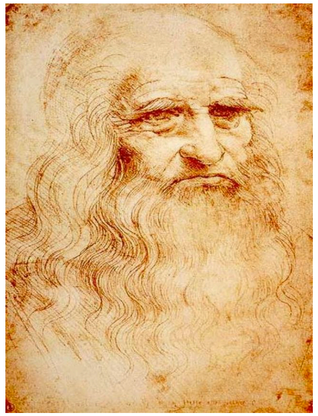
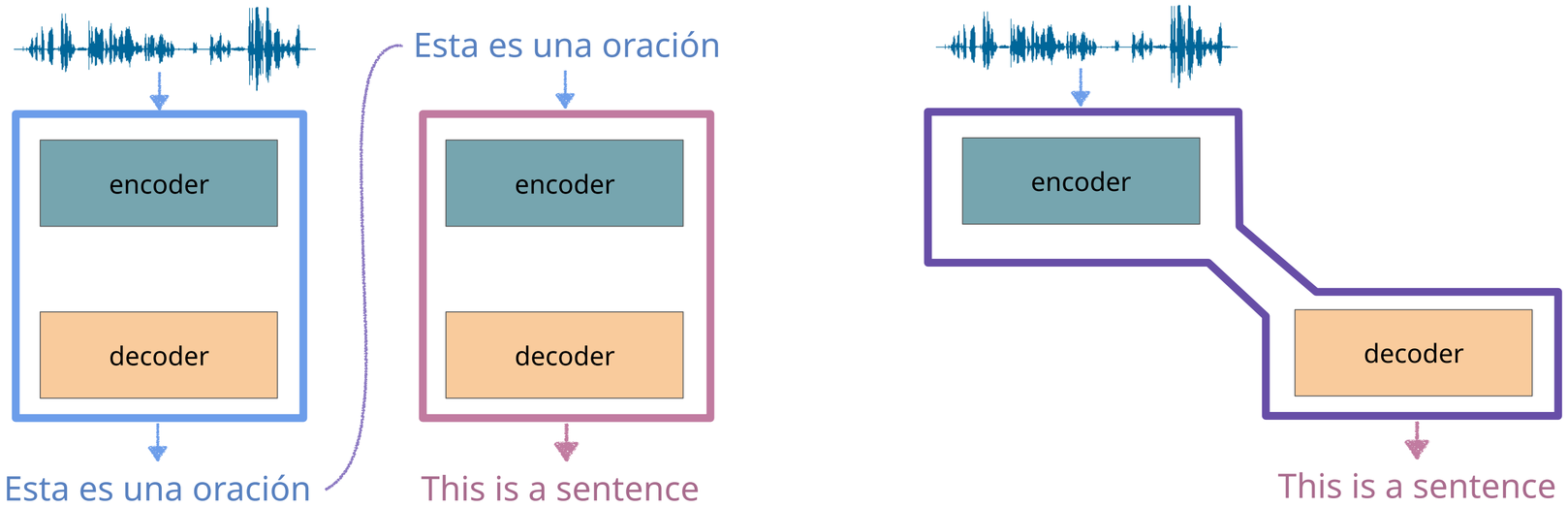
Phone Features Improve Speech Translation

Elizabeth Salesky



Alan W Black

Speech Translation



ASR

MT

Cascade:
2 models

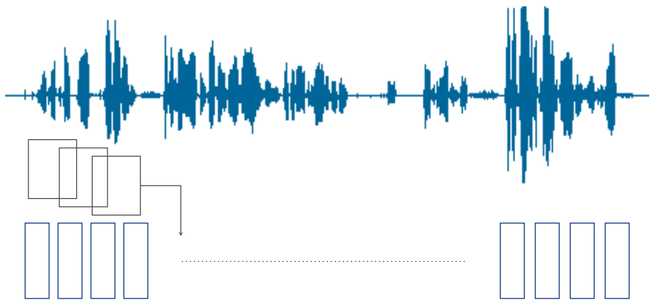
ST

End-to-End:
1 model

Challenges of Speech Input

① Length

c h a r a c t e r s → long sequences



Discretized audio – speech frames

10x longer 😱

Impacts:

- Memory
- Distance between dependencies
- Training efficiency

Performance Impact:

- **End-to-End** model performance varies based on dataset & size, language pair...
- **Cascaded** models perform better in many settings

Challenges of Speech Input

- **Low-Resource Settings**
  - Larger difference in performance between architectures
  - End-to-end models do not see enough data to learn variation

TEXT: p → p

SPEECH: p → … ② Variation in frame values

frames

p → … ③ Variation in number of frames per phone

frames

This Work

① End-to-End vs Cascade comparison

- Single dataset with much previous work:

  Fisher Spanish-English

- Compare multiple resource settings:

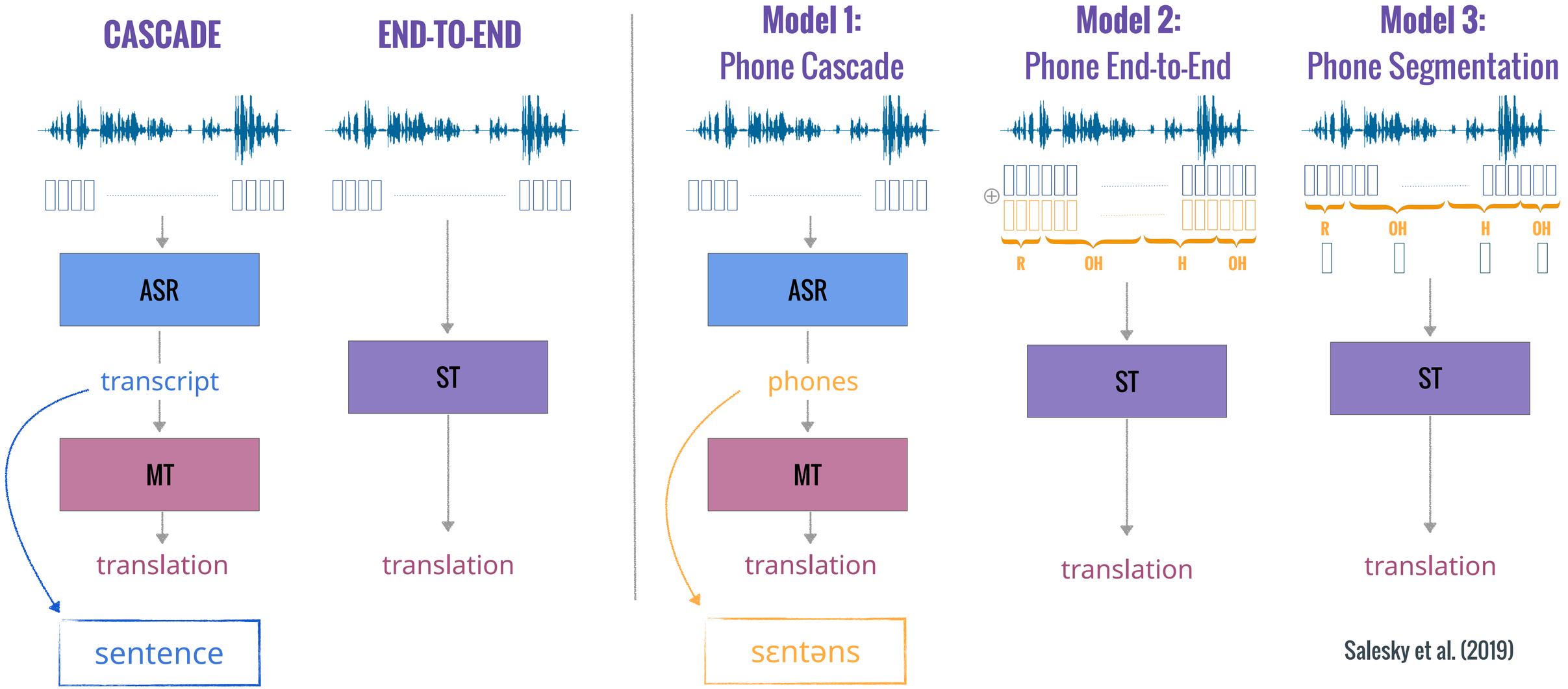
  HIGH (160hr) MED (40hr) LOW (20hr)

② Phone Features to address challenges of speech input

- Compare architectures:

  End-to-End Cascade

Models with Phone Features



Salesky et al. (2019)

Phone Features

See paper or Q&A for more!

Alignment Quality	WER ↓	ASR Supervision
Gold	–	*Gold transcript*
High	23.2	Salesky et al. (2019)
Med	30.4	Seq2Seq ASR
Low	35.5	Kaldi HMM/GMM

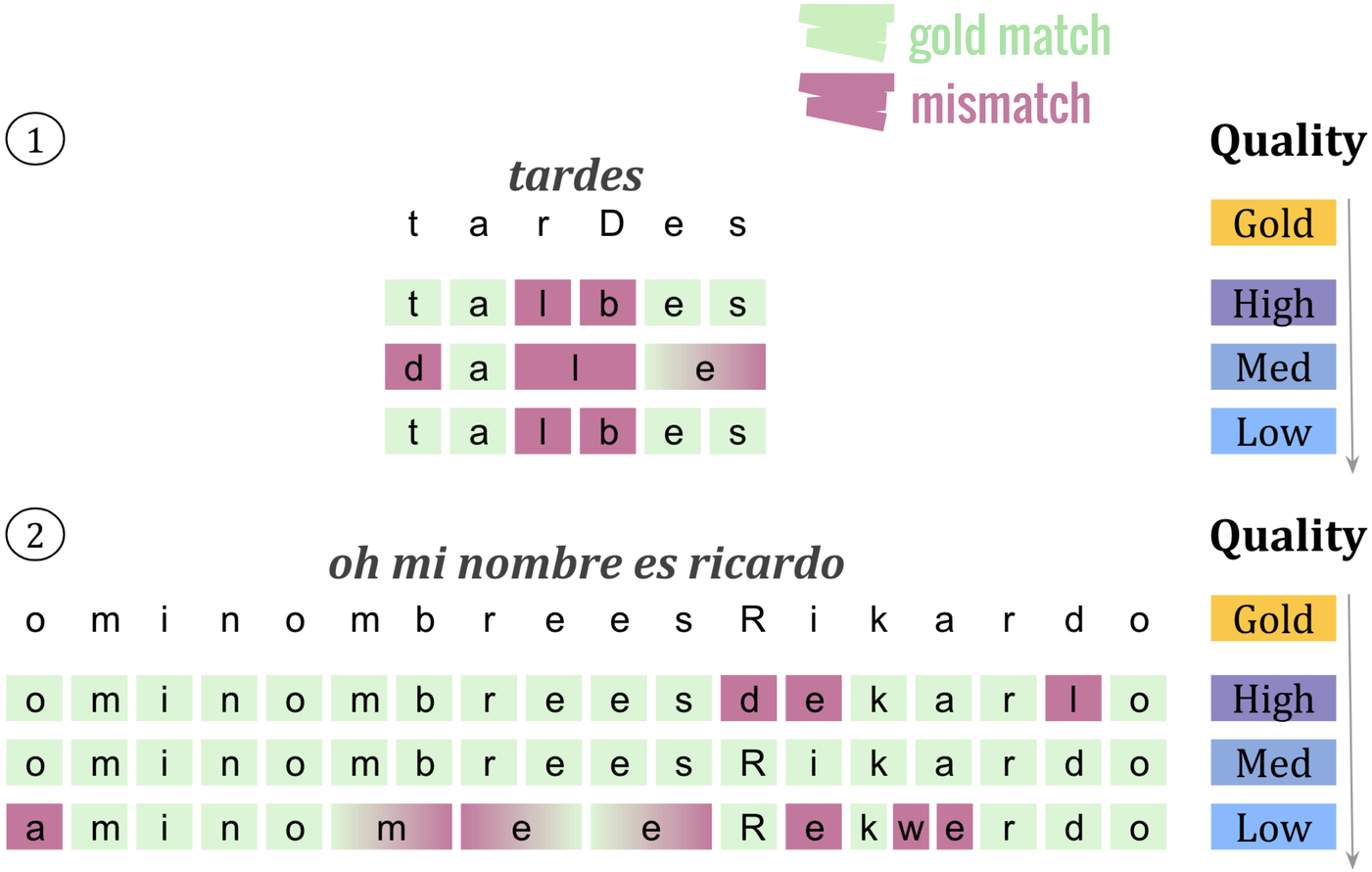
Mapping between phone quality and the ASR models used for alignment generation, with the models' WER on Fisher Spanish test

Phone Features

See paper or Q&A for more!

Alignment Quality	WER ↓	ASR Supervision
Gold	–	*Gold transcript*
High	23.2	Salesky et al. (2019)
Med	30.4	Seq2Seq ASR
Low	35.5	Kaldi HMM/GMM

Mapping between phone quality and the ASR models used for alignment generation, with the models' WER on Fisher Spanish test



Note: uniqued for clarity

① End-to-End vs Cascaded

End-to-End vs Cascaded Models

*Cascade

BPE targets: +2-4↑

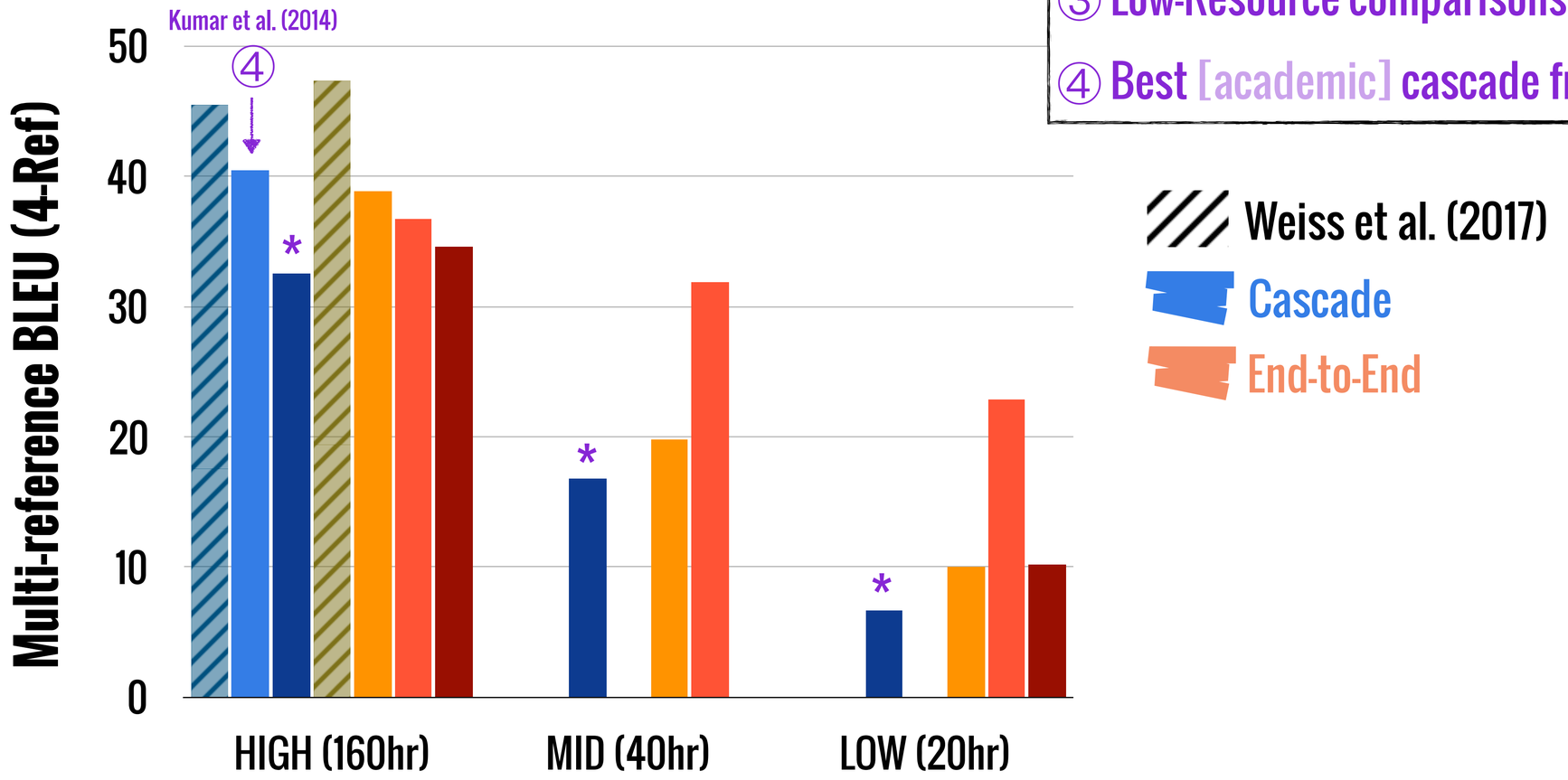
Beam search: +4-8 ↑

① Best end-to-end > Best cascade

② Architecture comparisons lacking

③ Low-Resource comparisons lacking

④ Best [academic] cascade from 2014



End-to-End vs Cascaded Models

Cascade

BPE targets: +2-4↑

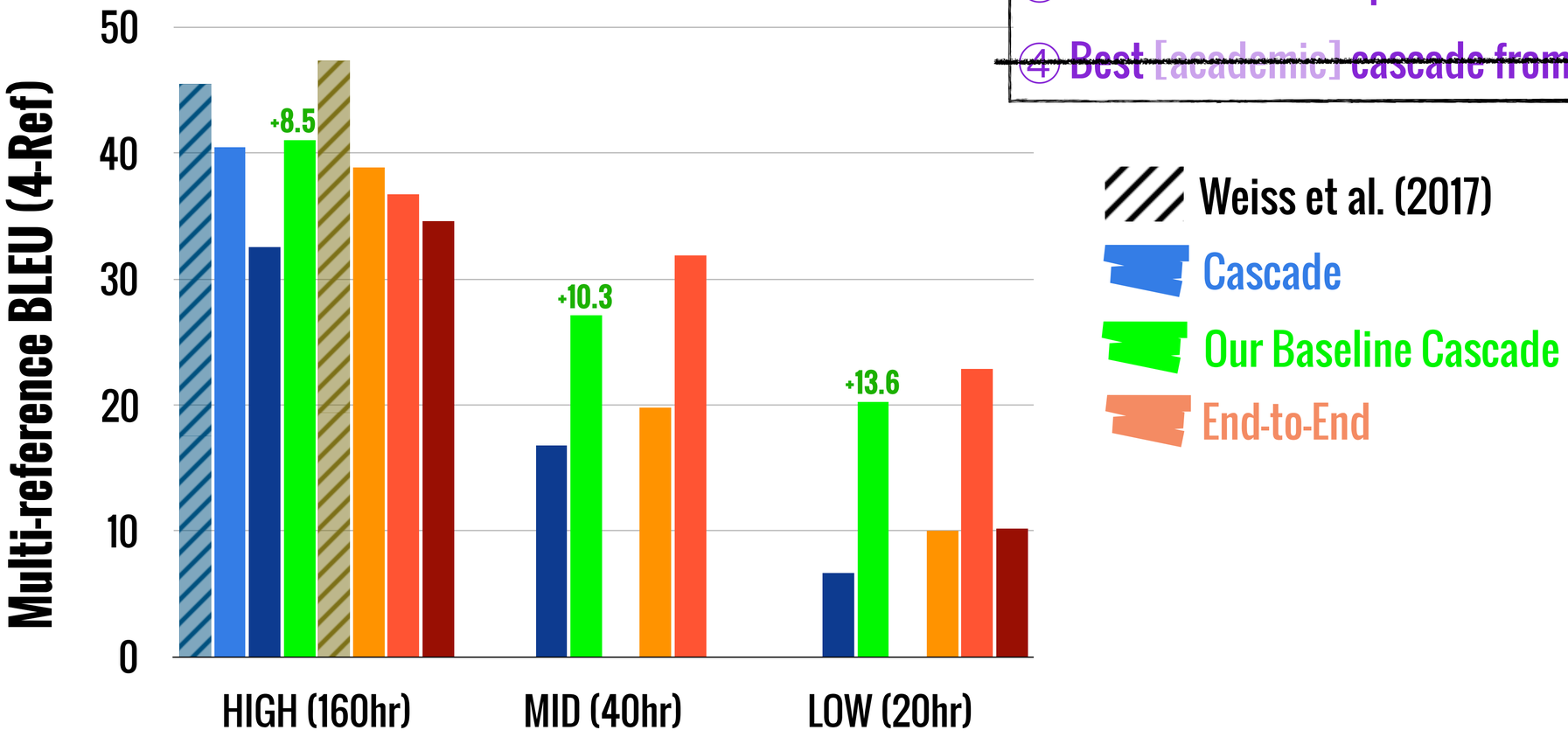
Beam search: +4-8 ↑

① Best end-to-end > Best cascade

② Architecture comparisons lacking

③ Low-Resource comparisons lacking

④ ~~Best [academic] cascade from 2014~~



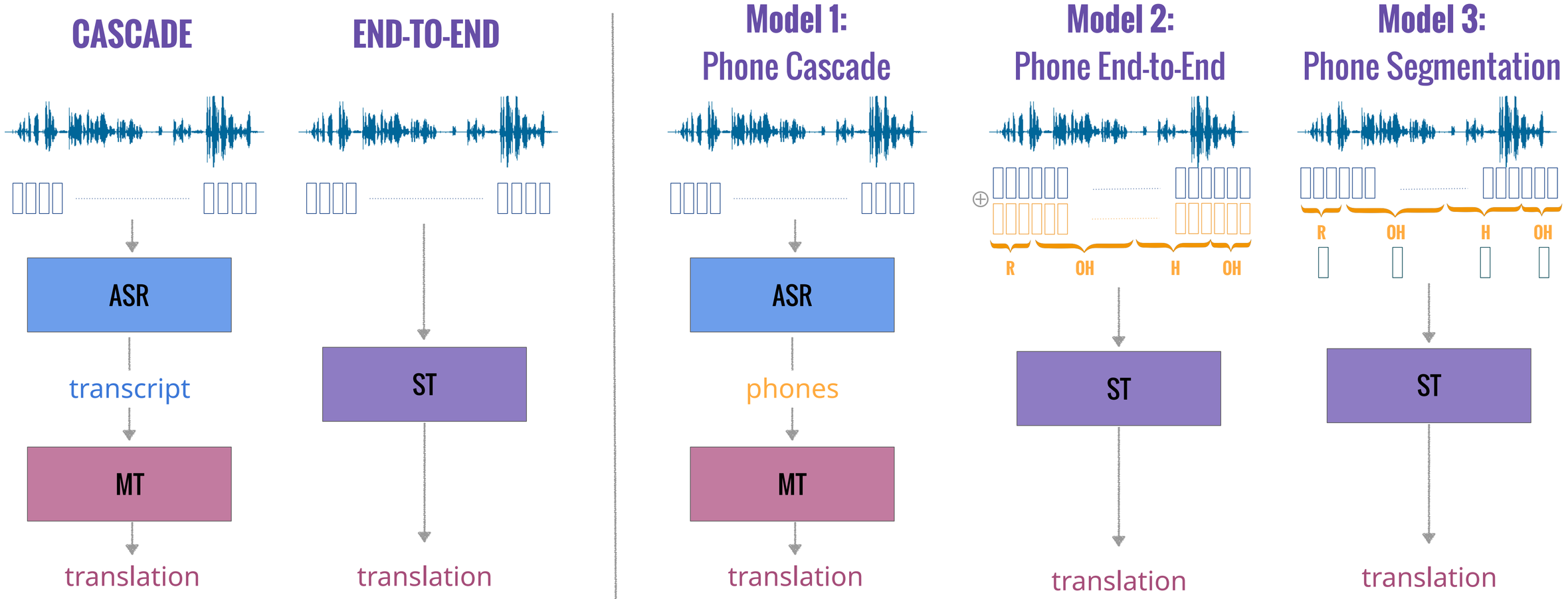
PSA

- Tuning models & parameters matters
  - Can change relative conclusions when making model comparisons

② Phone Features

Recall

Models with Phone Features



Salesky et al. (2019)

Recall

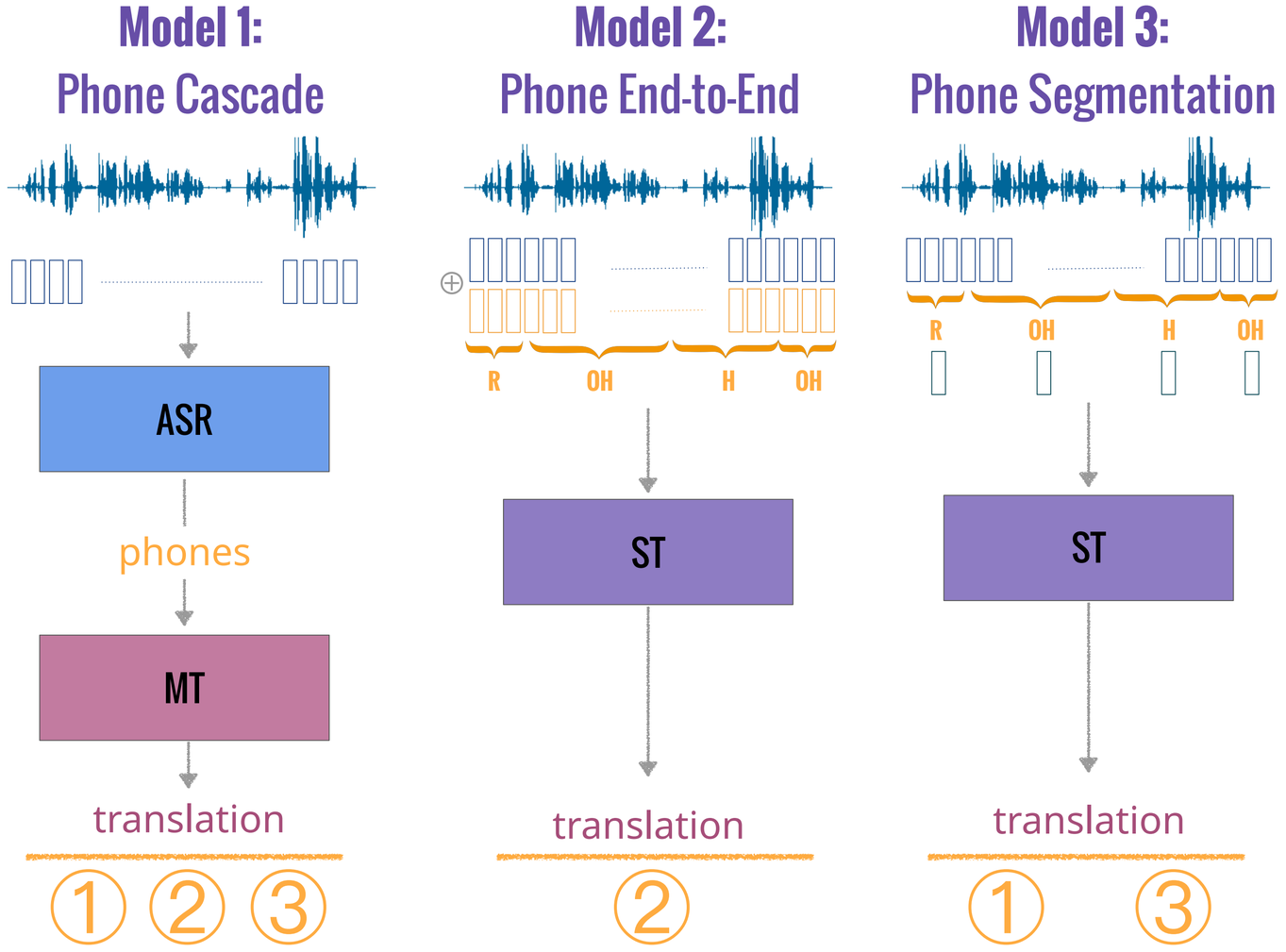
Models with Phone Features

Speech Input Challenges

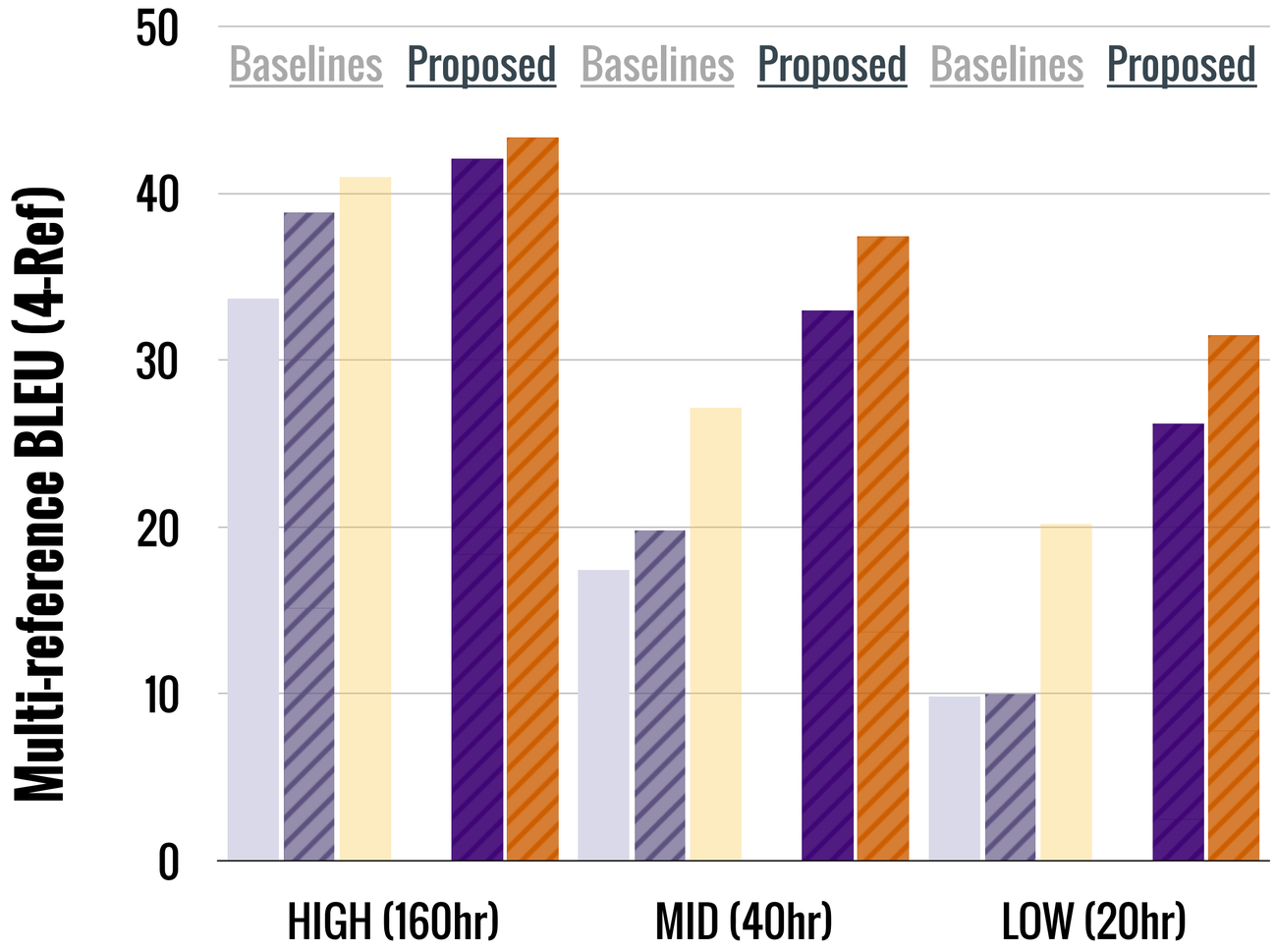
① Length

② Variation in frame values

③ Variation in number of frames per phone



Results with Phone Features



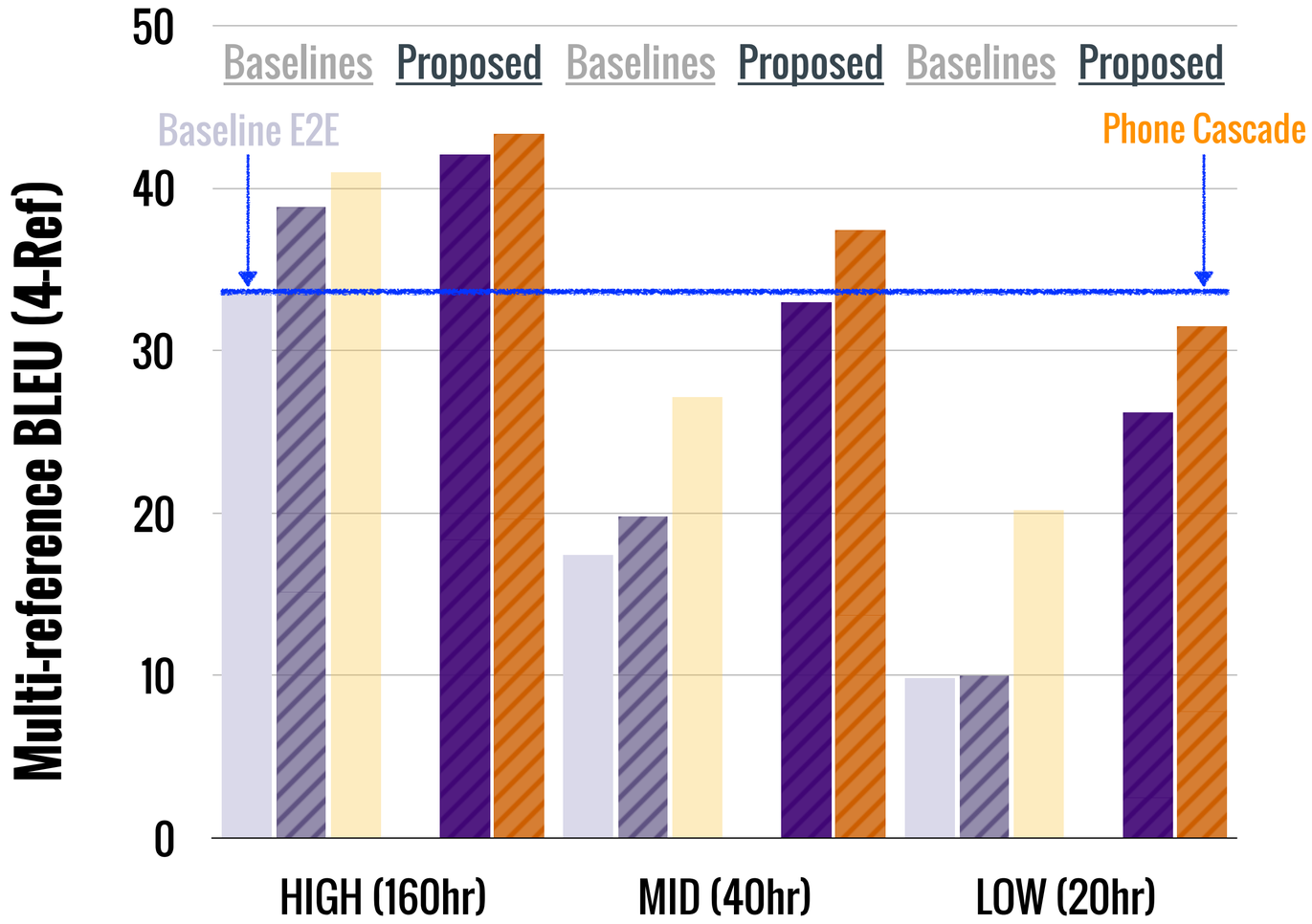
① Significant performance improvements
- Improvements of 10.9-22.1 over baseline E2E

Phone Features

Cascade

End-to-End

Results with Phone Features



① Significant performance improvements

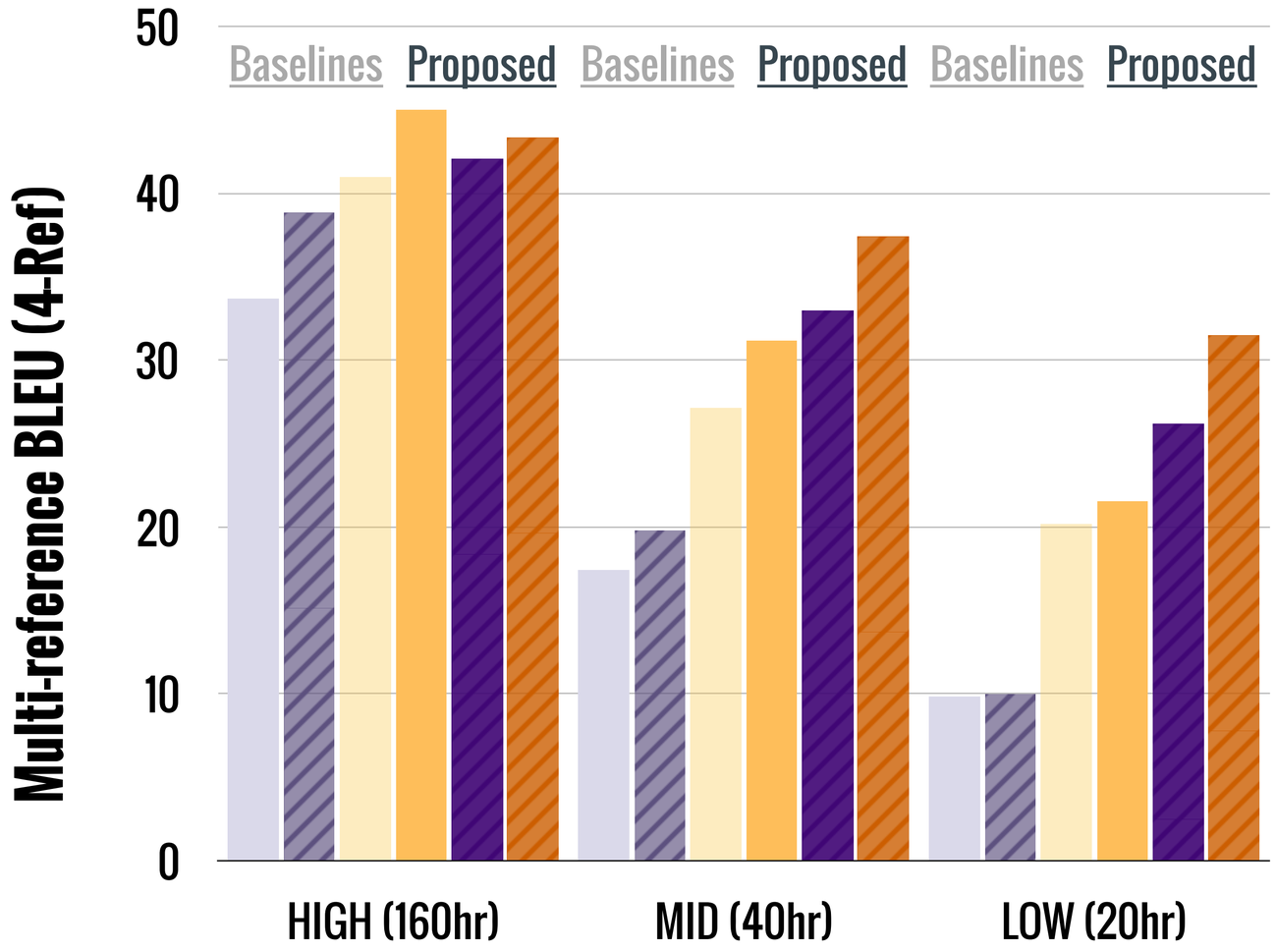
② More data efficient

Phone Features

Cascade

End-to-End

Results with Phone Features



① Significant performance improvements

② More data efficient

③ Benefits of phones remain over SOTA ASR

Phone Features

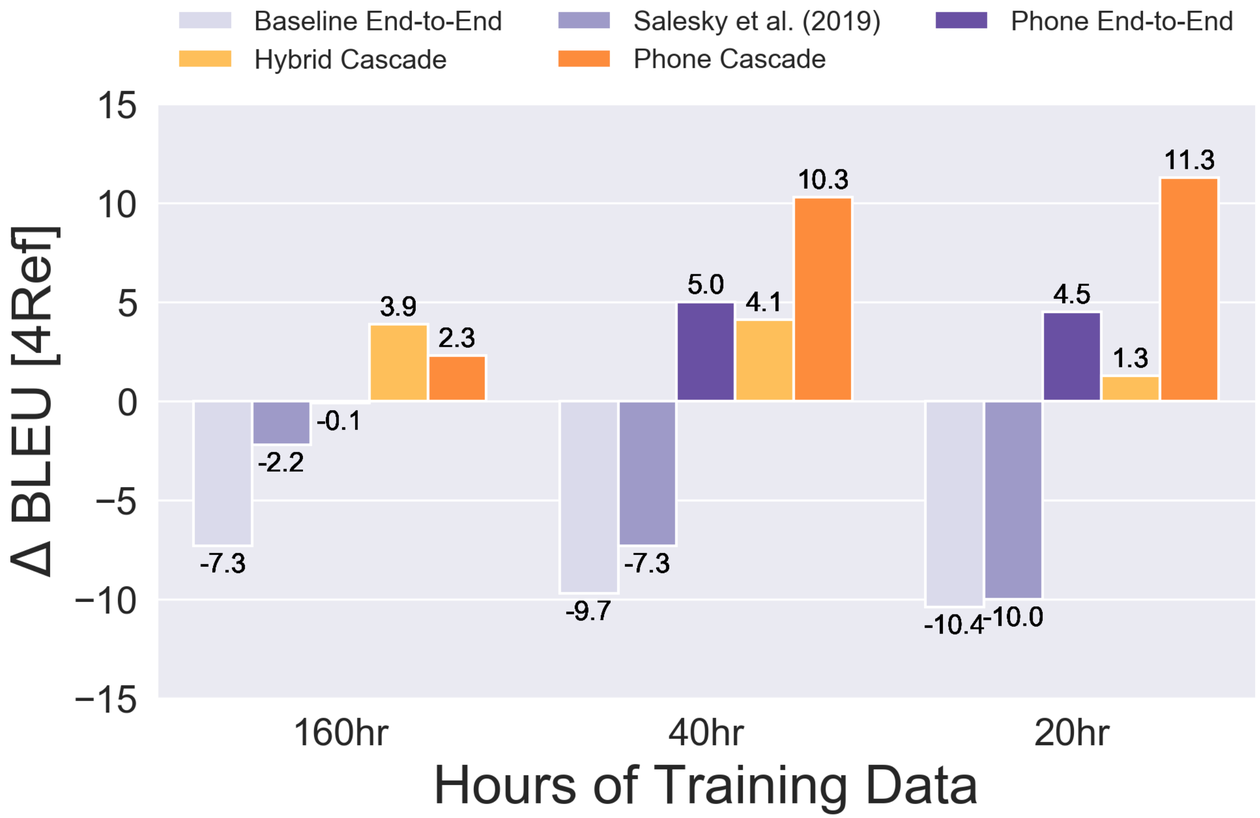
Cascade

End-to-End

Details: Zoom-In

Model	HIGH	MID	LOW
Baseline End-to-End	118hr	40hr	22hr
Salesky et al. (2019)	41hr	13hr	10hr
Baseline Cascade	76hr	19hr	12hr
Phone Cascade	57hr	39hr	27hr
Phone End-to-End	42hr	20hr	13hr
Hybrid Cascade	47hr	34hr	24hr

Training Time

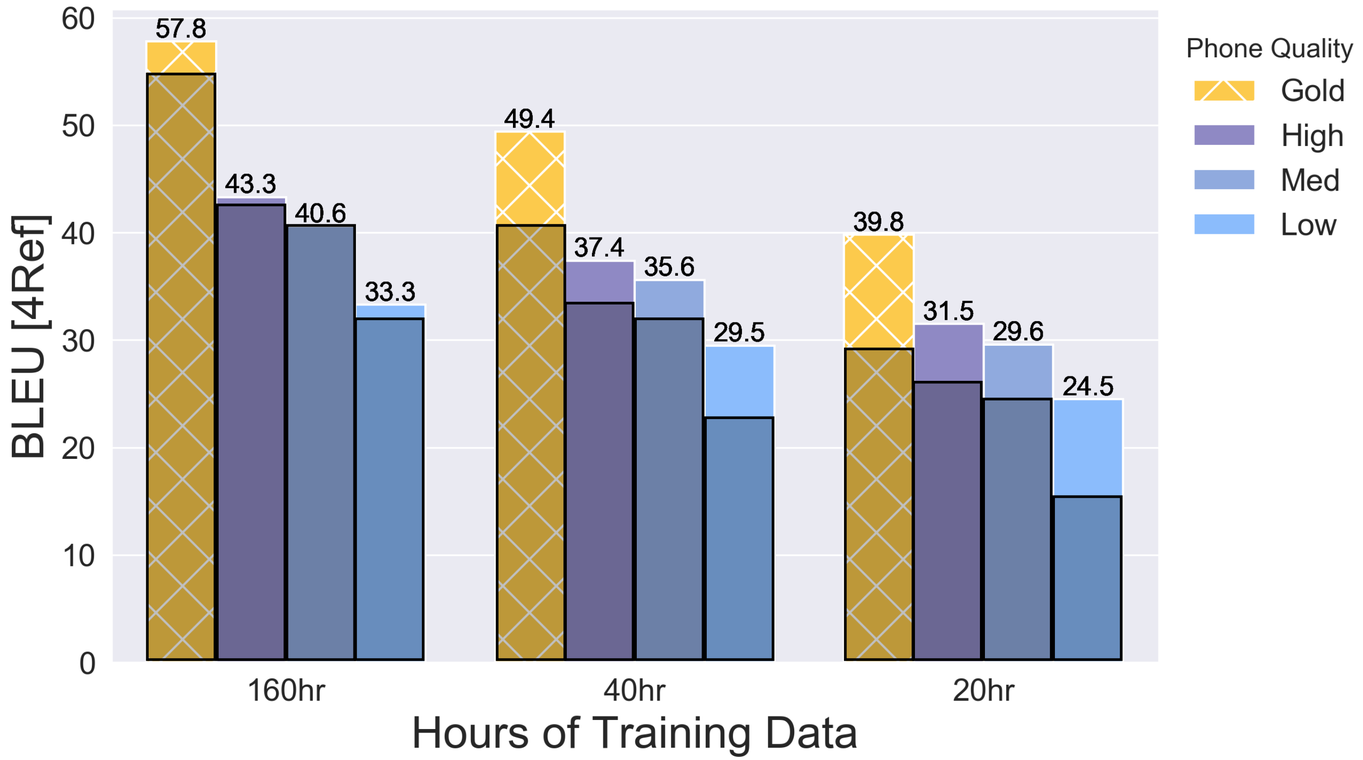


(shown relative to best baseline: Baseline Cascade)

Feature Quality

Phone Cascade

Phone End-to-End



Conclusion

- Phone features are very effective, at multiple resource settings!
  - Build models with intuitions from phone features
- Performance on high-resource settings ⇏ performance on low-resource
  - Test models on multiple resource settings
- Cascades are competitive and often better than current E2E models
  - Compare against strong cascaded baselines