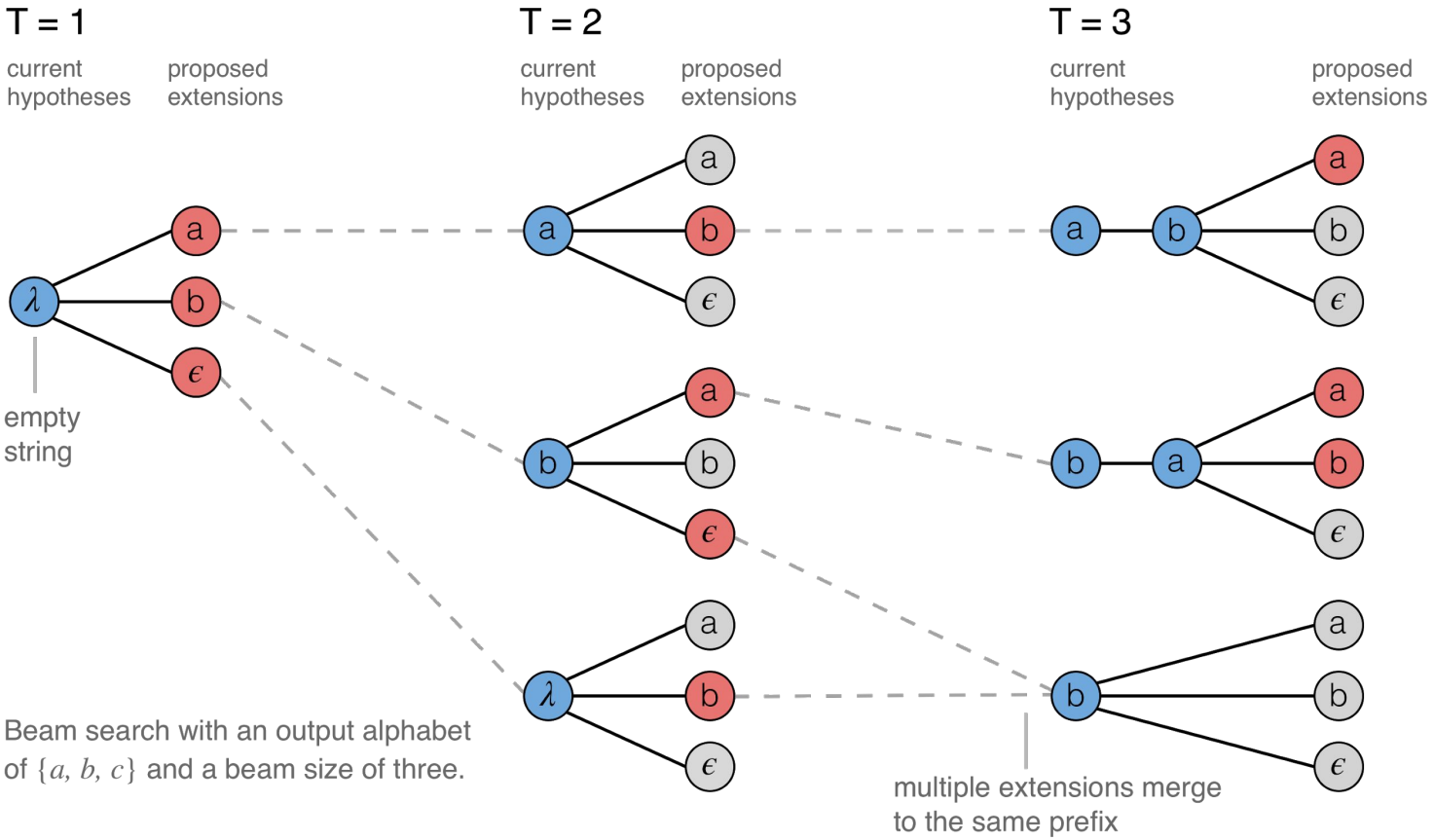
- Sum over all possible alignments
- Monotonic alignments

Like Seq2Seq

- No conditional independence assumption
- Output can be longer than input



Inference: RNN Sequence Transducer



Inference: RNN Sequence Transducer

Like Seq2Seq

- Decode with beam search over outputs (and ϵ)
- Stop when `beam size` complete hypotheses with scores greater than anything left in the beam

Like CTC

- Merge alignments with the same prefix
- Hypothesis complete after processing all the input
    - Consume input step when ϵ is proposed

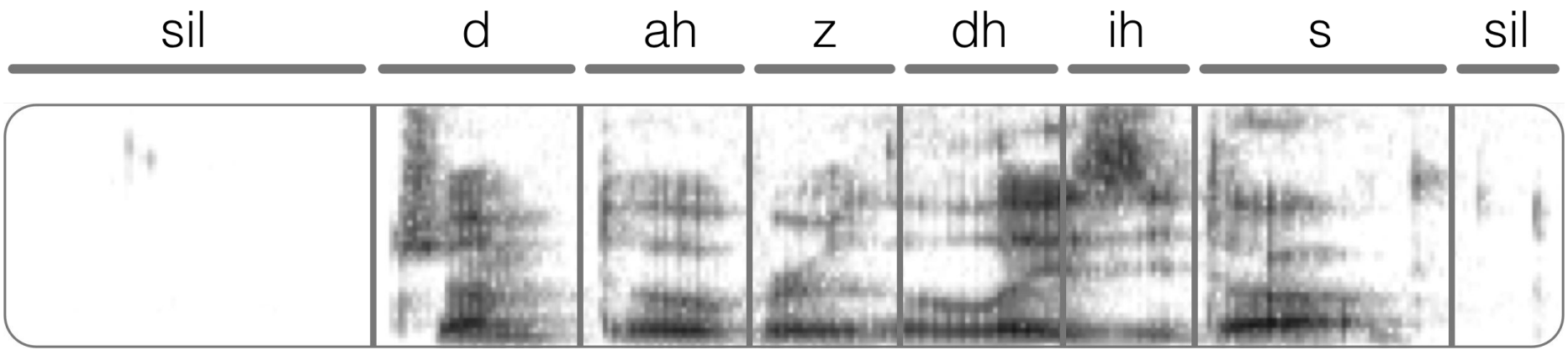
Learning to Align

Alignment Demos

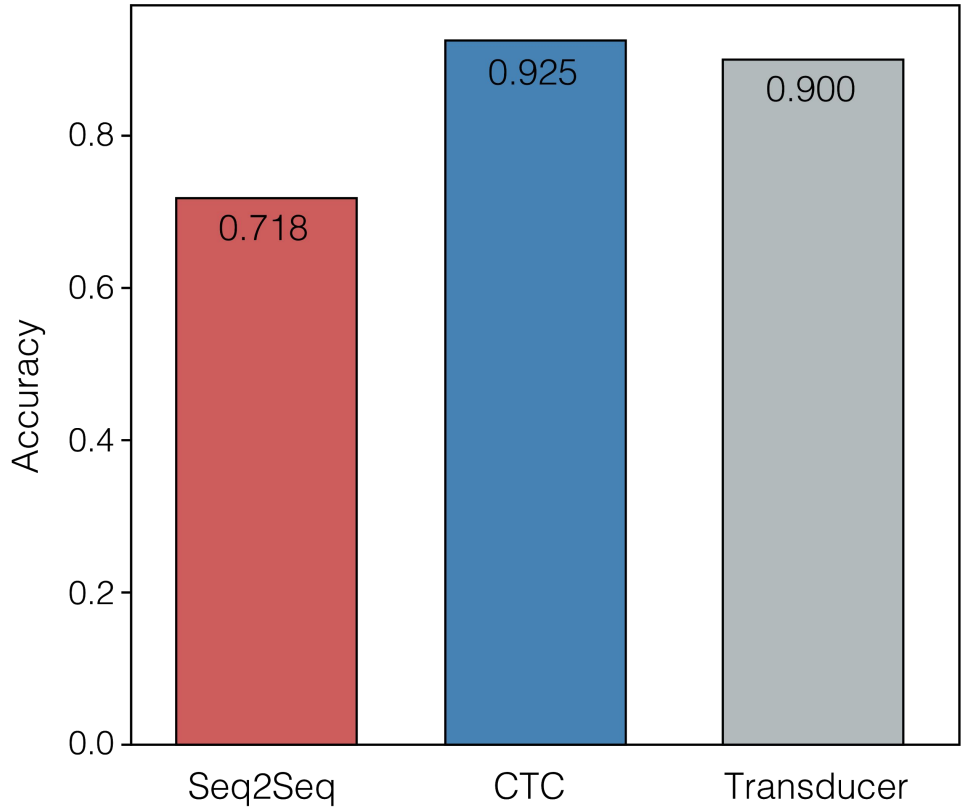
- CTC: file:///Users/awni/Desktop/misc_visualizations/vis_alignments/ctc.html
- Seq2Seq: file:///Users/awni/Desktop/misc_visualizations/vis_alignments/seq2seq.html
- RNN Transducer: file:///Users/awni/Desktop/misc_visualizations/vis_alignments/trans.html

Alignment Quality

- TIMIT: phonemes are aligned to input (dense labeling)
- Compute an alignment for each model
- Correct if phoneme aligned within labeled range

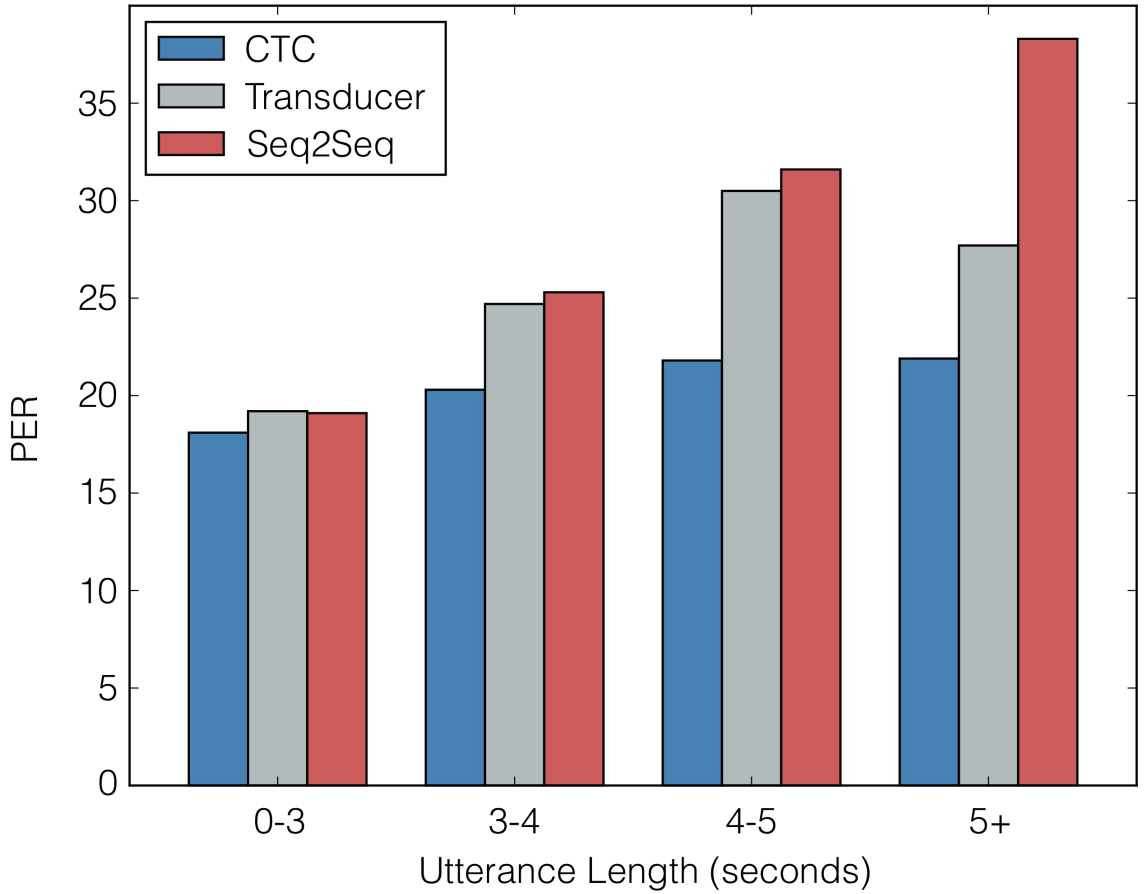


Alignment Quality

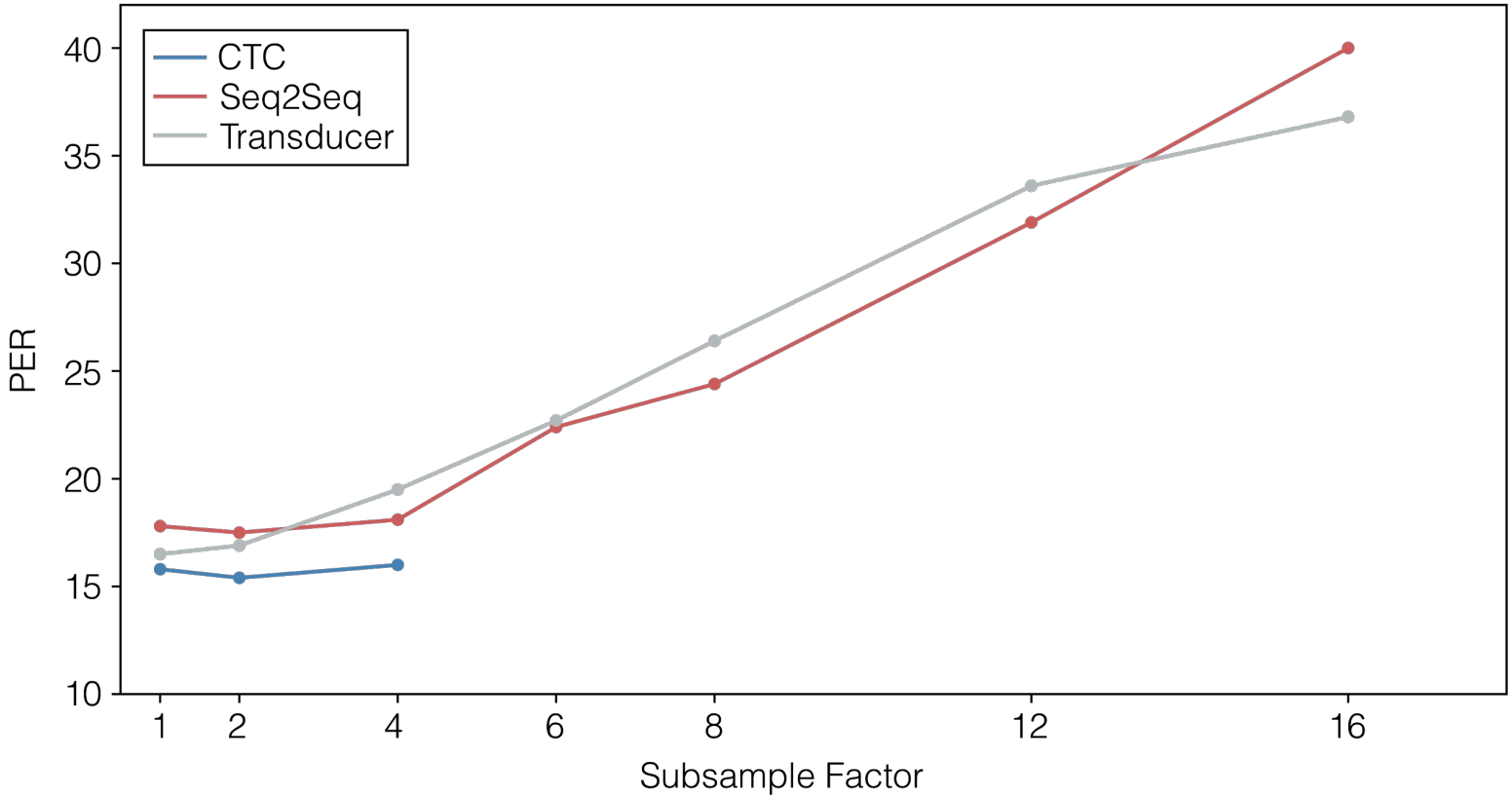


Utterance Length

- Train on utterances with length < 3 seconds
- Evaluate models on longer utterances



Subsampling



Thanks!