

# End-to-End Speech Translation

Jan Niehues, Elizabeth Salesky, Marco Turchi and Matteo Negri

*EACL 2021*

# Speakers



Jan Niehues,  
*Maastricht University*  
[jan.niehues@maastrichtuniversity.nl](mailto:jan.niehues@maastrichtuniversity.nl)



Elizabeth Salesky,  
*Johns Hopkins University*  
[esalesky@jhu.edu](mailto:esalesky@jhu.edu)



Marco Turchi,  
*Fondazione Bruno Kessler*  
[turchi@fbk.eu](mailto:turchi@fbk.eu)



Matteo Negri,  
*Fondazione Bruno Kessler*  
[negri@fbk.eu](mailto:negri@fbk.eu)

# Outline

<i>Sec 1: Introduction</i>	1.1: Task definition 1.2: Challenges in translation of speech 1.3: Traditional cascade approach
<i>Sec 2: End-to-End</i>	2.1: State-of-the-art 2.2: Input representations 2.3: Architecture & modifications 2.4: Output representations
<i>Sec 3: Leveraging Data Sources</i>	3.1: Available data 3.2: Techniques: Multi-task learning Transfer-learning & pretraining Knowledge distillation 3.3: Alternate data representations
<i>Sec 4: Evaluation</i>	4.1: Automatic Metrics 4.2: Utterance segmentation 4.3: Mitigating error – gender bias
<i>Sec 5: Advanced Topics</i>	5.1: Utterance segmentation 5.2: Multilingual ST 5.3: Under-resourced languages
<i>Sec 6: Real-world</i>	6.1: Automatic generation of subtitles 6.2: Simultaneous translation
<i>Sec 7: Conclusion</i>	🏆

*Sec 1:*

# Introduction

**Task definition**

**Challenges in translation of speech**

**Traditional cascade approaches**

*Sec 1.1*

# Task Definition

# Speech Translation - Task

Speech input



= *Welcome to this tutorial*

ST system

Textual translation



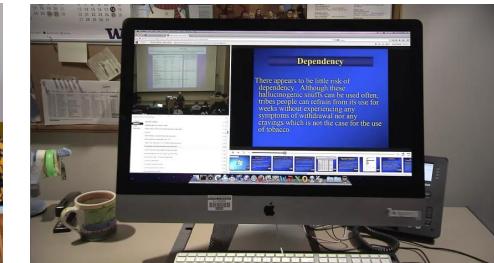
*Willkommen zu  
diesem Tutorial*

Spoken translation



# Speech Translation - Motivation

- Break language barriers to communicate, spread information and culture
  - Work
    - Meetings
  - Education and training
    - Lectures, conferences
  - Entertainment
    - Youtube, social media, cinema, tv
  - Everyday communication
    - Tourism, medical care, telephone conversations



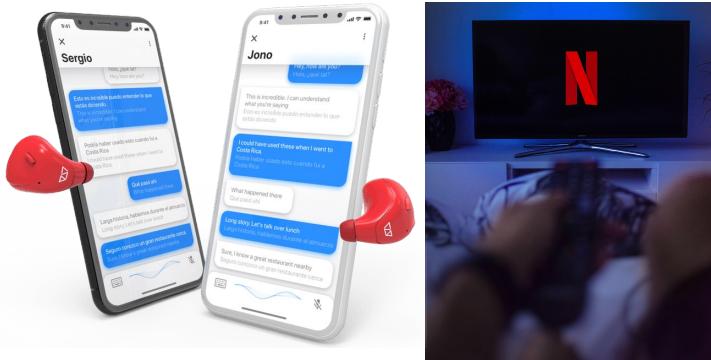
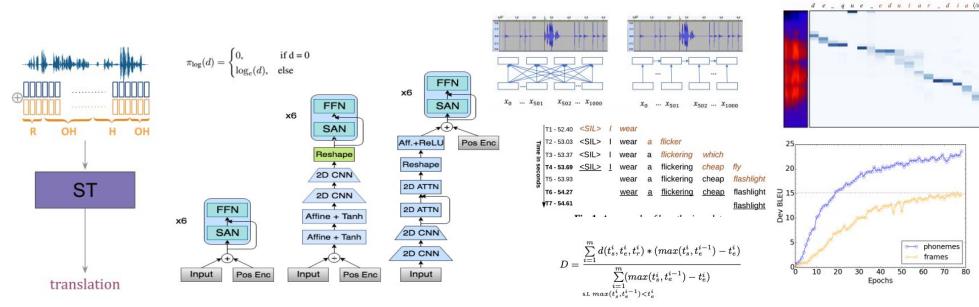
# Speech Translation - Motivation

- Room for advanced research...

- 99% of this tutorial

- ...and for applications

- Wearable devices
  - Video subtitling
  - Live captioning
  - Human-machine communication



# Speech Translation - History (before e2e)

## Late '80s: first proofs of concept

Constraints to control language ambiguity (phonetics, syntax, semantics)

- Restricted vocabulary
- Controlled speaking style
- Narrow domain
- Offline processing



## '90s: Less constraints (vocabulary, speaking style)

First spontaneous ST systems (C-STAR, Verbmobil, Nespole,...)

## 2003-2006: Less constraints (domain)

First open-domain ST systems (STR-DUST, TC-STAR, GALE)

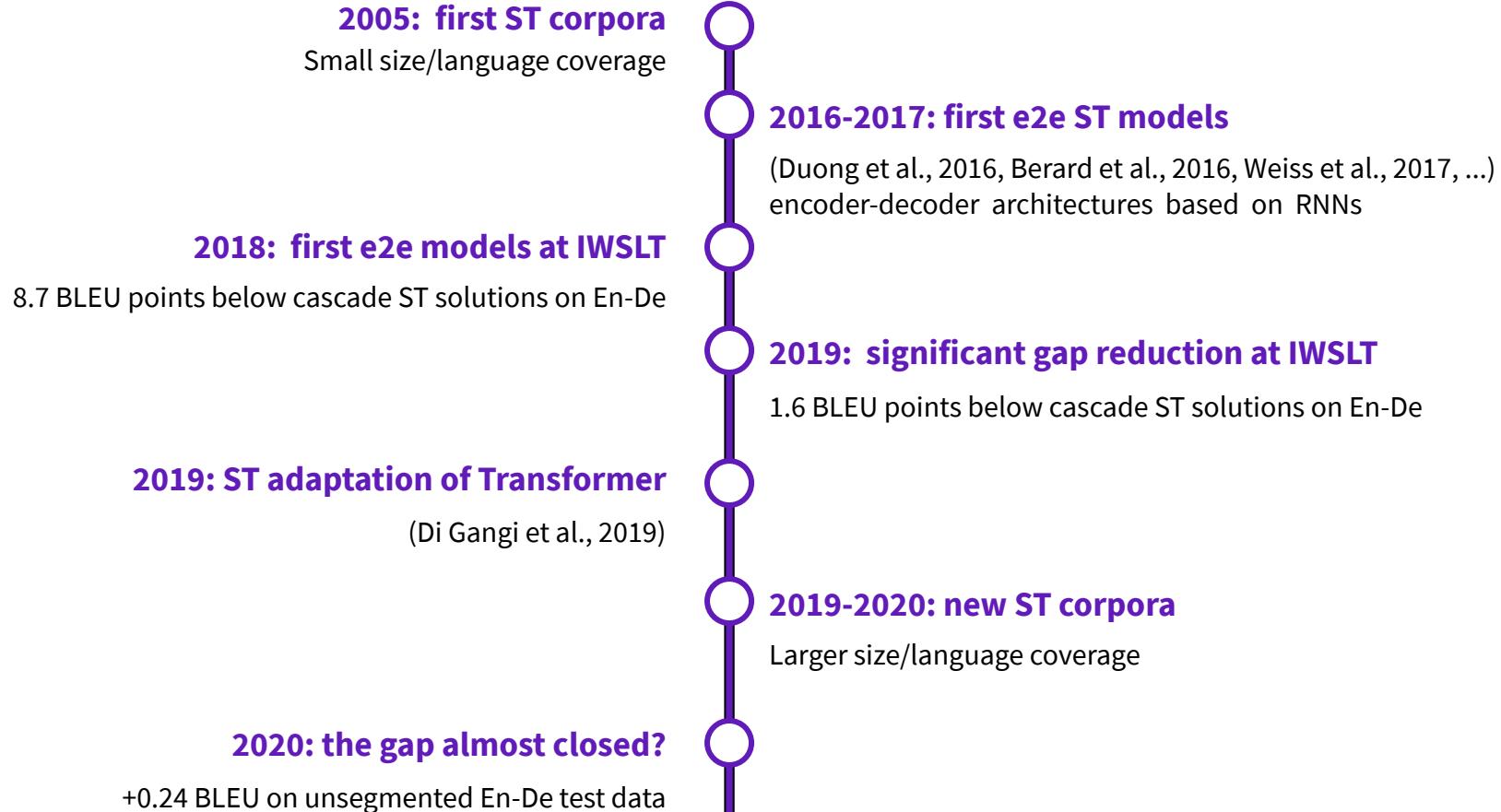
- different scenarios (broadcast news, parliamentary speeches, academic lectures)
- different languages (Zh, Ar, Es)



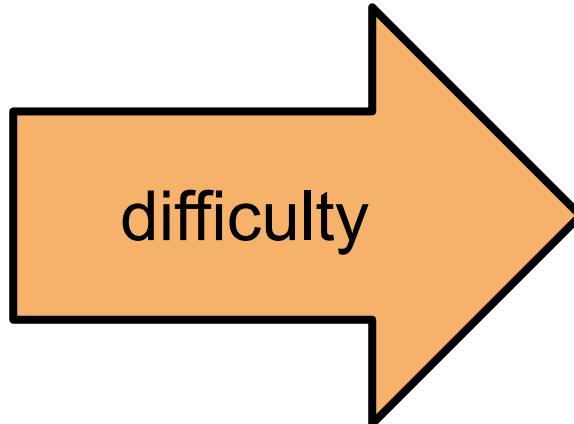
## 2006: Less constraints (operating conditions)

First simultaneous translator  
(real-time translation of spontaneous lectures and presentations)

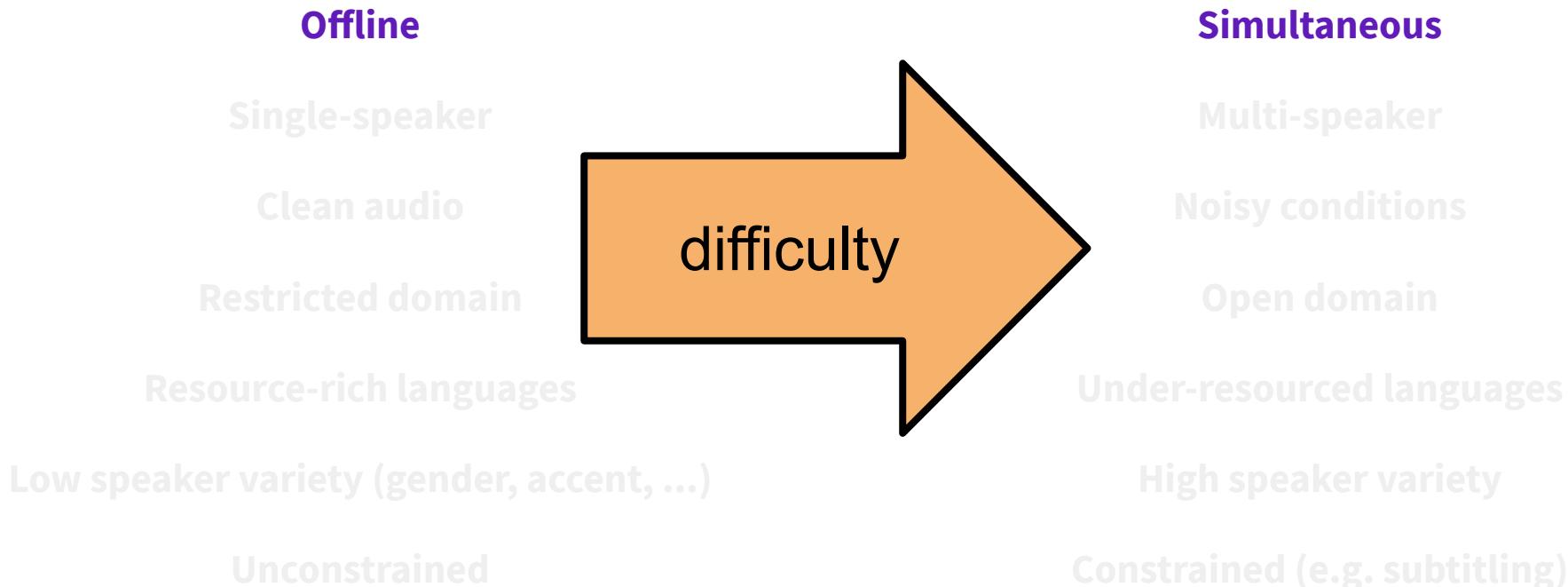
# Speech Translation - History (the e2e era)



# Speech Translation - a Multi-faceted Problem



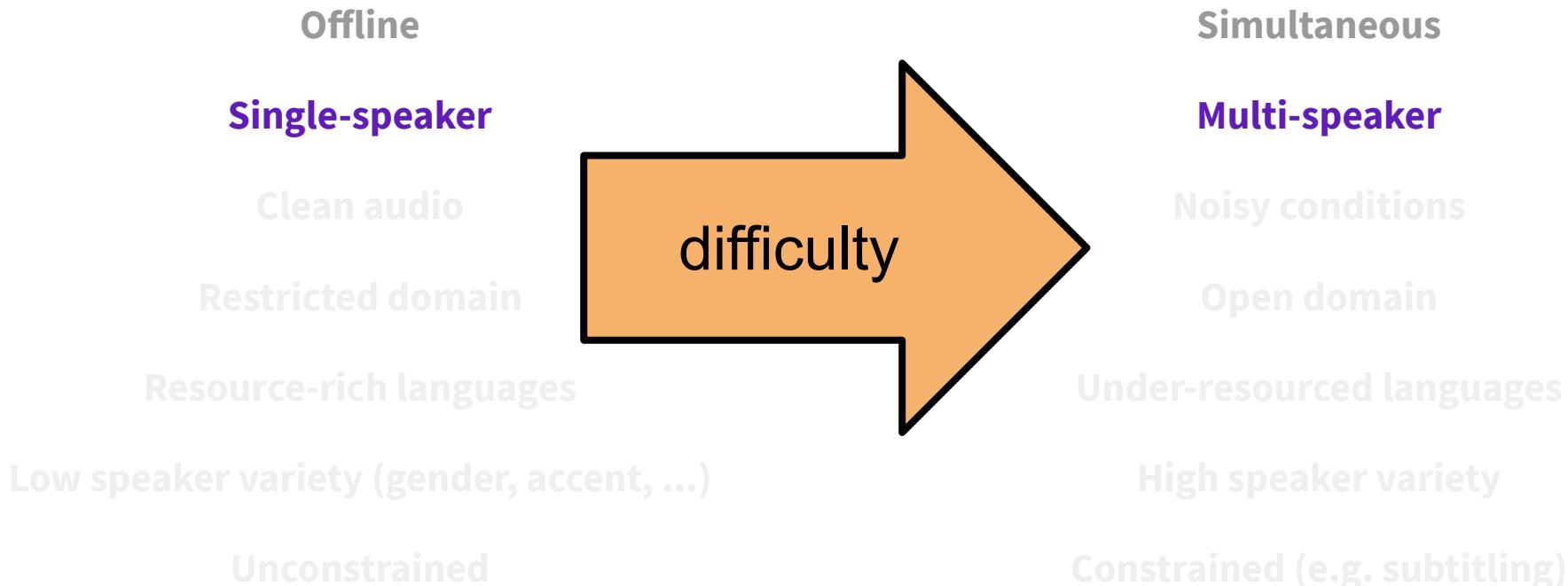
# Speech Translation - a Multi-faceted Problem



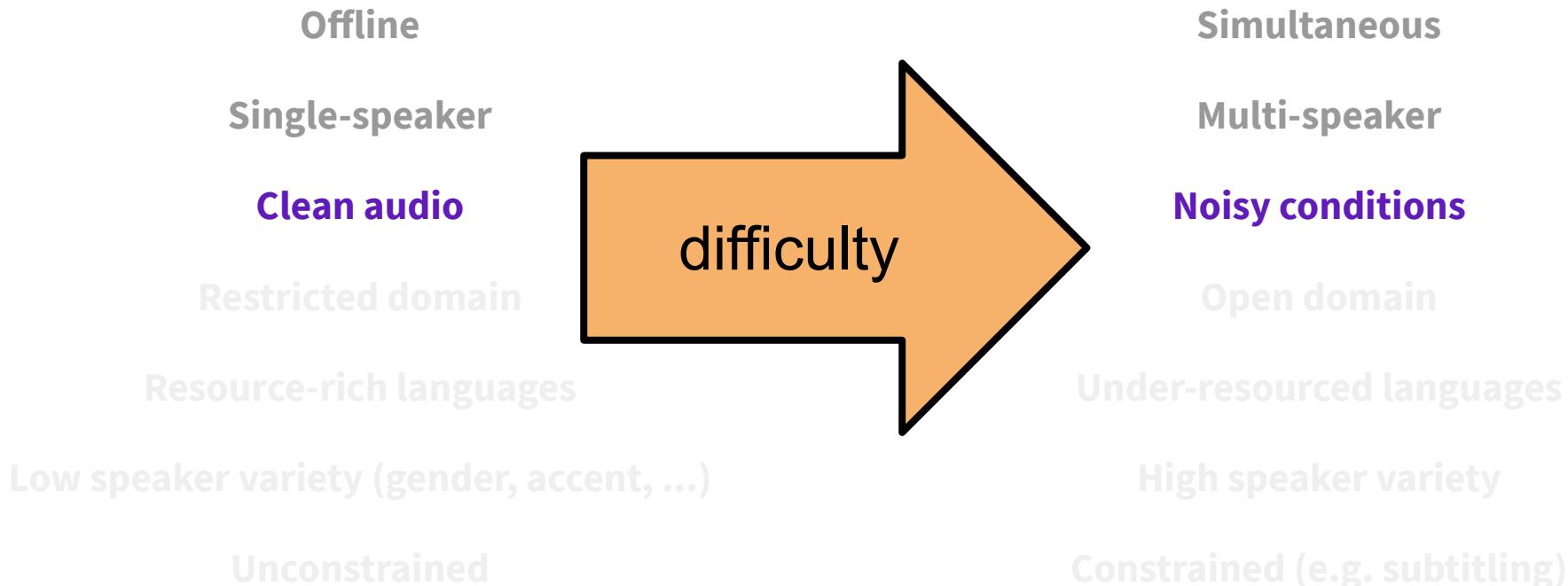
...

...

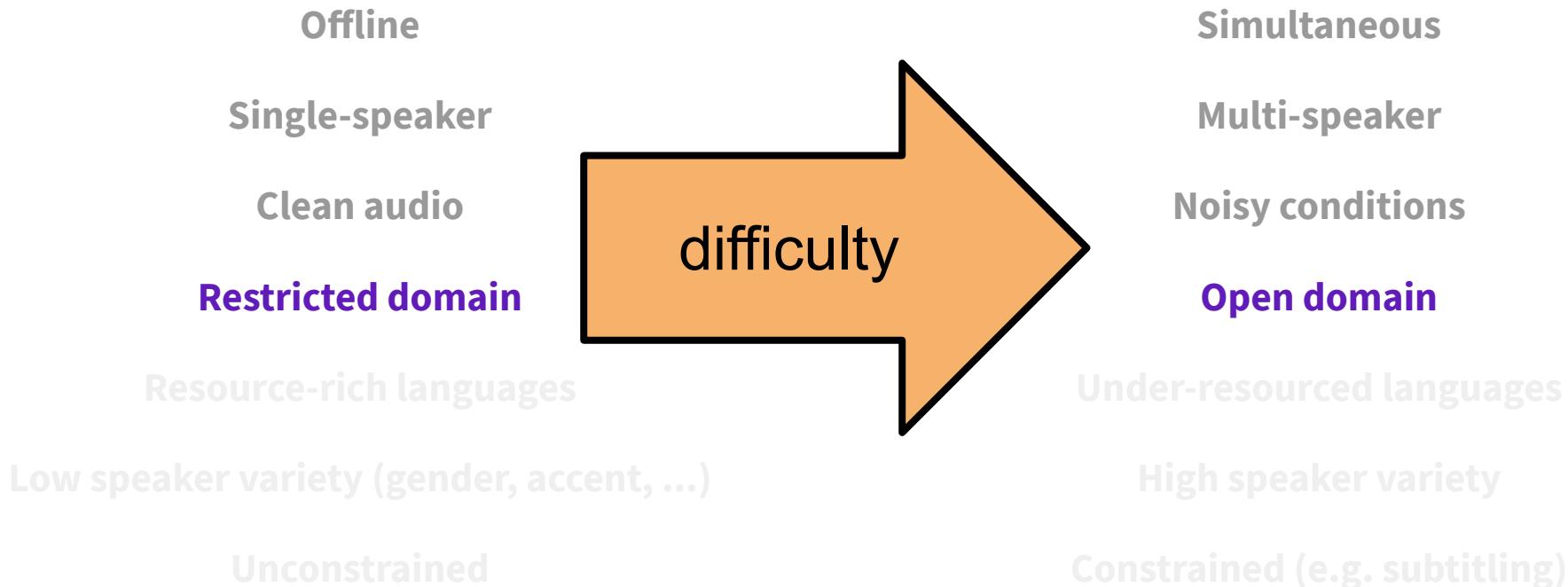
# Speech Translation - a Multi-faceted Problem



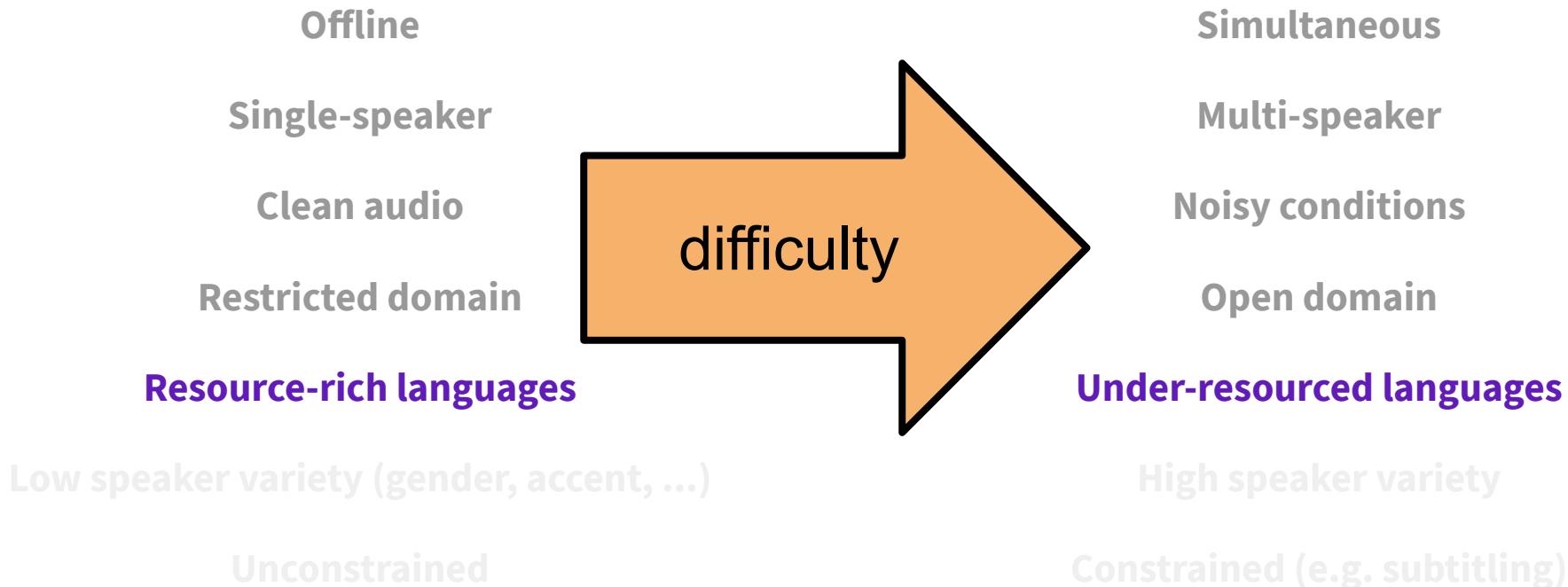
# Speech Translation - a Multi-faceted Problem



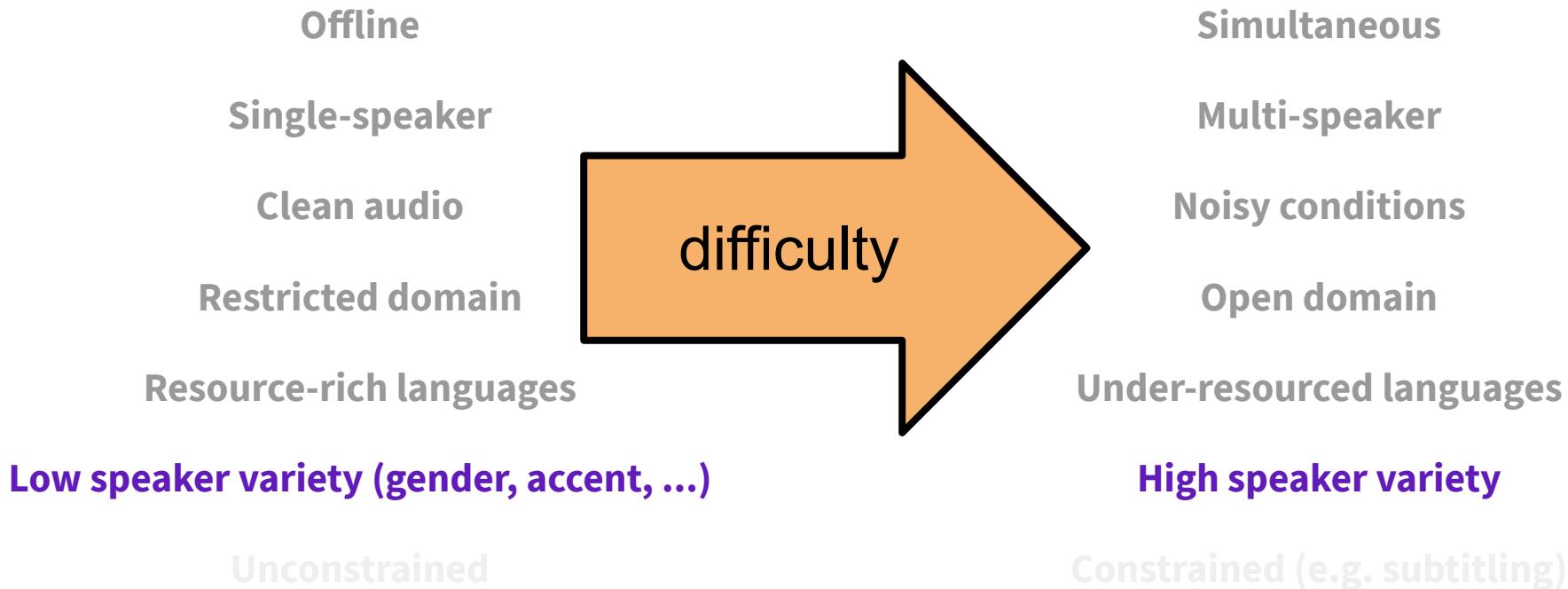
# Speech Translation - a Multi-faceted Problem



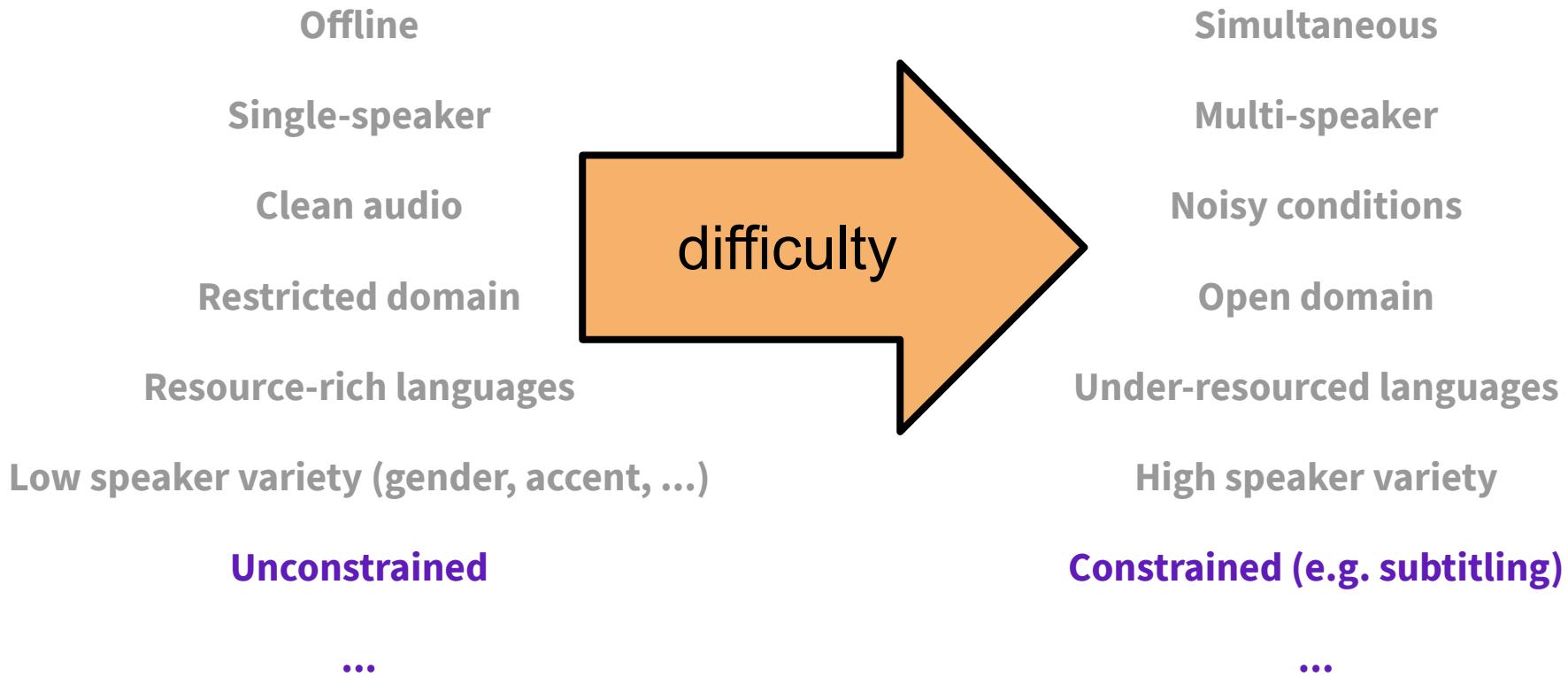
# Speech Translation - a Multi-faceted Problem



# Speech Translation - a Multi-faceted Problem



# Speech Translation - a Multi-faceted Problem



*Sec 1.2*

# Challenges in Translation of Speech

# Challenges in translation of speech

- Audio challenges
  - Multiple speaker
    - e.g. Meetings
    - Challenges:
      - Overlapping voice
  - Background noise
  - Audio segmentation



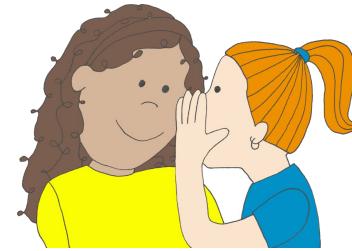
# Challenges in translation of speech

- Audio challenges
- Text-Speech mismatch
  - Disfluencies
    - Hesitations: “uh”, “uhm”, “hmm”,
    - Discourse markers: “you know”, “I mean”,...
    - Repetitions: “It had, it had been a good day”
    - Corrections: “no, it cannot, I cannot go there”
  - No punctuation
    - Let's eat Grandpa !
    - Let's eat, Grandpa !



# Challenges in translation of speech

- Audio challenges
- Text-Speech mismatch
- Error propagation
  - ASR errors worse after translation
    - More difficult to compensate by human
    - MT adds additional errors



Reden (engl. speeches)



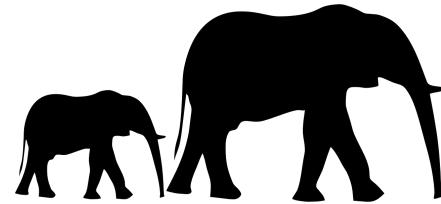
Reben (engl. vines)

# Challenges in translation of speech

- Audio challenges
- Text-Speech mismatch
- Error propagation
- Data
  - End-to-End data:
    - Growing amount but still limited
    - Integration of other data types
      - Speech transcripts
      - Parallel data

# Challenges in translation of speech

- Audio challenges
- Text-Speech mismatch
- Error propagation
- Data
- Partial information
  - Online: Translate during production of speech
  - Generate translation before full sentence is known



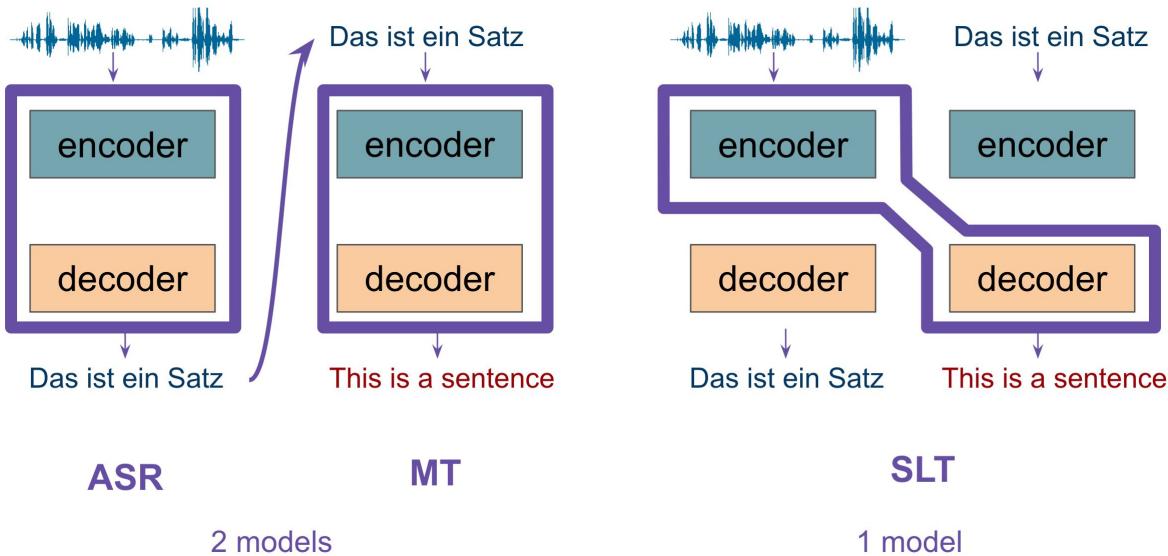
Speech  A horizontal green bar divided into two equal segments by a vertical line.

Translation  A horizontal blue bar divided into two equal segments by a vertical line.

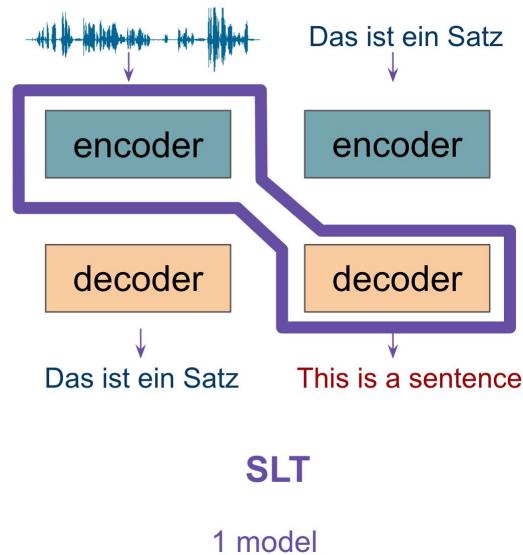
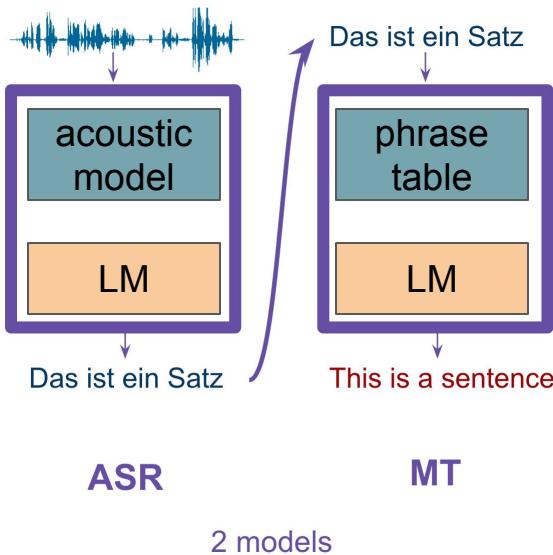
*Sec 1.3*

# Traditional cascade approach

# Traditional cascade approach



# Traditional cascade approach



*Modular, pipeline approach*  
*ASR, MT: isolated objectives*

(Waibel et al. 1991; Vidal, 1997; Ney, 1999; Saleem et al. 2004;  
Matusov et al. 2005; Bertoldi and Federico, 2005; Quan et al. 2005;  
Kumar et al. 2014; IWSLT Eval Campaigns 2004—)

# Data Used

- Datasets with parallel speech + translations arose with E2E models
- Traditionally, cascades used separate datasets for their component models
- **IWSLT Evaluation Campaigns (2004-present)**: ASR, MT, ST tasks

$\oplus$  many more data sources

$\ominus$  data is from different domains

# Modular Models

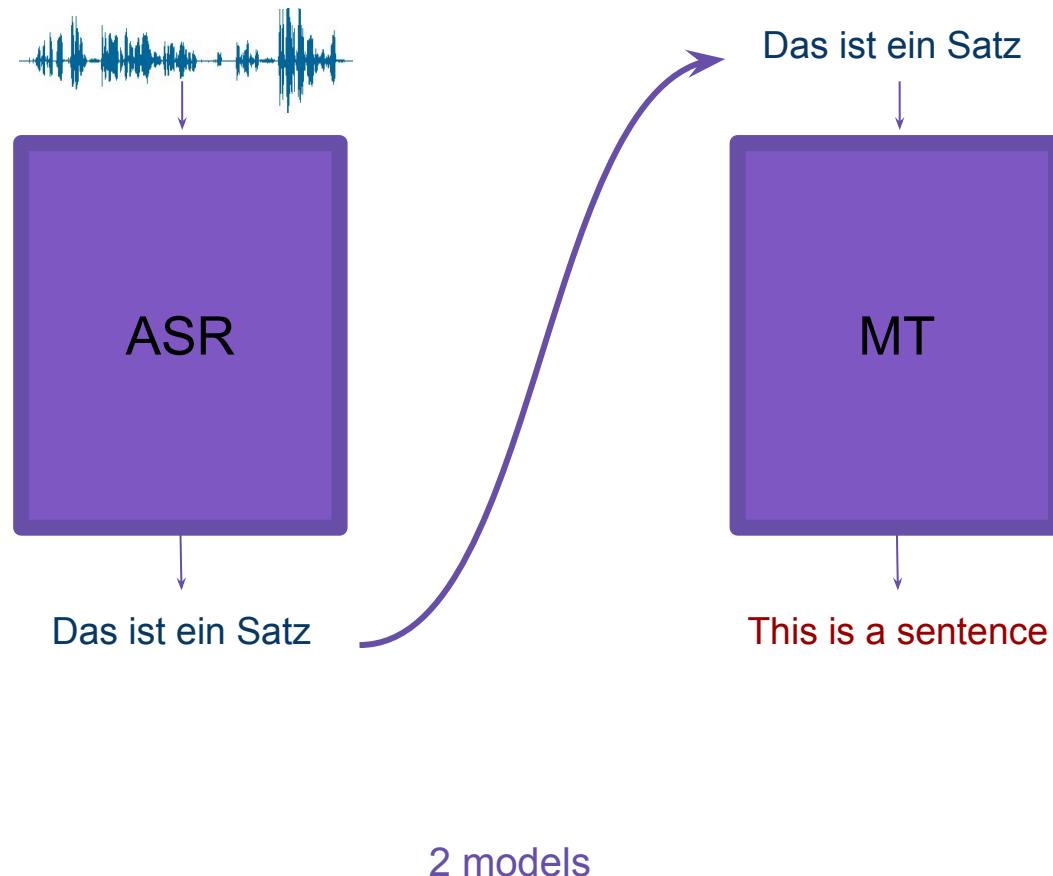
***Domain challenge:*** mismatch between ASR output and MT input

**ASR output:**

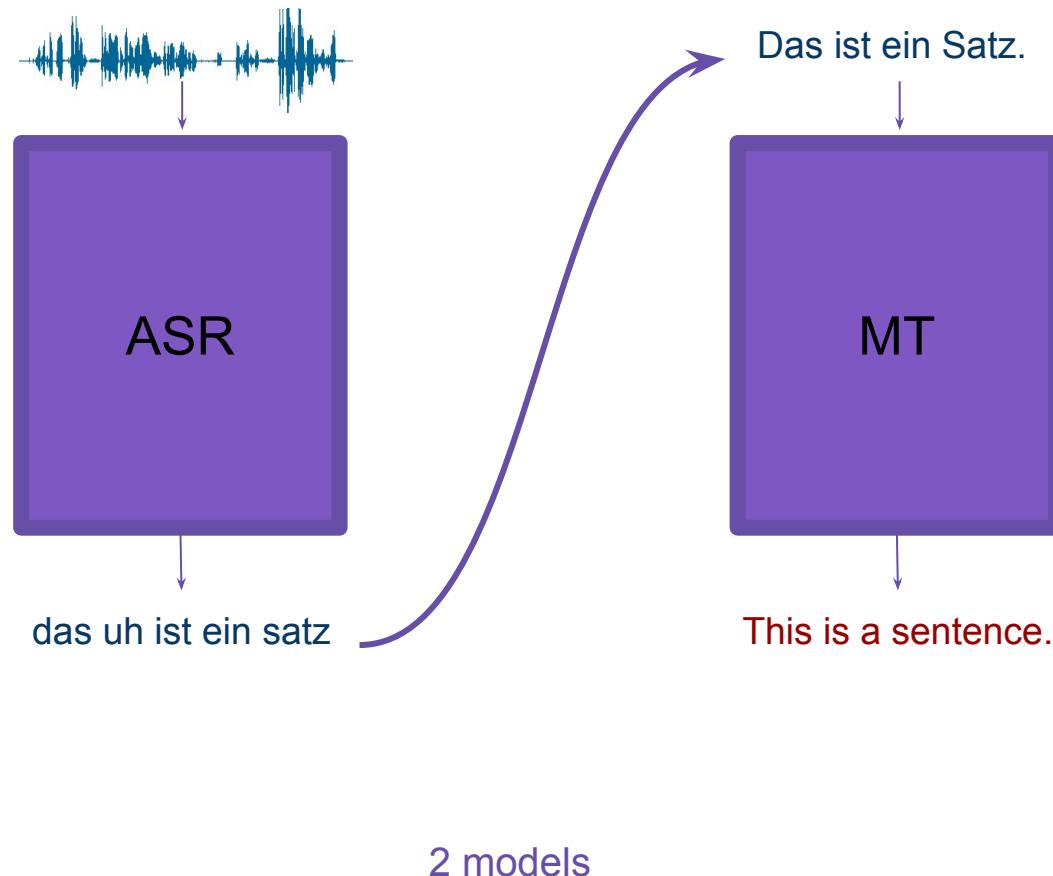
- lowercase, punctuation removed
- disfluencies (um, uh, ..., repetitions, false starts)
- ASR errors

→ *Differing training data domains, train-test mismatch:  
requires adaptation!*

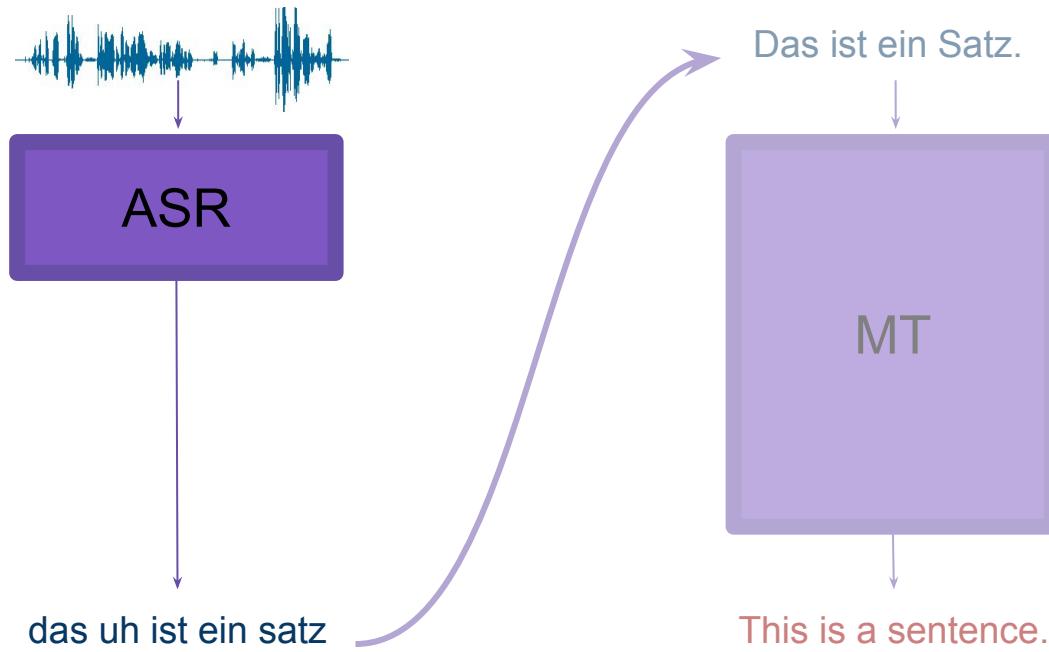
# Modular Models



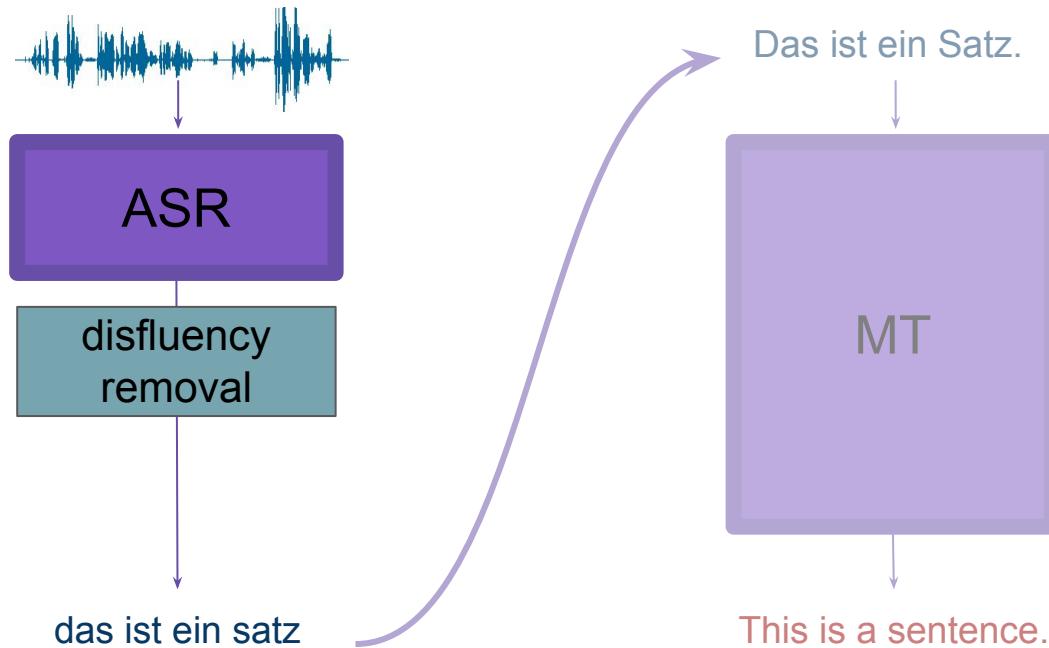
# Modular Models



# Modular Models

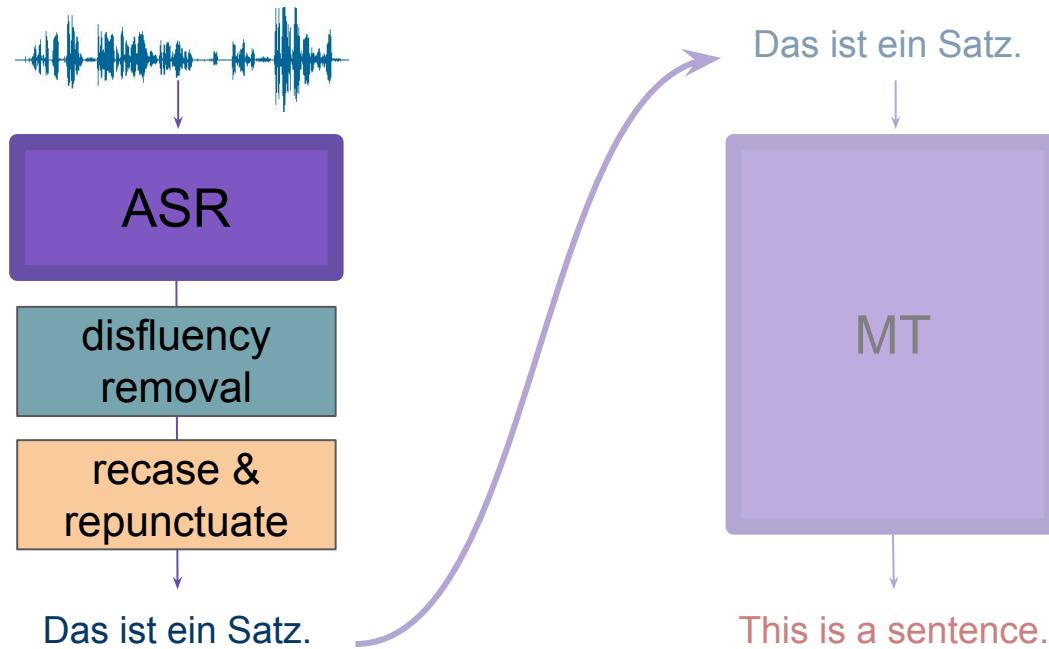


# Modular Models



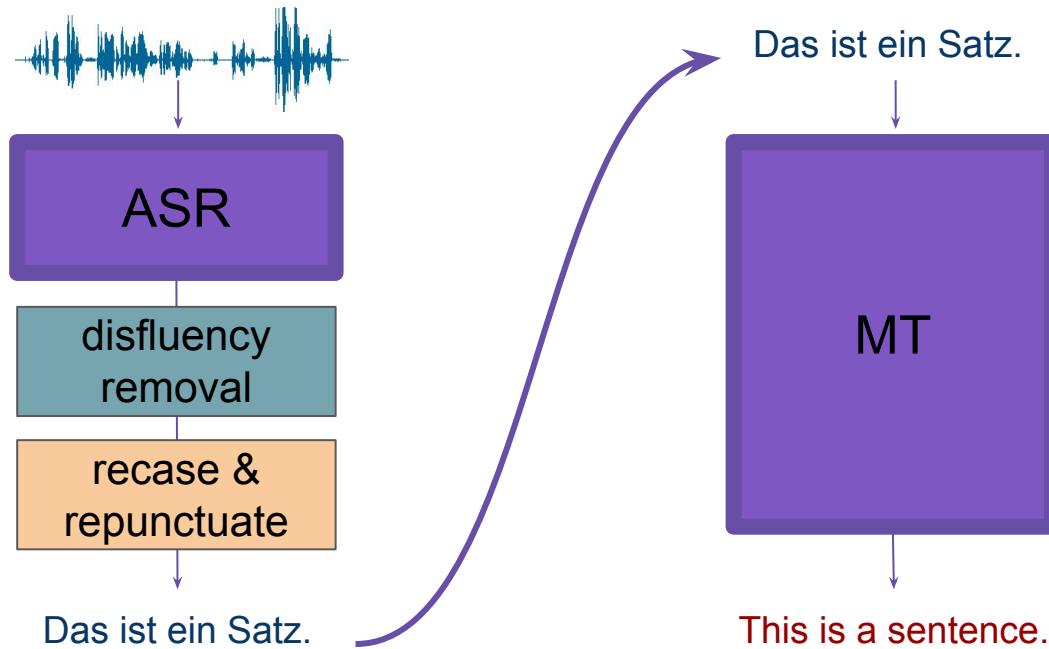
(Wang et al. 2010; Cho et al. 2013/2014)

# Modular Models

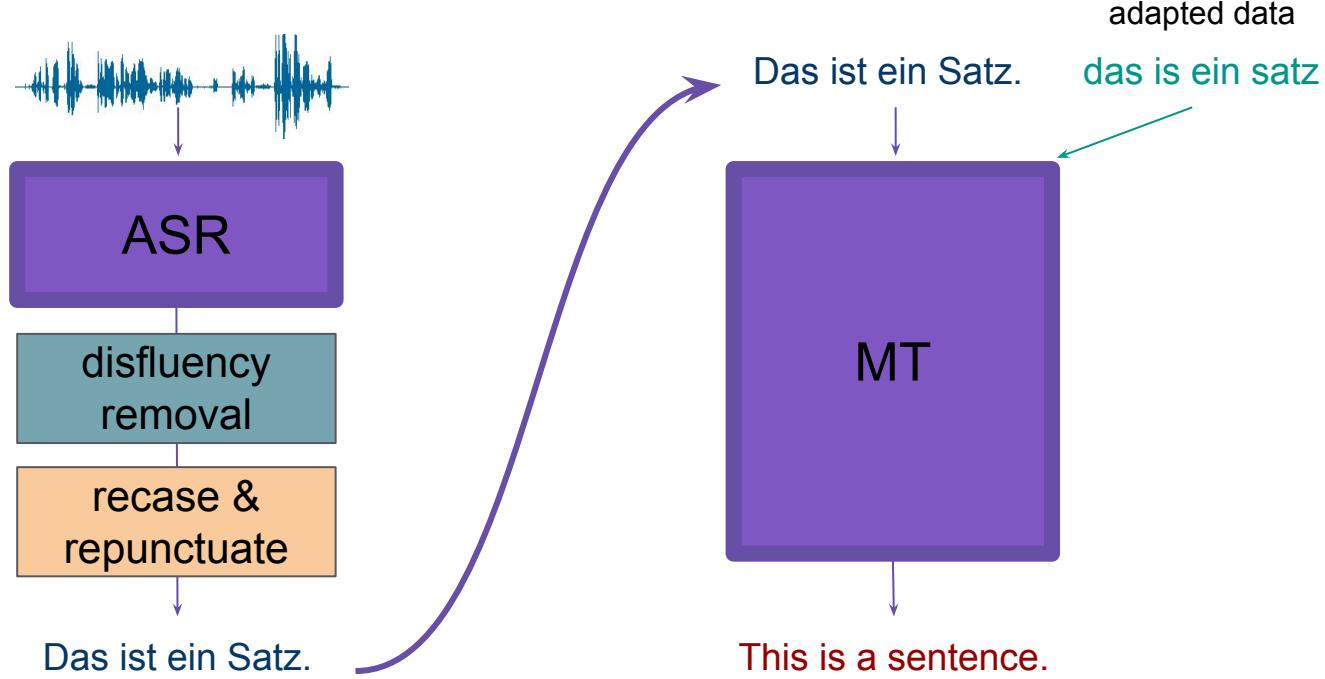


(Cho et al. 2012; Cho et al. 2017)

# Modular Models

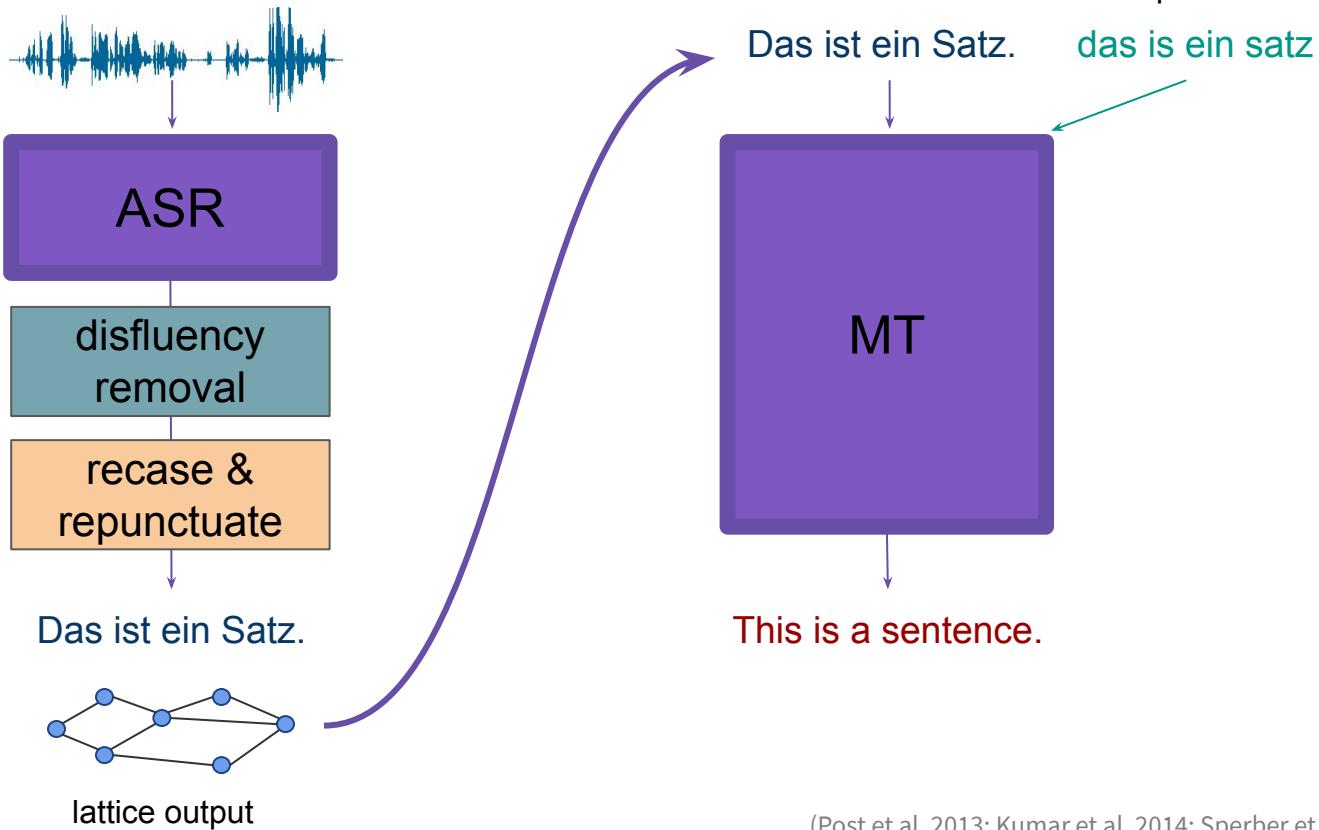


# Modular Models



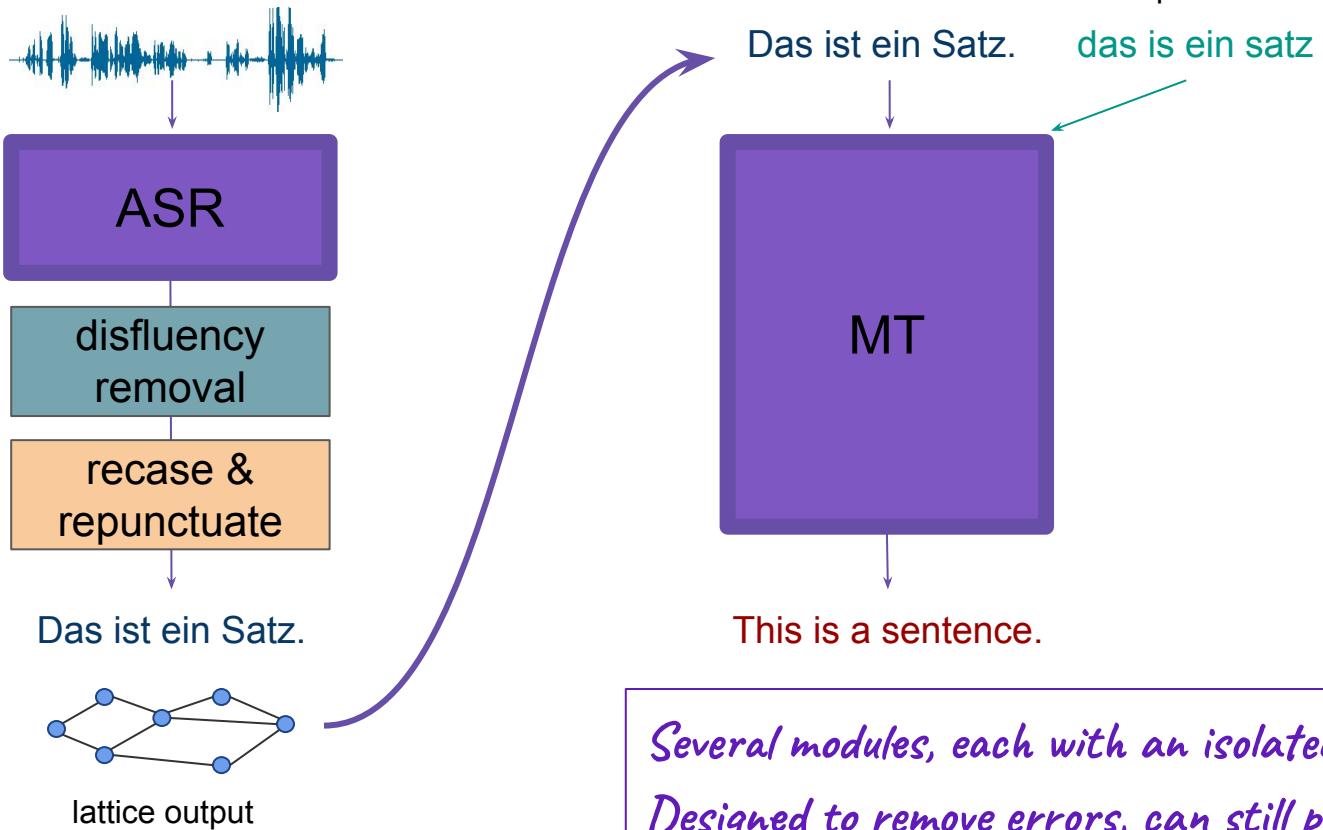
(Tsvetkov et al. 2014; Ruiz et al. 2015;  
Sperber et al. 2017)

# Modular Models



(Post et al. 2013; Kumar et al. 2014; Sperber et al. 2017)

# Modular Models



*Sec 2:*

# End-to-End

**Current state**

**Input representations**

**Architecture modifications**

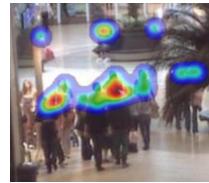
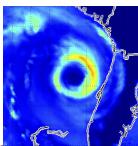
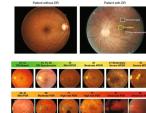
**Output representations**

*Sec 2.1*

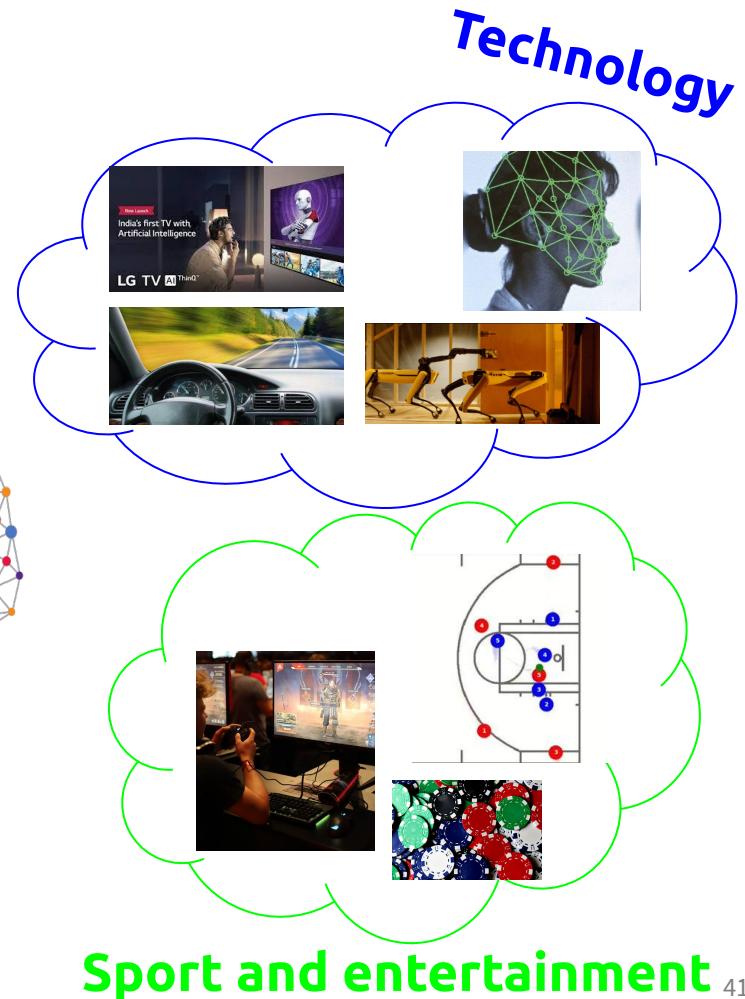
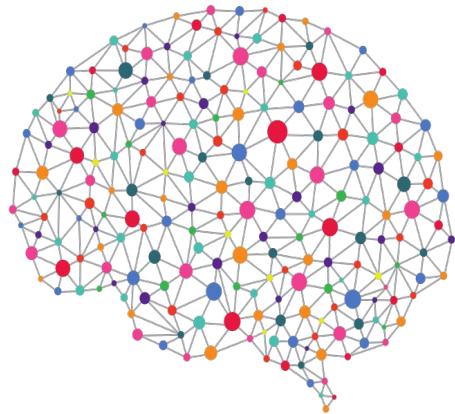
# Current state

# The AI Revolution

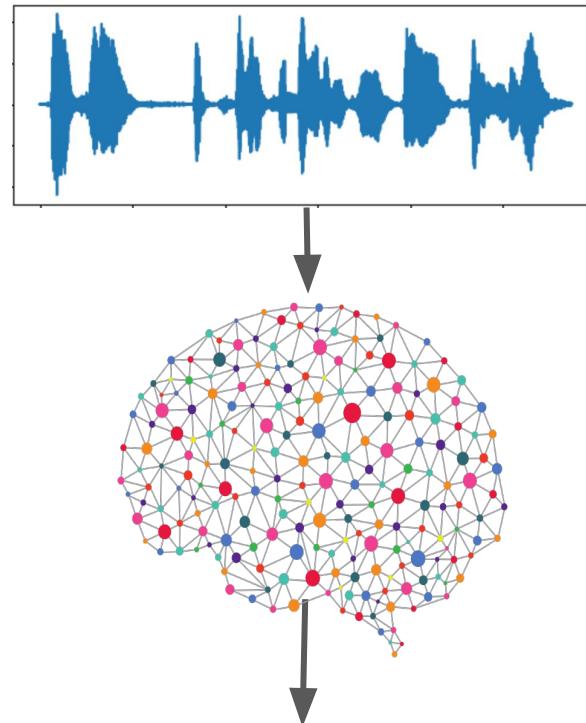
Health



Daily life



# End-to-end SLT (Bérard et al., 2016; Weiss et al., 2017)



What a wonderful tutorial!

# Definition of end-to-end approach

*IWSLT 2020 (Ansari et al., 2020)*

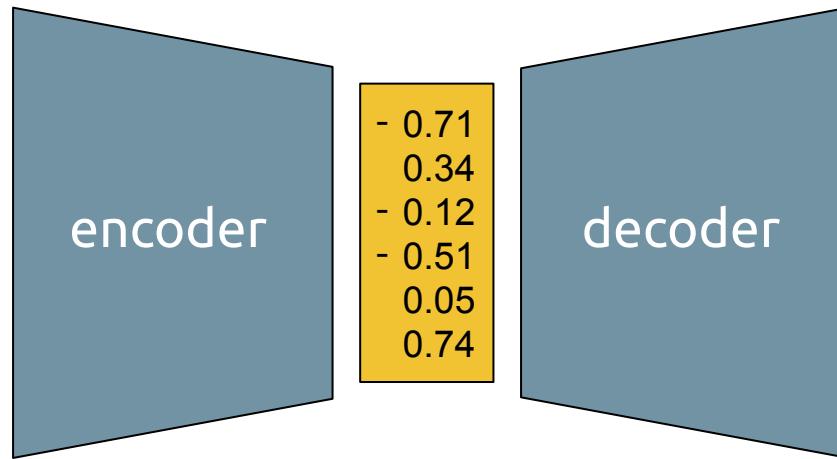
End-to-end model:

- No intermediate discrete representations (transcripts like in cascade or multiple hypotheses like in rover technique)
- All parameters/parts that are used during decoding need to be trained on the end2end task (may also be trained on other tasks → multitasking ok, LM rescoring is not ok)

Other definitions are possible depending on the application

# end-to-end speech translation (e2e)

Spanish Audio



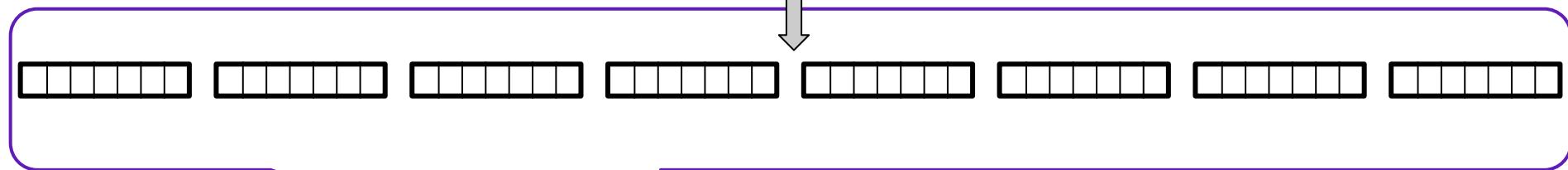
English  
Translated text

What a wonderful tutorial!

# end-to-end speech translation (e2e)

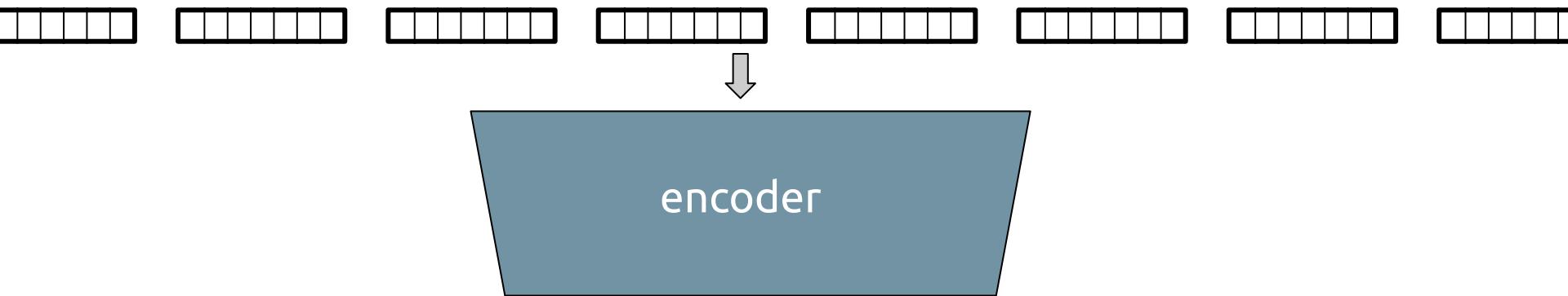


# end-to-end speech translation (e2e)

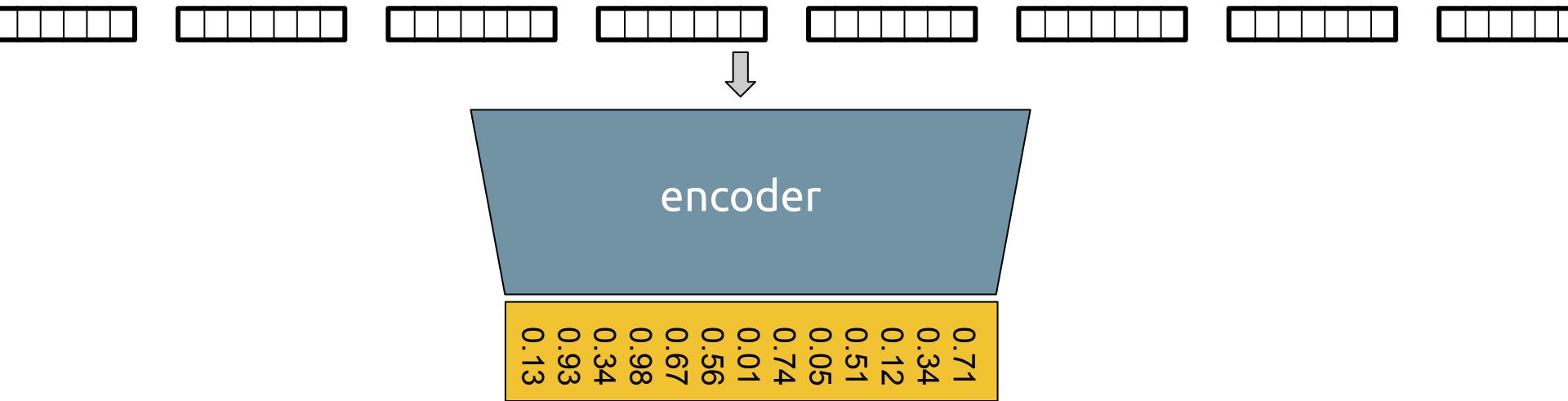


*Audio Representation*

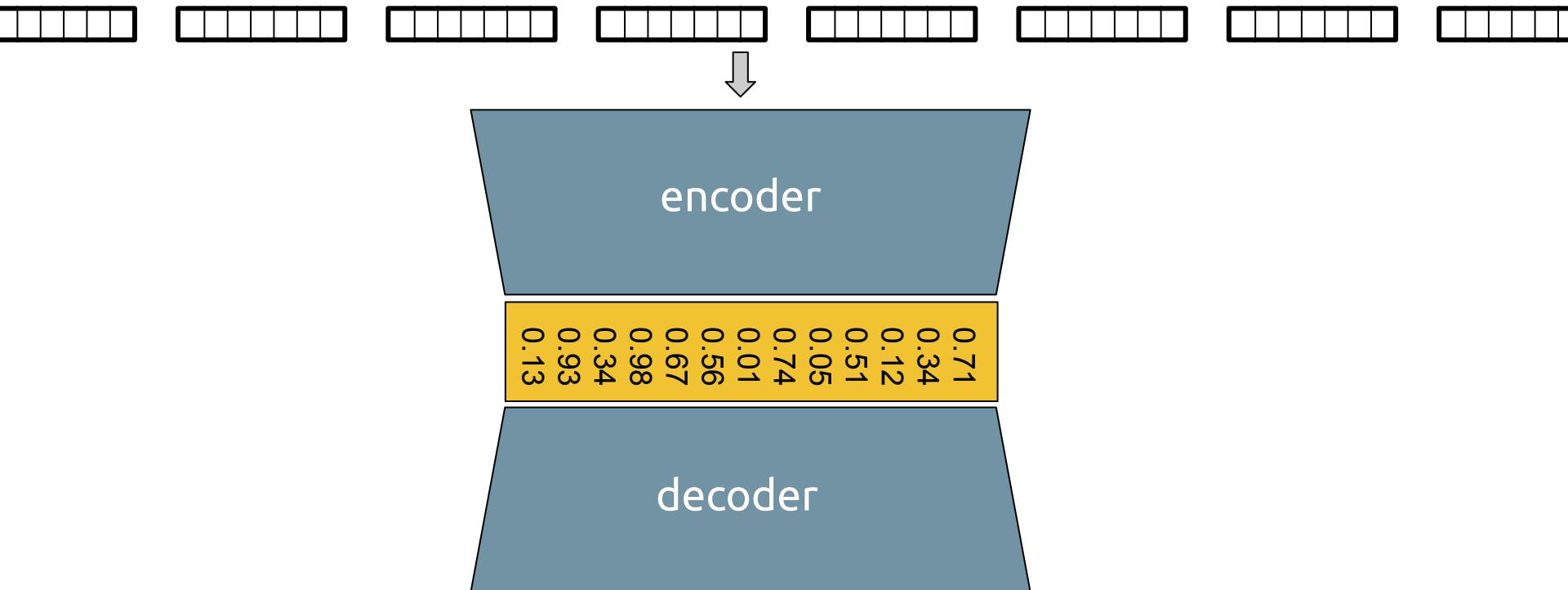
# end-to-end speech translation (e2e)



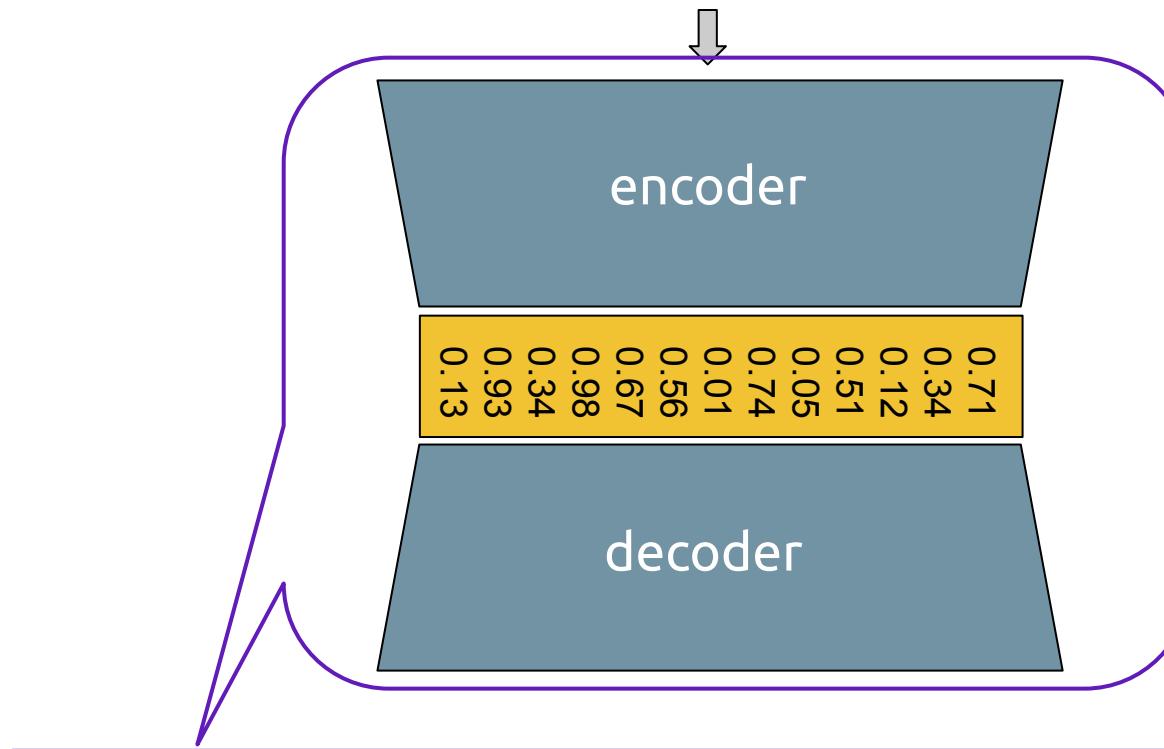
# end-to-end speech translation (e2e)



# end-to-end speech translation (e2e)

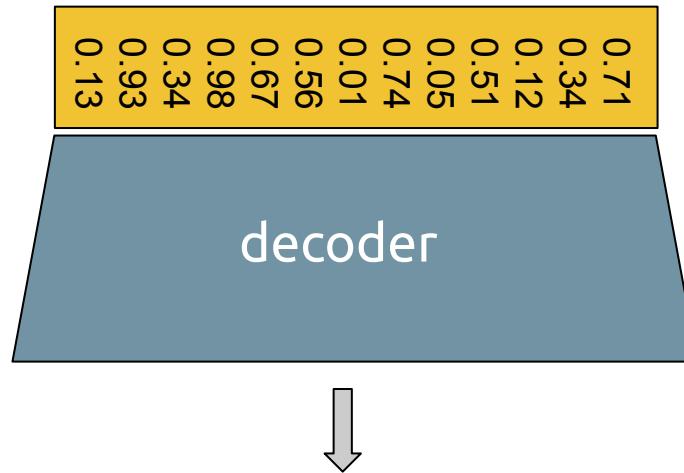


# end-to-end speech translation (e2e)



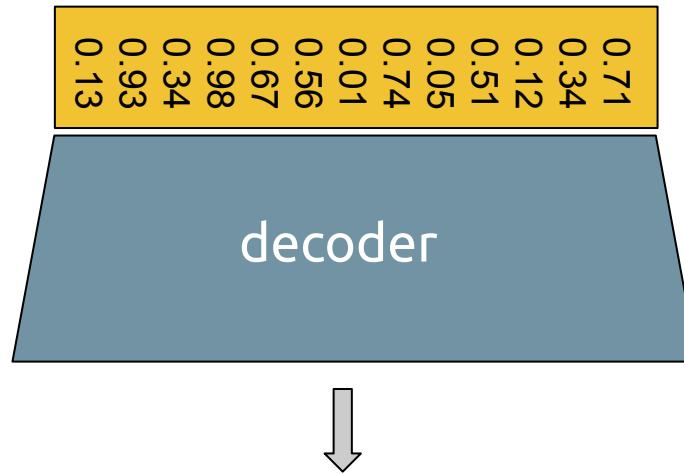
*System Architectures*

# end-to-end speech translation (e2e)



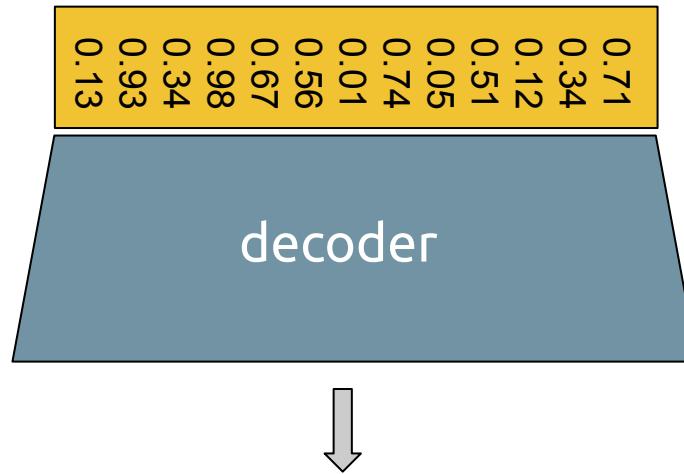
W h a t <space> a <space> w o n d e r f u l <space> t u t o r i a l !

# end-to-end speech translation (e2e)



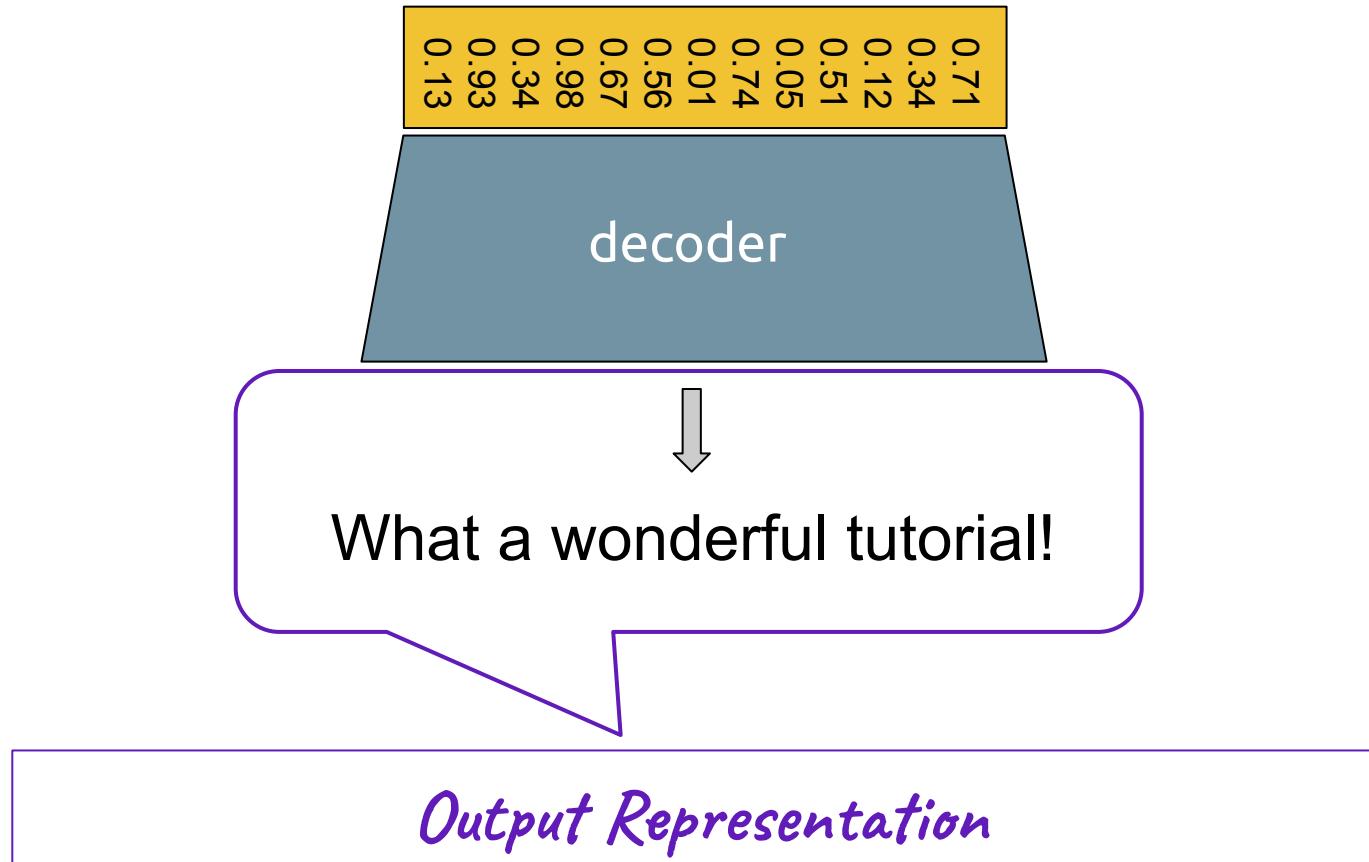
Wh @at a w @on @der @fu @l tut @or @ial!

# end-to-end speech translation (e2e)

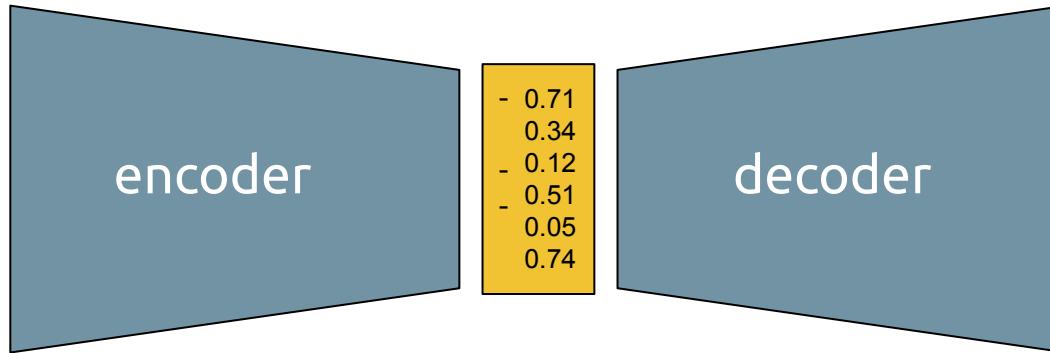


What a wonderful tutorial!

# end-to-end speech translation (e2e)



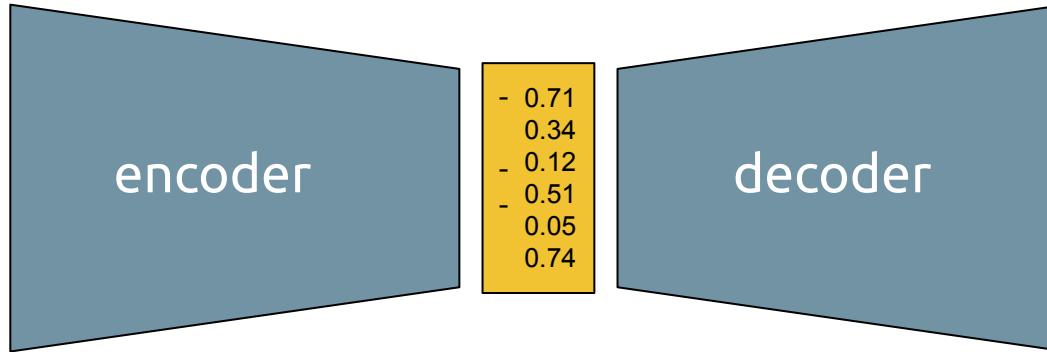
# Sequence-to-Sequence Model



## Pros:

- Direct access to the audio during translation
- No error propagation
- One system to maintain

# Sequence-to-Sequence Model



## Pros:

- Direct access to the audio during translation
- No error propagation
- One system to maintain

## Cons:

- Less consolidated technology
- Scarcity of training data
- Non-monotonic alignments audio-text

# Cascade vs End-to-End Systems

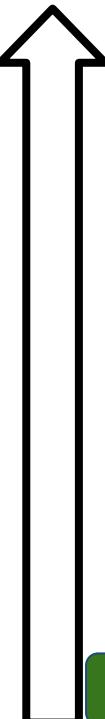
## Cascade

- ✓ 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

# Cascade vs End-to-End Systems

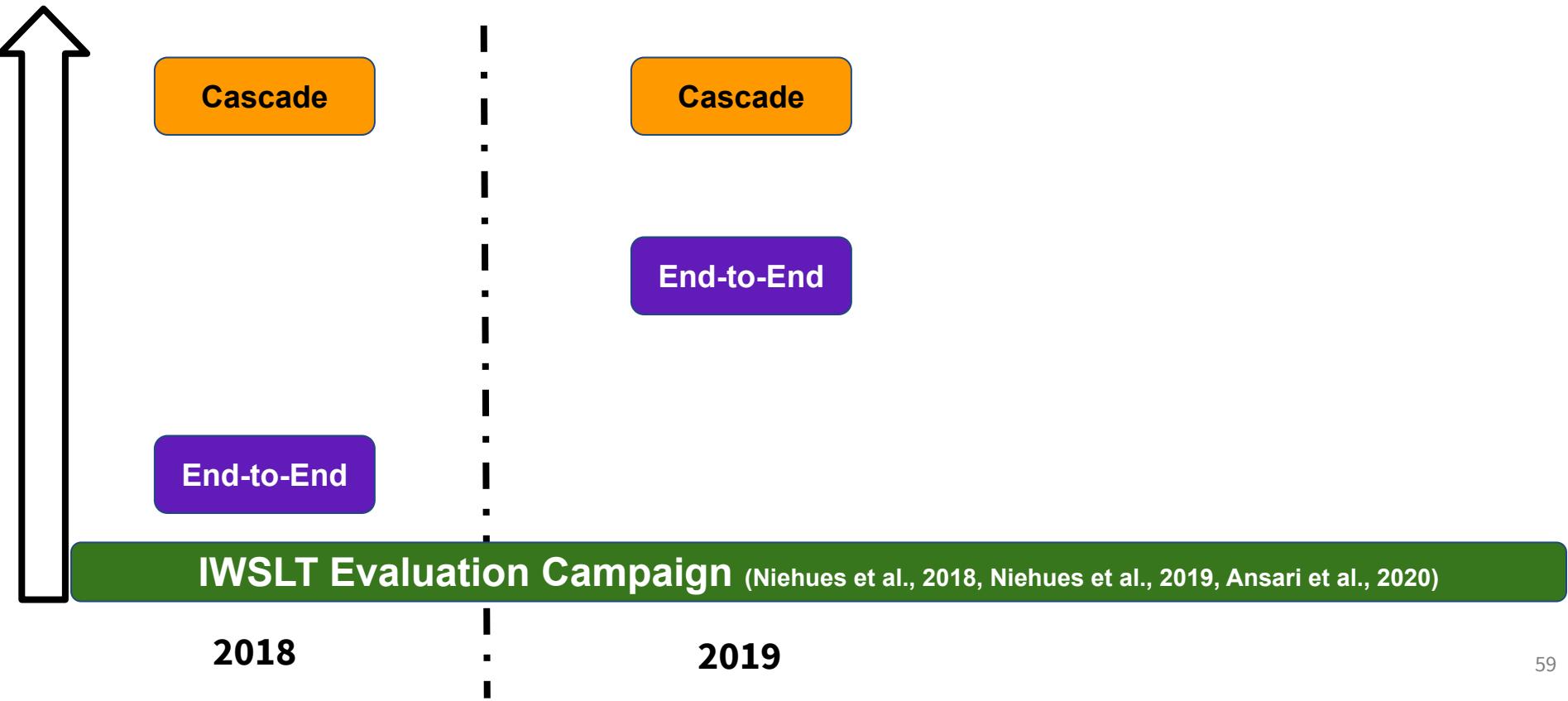


Cascade

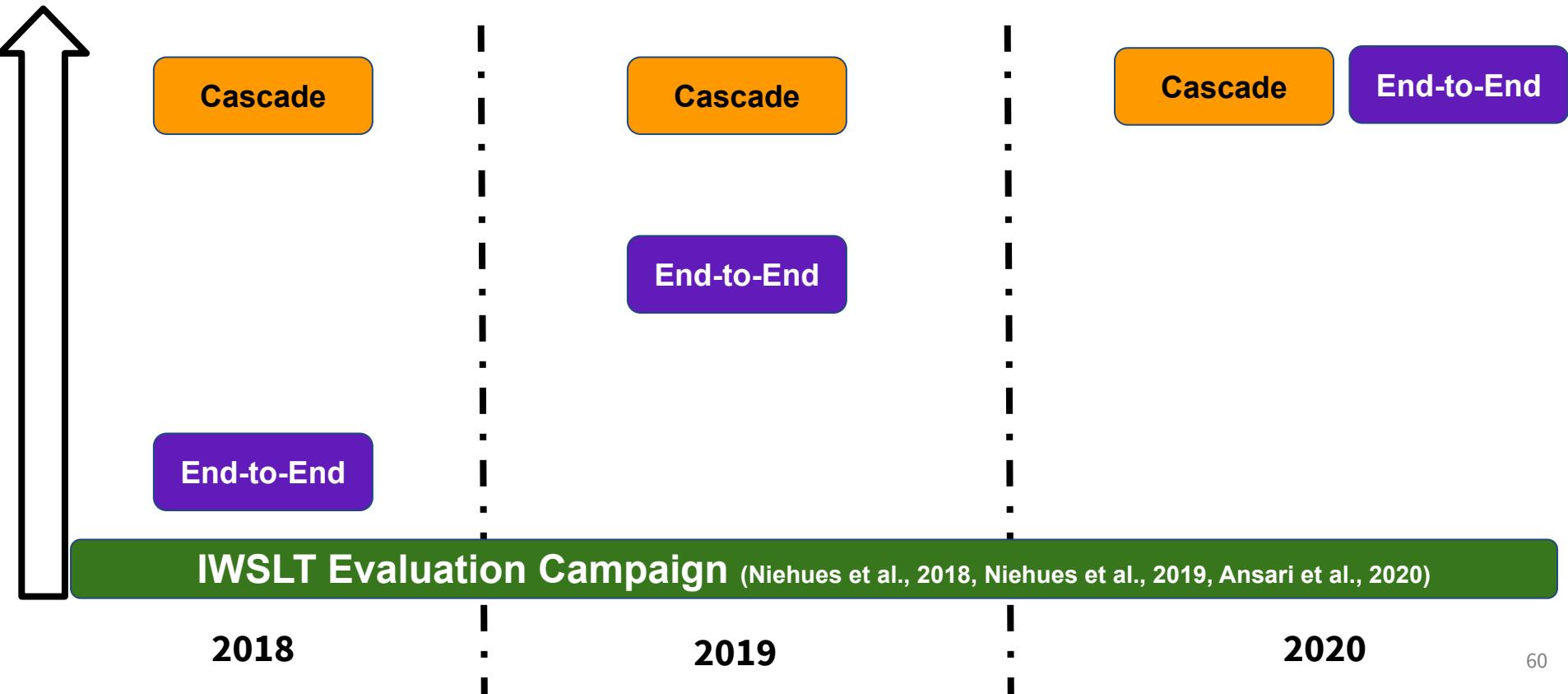
End-to-End

IWSLT Evaluation Campaign (Niehues et al., 2018, Niehues et al., 2019, Ansari et al., 2020)

# Cascade vs End-to-End Systems



# Cascade vs End-to-End Systems



# Cascade vs End-to-End Systems

Most of the papers (Weiss et al., 2017, Jia et al., 2019, Di Gangi et al., 2019) about end-to-end SLT system mention the following advantages over the cascade:

- No error propagation:

End-to-end naturally avoids compounding errors between the ASR and MT systems.

# Cascade vs End-to-End Systems

Most of the papers (Weiss et al., 2017, Jia et al., 2019, Di Gangi et al., 2019) about end-to-end SLT system mention the following advantages over the cascade:

- No error propagation:

End-to-end naturally avoids compounding errors between the ASR and MT systems

- Direct access to the audio:

End-to-end better manipulates paralinguistic and non-linguistic information during translation

*The correctness of these statements taken for granted*

# Cascade vs End-to-End Systems

Key questions:

*Is it true that end-to-end avoids error propagation?*

*To what extent does accessing the audio help? How? When?*

# Cascade vs End-to-End Systems

Key questions:

*Is it true that end-to-end avoids error propagation?*

*To what extent does accessing the audio help? How? When?*

No answers in this tutorial!

# No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.

# No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.
- For a direct comparison of the Cascade and e2e, the intermediate representations cannot be used (transcript vs. null)

# No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.
- For a direct comparison of the Cascade and e2e, the intermediate representations cannot be used (transcript vs. null)
- Difficult to disentangle the impact of various components in e2e (two tasks collapsed into one)

# No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.
- For a direct comparison of the Cascade and e2e, the intermediate representations cannot be used (transcript vs. null)
- Difficult to disentangle the impact of various components in e2e (two tasks collapsed into one)
- Not a consolidated architecture in end-to-end technology

# No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.
- For a direct comparison of the Cascade and e2e, the intermediate representations cannot be used (transcript vs. null)
- Difficult to disentangle the impact of various components in e2e (two tasks collapsed into one)
- Not a consolidated architecture in end-to-end technology

Possible opening:

Sperber et al., (2019) consider the encoder output as an intermediate representation and pose the attention on the presence of errors in it

# **Direct access to the audio**

Open issues:

- Better encoder technology results in better translation performance (not enough)

# Direct access to the audio

Open issues:

- Better encoder technology results in better translation performance (not enough)
- Not clear what aspects of the audio can help (e.g. prosody, emotions, tone, pauses)

# Direct access to the audio

Open issues:

- Better encoder technology results in better translation performance (not enough)
- Not clear what aspects of the audio can help (e.g. prosody, emotions, tone, pauses)
- Audio understanding capability can only be analyzed in the final translation (no transcripts)

# Direct access to the audio

Open issues:

- Better encoder technology results in better translation performance (not enough)
- Not clear what aspects of the audio can help (e.g. prosody, emotions, tone, pauses)
- Audio understanding capability can only be analyzed in the final translation (no transcripts)
- Lack of *ad hoc* test sets to measure the impact of prosody, emotions, ...

# Direct access to the audio

Open issues:

- Better encoder technology results in better translation performance (not enough)
- Not clear what aspects of the audio can help (e.g. prosody, emotions, tone, pauses)
- Audio understanding capability can only be analyzed in the final translation (no transcripts)
- Lack of *ad hoc* test sets to measure the impact of prosody, emotions, ...

Possible openings:

Karakanta et al. (2020): the direct access to the audio pauses improves subtitles' quality

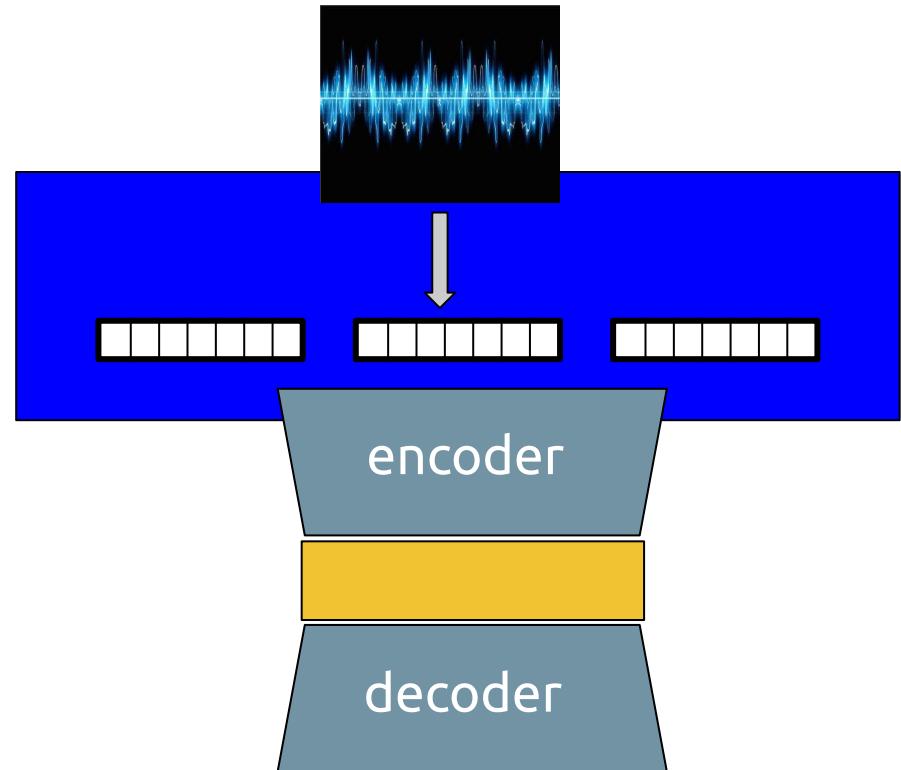
Gaido et al. (2020): vocal characteristics can guide e2e systems in modeling gender (but opens ethical issues!)

*Sec 2.2*

# Input representations

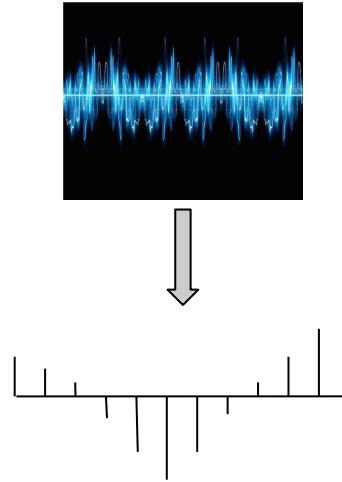
# From text translation to speech translation

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



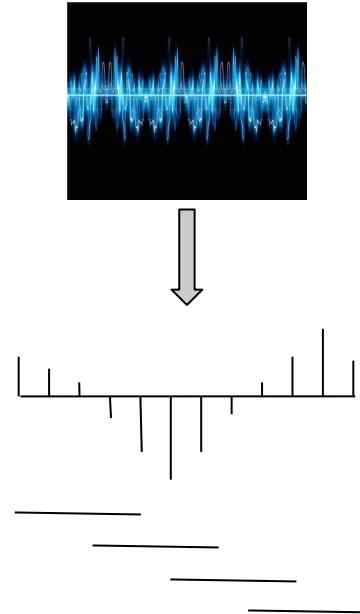
# Audio representation

- 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

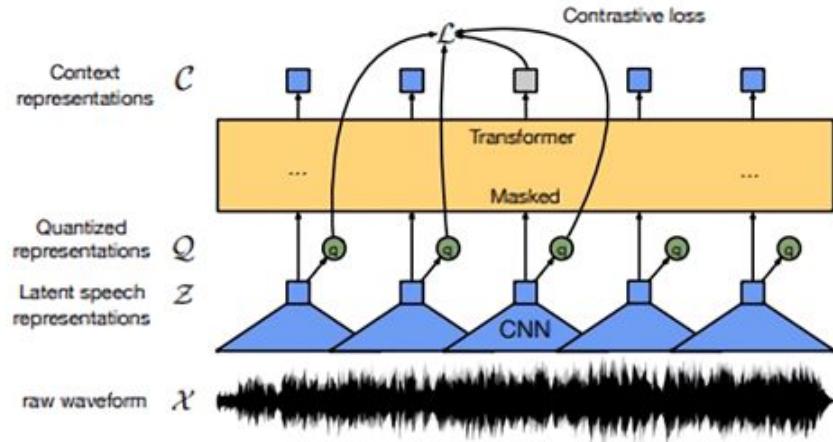


# Audio representation

- 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

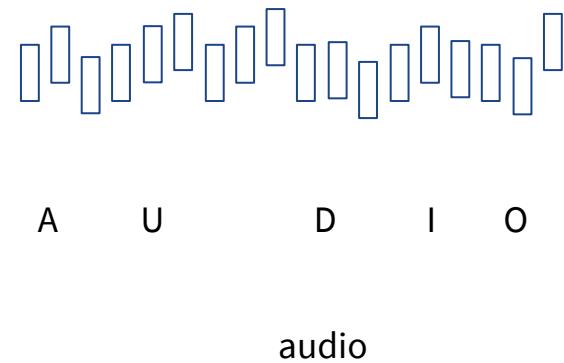
- Input features:
  - Signal processing:
  - Deep Learning:
    - Self-supervised Learning
      - Predict frame based on context
    - E.g. Wav2Vec 2.0 (Baevski et al., 2020)



Baevski et al. 2020

# Challenges

- Variation
  - Many different ways to speech same sentence
  - Data augmentation
- Sequence Length
  - IWSLT test set 2020
    - Segments: 1804
    - Words: 32.795
    - Characters: 149.053
    - Features: 1.471.035
  - Architectural changes

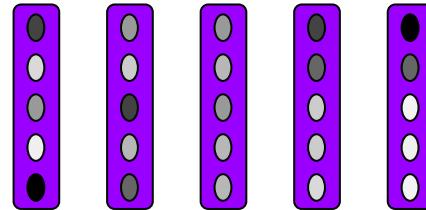


# Data augmentation

- Limited training data
- Generate synthetic training data
- 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)

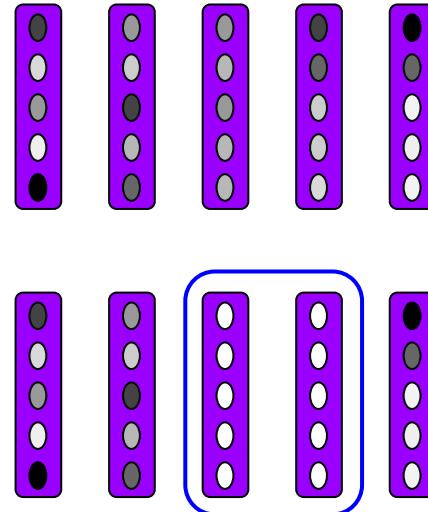
# SpecAugment

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



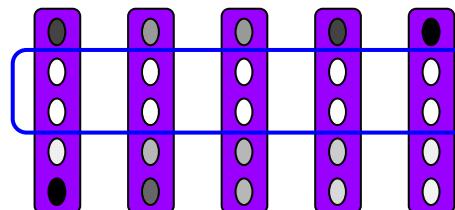
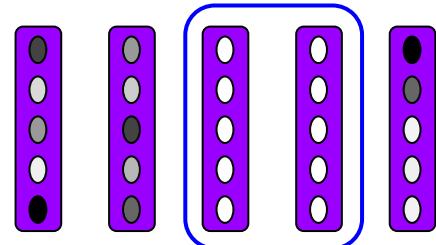
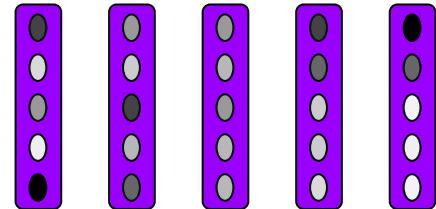
# SpecAugment

- Directly applied on audio features
- Idea:
  - Mask information
- *Time masking*
  - Set several consecutive feature vector to zero



# SpecAugment

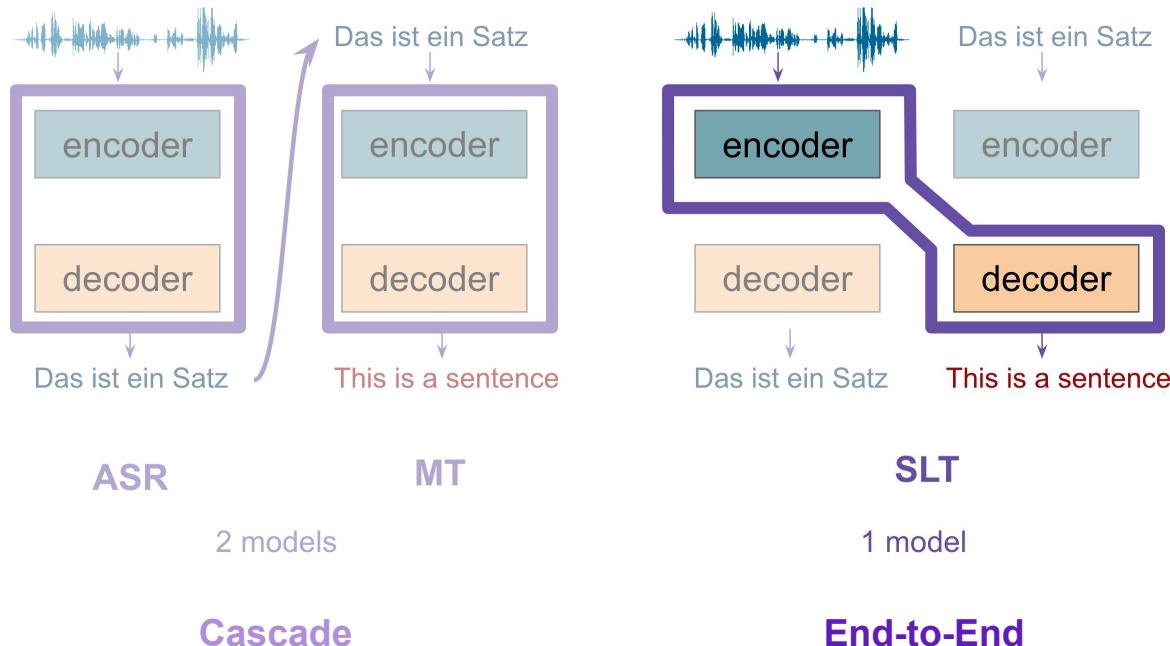
- Directly applied on audio features
- Idea:
  - Mask information
- *Time masking*
  - Set several consecutive feature vector to zero
- *Frequency masking*
  - Mask consecutive frequency channels



*Sec 2.3*

# Architecture & Modifications

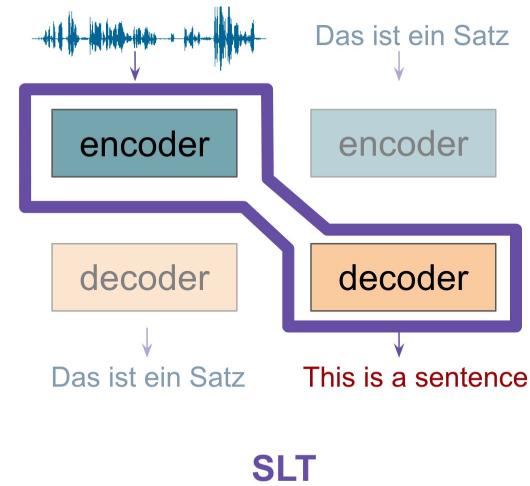
# End-to-End Architecture



# End-to-End Architecture

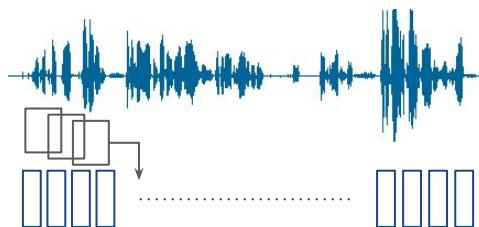
LSTM or Transformer  
Encoder-Decoder Models

*However, speech ≠ text*

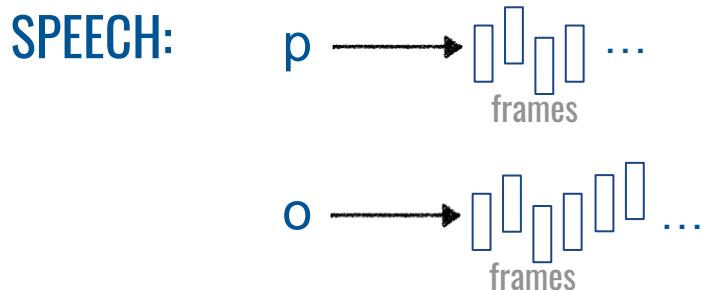


End-to-End

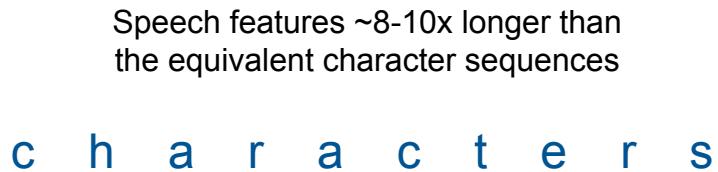
# Speech vs. Text



Discretized audio — speech frames



Each feature vector is unique,  
Number of feature vectors per phone varies



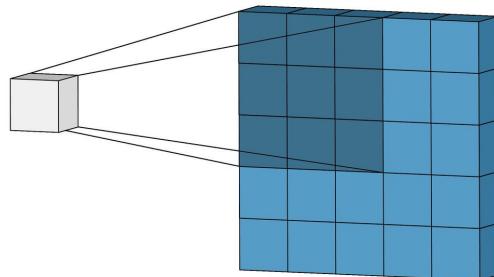
# Challenges

- Sequence length:
  - increased memory requirements
  - greater distance between dependencies
- Redundancy:
  - adds task for model to learn
- Variation:
  - requires more data for model to learn correspondences

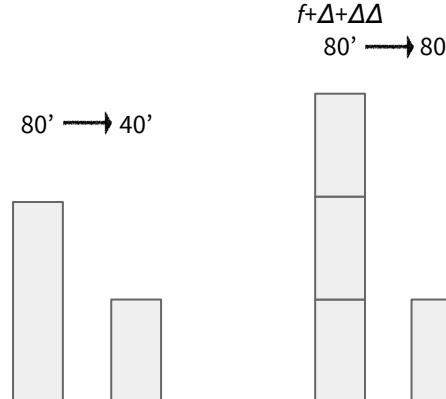
# Dimensionality Reduction

Two directions: ① temporal and ② feature dimension

Convolutional layers enable *fixed-length downsampling*



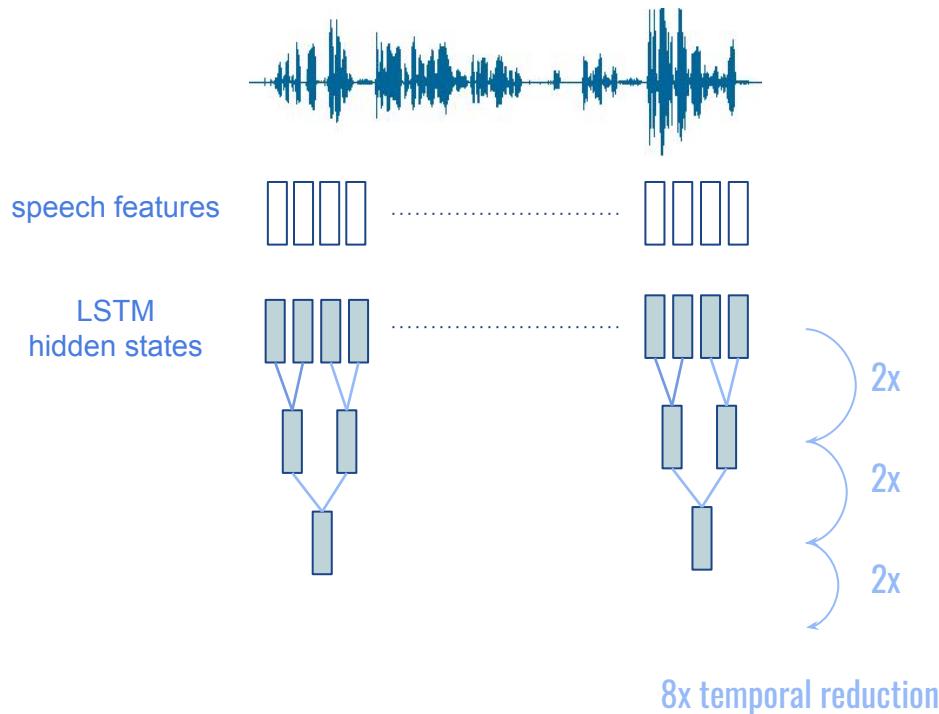
Scale sequence length and feature dimension linearly by a factor corresponding to the convolutional kernel size and stride length



Conv1D, ConvLSTM layers

(Weiss et al. 2017;  
Bansal et al. 2018)

# 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)

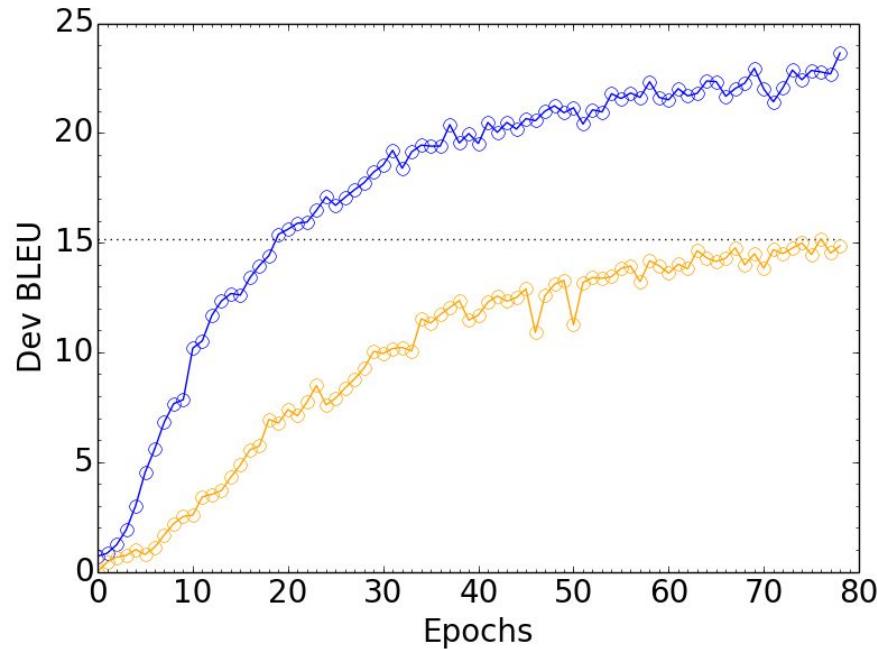
Pyramidal encoder in ST:  
(Weiss et al. 2017; Salesky et al. 2019;  
Sperber et al. 2019; Salesky et al. 2020)

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

# Dimensionality Reduction Impact

*Improved training efficiency!*

- Reduces memory footprint
- Faster convergence
- Improved results



(Salesky et al. 2019)

# Encoder and Decoder Depth

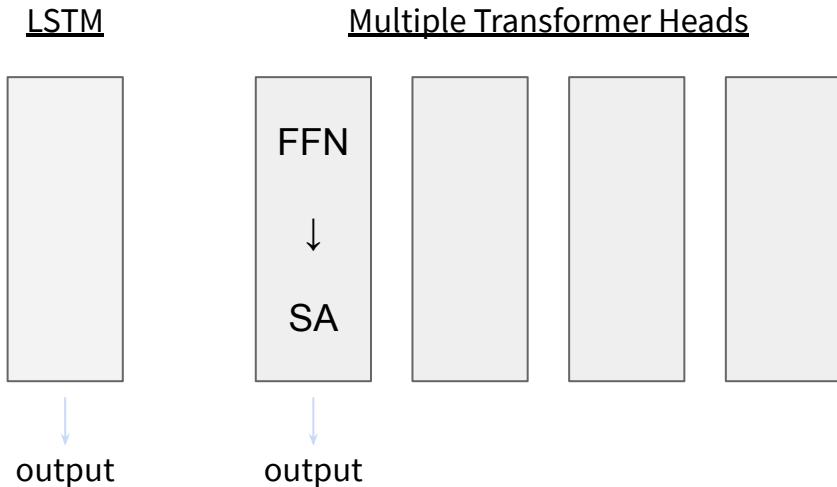
**MT**: typically same depth for encoder and decoder

**ST**: empirically, deeper encoders than decoders perform better!

→ *more parameters allocated to learning more complicated associations between inputs*

Models	Test WER
CTC [19]	17.4
CTC/LM + speed perturbation [19]	13.7
12Enc-12Dec (Ours)	14.2
Stc. 12Enc-12Dec (Ours)	12.4
Stc. 24Enc-24Dec (Ours)	11.3
Stc. 36Enc-12Dec (Ours)	<b>10.6</b>

# LSTM → Transformer



## Transformer-S

- 2D Convolutions
- Distance penalty for attention
- 2D self-attention

...

## Conv-Transformer

(DiGangi et al. 2019; Huang et al. 2020)

*Sec 2.4*

# Output representations

# Output representation

Bla

Word (<https://arxiv.org/pdf/1803.09164.pdf>) Bansal, Sameer, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. "Low-resource speech-to-text translation." *arXiv preprint arXiv:1803.09164* (2018).

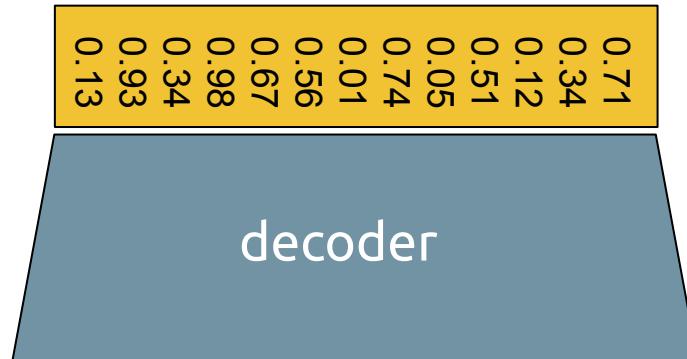
Char (Weiss, Berard)

Bpe (??) Sperber, Matthias, Ngoc Quan Pham, Thai Son Nguyen, Jan Niehues, Markus Müller, Thanh-Le Ha, Sebastian Stüker, and Alex Waibel. "Kit's iwslt 2018 slt translation system." (2018).

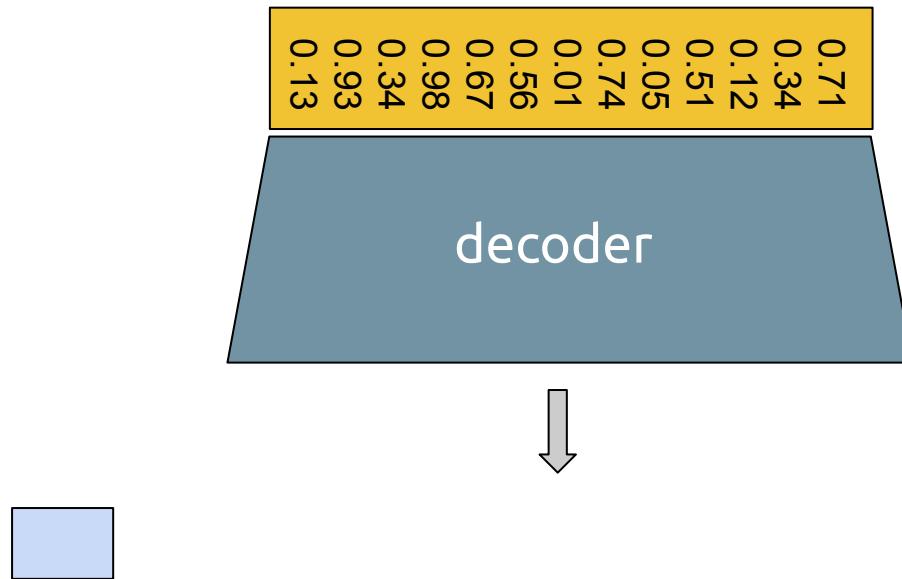
~~Automatic generation of BPE (DPE)~~

Comparison (FBK) Di Gangi, Mattia A., Marco Gaido, Matteo Negri, and Marco Turchi. "On Target Segmentation for Direct Speech Translation." In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pp. 137-150. 2020.

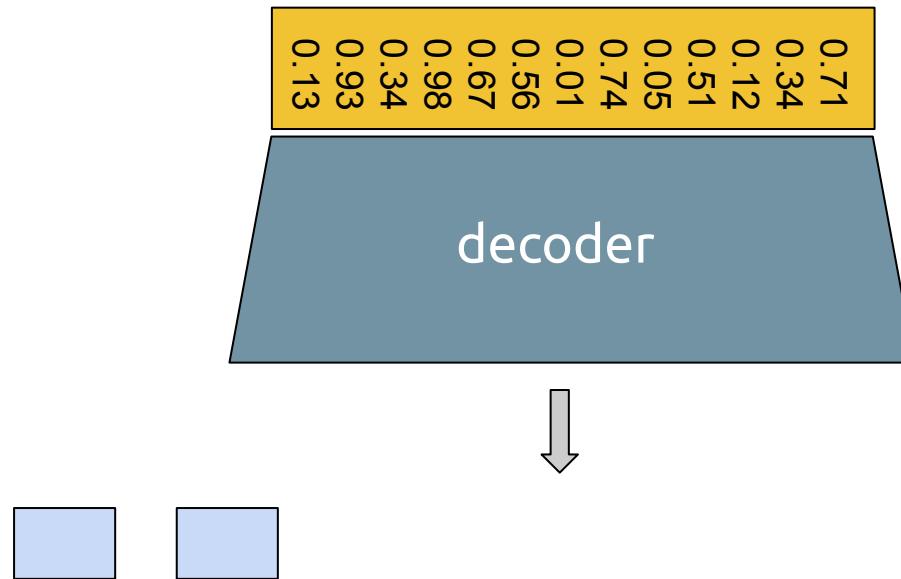
# Output representation



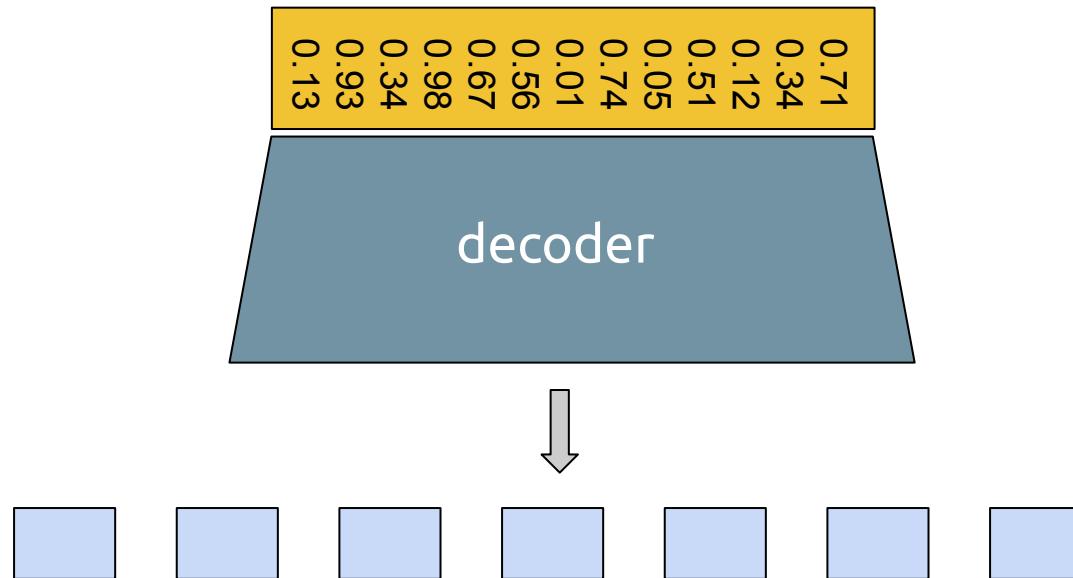
# Output representation



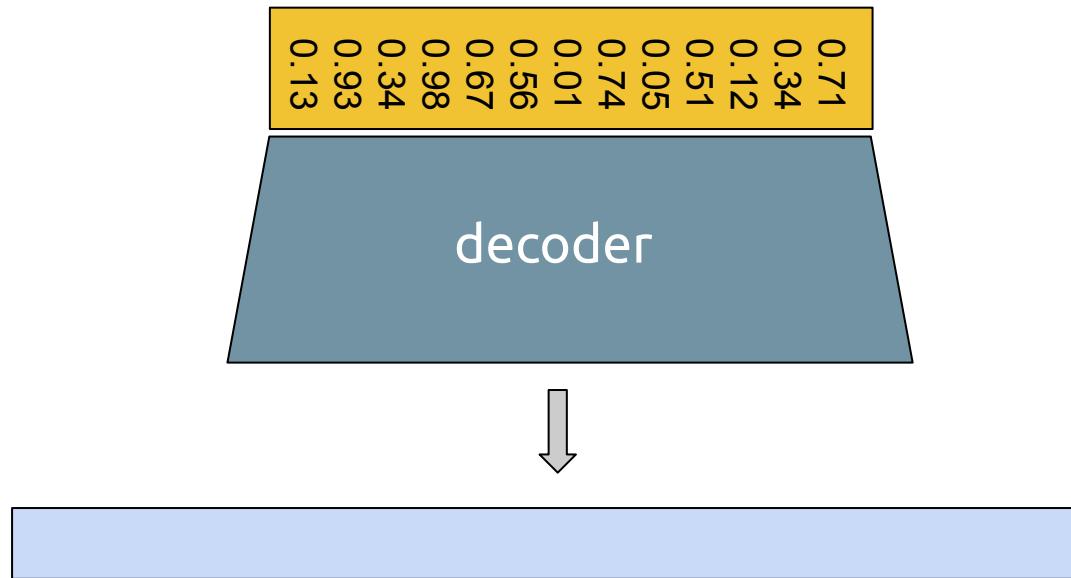
# Output representation



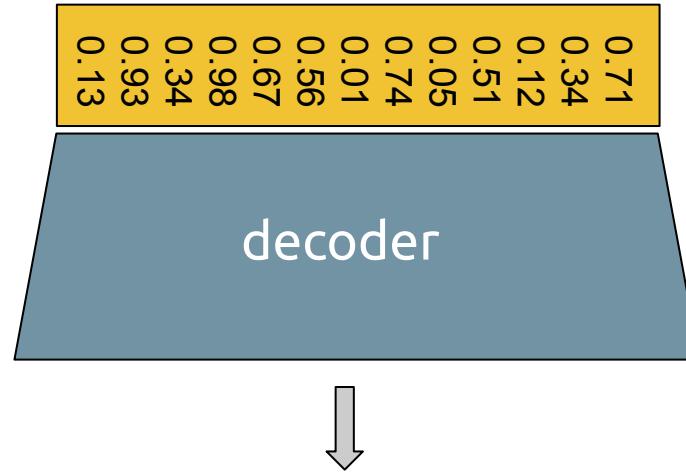
# Output representation



# Output representation



# Output representation



What a wonderful tutorial!

# Output representation

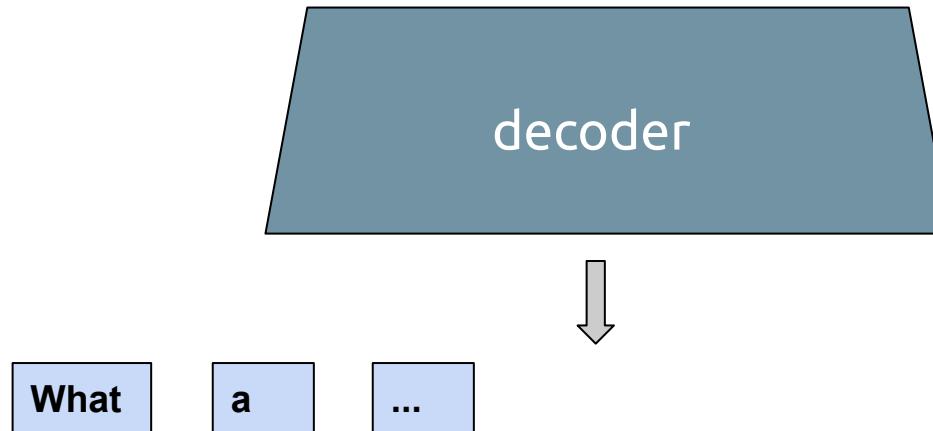
- Word (Bansal et al., 2018)
- Byte Pair Encoding (BPE) (Sperber et al., 2018)
- Character (Bérard et al., 2016; Weiss et al., 2017)

# Output representation: Word

- Words as atomic unit
- Applicable only for small and high-repetitive datasets
- Tested in low-resource speech-to-text translation

# Output representation: Word

- Words as atomic unit
- Applicable only for small and high-repetitive datasets
- Tested in low-resource speech-to-text translation



# Output representation: BPE

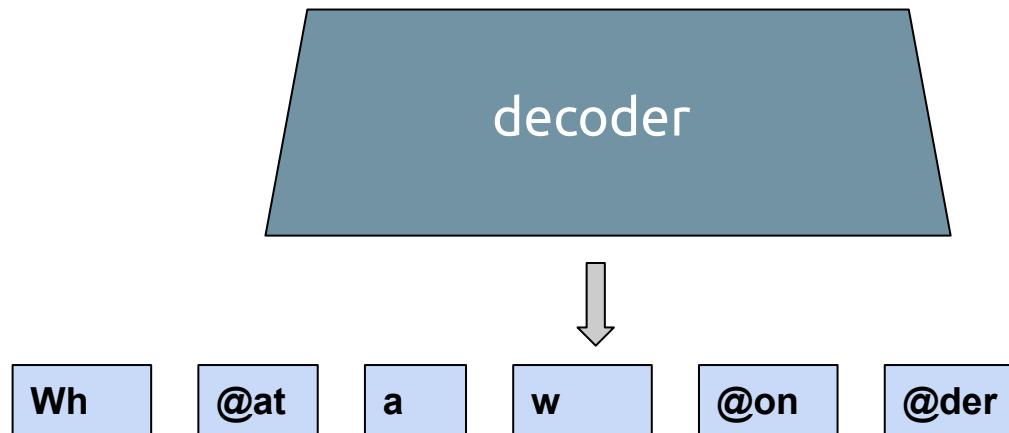
- Introduced in Neural Machine Translation to fit a large vocabulary in memory
- Each target sentence splits in sub-word units
- Iterative approach merging the most frequently co-occurring characters or character sequences
- Widely used in several NLP tasks

# Output representation: BPE

- Training and test data are split based on a learned vocabulary
- After translation, BPEs converted into words

# Output representation: BPE

- Training and test data are split based on a learned vocabulary
- After translation, BPEs converted into words

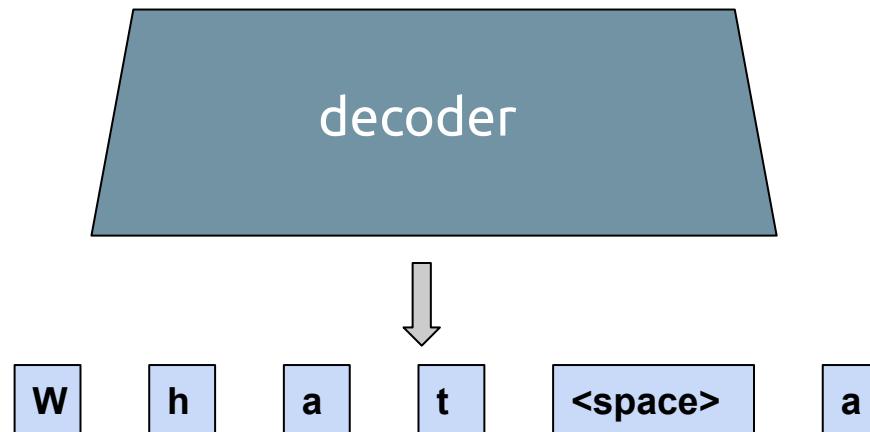


# Output representation: Characters

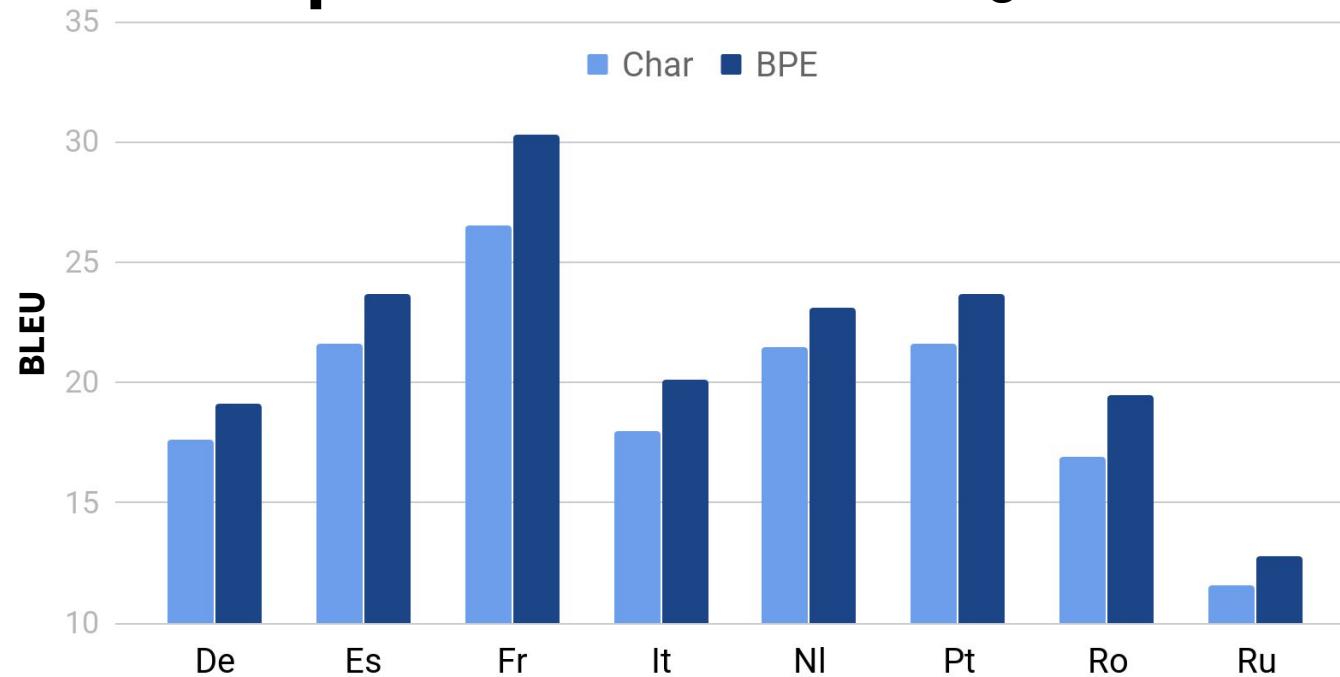
- Each sentence splits in characters with a special symbol for the empty space
- Training and test data are split
- After translation, characters converted into words

# Output representation: Characters

- Each sentence splits in characters with a special symbol for the empty space
- Training and test data are split
- After translation, characters converted into words

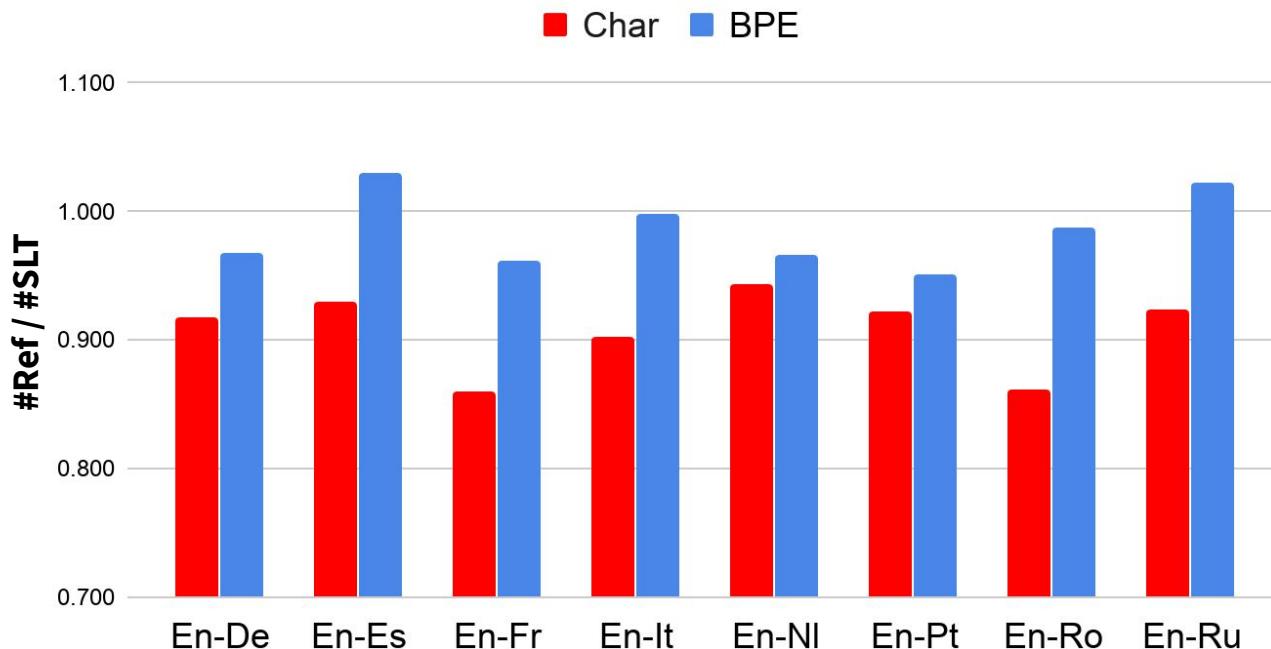


# Translation performance (Di Gangi et al., 2020)



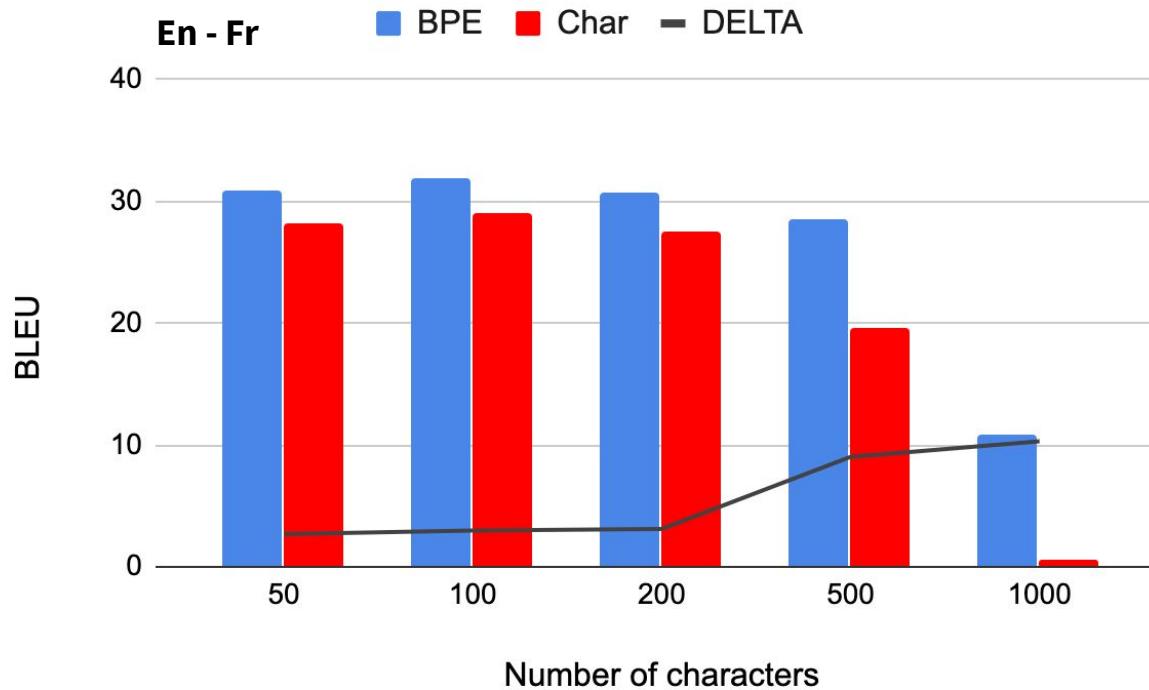
*BPE outperforms Characters in all languages*

# Length comparison



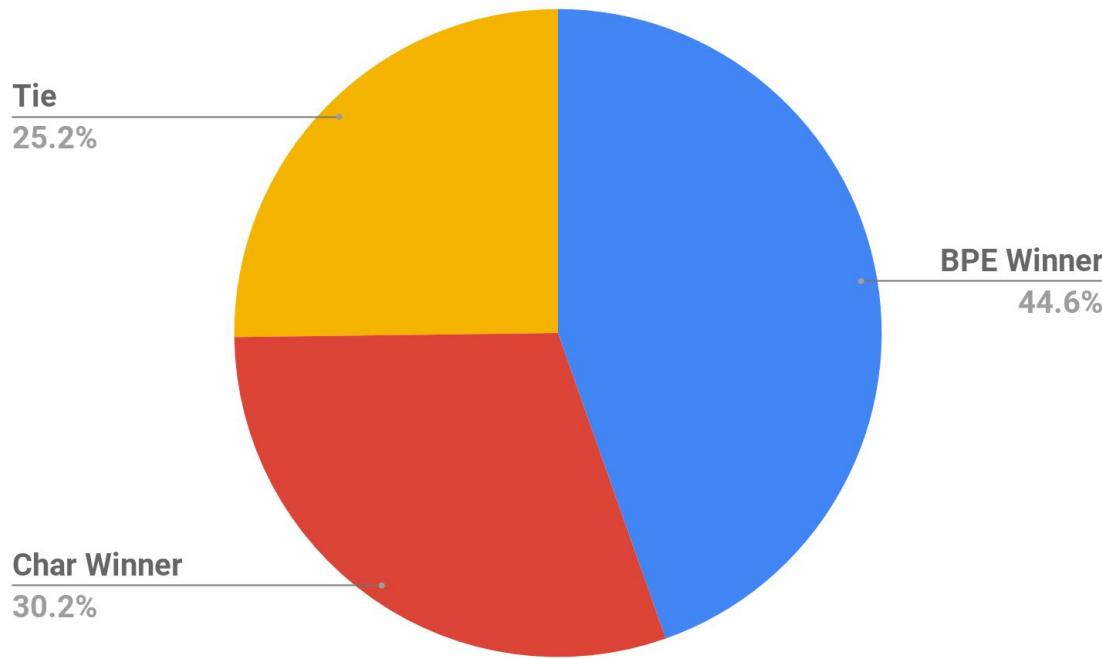
*BPE produces longer sentences*

# Translation quality by sent. length



*BPE better on longer sentences*

# Sentence Level Comparison



*Chars better on lower quality translations*

*Sec 3:*

# Leveraging Data Sources

**Available data**

**Techniques**

Multi-task learning

Transfer learning and pretraining

Knowledge distillation

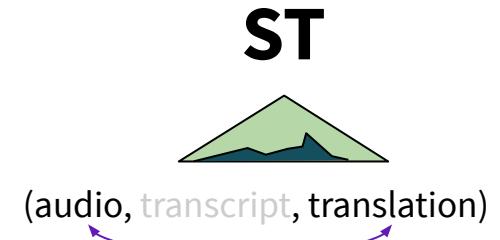
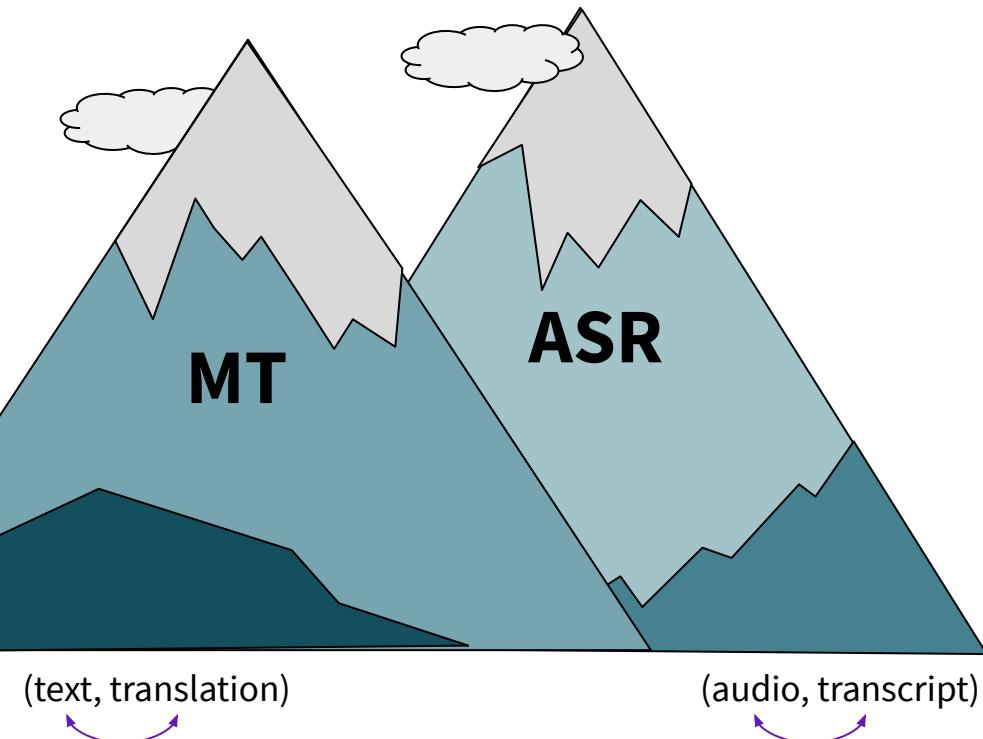
**Alternate data representations**

---

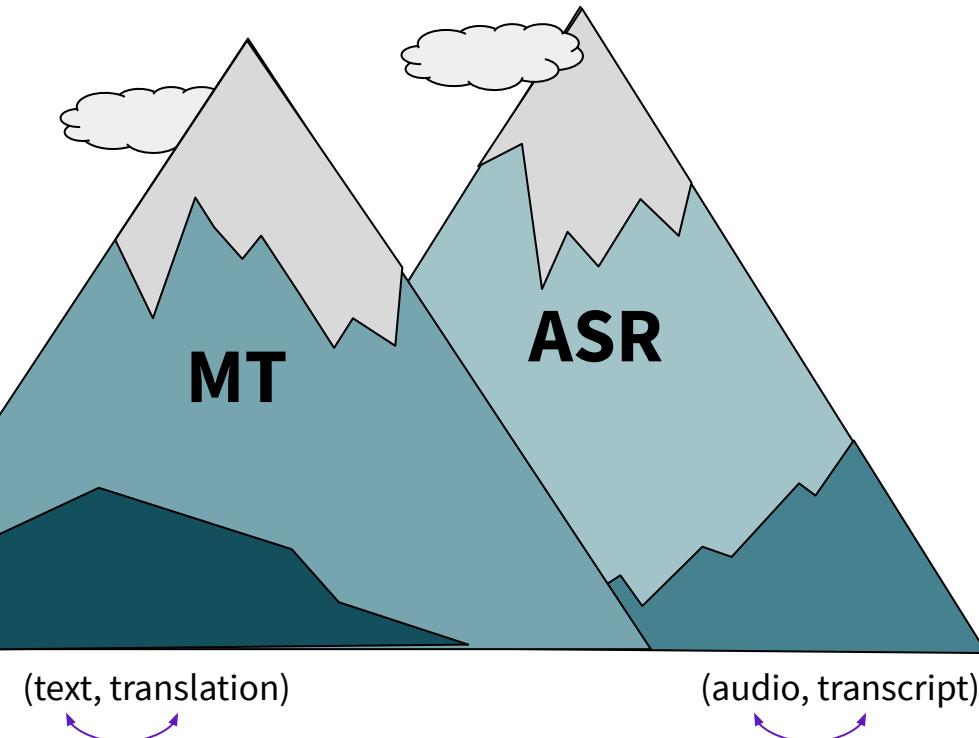
*Sec 3.1*

# Available Data

# Available data



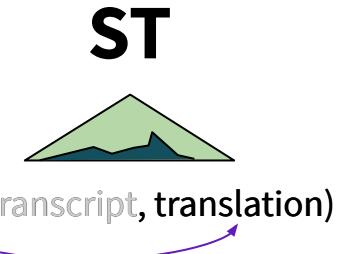
# Available data



**Question: Why so few data?**  
**Answer: High creation costs!**

1. Find good data (e.g. audio+transcr+transl., free)
2. Download and clean
3. Segment transcripts and translations
4. Align transcripts and translations
5. Align transcripts and audio
6. Filter wrong/poor alignments
7. Pack in suitable format, extract features

MuST-C (Cattoni et al., 2021)



# Available data ( $\geq$ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
Multilingual TEDx	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

# Available data ( $\geq$ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
<b>MuST-C</b>	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
<b>CoVoST</b>	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
<b>Europarl-ST</b>	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
<b>LibriVoxDeEn</b>	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
<b>MaSS</b>	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
<b>BSTC</b>	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
<b>Multilingual TEDx</b>	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

Half of these corpora were built in the last 2 years

# Available data ( $\geq$ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
<b>How2</b>	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
<b>IWSLT 2018</b>	(Niehues et al., 2018)	En→De 273hrs	TED talks
<b>LIBRI-TRANS</b>	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
<b>MuST-C</b>	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
<b>CoVoST</b>	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
Multilingual TEDx	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

*Trend (1): increasing data size (>200 hours of translated speech)*

# Available data ( $\geq$ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
<b>MuST-C</b>	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
<b>CoVoST</b>	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
<b>Europarl-ST</b>	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
<b>MaSS</b>	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
<b>Multilingual TEDx</b>	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

*Trend (2): more language directions*

# Available data ( $\geq$ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
<b>CoVoST</b>	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
<b>Europarl-ST</b>	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
<b>MaSS</b>	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
<b>Multilingual TEDx</b>	(Salesky et al., 2021)	8 lang.→6 lang. 11-69hrs	TED talks

*Trend (3): multilinguality + non-English speech*

# Available data ( $\geq$ 20 hrs of speech)

(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
MuST-C	(Cattoni et al., 2021)	En→ 14 lang. (237-504hrs)	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
<b>Multilingual TEDx</b>	(Salesky et al., 2021)	<b>8 lang.→6 lang. 11-69hrs</b>	TED talks

*Trend (4): same segmentation across datasets*

# Available data ( $\geq$ 20 hrs of speech)

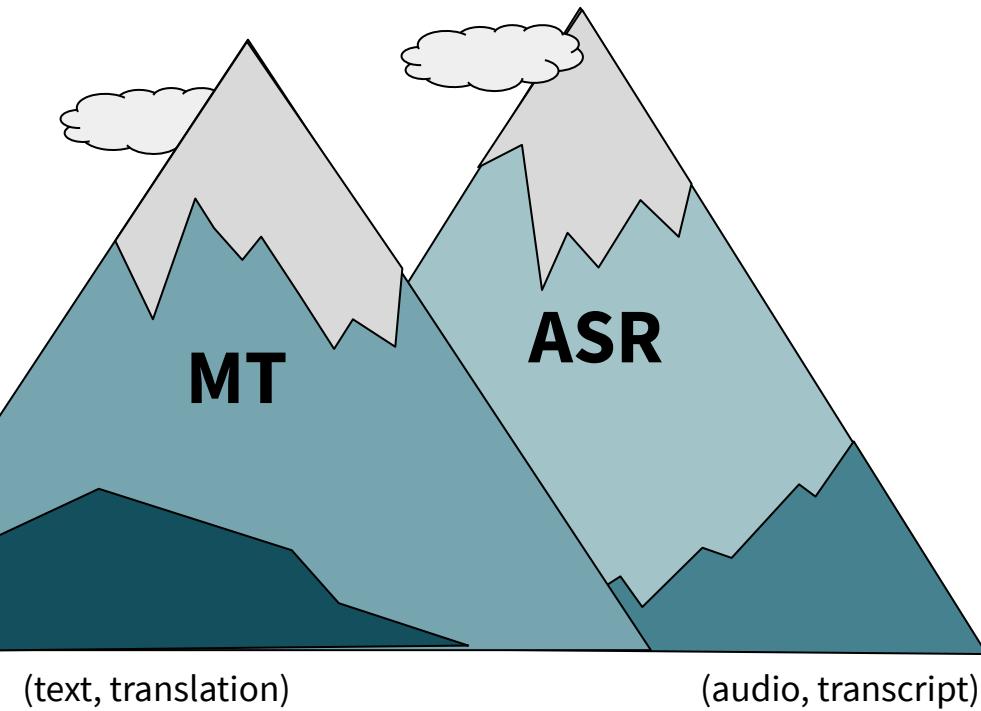
(no name)	(Tohyama et al., 2005)	En↔Jp 182hrs	simult. interpret.
(no name)	(Paulik and Waibel, 2009)	En→Es 111 Es→En 105hrs	simult. interpret.
Fisher	(Post 2013)	Es→En 160hrs	phone conversations
STC	(Shimizu et al., 2014)	En↔Jp 22hrs	simult. interpret.
How2	(Sanabria et al., 2018)	En→Pt 300hrs	instructional videos
IWSLT 2018	(Niehues et al., 2018)	En→De 273hrs	TED talks
LIBRI-TRANS	(Kocabiyikoglu et al., 2018)	En→Fr 236hrs	read audiobooks
<b>MuST-C</b>	(Cattoni et al., 2021)	<b>En→ 14 lang. (237-504hrs)</b>	TED talks
CoVoST	(Wang et al., 2020)	En→15 lang. (929hrs), 21 lang.→En (30-311hrs)	read, Common Voice
Europarl-ST	(Iranzo-Sanchez et al., 2020)	9 lang. (72 dir., 10-90hrs)	EP proceedings
LibriVoxDeEn	(Beilharz et al., 2020)	De→En 100hrs	read audiobooks
MaSS	(Zanon Boito et a., 2020)	8 lang. (56 dir.) 20hrs	Bible readings
BSTC	(Baidu, 2020)	Zh→En 50hrs	simult. interpret.
<b>Multilingual TEDx</b>	(Salesky et al., 2021)	<b>8 lang.→6 lang. 11-69hrs</b>	TED talks

*Trend (5): common test data across language pairs*

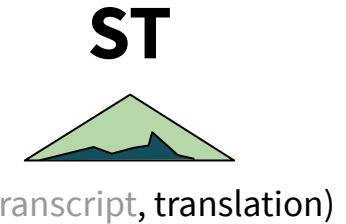
*Sec 3.2*

# Techniques

# Recap: Available data



Can we make use of this large amount of data?



# Multi-task learning

Definition:

*“Multi-task learning improves generalization by leveraging the domain-specific information contained in the training signals of related tasks”*

— Caruana, R. (1998)

# Transfer Learning

Definition:

*“Transfer learning and domain adaptation refer to the situation where what has been learned in one setting ... is exploited to improve generalization in another setting”*

— Page 526, [Deep Learning](#), 2016.

# Transfer Learning

Definition(2):

*“Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned.”*

— [Chapter 11: Transfer Learning, Handbook of Research on Machine Learning Applications, 2009.](#)

# Transfer Learning

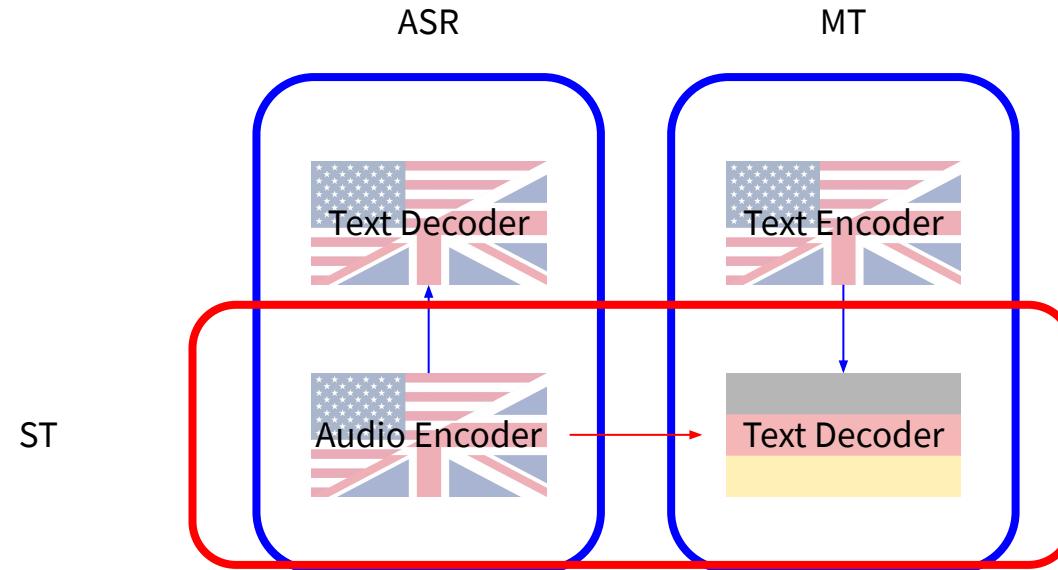
Definition:

*“In transfer learning, we first train a base network on a base dataset and task, and then we repurpose the learned features, or transfer them, to a second target network to be trained on a target dataset and task. This process will tend to work if the features are general, meaning suitable to both base and target tasks, instead of specific to the base task.”*

— [How transferable are features in deep neural networks?](#) Nips 2014

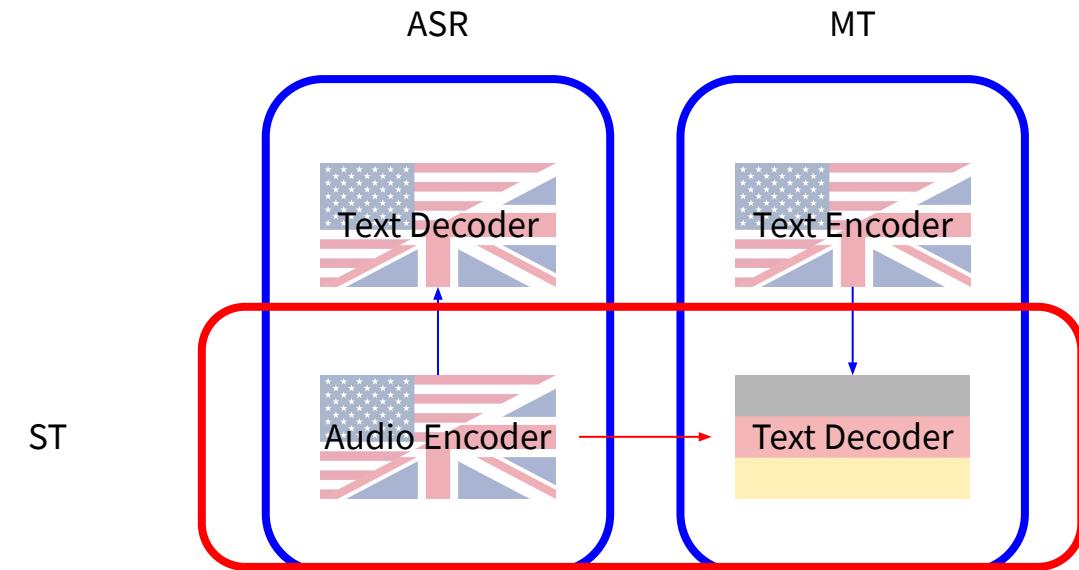
This form of transfer learning used in deep learning is called inductive transfer.

# Setting



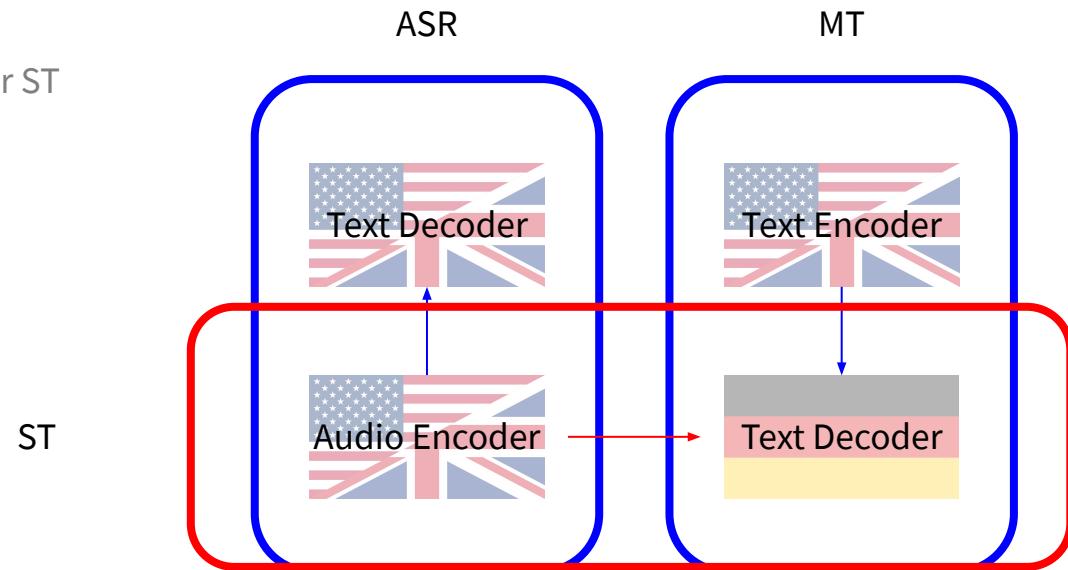
# Setting

- Multi-task
  - Train all three tasks jointly



# Setting

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

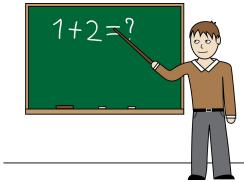


# Setting

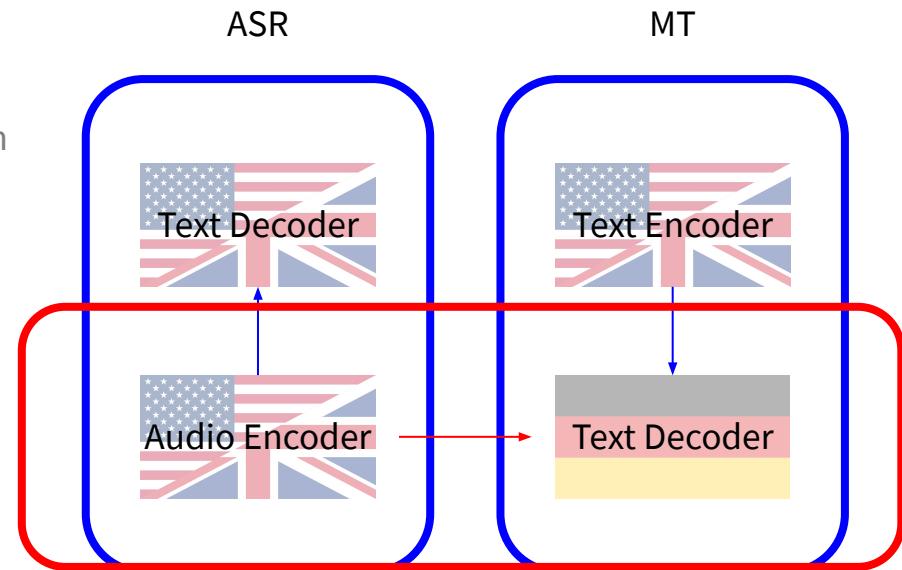
- Multi-task
- Pre-training
- Knowledge distillation
  - Take MT model
  - Train ST based on training signal from MT



ST



MT

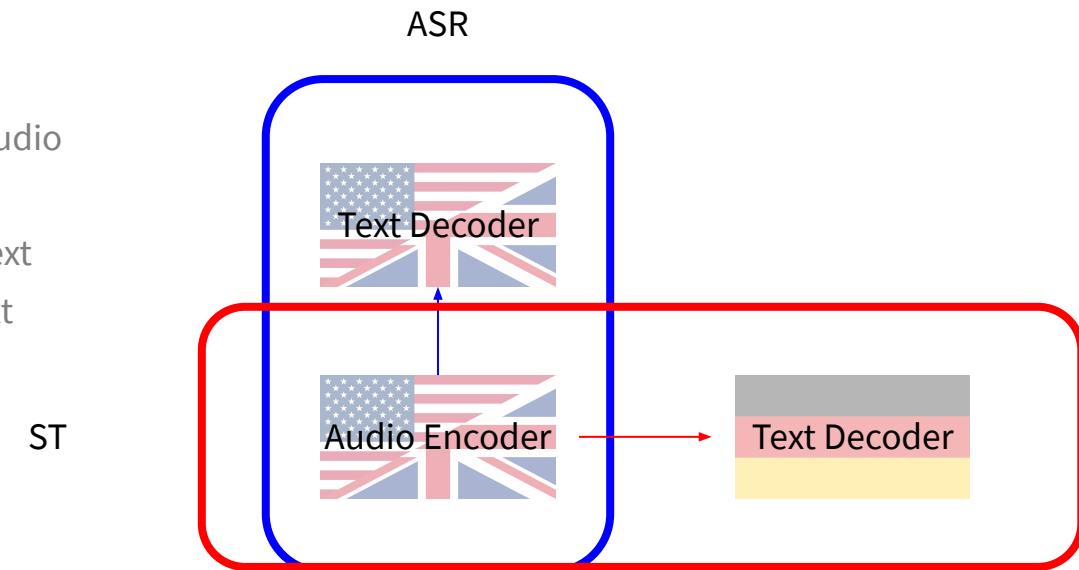


*Sec 3.2.1*

# Multi-task Learning

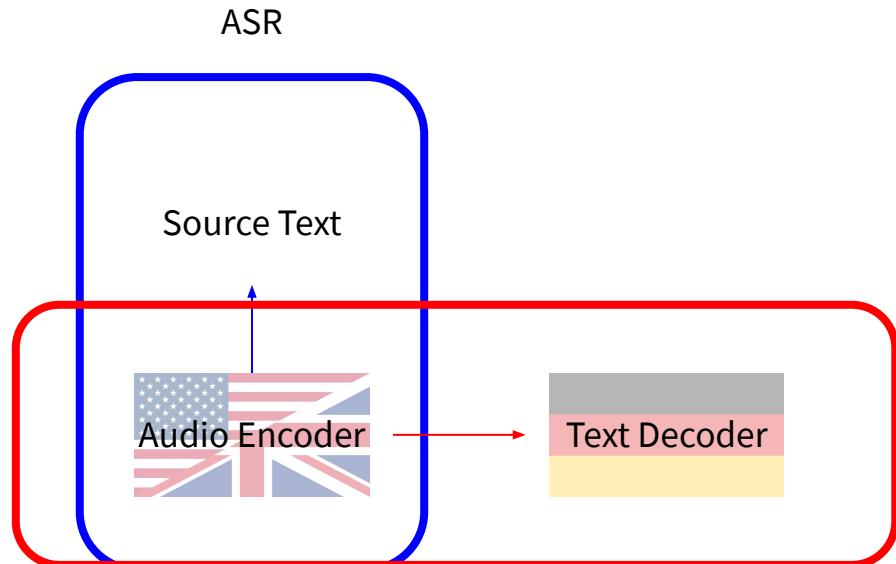
# Multi-task learning

- Baseline
  - No changes to the architecture
- ST+ASR
  - One encoder
    - Source Language audio
  - Two decoder
    - Source Language text
    - Target language text
  - (Weis et al, 2017)



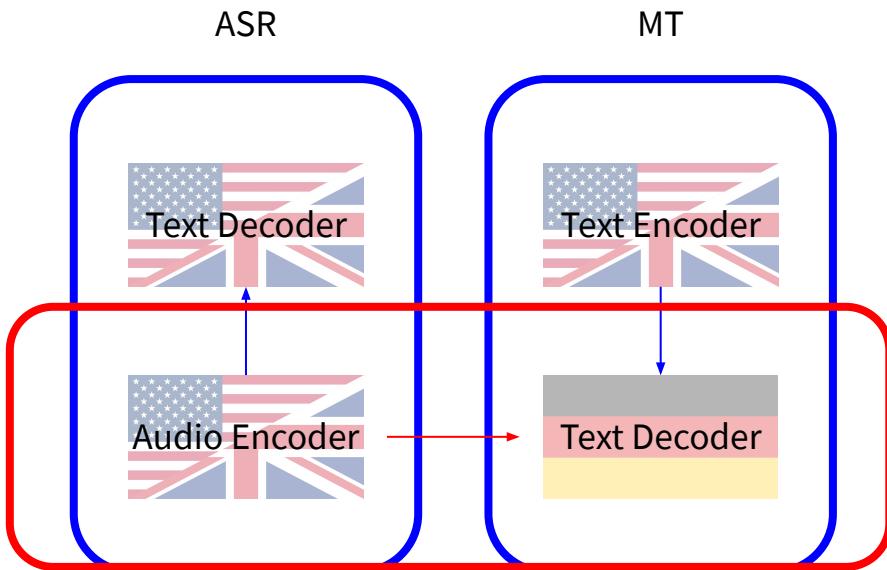
# Multi-task learning

- Baseline
  - No changes to the architecture
- ST+ASR
  - One encoder
    - Source Language audio
  - Two decoder
    - Source Language text
    - Target language text
  - (Weis et al, 2017)
- ASR using CTC loss on encoder
  - (Hori et al, 2017)
  - (Bahra et al, 2019)



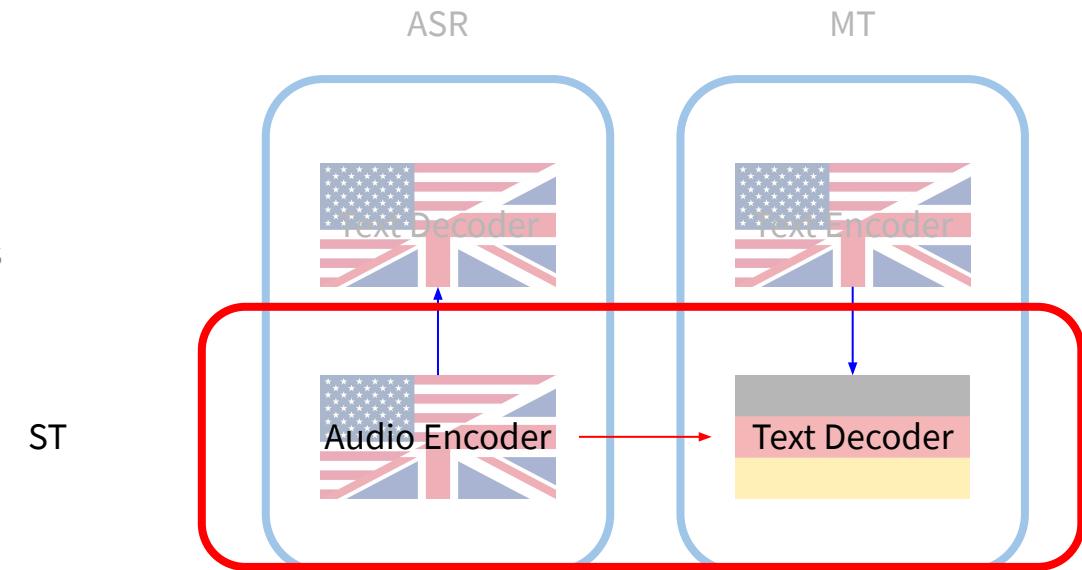
# Multi-task learning

- Baseline
  - No changes to the architecture
- ST+ASR
- ST+ASR+MT
  - Two encoder
    - Source Language audio
    - Source Language text
  - Two decoder
    - Source Language text
    - Target language text ST
  - (Berard et al, 2018)



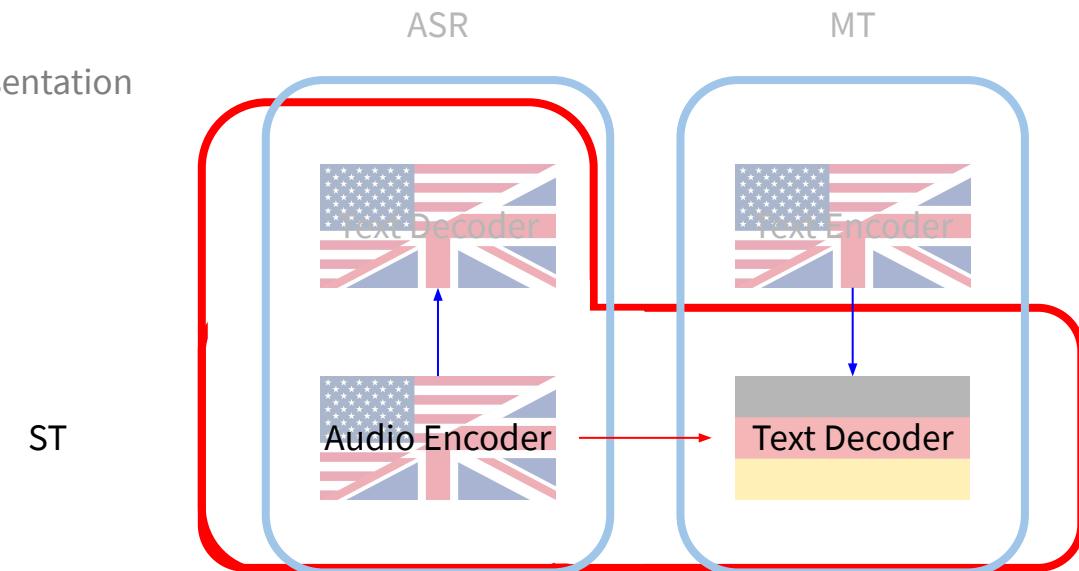
# Multi-task learning

- Baseline
  - No changes to the architecture
- ST+ASR
- ST+ASR+MT
- Inference:
  - Direct translation
  - No use of additional parts



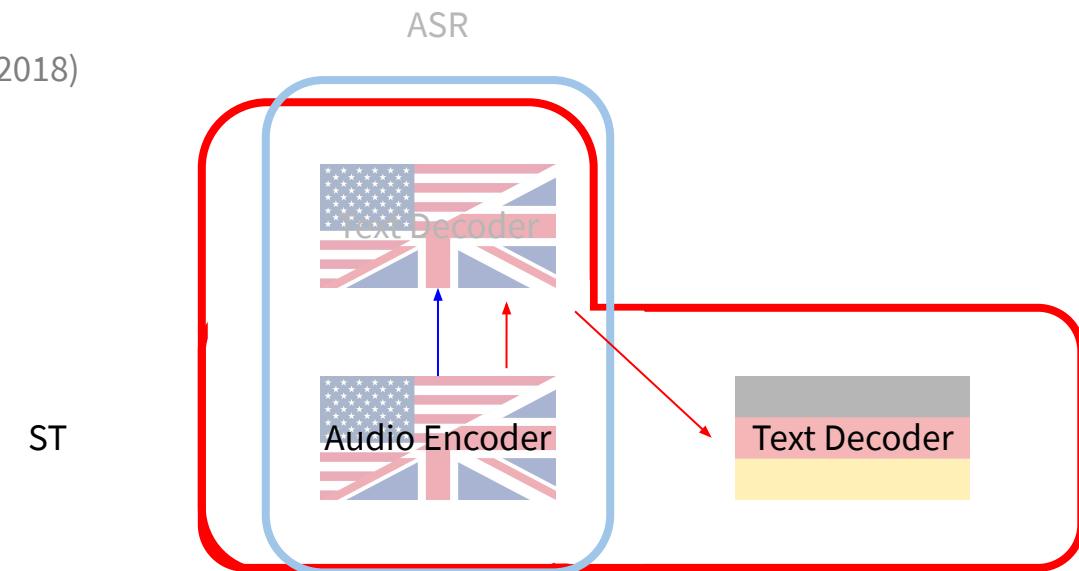
# 2-stage models

- Make use of additional model also during decoding
- *Simplify task*
  - using intermediate representation
- Comparison to cascade:
  - Full pipeline is trained
- Methods:
  - Adapt architecture
  - Preprocess data



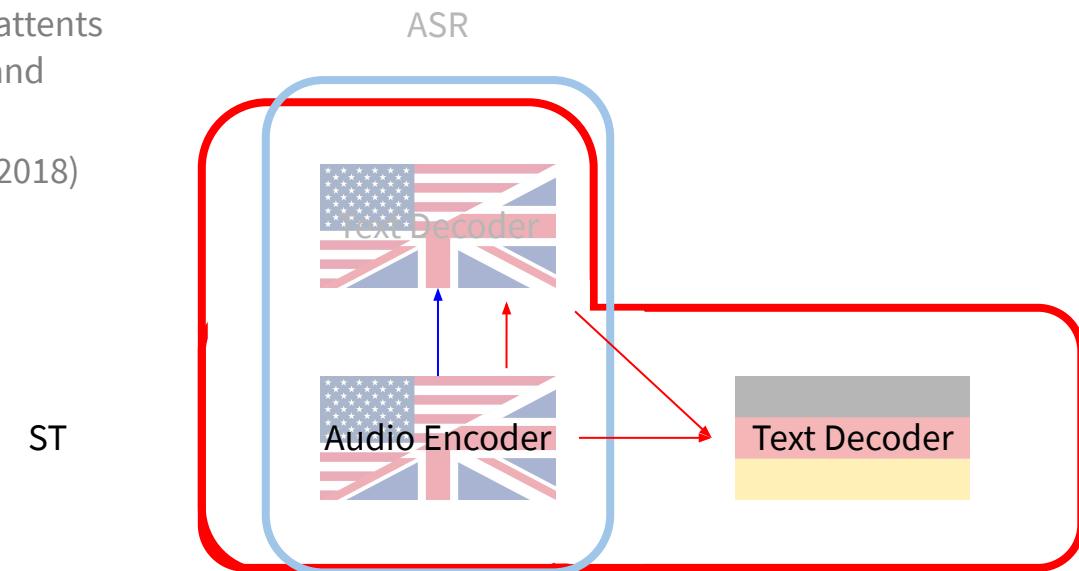
# 2-stage models

- Cascade:
  - Target language decoder attents to source text decoder
  - (Anastasopoulos Chiang, 2018)



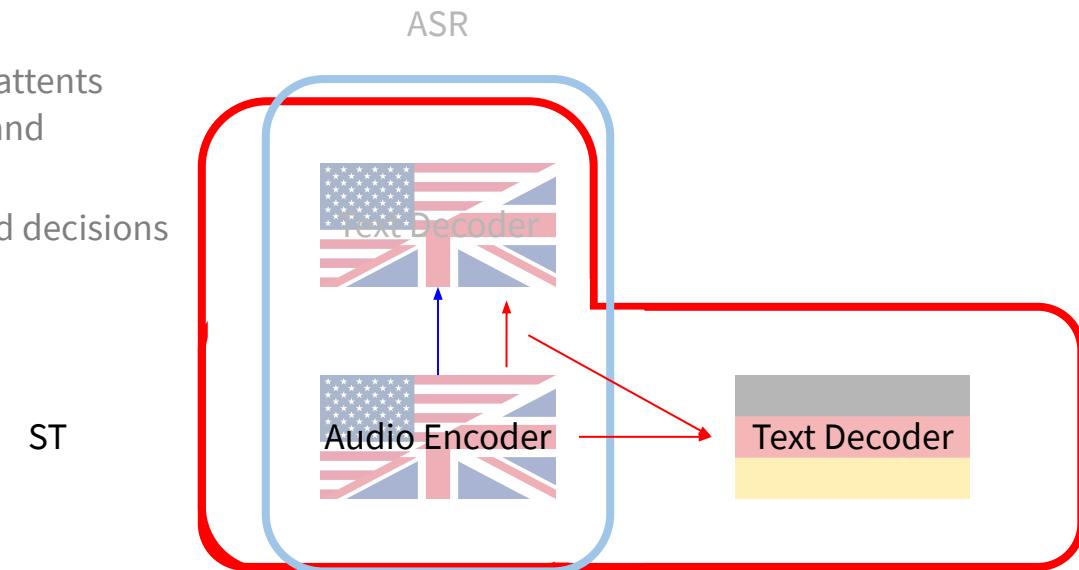
# 2-stage models

- Cascade:
- Triangle:
  - Target language decoder attents to source audio encoder and source text decoder
  - (Anastasopoulos Chiang, 2018)



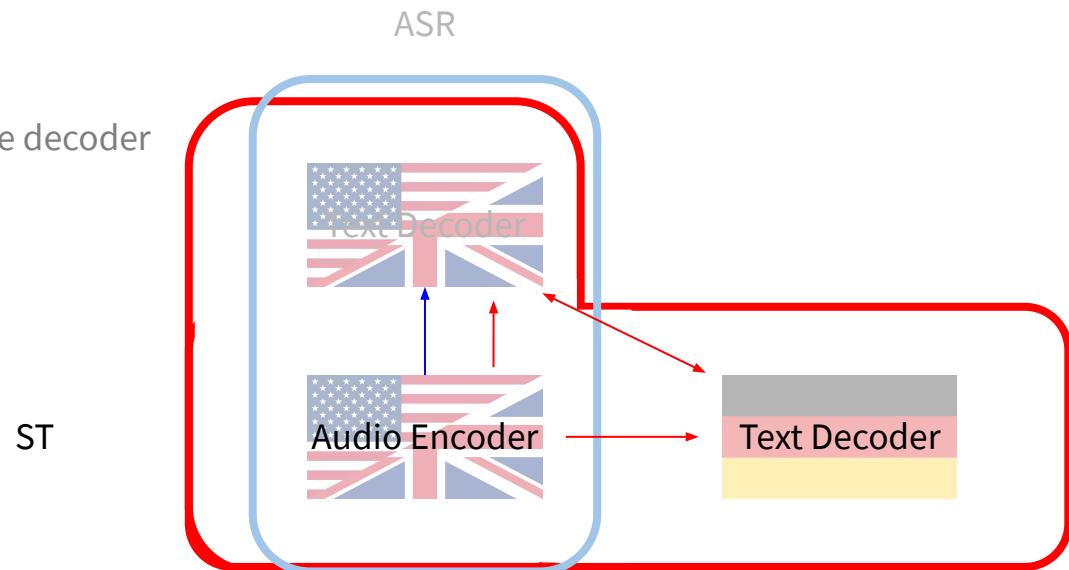
# 2-stage models

- Cascade:
- Triangle:
- Shared context vector
  - Target language decoder attends to source audio encoder and ASR context vectors
  - No direct influence of hard decisions of source text decoder
  - (Sperber et al, 2019)



# 2-stage models

- Cascade:
- Triangle:
- Shared context vector
- Dual Decoder
  - Source and target language decoder run in parallel
  - Attend to each other
  - (Le et al, 2020)

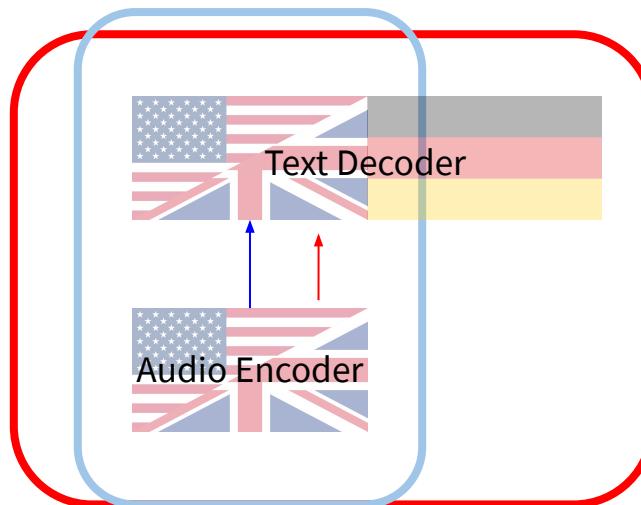


# 2-stage models

- Cascade:
- Triangle:
- Shared context vector
- Dual Decoder
- Concat
  - Single decoder generates source and target language
  - Output is concatenation
  - (Sperber et al, 2020)

ST

ASR



*Sec 3.2.2*

# Transfer Learning & Pretraining

-) pre-training encoder (<https://arxiv.org/abs/1809.01431>) Bansal, Sameer, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation." *arXiv preprint arXiv:1809.01431* (2018).

## Transfer learning and pre-training

-) decoder (<https://arxiv.org/pdf/1802.04200.pdf>) Bérard, A., Besacier, L., Kocabiyikoglu, A.C. and Pietquin, O., 2018, April. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6224-6228). IEEE.

-) KD (<https://arxiv.org/abs/1904.08075>) Kim, Yoon, and Alexander M. Rush. "Sequence-Level Knowledge Distillation." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317-1327. 2016.

Liu, Y., Xiong, H., Zhang, J., He, Z., Wu, H., Wang, H. and Zong, C., 2019. End-to-End Speech Translation with Knowledge Distillation}}}. *Proc. Interspeech 2019*, pp.1128-1132.

Confronto Gaido, M., Di Gangi, M.A., Negri, M. and Turchi, M., 2020. On Knowledge Distillation for Direct Speech Translation. *arXiv preprint arXiv:2012.04964*.

-) use of word2vc and BART (<https://arxiv.org/pdf/2010.12829.pdf>

<https://arxiv.org/pdf/2006.12124.pdf>) Wu, Anne, Changhan Wang, Juan Pino, and Jiatao Gu. "Self-Supervised Representations Improve End-to-End Speech Translation}})." *Proc. Interspeech 2020* (2020): 1491-1495.

Li, Xian, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. "Multilingual Speech Translation with Efficient Finetuning of Pretrained Models." *arXiv e-prints* (2020): arXiv-2010.

# How to leverage no-SLT data

- SLT data is of minimal quality compared to ASR and MT
- ASR and MT data are degrees of magnitude larger than the SLT data
- Pre-trained language models are an additional source of information
- Important question:

**How can the SLT use other sources of information?**

# How to leverage no-SLT data

- Transfer learning
- Pre-training SLT components:
  - encoder (Bansal et al., 2018),
  - decoder (Bérard et al., 2018)
- Knowledge distillation (Kim and Rush, 2016, Liu et al., 2019)

# XXXXXX

- Transfer learning
- Pre-training SLT components:
  - encoder (Bansal et al., 2018),
  - decoder (Bérard et al., 2018)

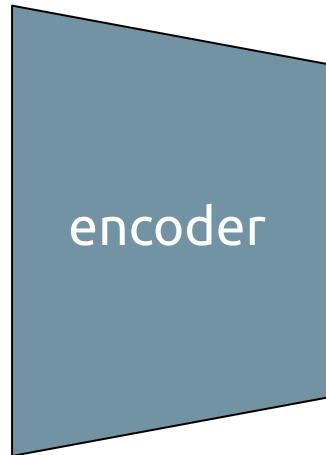
# Pre-training SLT components

Pre-training components of the SLT systems on different tasks

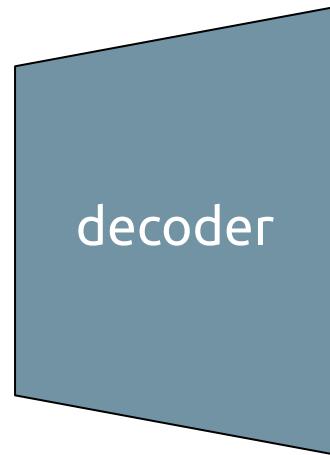
- Encoder pre-training (Bansal et al., 2018) <--> Automatic Speech Recognition
- Decoder pre-training (Bérard et al., 2018) <--> Machine Translation

# Encoder Pre-training

Spanish Audio



-	0.71
0.34	
-	0.12
-	0.51
0.05	
0.74	

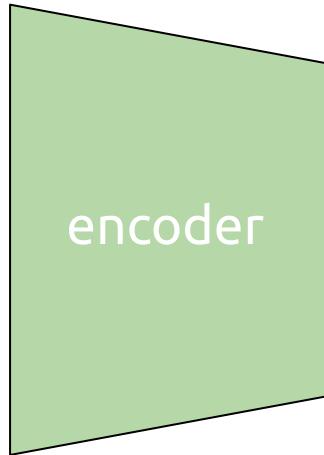


English text

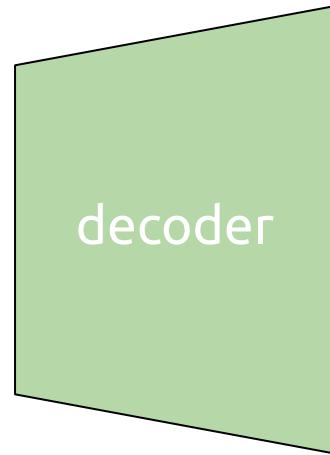
What a wonderful tutorial!

# Encoder Pre-training

Spanish Audio



- 0.71  
0.34  
- 0.12  
- 0.51  
0.05  
0.74



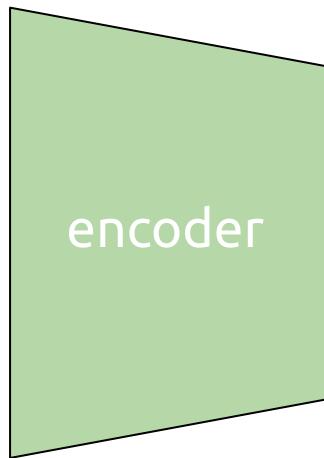
Spanish text

¡Qué maravilloso tutorial!

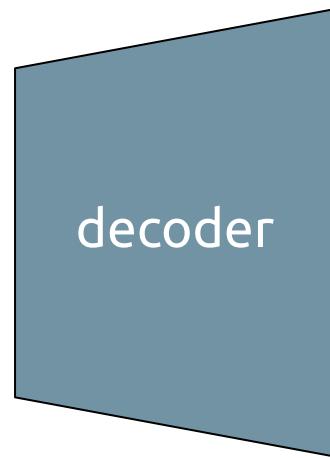
Training an ASR using the same SLT architecture

# Encoder Pre-training

Spanish Audio



- 0.71
0.34
- 0.12
- 0.51
0.05
0.74



English text

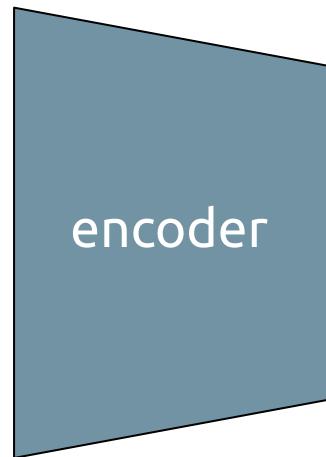
What a wonderful tutorial!

Training an ASR using the same SLT architecture

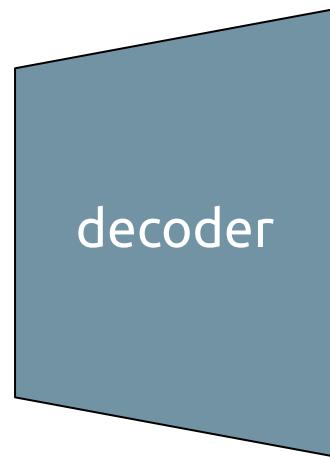
Training the SLT system initializing the encoder with the trained ASR encoder

# Decoder Pre-training

Spanish Audio



- 0.71  
0.34  
- 0.12  
- 0.51  
0.05  
0.74



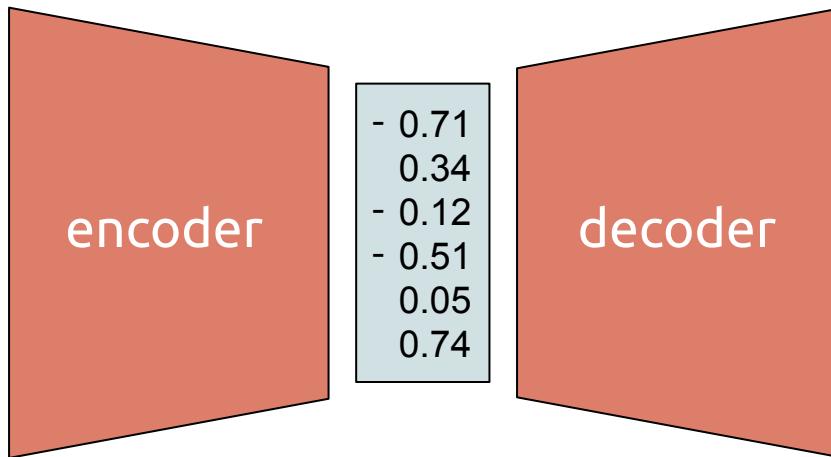
English text

What a wonderful tutorial!

# Decoder Pre-training

Spanish text

¡Qué maravilloso  
tutorial!



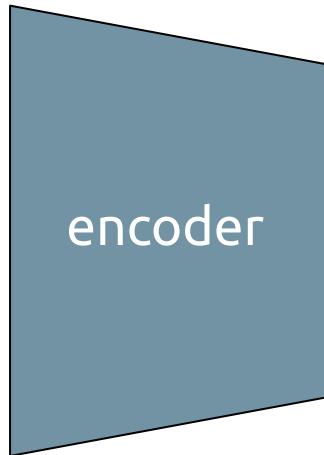
English text

What a wonderful  
tutorial!

Training an MT system using the same SLT architecture

# Decoder Pre-training

Spanish Audio



- 0.71  
0.34  
- 0.12  
- 0.51  
0.05  
0.74



English text

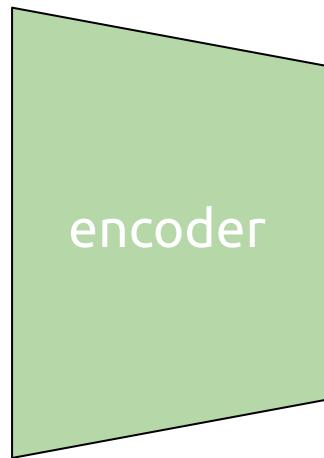
What a wonderful tutorial!

Training an MT system using the same SLT architecture

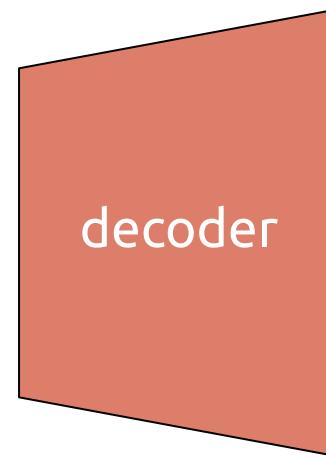
Training the SLT system initialising the decoder with the trained MT decoder

# Encoder-Decoder Pre-training

Spanish Audio



- 0.71  
0.34  
- 0.12  
- 0.51  
0.05  
0.74



English text

What a wonderful tutorial!

Training the SLT system initializing:

- the encoder with the trained ASR encoder
- the decoder with the trained MT decoder

# Exploiting unlabelled data

Following the trends in MT and text generation, exploiting unlabelled data

# Exploiting unlabelled data

Following the trends in MT and text generation, exploiting unlabelled data

Integration of:

- Encoder pre-training based on a general-purpose acoustic models: wav2vect (Ly et al., 2020)
- Decoder pre-training based on general-purpose language models: BERT or mBART (Wu et al., 2020)

# Exploiting unlabelled data

Following the trends in MT and text generation, exploiting unlabelled data

Integration of:

- Encoder pre-training based on a general-purpose acoustic models: wav2vect (Ly et al., 2020)
- Decoder pre-training based on general-purpose language models: BERT or mBART (Wu et al., 2020)

Useful in low-resourced and zero-shot conditions

*Sec 3.2.3*

# Knowledge Distillation

# Knowledge distillation

E2E SLT

# Knowledge distillation

**E2E SLT  
(Student)**

# Knowledge distillation

**E2E SLT  
(Student)**

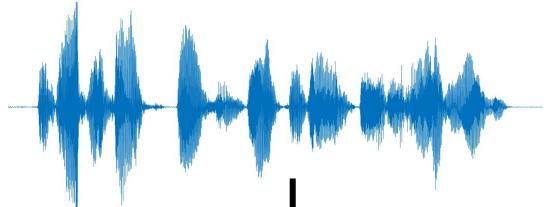
**MT**

# Knowledge distillation

E2E SLT  
(Student)

MT  
(Teacher)

# Knowledge distillation



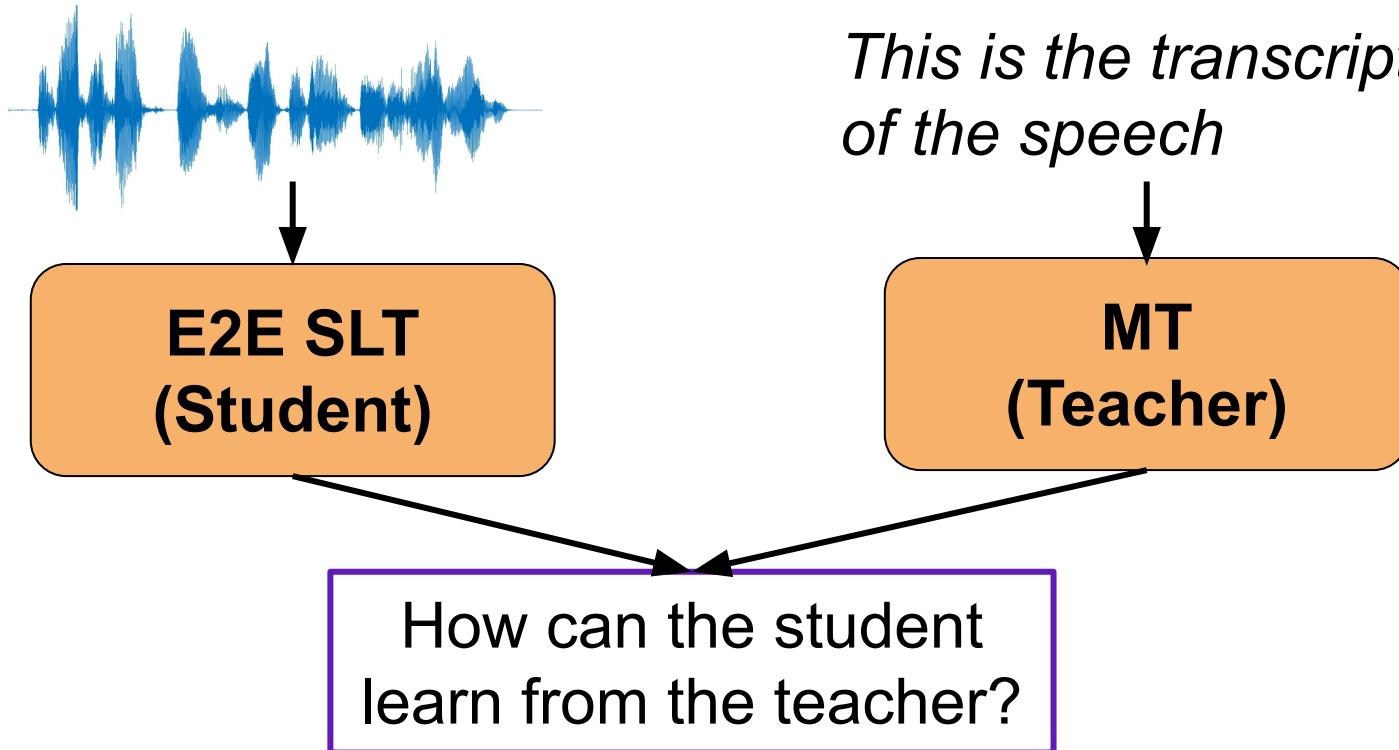
**E2E SLT  
(Student)**

*This is the transcript  
of the speech*



**MT  
(Teacher)**

# Knowledge distillation



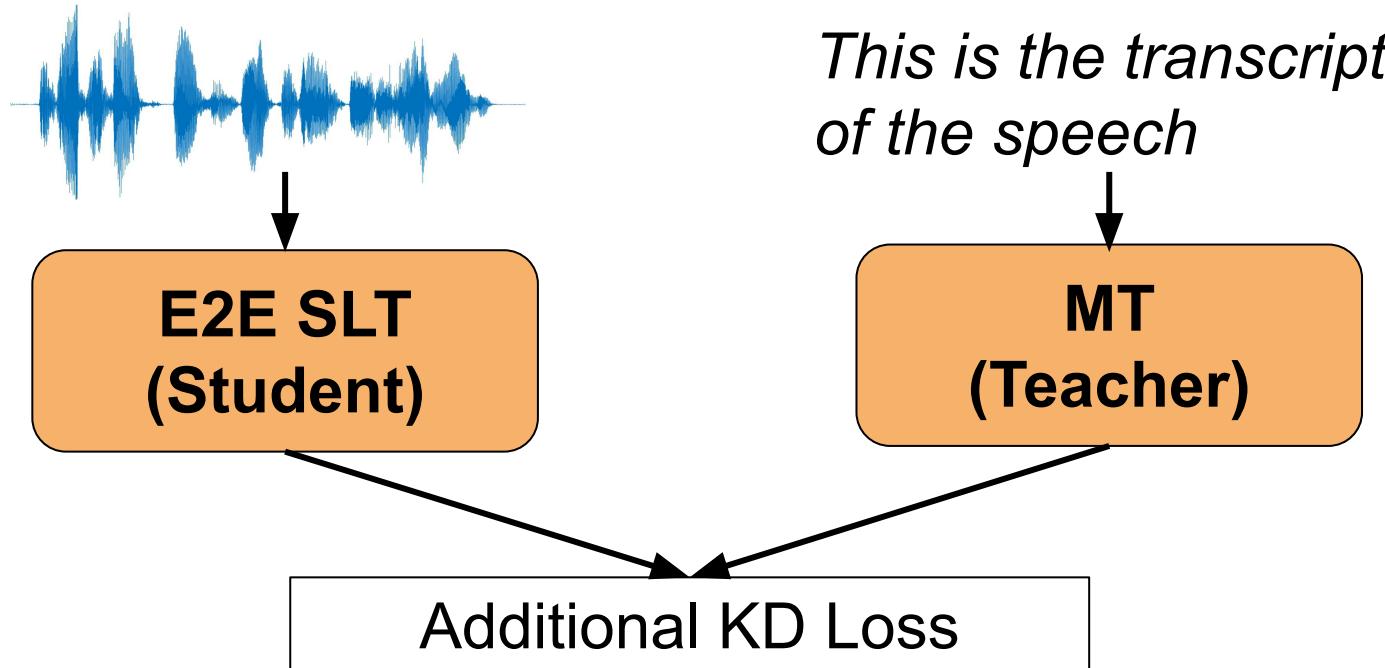
# Knowledge Distillation

Knowledge distillation for sequences (Kim and Rush, 2016)

- Word-Level KD
- Sequence KD
- Sequence Interpolation KD
- Requirements:
  - ASR data
  - Pre-trained MT system

# Word-Level KD

- Proposed by Liu et al. (2019)

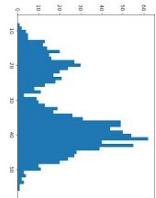


# Word-Level KD

E2E SLT  
(Student)

MT  
(Teacher)

During  
training

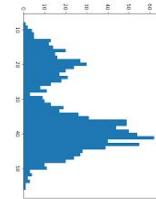
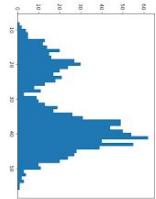


# Word-Level KD

E2E SLT  
(Student)

MT  
(Teacher)

During  
training

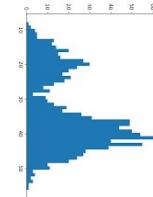
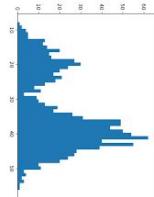


# Word-Level KD

E2E SLT  
(Student)

MT  
(Teacher)

During  
training



$$KL(ST_1, MT_1)$$

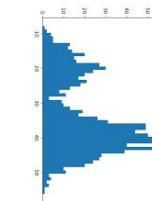
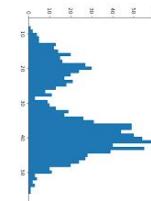
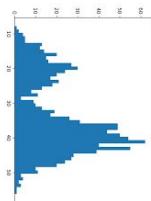
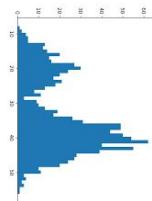


# Word-Level KD

E2E SLT  
(Student)

MT  
(Teacher)

During  
training



$$KL(ST_1, MT_1)$$



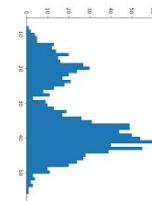
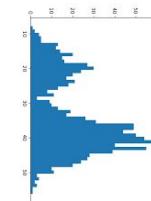
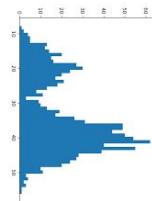
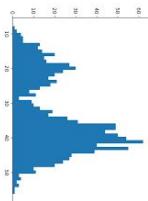
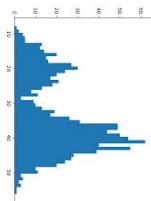
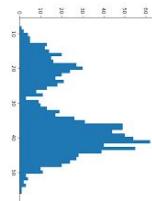
$$KL(ST_2, MT_2)$$

# Word-Level KD

E2E SLT  
(Student)

MT  
(Teacher)

During  
training



$$KL(ST_1, MT_1)$$



$$KL(ST_2, MT_2)$$



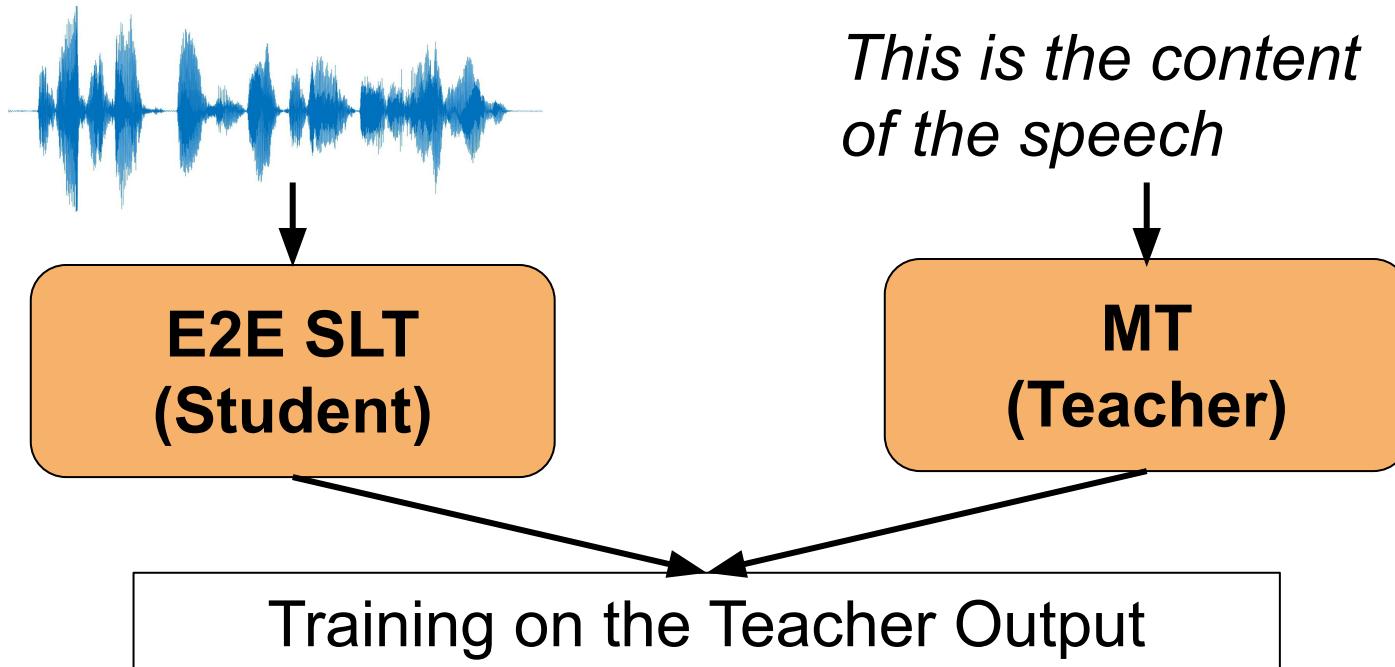
...

# Word-Level KD

- Training with SLT and KD losses
- Goal:
  - matching the output of SLT ground-truth
  - matching also the output probabilities of teacher model

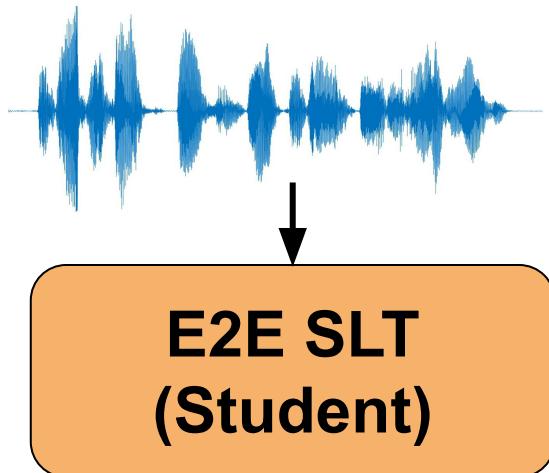
# Sequence Level KD (Seq-KD)

- The output of the teacher is used as reference

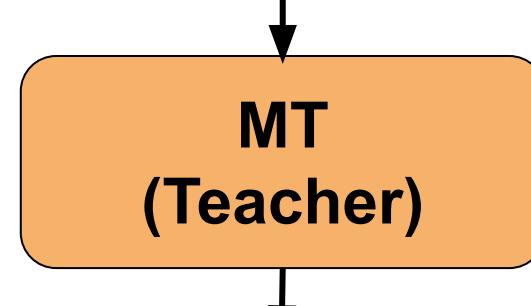


# Sequence Level KD (Seq-KD)

- The output of the teacher is used as reference



*This is the content  
of the speech*



Questo e' il contenuto  
del discorso

# Sequence Level KD (Seq-KD)

- The output of the teacher is used as reference

**E2E SLT  
(Student)**

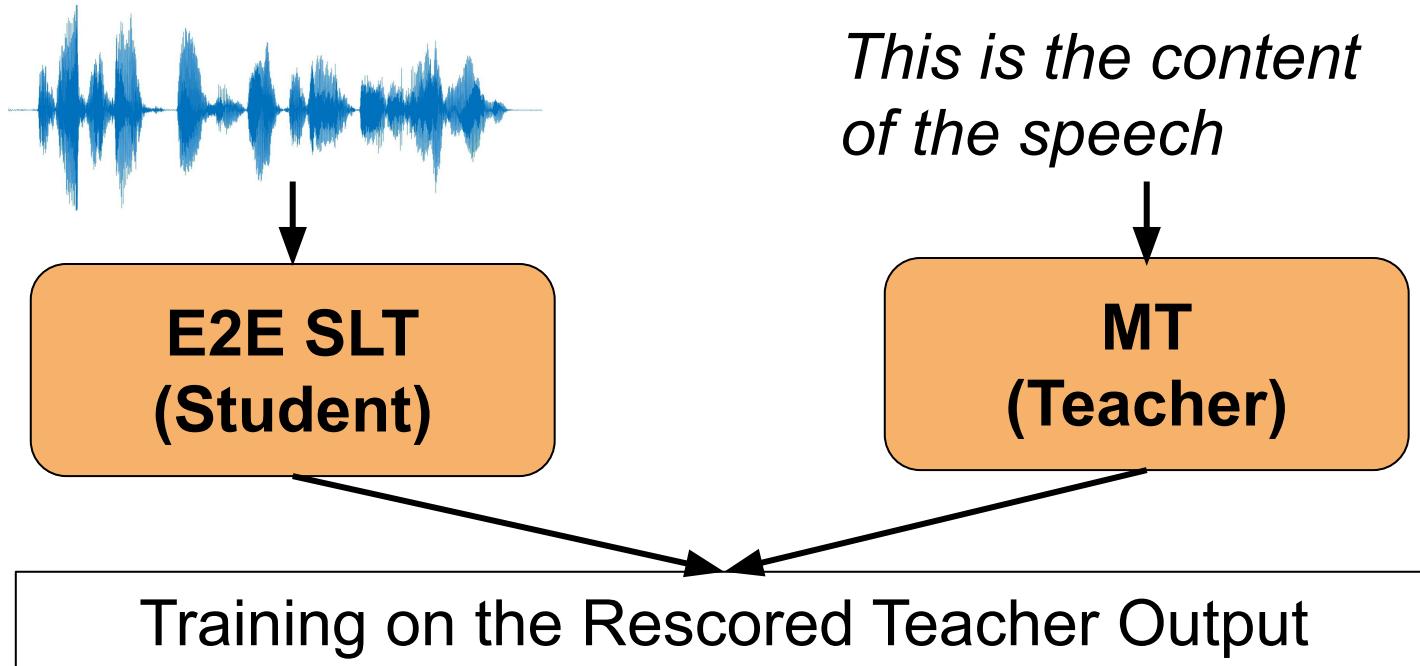
**MT  
(Teacher)**



Questo e' il contenuto  
del discorso

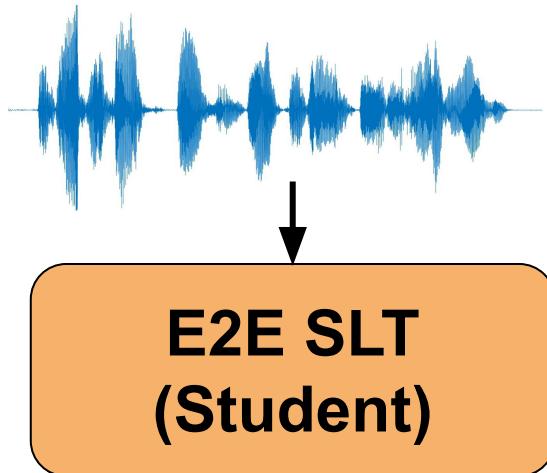
# Sequence Interpolation (Seq-Inter)

- The n-best of the teacher are rescored

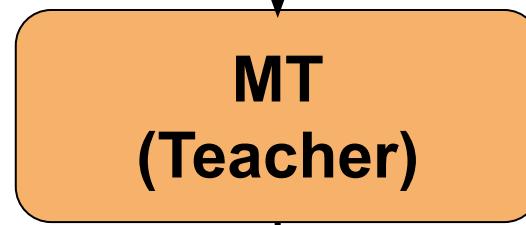


# Sequence Interpolation (Seq-Inter)

- The n-best of the teacher are rescored



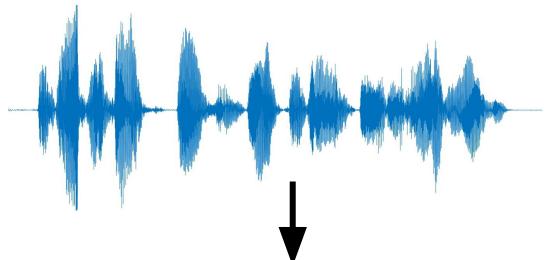
*This is the content  
of the speech*



Questo e' il contenuto del discorso  
Questo e' il contenuto dell'audio  
Questo e' il contenuto

# Sequence Interpolation (Seq-Inter)

- The n-best of the teacher are rescored



**E2E SLT  
(Student)**

*Re-ranked n-best*

*This is the content  
of the speech*

**MT  
(Teacher)**

Questo e' il contenuto dell'audio  
Questo e' il contenuto del discorso  
Questo e' il contenuto

# Sequence Interpolation (Seq-Inter)

- The n-best of the teacher are rescored

E2E SLT  
(Student)

MT  
(Teacher)



Questo e' il contenuto  
dell'audio

# Sequence Interpolation (Seq-Inter)

How to rescore:

- BLEU using SLT data for which there is the reference
- Other methods: e.g. quality estimation (using ASR data)

# Sequence Interpolation (Seq-Inter)

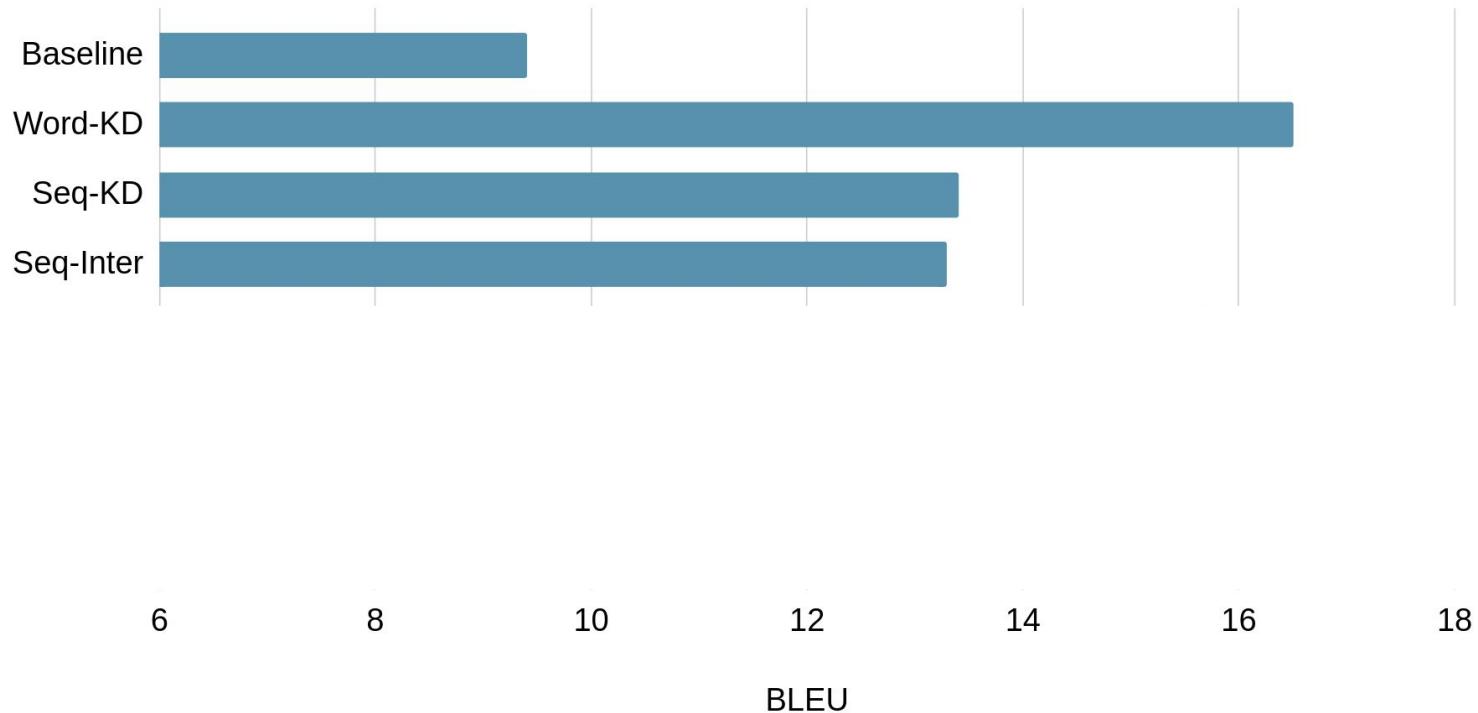
How to rescore:

- BLEU using SLT data for which there is the reference
- Other methods: e.g. quality estimation (using ASR data)

Goal:

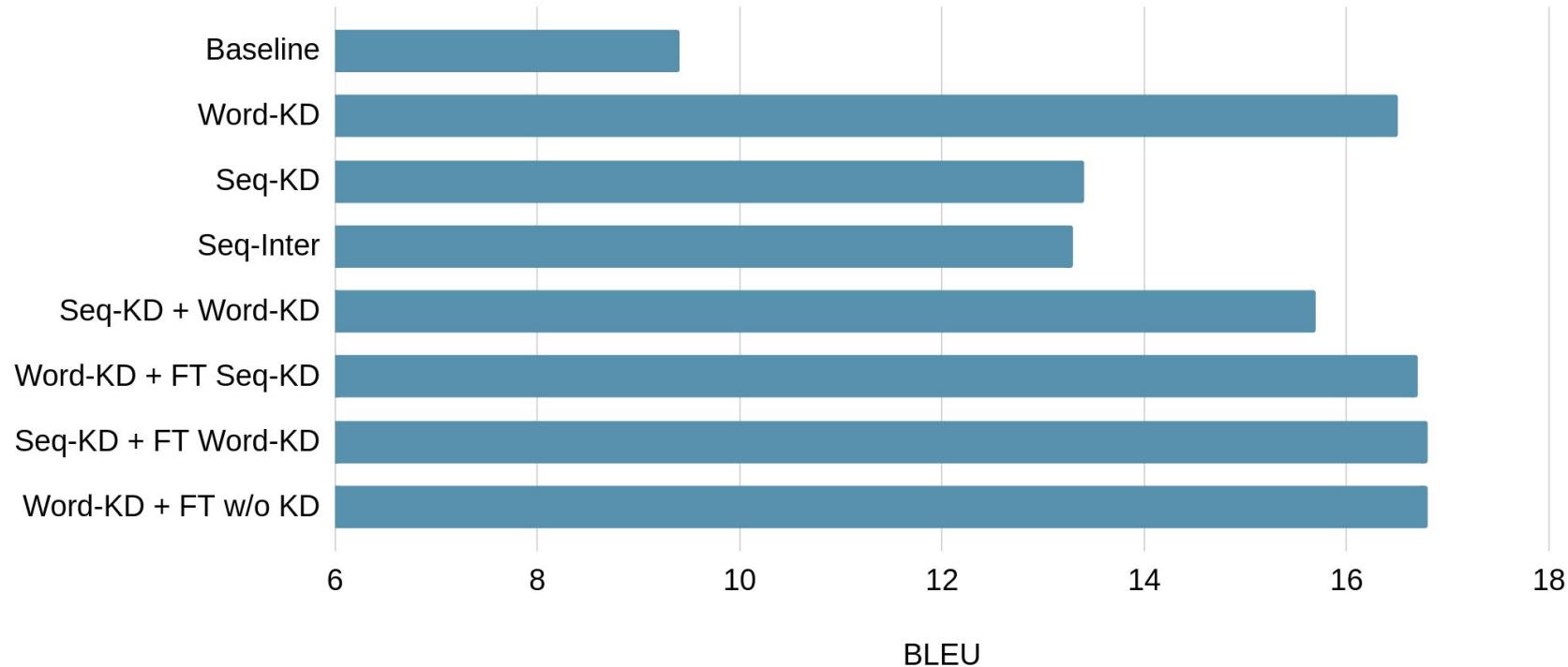
- To add knowledge from the teacher
- To reduce the lexical variability in the data (MT outputs have less variability)

# KD Methods (Gaido et al., 2020)



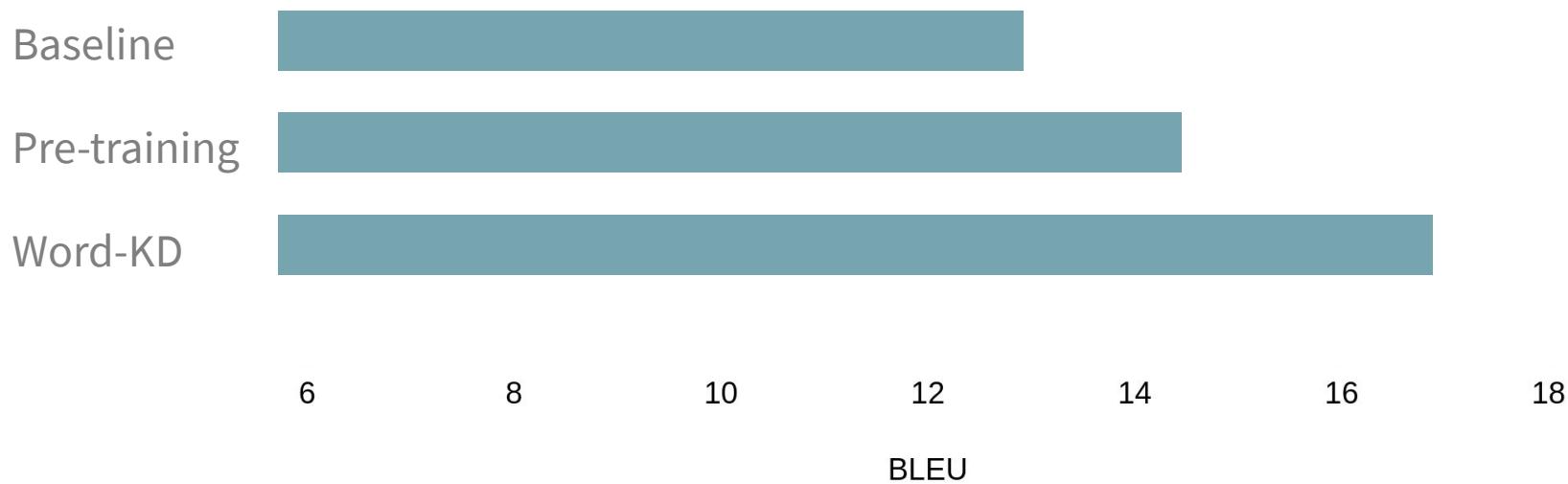
Word KD works the best

# KD Methods (Gaido et al., 2020)



Word KD with a fine-tuning slightly improves over word KD

# Pre-training vs KD (Liu et al., 2019)

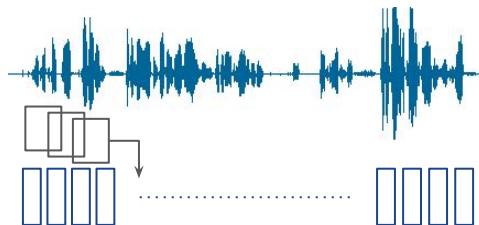


*KD outperforms pre-training*

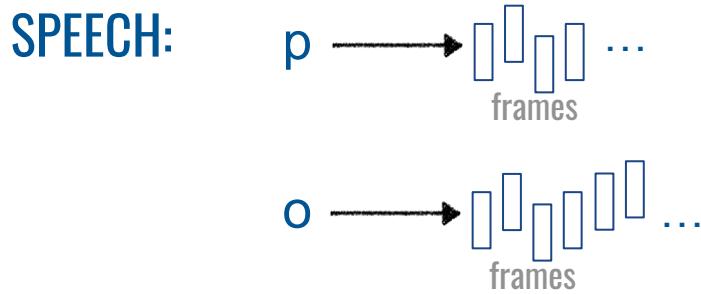
*Sec 3.3*

# Alternate Data Representations

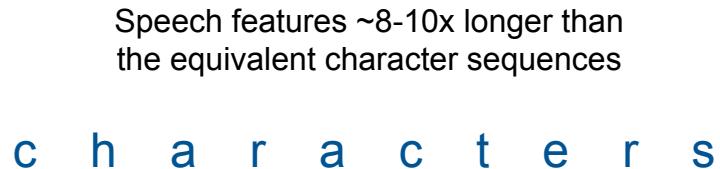
# [Recall] Speech vs. Text



Discretized audio — speech frames



Each feature vector is unique,  
Number of feature vectors per phone varies



Challenges:

- Sequence length
- Sequence redundancy
- Speech feature variation

# A Closer Look



speech features



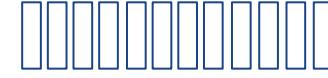
EH EH EH EH EH S S S S S S S S T T T AH AH AH AH

EH

S

T

AH



OHOHOHOH N N N N N N N N

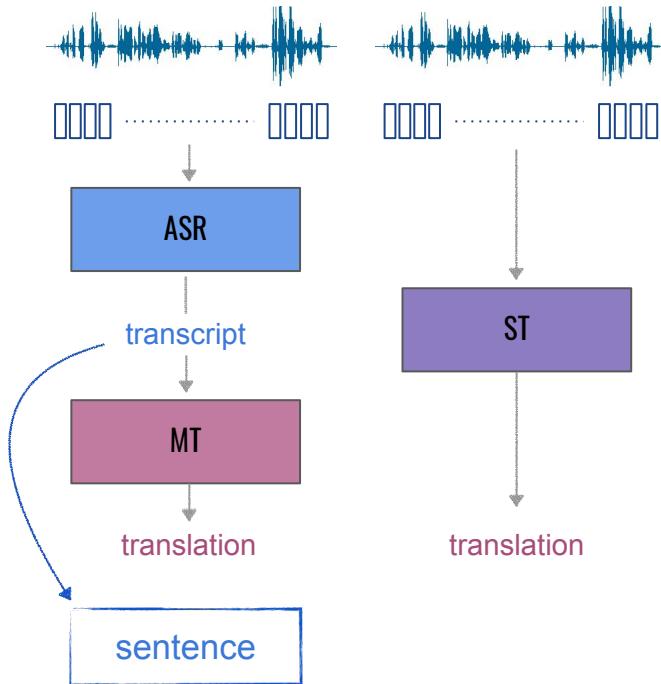
OH

N

[Esta es una oración]

# ST Architectures

CASCADE

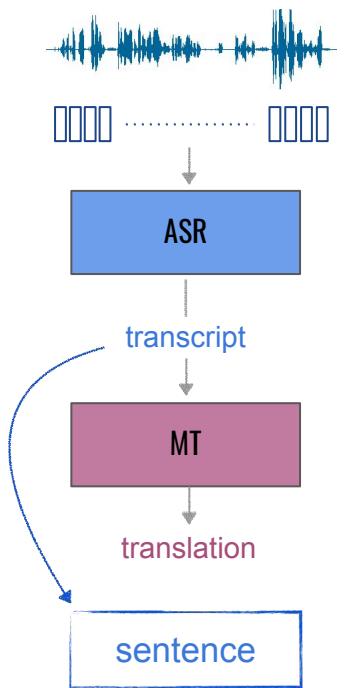


END-TO-END

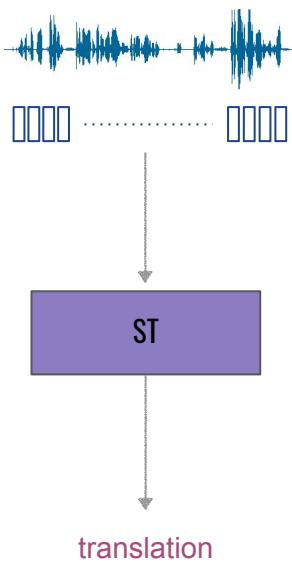


# ST Architectures

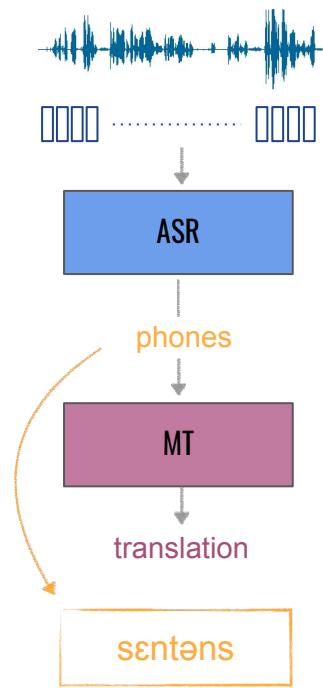
CASCADE



END-TO-END



Phone Cascade



*Recall: Redundancy*

Translating redundant phone sequences:

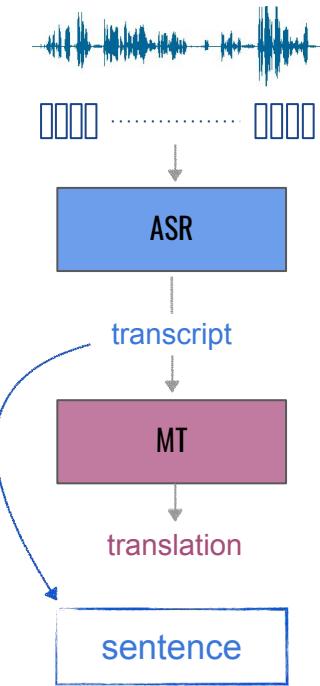
EH EH EH EH EH S S S S S S S S T T AH AH AH AH AH

performs 13% worse than uniques:

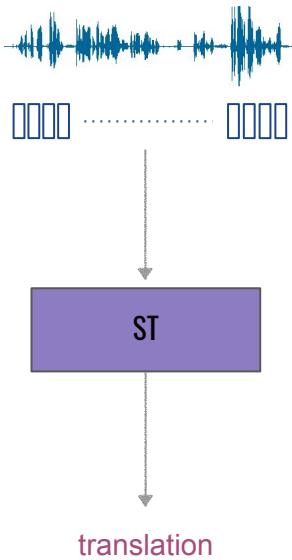
EH S T AH

# ST Architectures

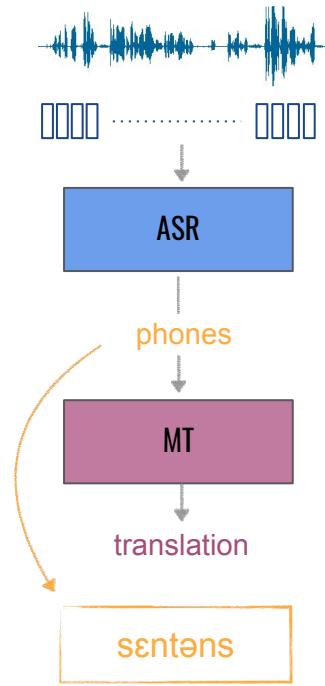
CASCADE



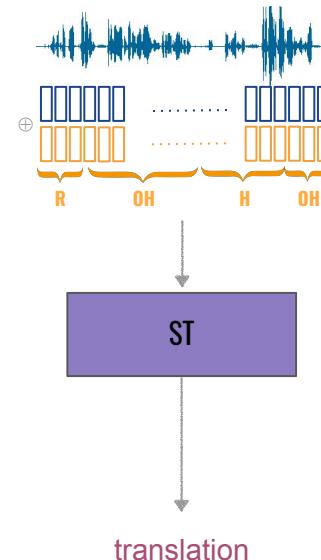
END-TO-END



Phone Cascade

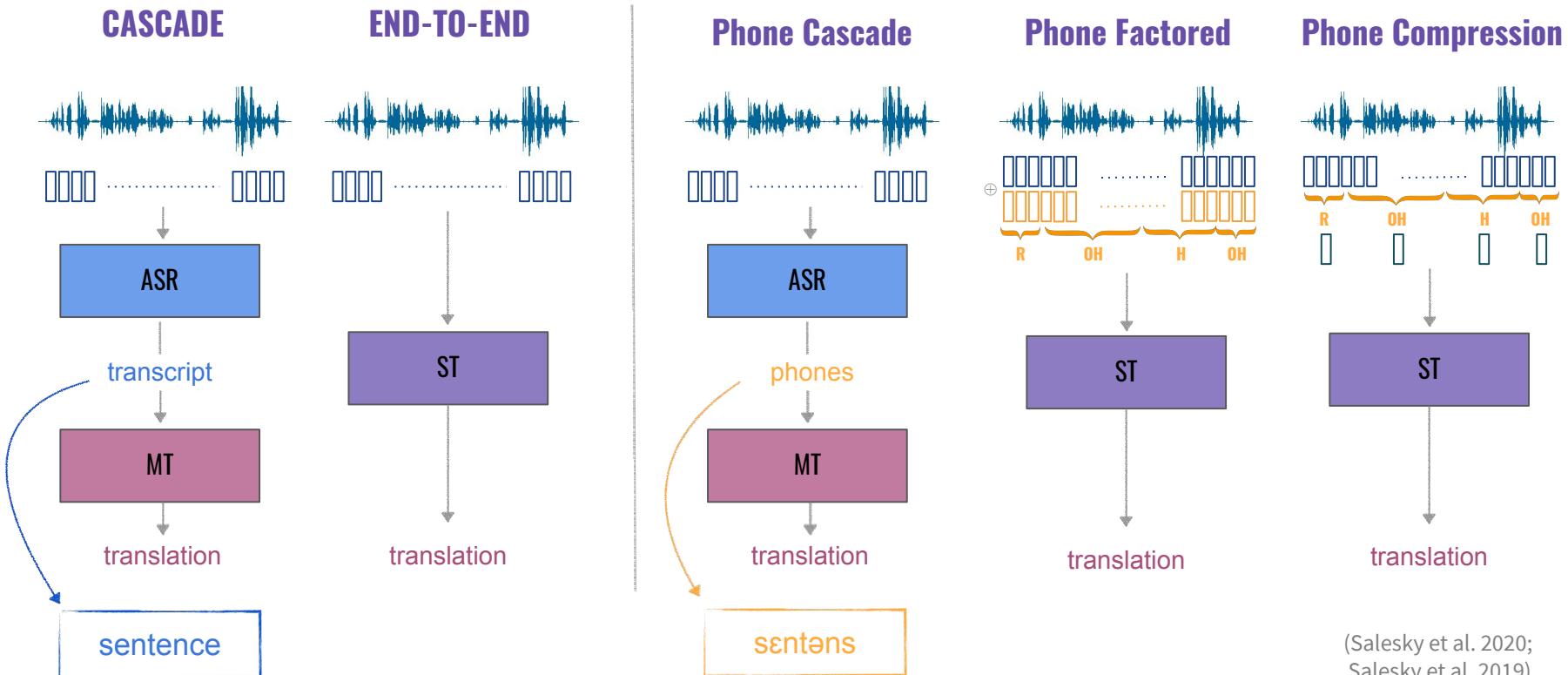


Phone Factored



(Salesky et al. 2020)

# ST Architectures

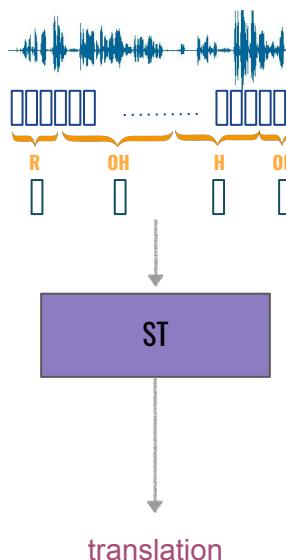


(Salesky et al. 2020;  
Salesky et al. 2019)

197

# Methods

## Phone Compression



### Detecting ‘phone’ units:

- ASR alignment\* (Salesky et al. 2019)
- Adaptive feature selection (AFS)\* (Zhang et al. 2020)
- CTC loss applied in encoder (Gaido et al. 2021)

\*require an additional model

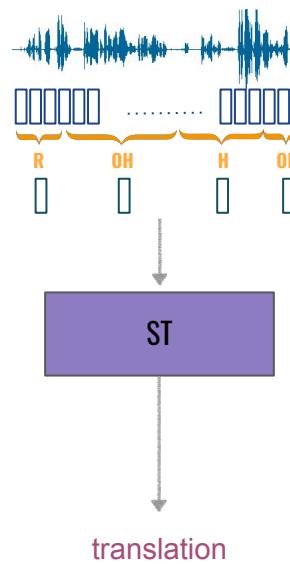
### Compression:

- Averaging
- Skip (select key-frame only)
- Softmax
- Weighted projection

(Salesky et al. 2019; Zhang et al. 2020;  
Gaido et al. 2021)

# Methods

## Phone Compression



How CTC collapsing works

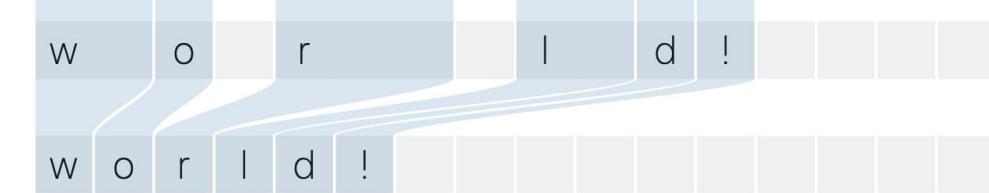
For an input,  
like speech



Predict a  
sequence of  
tokens



Merge repeats,  
drop ε



Final output



(Hannun et al. 2017) —  
<https://distill.pub/2017/ctc>

# Results

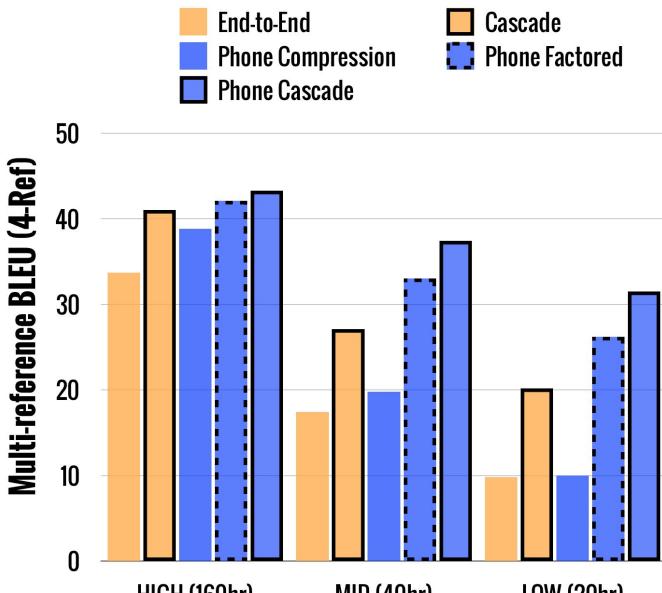
## Larger datasets

- Librispeech English—French
- MuST-C English—German+
- ~400 hours of speech with translations, transcripts

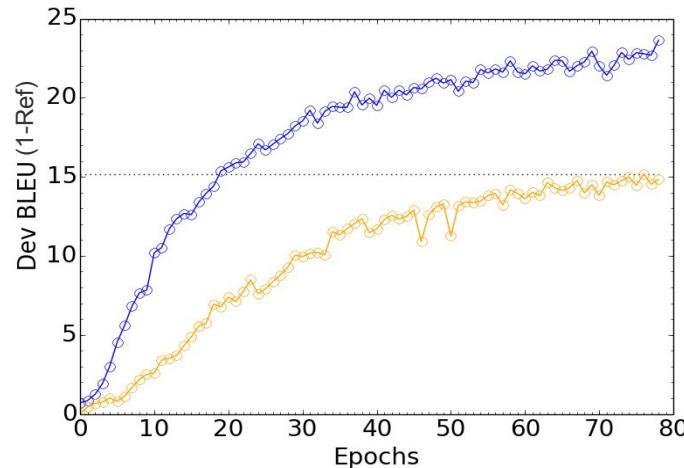
## Performance Improvements

- Improvements of 1-2 BLEU
- Computation reduction:
  - AFS: temporal reduction by 80%
  - CTC: overall computation reduced by ~10%
- Training and inference time reductions

# Results



Fisher Spanish—English  
(160 hours)



(Salesky et al. 2019; Salesky et al. 2020)

*Sec 4:*

# Evaluation

**Automatic Metrics**

**Utterance segmentation**

**Mitigating error due to speaker variation**

*Sec 4.1*

# Automatic Metrics

# Evaluation

- Motivated by evaluation in machine translation
  - Automatic evaluation
    - Cheap
    - Fast
  - Human evaluation
    - Gold standard
    - Subjective
    - Expensive, time-consuming

# Automatic metrics

- Reuse Text MT-based metrics
  - *BLEU*
    - Compare reference translation to output
- Multi-task system
  - *Word error rate (WER)* of transcription
    - Single correct output
    - Often calculated ignoring punctuation and case

# BLEU

- Compare Hypothesis to reference translation
  - Geometric mean of n-gram precision (1 to 4-grams)
  - Using case- and punctuation information

Reference: BLEU is a MT metric

Hypothesis: BLEU is my metric

# BLEU

- Compare Hypothesis to reference translation
  - Geometric mean of n-gram precision (1 to 4-grams)
  - Using case- and punctuation information

Reference: BLEU is a MT metric

Hypothesis: BLEU is my metric

1-gram: 3/4

2-gram: 1/3

3-gram: 0/2

4-gram: 0/1

$$\text{BLEU} = \sqrt[4]{3/4 * 1/3 * 0 * 0 * \text{BP}}$$

# BLEU

- Compare Hypothesis to reference translation
  - Geometric mean of n-gram precision (1 to 4-grams)
  - Using case- and punctuation information
- Aggregated scores over large dataset
- “*Brevity penalty*” to account for recall

Reference: BLEU is a MT metric

Hypothesis: BLEU is my metric

1-gram: 3/4

2-gram: 1/3

3-gram: 0/2

4-gram: 0/1

$$\text{BLEU} = \sqrt[4]{3/4 * 1/3 * 0 * 0 * \text{BP}}$$

# Word error rate (WER)

- Align reference and hypothesis
    - Calculate insertions, deletions and substitutions
    - Divide by reference length
  - Often ignoring case and punctuation
- Reference: WER is an ASR metric
- Hypothesis: WER is my \*\*\* metric

# Word error rate (WER)

- Align reference and hypothesis
    - Calculate insertions, deletions and substitutions
    - Divide by reference length
  - Often ignoring case and punctuation
- Reference: WER is an ASR metric
- Hypothesis: WER is my \*\*\* metric
- Alignment: S D

# Word error rate (WER)

- Align reference and hypothesis
  - Calculate insertions, deletions and substitutions
  - Divide by reference length
- Often ignoring case and punctuation

Reference: WER is an ASR metric

Hypothesis: WER is my \*\*\* metric

Alignment:        S      D

$$\text{WER} = \frac{S+D+I}{N} = \frac{2}{5}$$

*Sec 4.2*

# Utterance Segmentation

# Utterance segmentation

Bla

How to evaluate different STL systems having different segmentations:

-) Matusov approach

(<https://www-i6.informatik.rwth-aachen.de/publications/download/344/Matusov-IWSLT-2005.pdf>)

Matusov, Evgeny, Gregor Leusch, Oliver Bender, and Hermann Ney. "Evaluating machine translation output with automatic sentence segmentation." In *International Workshop on Spoken Language Translation (IWSLT) 2005*. 2005.

-) Concatenation

-) ...

# **Utterance segmentation**

SLT evaluation has an additional level of complexity compared to machine translation.

# Utterance segmentation

SLT evaluation has an additional level of complexity compared to machine translation.

Machine Translation:

Document:

This is an audio signal.  
In the training data it  
was split using strong  
punctuation. Three  
sentences in total.

# Utterance segmentation

SLT evaluation has an additional level of complexity compared to machine translation.

Machine Translation:

Document:

This is an audio signal.  
In the training data it  
was split using strong  
punctuation. Three  
sentences in total.

Source sentences:

This is an audio signal.

In the training data it was split  
using strong punctuation.

Three sentences in total!

# Utterance segmentation

SLT evaluation has an additional level of complexity compared to machine translation.

Machine Translation:

Document:

This is an audio signal.  
In the training data it  
was split using strong  
punctuation. Three  
sentences in total.

Source sentences:

This is an audio signal.

In the training data it was split  
using strong punctuation.

Three sentences in total!

Reference sentence:

Questo e' un segnale audio.

Nei dati di training e' stato  
diviso usando la  
punteggiatura forte.

Tre frasi in totale!

# Utterance segmentation

SLT evaluation has an additional level of complexity compared to machine translation.

Machine Translation:

Source sentences:

This is an audio signal.

In the training data it was split using strong punctuation.

Three sentences in total!

MT sentences:

Questo è un segnale audio.

Nei dati di allenamento è stato suddiviso utilizzando una forte punteggiatura.

3 frasi in totale!

Reference sentence:

Questo e' un segnale audio.

Nei dati di training e' stato diviso usando la punteggiatura forte.

Tre frasi in totale!

# Utterance segmentation

SLT evaluation has an additional level of complexity compared to machine translation.

Machine Translation:

Source sentences:

This is an audio signal.

In the training data it was split using strong punctuation.

Three sentences in total!

MT sentences:

Questo è un segnale audio.

Nei dati di allenamento è stato suddiviso utilizzando una forte punteggiatura.

3 frasi in totale!

Reference sentence:

Questo e' un segnale audio.

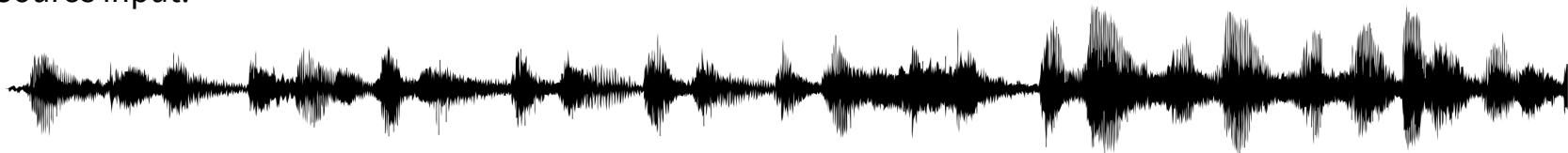
Nei dati di training e' stato diviso usando la punteggiatura forte.

Tre frasi in totale!

# Utterance segmentation

Spoken Language Translation:

Source input:

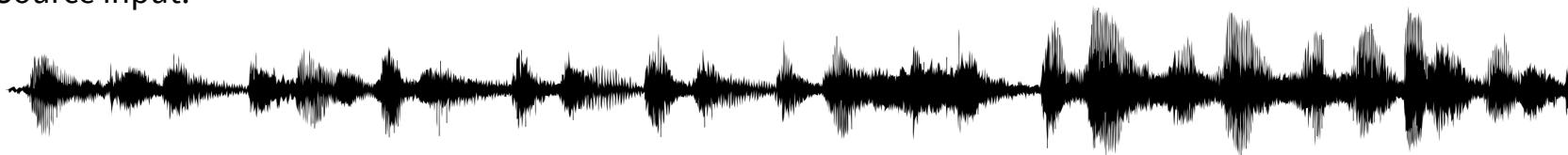


this is an audio signal in the training data it was split using strong punctuation into three sentences in total

# Utterance segmentation

Spoken Language Translation:

Source input:



this is an audio signal in the training data it was split using strong punctuation three sentences in total

Reference sentences:

This is an audio signal.

In the training data it was split using strong punctuation.

Three sentences in total!



# Utterance segmentation

SLT outputs depend on the segmentation of the audio input:

This is an audio

Signal in the training data was split.

Using strong punctuation, 3 sentences in total!

# Utterance segmentation

SLT outputs depend on the segmentation of the audio input:

This is an audio

Signal in the training data was split.

Using strong punctuation, 3 sentences in total!

This is an audio signal in the training data.

It was split using strong punctuation.

Three sentences in total!

# Utterance segmentation

SLT outputs depend on the segmentation of the audio input:

This is an audio

Signal in the training data was split.

Using strong punctuation, 3 sentences in total!

This is an audio signal in the training data.

It was split using strong punctuation.

Three sentences in total!

This is a signal.

In the training data.

It was split in three sentences.

# Utterance segmentation

SLT outputs depend on the segmentation of the audio input:

This is an audio

Signal in the training data was split.

Using strong punctuation, 3 sentences in total!

This is an audio signal in the training data.

It was split using strong punctuation.

Three sentences in total!

This is a signal.

In the training data.

It was split in three sentences.

This is

Signal. In the training data

it was split using strong punctuation.

Three sentences

in total!

# Utterance segmentation

SLT outputs depend on the segmentation of the audio input:

This is an audio

Signal in the training data was split.

Using strong punctuation, 3 sentences in total!

This is an audio signal in the training data.

It was split using strong punctuation.

Three sentences in total!

This is a signal.

In the training data.

It was split in three sentences.

This is

Signal. In the training data

it was split using strong punctuation.

Three sentences

in total!

Reference sentences:

This is an audio signal.

In the training data it was split using strong punctuation.

Three sentences in total! 226

# **SLT output - reference alignment**

1. How to compare the automatically split SLT outputs with the manually split references?
2. How to compare different systems splitting the SLT outputs in different ways?

# SLT output - reference alignment

1. How to compare the automatically split SLT outputs with the manually split references?
2. How to compare different systems splitting the SLT outputs in different ways?

Issues:

- Different number of sentences
- Truncated SLT sentences
- Insertion of additional text in the SLT outputs
- Missing large parts in the SLT outputs

# Concatenation

SLT output:

This is

Signal. In the training data

it was split using strong punctuation.

Three sentences

in total!

Reference sentences:

This is an audio signal.

In the training data it was split using strong punctuation.

Three sentences in total!

# Concatenation

SLT output:

This is Signal . In the training data it was split using strong punctuation . Three sentences in total !

Reference sentences:

This is an audio signal . In the training data it was split using strong punctuation . Three sentences in total !

The concatenated STL outputs (references) are considered as a single sentence.

Automatic metrics applied on two strings.

Much less precise than working at segment level, but fast to implement

# Automatic re-segmentation algorithm

SLT output:

This is Signal . In the training data it was split using strong punctuation . Three sentences in total !

Reference sentences:

This is an audio signal . In the training data it was split using strong punctuation . Three sentences in total!

# Automatic re-segmentation algorithm

SLT output:

This is Signal . In the training data it was split using strong punctuation . Three sentences in total !

Reference sentences:

This is an audio signal . <eos> In the training data it was split using strong punctuation . <eos> Three sentences in total ! <eos>

# Automatic re-segmentation algorithm

This is Signal . In the training data it was split using strong punctuation . Three sentences in total !

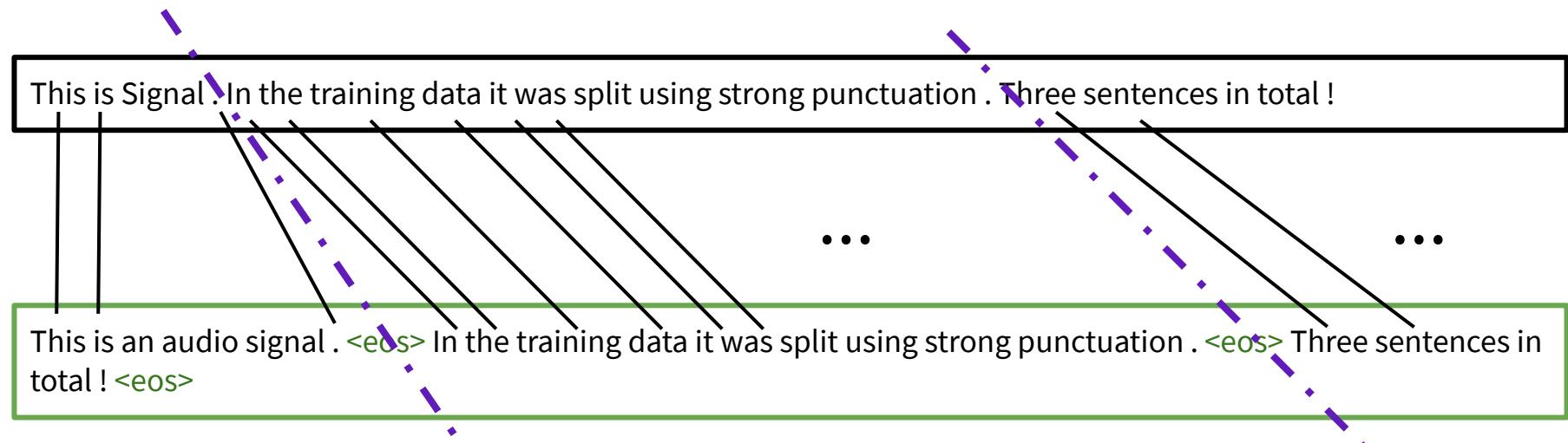
This is an audio signal . <eos> In the training data it was split using strong punctuation . <eos> Three sentences in total ! <eos>

# Automatic re-segmentation algorithm

This is Signal : In the training data it was split using strong punctuation . Three sentences in total !

This is an audio signal . <eos> In the training data it was split using strong punctuation . <eos> Three sentences in total ! <eos>

# Automatic re-segmentation algorithm



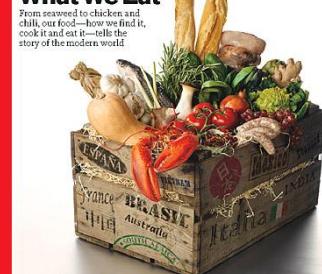
Based on the word alignments and <eos>, the SLT output and reference are segmented.

Alignment and segmentation in one step using the Levenshtein distance (Matuzov et al., 2015).

New segments used to compute the automatic metrics.

*Sec 4.3*

# Mitigating error — Gender bias

**We Are  
What We Eat**

# Gender and data



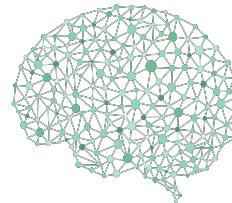
We Are  
What We Eat



# Gender and data



# Gender and data



- ~ 70% of the TED speakers is male
- Most of the ASR and MT data are generated by male speakers



# Gender and translation

- How do languages convey the gender of a referred entity?

English:  
Natural Gender Language

- Pronouns (he/she)
- Lexical gender (boy/girl)
- Gender-marked titles  
(actor/actress)

**she** is a good friend  
**he** is a good friend



I'm a good friend

Italian/French:  
Grammatical Gender Languages

- Overtly express feminine/masculine gender on numerous POS

è una buona amica (Fem.)  
è un\_ buon\_ amico» (Masc.)

# Gender bias: a technical and ethical problem

<i>"I'm a good friend"</i>	Correct Italian translation	Most probable automatic translation
M: "Sono <u>un</u> _buon_ <u>amico</u> "	✓	✓
F: "Sono <u>una</u> <u>buona</u> <u>amica</u> "	✓	

# Gender bias: a technical and ethical problem

<i>"I'm a good friend"</i>	Correct Italian translation	Most probable automatic translation
M: "Sono <u>un</u> _buon_ <u>amico</u> "	✓	✓
F: "Sono <u>una</u> <u>buona</u> <u>amica</u> "	✓	

*Independently from the speaker*



# Gender bias: a technical and ethical problem

<i>"I'm a good friend"</i>	Correct Italian translation	Most probable automatic translation
M: "Sono un_ buon_ amico"	✓	✓
F: "Sono una_ buona_ amica"	✓	

*Independently from the speaker*



Heart surgeon

Bias in the training data...  
...pushes systems towards a “male default”...  
...amplifying social asymmetries!



Nurse

# Gender bias and automatic translation

- **Machine Translation** (text-to-text)  
→ textual input does NOT always provide gender clues
- **Speech Translation** (speech-to-text)  
→ audio input can provide gender clues

I'm a good friend

I'm a good friend



*Are ST systems able to exploit audio information to translate gender?*

# Gender bias and ST - exploiting audio features

- Bentivogli et al., “*Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus*”, ACL 2020
  - **MuST-SHE: a benchmark for the analysis of gender translation in MT and ST**

- **Derived from MuST-C** (2 language directions En→It, En→Fr)
- **Gender-sensitive design:** each segment contains 1+ English gender-neutral word translated into the corresponding masculine/feminine target word(s)
- **2 gender phenomena:** info-in-audio (*I'm a good friend*), info-in-content (*she is a good...*)

# Gender bias and ST - exploiting audio features

- Bentivogli et al., “*Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus*”, ACL 2020
  - MuST-SHE: a benchmark for the analysis of gender translation in MT and ST
  - **Gender-sensitive evaluation methodology based on “gender swapping”**

- BLEU/Accuracy scores computed against **correct** and **wrong** references
  - Src: *I have been to London* (female speaker)
  - C-Ref: *Io sono stata a Londra*,
  - W-Ref: *Io sono stato a Londra*
- Difference between correct and wrong reference as a measure of gender translation performance (the higher the better -- lower bias!)

# Gender bias and ST - exploiting audio features

- Bentivogli et al., “*Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus*”, ACL 2020
  - MuST-SHE: a benchmark for the analysis of gender translation in MT and ST
  - Gender-sensitive evaluation methodology based on “gender swapping”
  - **Comparison between end-to-end and cascade ST approaches**

- Translation quality (BLEU): cascade better than e2e
- Gender translation (BLEU+gender swapping): the two perform on par
- Gender translation (Accuracy+gender swapping) on info-in-audio samples:
  - **e2e much better than simple cascade**
    - leveraging audio features  $\Rightarrow$  ethical issues (vocally impaired, transgender)?

# Gender bias and ST - exploiting speakers' info

- Gaido et al., “*Breeding Gender-aware Direct Speech Translation Systems*”, Coling 2020
  - **MuST-Speakers: annotation of MuST-SHE with speakers' gender information**

# Gender bias and ST - exploiting speakers' info

- Gaido et al., “*Breeding Gender-aware Direct Speech Translation Systems*”, Coling 2020
  - MuST-Speakers: annotation of MuST-SHE with speakers’ gender information
  - **Comparison of different e2e ST systems**

- **Base**: Generic, “gender-unaware” ST model
- **Multi-gender**: single model informed of the speaker’s gender via pre-pended gender tokens
- **Gender-specialized**: two models, fine-tuned on utterances spoken by men/women
- Overall translation quality (BLEU): small differences
- Gender translation (Accuracy+gender swapping) on info-in-audio samples (*I'm a good friend*):
  - **Specialized >> Multi-gender >> Base**

*Sec 5:*

# Advanced topics

**Utterance segmentation**

**Multilingual ST**

**Under-resourced languages**

*Sec 5.1*

# Utterance Segmentation

# Utterance segmentation - Problem

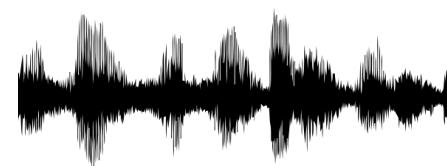
- **Mismatch between training and evaluation data**
  - Training corpora: “sentence-level” split of continuous speech



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



This is an audio signal.



In the training data it was split using strong punctuation.



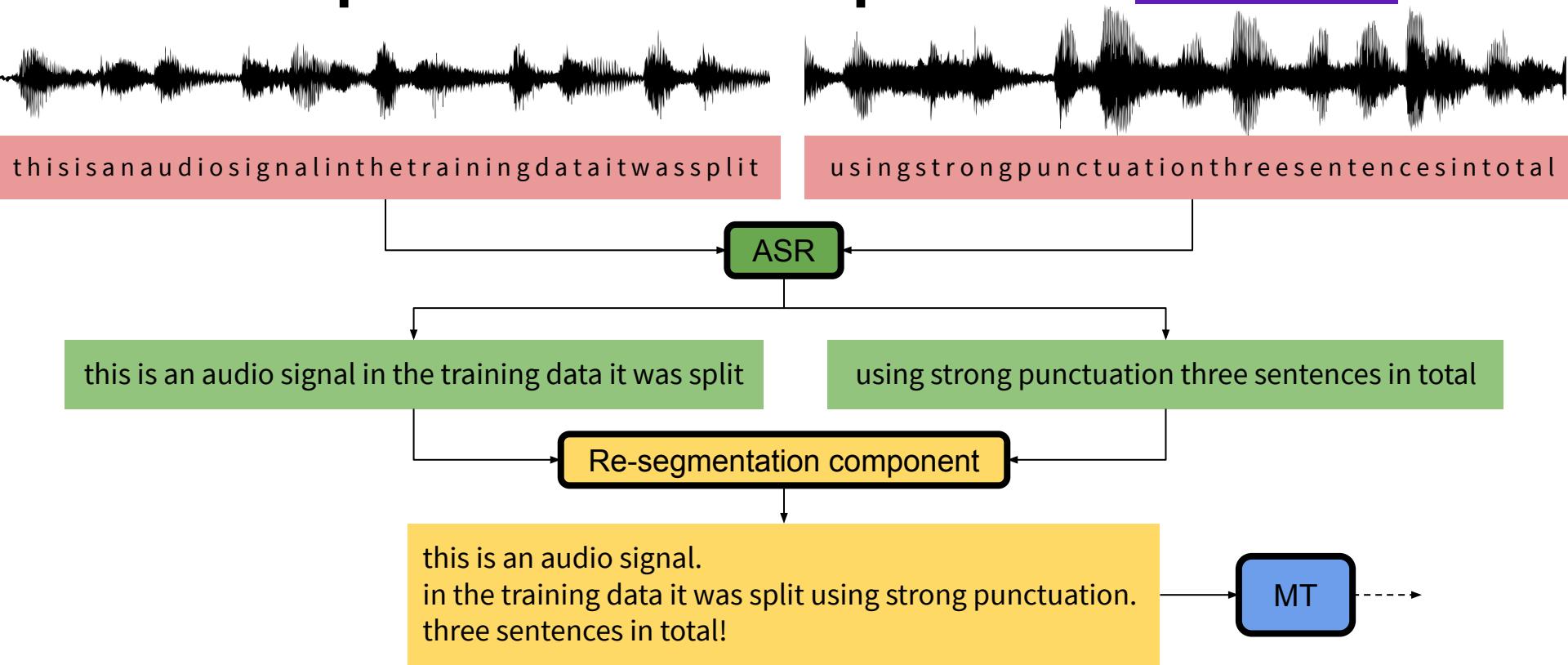
Three sentences in total!

- At run-time: unsegmented continuous speech



this is an audio signal in the training data it was split using strong punctuation three sentences in total

# How to split continuous speech in cascade ST?



# How to split continuous speech in e2e ST?



this is an audio signal in the training data it was split using strong punctuation three sentences in total

# Solution 1: Split on silences (via VAD)



this is an audio signal in the training data it was split using strong punctuation three sentences in total



this is an audio signal



in the training data it was split using strong punctuation



three sentences

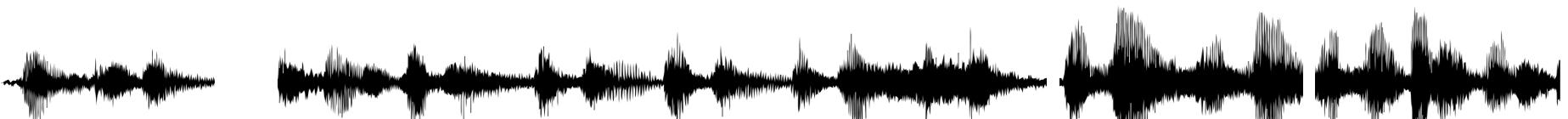


in total

# Solution 1: Split on silences (via VAD)



this is an audio signal in the training data it was split using strong punctuation three sentences in total



this is an audio signal

in the training data it was split using strong punctuation

three sentences

in total

*Advantage: silences as a proxy of sentence boundaries*

*Drawback: variable segments' length (including very short and very long ones)*

# Solution 2: Split based on fixed audio duration



this is an audio signal in the training data it was split using strong punctuation three sentences in total



this is an audio signal in the tra



ining data it was split using strong pu



nctuation three sentences in total

# Solution 2: Split based on fixed audio duration



this is an audio signal in the training data it was split using strong punctuation three sentences in total



this is an audio signal in the tra



ining data it was split using strong pu



nctuation three sentences in total

*Advantage: uniform segment length*

*Drawback #1: split points are likely to break the input in critical positions*

*Drawback #2: non-speech frames are kept in the input*

# Solution 3: Split on silences & segments' length

Potapczyk and Przybysz: “SRPOL’s system for the IWSLT 2020 end-to-end speech translation task”, IWSLT 2020



this is an audio signal in the training data it was split



this is an audio signal

in the training data it was split



using strong punctuation three sentences in total



using strong punctuation



three sentences in total

# Solution 3: Split on silences & segments' length

Potapczyk and Przybysz: “SRPOL’s system for the IWSLT 2020 end-to-end speech translation task”, IWSLT 2020



this is an audio signal in the training data it was split



using strong punctuation three sentences in total



this is an audio signal

in the training data it was split



using strong punctuation



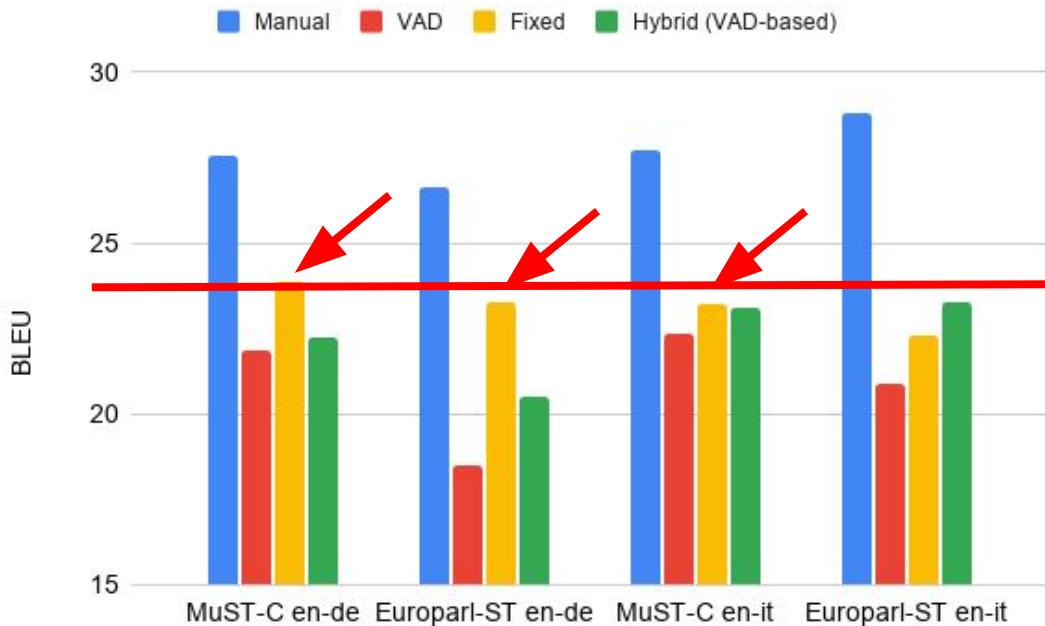
three sentences in total

Advantages: closer to sentence-like splits, uniform segment length

Drawback #1: manually-detected silences (non scalable/reproducible)

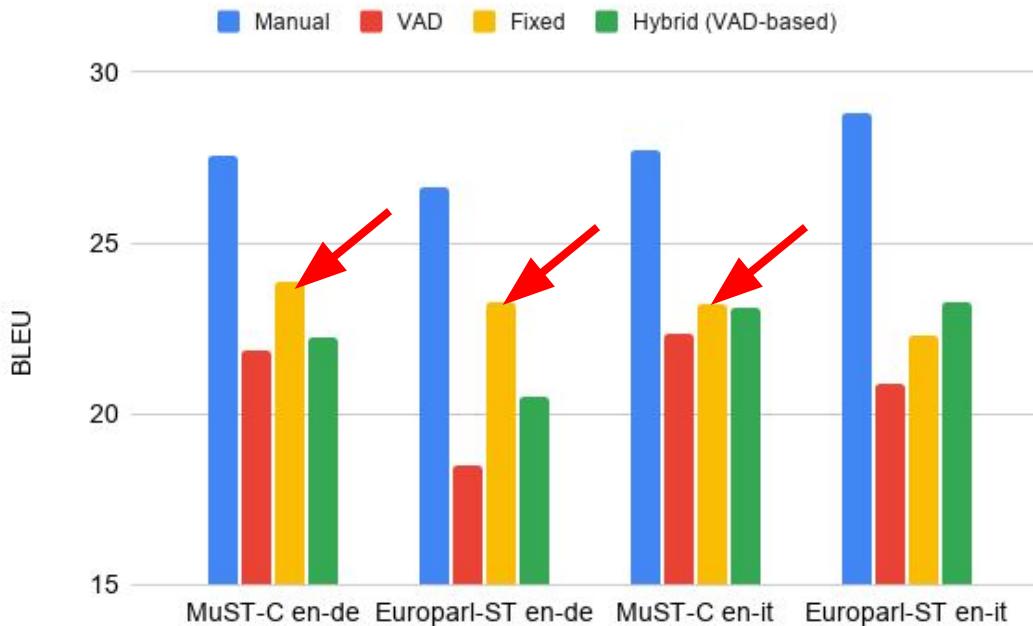
Drawback #2: full audio required for splitting (not applicable to audio streams)

# Utterance segmentation - An open problem



*Large room for improvement compared to manual segmentation*

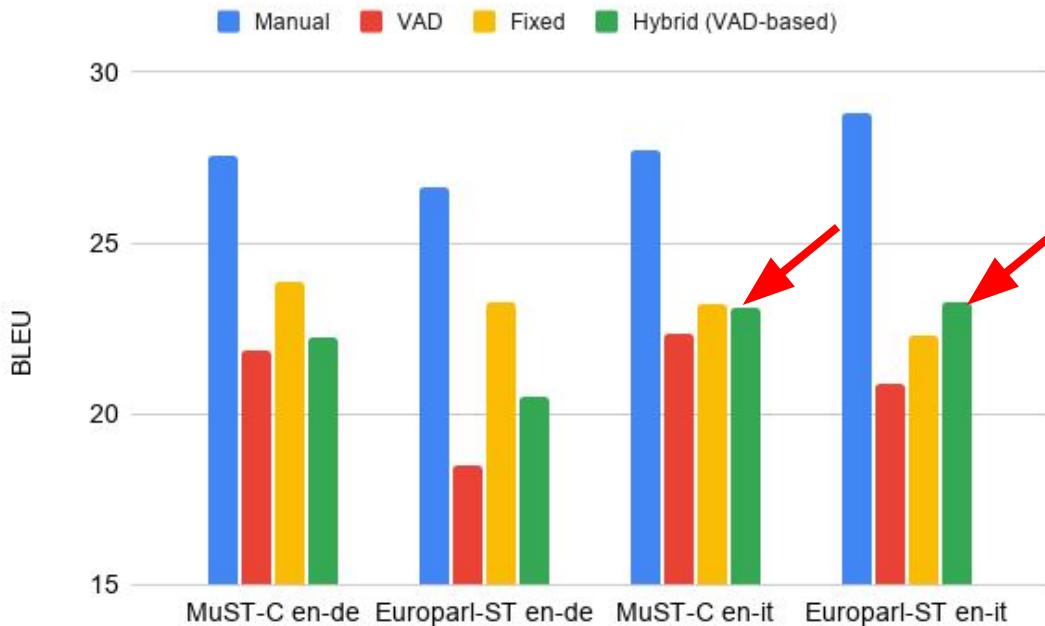
# Utterance segmentation - An open problem



*FIXED length surprisingly good*

→ segments' length is more important than precise split times

# Utterance segmentation - An open problem



Fully automatic hybrid segmentation?

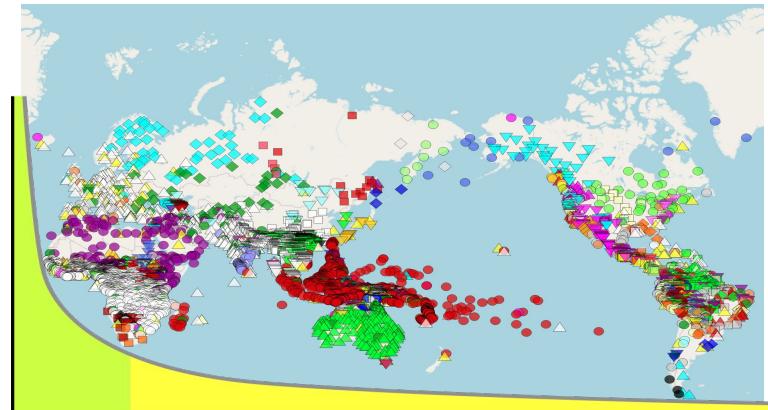
→ better than VAD, better than FIXED on one language pair

*Sec 5.2*

# Multilingual ST

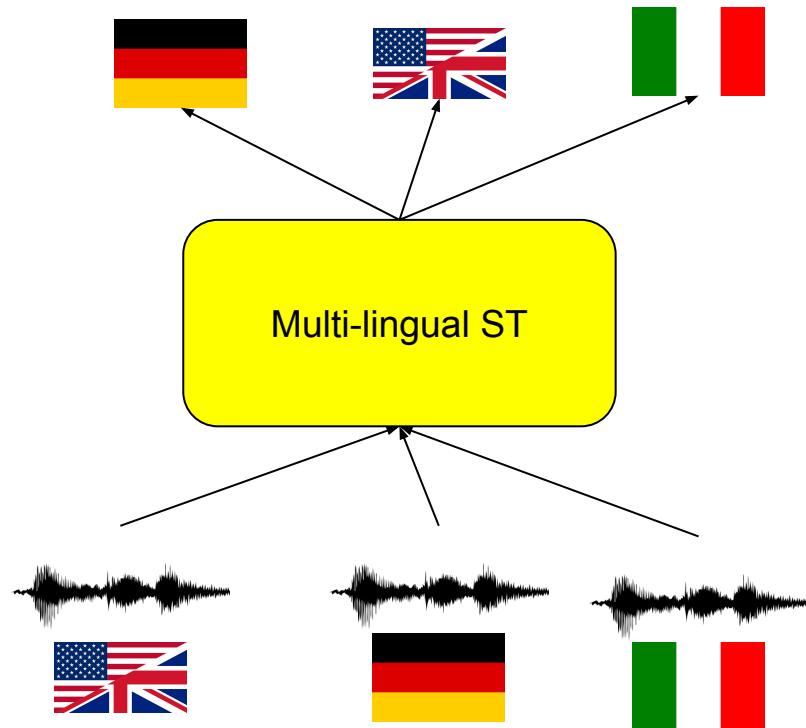
# Multilingual ST

- Most research focuses on few languages
- More than *7,000 languages* in the world
- Challenges:
  - Scale to many languages
  - Limited resources



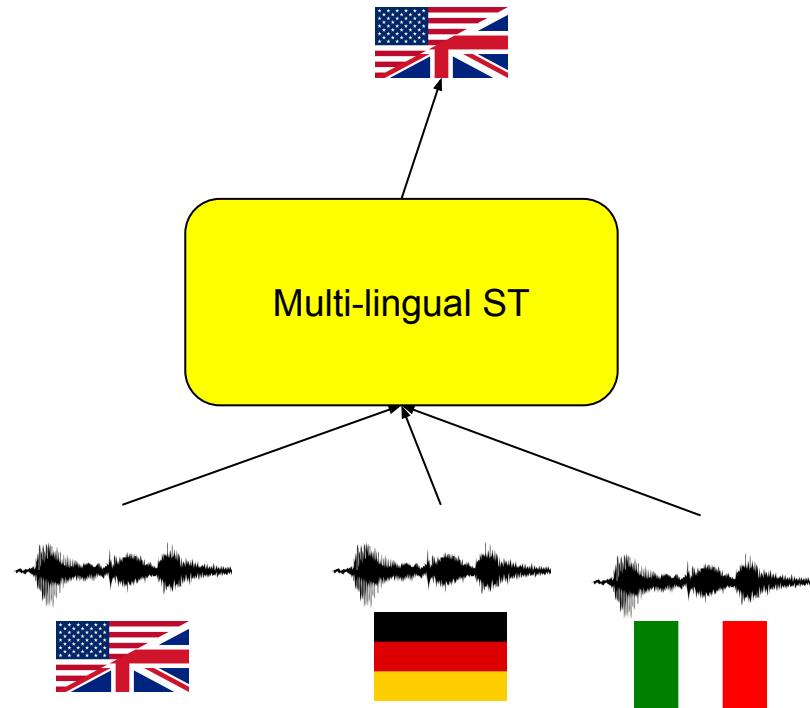
# Multilingual ST

- Idea:
  - *Single model for many languages*
  - Motivated by text translation
- Advantages:
  - Less training data necessary
  - Handle several languages by single model
  - Zero-shot direction:
    - Translate between languages without training data



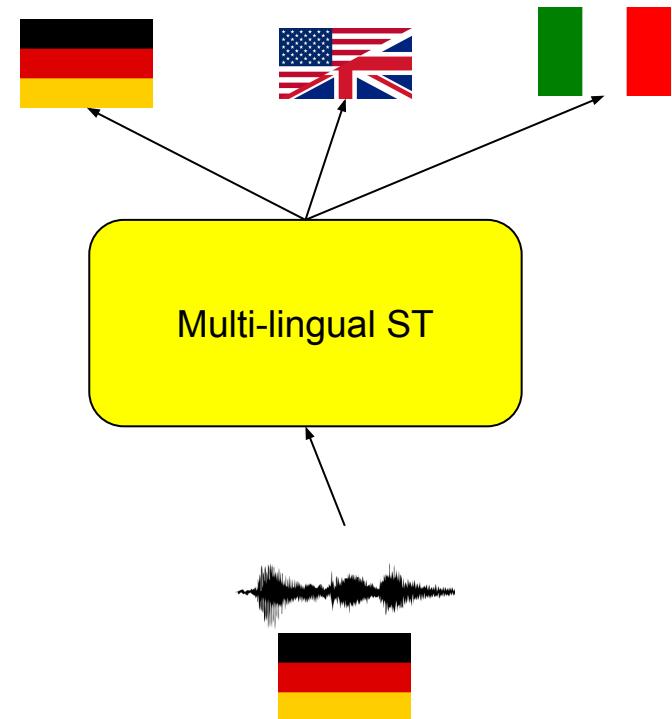
# Multilingual ST

- Scenarios:
  - Many-to-One



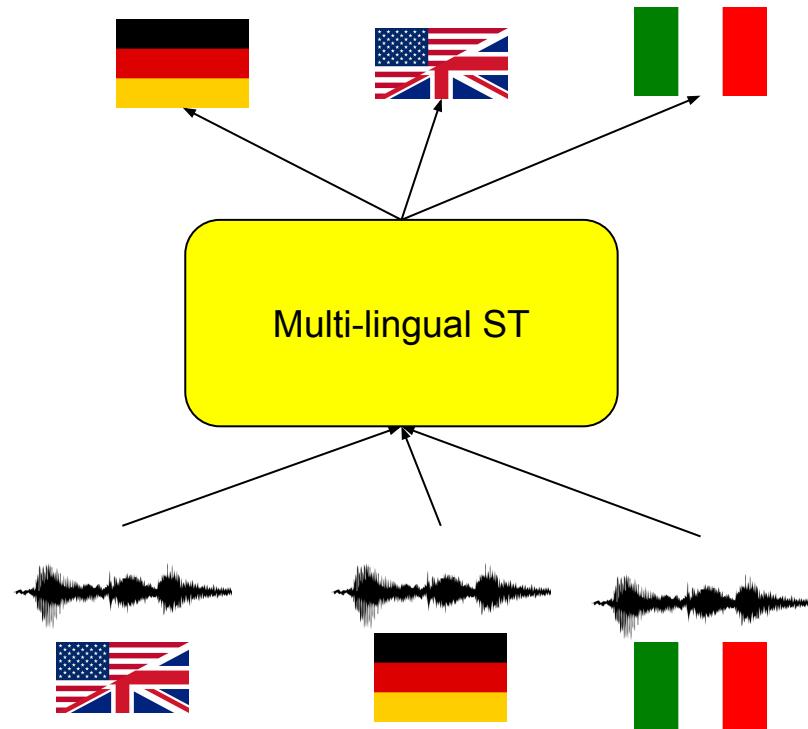
# Multilingual ST

- Scenarios:
  - Many-to-One
  - One-to-Many



# Multilingual ST

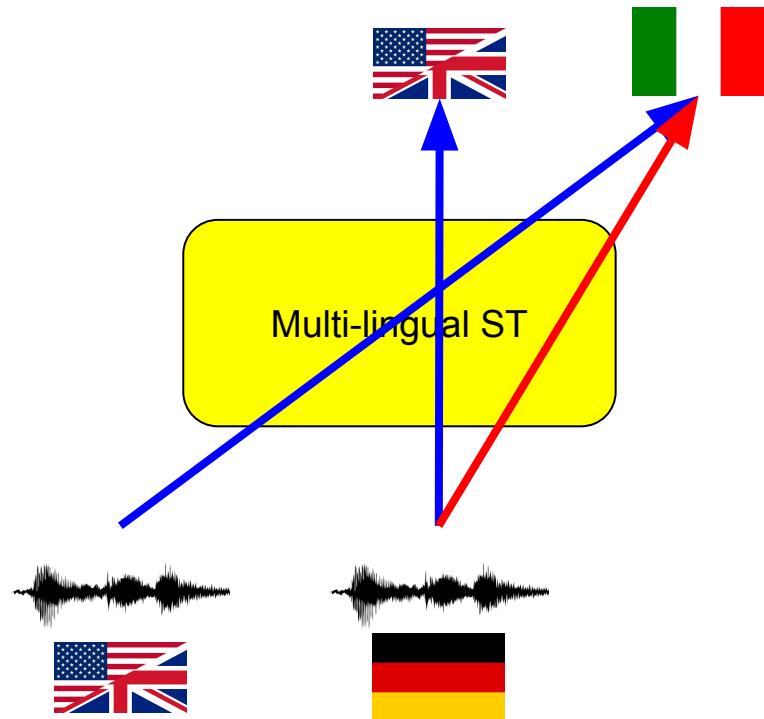
- Scenarios:
  - Many-to-One
  - One-to-Many
  - Many-to-Many



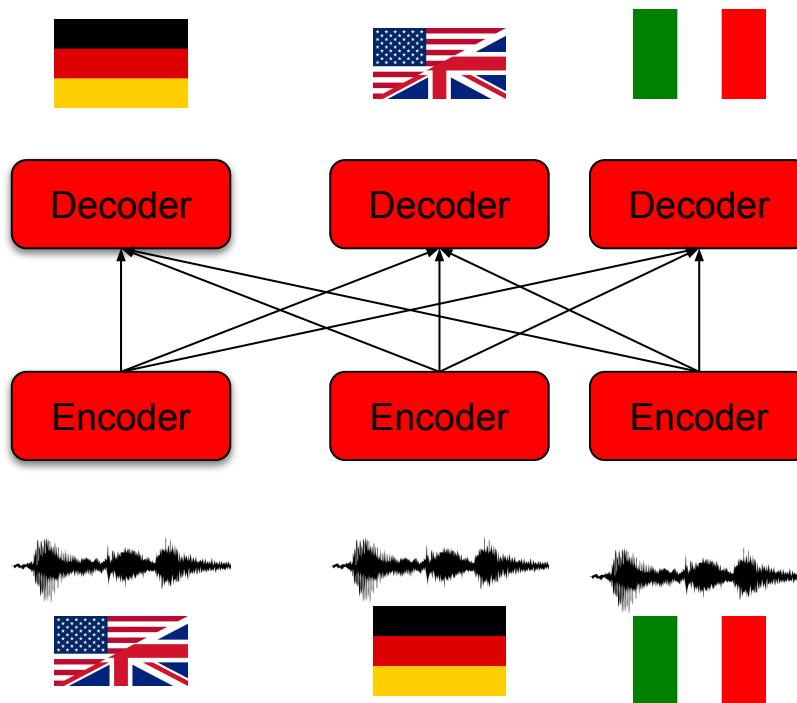
# Multilingual ST

- Scenarios:
  - Many-to-One
  - One-to-Many
  - Many-to-Many
- Zero-shot:
  - No training data in test language pair

Training direction        
Test direction     



# Multilingual ST - Architecture



Individual encoder and decoder for each language  
(e.g. Escolano et al. 2020)

# Multilingual ST - Architecture

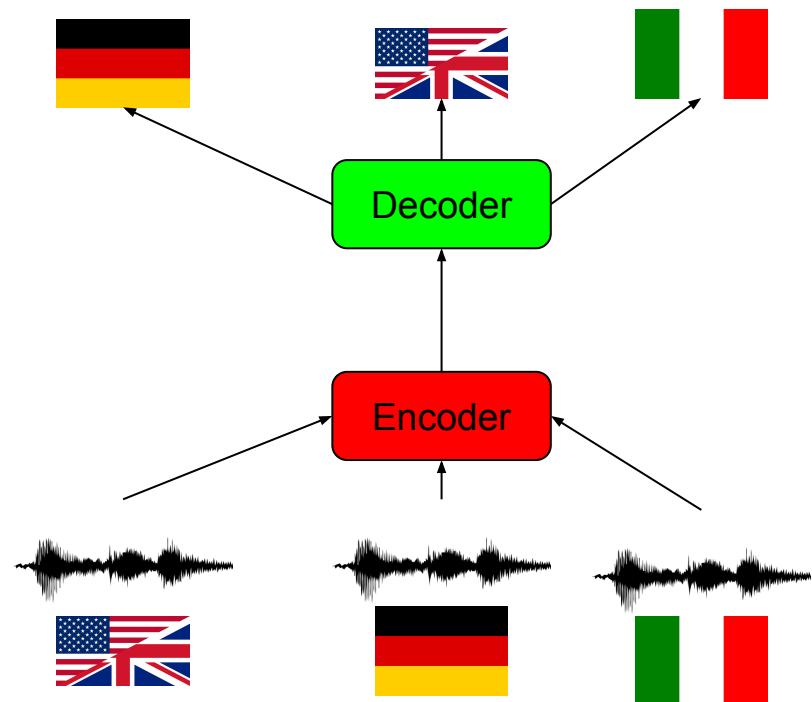
Joint encoder and decoder

Di Gangi et al., 2019

Inaguma et al., 2019

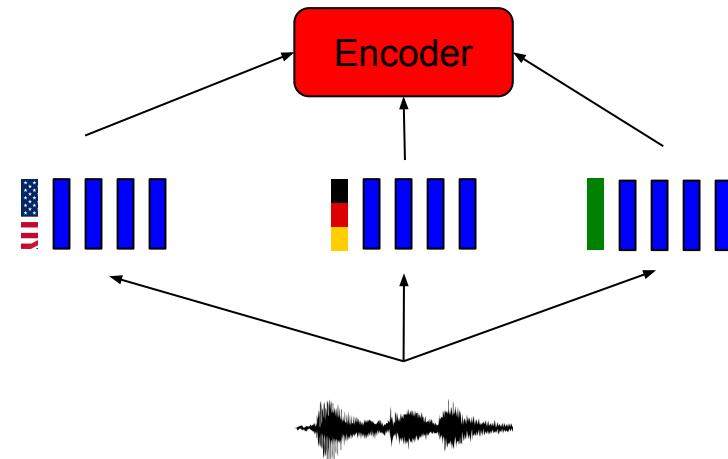
Challenge:

*How to model different languages?*



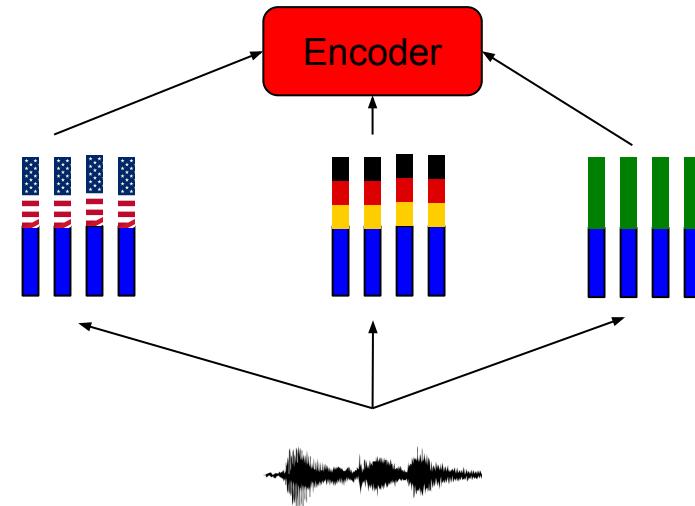
# Multilingual ST - Language representation

- Encoder
  - Concat
    - Append learned language embedding for target language to audio features



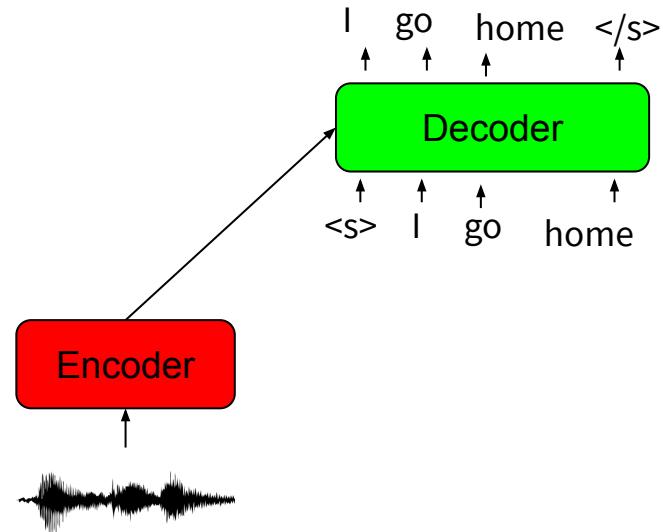
# Multilingual ST - Language representation

- Encoder
  - Concat
    - Append learned language embedding for target language to audio features
  - Merge
    - Repeat language embedding for target language at each time step



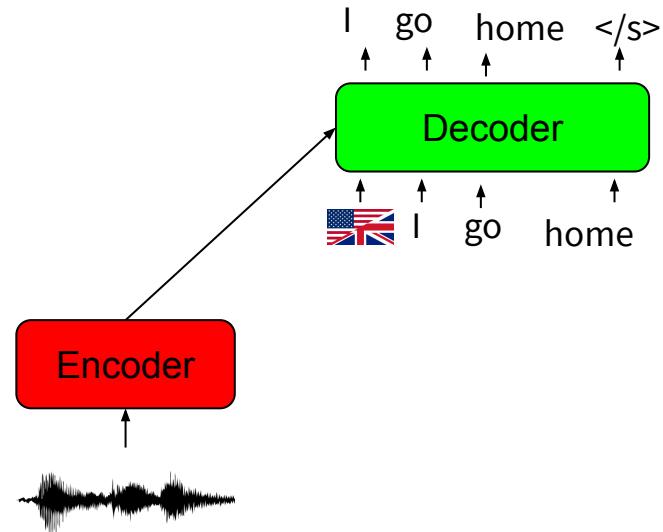
# Multilingual ST - Language representation

- Encoder
- Decoder



# Multilingual ST - Language representation

- Encoder
- Decoder
  - Replace Begin of sentence by sentence embedding

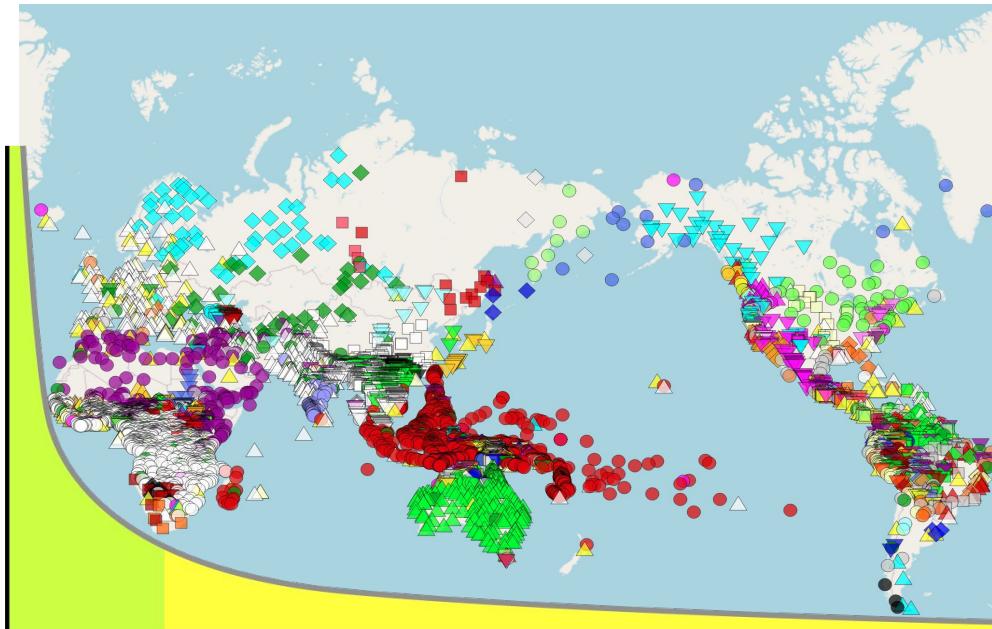


*Sec 5.3*

# Under-resourced Languages

# Under-resourced languages

*More than 7,000 languages spoken today*



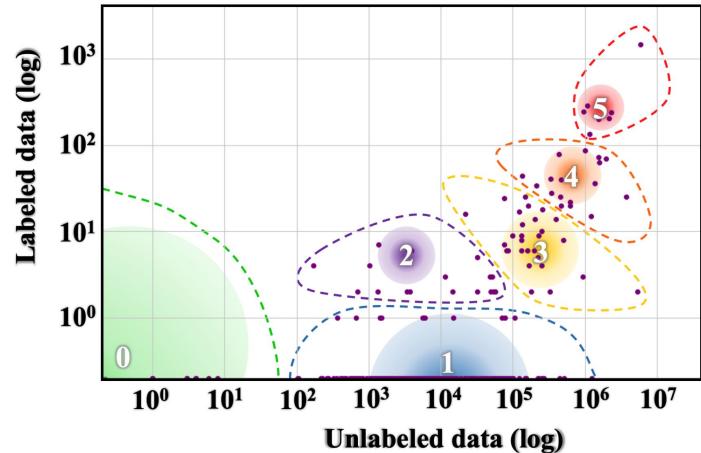
# Under-resourced languages

*What makes a language under-resourced?*

- Data availability: labeled data, unlabeled data, quality and representation
- Data domain: coverage and representation
- Noisy and/or opaque orthographies
- Unwritten languages
- Typological coverage:
  - Unique phonetic and phonological systems
  - Dialectal variation
  - Code-switching
  - Representation of non-native speakers

# Taxonomy

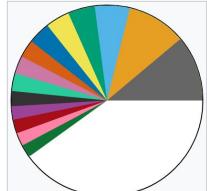
0. Exceptionally limited resources: pretraining exacerbates situation
1. Some amount of unlabeled data
2. Small set of labeled data created
3. Unlabeled data enables pretraining, but limited labeled data
4. Large amount of unlabeled data, high quality but limited labeled
5. High-resource languages



Language resource distribution of Joshi et al. (2020). The size and colour of a circle represent the number of languages and speakers respectively in each category. Colours (on the VIBGYOR spectrum; **Violet–Indigo–Blue–Green–Yellow–Orange–Red**) represent the total speaker population size from low (violet) to high (red).

(Joshi et al. 2020)

# Languages: Examples



Distribution of the 55,991,866 articles in different language editions (as of 9 March 2021).<sup>[4]</sup> The majority of the articles in Swedish, Cebuano, and Waray were created by Lsjbot.<sup>[6]</sup>

English	(11.2%)
Cebuano	(9.9%)
Swedish	(6%)
German	(4.5%)
French	(4.1%)
Dutch	(3.7%)
Russian	(3%)
Italian	(3%)
Spanish	(3%)
Polish	(2.6%)
Waray	(2.3%)
Vietnamese	(2.3%)
Japanese	(2.2%)
Egyptian Arabic	(2.2%)
Other	(40%)

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.0B	88.17%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	1.0B	8.93%
2	Zulu, Konkani, Lao, Maltese, Irish	19	300M	0.76%
3	Indonesian, Ukrainian, Cebuano, Afrikaans, Hebrew	28	1.1B	1.13%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	1.6B	0.72%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Number of languages, number of speakers, and percentage of total languages for each language class

## 0. Dahalo:

[Recorded Swadesh list](#)

## 1. Cherokee:

[Bible](#); [15k sentences parallel text](#); Tatoeba; Ubuntu

## 2. Zulu:

[Recorded word lists](#); Tatoeba; Ubuntu

## 3. Cebuano:

[Recorded word lists](#); [BABEL](#); [Bible](#); Wikipedia; Tatoeba; Ubuntu

## 4. Korean:

[Bible](#); Wikipedia; OpenSLR [40](#), [58](#), [97](#); Tatoeba; Ubuntu

## 5. English:

▽

# ST: Resources Required

*Two steps where resources are required: ① for training and ② for corpus creation*

## Labeled data:

parallel speech and translations, segmented

## Availability:

MuST-C (1); mTEDx (8); CoVoST (21)

## Unlabeled data:

monolingual source language speech;  
monolingual target language text

Bible (~1000); Wikipedia (285);  
linguistic resources often <2 hours

## Pronunciation lexicons:

Use: alignment, hybrid ASR models; alternate data representations; CTC loss and/or compression

Hand-created lexicons often unreleased;  
Wikipron (117); Epitran (63)

(# source languages)

# Pretrained Models

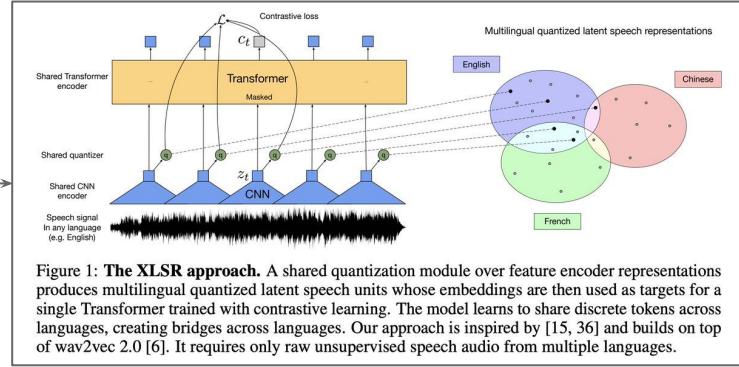
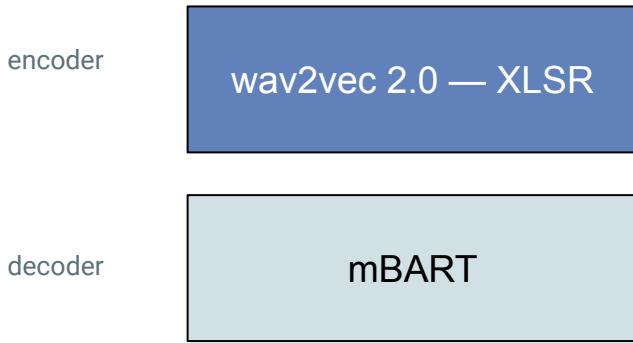


Figure 1: **The XLSR approach.** A shared quantization module over feature encoder representations produces multilingual quantized speech units whose embeddings are then used as targets for a single Transformer trained with contrastive learning. The model learns to share discrete tokens across languages, creating bridges across languages. Our approach is inspired by [15, 36] and builds on top of wav2vec 2.0 [6]. It requires only raw unsupervised speech audio from multiple languages.

(Baevski et al. 2020; Liu et al. 2020;  
Li et al. 2021)

## Methods previously discussed:

pretraining + finetuning, knowledge distillation,  
alternate data representations

## Dependences on shared features:

in-vocabulary orthography, phone inventories,  
use of same model architecture

Unless we assess on under-resourced  
languages, we will not know how well  
methods apply!

*Sec 6:*

# Real-world Applications

**Automatic generation of subtitles**

**Interpretability**

**Simultaneous translation**

*Sec 6:*

# Real-world Applications

**Automatic generation of subtitles**

**Simultaneous translation**

*Sec 6.1*

# Automatic Generation of Subtitles

# Automatic subtitling - Motivation



- Explosion of audio-visual content available (Cinema, OTT platforms, social media,...)
  - Need: offer high-quality subtitles into dozens of languages in a short time
  - Problem: human subtitling is slow and costly (1-15\$/min)
  - Goal: automatic solutions to reduce human workload and costs

# 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  $\approx$  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.
```

# Segmenting into proper subtitles

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

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

# Segmentation approaches

Manual template

This kind of harassment keeps women <eol> from accessing the internet – <eob>

MT

Ce harcèlement empêche les femmes <eol> d'accéder à Internet, <eob>

# Segmentation approaches

Manual template

This kind of harassment keeps women <eol> from accessing the internet – <eob>



MT

Ce harcèlement empêche les femmes <eol> d'accéder à Internet, <eob>

Previous works focused only on length-matching given the template

(Matusov et al., 2019;  
Lakew et al., 2019)

# Segmentation approaches

Manual template

This kind of harassment keeps women <eol> from accessing the internet – <eob>

MT

Cascade



→ ASR

this kind of harassment  
keeps woman from  
accessing internet

→ MT

Ce harcèlement empêche  
les femmes <eol>  
d'accéder à Internet,  
<eob>

# Segmentation approaches

Manual template

This kind of harassment keeps women <eol> from accessing the internet – <eob>

MT

Cascade



→ ASR

this kind of harassment keeps woman from accessing internet

MT

Ce harcèlement empêche les femmes <eol> d'accéder à Internet, <eob>

E2E



→ ST

# Segmentation approaches

Manual template

*Costly!*

This kind of harassment keeps women <eol> from accessing the internet – <eob>

MT

Cascade



→ ASR →

this kind of harassment keeps woman from accessing internet

MT

Ce harcèlement empêche les femmes <eol> d'accéder à Internet, <eob>

E2E



→ ST

*Audio info (e.g. duration) is available to ST*

*Audio info (e.g. duration) is lost*

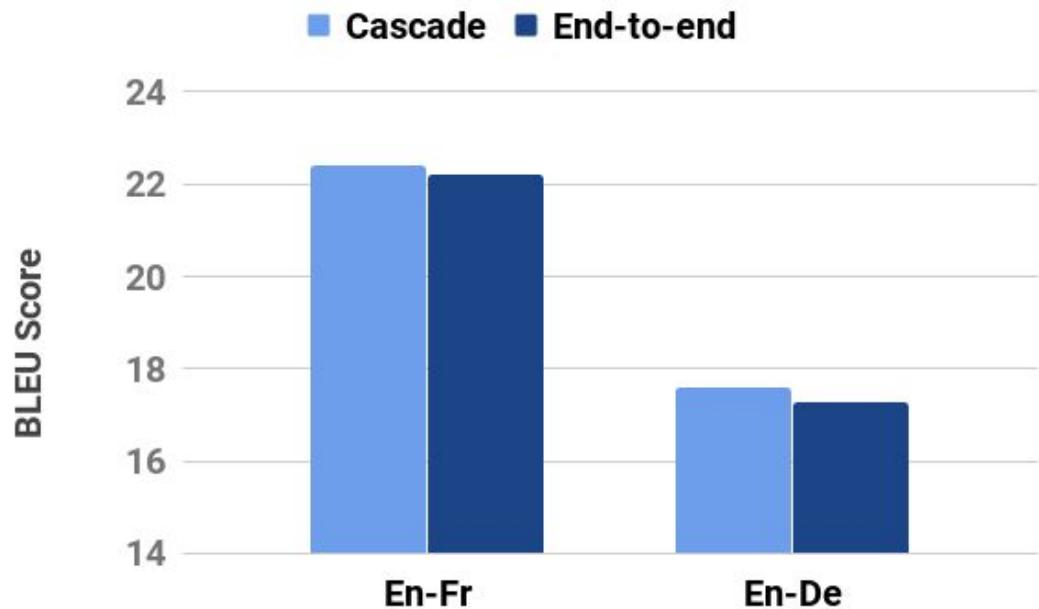
# Automatic subtitling - Data

- **OpenSubtitles** (Lison and Tiedemann, 2016) -- 60 languages
  - Variable quality (professional/amateur subt., automatic sentence-level alignm.):
  - No information about subtitle breaks
  - No alignment with audio (mostly copyright-protected videos)
- **JESC** (Pryzant et al., 2018) -- Ja-En
  - Automatic alignments (caption level = only subtitles with matching timestamps)
  - No alignment with audio
- **Must-Cinema** (Karakanta et al., 2020) -- En→ 7 languages
  - Derived from MuST-C (TED talks)
  - Annotated with subtitle breaks
  - Audio-transcript-translation alignments

# E2E subtitling: experiments on En-Fr/De

- **Doable?**

- Translation quality

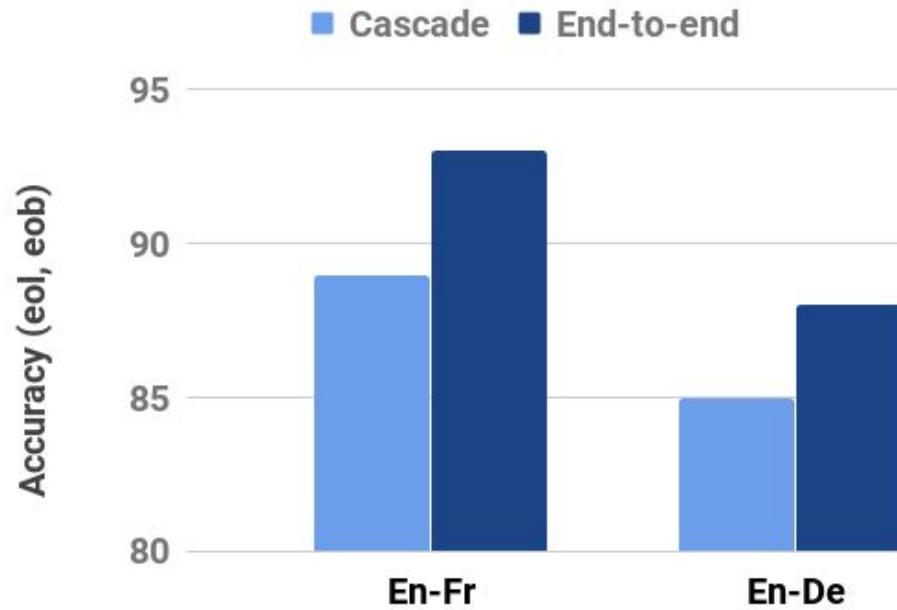


*No gap between Cascade and E2E*

# E2E subtitling: experiments on En-Fr/De

- **Effective?**

- Segmentation (<eol> and <eob> insertion)



*E2E exploits acoustic information (pause duration) to insert breaks*

*Sec 6.2*

# Interpretability

# Interpretability

Demonstrating and computationally proof the main differences between e2e and cascade:

- ) error propagation (can the e2e have errors in the internal representation?) No discrete representation
- ) showing that e2e better manipulates paralinguistic and non-linguistic information during translation

Di Gangi, Mattia A., Matteo Negri, and Marco Turchi. "Adapting transformer to end-to-end spoken language translation." In *INTERSPEECH 2019*, pp. 1133-1137. International Speech Communication Association (ISCA), 2019.

Jia, Ye, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model}]." *Proc. Interspeech 2019* (2019): 1123-1127.

Sperber, Matthias, Graham Neubig, Jan Niehues, and Alex Waibel. "Attention-passing models for robust and data-efficient end-to-end speech translation." *Transactions of the Association for Computational Linguistics* 7 (2019): 313-325.

# Interpretability

Most of the papers (Weiss et al., 2017, Jia et al., 2019, Di Gangi et al., 2019) about end-to-end SLT system mention the following advantages over the cascade:

- No error propagation:

End-to-end naturally avoids compounding errors between the ASR and MT systems.

# Interpretability

Most of the papers (Weiss et al., 2017, Jia et al., 2019, Di Gangi et al., 2019) about end-to-end SLT system mention the following advantages over the cascade:

- No error propagation:

End-to-end naturally avoids compounding errors between the ASR and MT systems

- Direct access to the audio:

End-to-end better manipulates paralinguistic and non-linguistic information during translation

The correctness of these statements given for granted

# Interpretability

Key questions:

How to empirically validate these hypotheses?

Is it true that end-to-end avoid error propagation?

To what extent, does accessing the audio help? How? When?

No answers in this tutorial!

# No error propagation

- Overall quality is not enough
- Cascade and e2e intermediate representations not comparable (transcript vs. null)
- End-to-end output depends on a mix of factors in the same model difficult to separate (two tasks collapsed into one)
- Not a consolidated architecture in end-to-end technology

Possible opening:

Sperber et al., (2019) consider the encoder output as an intermediate representation and pose the attention on the presence of errors in it

# Direct access to the audio

- Better encoder technology results in better translation performance (not enough)
- Not clear what aspects of the audio can help (e.g. prosody, emotions, tone, pauses)
- Audio understanding capability can only be analyzed in the final translation (no transcripts)
- Lack of *ad hoc* test sets to measure the impact of prosody, emotions, ...
- Extrinsic evaluations (e.g. male/female audio recognition) should not ignore the translation aspects of the problem

Possible openings:

Karakanta et al. (2020): the direct access to the audio pauses improves subtitle quality

Gaido et al. (2020): vocal characteristics can guide e2e systems in modeling gender (but opens ethical issues!)

*Sec 6.2*

# Simultaneous ST

# Simultaneous Translation

- Generate translation while speaker speaks
- Tradeoff:
  - *More context* improves speech 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 language
    - SOV vs SVO

German	Ich	melde	mich	zum	E2E	Tutorial	an
Gloss	I	register/ cancel	myself	to	E2E	tutorial	
English	I	????					

# Simultaneous Translation

- Approaches:
  - Learn optimal segmentation strategies
    - Create segments that optimizing tradeoff between segment length and translation quality
  - Advantages:
    - No changes to the system
  - Disadvantage:
    - Shorter context during translation
  - Mainly used in cascaded approaches (e.g. Oda et al., 2014)

Example:

Ich melde mich

zur Konferenz an

# Simultaneous Translation

- Approaches:
  - Learn optimal segmentation strategies
  - Re-translate / Iterative -update
    - Directly output first hypothesis
    - If more context is available:
      - Update with better hypothesis
  - Cascade
    - (Niehues et al, 2018; Arivazhagan et al, 2020)
  - End-to-end
    - (Weller et al, 2021)

Example:

Ich  
I

Ich melde mich  
I register

Ich melde mich von  
I cancel my  
registration for

# Re-translation

- Challenge:
  - Flickering
- Ideas:
  - Output masking
    - Do not output last tokens
  - Constrained decoding:
    - Fixed part of the previous translation

Example:

Ich  
I

Ich melde mich  
I register

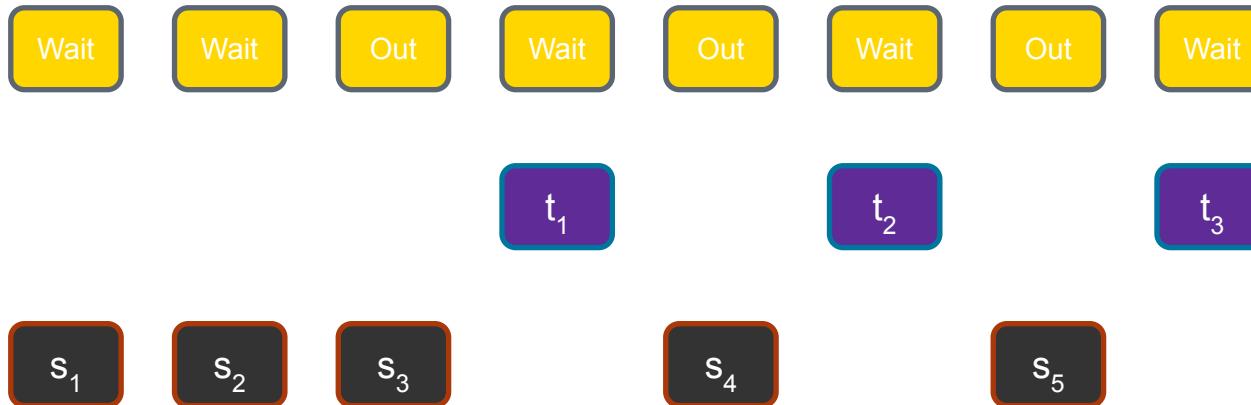
Ich melde mich von  
I cancel my  
registration for

# Simultaneous Translation

- Approaches:
  - Learn optimal segmentation strategies
  - Re-translate
  - Stream decoding
    - Dynamically learn when to generate a translation
    - At each time step:
      - Decided to output word
      - Wait for additional input

# Stream decoding

- Methods:
  - Fixed schedule (Ma et al, 2019)
    - Wait-k policy



# Stream decoding

- Challenges:
  - Assumes constant rate between input and output
    - Speaking speed varies
- Ideas:
  - Estimate word boundaries on the source side (Ma et al. 2020)
    - Predict using CTC Loss (Ren et al, 2020)

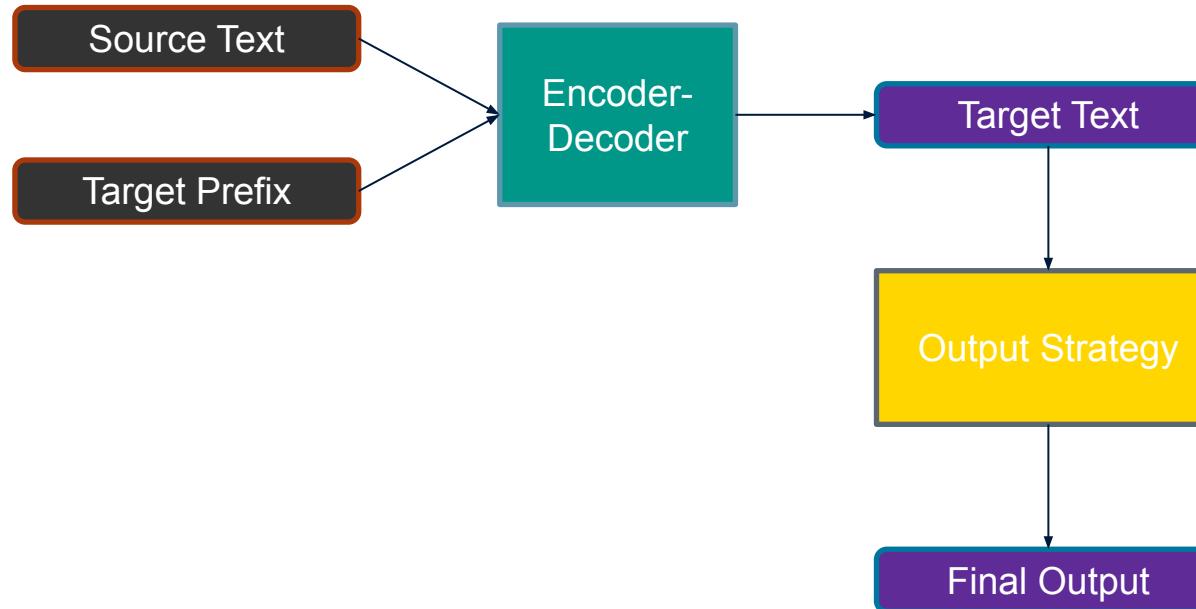
# Stream decoding

- Methods:
  - Fixed schedule (Ma et al, 2019)
  - Dynamic decision (Cho et al, 2016; Gu et al, 2017; Dalvi et al, 2018)
    - End-to-end:
      - Estimate output probability based on confidence



# Stream decoding using Retranslation

- 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	∅	All model trains	∅
1,2	∅	All models art	All
1,2,3	All	All models are wrong	All models
1,2,3,4	All models		
...			

*Sec 7:*

# Conclusion

# Recap

- Introduction
- End-to-End Models
- Leveraging Data Sources
- Evaluation
- Advanced Topics
- Real-World

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

The screenshot shows a website for a "Speech Translation Tutorial". The header includes links for "ST Tutorial", "Overview", "Materials", "Data & Resources", and a search icon. The main content area is titled "Speech Translation Tutorial" and notes that it will be presented at "EACL 2021". Below this is a section titled "Abstract" which defines speech translation and describes its recent successful applications through joint modeling. A larger text block below provides an overview of the tutorial's content, mentioning the introduction of cutting-edge research, an overview of data sources and model architectures, and discussions on high- and low-resource languages, proposed solutions, and challenges.

# References

<http://st-tutorial.github.io/materials>

Links to:

- All cited papers in this tutorial:  
bibtex and links to papers
- Individual section videos and slides

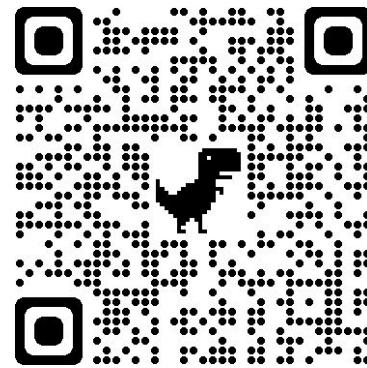


# Resources

<http://st-tutorial.github.io/resources>

Links to:

- Available data
- Available toolkits and code
- ST communities:
  - [SIGSLT](#)
  - [iwslt.org](#)





# Thank you!



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

Jan Niehues,  
*Maastricht University*  
[jan.niehues@maastrichtu...](mailto:jan.niehues@maastrichtu...)  
[niversity.nl](http://niversity.nl)

Marco Turchi,  
*Fondazione Bruno Kessler*  
[turchi@fbk.eu](mailto:turchi@fbk.eu)

Elizabeth Salesky,  
*Johns Hopkins University*  
[esalesky@jhu.edu](mailto:esalesky@jhu.edu)

Matteo Negri,  
*Fondazione Bruno Kessler*  
[negri@fbk.eu](mailto:negri@fbk.eu)