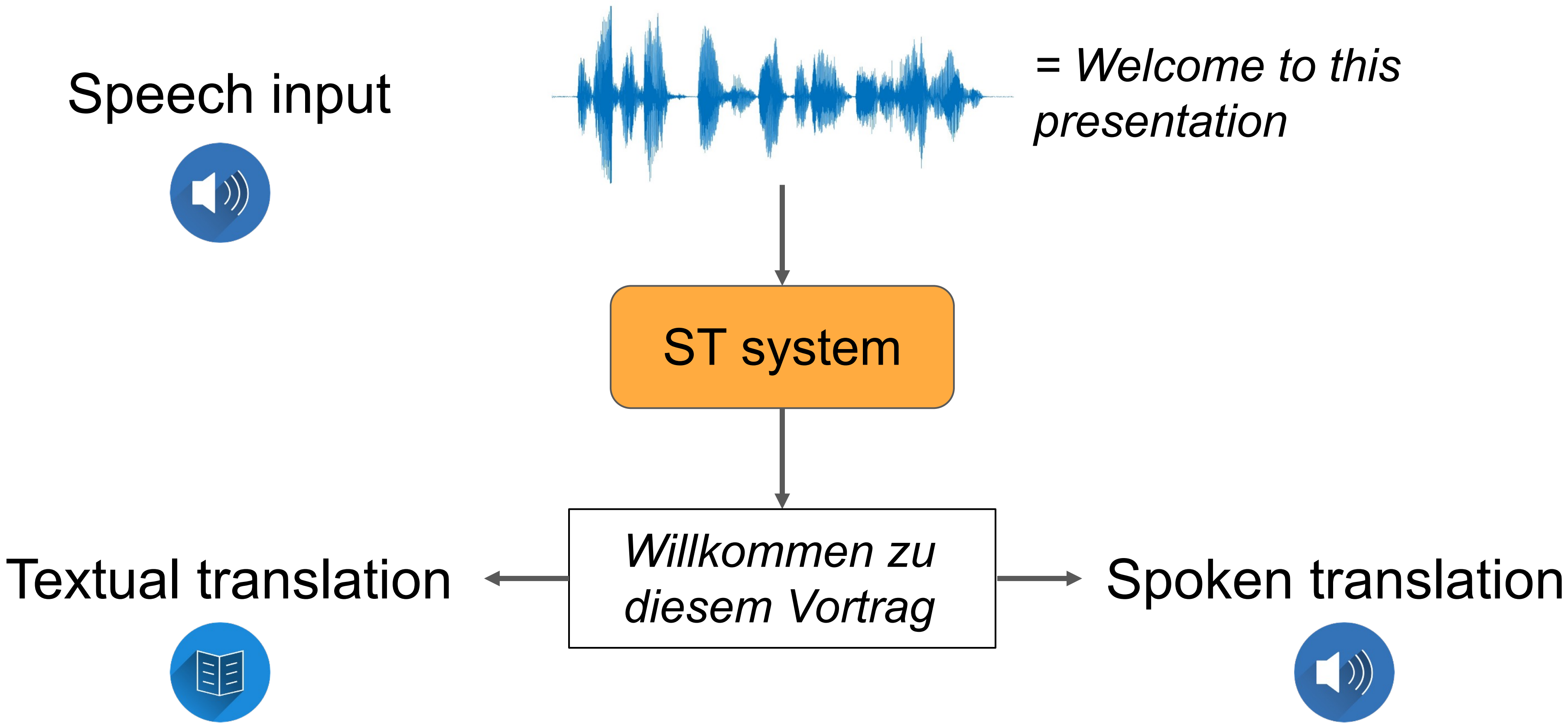
Karlsruhe Institute of Technology

End to End Speech to Speech Translation

Jan Niehues

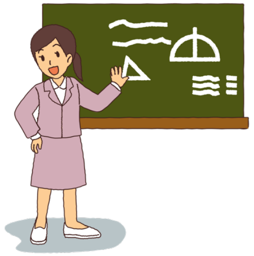
Speech Translation - Task



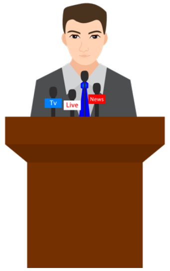
Motivation

Karlsruhe Institute of Technology

- Globalized world enables interaction between people from many cultures



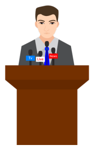




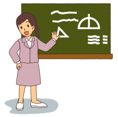
- Language barrier still main issue
  - Human interpretation or broken English
  - Complement by automated speech translation?


Different Application Scenarios

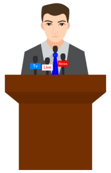
- Sequence
  - Consecutive translation
  - Simultaneous translation




- Number of speakers
  - Single/Multiple speaker

- Online/Offline systems
  - Latency: Time passes between speech & translation



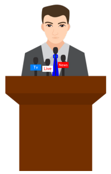
- Output Modality

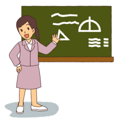

History

1991 Janus: First speech translation system for limited domains

1990s: Limited domain consecutive translation (e.g. Verbmobil)

2004-2007: Open-Domain Continuous Translation (TC-Star (European Parliament))





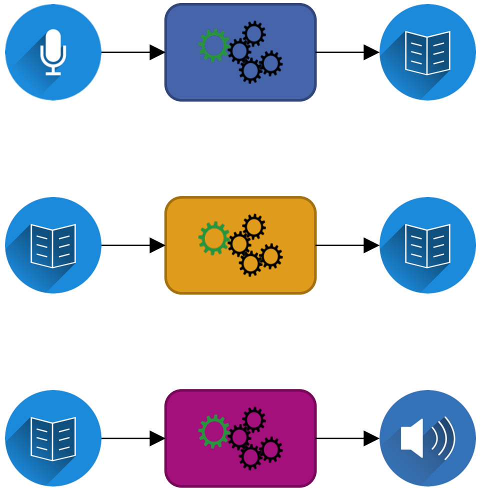
2012: KIT: Simultaneous translation of lectures
2015: Rise of deep learning

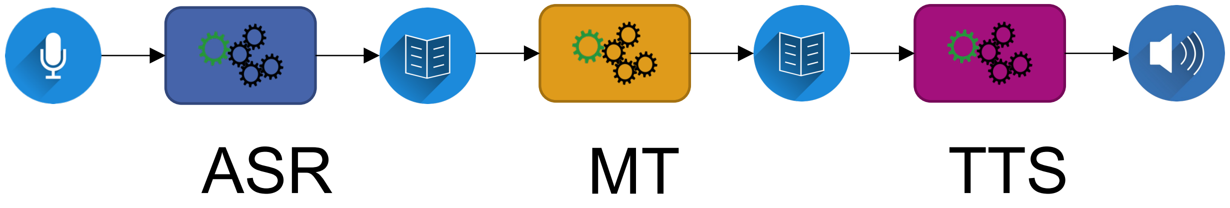
2016: Translator apps (Microsoft)
2019: PowerPoint integration

Automated Speech Translation

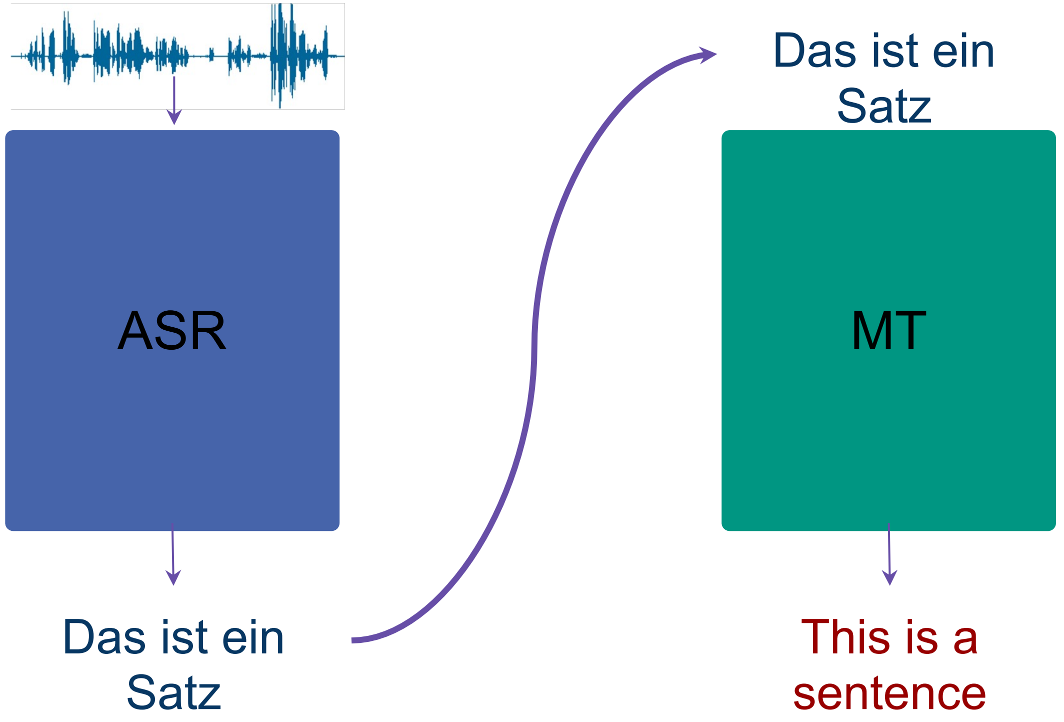
Karlsruhe Institute of Technology

- Basic Technology
  - Automatic speech recognition (ASR)
    - Transcript audio into source language text
  - Machine translation (MT)
    - Translate from source language to target language
  - Text-to-Speech (TTS)
- Serial combination of several components

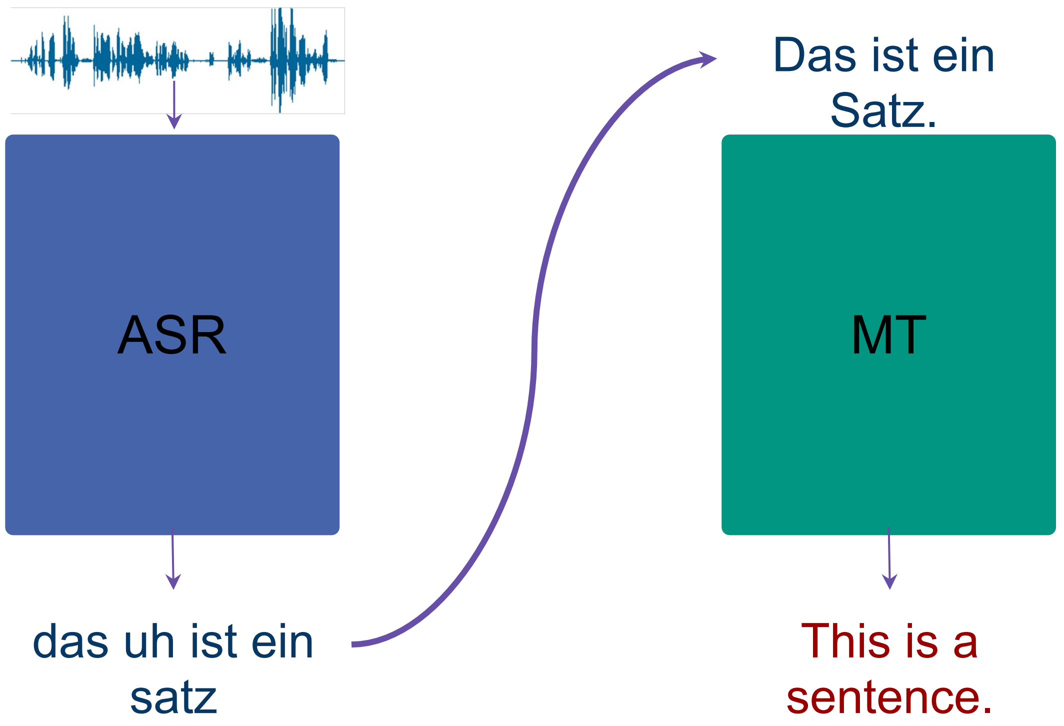




Cascaded Combination

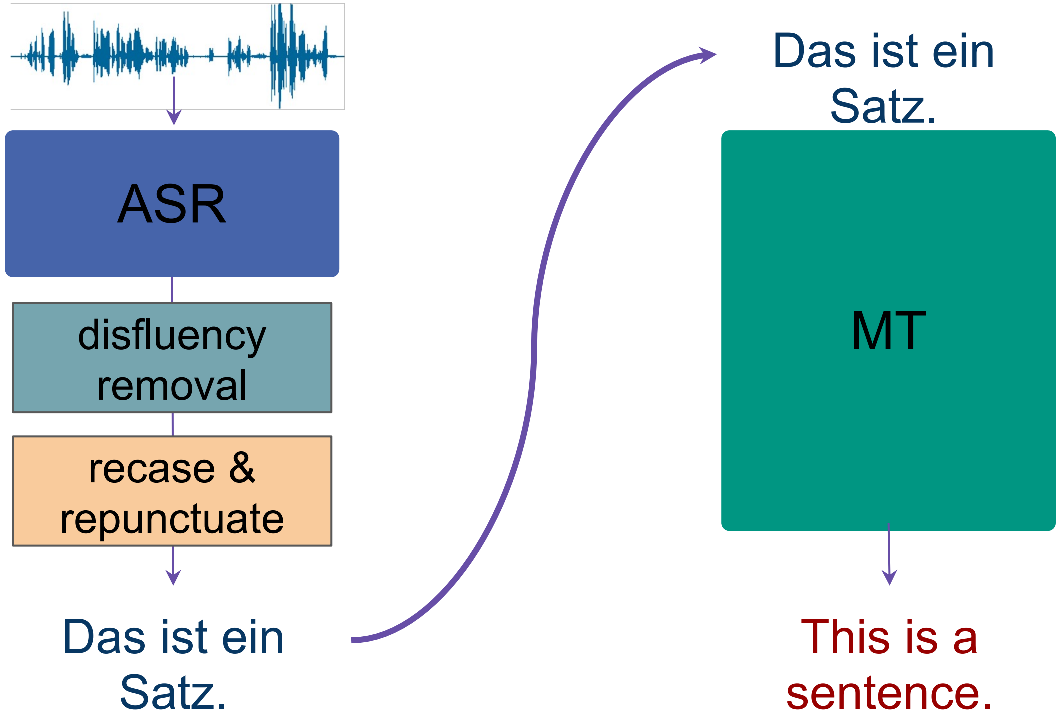


Cascaded Combination



Cascaded Combination

Cho et al., 2017



Challenges - Cascade

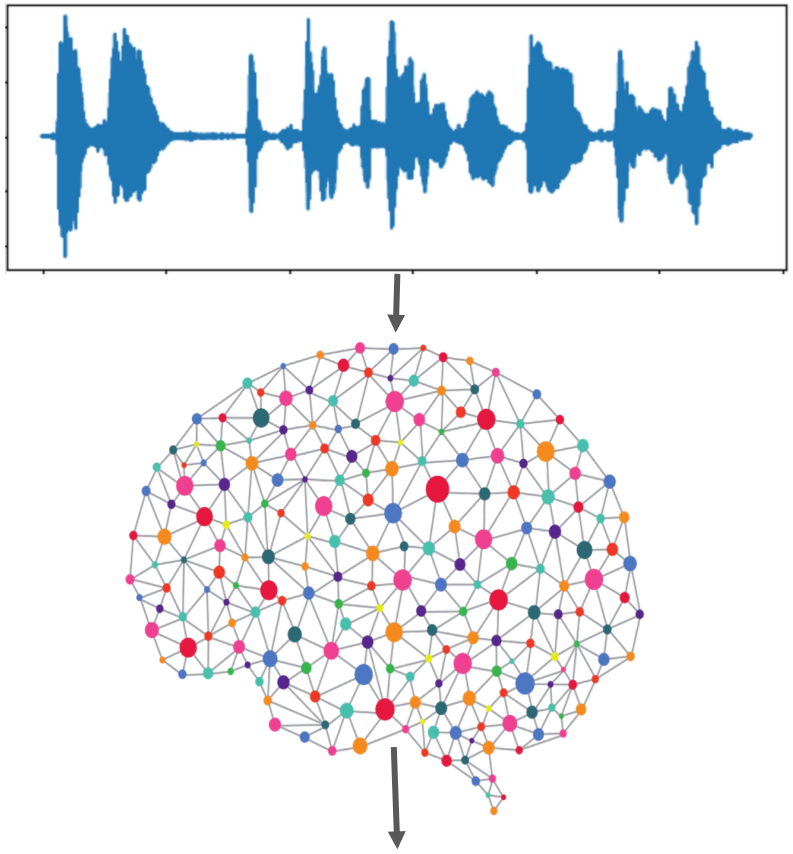
- Error propagation
  - ASR errors worse after translation
  - More difficult to compensate by human
  - MT adds additional errors

Re**d**en (engl. speeches)

Re**b**en (engl. vines)

- Opportunity:
  - Similar technology for ASR and MT

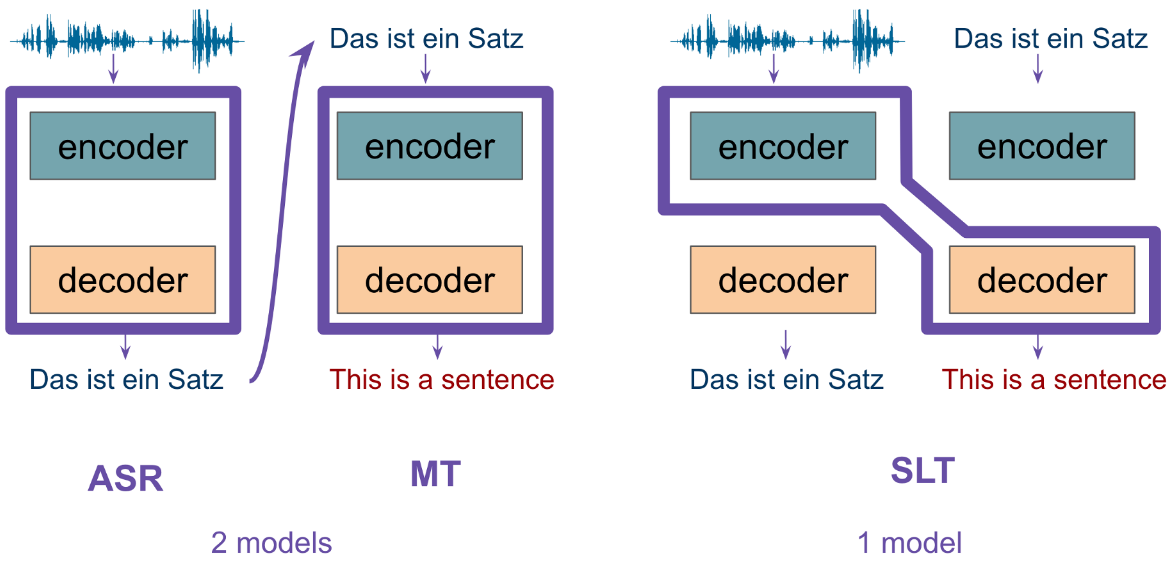
End-to-end SLT



What a wonderful world!

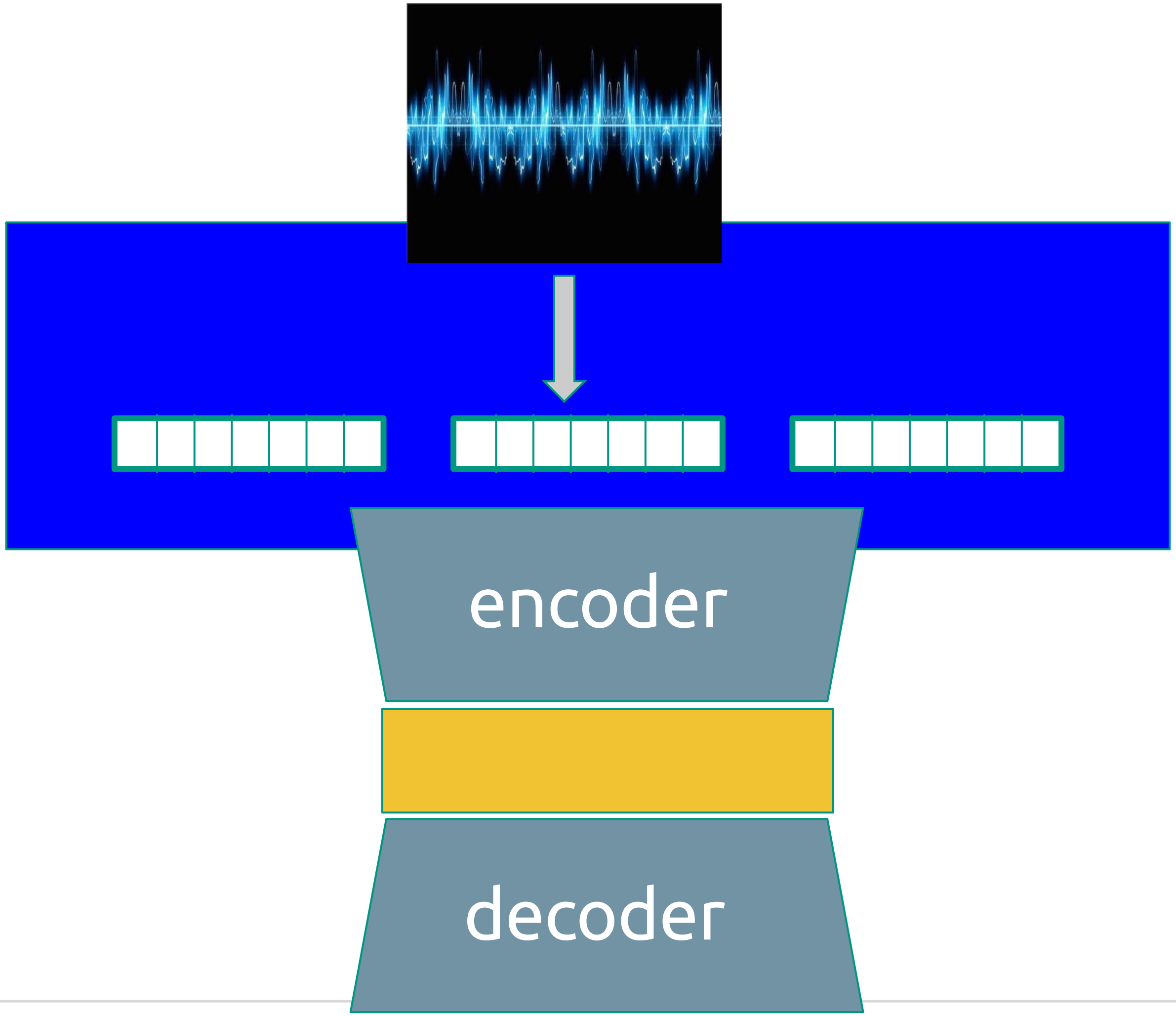
(Bérard et al., 2016;
Weiss et al., 2017)

End-to-End Speech Translation



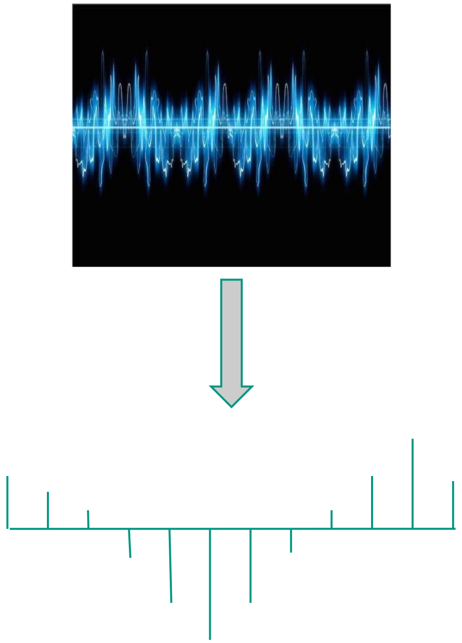
From text translation to speech translation

- Encoder-decoder models:
  - Can apply similar techniques
- Main differences to text translation
  - Input: Audio signal
    - Continuous
    - Longer



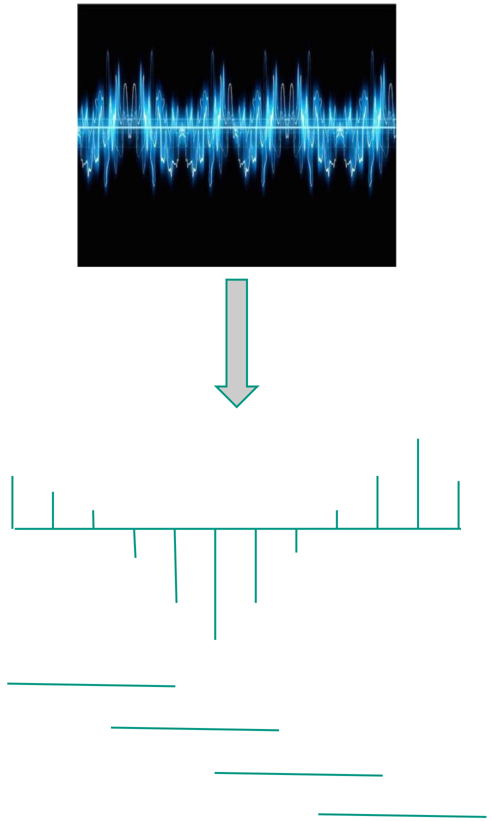
Audio representation

- Following best-practice from ASR
- Sampling
  - Measure Amplitude of signal at time t
  - Typically 16 kHz



Audio representation

- Following best-practice from ASR
- Sampling
  - Measure Amplitude of signal at time t
  - Typically 16 kHz
- Windowing
  - Split signal in different windows
    - Length: ~ 20-30 ms
    - Shift: ~ 10 ms
- Result:
  - One representation every 10 ms

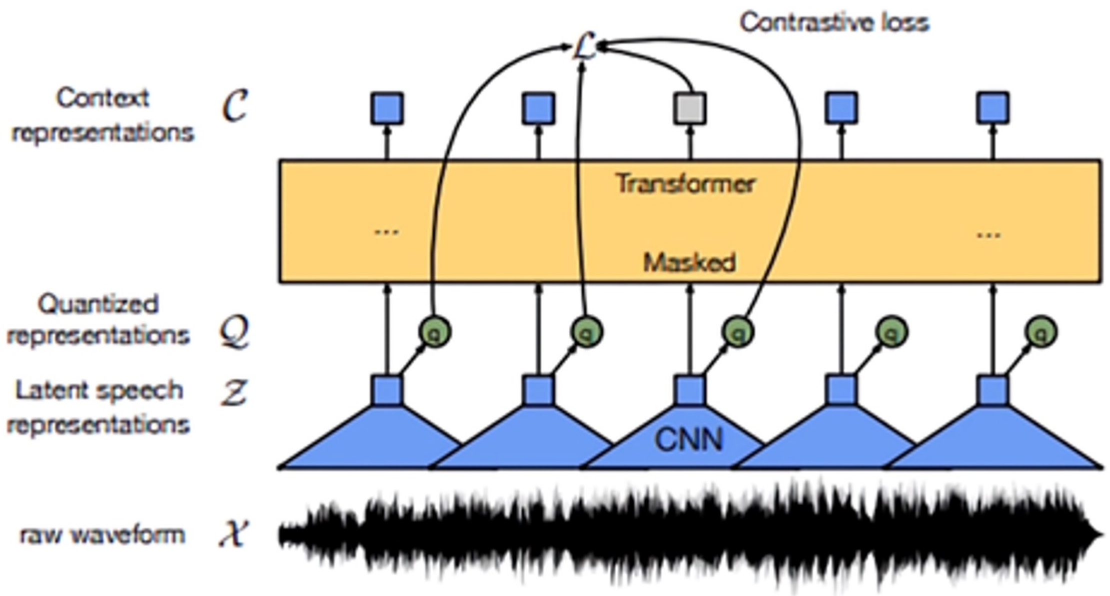


Audio representation

- Input features:
  - Signal processing:
    - Most common:
      - Mel-Frequency Cepstral Coefficients (MFCC)
      - Log mel-filterbank features (FBANK)
    - Idea:
      - Analyse frequencies of the signal
    - Steps:
      - Discrete Fourier Transformation
      - Mel filter-banks
      - Log scale
      - (Inverse Discrete Fourier Transformation)
    - Size:
      - 20-100 features per frame

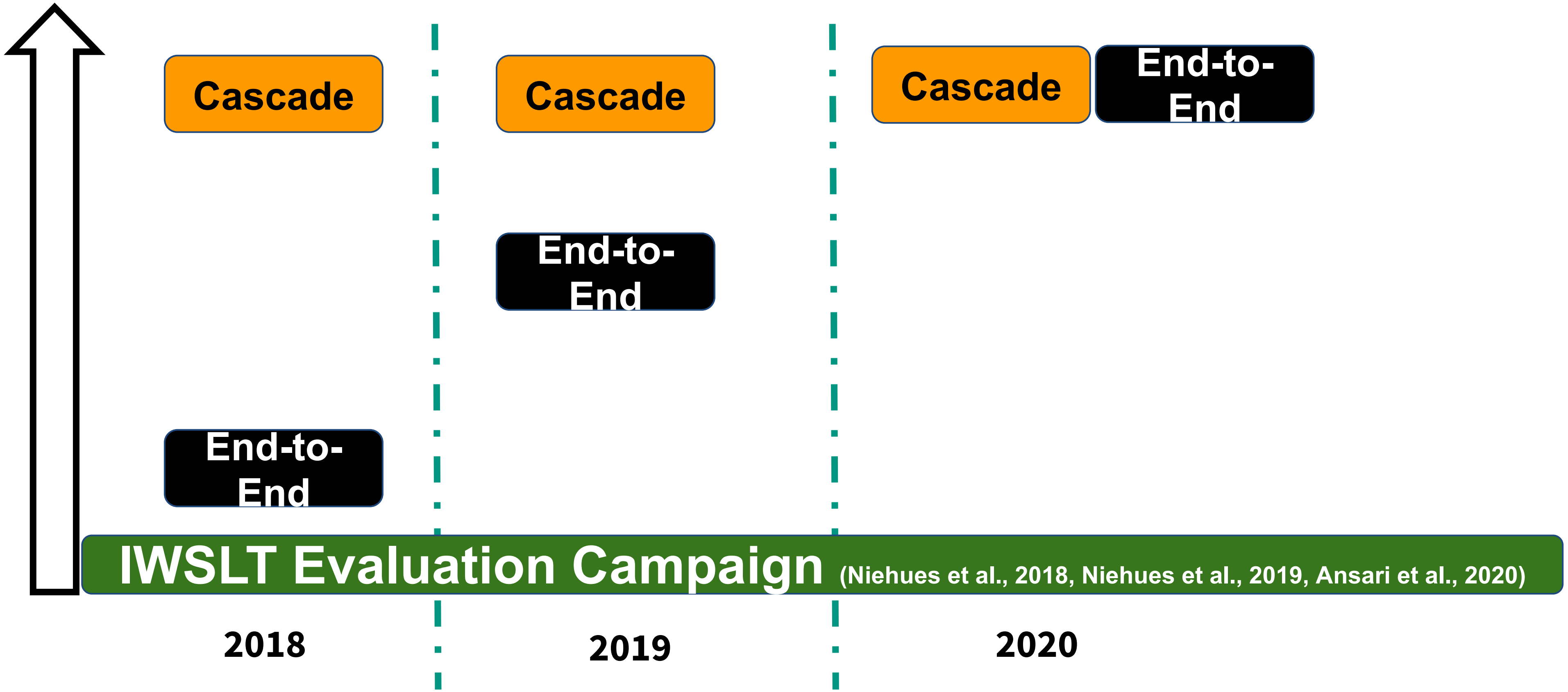
Audio representation

- Input features:
  - Signal processing:
  - Deep Learning:
    - Self-supervised Learning
      - Predict frame based on context
    - E.g. Wav2Vec 2.0

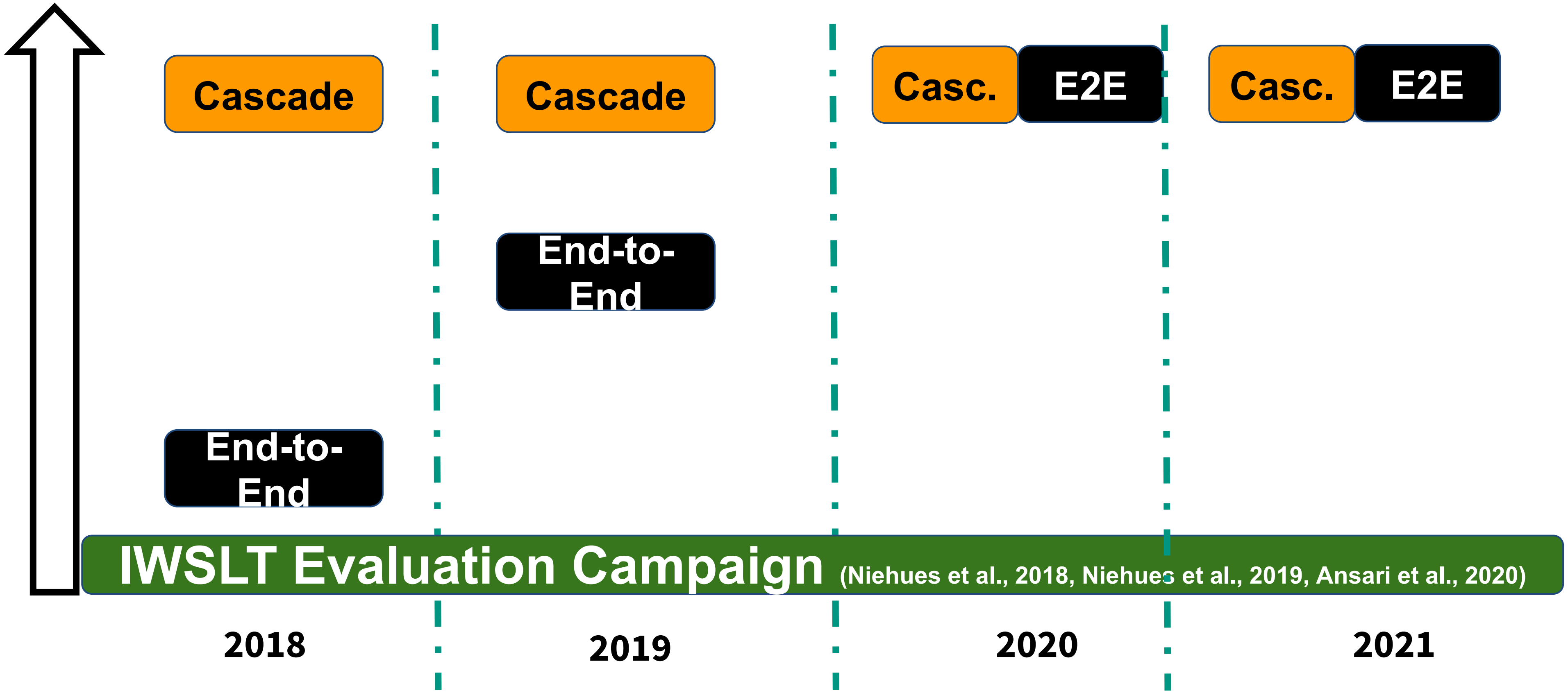


Baevski et al. 2020

Cascade vs End-to-End Systems



Cascade vs End-to-End Systems



Cascade vs End-to-End Systems

Karlsruhe Institute of Technology

Cascade

- ✓ Large corpora for ASR and MT
- ✓ Less complex tasks
- Error propagation
- Information loss
- Higher latency

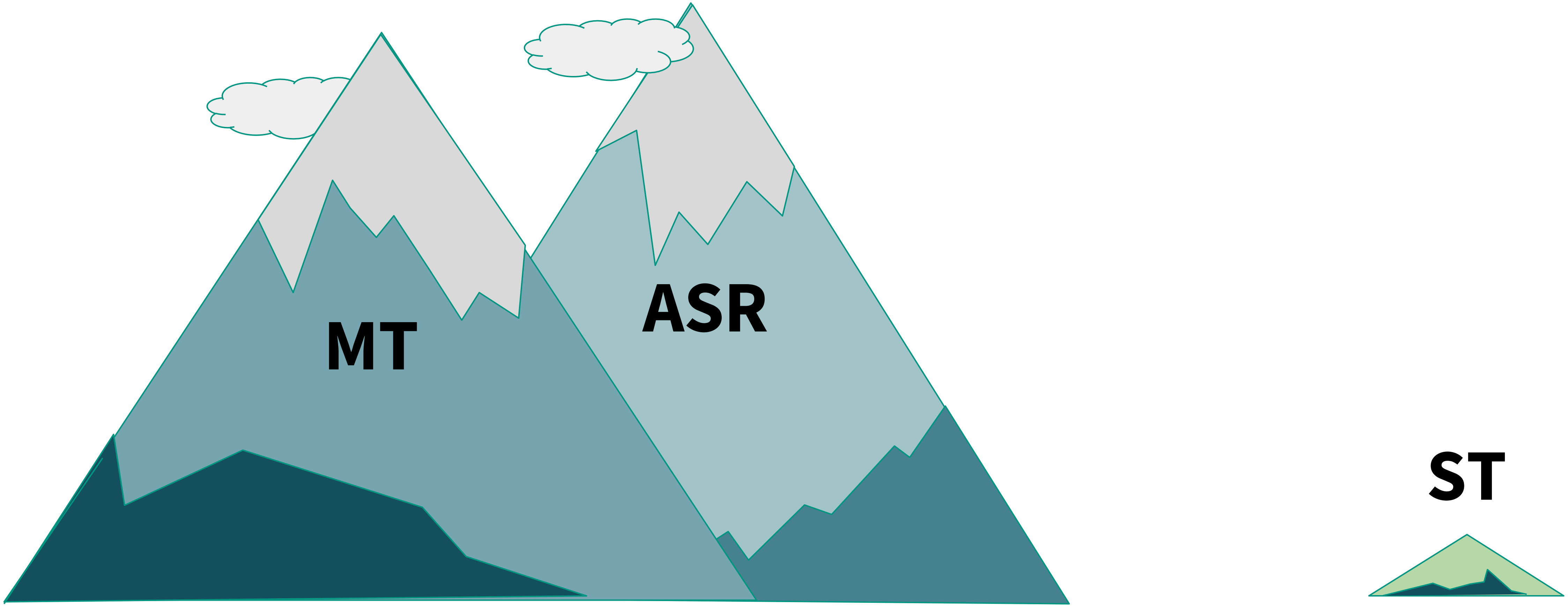
End-to-End

- ✓ Access to all audio information
- ✓ Reduced latency
- ✓ Easier management
- Small corpora
- More complex task

Challenges

- Data
  - Other data sources
  - Pre-trained models
- Audio
  - Input length
  - High variability
  - Unsegmented
- Output
  - Audio
  - Low latency
  - Additional constraints

Available data

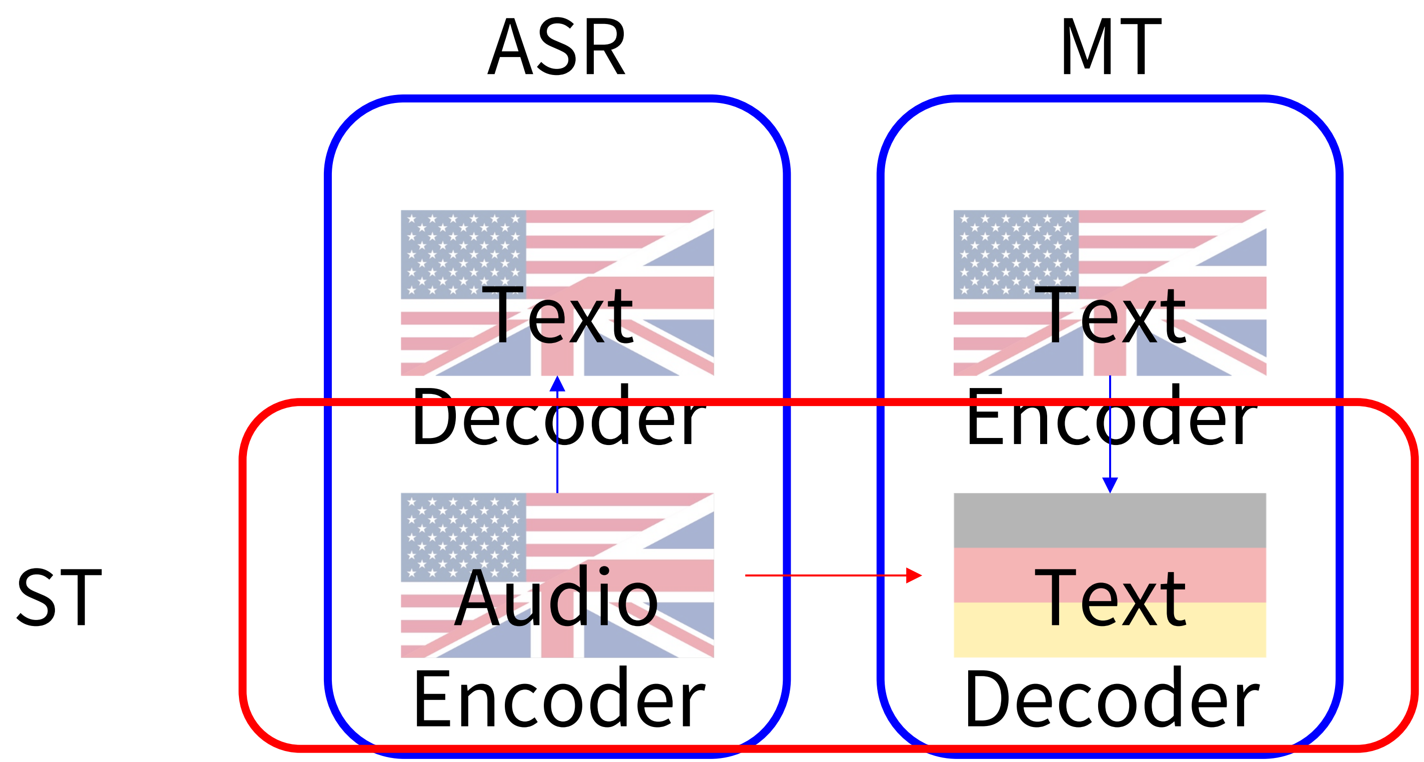


(text, translation)

(audio, transcript)

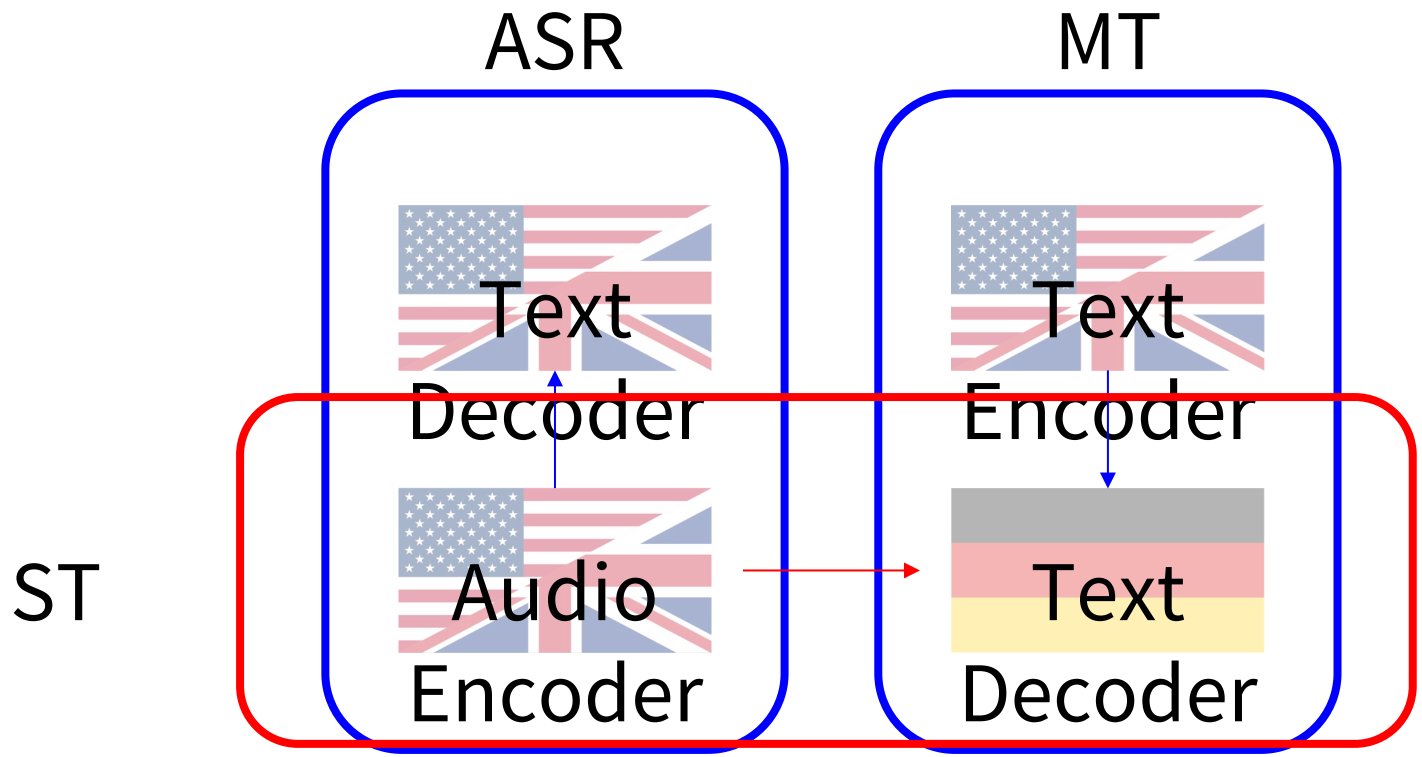
(audio, transcript, translation)

Integration of additional data sources



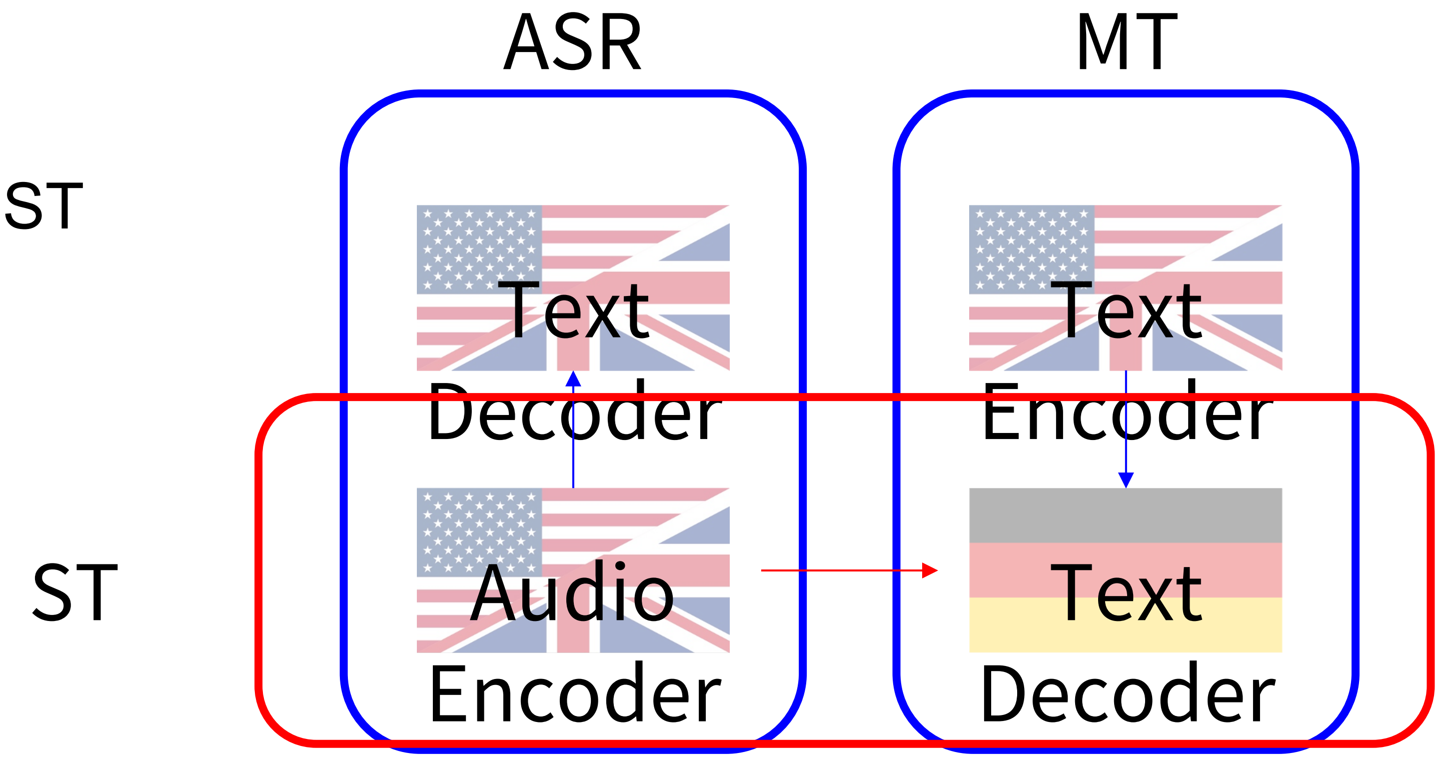
Setting

- Multi-task
  - Train all three tasks jointly



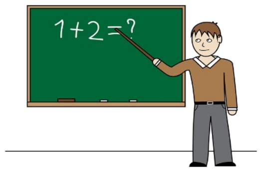
Setting

- Multi-task
- Pre-training
  - Train ASR and MT
  - Reuse part of the model for ST

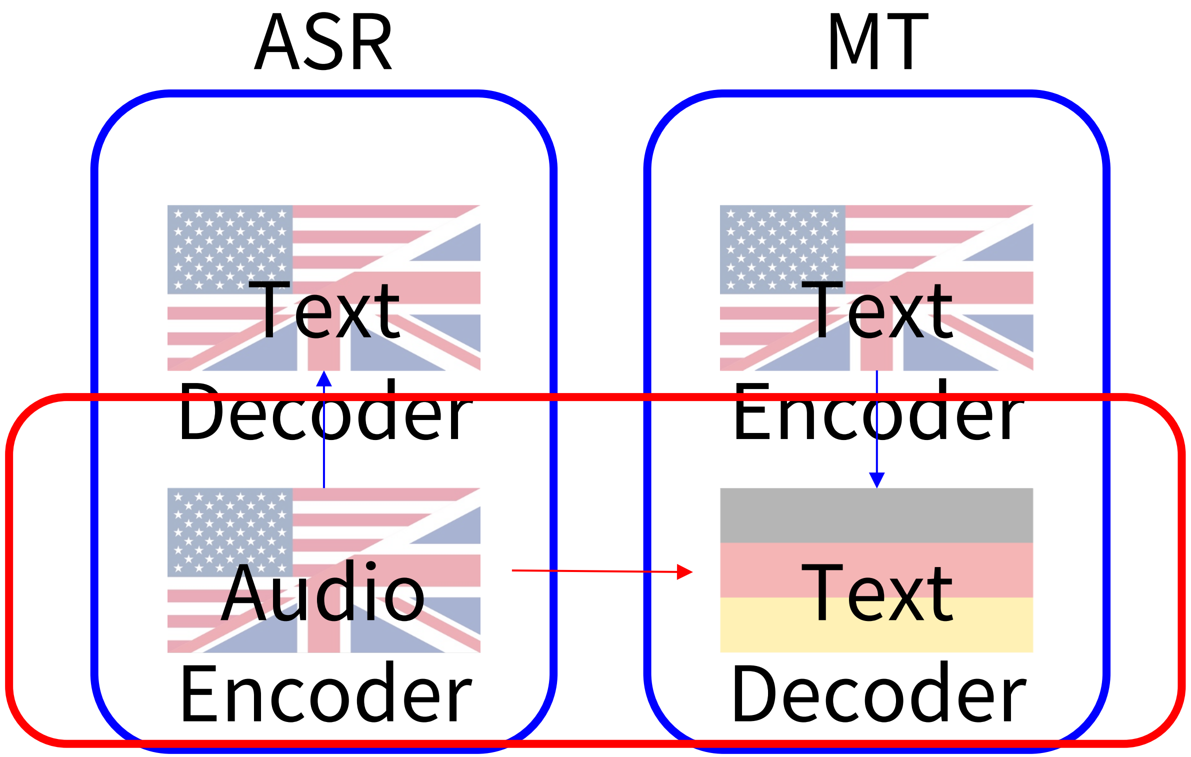


Setting

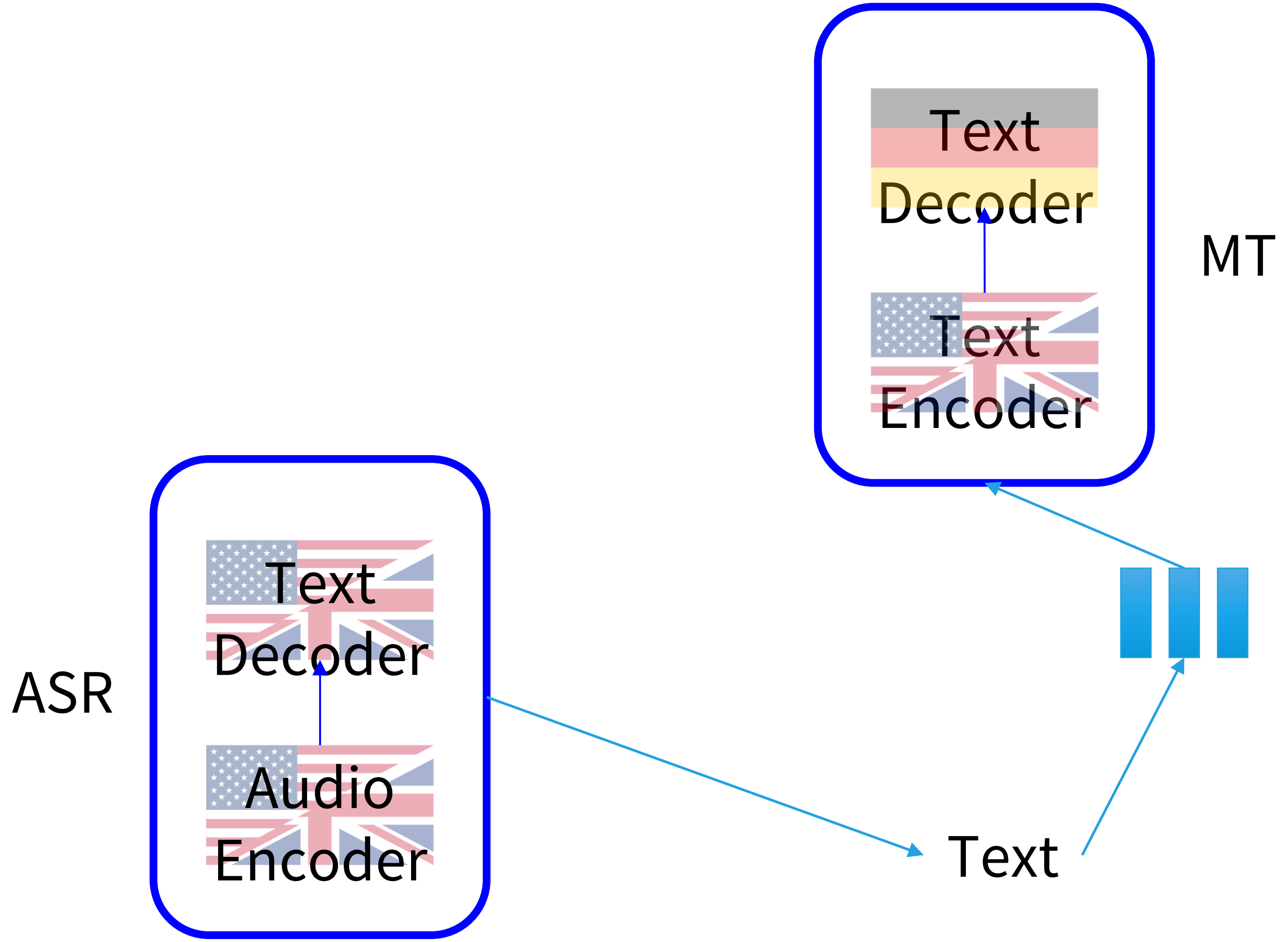
- Multi-task
- Pre-training
- Knowledge distillation
  - Take MT model
  - Train ST based on training signal from MT



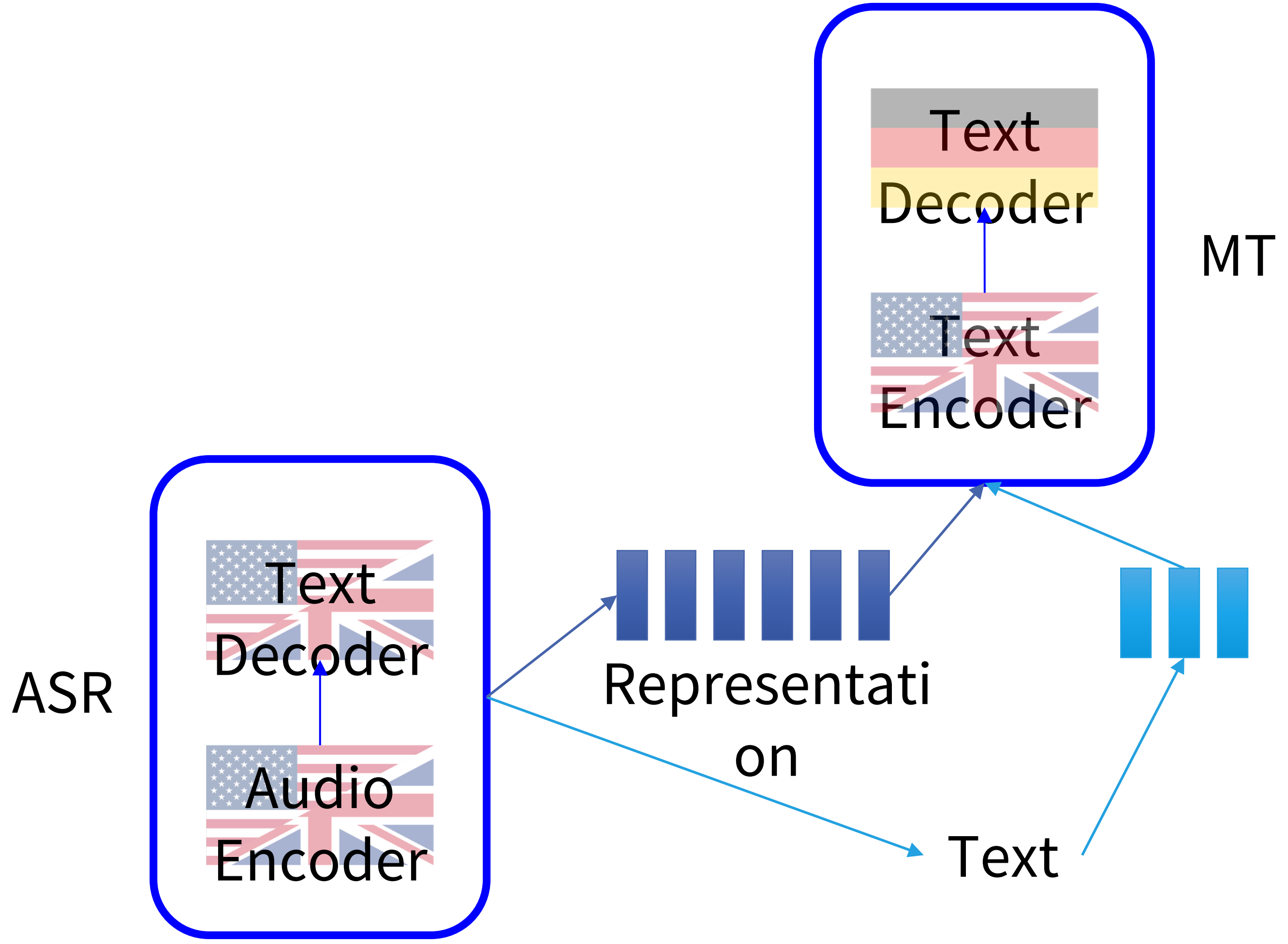
ST



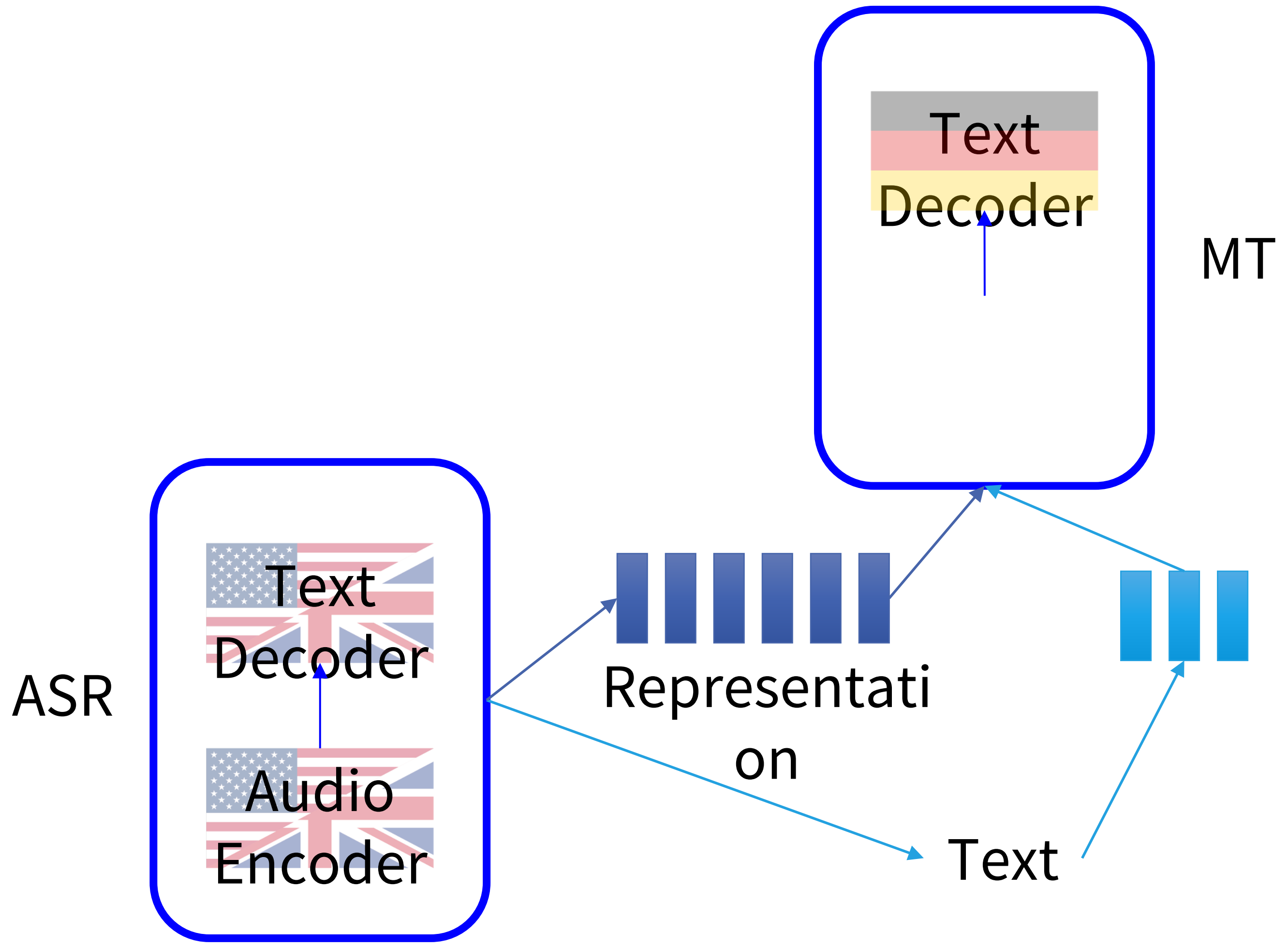
Integrating pre-trained models



Integrating pre-trained models

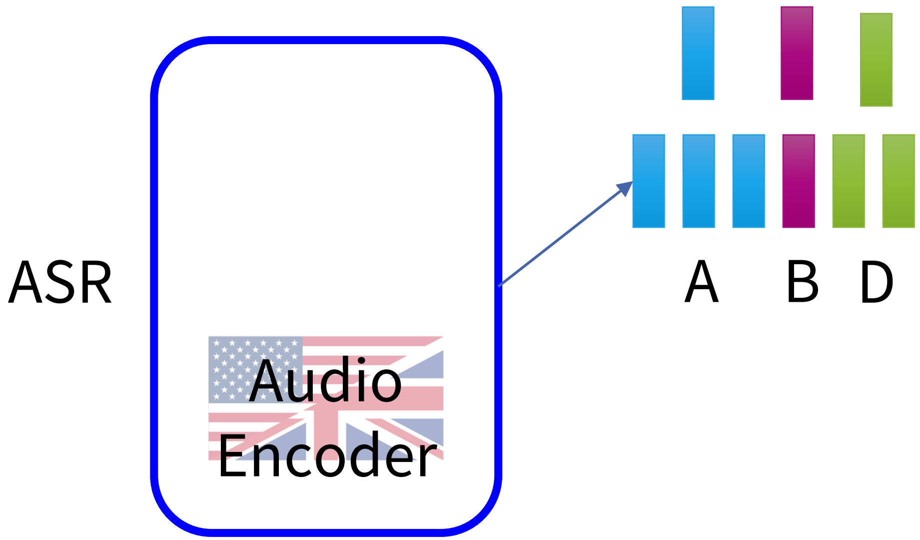


Integrating pre-trained models



Compression Layer

- CTC compression (Gaido et al, 2021)
  - Collapse adjacent representations with same index by averaging
  - Remove redundant and uninformative vectors



Challenges

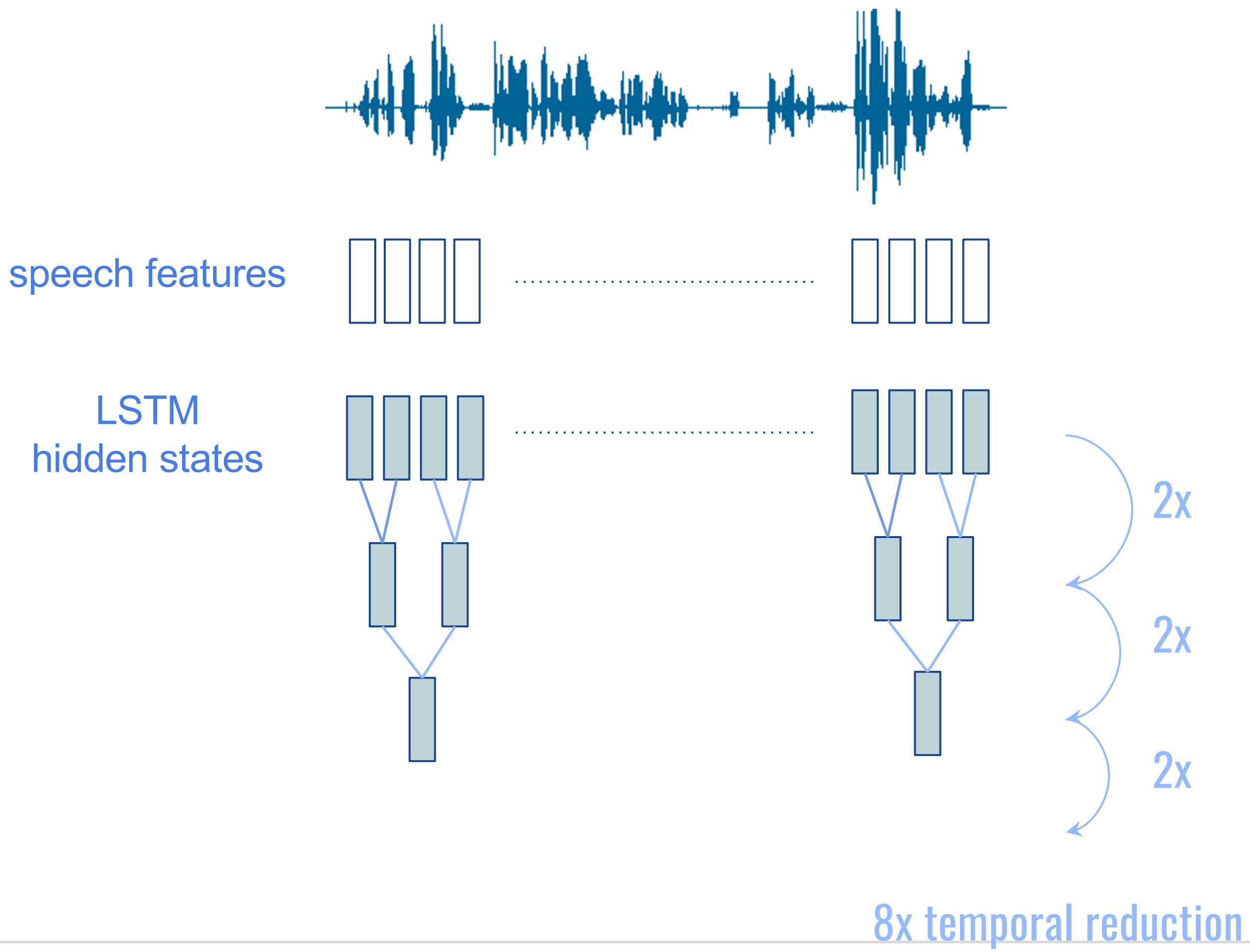
- Data
  - Other data sources
  - Pre-trained models
- Audio
  - Input length
  - High variability
  - Unsegmented
- Output
  - Audio
  - Low latency
  - Additional constraints

Challenges

- Data
  - Other data sources
  - Pre-trained models
- Audio
  - Input length
  - High variability
  - Unsegmented
- Output
  - Audio
  - Low latency
  - Additional constraints

- Sequence Length
  - IWSLT test set 2020
    - Segments: 1804
    - Words: 32.795
    - Characters: 149.053
    - Features: 1.471.035

Pyramidal Encoder



- Motivation: do not need attention to the granularity of speech features
- Reduce dimensionality *through* encoder

  - concatenation
  - sum
  - skip
  - linear projection

Linear projection, ASR:
(Zhang et al. 2017; Sperber et al. 2018)

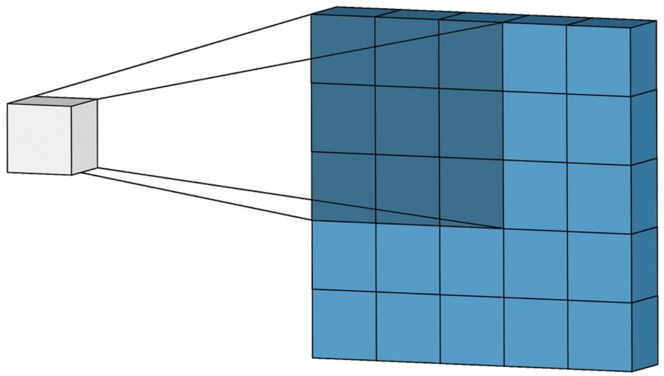
Pyramidal encoder in ST:
(Weiss et al. 2017; Salesky et al. 2019;
Sperber et al. 2019; Salesky et al. 2020)

Listen, Attend, and Spell
(Chan et al. 2015)

Dimensionality Reduction

Two directions: ① temporal and ② feature dimension

Convolutional layers enable *fixed-length downsampling*



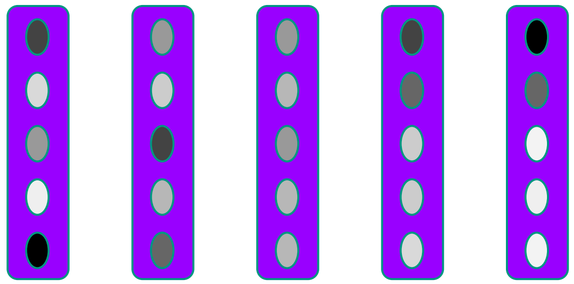
Scale sequence length and feature dimension linearly by a factor corresponding to the convolutional kernel size and stride length

Challenges

- Data
  - Other data sources
  - Pre-trained models
- Audio
  - Input length
  - High variability
  - Unsegmented
- Output
  - Audio
  - Low latency
  - Additional constraints

- Variation
  - Many different ways to speech same sentence
  - Limited training data
- Data augmentation
  - ASR investigated several possibilities
    - Noise injection (Hannun et al., 2014)
    - Speed perturbation (Ko et al., 2015)
  - Successful technique in deep learning ASR
    - SpecAugment (Spark et al., 2019)
    - Also applied in ST (Bahar et al, 2019)

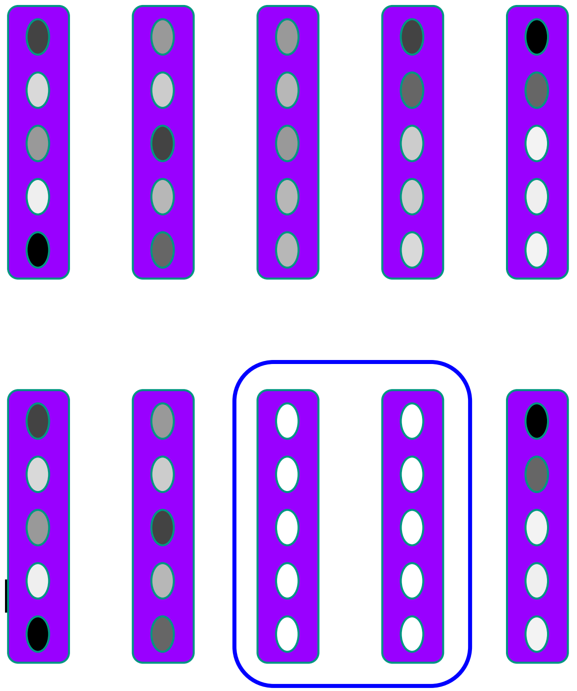
SpecAugment



- Directly applied on audio features
- Idea:
  - Mask information

SpecAugment

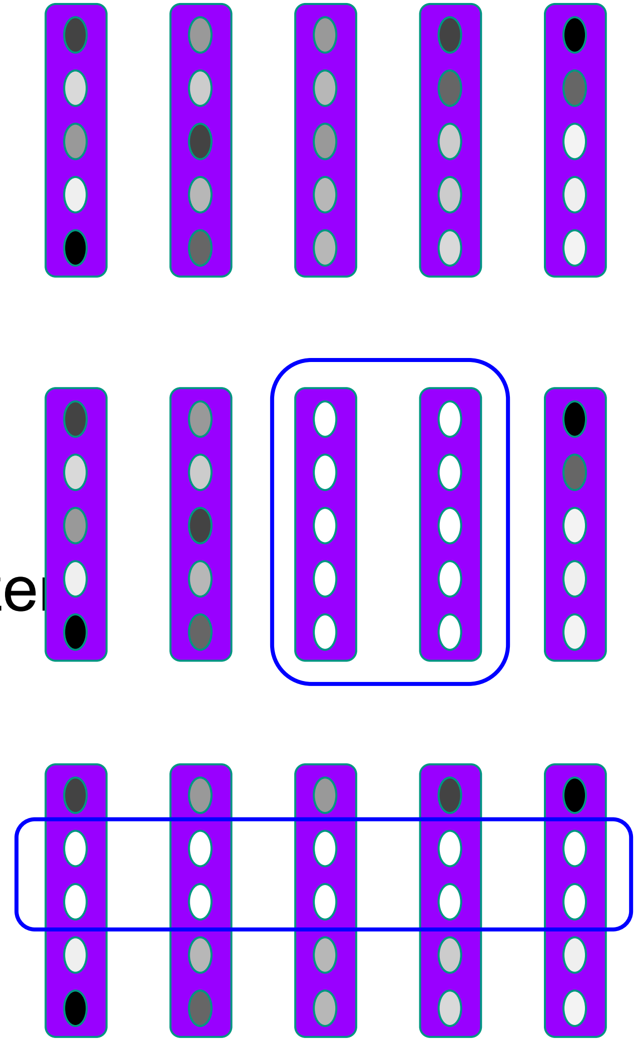
- Directly applied on audio features
- Idea:
  - Mask information



- Time masking
  - Set several consecutive feature vector to zer

SpecAugment

- Directly applied on audio features
- Idea:
  - Mask information



- Time masking
  - Set several consecutive feature vector to ze
- Frequency masking
  - Mask consecutive frequency channels

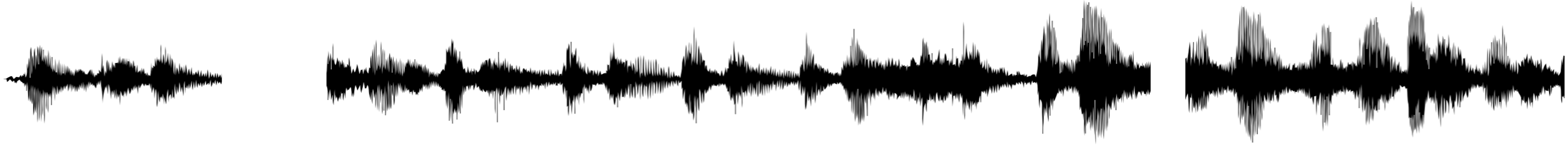
Challenges

- Data
  - Other data sources
  - Pre-trained models
- Audio
  - Input length
  - High variability
  - Unsegmented
- Output
  - Audio
  - Low latency
  - Additional constraints

- No segmentation in audio signal
- Segment audio
  - Using voice activity detection
  - Supervised classification

Utterance segmentation - Problem

- **Mismatch between training and evaluation data**
  - Training corpora: “sentence-level” split of continuous speech



This is an audio signal.

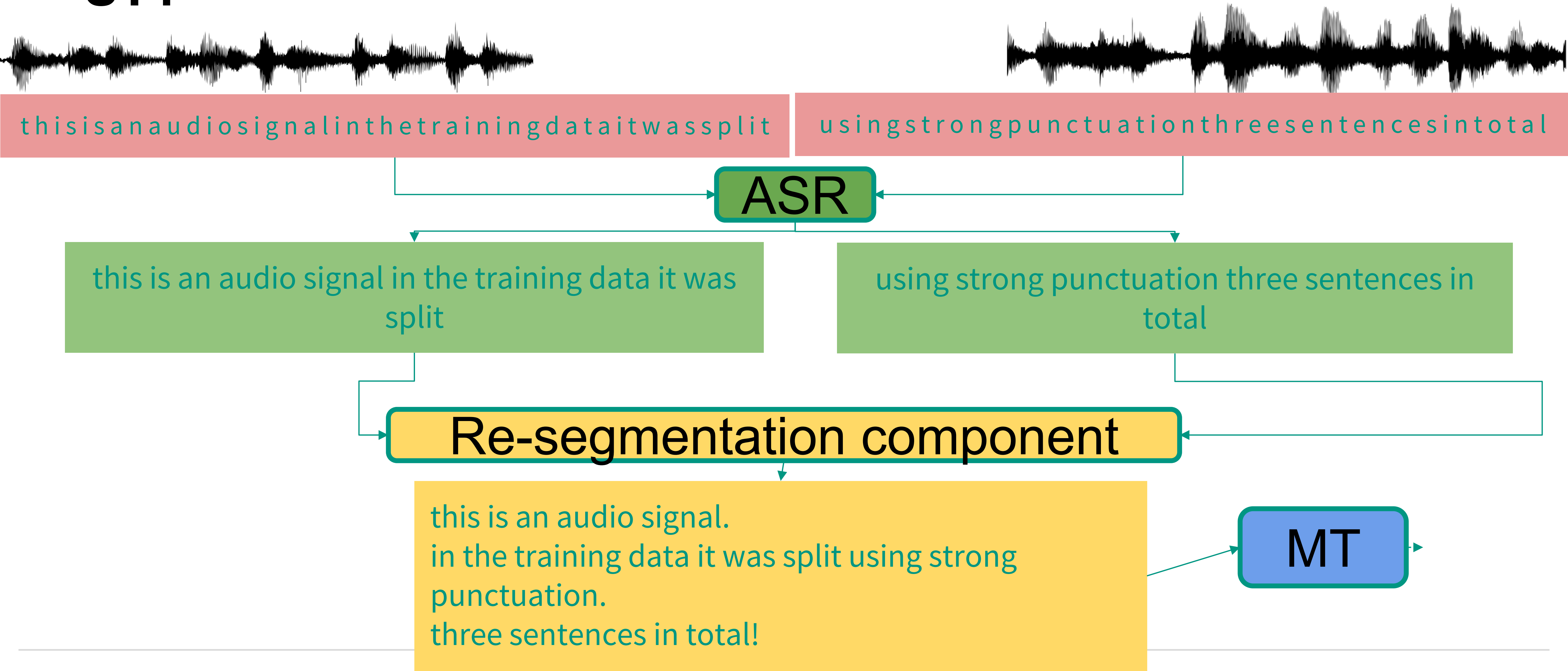
In the training data it was split using strong punctuation.

Three sentences in total!

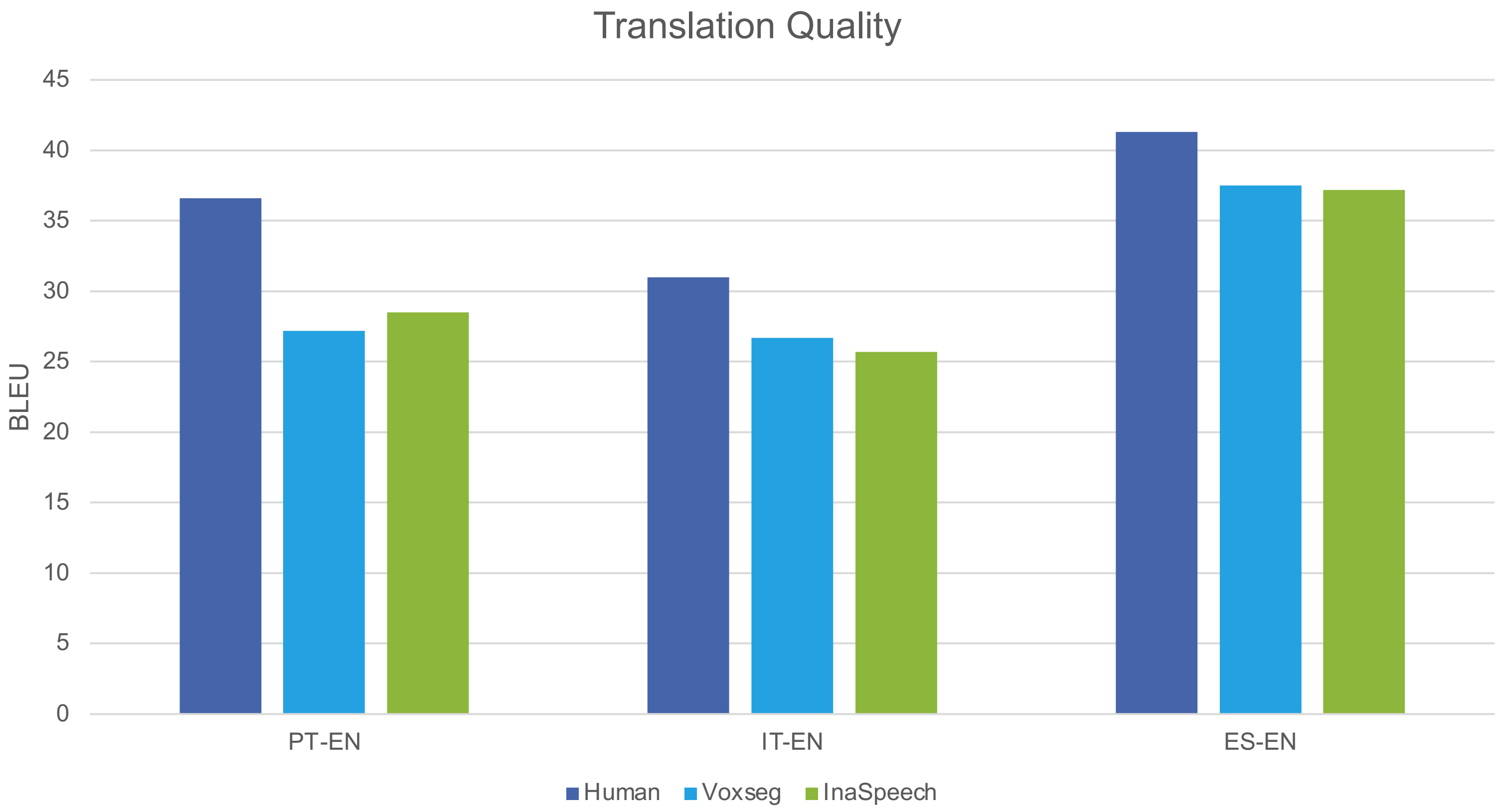
Utterance segmentation - Problem

- **Mismatch between training and evaluation data**
  - Training corpora: “sentence-level” split of continuous speech
  - At run-time: unsegmented continuous speech

How to split continuous speech in <u>cascade</u> ST?

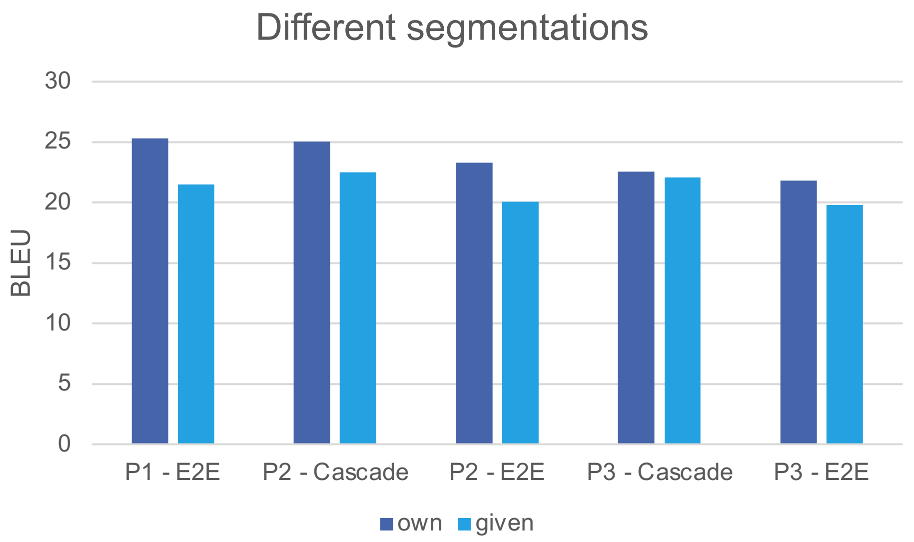


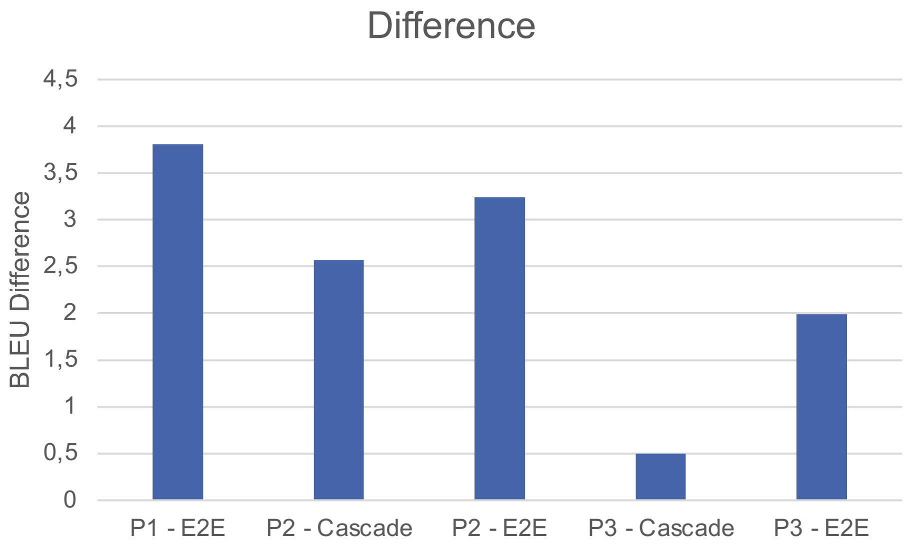
Unsegmented audio



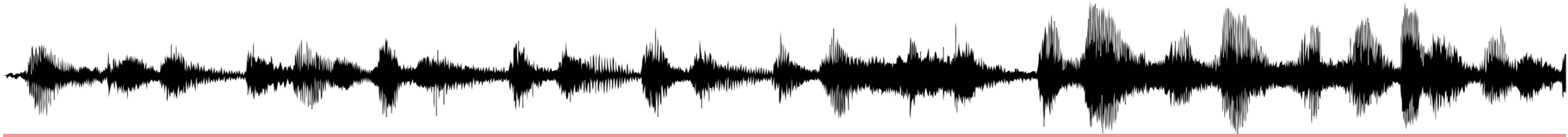
Unsegmented audio – IWSLT 2020

Karlsruhe Institute of Technology



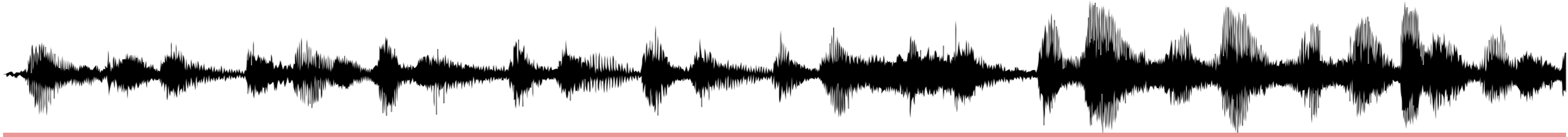


How to split continuous speech in <u>e2e</u> ST?

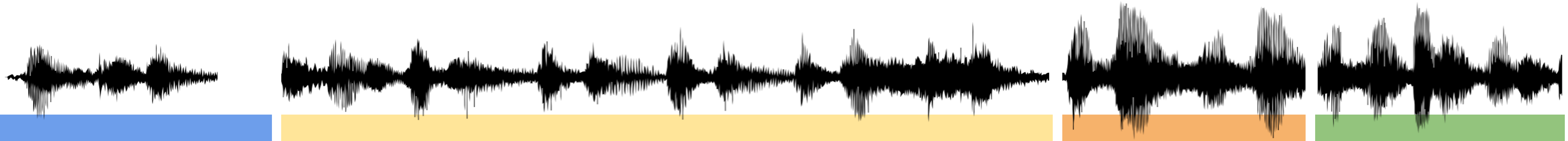


thisisanaudiosignalinthetrainingdataitwassplitusingstrongpunctuationthreesentencesintotal

Solution 1: Split on silences (via VAD)



t h i s i s a n a u d i o s i g n a l i n t h e t r a i n i n g d a t a i t w a s s p l i t u s i n g s t r o n g p u n c t u a t i o n t h r e e s e n t e n c e s i n t o t a l



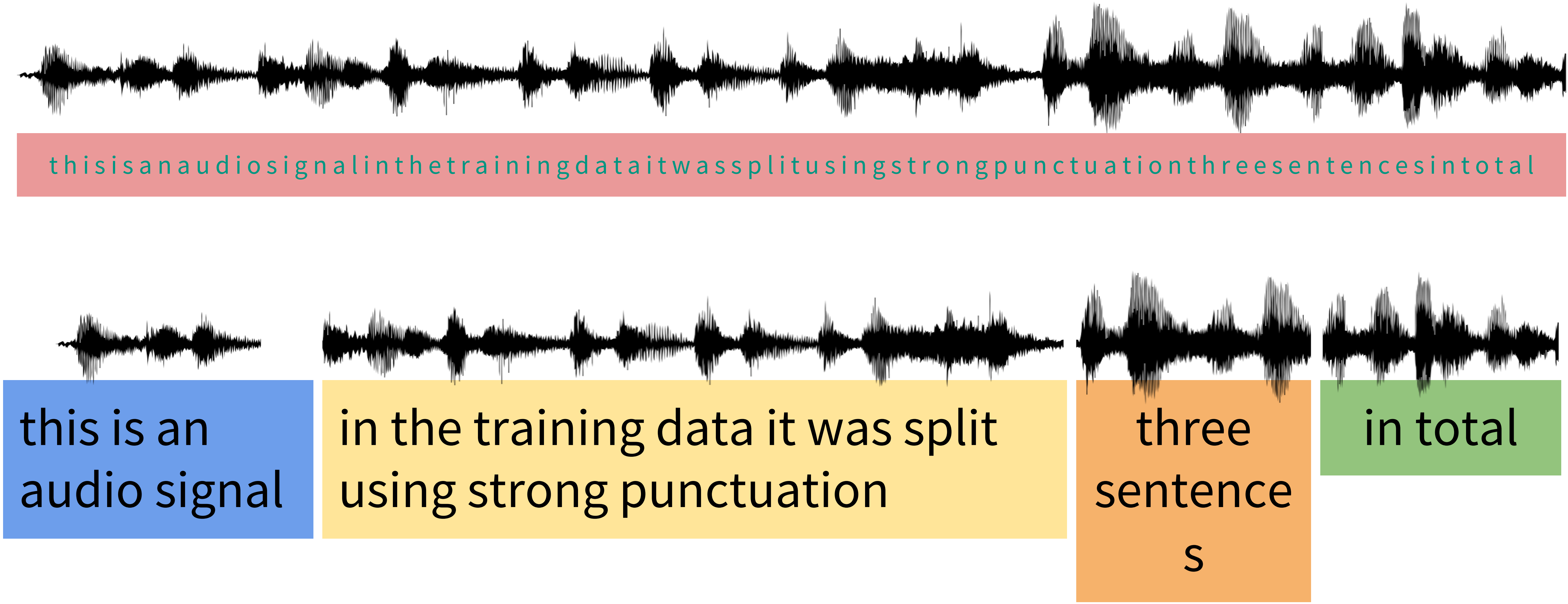
this is an audio signal

in the training data it was split using strong punctuation

three sentence s

in total

Solution 1: Supervised classification (SHAS)

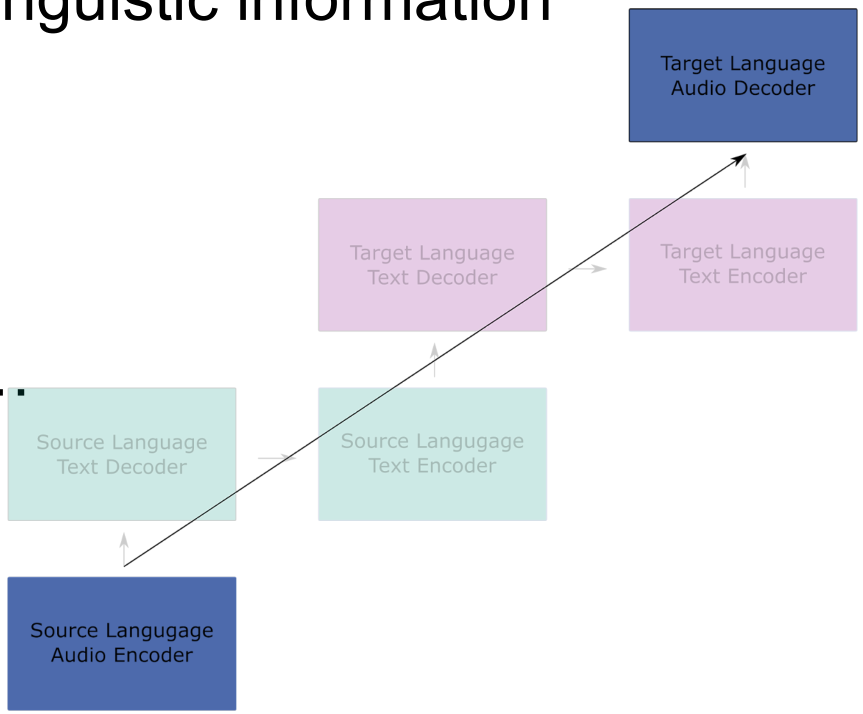


Challenges

- Data
  - Other data sources
  - Pre-trained models
- Audio
  - Input length
  - High variability
  - Unsegmented
- Output
  - Audio
  - Low latency
  - Additional constraints

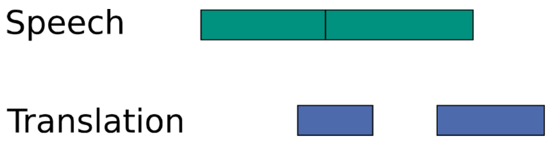
Speech output

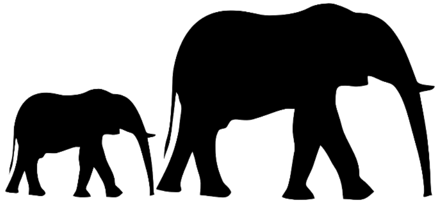
- Jointly train ASR, MT and TTS
- Opportunities:
  - Retaining paralinguistic and non-linguistic information
    - Maintain source speaker voice
    - Emotion
    - Prosody
  - Fluent pronunciations of names, …
- First approach:
  - Jia et al, 2019



Low latency

- Partial information
  - Online: Translate during production of speech
  - Generate translation before full sentence is known





Challenges – Simultaneous Translation

- Generate translation while speaker speaks
- Tradeoff:
  - More context improves speech recognition and machine translation
    - Wait as long as possible
  - Low latency is important for user experience
    - Generate translation as early as possible
- Challenge:
  - Different word order in the languages
    - SOV vs SVO

German	Ich	melde	mich	zur	Summer	School	an
Gloss	I	register/ cancel	myself	to	summer	School	
English	I	????					

Simultaneous Translation

- Approaches:
  - Learn optimal segmentation strategies
  - Re-translate
    - Update previous translation with better once
  - Stream decoding
    - Dynamically learn when to generate a translation

Re-translate

- Directly output first hypothesis
- If more context is available:
  - Update with better hypothesis

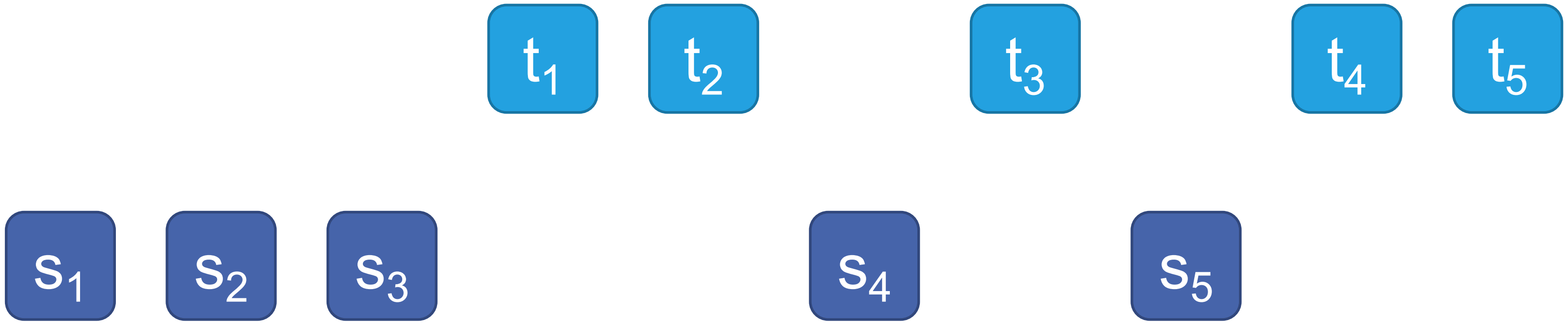
- Example:
  - Ich melde mich
  - I register

  - Ich melde mich von der Klausur ab
  - I withdraw from the exam

Niehues et al, 2016

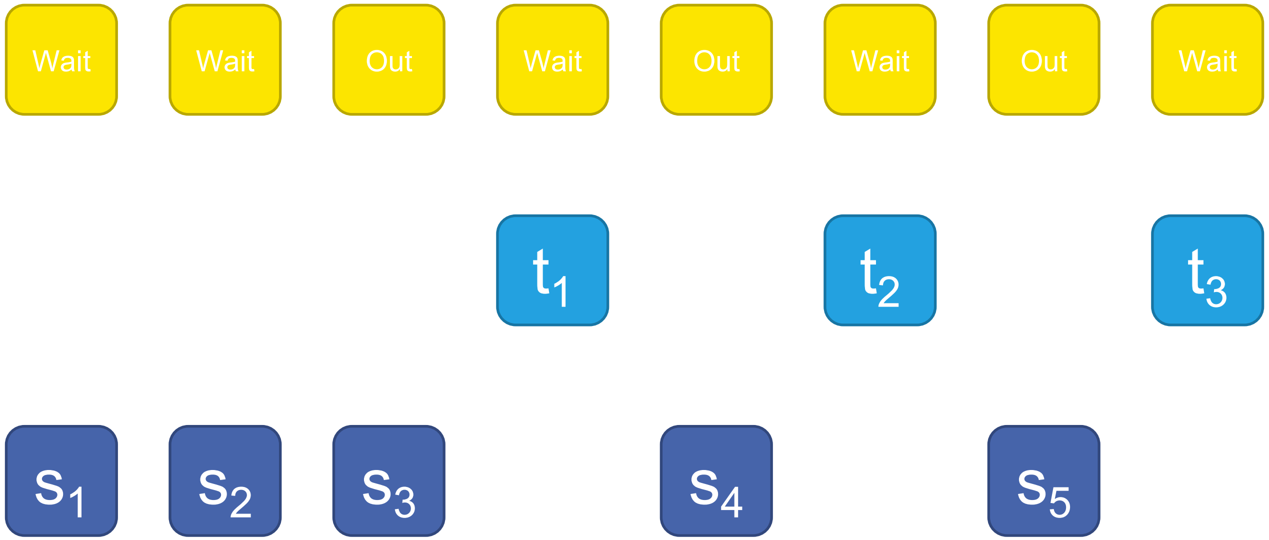
Stream decoding

- Idea:
  - At each time step:
    - Decided to output word
    - Wait for additional input
  - (Kolss et at., 2008)



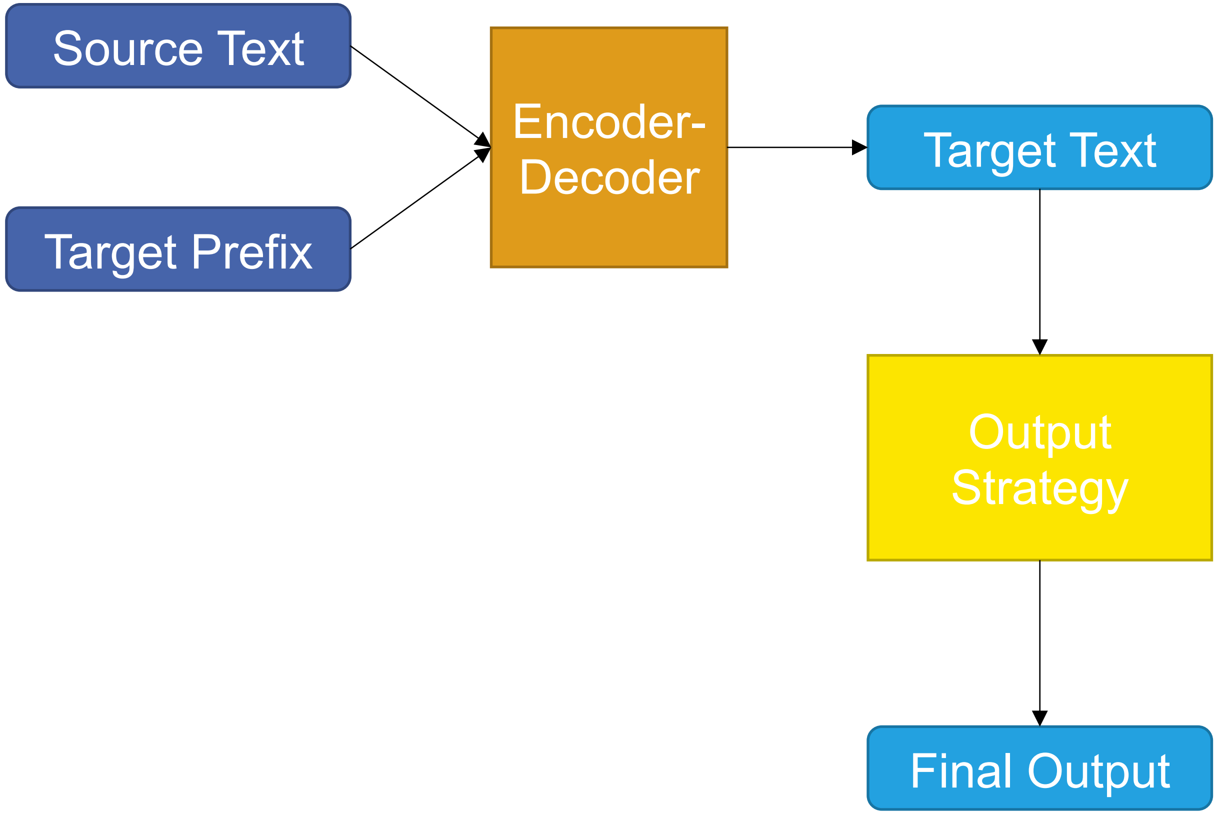
Stream decoding - Decoder

- Methods:
  - Dynamic decision Cho et al, 2016; Gu et al, 2017; Dalvi et al, 2018
  - Fixed schedule (Ma et al, 2019)
    - Wait-k policy



Relation to re-translate

- Decoding with fixed target prefix



Stream decoding strategies

- Local agreement [Liu et al, 2020]
  - Output if previous and current output agree on prefix
  - Variation [Yao et al., 2020]:
    - Predict the next source word instead of relying on the previous input

Input	Prefix	Target Text	Final Output
1	Ø	All model trains	Ø
1,2	Ø	All models art	All
1,2,3	All	All models are wrong	All models
1,2,3,4	All models		
…			

What is special about Subtitling?

- Importance of time
- Text needs to satisfy spatial and temporal constraints

In and out times based on speech rhythm

Length:
max. 2 lines (of ≈ length)
max. 42 characters/line

Reading speed:
max. 21 characters/second

Segmenting into proper subtitles

This kind of harassment keeps women <eob> from accessing the internet – <eol> essentially, knowledge. <eob>

```
10
00:00:31,066 --> 00:00:34,390
This kind of harassment keeps women
11
00:00:34,414 --> 00:00:36,191
from accessing the internet --
essentially, knowledge.
```

Evaluation campaign

- International Conference of Spoken Language Translation (IWSLT)
  - Largest evaluation campaign on Spoken Language translation
    - 4 tracks
    - 22 teams

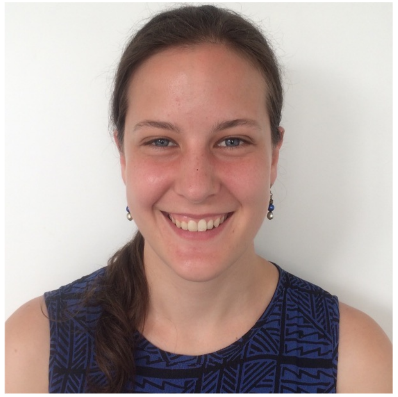
- Next event:
  - IWSLT 2023 collocated with ACL (Toronto)

Anastasopoulos et al, 2022

https://www.iwslt.org/

Tutorial EACL 2021: End-to-End Speech translation

Jan Niehues,
Maastricht University
jan.niehues@maastricht university.nl



Elizabeth Salesky,
Johns Hopkins University
esalesky@jhu.edu

Marco Turchi,
Fondazione Bruno Kessler
turchi@fbk.eu

Matteo Negri,
Fondazione Bruno Kessler
negri@fbk.eu

https://st-tutorial.github.io/

SIG-SLT Talk Series

- Month virtual presentation by international research
  - Speech Translation
- Join Google groupe for more information
  - https://iwslt.org/sigslt/

End-to-End Speech to Speech Translation

- Questions

- Contact:
  - jan.niehues@kit.edu
  - https://ai4lt.anthropomatik.kit.edu/