ESPnet-ST: All-in-One Speech Translation Toolkit

Hirofumi Inaguma[1] Shun Kiyono[2] Kevin Duh[3]
Shigeki Karita[4] Nelson Yalta[5]
Tomoki Hayashi[6,7] Shinji Watanabe[3]

[1]Kyoto university [2]RIKEN AIP [3]Johns Hopkins University
[4]NTT Communication Science Laboratories
[5]Waseda University [6]Nagoya University
[7]Human Dataware Lab. Co., Ltd.

https://github.com/espnet/espnet

Speech translation: Cascade vs. End-to-end

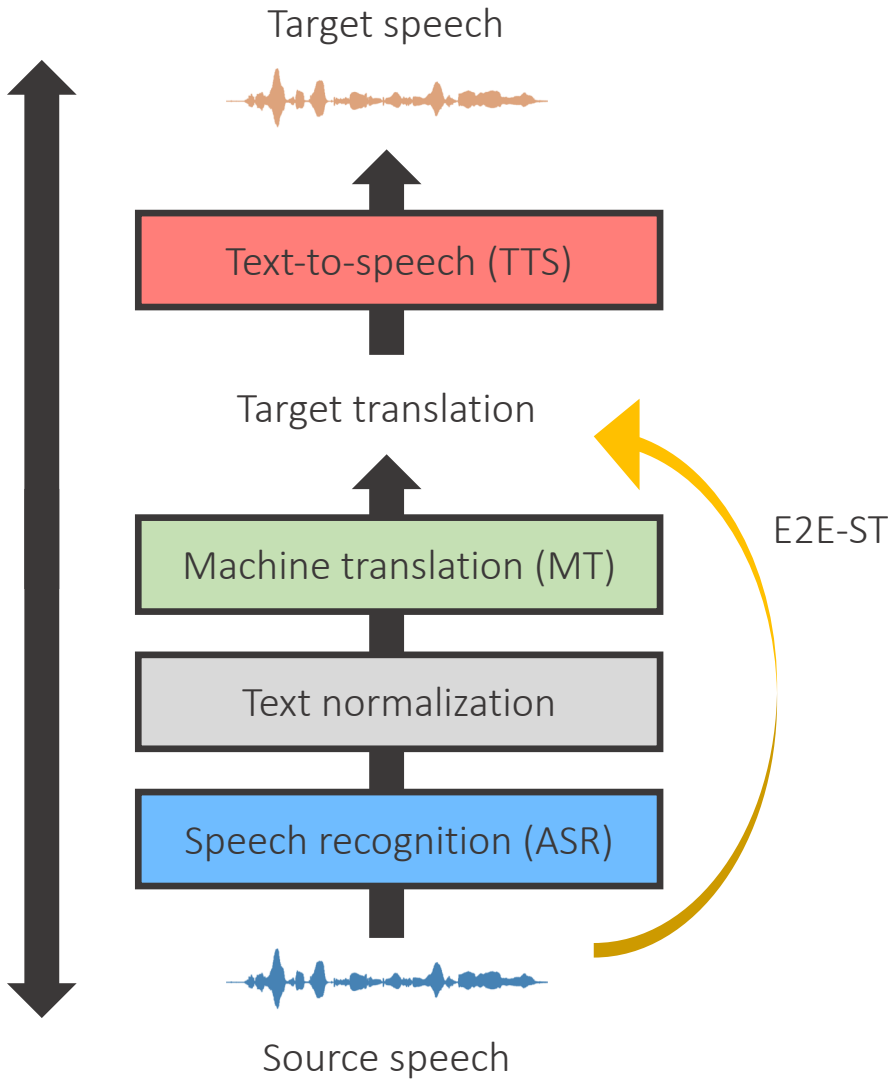
Cascaded speech translation system

- Modularized training (ASR->MT)
- Still better performance than E2E

End-to-end speech translation (E2E-ST)

- Low-latency during inference
- Mitigate error propagation from ASR
- Easy implementation
- Endangered language documentation

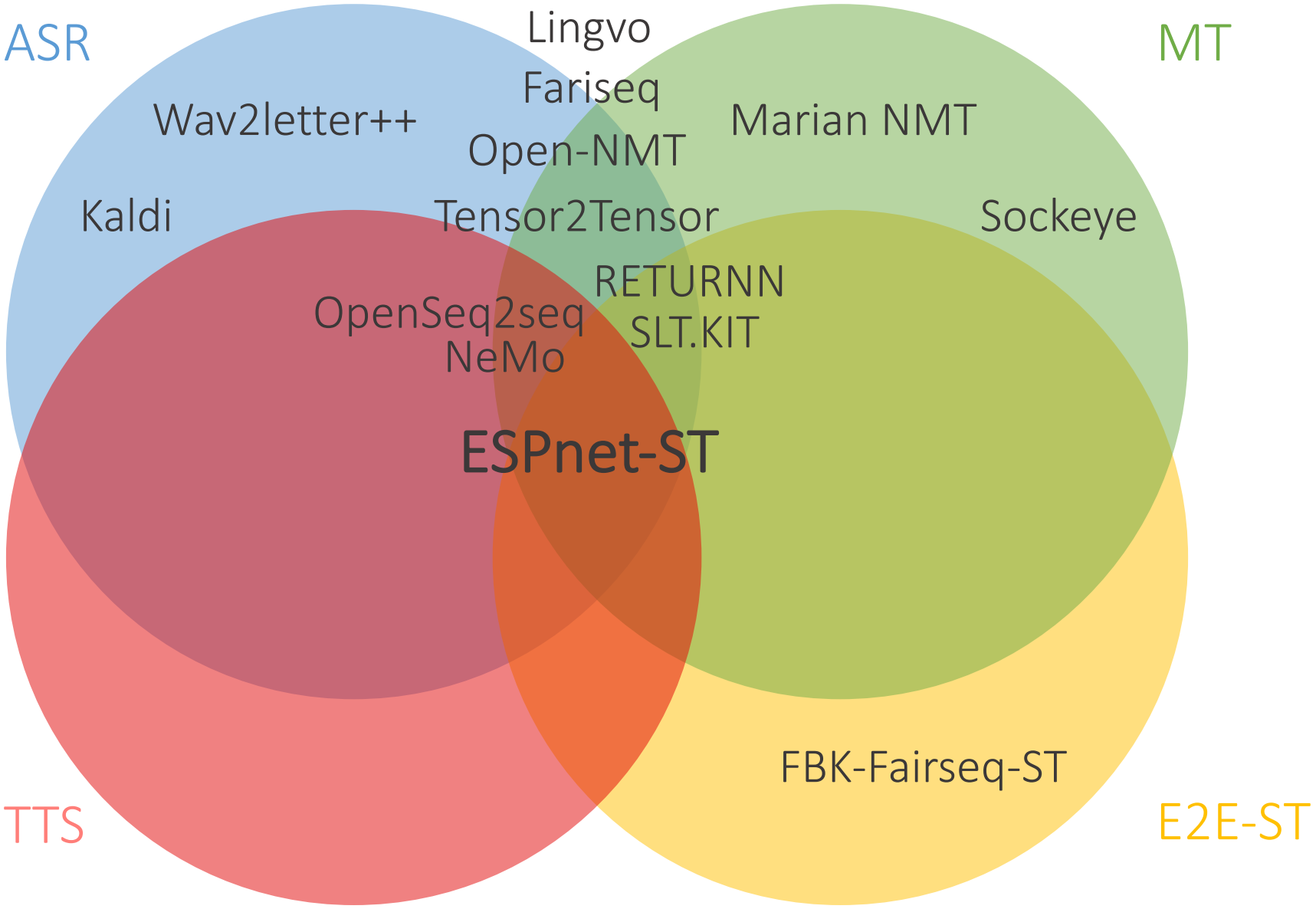
Simultaneous interpretation



Key features

- ASR/LM/MT/E2E-ST/TTS in a single unified framework
  - ➢PyTorch backend
  - ➢Both cascade and E2E speech translation systems are supported
- Reproducible SOTA results on most corpora
- Support various recipes
  - ➢All you need is to command ./run.sh
  - ➢Data downloading, data preprocessing, feature extraction, dataset construction, training, decoding
- Provide pre-trained ASR/LM/E2E-ST/MT/TTS models

Comparison with other seq2seq frameworks



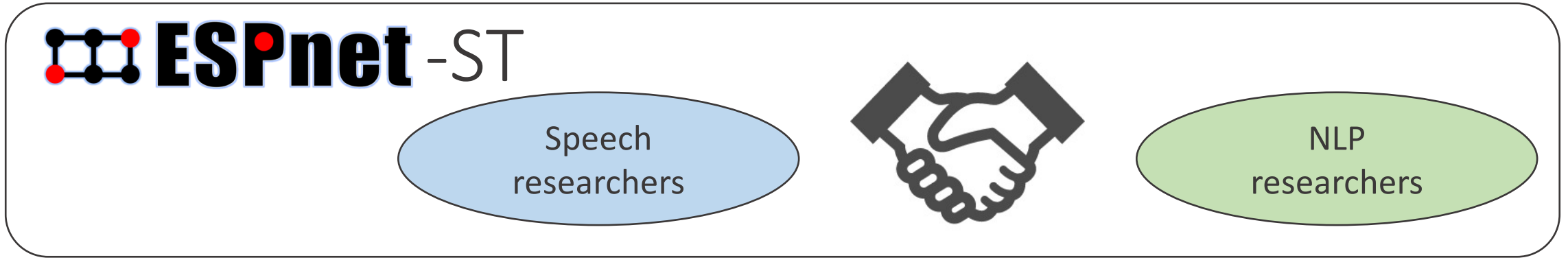
Comparison with other seq2seq frameworks

Toolkit	Supported task						Example (w/ corpus pre-processing)						Pre-trained model
	ASR	LM	E2E-ST	Cascade-ST	MT	TTS	ASR	LM	E2E-ST	Cascade-ST	MT	TTS	
ESPnet-ST (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Lingvo[1]	✓	✓	✓♣	✓♣	✓	✓♣	✓	✓	–	–	✓	–	–
OpenSeq2seq[2]	✓	✓	–	–	✓	✓	✓	✓	–	–	✓	–	✓
NeMo[3]	✓	✓	–	–	✓	✓	✓	✓	–	–	✓	–	✓
RETURNN[4]	✓	✓	✓	–	✓	–	–	–	–	–	–	–	✓
SLT.KIT[5]	✓	–	✓	✓	✓	–	✓	–	✓	✓	✓	–	✓
Fairseq[6]	✓	✓	–	–	✓	–	✓	✓	–	–	✓	–	✓
Tensor2Tensor[7]	✓	✓	–	–	✓	–	–	–	–	–	✓	–	✓◇
OpenNMT-{py, tf}[8]	✓	✓	–	–	✓	–	–	–	–	–	–	–	✓
Kaldi[9]	✓	✓	–	–	–	–	✓	✓	–	–	–	–	✓
Wav2letter++[10]	✓	✓	–	–	–	–	✓	✓	–	–	–	–	✓

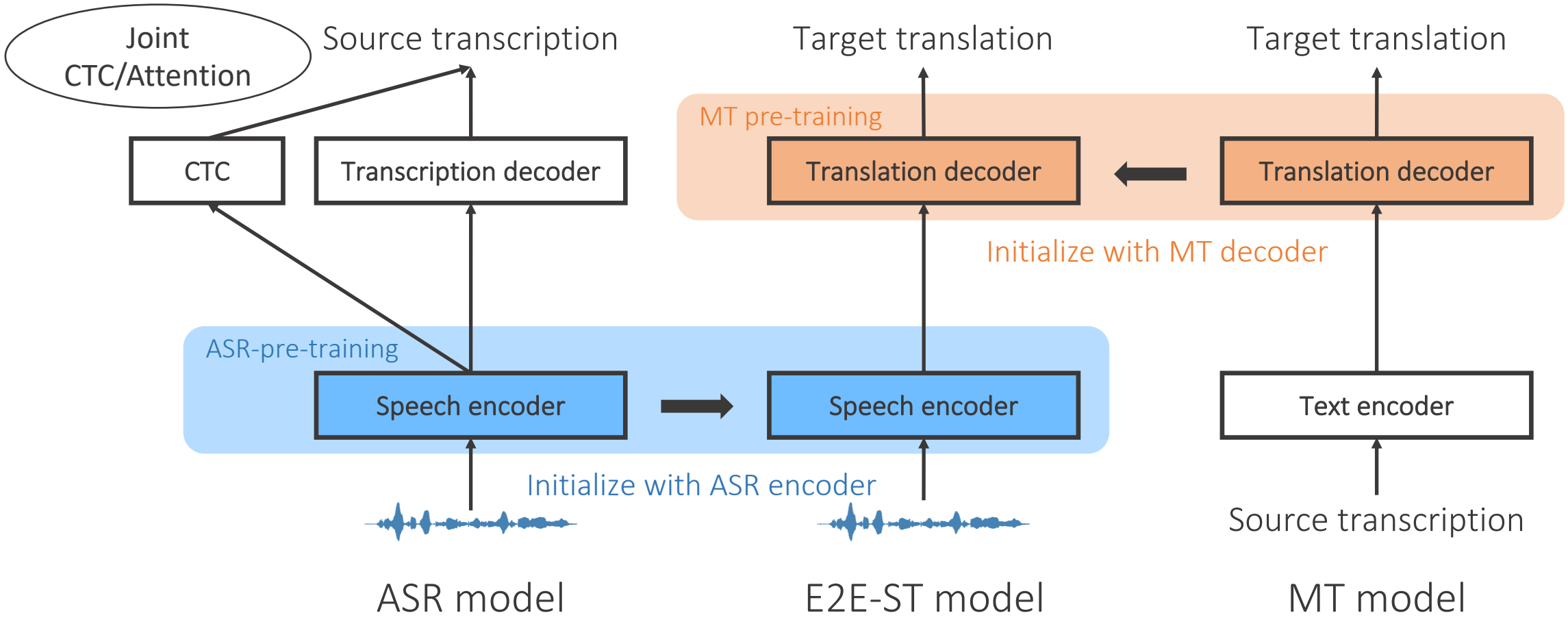
Table 1: Framework comparison on supported tasks in January, 2020. ♣Not publicly available. ◇Available only in Google Cloud storage. [1](Shen et al., 2019) [2](Kuchaiev et al., 2018) [3](Kuchaiev et al., 2019) [4](Zeyer et al., 2018) [5](Zenkel et al., 2018) [6](Ott et al., 2019) [7](Vaswani et al., 2018) [8](Klein et al., 2017) [9](Povey et al., 2011) [10](Pratap et al., 2019)

Why is the unified framework important?

- E2E-ST model is simple, but difficult to optimize
  - ➢Initialization with well-trained ASR/MT models (***pre-training***)
  - ➢Auxiliary ASR/MT objective (***multi-task learning***)
- Combining multiple toolkits for E2E-ST is difficult
  - ➢Parameter namespace does not match across toolkits
  - ➢I/O interfaces are different across toolkits



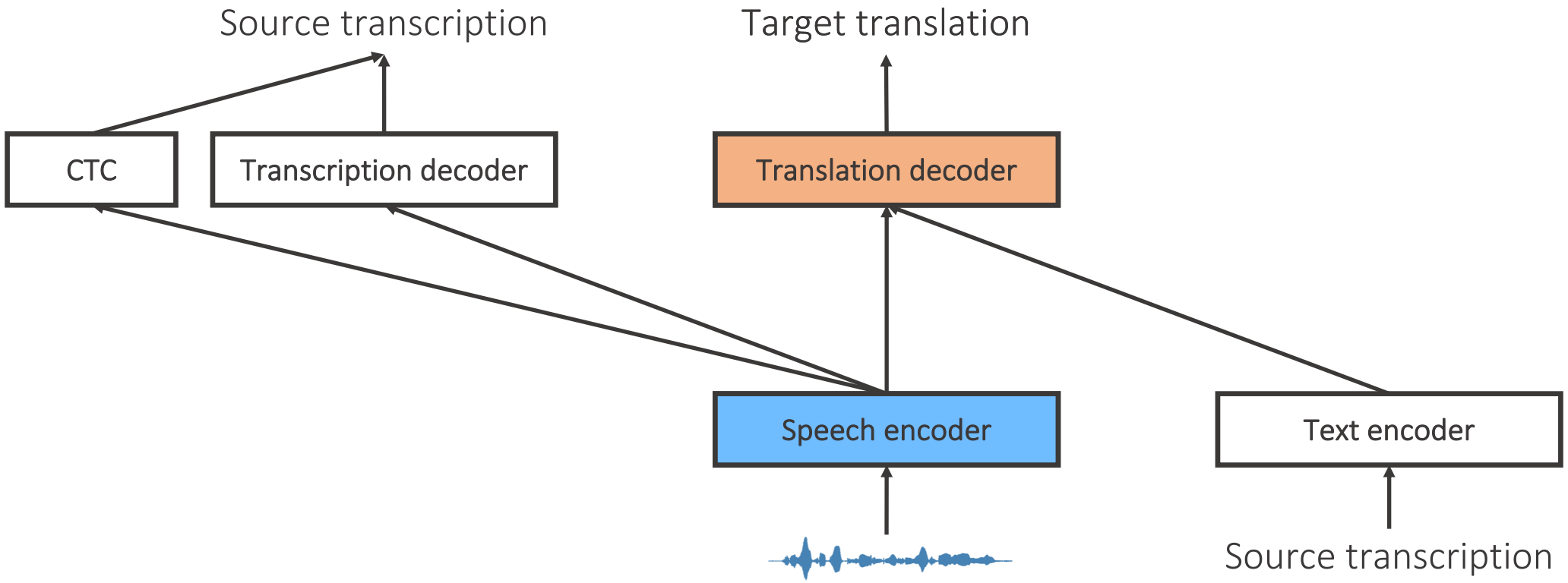
Pre-training/transfer learning from ASR/MT models



*Multilingual training is also supported

**Speech data augmentation is also supported (speed perturbation, SpecAugment)

Multi-task learning with auxiliary ASR/MT objective



- Objective function

$$\mathcal{L}_{\text{total}} = (1 - \lambda_{\text{asr}} - \lambda_{\text{mt}})\mathcal{L}_{\text{st}} + \lambda_{\text{asr}}\mathcal{L}_{\text{asr}} + \lambda_{\text{mt}}\mathcal{L}_{\text{mt}}$$

History of ESPnet

Kaldi [Povey et al., 2011]

- Most famous ASR toolkit for conventional HMM-based hybrid system

ESPnet [Watanabe et al., 2018]

- Designed for E2E-ASR systems started from Kaldi-like data preparation
- More than 2.4k stars on GitHub
- Transformer ASR reached SOTA results on Libirspeech dataset [Karita et al., 2019]
- Most ASR corpora (w/ pre-trained models + demo) are covered (38 ASR corpora@June 2020)

ESPnet-TTS [Hayashi et al., 2020]

- Extended to end-to-end TTS task
- Most TTS corpora (w/ pre-trained models + demo) are covered (12 TTS corpora@June 2020)

Recipes for multiple corpora

- Existing toolkits do not include corpus preprocessing for ST tasks
  - Manual alignment across ASR/ST/MT tasks are necessary (it's non-trivial!)

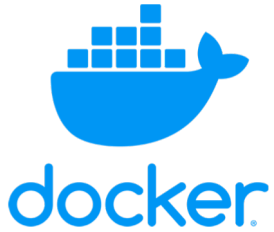
- Most benchmark ST corpora are supported (including corresponding ASR/MT recipes)

- Fisher-Callhome Spanish (Es->En, 160h)
- Libri-trans (En->Fr, 100h)
- Must-C (En->{De, Pt, Fr, Es, Ro, Ru, Nl, It}, 400h)
- How2 (En->Pt 300h)
- ST-TED (En->De, 200h)
- Mboshi-French (Mboshi->Fr, 4h)

- MT corpora
  - All ST corpora
  - IWSLT16 (En<->De)

Installation

- All relevant tools are automatically installed with Makefile
  Docker is also supported

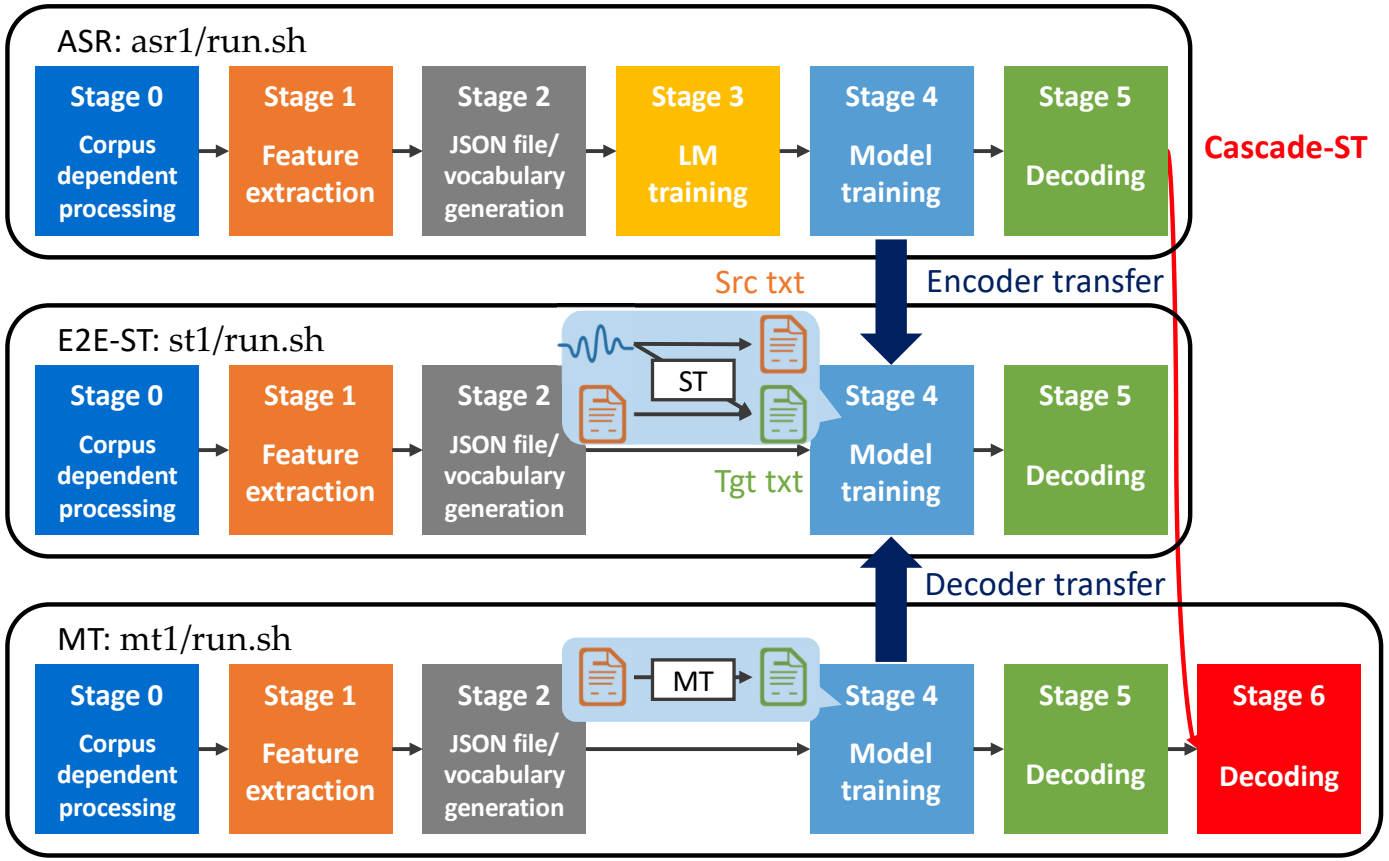


- Neural network library
  - ➢PyTorch
  - ➢warpctc-pytorch

- Preprocessing toolkit
  - ➢Kaldi (speech feature extraction)
  - ➢Moses (text pre-processing, BLEU calculation)
  - ➢Sentencepiece (vocabulary construction) etc.

Stage-by-stage processing

```
root
├── egs
│   └── corpus-name
│       ├── asr1  (ASR recipe)
│       ├── st1  (E2E-ST recipe)
│       │   ├── conf  (configuration files)
│       │   ├── local  (corpus-dependent)
│       │   ├── data  (pre-processed data)
│       │   ├── dump  (speech features)
│       │   ├── exp  (model directories)
│       │   └── run.sh  (main script)
│       └── mt1  (MT/Cascade-ST recipe)
├── espnet  (model definition)
└── tools  (tool installation)
```



Overall directory structure

Other features

Experiment manager

- Tensorboard
- Chainer reporter
- Attention plotting at every epoch
- Loss/Accuracy/BLEU/PPL plotting at every epoch

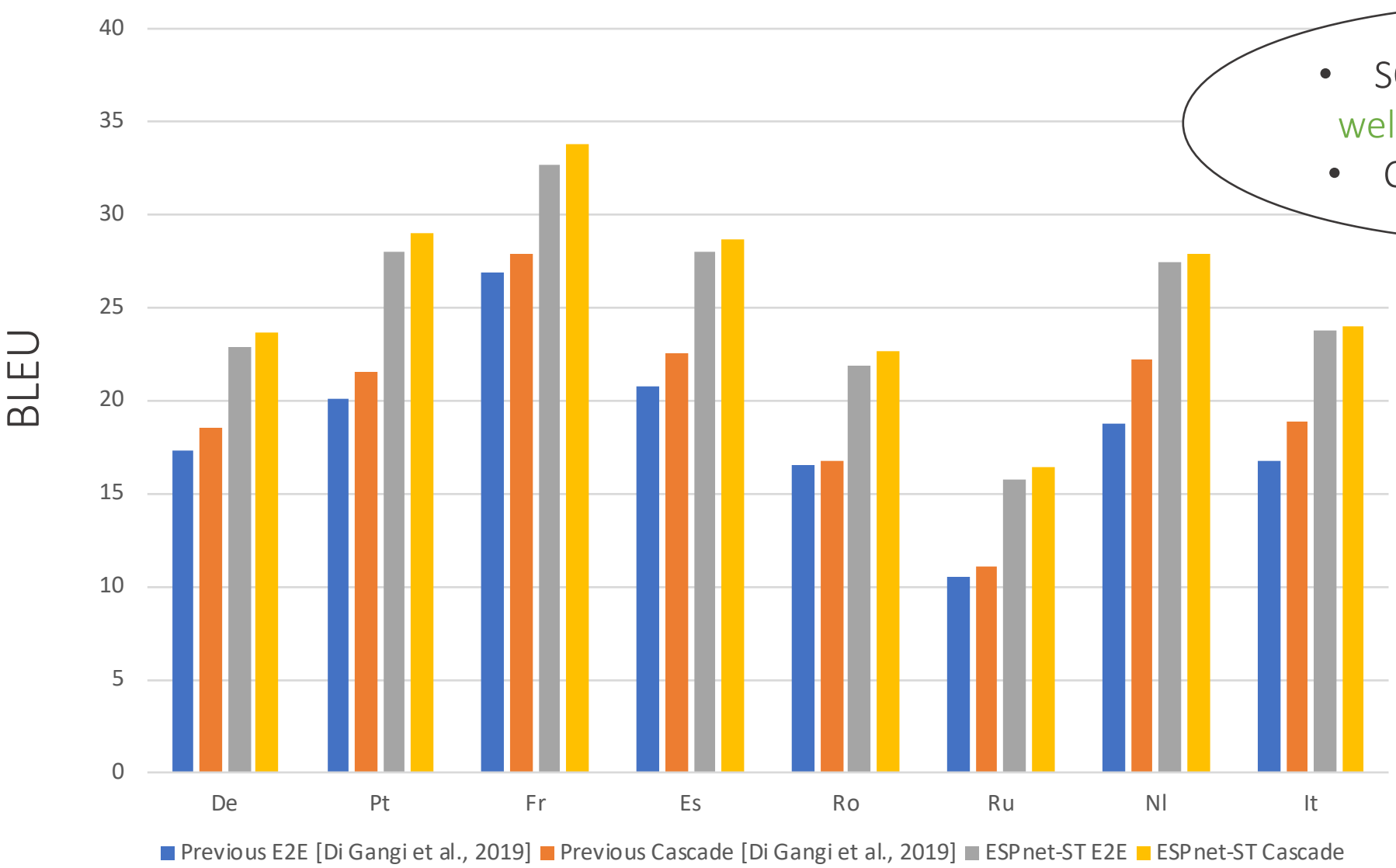
Large-scale training/decoding

- Job scheduler with SLURM/Grid Engine etc.
- Multi-GPU training
- Mixed precision training with apex
- Batch beam search decoding (on multiple CPUs)

Supported model

Task	LSTM/GRU	Transformer	Others
ASR	○	○	CTC, Joint CTC/Attention, RNN transducer, Transformer transducer, Lightweight and dynamic conv.
LM	○	○	N-gram
E2E-ST	○	○	-
MT	○	○	-
TTS	○ (Tacotron2)	○	Multi-speaker TTS, voice conversion, FastSpeech (non-autoregressive TTS)

ST results: Must-C (En->8 languages)

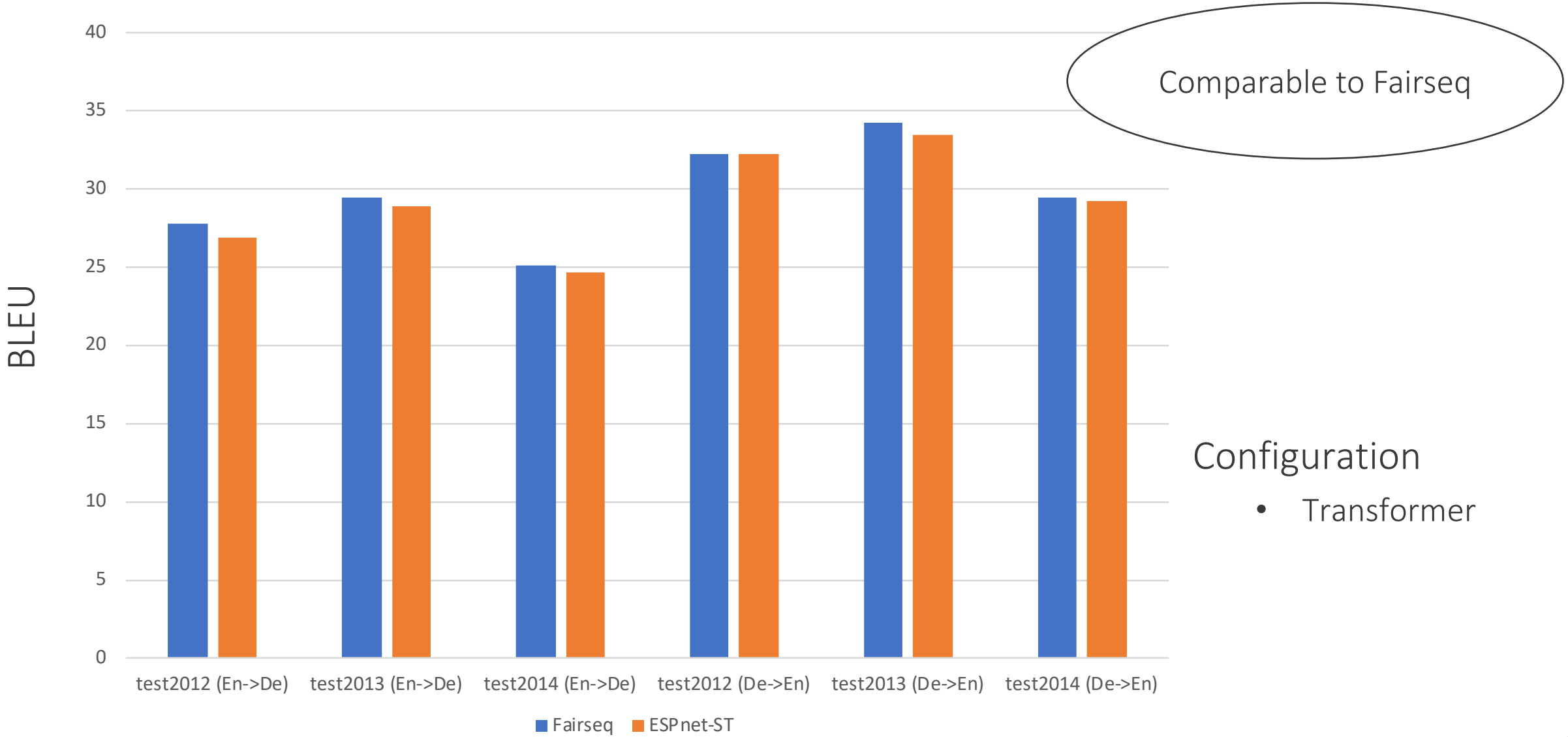


- SOTA results only with well-known techniques
- Cascade is still better

Configuration

- Transformer
- ASR pre-training
- MT pre-training
- SpecAugment
- No external data

Text-based MT results: IWSLT16 (En<->De)



Comparable to Fairseq

Configuration

- Transformer

Future work

- Support more MT corpora (e.g., WMT)
- Semi-supervised training
  - ➢Self-supervised learning with pseudo labelling
  - ➢Back translation
- Simultaneous translation

Get started ESPnet -ST

https://github.com/espnet/espnet

E2E-ST demo