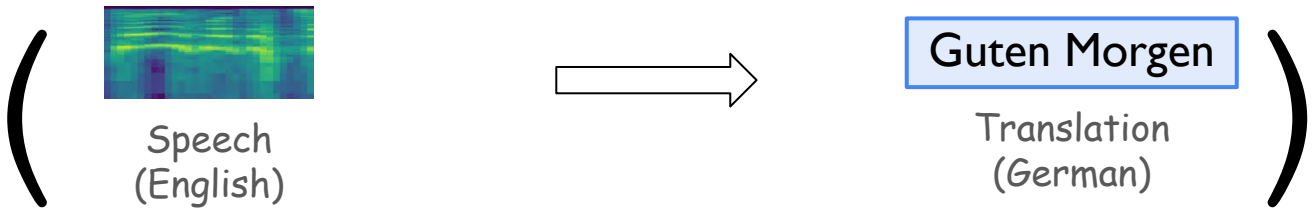
Revisiting End-to-End Speech-to-Text Translation From Scratch

Biao Zhang[1], Barry Haddow[1], Rico Sennrich[2,1]

[1]University of Edinburgh [2]University of Zurich

The Thirty-ninth International Conference on Machine Learning (ICML 2022)

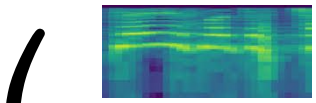
End-to-End Speech-to-Text Translation Is Challenging



E2E ST aims at translating speech directly to a foreign text **without any intermediate outputs**, e.g., transcript

- ✘ Implicitly modeling ASR and MT via a single model is difficult
- ✘ Performance of the direct ST baseline lags far behind the cascade

Rescue: ASR/MT Pretraining and ST Finetuning



Speech (English)

Good Morning

Transcript (English)

Guten Morgen

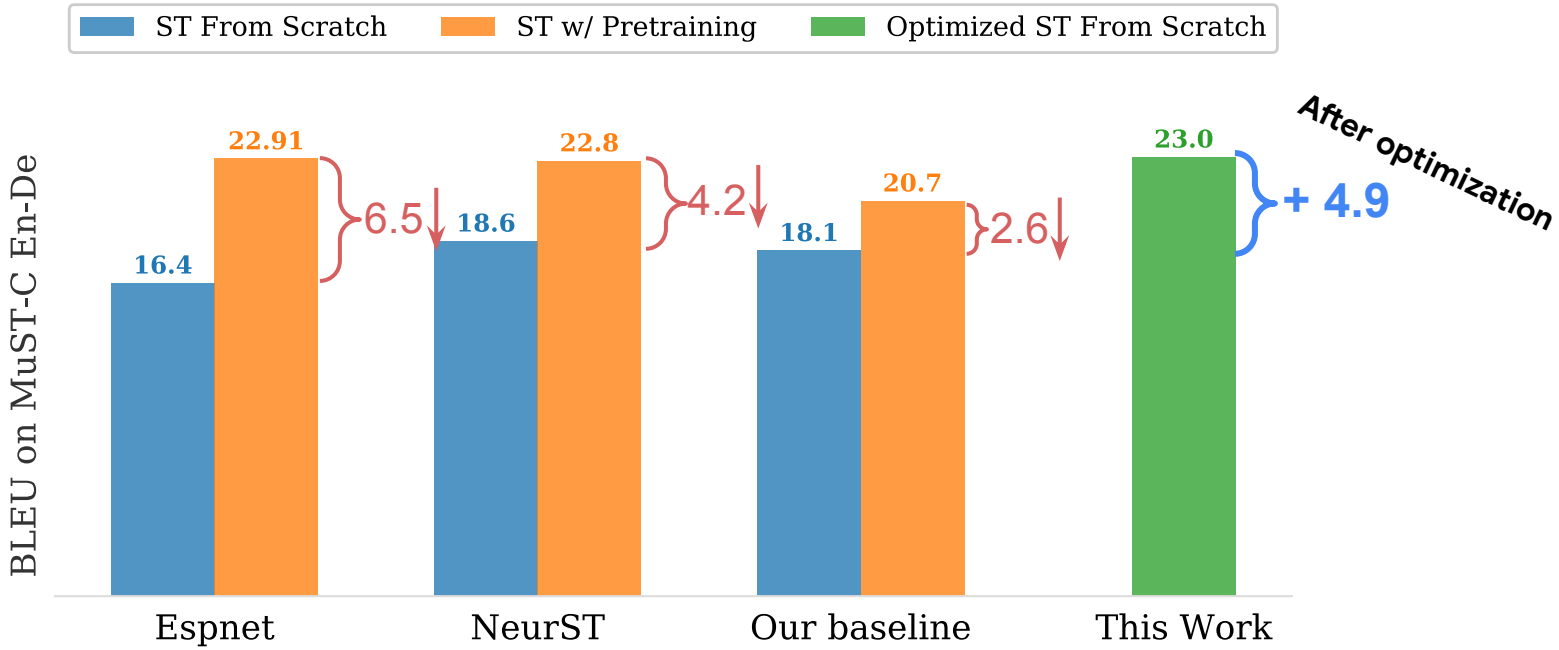
Translation (German)

Step 1: pretrain ST encoder/decoder with ASR/MT using **transcripts**

Step 2: finetune the model on direct speech-translation pairs

Inaguma et al., 2020; Wang et al., 2020; Zhang et al., 2020; Zhao et al., 2021

E2E ST Without Transcript Performs Poorly, Really?



ST from Scratch: train ST on **speech-translation pairs alone** and from scratch

Why Revisit E2E ST ***From Scratch***?

- ✔ Improve our understanding of pretraining in E2E ST
  - when and where ASR/MT pretraining really matters
- ✔ Transcript is not always available
  - > 3000 languages in our world have no written form
- ✔ Simplify the training pipeline, and develop useful inductive biases
  - using two-stage training complicates the modeling

Improving E2E ST Towards Training From Scratch

CTC Regularization

- Use translation as CTC labels
- No transcripts are used

Parameterized Distance Penalty (PDP)

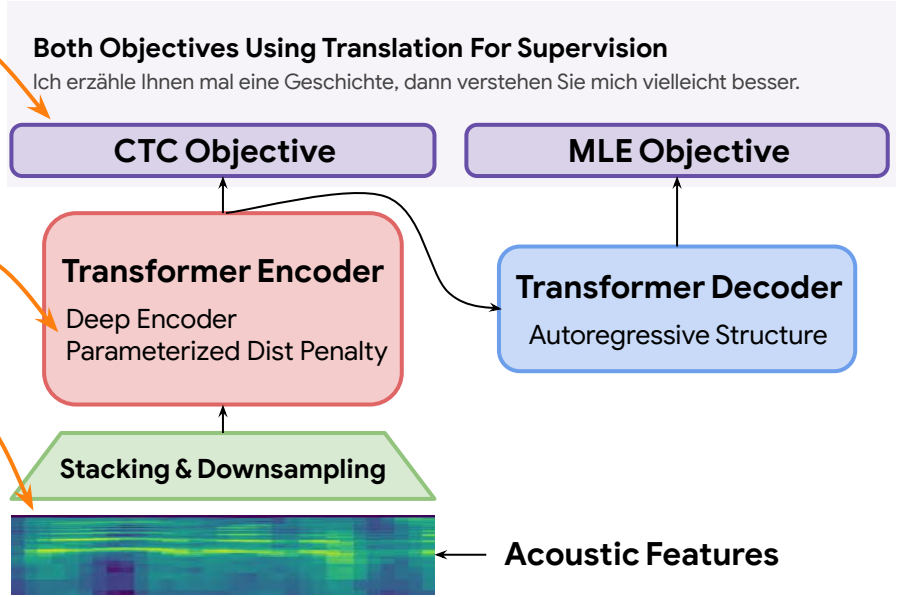
- Add freedom in local attention modeling

Neural Acoustic Feature Modeling

- Use raw waveform to retain local details

Hyperparameter Tuning

- Beam search; Model depth/width



Using speech-translation pairs alone **with no transcripts**

Improved Results: Different Techs Are Complementary

System	BLEU		Avg
NeurST (pretrain-finetune)	22.8		24.9
Baseline	18.1	+3.0	-
+ hyperparameter tuning	21.1	+0.7	-
+ PDP (R=512)	21.8	+0.9	-
+ CTC regularization	22.7	+0.3	-
+ neural acoustic model	**23.0**		**25.2**

Test performance on MuST-C En-De and average results on the other language pairs

Note all our models are trained with speech-translation pairs alone

To Summarize

- The quality gap between ST with and without transcripts is overestimated
- We figure out a set of practices for improving ST from scratch
  - deep post-LN encoder, wider feed-forward layer, ST-based CTC regularization and parameterized distance penalty, neural acoustic feature modeling
- Pretraining still matters
  - low-resource regime and large-scale external data available

Paper: https://arxiv.org/abs/2206.04571

Code: https://github.com/bzhangGo/st_from_scratch