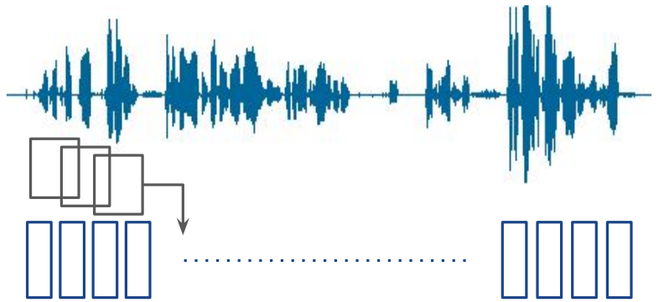
Sec 3.3

Alternate Data Representations

[Recall] Speech vs. Text

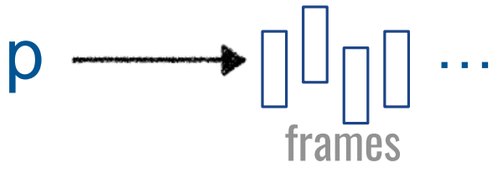


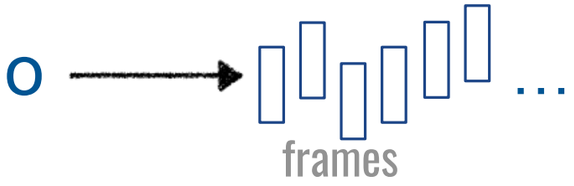
Discretized audio —— speech frames

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

c h a r a c t e r s

SPEECH:





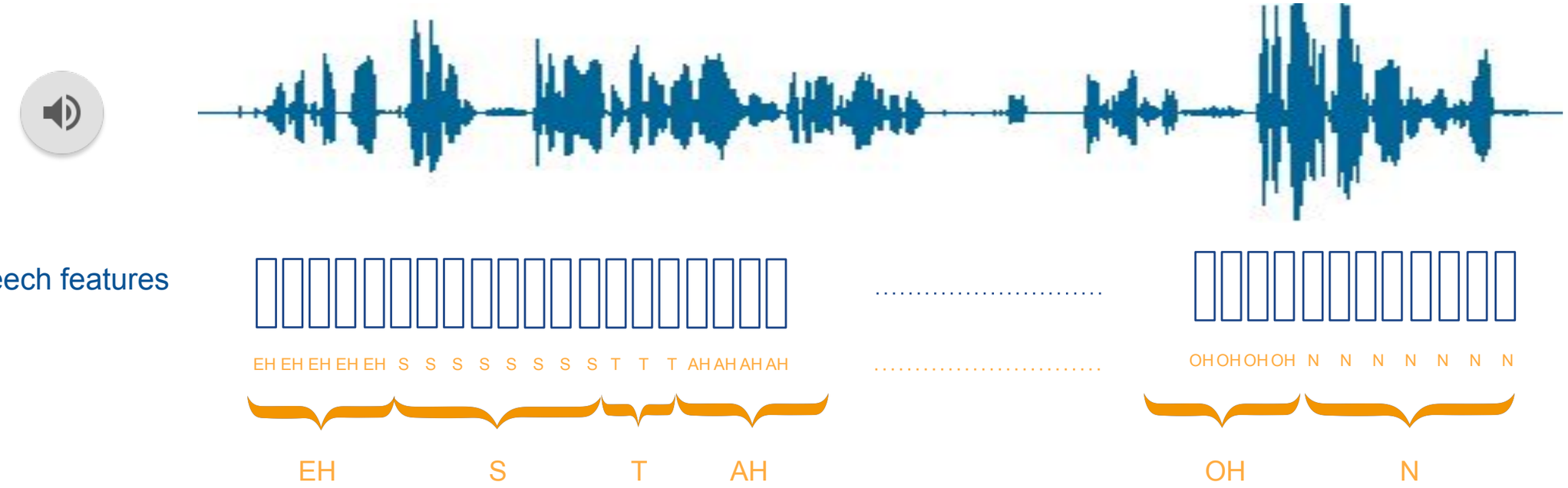
TEXT:

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

Challenges:

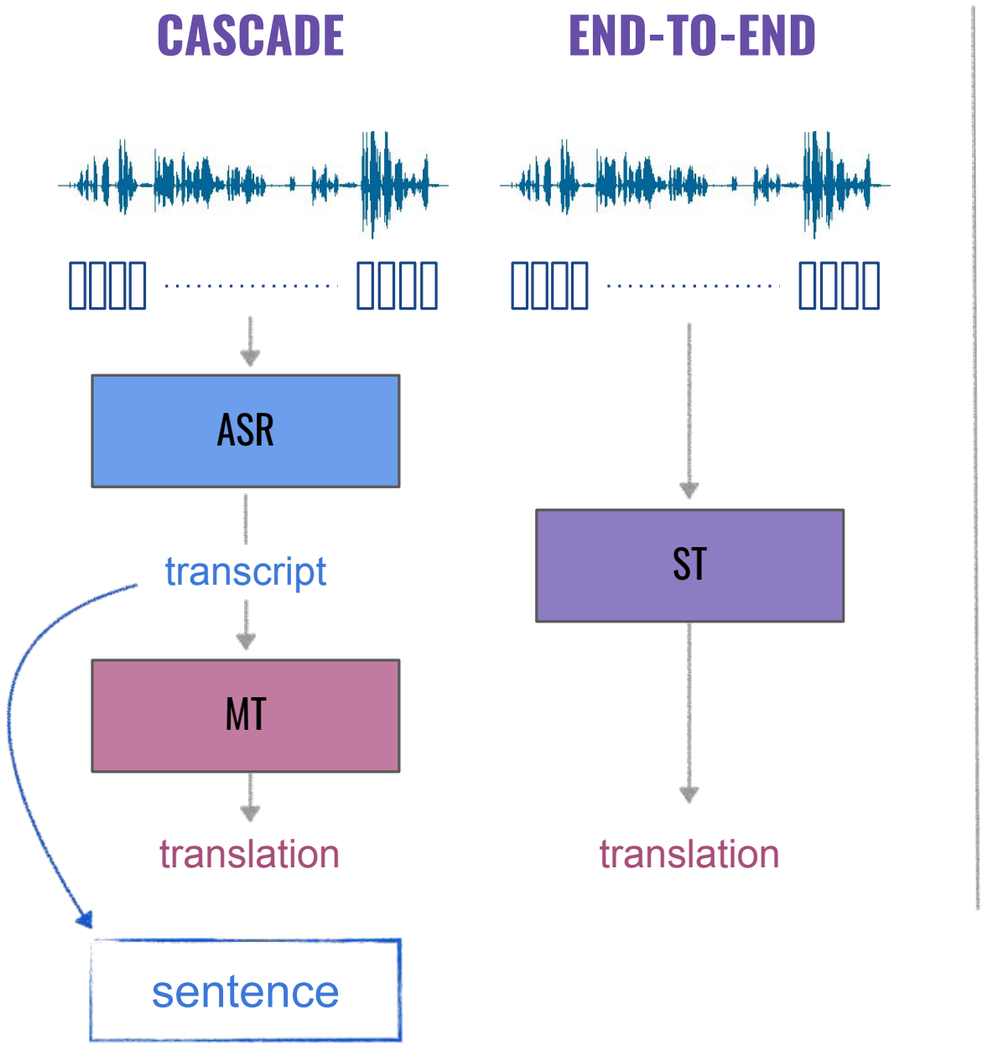
- *Sequence length*
- *Sequence redundancy*
- *Speech feature variation*

A Closer Look

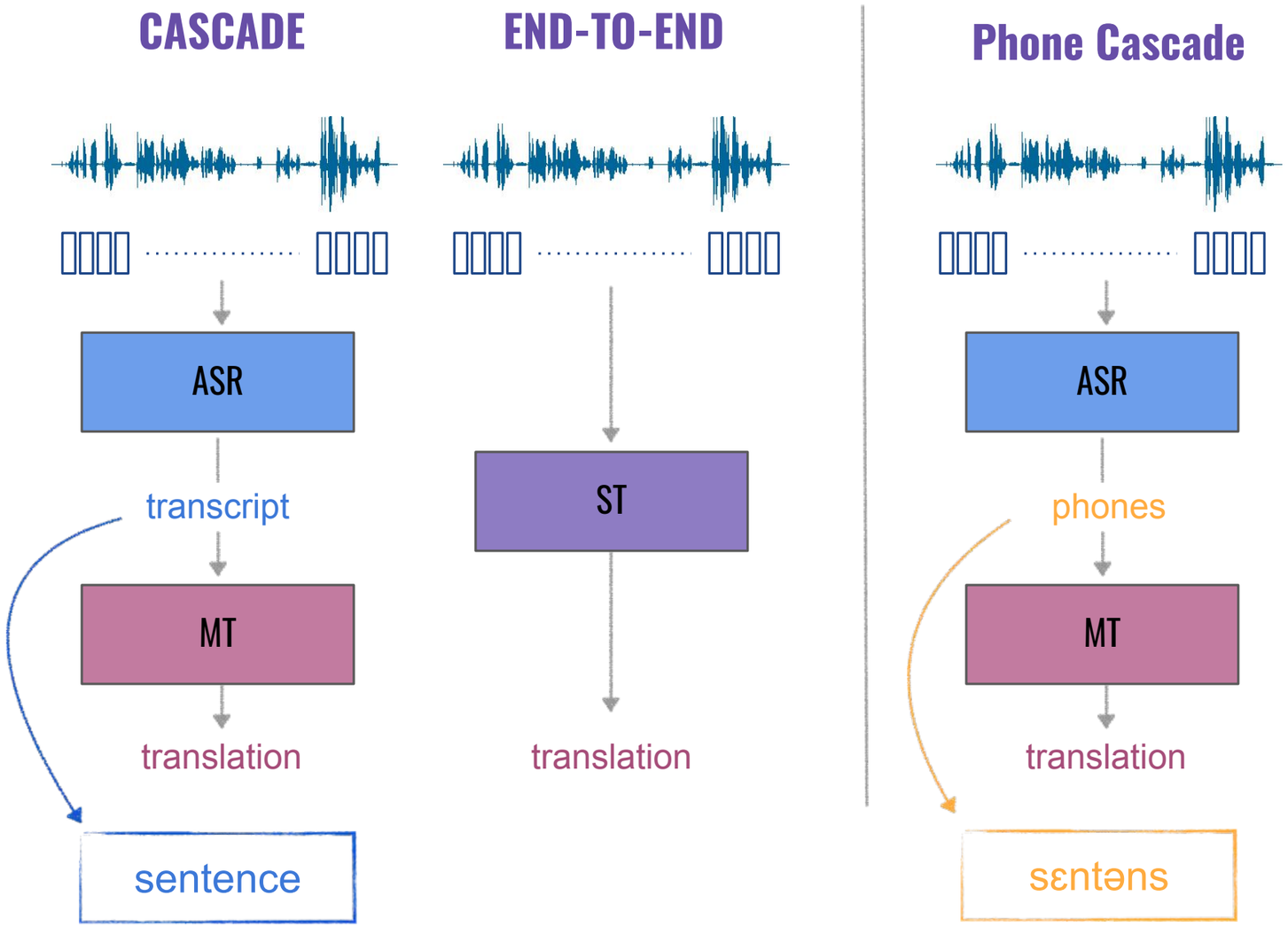


[Esta es una oración]

ST Architectures



ST Architectures



Recall: Redundancy

Translating redundant phone sequences:

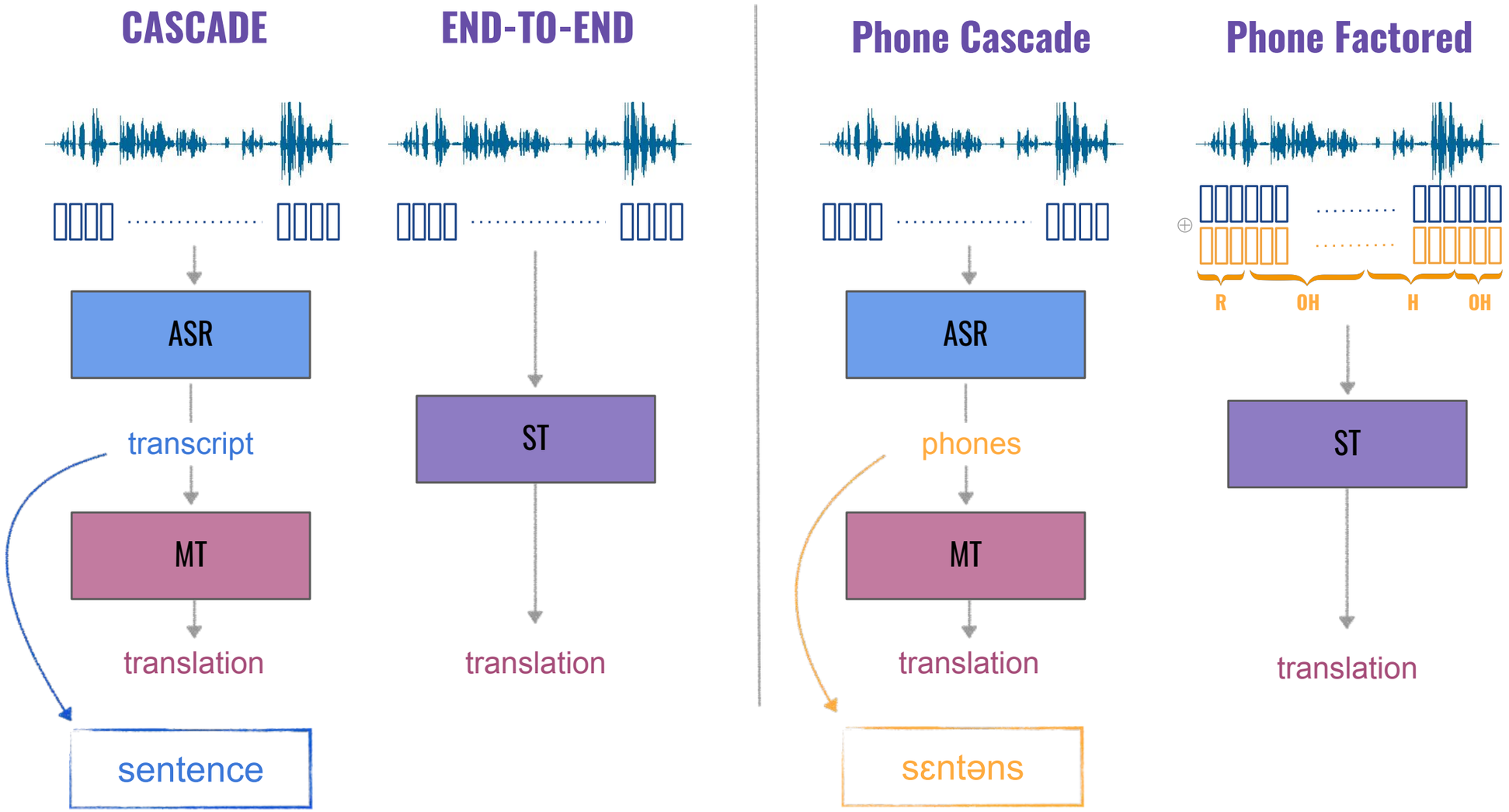
EH EH EH EH EH S S S S S S S T T T AH AH AH AH

performs 13% worse than uniqued:

EH S T AH

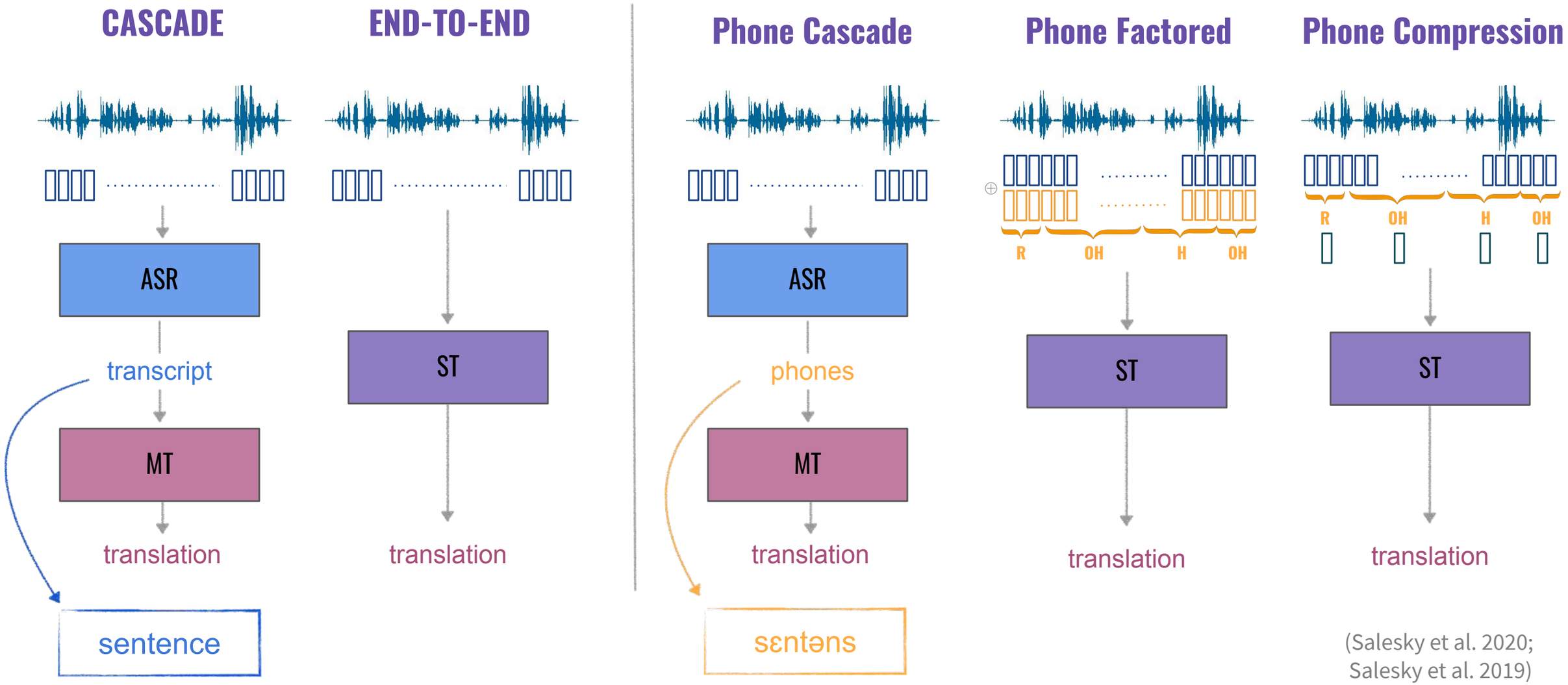
(Salesky et al. 2020)

ST Architectures



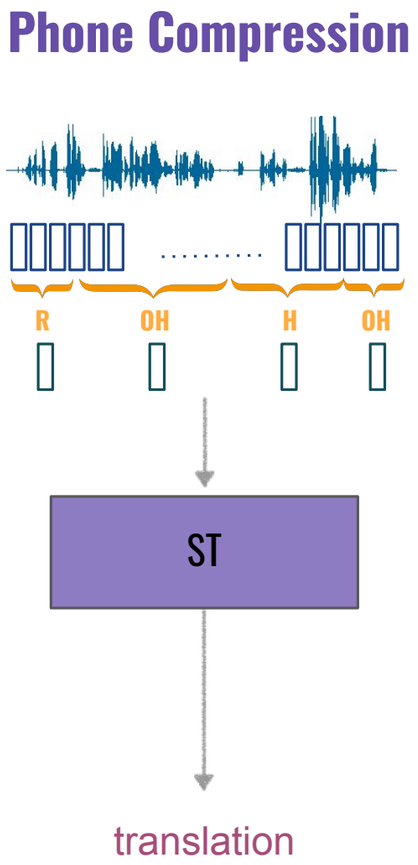
(Salesky et al. 2020)

ST Architectures



(Salesky et al. 2020; Salesky et al. 2019)

Methods



Detecting 'phone' units:

- ASR alignment* (Salesky et al. 2019)
- Adaptive feature selection (AFS)* (Zhang et al. 2020)
- CTC loss applied in encoder (Gaido et al. 2021)

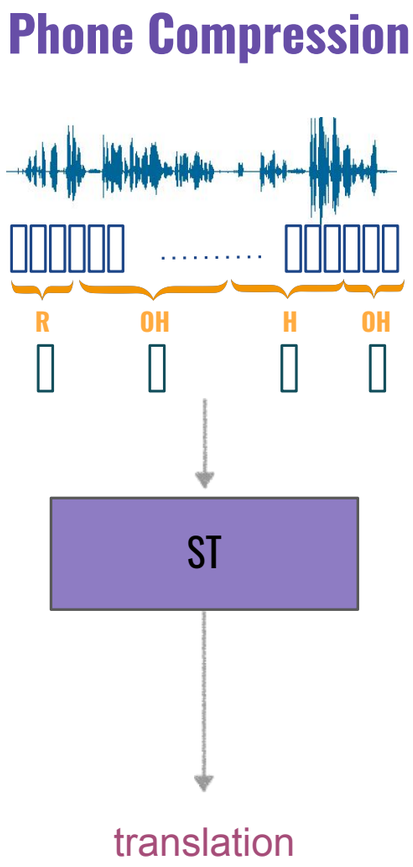
*require an additional model

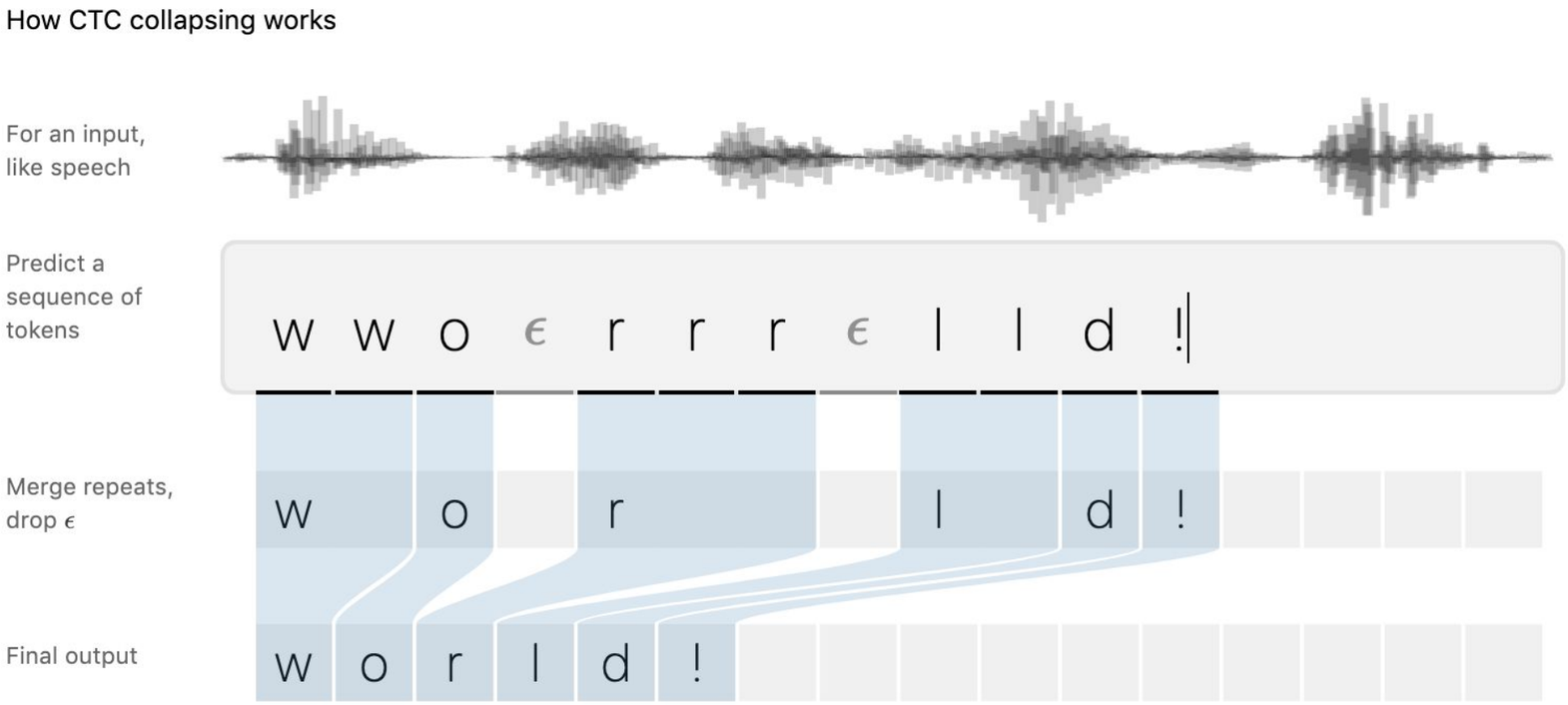
Compression:

- Averaging
- Skip (select key-frame only)
- Softmax
- Weighted projection

(Salesky et al. 2019; Zhang et al. 2020; Gaido et al. 2021)

Methods





(Hannun et al. 2017) — https://distill.pub/2017/ctc

Results

Larger datasets

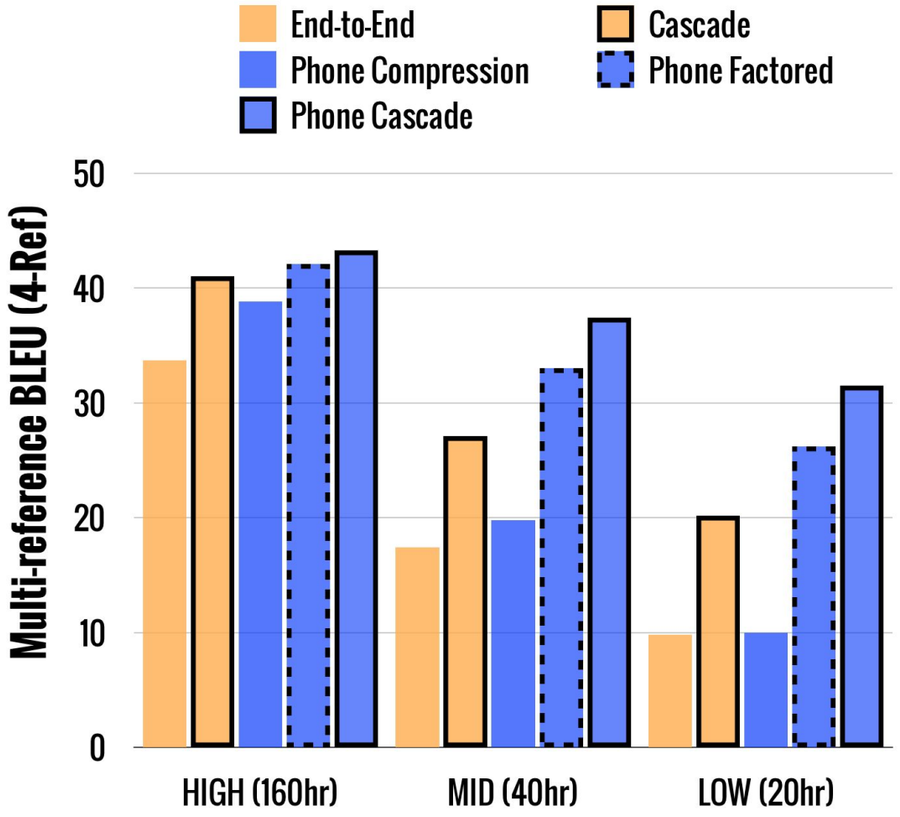
- Librispeech English—French
- MuST-C English—German+
- ~400 hours of speech with translations, transcripts

Performance Improvements

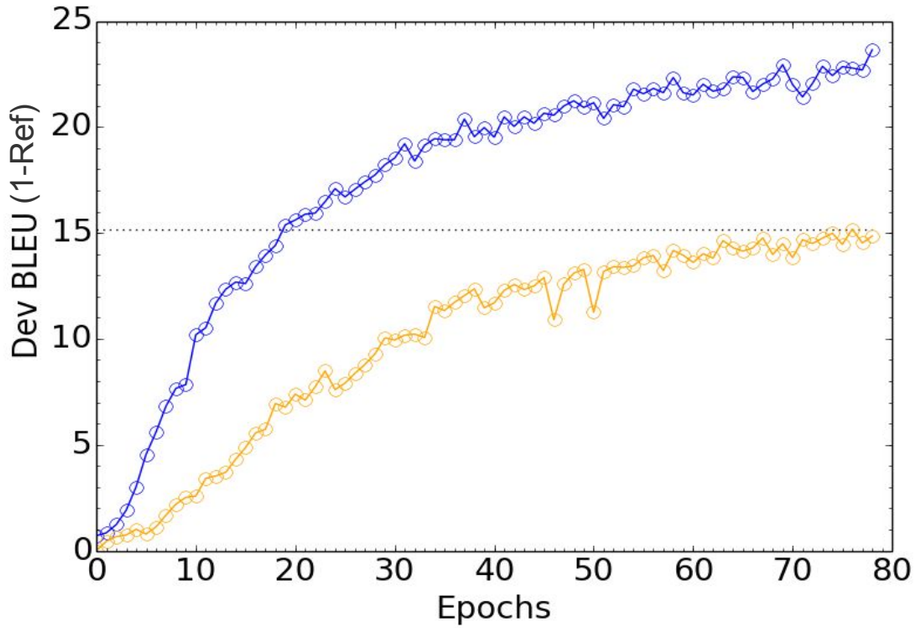
- Improvements of 1-2 BLEU
- Computation reduction:
  - *AFS*: temporal reduction by 80%
  - *CTC*: overall computation reduced by ~10%
- Training and inference time reductions

(Zhang et al. 2020; Gaido et al. 2021)

Results



Fisher Spanish—English
(160 hours)



(Salesky et al. 2019; Salesky et al. 2020)