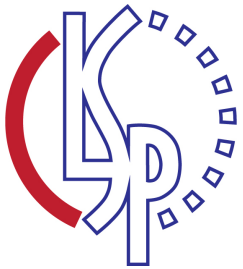
Speech Translation

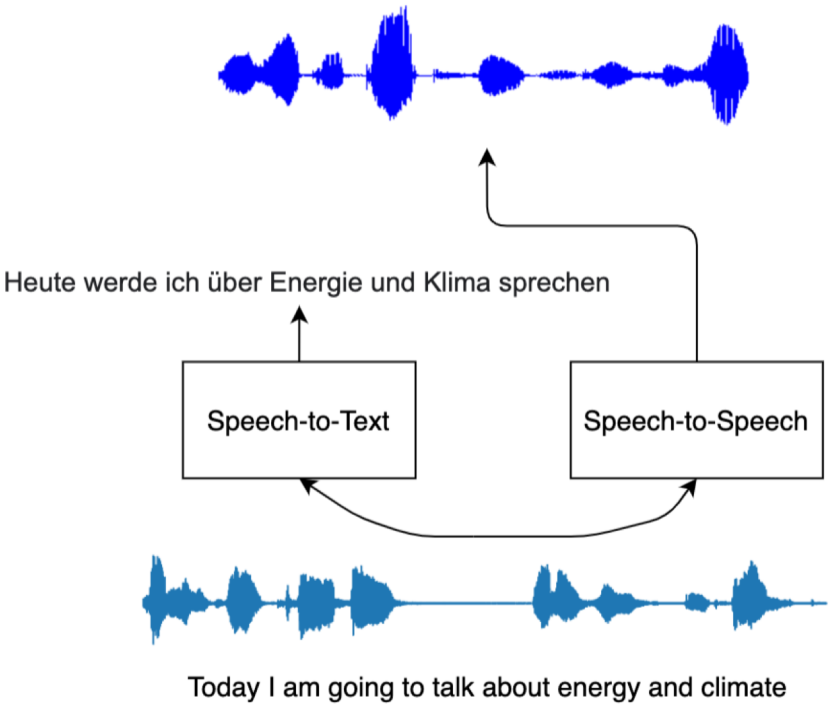
Xutai Ma

11/11/2020



Introduction

- What is speech translation?
  - Translate speech in source language to text / speech in target language



Introduction

Why/Where do we need speech translation?

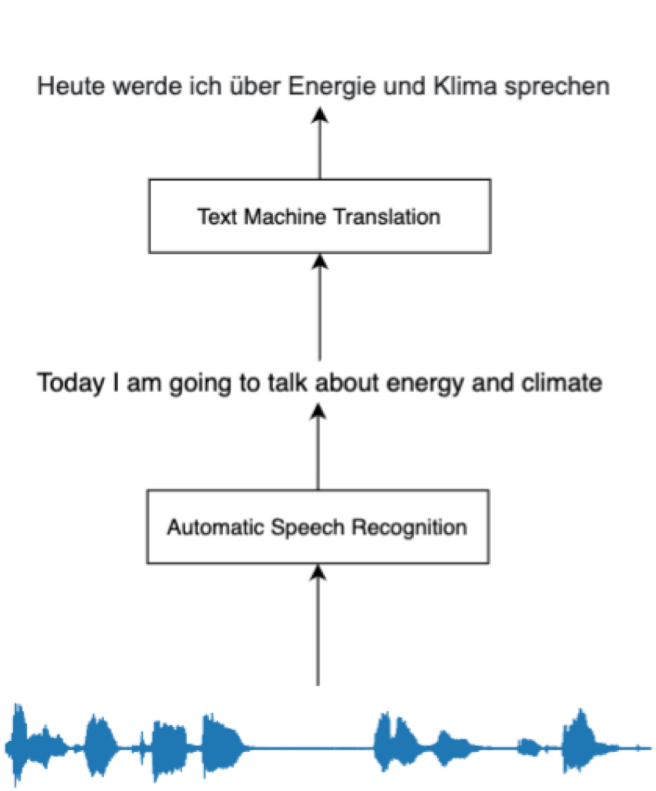
- International conferences (e.g., UN, EU)
- Live video translation (e.g., YouTube, streaming)
- Personal translator (e.g., international travels)
    - Google translate (Conversation)

How to do speech translation?

Cascade Speech Translation

- A pipeline of systems
  - Transcribe source speech into source text
  - Translated source text with a text MT model
  - If the output is speech, synthesize speech from target text

Cascade Speech Translation



(a) Speech-to-Text

(b) Speech-to-Speech

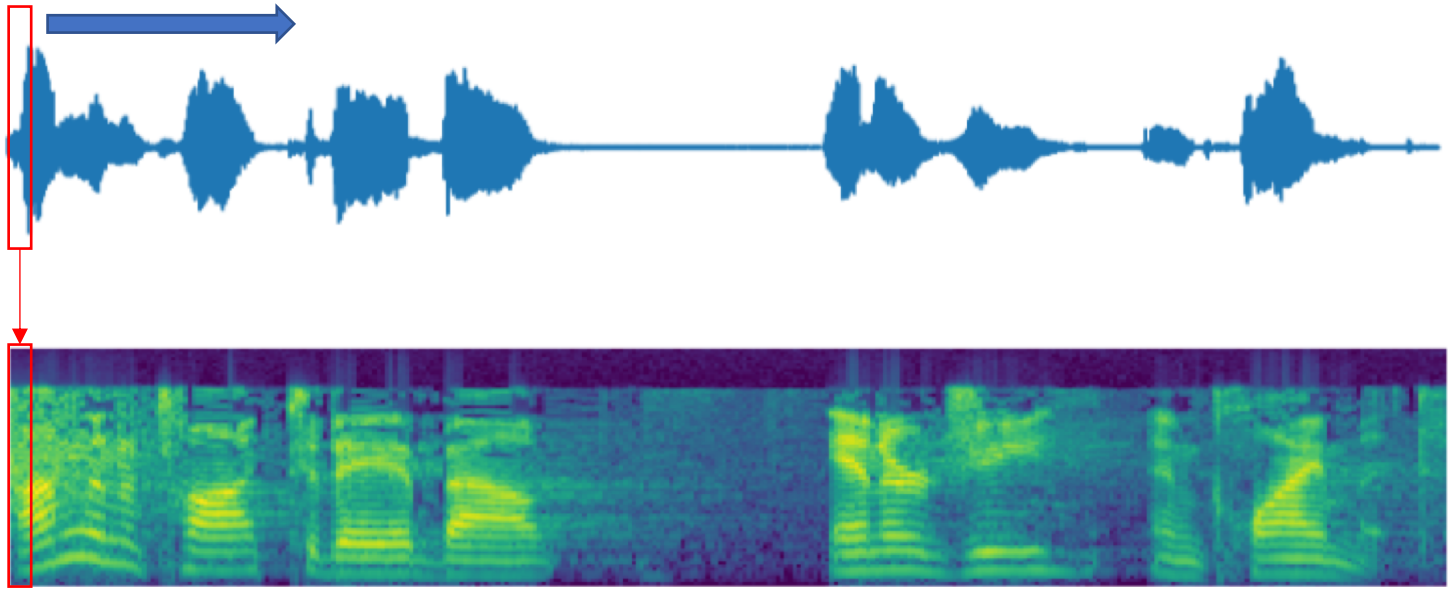
Waibel et al., 1991; Woszczyna et al., 1993; Vidal, 1997; Wang and Waibel, 1998; Takezawa et al., 1998; Ney, 1999; Bangalore and Riccardi, 2001; Fu-Hua Liu et al., 2003; Schultz et al., 2004; Matusov et al., 2005; Bertoldi and Federico, 2005; Zhang et al., 2005; Pérez et al., 2007; Sperber et al., 2017, 2019; Zhang et al., 2019; Beck et al., 2019; Black et al., 2002; Sumita et al., 2007

Background:
Speech Processing and Recognition

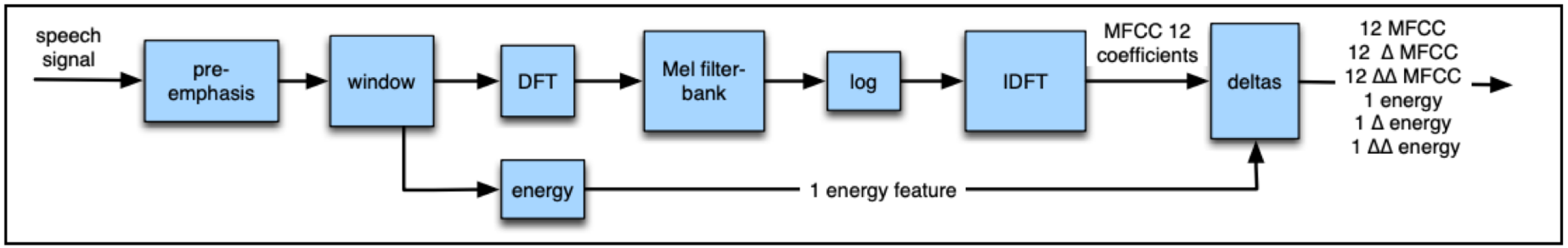
- Speech Processing
  - How to represent speech → feature extraction
- Automatic Speech Recognition (ASR)
  - Transcribe speech to text in one language
  - Seq2seq task, but input and output have the same order

Feature Extractions

- Short-Term Spectrum
  - (Mel-frequency cepstral coefficients) MFCC
- Convert speech samples to sequence of vectors



Feature Extractions



Automatic Speech Recognition

- Revisit noisy channel model

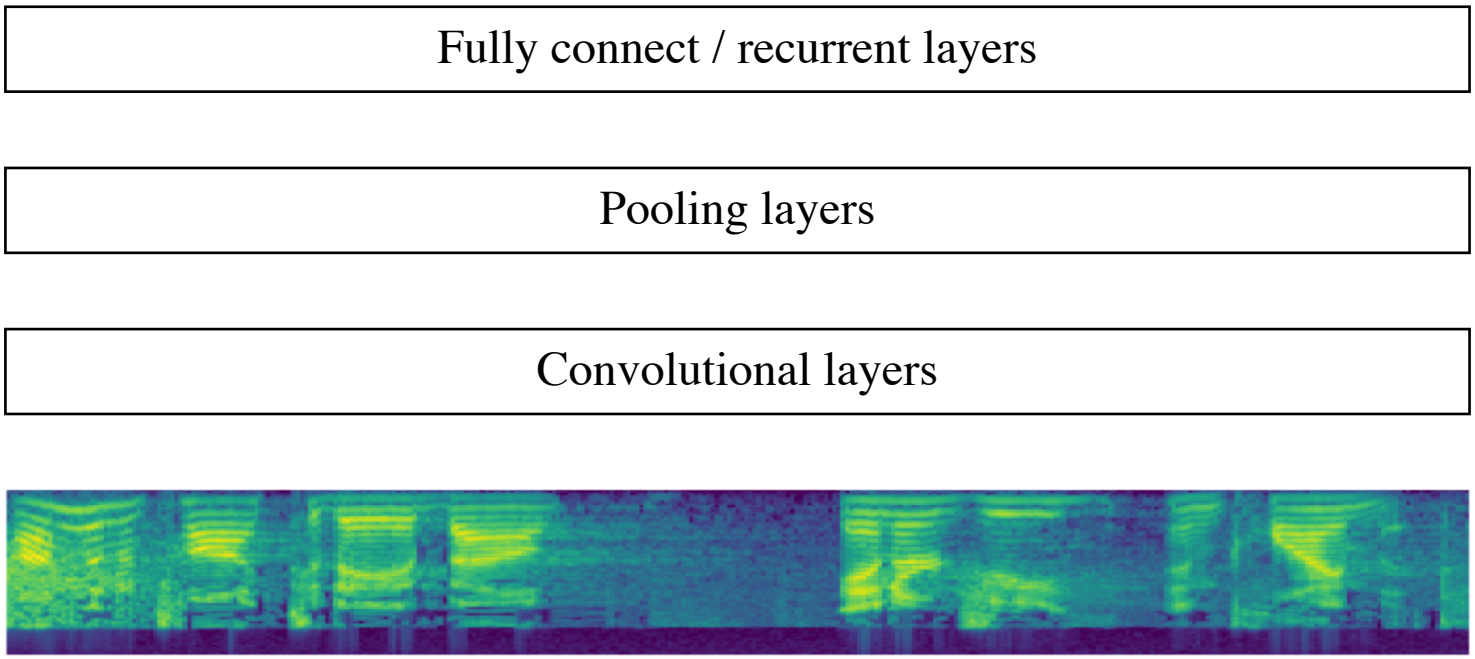
$$\hat{Y} = \text{argmax}\, P(Y|X) = \text{argmax}\, \underbrace{P(X|Y)}_{\text{Acoustic Model}} \overbrace{P(Y)}^{\text{Language Model}}$$

Automatic Speech Recognition

- Acoustic Model
  - Phone recognizer
  - Gaussian mixture model + hidden Markov model
  - Neural-based models

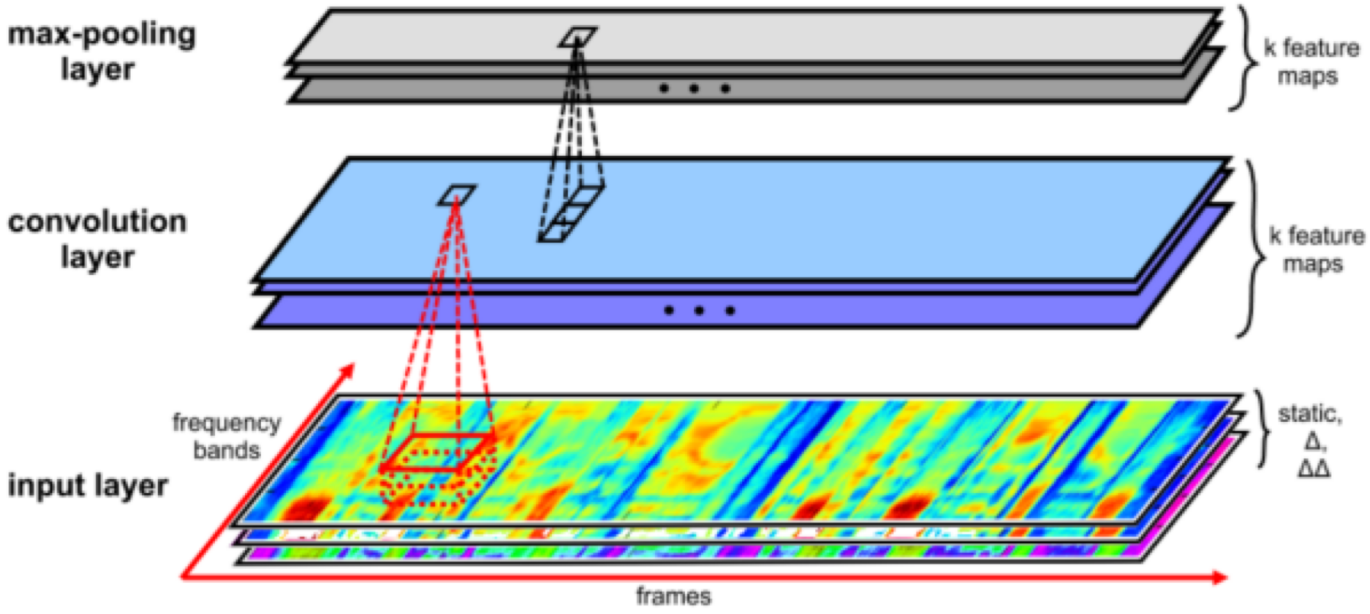
Automatic Speech Recognition

- Acoustic Model
  - Neural-based models



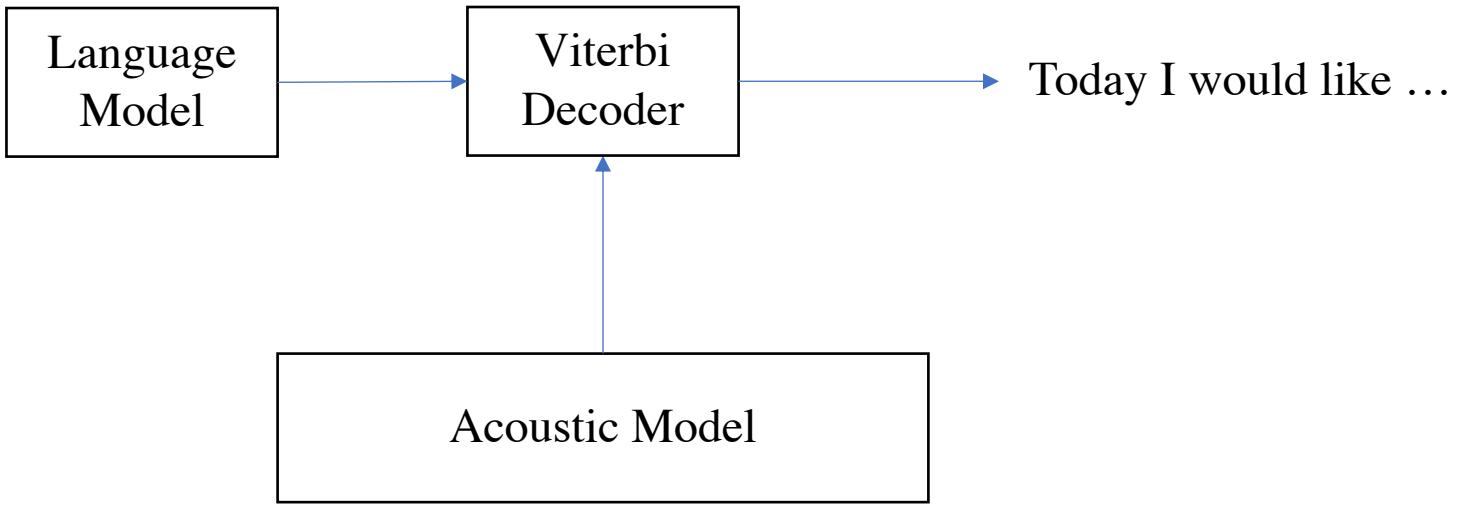
Automatic Speech Recognition

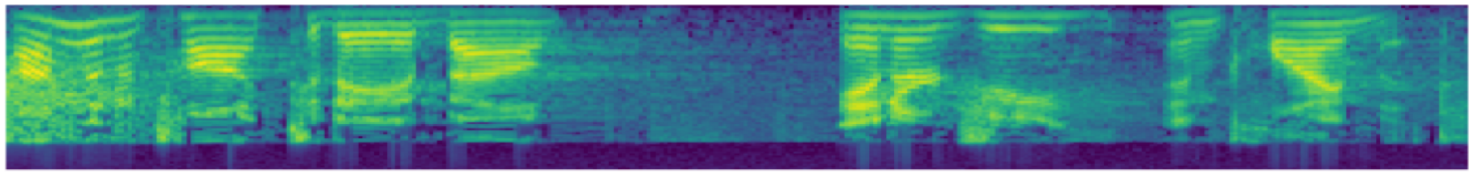
- Convolutional layers



Zhang, Ying, et al. "Towards end-to-end speech recognition with deep convolutional neural networks." *arXiv preprint arXiv:1701.02720* (2017).

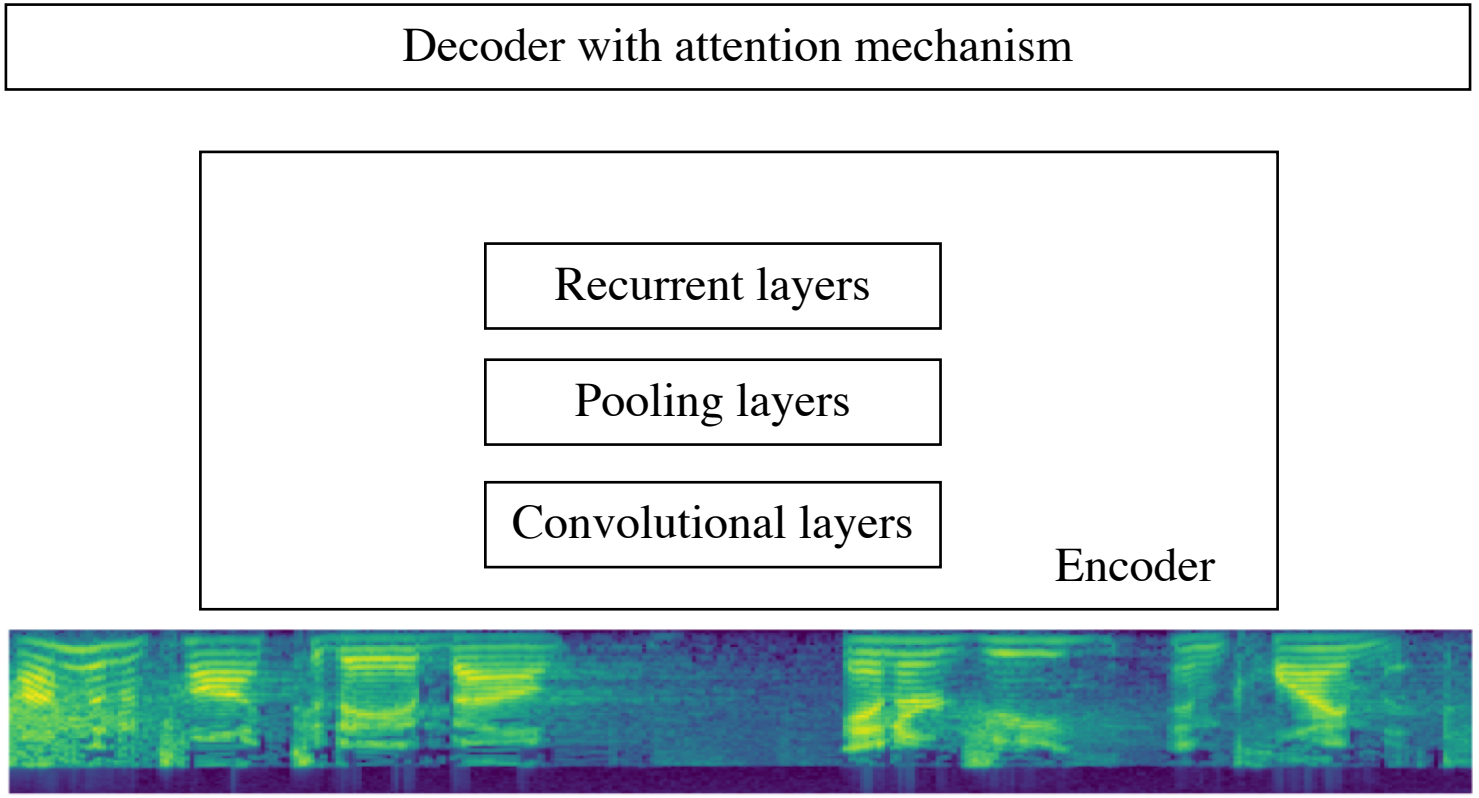
Automatic Speech Recognition



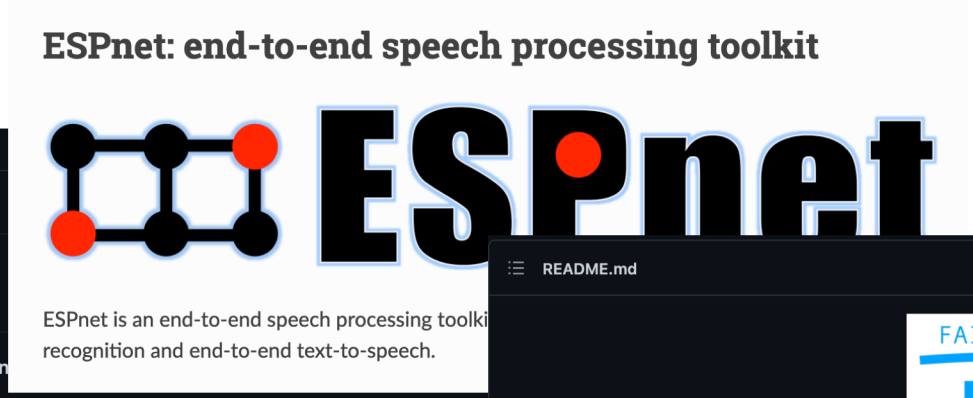


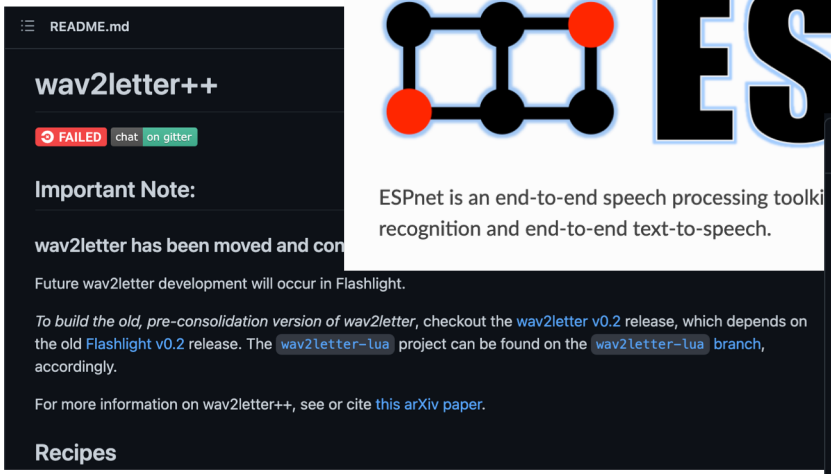
Automatic Speech Recognition

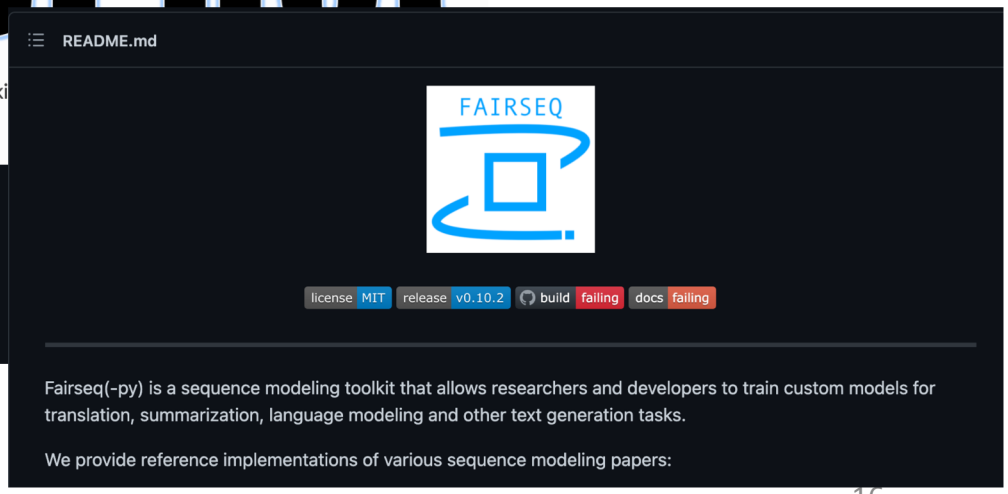
- Seq2Seq model



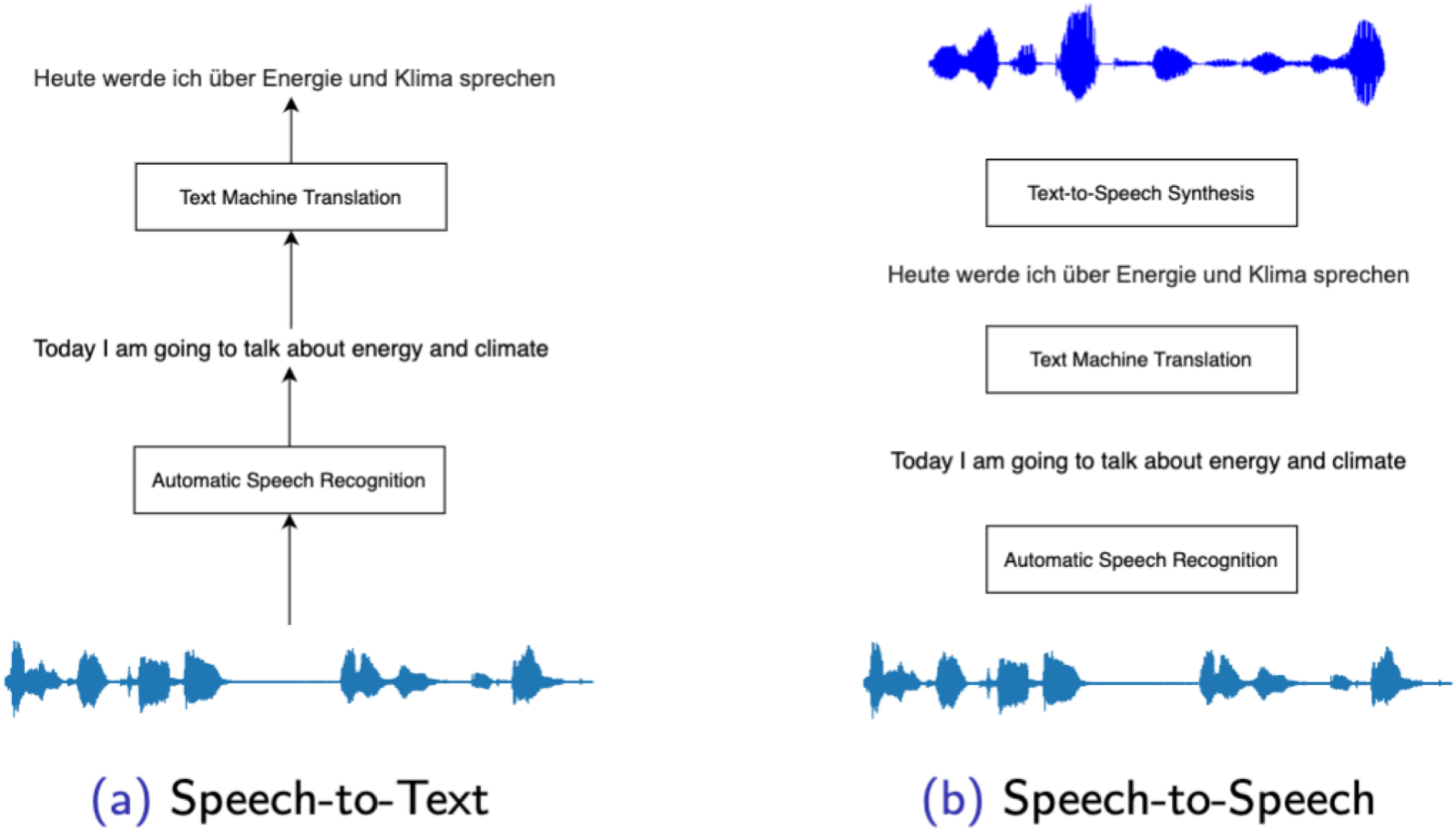
Automatic Speech Recognition







Cascade Speech Translation



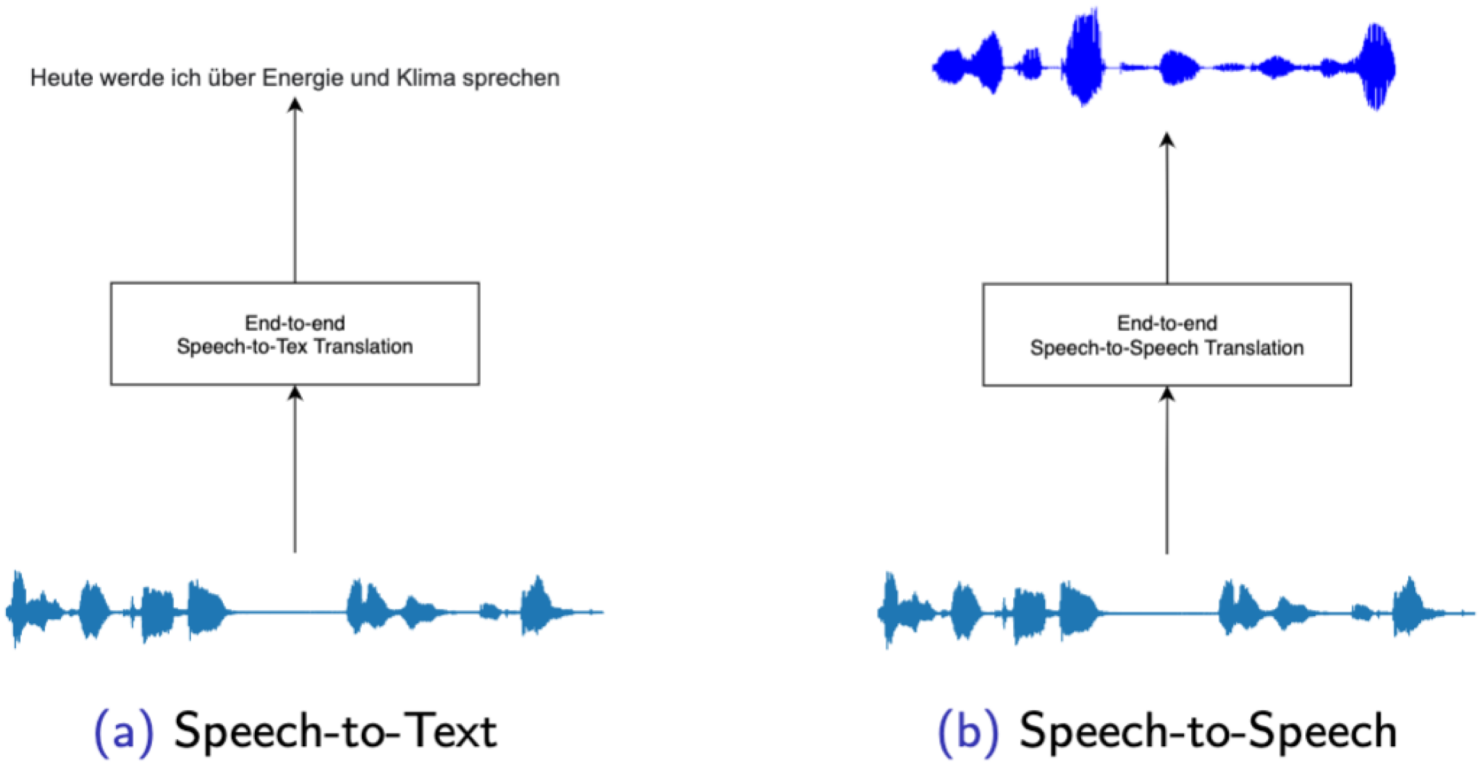
Waibel et al., 1991; Woszczyna et al., 1993; Vidal, 1997; Wang and Waibel, 1998; Takezawa et al., 1998; Ney, 1999; Bangalore and Riccardi, 2001; Fu-Hua Liu et al., 2003; Schultz et al., 2004; Matusov et al., 2005; Bertoldi and Federico, 2005; Zhang et al., 2005; Pérez et al., 2007; Sperber et al., 2017, 2019; Zhang et al., 2019; Beck et al., 2019; Black et al., 2002; Sumita et al., 2007

Cascade Speech Translation

- Pros
  - Easy to build (ASR + MT or ASR + MT + TTS)
  - More training data
    - Different data for ASR and MT

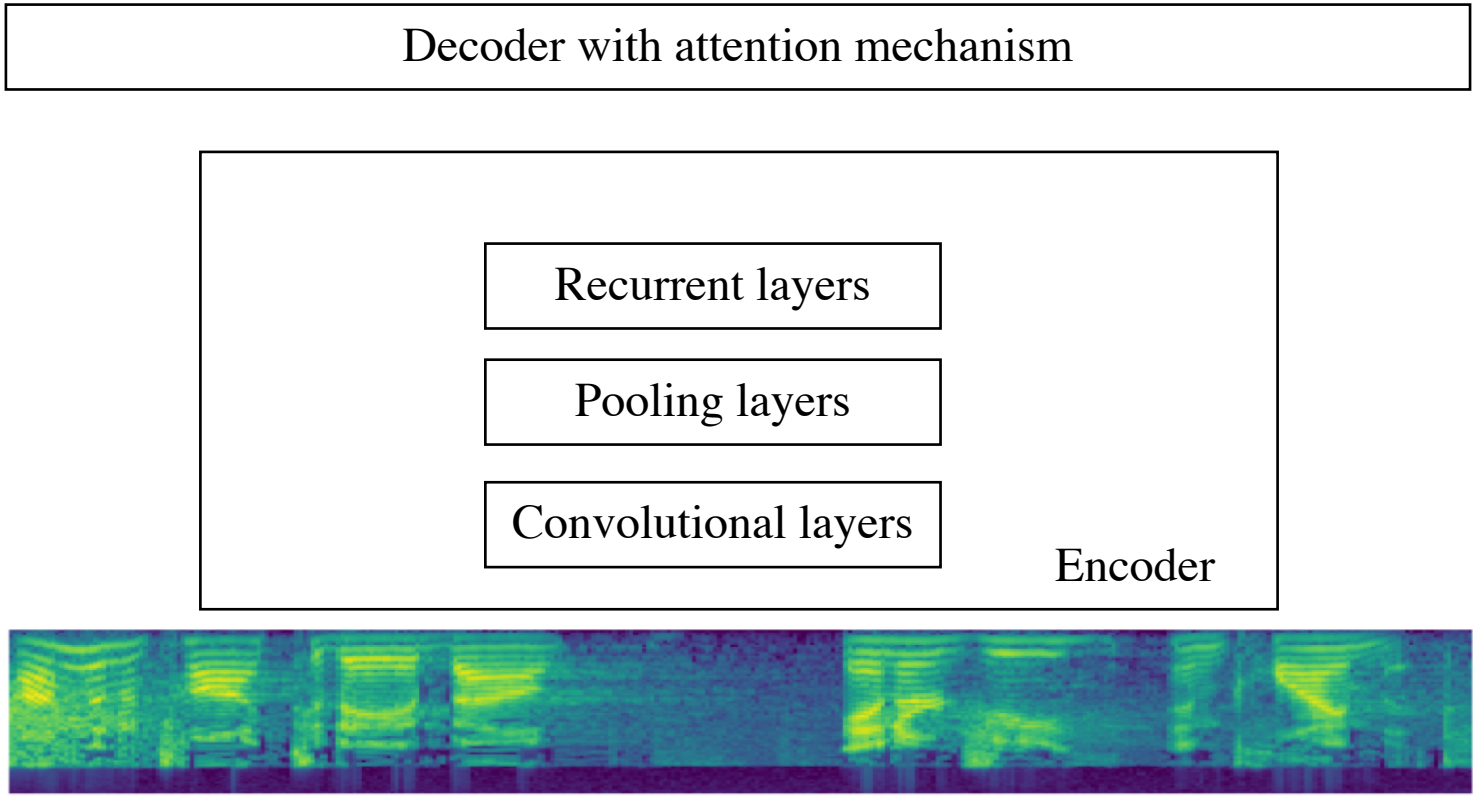
- Cons
  - Model size
  - Inference latency
  - Compounding errors

End-to-end Speech Translation

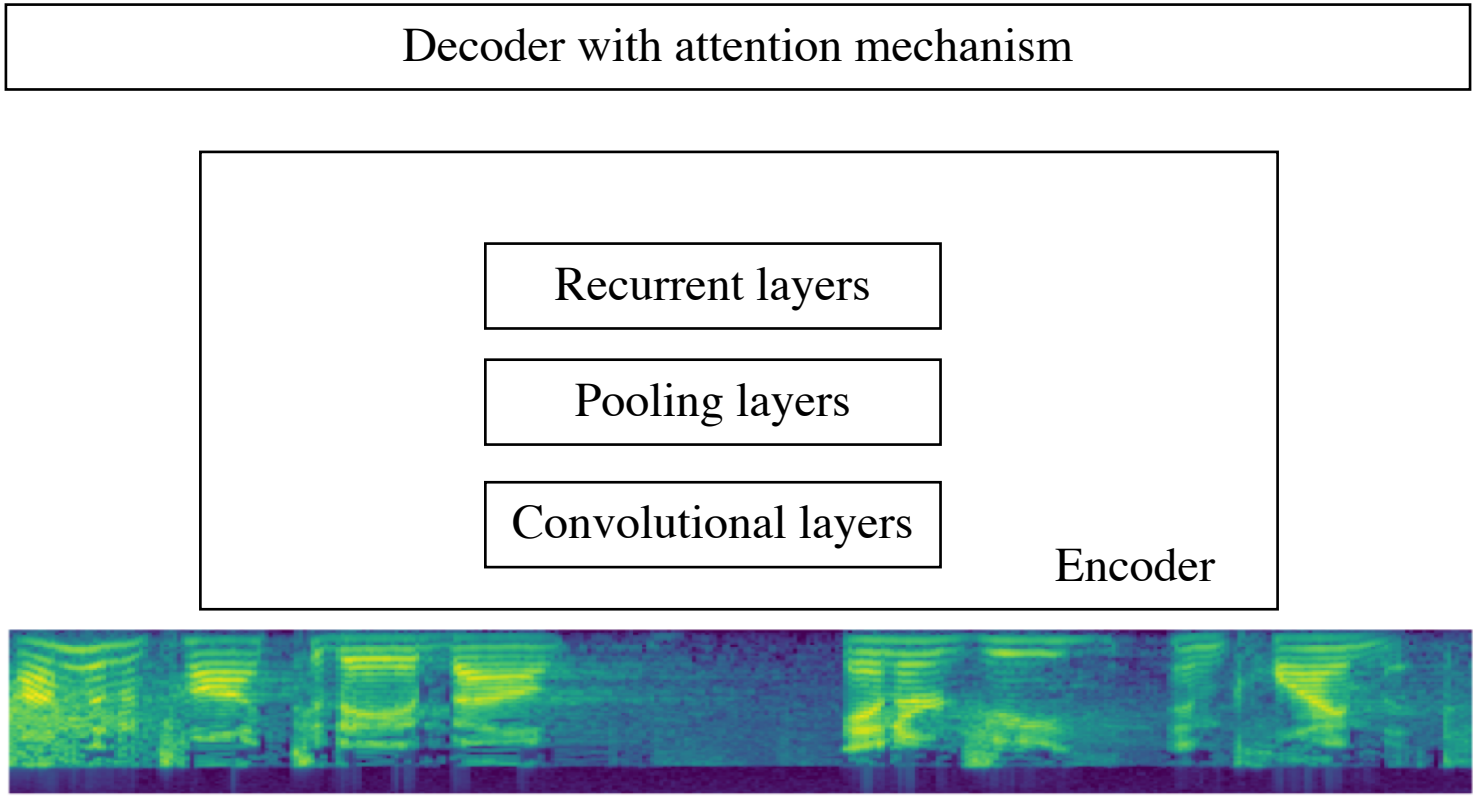


Duong et al. (2016); Berard et al. (2016); Weiss et al. (2017); Bansal et al. (2018); Di Gangi et al. (2019b); Pino et al. (2020); Inaguma et al. (2020)

Automatic Speech Recognition



End-to-End Speech Translation



End-to-End Speech Translation

- Pros
  - Small model size
  - Lower inference latency
  - No Compounding errors
- Cons
  - Data!

Data Scarcity

- Data is more difficult to collect and annotate
  - Parallel speech to text / speech

Tgt	#Talk	#Sent	Hours	src w	tgt w
De	2,093	234K	408	4.3M	4.0M
Es	2,564	270K	504	5.3M	5.1M
Fr	2,510	280K	492	5.2M	5.4M
It	2,374	258K	465	4.9M	4.6M
Nl	2,267	253K	442	4.7M	4.3M
Pt	2,050	211K	385	4.0M	3.8M
Ro	2,216	240K	432	4.6M	4.3M
Ru	2,498	270K	489	5.1M	4.3M

Speech Translation Dataset

	#sent
WMT 16 EN-DE	~4M
WMT 14 EN-FR	~14M

Machine Translation Dataset

Di Gangi, Mattia A., et al. "Must-c: a multilingual speech translation corpus." *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019.

Data Scarcity

- Data augmentation
- Multi-task
  - Share components with other models (ASR / MT)
- Multi-lingual training
- Pretrained components
- Self-learning
  - Train on synthesized data

Pino, Juan, et al. "Self-training for end-to-end speech translation." *arXiv preprint arXiv:2006.02490* (2020).
Inaguma, Hirofumi, et al. "Multilingual end-to-end speech translation." *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019.
Kano, Takatomo, Sakriani Sakti, and Satoshi Nakamura. "End-to-end speech translation with transcoding by multi-task learning for distant language pairs." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 1342-1355.

End-to-End Speech Translation

	Model	De	Pt	Fr	Es	Ro	Ru	Nl	It
E2E	Transformer + ASR encoder init.[1,♣]	17.30	20.10	26.90	20.80	16.50	10.50	18.80	16.80
	ESPnet-ST (Transformer)								
	ASR encoder/MT decoder init.	22.33	27.26	31.54	27.84	20.91	15.32	26.86	22.81
	+ SpecAugment	**22.91**	**28.01**	**32.69**	**27.96**	**21.90**	**15.75**	**27.43**	**23.75**
Cascade	Transformer ASR → Transformer MT[1]	18.5	21.5	27.9	22.5	16.8	11.1	22.2	18.9
	ESPnet-ST								
	Transformer ASR → Transformer MT	**23.65**	**29.04**	**33.84**	**28.68**	**22.68**	**16.39**	**27.91**	**24.04**

Inaguma, Hirofumi, et al. "ESPnet-ST: All-in-one speech translation toolkit." *arXiv preprint arXiv:2004.10234* (2020).

End-to-End Speech Translation

- Open source toolkits

FAIRSEQ S2T: Fast Speech-to-Text Modeling with FAIRSEQ

Changhan Wang[1], **Yun Tang**[1], **Xutai Ma**[1,2], **Anne Wu**[1], **Dmytro Okhonko**[1], **Juan Pino**[1]

[1]Facebook AI
[2]Johns Hopkins University

xutai_ma@jhu.edu
{changhan,yuntang,xutaima,annewu,oxo,juancarabina}@fb.com

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

Hirofumi Inaguma[1] **Shun Kiyono**[2] **Kevin Duh**[3] **Shigeki Karita**[4]
Nelson Yalta[5] **Tomoki Hayashi**[6,7] **Shinji Watanabe**[3]
[1] Kyoto University [2] RIKEN AIP [3] Johns Hopkins University
[4] NTT Communication Science Laboratories [5] Waseda University
[6] Nagoya University [7] Human Dataware Lab. Co., Ltd.
inaguma@sap.ist.i.kyoto-u.ac.jp

Simultaneous Speech Translation

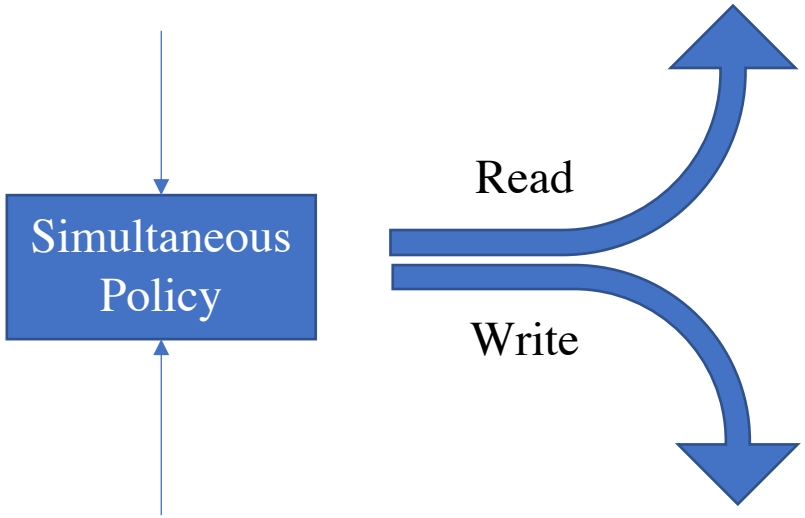
- Start the translation before read all the input speech

I am going to talk today about energy and climate.

Heute spreche ich zu Ihnen über Energie und Klima.

Simultaneous Speech Translation

Input I am going to talk today about energy and climate.

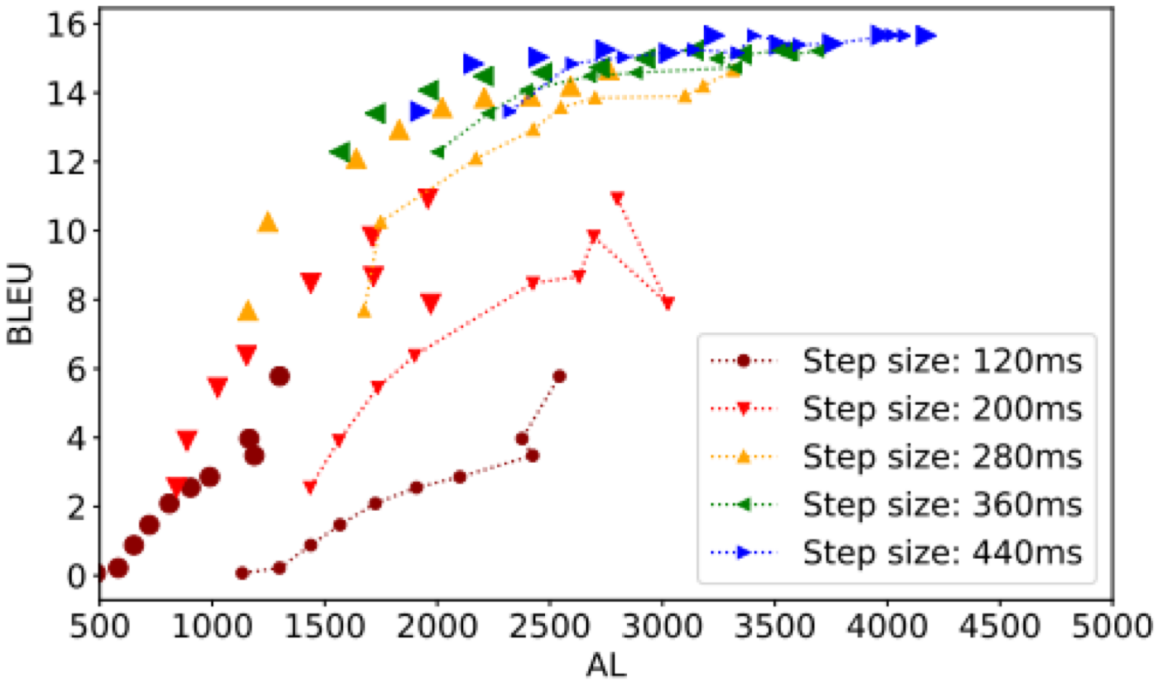


Output Heute spreche ich zu Ihnen über Energie und Klima.

Simultaneous Translation Policies

- Reinforcement learning (Gu et al. 2017; Luo et al. 2017; Lawson et al. 2018)
  - Less stable learning process.
- Fixed policy (Cho and Esipova 2016; Ma et al. 2019a)
  - Weaker performance, for instance Wait-K (Ma et al. 2019a).
- Monotonic attention (Raffel et al., 2017; Arivazha-gan et al., 2019; Ma et al., 2020)
  - The State of the art for the task.

Quality-Latency Trade-off

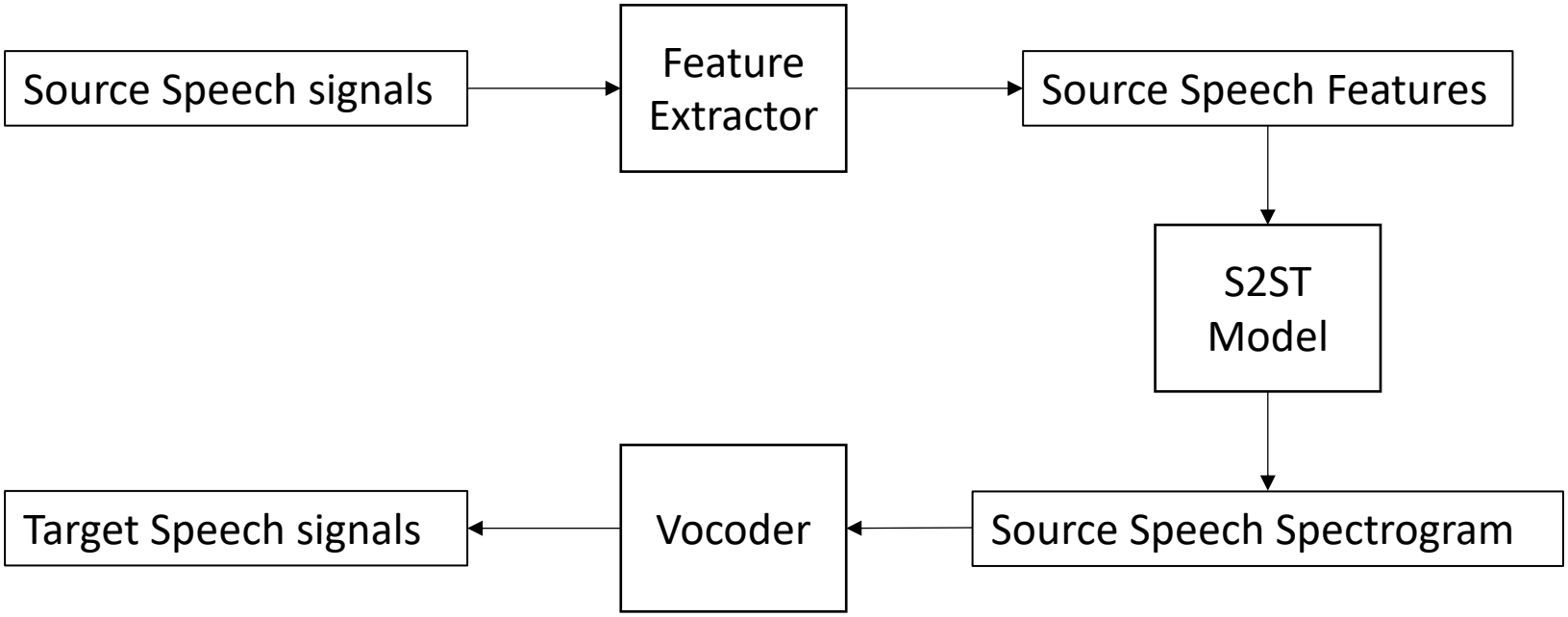


The lagging behind an oracle/perfect system

Speech-to-Speech Translation

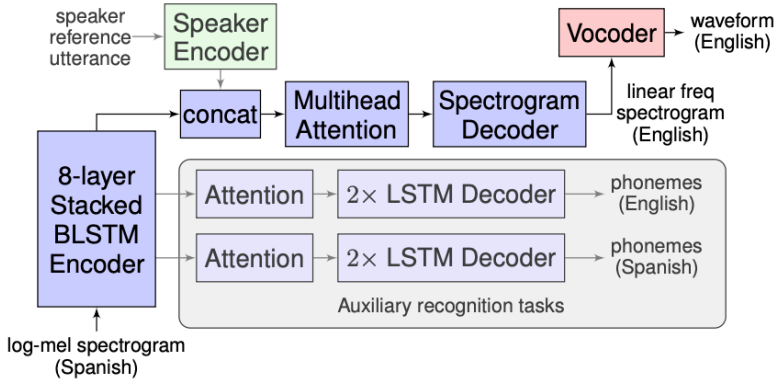
- Source speech → target speech
- Cascade
  - ST + Text-to-speech (TTS) system
- End-to-end (direct)
  - Directly generate target spectrogram
  - Preserve prosody, emphasis, emotion
  - Suffers more from data scarcity

Direct Speech-to-Speech Translation



Direct S2ST with Sequence-to-Sequence Model

- Speech encoder from ASR & ST
- Spectrogram decoder from TTS
- Multi-task learning
- Examples



Y. Jia et al., "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model." Proc. Interspeech 2019

Conclusion

- Speech translation is a new and challenging task
- End-to-end approach shows strength when overcome the data scarcity issue.
- More challenges: low latency, speech-to-speech, etc.