# ESPnet-ST-v2: Multipurpose Spoken Language Translation Toolkit

Brian Yan, Jiatong Shi, Yun Tang, Hirofumi Inaguma, Yifan Peng, Siddharth Dalmia, Peter Polák, Patrick Fernandes, Dan Berrebbi, Tomoki Hayashi, Xiaohui Zhang, Zhaoheng Ni, Moto Hira, Soumi Maiti, Juan Pino, Shinji Watanabe

## What's New?

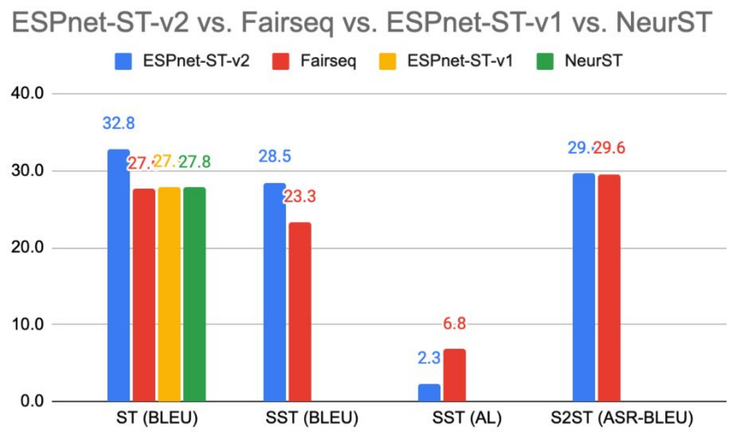
- New Tasks
  - Simultaneous speech-to-text translation,
  - Speech-to-speech translation
- New Core Architectures
  - CTC/Attention,
  - Multi-Decoder (E2E differentiable cascade),
  - Transducer
- New Auxiliary Techniques
  - Hierarchical Encoding,
  - Speech SSL Representations,
  - LLM Pre-trained Initializations,
  - MBR Ensembling,
  - Stable Hyp Detection

\+ Variety of Example Models

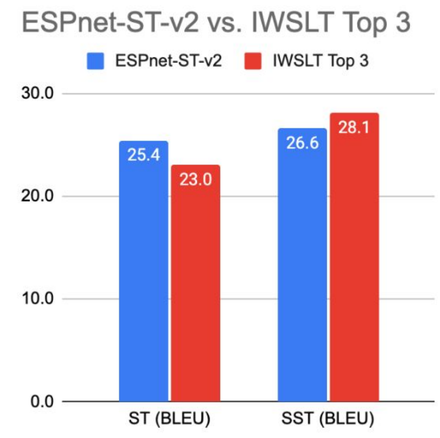
| Features | ESPnet-ST-v2 | ESPnet-ST-v1 | Fairseq-S2T | NeurST |
|---|---|---|---|---|
| **Offline ST** | ✓ | ✓ | ✓ | ✓ |
| End-to-End Architecture(s) | ✓ | ✓ | ✓ | ✓ |
| Attentional Enc-Dec | ✓ | ✓ | ✓ | ✓ |
| CTC/Attention | ✓ | - | - | - |
| Transducer | ✓ | - | - | - |
| Multi-Decoder | ✓ | - | - | - |
| Hierarchical Encoding | ✓ | - | - | - |
| Speech SSL Representations | ✓[1] | - | ✓ | - |
| Speech & Text Pre-training | ✓ | ✓ | ✓ | ✓ |
| Minimum Bayes Risk Ensemble | ✓ | - | - | - |
| Joint Speech/Text Pre-training | - | - | ✓ | - |
| Cascaded Architectures | ✓ | ✓ | ✓ | ✓ |
| **Simultaneous ST** | ✓ | - | ✓ | ✓[3] |
| End-to-End Architecture(s) | ✓ | - | ✓ | - |
| Contextual Block Encoders | ✓ | - | - | - |
| Blockwise Attn Enc-Dec | ✓ | - | - | - |
| Blockwise CTC/Attention | ✓ | - | - | - |
| Blockwise Transducer | ✓ | - | - | - |
| Wait-K Attn Enc-Dec | - | - | ✓ | - |
| Monotonic Attn Enc-Dec | - | - | ✓ | - |
| Stable Hypothesis Detection | ✓ | - | - | - |
| Cascaded Architectures | - | - | ✓ | ✓[3] |
| **Offline S2ST** | ✓ | - | ✓ | - |
| End-to-End Architecture(s) | ✓ | - | ✓ | - |
| Spec Enc-Dec (Translatotron) | ✓ | - | ✓ | - |
| Spec Multi-Dec (Translatotron 2) | ✓ | - | ✓ | - |
| Discrete Enc-Dec (Speech-to-Unit) | ✓ | - | ✓ | - |
| Discrete Multi-Decoder (UnitY) | ✓ | - | ✓ | - |
| Speech SSL Representations | ✓[1] | - | ✓ | - |
| Neural Vocoder Support | ✓[2] | ✓ | ✓ | - |

## Benchmarking

- **5 BLEU** better than other toolkits for ST and SST; on par for S2ST
- Large scale models (more data, more params) are **competitive with IWSLT systems**

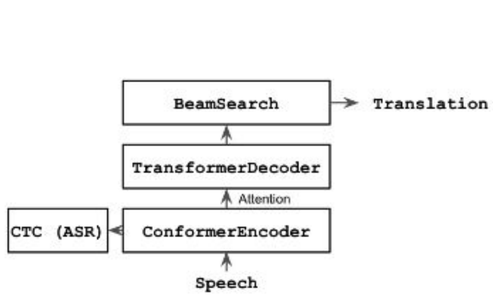



*Fairseq's latest S2ST implementation is not public as of June 2023




*Comparing vs. 2021 systems for ST, 2022 for SST

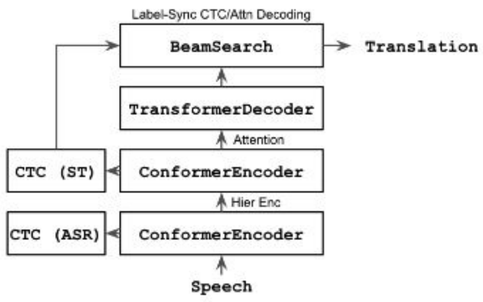
## Speech-to-Text Example Models




### Attentional Enc-Dec

+ simple, easy to use
- no hard alignments

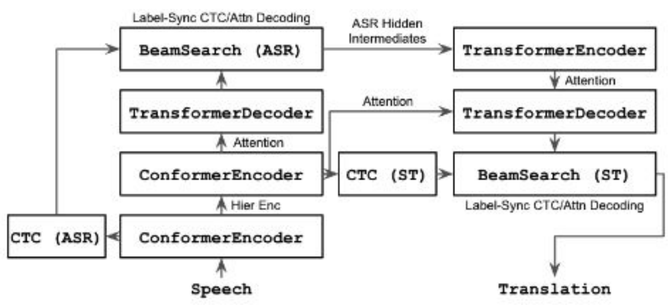
Translation Quality ★★
Streaming Compatibility ★★
Ease of Use ★★★★★




### CTC/Attention

+ hard & soft alignments
- cost of joint decoding

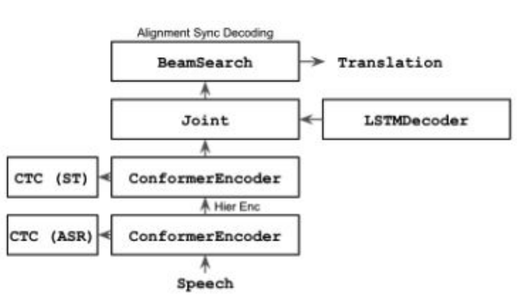
Translation Quality ★★★
Streaming Compatibility ★★★
Ease of Use ★★★




### Multi-Decoder CTC/Attn

+ E2E differentiable cascade
- cost of cascaded inference

Translation Quality ★★★★
Streaming Compatibility ★
Ease of Use ★★

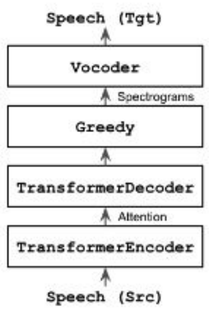



### Transducer

+ autoregressive hard alignment
- more difficult to train

Translation Quality ★★
Streaming Compatibility ★★★★
Ease of Use ★

## Speech-to-Speech Example Models

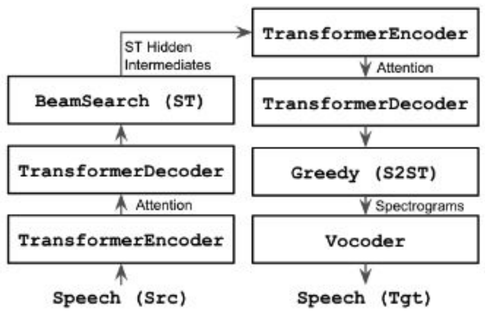



### Spectral Attn Enc-Dec

*aka Translatotron*

+ simple, easy to use
- low quality, slower training

Translation Quality ★★
Ease of Use ★★

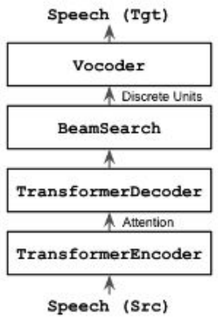



### Spectral Multi-Decoder

*aka Translatotron2*

+ good quality
- slower training, cascaded inf.

Translation Quality ★★★
Ease of Use ★

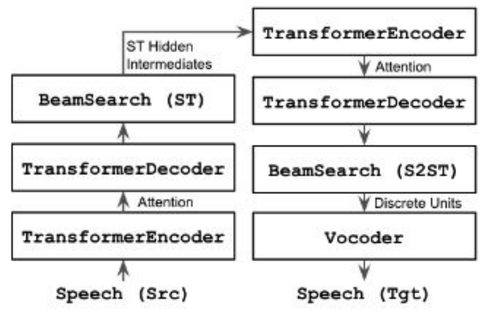



### Discrete Attn Enc-Dec

*aka Speech-to-Unit*

+ good quality, faster training
- reliant on SSL

Translation Quality ★★★
Ease of Use ★★★★




### Discrete Multi-Decoder

*aka UnitY*

+ best quality, faster training
- reliant on SSL, cascaded inf.

Translation Quality ★★★★★
Ease of Use ★★★

## Links

### Code

https://github.com/espnet/espnet

### Paper

https://arxiv.org/abs/2304.04596

### Demo

http://bit.ly/3XHX9OT

Carnegie Mellon University
Language Technologies Institute
CMU-LTI WAVLab
Meta
{byan, jiatongs}@cs.cmu.edu