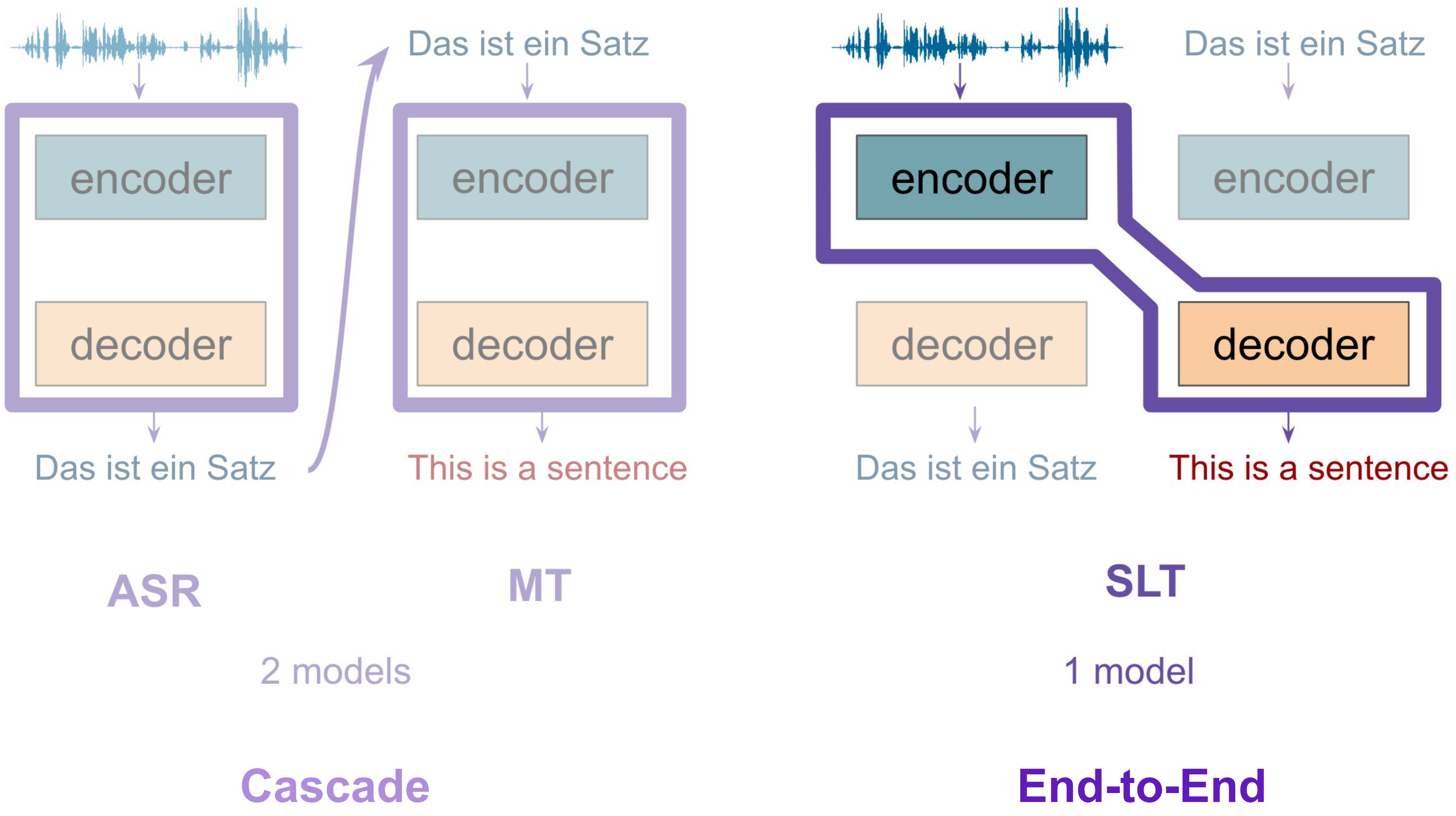
Sec 2.3

Architecture & Modifications

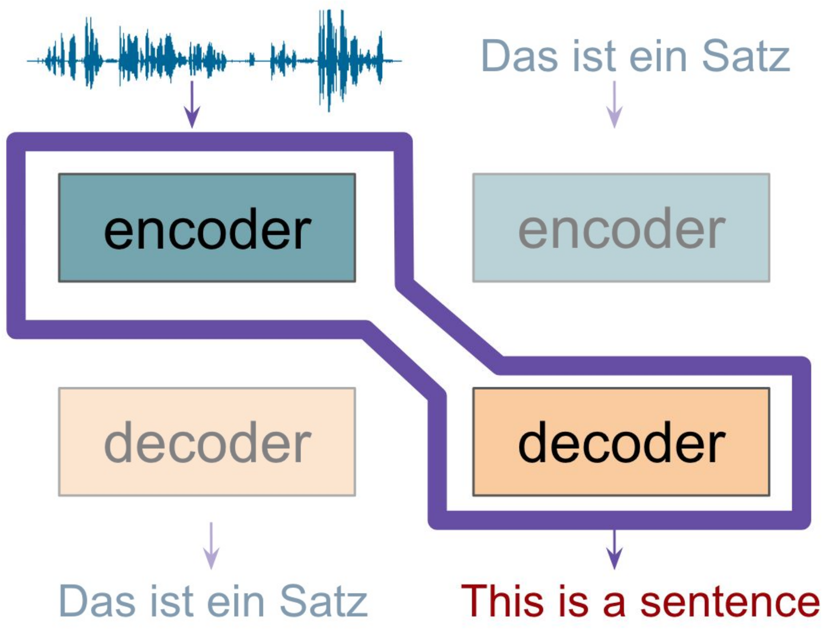
End-to-End Architecture



End-to-End Architecture

LSTM or Transformer
Encoder-Decoder Models

However, speech ≠ text

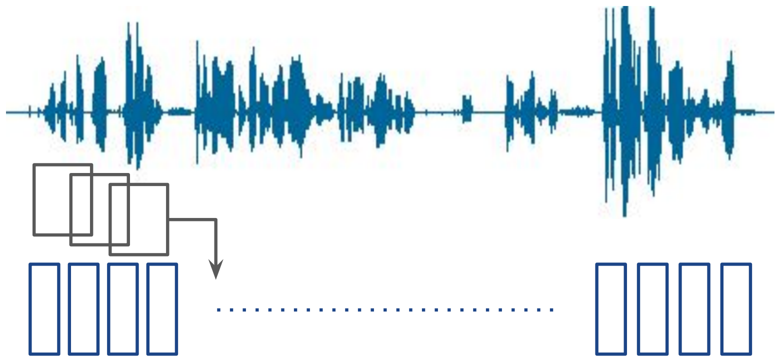


SLT

1 model

End-to-End

Speech vs. Text

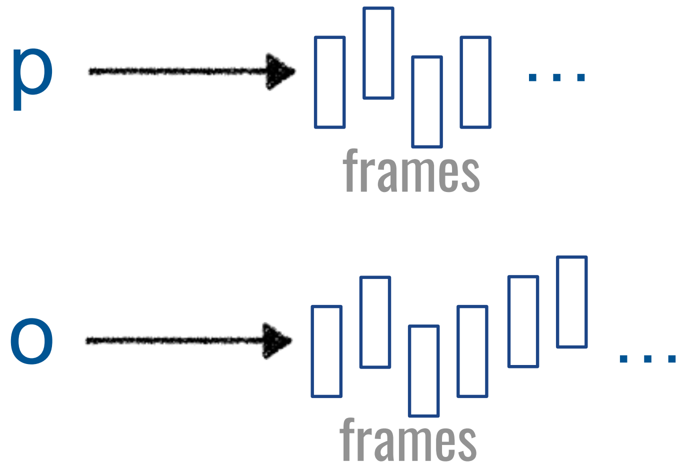


Discretized audio — speech frames

Speech features ~8-10x longer than
the equivalent character sequences

c h a r a c t e r s

SPEECH:



TEXT:

p → p

Each feature vector is unique,
Number of feature vectors per phone varies

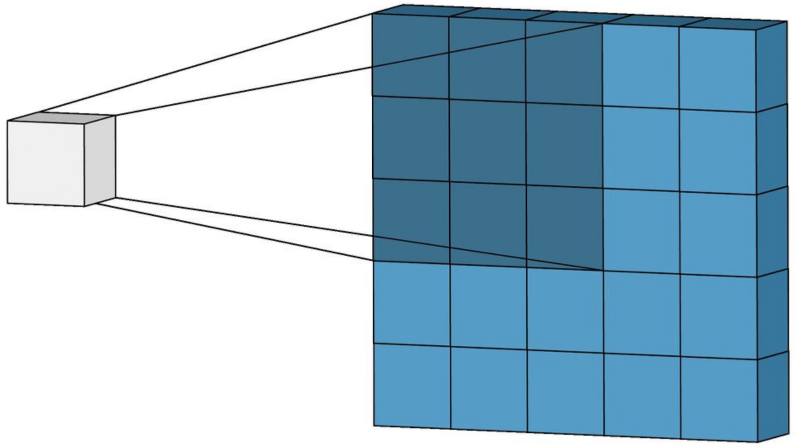
Challenges

- Sequence length:
  - increased memory requirements
  - greater distance between dependencies
- Redundancy:
  - adds task for model to learn
- Variation:
  - requires more data for model to learn correspondences

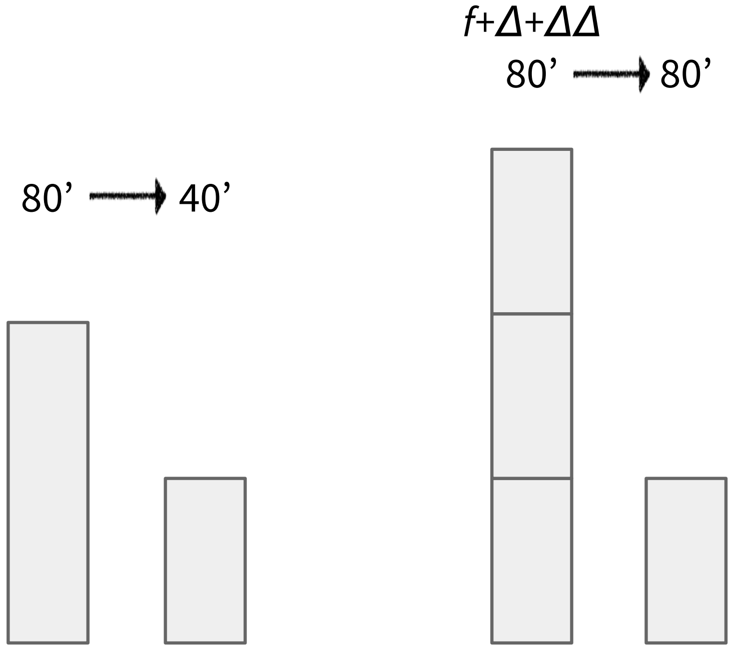
Dimensionality Reduction

Two directions: ① temporal and ② feature dimension

Convolutional layers enable *fixed-length downsampling*



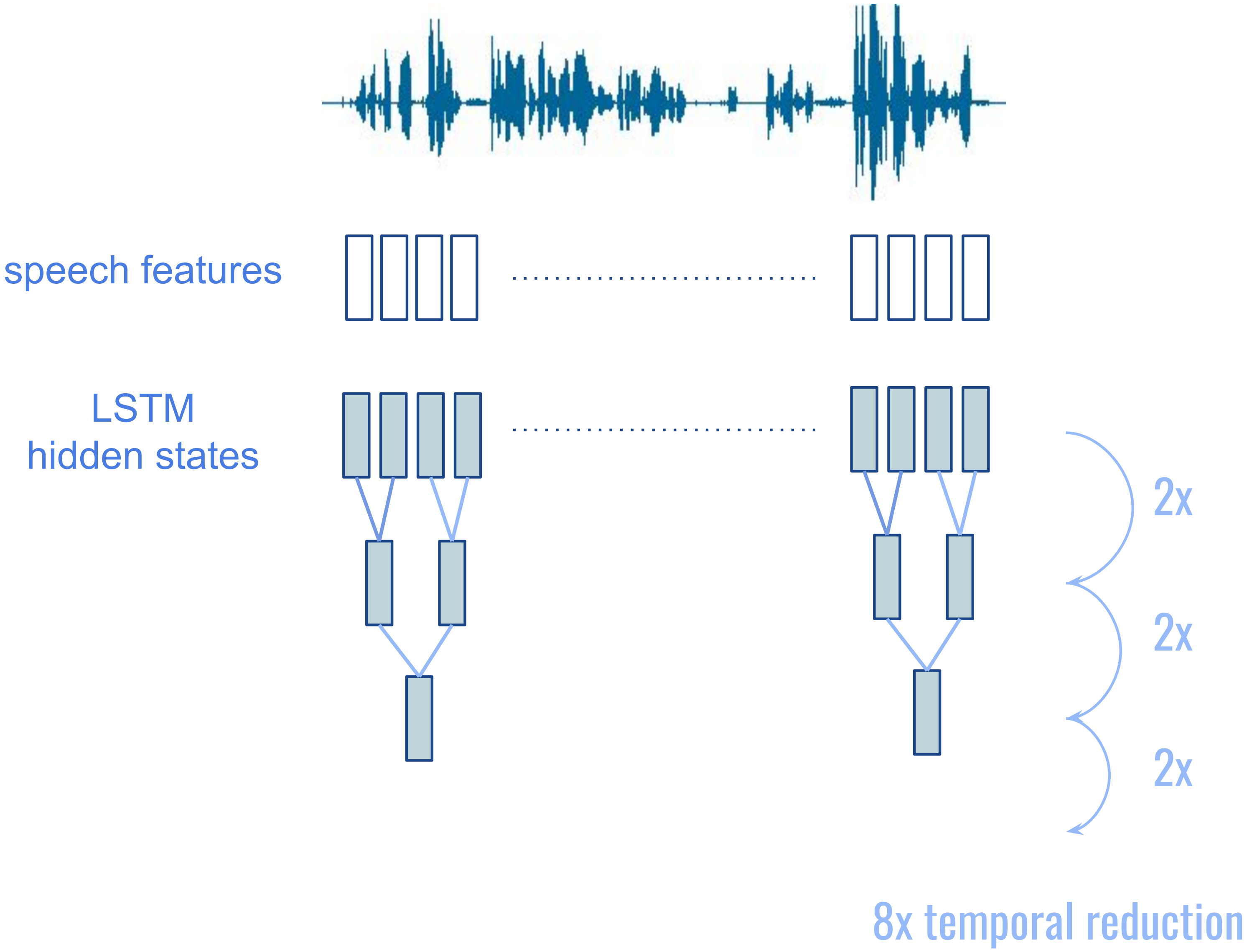
Scale sequence length and feature dimension linearly by a factor corresponding to the convolutional kernel size and stride length



Conv1D, ConvLSTM layers

(Weiss et al. 2017;
Bansal et al. 2018)

Pyramidal Encoder



- Motivation: do not need attention to the granularity of speech features
- Reduce dimensionality *through* encoder

- concatenation
- sum
- skip
- linear projection

Linear projection, ASR:
(Zhang et al. 2017; Sperber et al. 2018)

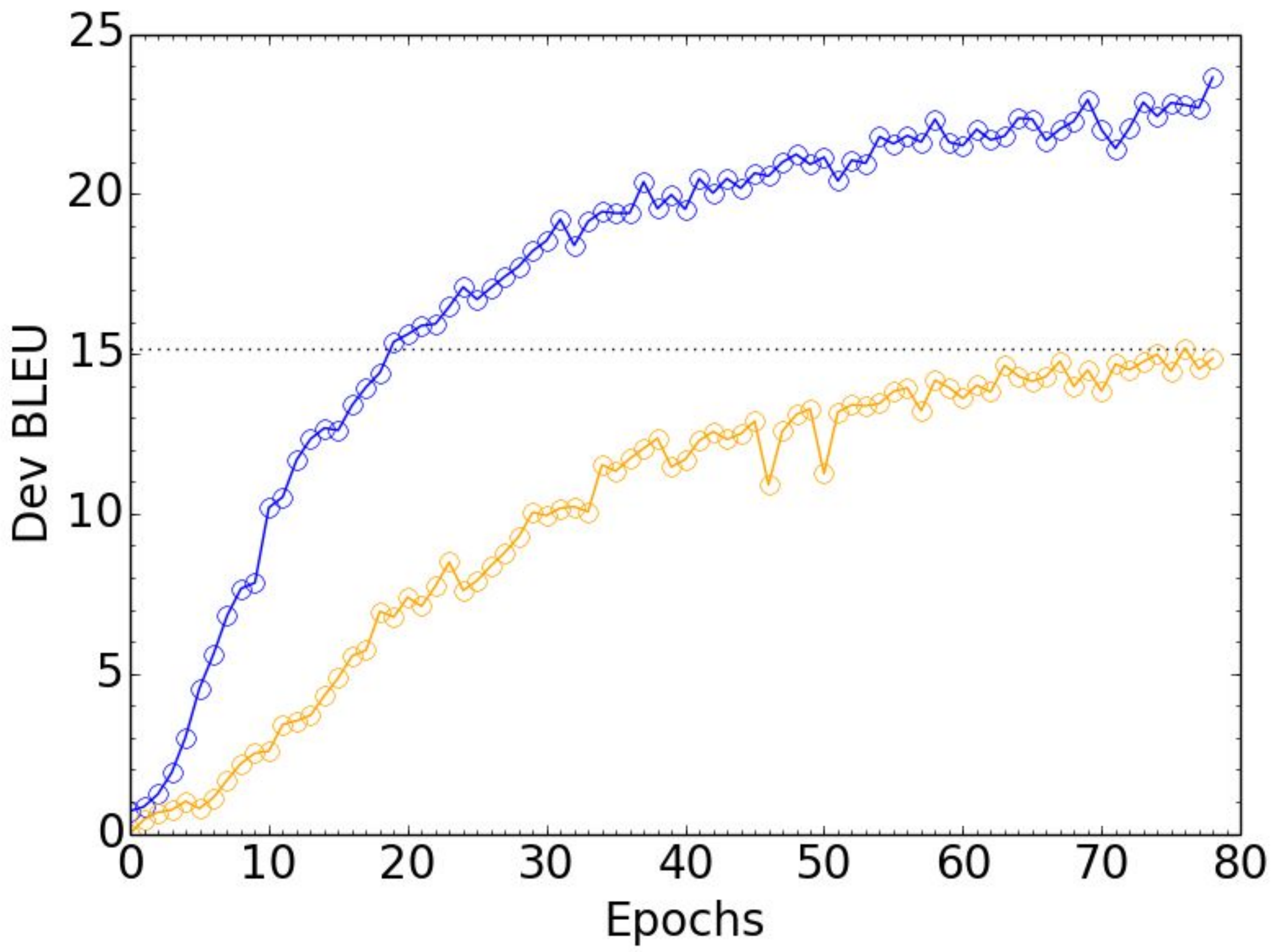
Pyramidal encoder in ST:
(Weiss et al. 2017; Salesky et al. 2019; Sperber et al. 2019; Salesky et al. 2020)

Listen, Attend, and Spell
(Chan et al. 2015)

Dimensionality Reduction Impact

Improved training efficiency!

- Reduces memory footprint
- Faster convergence
- Improved results



(Salesky et al. 2019)

Encoder and Decoder Depth

MT: typically same depth for encoder and decoder

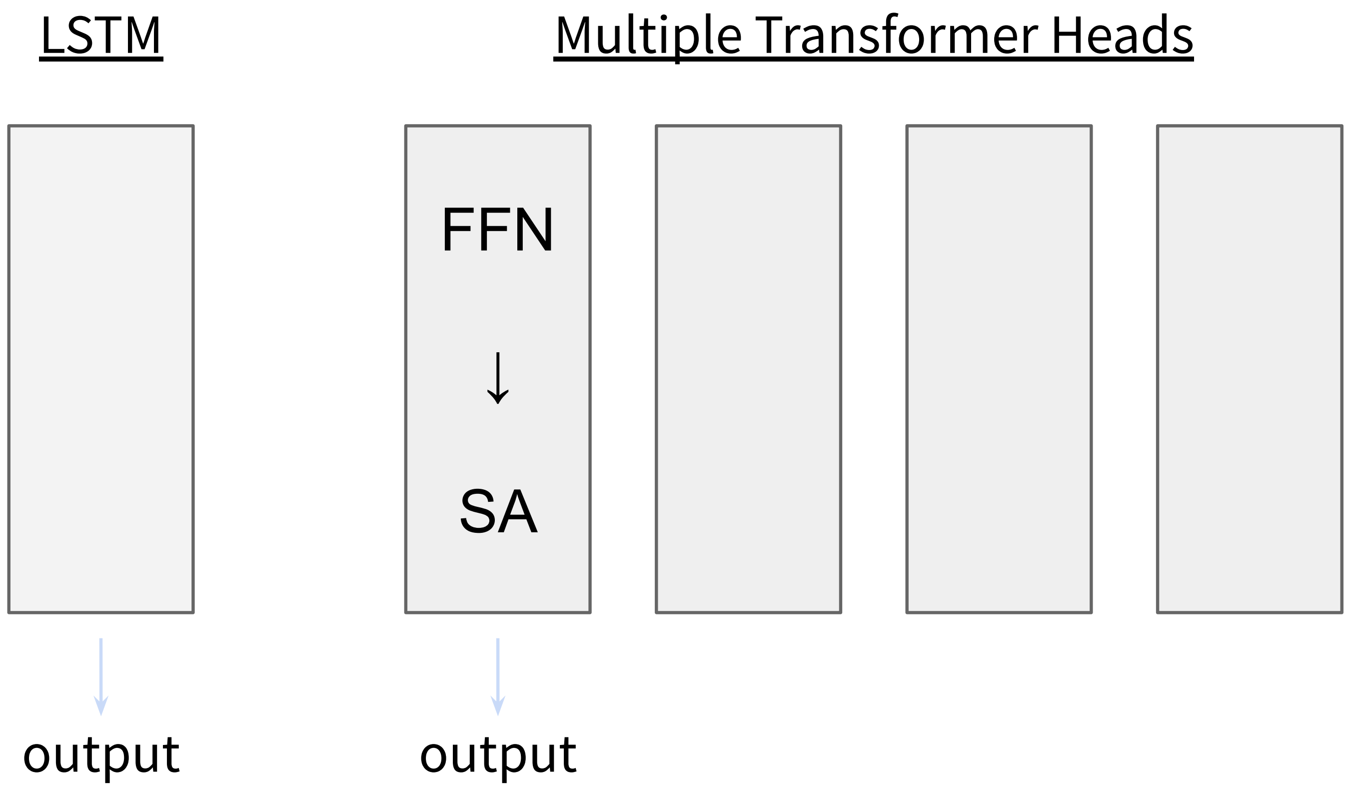
ST: empirically, deeper encoders than decoders perform better!

⟶ *more parameters allocated to learning more complicated associations between inputs*

Models	Test WER
CTC [19]	17.4
CTC/LM + speed perturbation [19]	13.7
12Enc-12Dec (Ours)	14.2
Stc. 12Enc-12Dec (Ours)	12.4
Stc. 24Enc-24Dec (Ours)	11.3
Stc. 36Enc-12Dec (Ours)	**10.6**

(Zhang et al. 2017; Pham et al. 2018)

LSTM → Transformer



Transformer-S

- 2D Convolutions
- Distance penalty for attention
- 2D self-attention

…

Conv-Transformer

(DiGangi et al. 2019; Huang et al. 2020)