Speech Translation

An Introduction

Ian Neidel (ian.neidel@yale.edu)
CS 677: Advanced Natural Language Processing
Dr. Dragomir Radev
November 4, 2021

Why Speech Translation

- Videos
  - Internet: e.g. Youtube, Facebook
  - Television shows or movies
  - Lectures
- Real life
  - Telephone calls or meetings
  - Tourist interactions
  - Medical care
  - Use with authorities or international crisis response

Why Speech Translation

- Integration of ASR, MT, TTS for cascaded systems
- Interesting concepts for end-to-end Speech Translation
  - Corpus augmentation, unique representations, etc.
- Challenges
  - Disfluencies, segmentation, simultaneous translation, etc.

Overview

- Background
- Cascaded Speech Translation
- End-to-end Speech Translation
- Comparison
- Papers
  - Translatotron for speech -> speech
  - Fine-tuning pretrained models for speech->text
  - Translatotron 2.0 if time

Problem Varieties

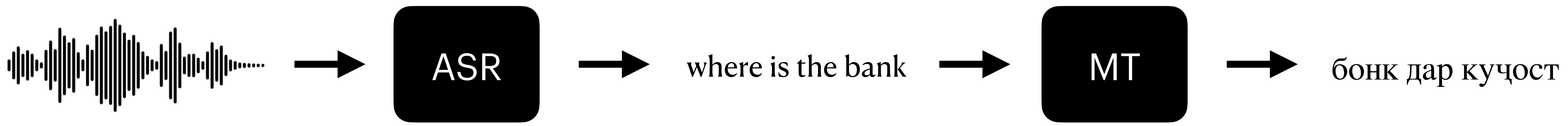
- Sequence
  - Consecutive translation
  - Simultaneous translation
- Number of speakers
  - Single (presentation)
  - Multiple (meeting)
- Output
  - Text
  - Audio
- Model use
  - Online, offline, device capabilities, etc.

Problem Varieties

- Difficulty:
  - Audio quality
  - Speed requirements
  - Domain size
  - Resource availability
  - Speaker variety

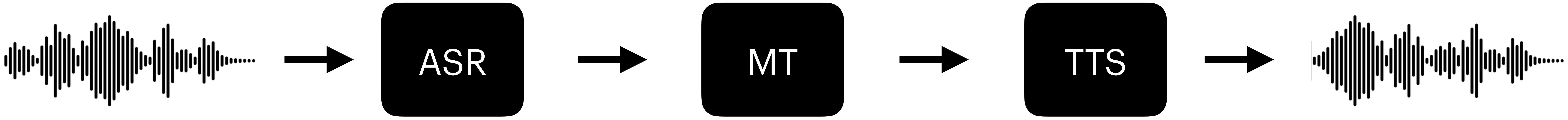
Cascaded Approach

- Combination of several models
  - Automatic Speech Recognition (ASR)
  - Machine Translation (MT)
  - Text-to-Speech (TTS)



Cascaded Approach

- Combination of several models
    - Automatic Speech Recognition (ASR)
    - Machine Translation (MT)
    - Text-to-Speech (TTS)



where is the bank

бонк дар куҷост

Cascaded Approach

- Combination of several models
    - Automatic Speech Recognition (ASR)
    - Segmentation
    - Machine Translation (MT)
    - Text-to-Speech (TTS)

Cascaded Approach

- Combination of several models
    - Automatic Speech Recognition (ASR)
    - Segmentation
    - Machine Translation (MT)
    - Text-to-Speech (TTS)

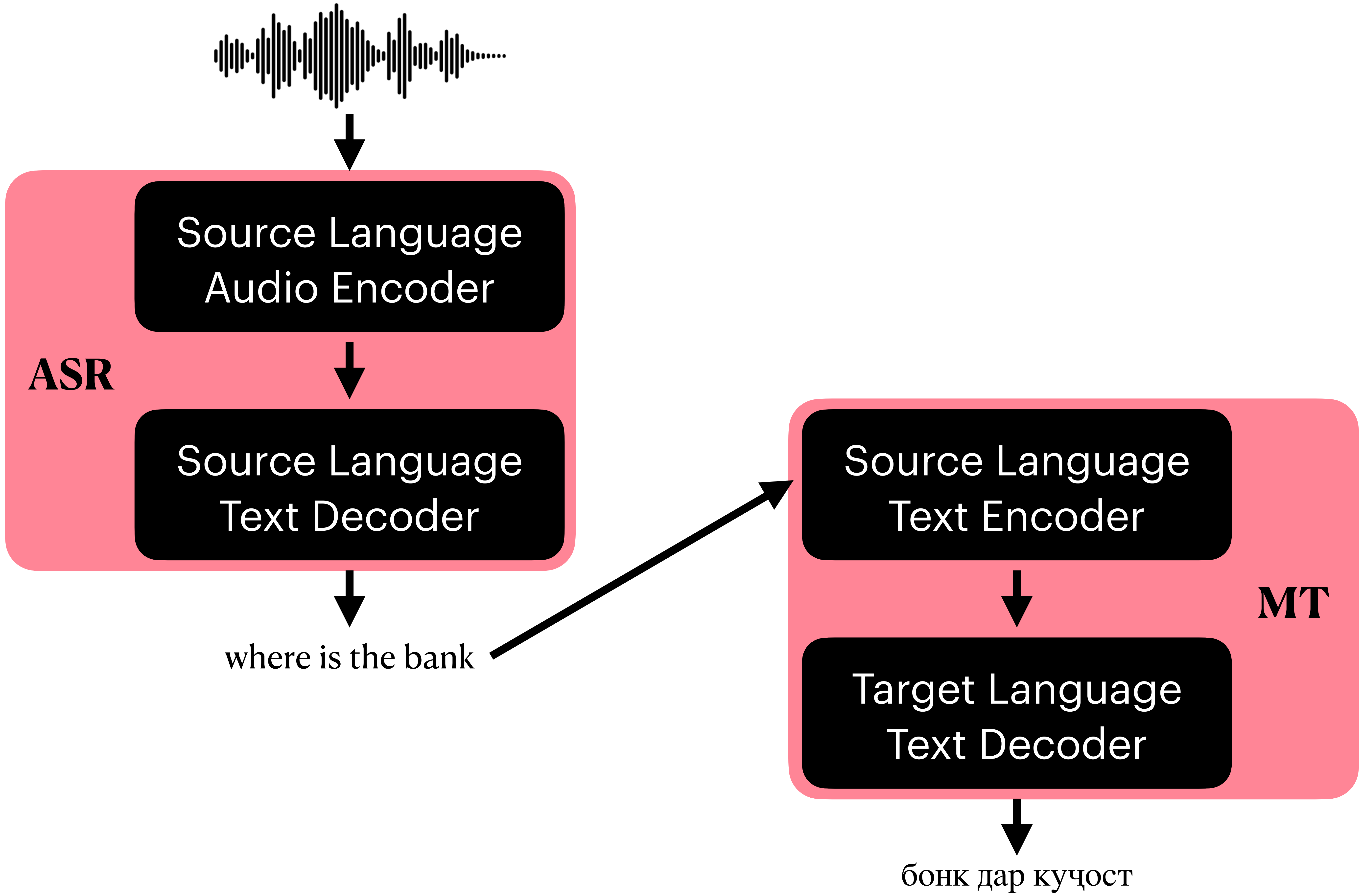
Cascaded Approach

- Advantages:
  - Modularity
  - Quantity of data for each part
  - Easy to incorporate new cutting-edge systems

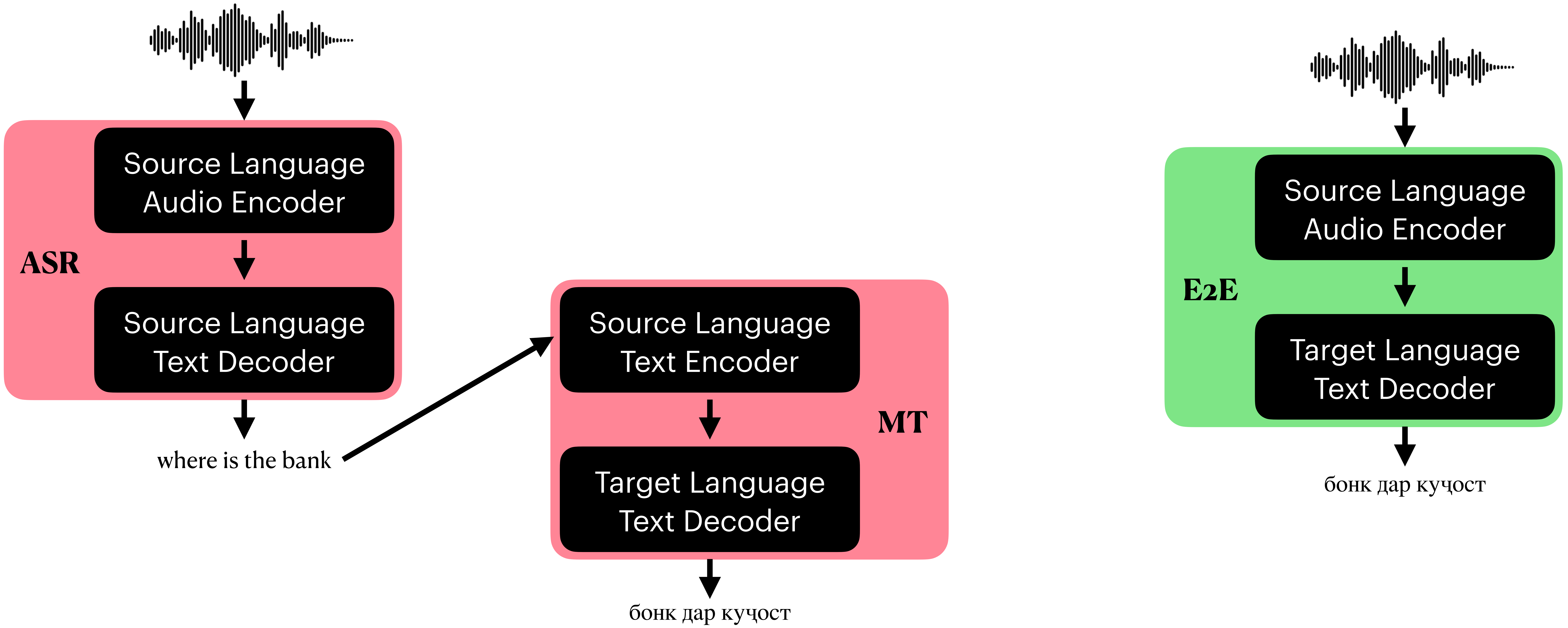
Cascaded Approach

- Issues:
  - Error Propagation
    - Ignore the errors
    - Test options (n-best lists; lattices)
    - Create more robust systems (e.g. train with noise)

End-to-end Speech Translation

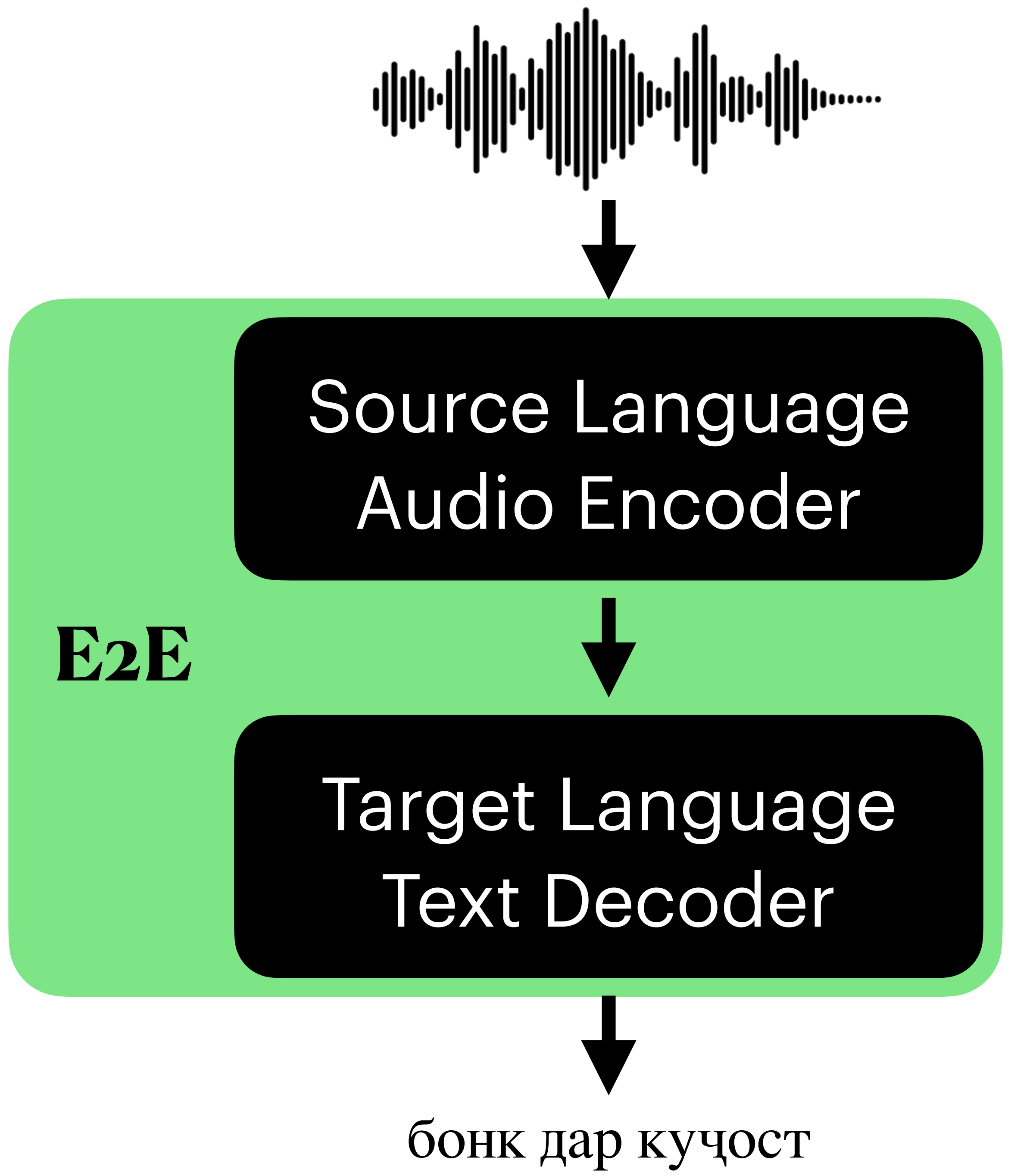


End-to-end Speech Translation



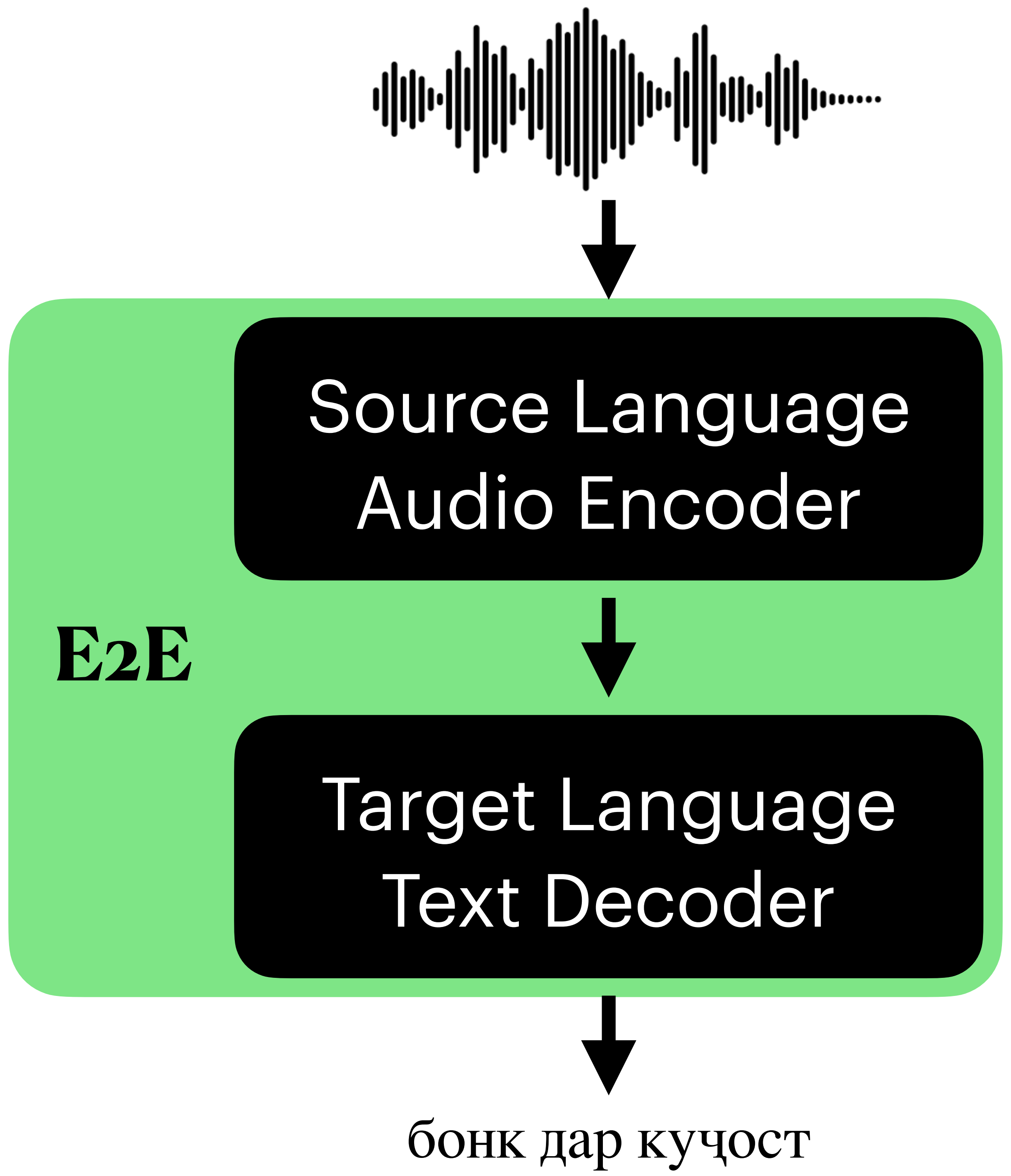
End-to-end Speech Translation

- Shown to be possible to learn audio source -> target text
  - Duong et al. 2016
- As of 2018, worse performance although great promise
- As of 2021, the gap has been closed for some languages



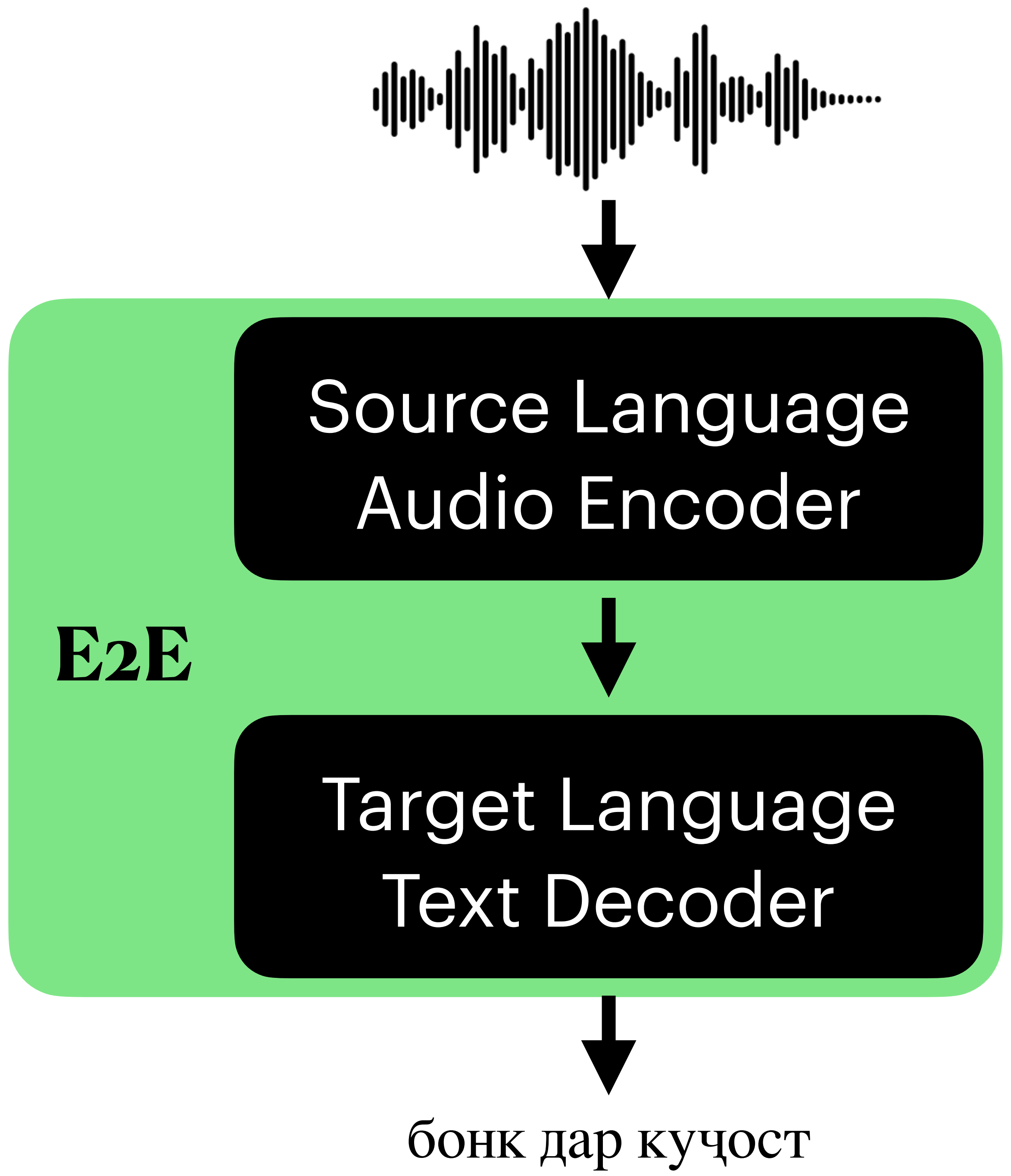
End-to-end Speech Translation

- Challenges
  - Audio signal input
    - Long sequences, different dependencies
    - May have to use ASR techniques
  - Very little data
  - Mapping is less straightforward
    - Use source transcript as intermediate during training?

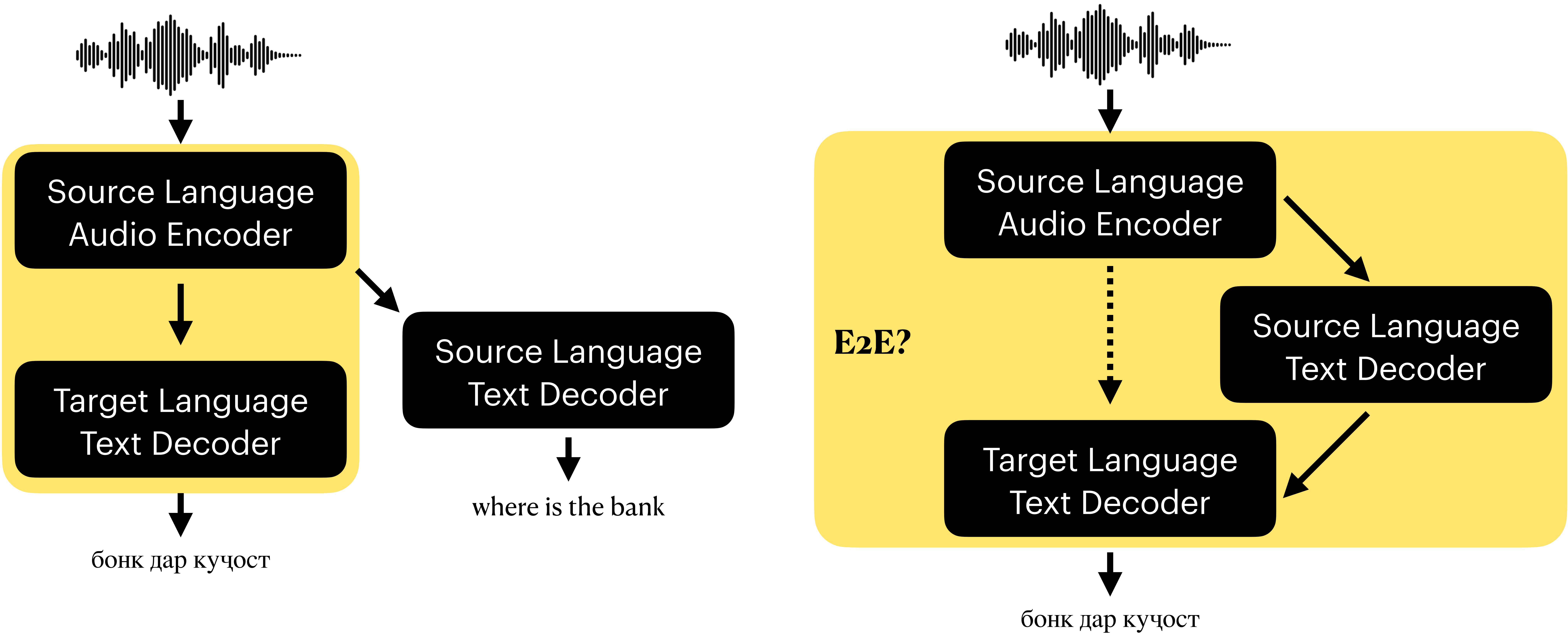


End-to-end Speech Translation

- Approaches
  - Synthetic data
    - TTS of parallel corpora for audio
  - Multitasking
    - Allow network to produce many outputs
      - e.g. as a MT, ASR system
  - Pre-training
    - Train encoder with ASR, train new encoder for ST



End-to-end Speech Translation



End-to-end Speech Translation

- Other options
  - Stack decoders that attend to source language hidden states
  - Shared context vectors

End-to-end Speech Translation

- Challenges:
  - Data efficiency
  - Small datasets

Other Considerations

- Importance of segmentation:
  - Audio is continuous
    - Out of the box MT at sentence+ level
  - No explicit punctuation in audio
    - Semantic differences
      - “I love cooking, my cats, and my dogs” | “I love cooking my cats and my dogs”
      - “I’m sorry. I love you.” | “I’m sorry I love you”
  - Can be done after ASR, as part of MT, after MT

Other Considerations

- Simultaneous translation:
  - Reducing latency improves experience => translate as soon as possible
  - Context improves accuracy of ASR and MT => wait as much as possible
- How to deal with different word orders?
  - e.g. SOV, VSO, VOS

German	Ich	melde	mich	zur	Konferenz	an
Gloss	I	Register/ cancel	Myself	To	Conference	
English	I	?				

Other Considerations

- Simultaneous translation:
  - Train models to optimally segmenting input
    - Loss with segment length and quality metrics
  - Stream decoding
    - Output word or wait for next
    - Use a decision model or fixed schedule
  - Update translations on the go

Other Considerations

- Spontaneous speech:
  - Disfluencies
    - Filled pauses e.g. “She was, uh, certain about it”
    - Repetition e.g. “He he really wanted to go”
    - Insertion e.g. “He wanted really wanted to go”
    - Trailing off e.g. “And then…”
    - Error e.g. “They misunderestimated me”
    - Filler words, etc.
  - Where to deal with disfluencies?
    - Special model/MT/…

History

- 80's: proof of concept with restricted domain, controlled speaking style
- 90's: spontaneous ST systems
- 2003-6: open domain ST systems and new languages e.g. Zh, Ar
- 2005: first ST corpora
- 2006: simultaneous ST
- 2016: first E2E ST models
- 2018-19: E2E 8.7—1.6 BLEU pts below cascade ST for En-De
- 2020: 0.2 BLEU pts above cascade ST for En-De

End-to-end benefits

- Prevent error propagation
- Preserve information e.g. through prosody

Speech transcription	those are their expectations of who you are not yours
Target reference	那 是 他们 所期望的 你的 样子 而不是 你自己的 期望 *that is they expected your appearance not yourself expectation*
Cascade-ASR	those are **there** expectations **to do** you are not yours
Cascade-Translation	那些 都是 希望 做到的， 你 不是 你的 。 *those are expect achievement you not yours*
FAT-ST	这些 是 他们 对 你的 期 望， 而不是 你的 期望 。 *these are they to your expectation not your expectation*

Example from https://arxiv.org/abs/2102.05766

End-to-end benefits

- Prevent error propagation
- Preserve information e.g. through prosody

English	Japanese
this is my niece , lucy	*kochira wa suekko no lucy desu* こちら は 姪っ子 の ルーシー です 。
this is my niece , lucy	*lucy, kono ko ga watashi no suekko desu* ルーシー 、 この 子 が 私 の 姪っ子 です 。
will you have /cheese or /jam	*chiizu toka jamu toka, dore ni shimasu ka* チーズ とか ジャム とか、 どれ に します か ？
will you have /cheese or \jam	*chiizu ka jamu, docchi ni shimasu ka* チーズ か ジャム、 どっち に します か ？

Example from https://arxiv.org/pdf/2004.06358.pdf

Discussion Questions

- Should having access to the original audio improve translation? How? When?

Discussion Questions

- Should having access to the original audio improve translation? How? When?
- Does end-to-end speech translation avoid error propagation? How can we compare error propagation between E2E and cascaded models?

Input

- Use sampling or windowing
- Mel-Frequency Cepstral Coefficients (MFCC)
- Log mel-filterbank features (FBANK)
- Sequence length issues:
  - IWSLT test set 2020
    - Segments: 1804
    - Words: 33,795
    - Characters: 149.053
    - Features: 1,471,035

Output

- Words
- Byte Pair Encodings (BPE)
- Characters

Datasets

Dataset	Paper	Languages and Duration	Domain
(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–CALLHOME	(Post et al. 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	(Boito et al. 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

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

Data Augmentation

- From ASR:
  - Noise injection
  - Speed perturbation
  - Time masking
  - Frequency masking

Paper 0:
Direct speech-to-speech translation with a sequence-to-sequence model

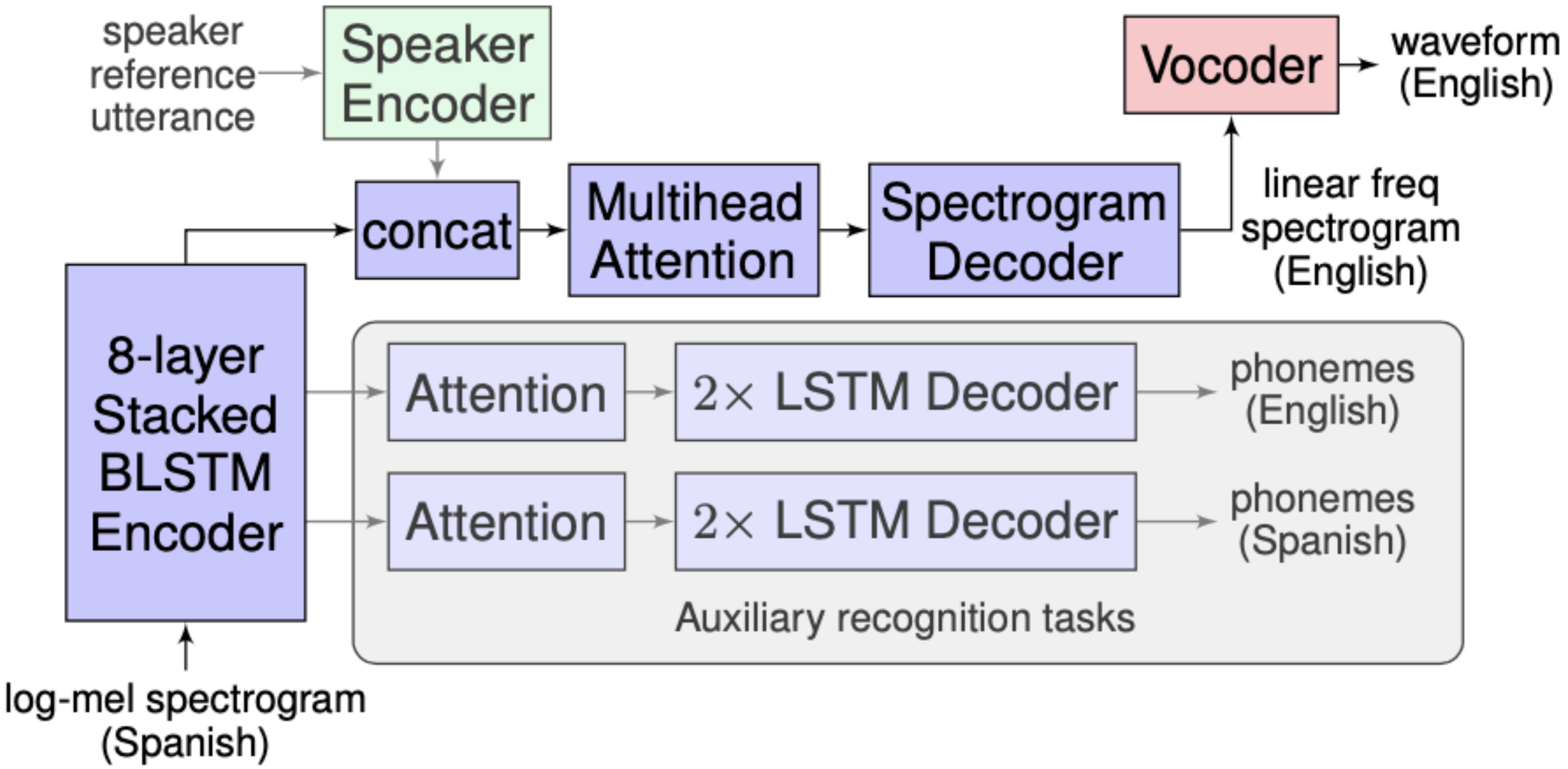
Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, Yonghui Wu
(Google; Interspeech 2019)

https://arxiv.org/abs/1904.06037

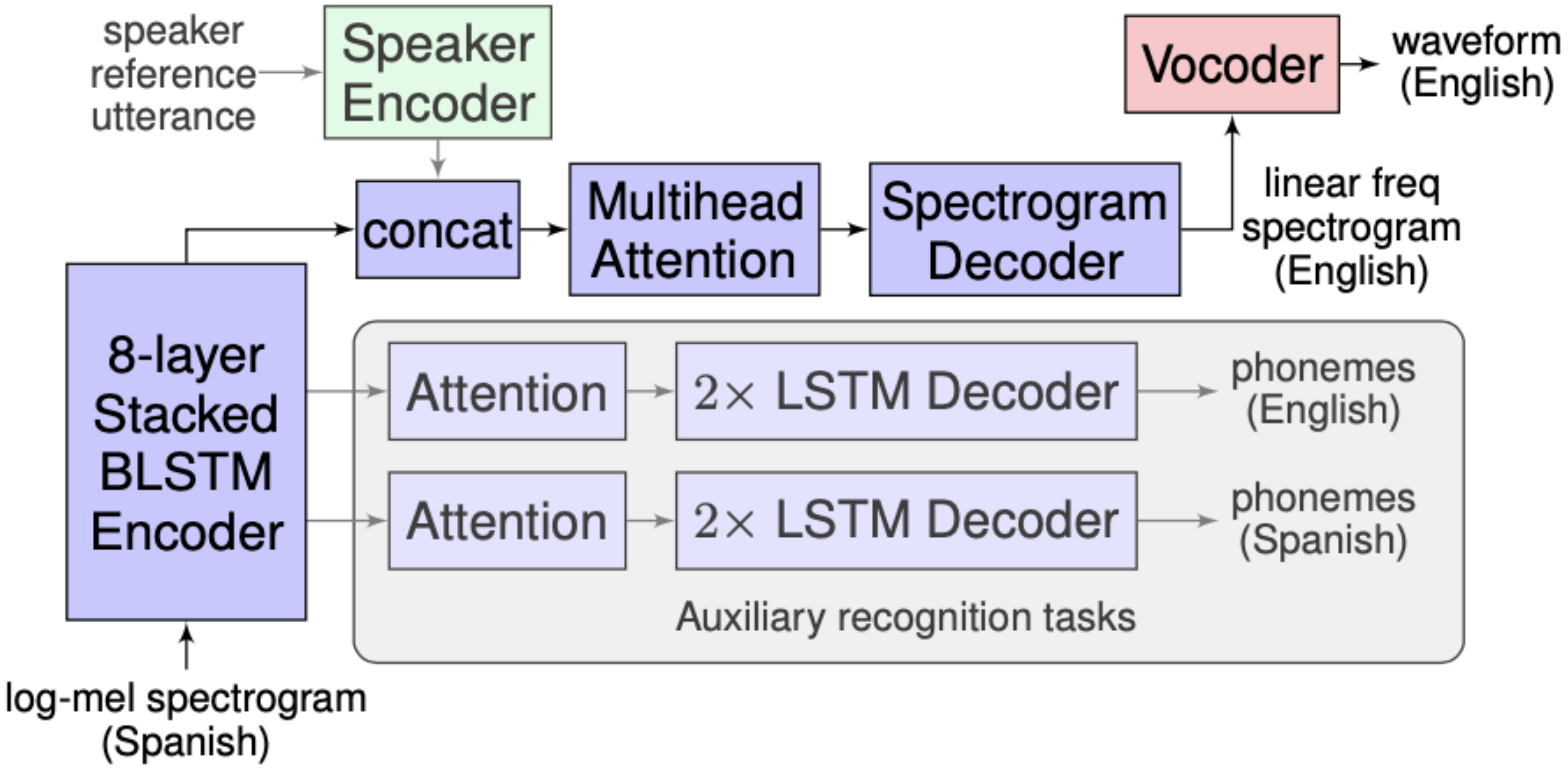
Outline

- Translatotron
  - First end-to-end speech-speech model
  - No intermediate text representation
  - Many-to-many speaker configurations
  - Similar to Tacotron 2 TTS
  - Performs worse than cascaded but is proof of concept and demonstrates benefits

S2S Translation Model



S2S Translation Model



- Separately trained components
  - Attention-based sequence-to-sequence network
  - Vocoder
  - Optional speaker encoder

S2S Translation Model

- S2S encoder
  - Stack maps 80-channel log-mel spectrogram input features into hidden states
  - Passed through an attention based alignment mechanism to condition an autoregressive decoder
  - Predicts 1025-dim log spectrogram frames corresponding to the translated speech
  - Auxiliary decoders, each with their own attention components, predict source and target phoneme sequences

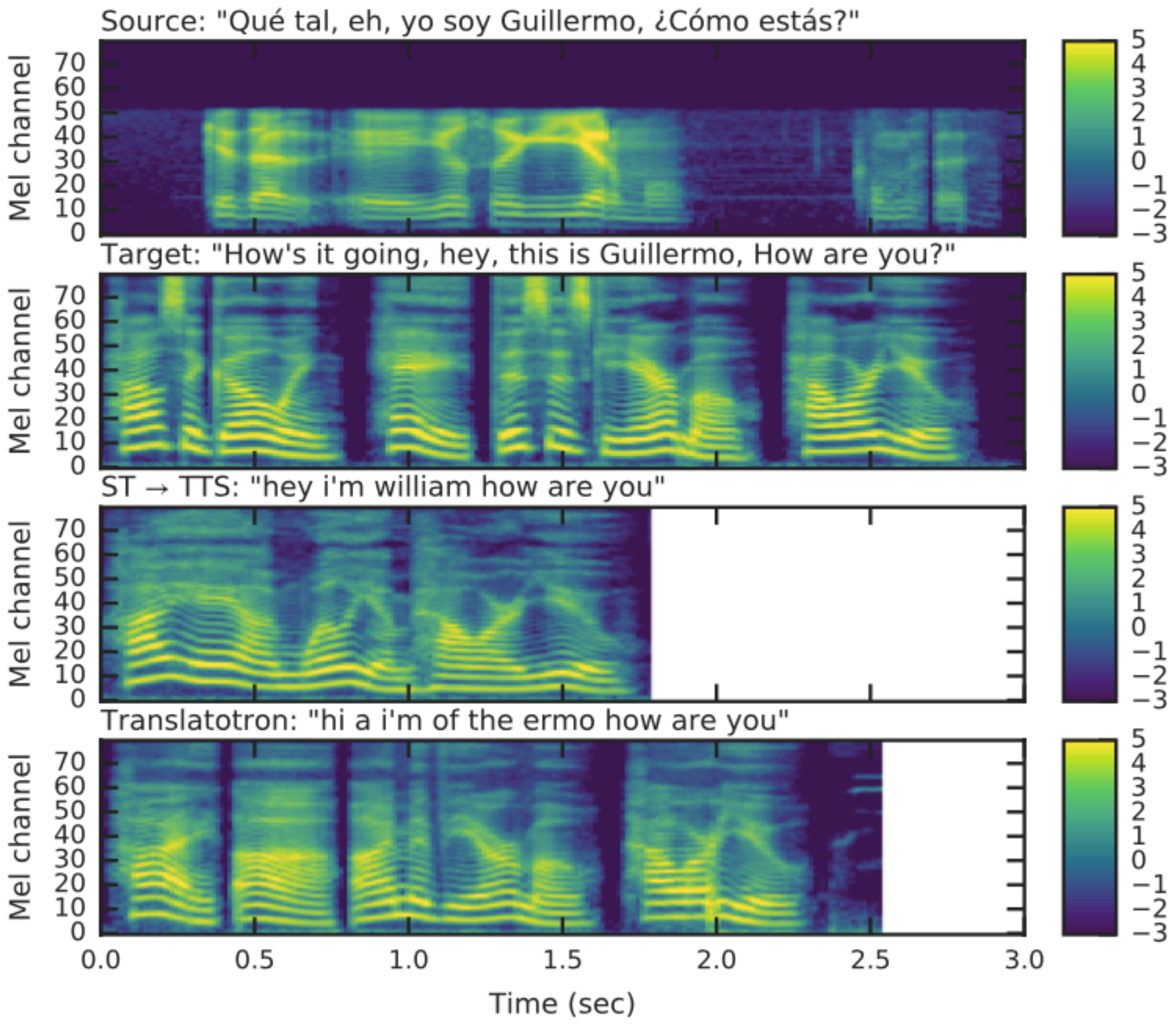
Training

- Multitask training
  - Auxiliary decoder networks to predict phoneme sequences for source/target speech
  - Predict source/target transcripts
- English-Spanish conversational dataset
  - Parallel text and read speech
- English-Spanish Fisher Dataset
  - From telephone conversations

Evaluation

- Run ASR on the output audio
- BLEU score with reference transcription
- This acts a lower bound on performance
- Mean opinion score on speech naturalness and voice transfer

Example



Source: "Qué tal, eh, yo soy Guillermo, ¿Cómo estás?"
Target: "How's it going, hey, this is Guillermo, How are you?"
ST → TTS: "hey i'm william how are you"
Translatotron: "hi a i'm of the ermo how are you"
Mel channel
Time (sec)

Listen to Results

- https://ai.googleblog.com/2019/05/introducing-translatotron-end-to-end.html
- https://google-research.github.io/lingvo-lab/translatotron/

Qualitative results

- Can reproduce disfluencies
- Reproduces sounds much better
  - Preserves Guillermo rather than William
  - However, preserves some Spanish sounds e.g. in Dan
  - May have a bias for cognates “pasejos” -> “passages” not “tickets”

Quantitative Results

Auxiliary loss	dev1	dev2	test
None	0.4	0.6	0.6
Source	7.4	8.0	7.2
Target	20.2	21.4	20.8
Source + Target	24.8	26.5	25.6
Source + Target (1-head attention)	23.0	24.2	23.4
Source + Target (encoder pre-training)	30.1	31.5	31.1
ST [19] $\rightarrow$ TTS cascade	39.4	41.2	41.4
Ground truth	82.8	83.8	85.3

Auxiliary loss	BLEU	Source PER	Target PER
None	0.4	-	-
Source	42.2	5.0	-
Target	42.6	-	20.9
Source + Target	42.7	5.1	20.8
ST [21] $\rightarrow$ TTS cascade	48.7	-	-
Ground truth	74.7	-	-

Speaker Emb	BLEU	MOS-naturalness	MOS-similarity
Source	33.6	3.07 ± 0.08	1.85 ± 0.06
Target	36.2	3.15 ± 0.08	3.30 ± 0.09
Random target	35.4	3.08 ± 0.08	3.24 ± 0.08
Ground truth	59.9	4.10 ± 0.06	-

Takeaways

- Auxiliary loss required
  - Without them it can synthesize simple words/phrases, but mostly synthesizes plausible sounds
  - Issues attending to input, demonstrates the difficulty of S2ST
  - Transcripts improve speech translation training, not needed during inference
  - Auxiliary loss for phonemes improves attention

Discussion Questions

- Does using multitask learning reduce the benefits of an end-to-end system?

Discussion Questions

- Does using multitask learning reduce the benefits of an end-to-end system?
- What kind of (cross-lingual?) voice transfer would we want?

Translatotron 2.0

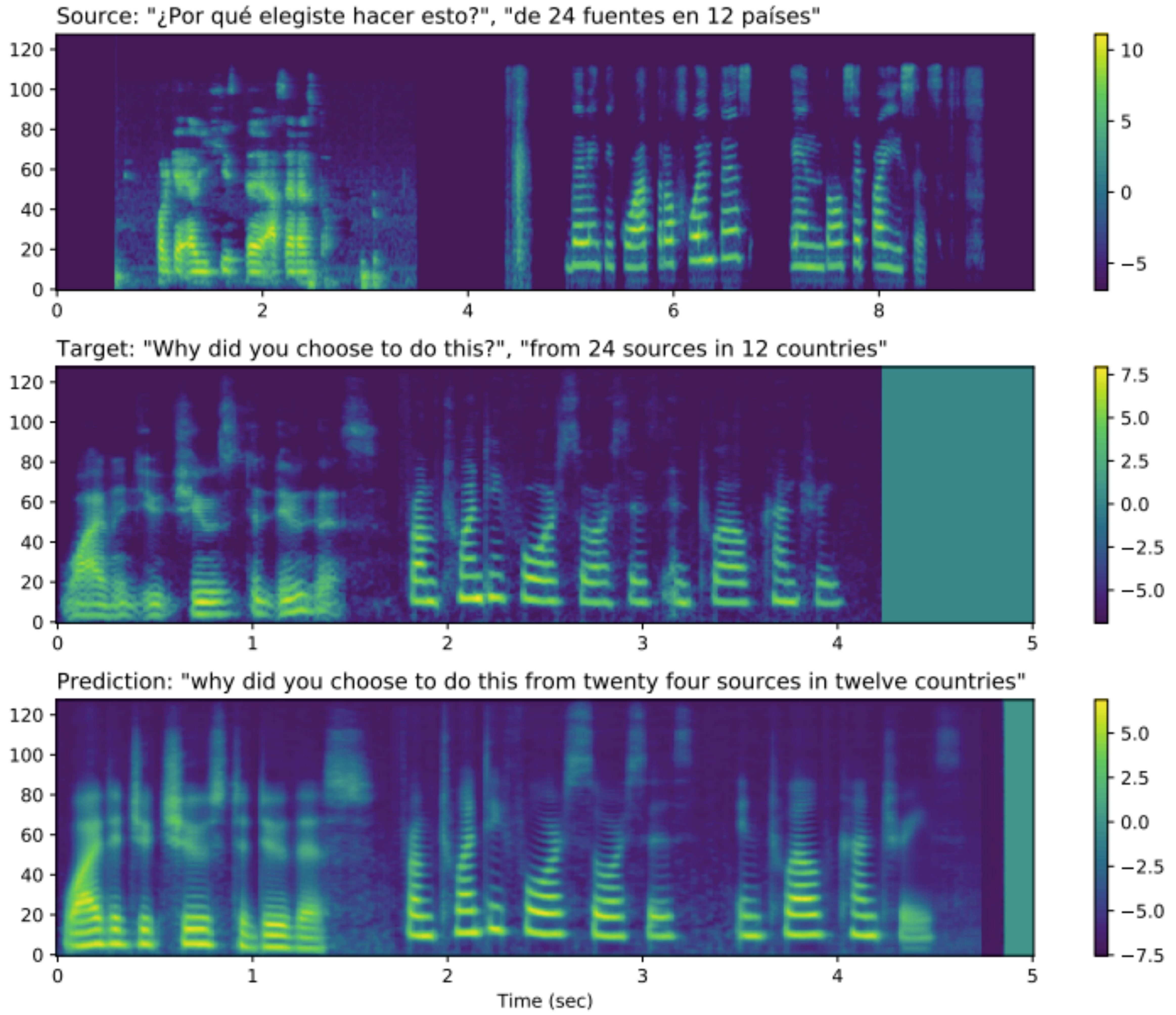
Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, Roi Pomerantz of Google, 2021

https://arxiv.org/abs/2107.08661

- Changes
  - the output from the auxiliary target phoneme decoder is used as an input to the spectrogram synthesizer
  - the spectrogram synthesizer is duration-based, while still keeping the benefits of the attention mechanism
  - Removed target voice transferral
- Large improvement on Translatotron 1.0; on par with cascaded models

Translatotron 2.0

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, Roi Pomerantz of Google, 2021



https://google-research.github.io/lingvo-lab/translatotron2/

Paper 1:
Multilingual Speech Translation with Efficient Finetuning of Pretrained Models

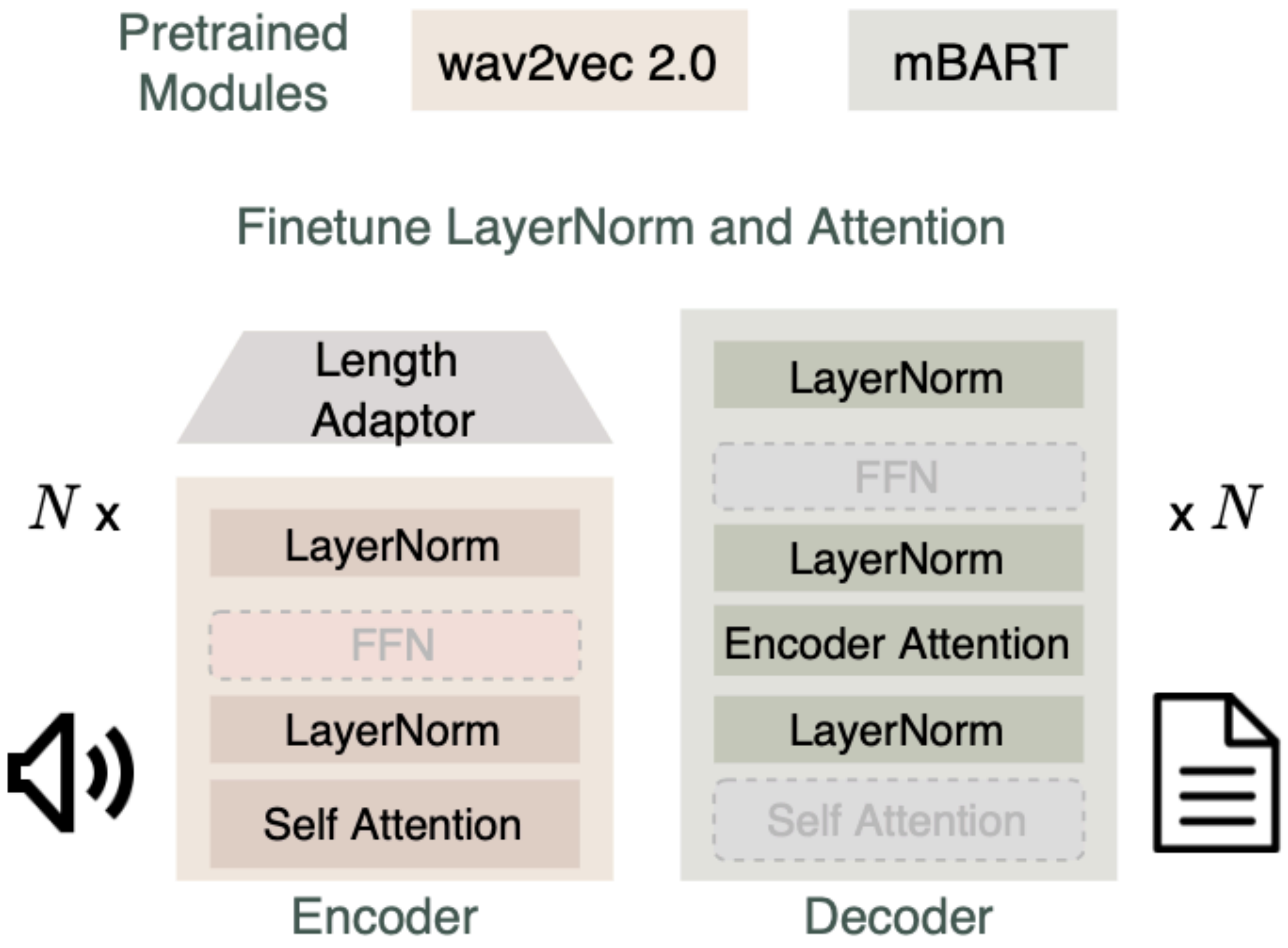
Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, Michael Auli (Facebook AI team; ACL 2021)

https://aclanthology.org/2021.acl-long.68.pdf

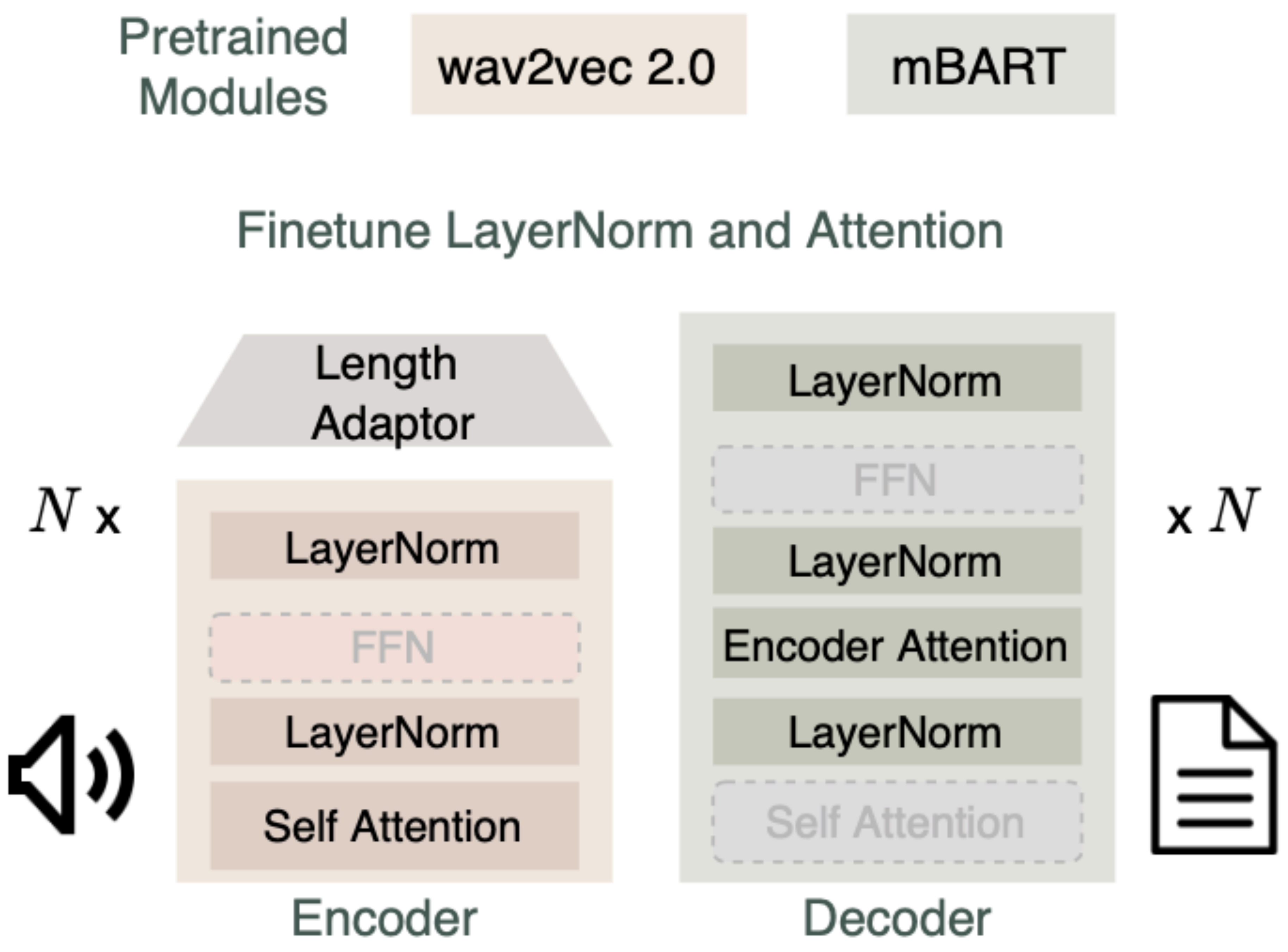
Overview

- Fine tune pre-trained modules to improve data efficiency
- Can achieve zero-shot learning, significantly lowering costs
- Achieves SOTA for 34 translation directions, surpasses cascaded ST in 23
- Creates a many-many multilingual model

Architecture

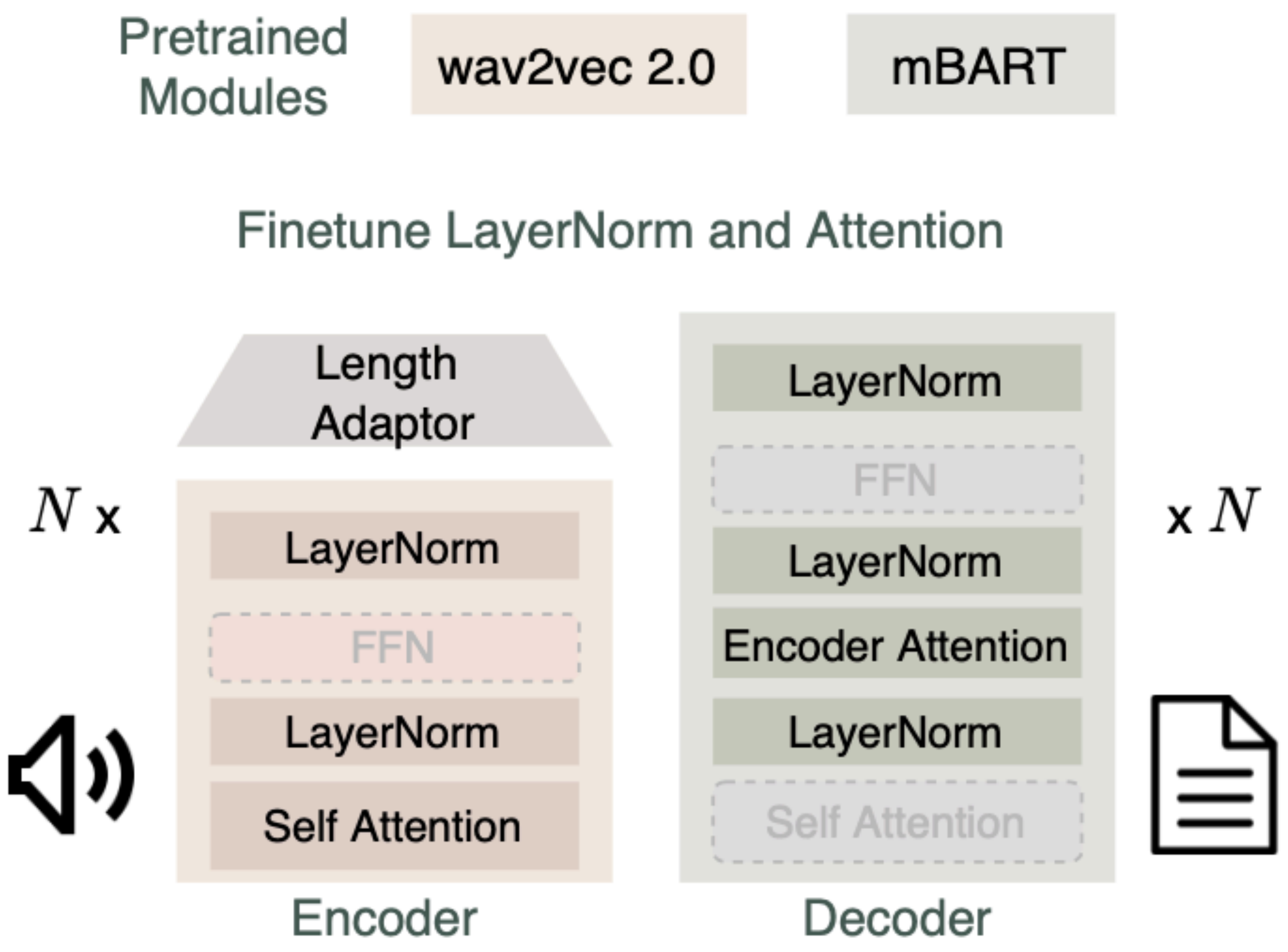


Architecture



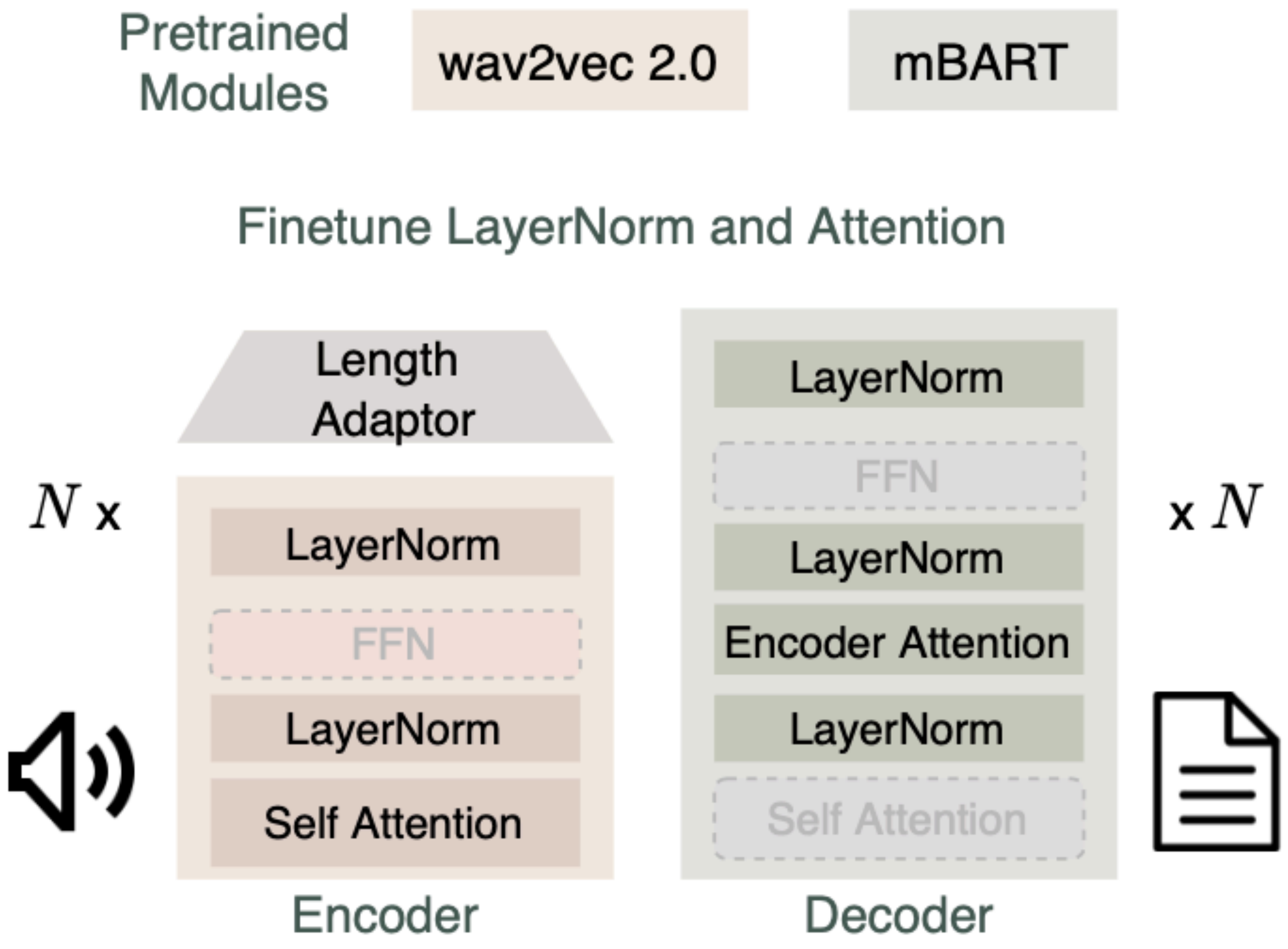
- wav2vec 2.0
- Learned to create high quality speech representations from unlabelled audio data
  - Feature encoder built from temporal convolution layers
  - Transformer based context encoder encoder
  - Trained on masked speech input and must solve a contrastive task

Architecture



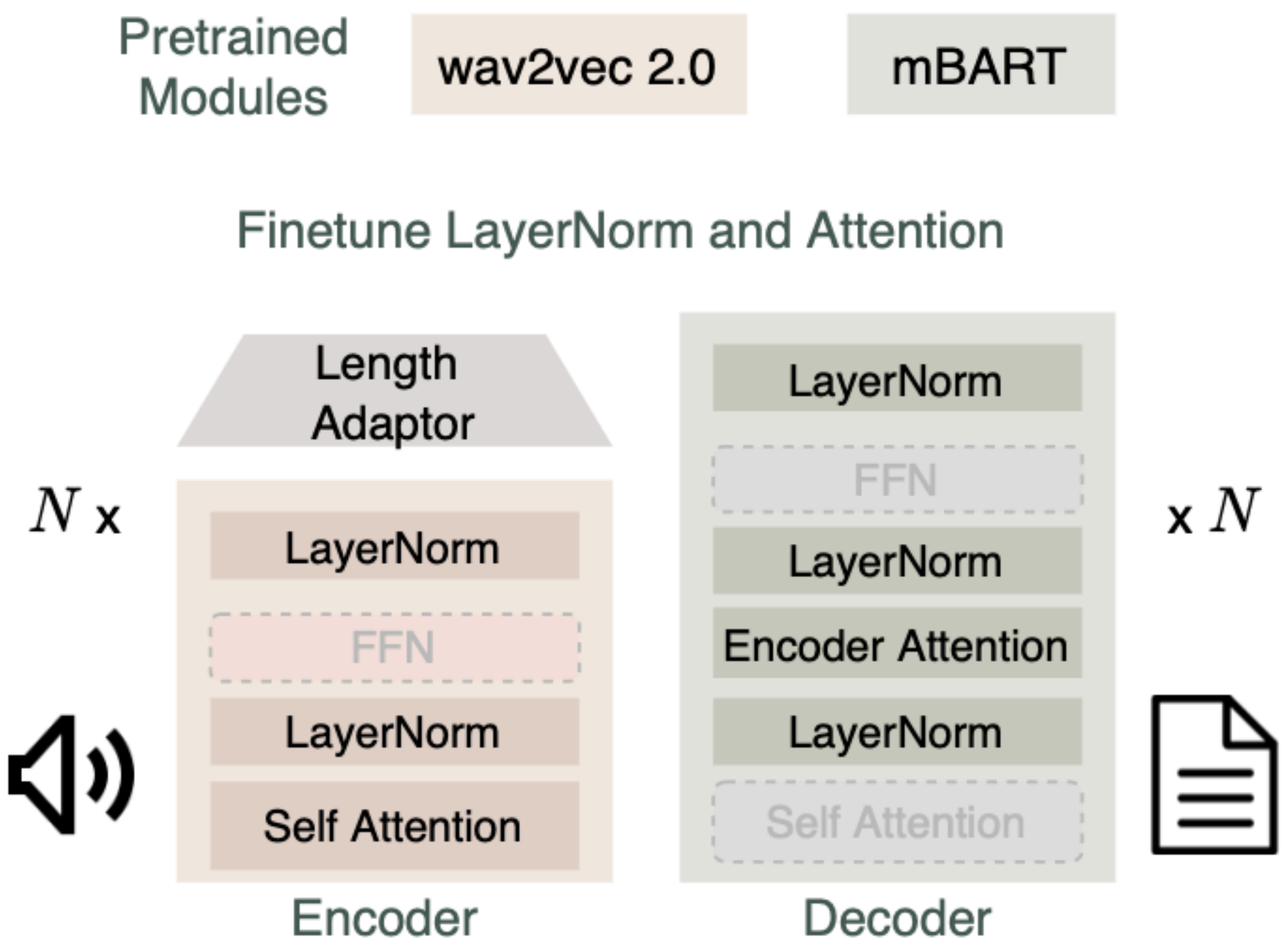
- mBART
  - Sequence-to-sequence generative pre-training scheme or a denoising autoencoder
  - For text x that is "noised" to be g(x), it reconstructs x for monolingual data for many languages (or across languages)
  - g does random span masking, order permutation

Architecture



- Length adaptor
  - Aligns two modules due to their different modalities
  - Projection and downsampling using 1d convolutional layers to shrink speech sequence (encoder output)

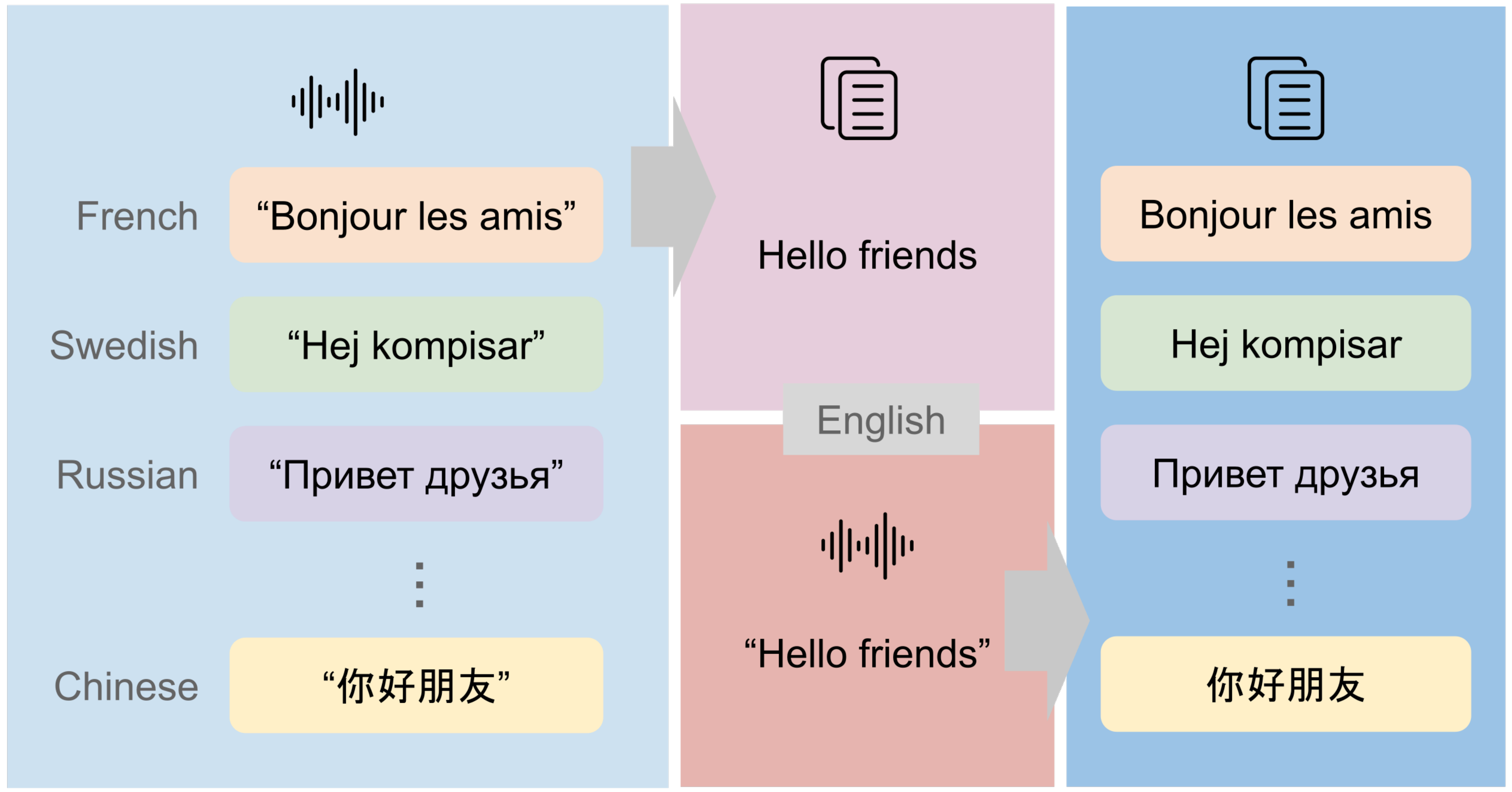
Architecture



- LNA finetuning
  - Fine tune only layer normalization and multi-head attention parameters
  - Attention was pre-trained on text-to-text so must be adapted to ST

Data

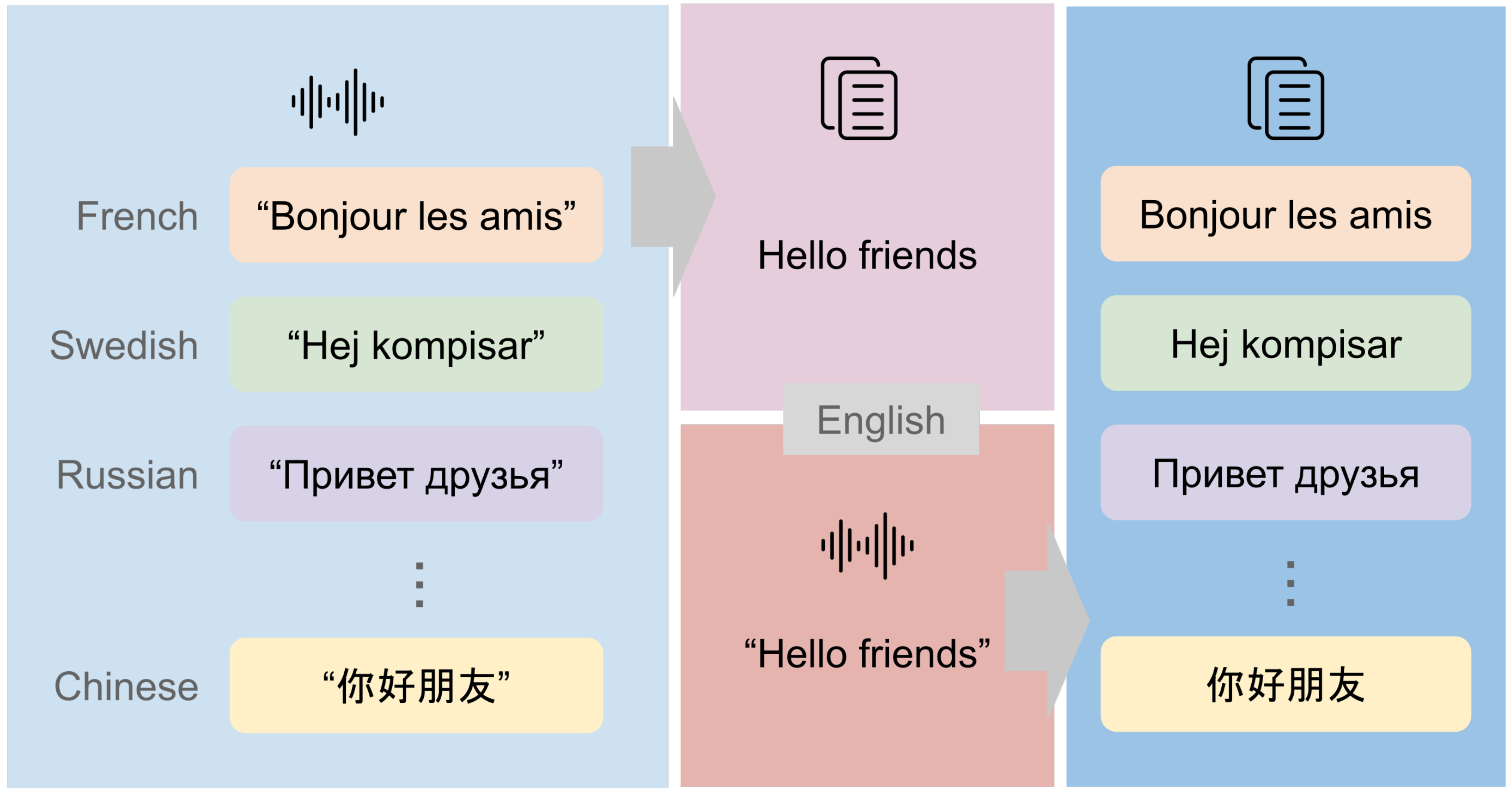
- Uses CoVoST 2.0
- Speech to text translations from English into 15 languages
- Speech to text translations from 21 languages into English
- e.g. Tamil, Chinese, Catalan, Arabic
- 2,880 hours of speech across 78K speakers



From https://github.com/facebookresearch/covost

Data

- CoVoST 2.0 is a good testbed for different resource levels
  - 4 X-En languages with 10-20hrs
  - 11 with <4hrs of data
- Uses EuroParl ST as it has non-English language pairs
  - de, en, es, fr, it, pt
  - Can assess zero-shot performance



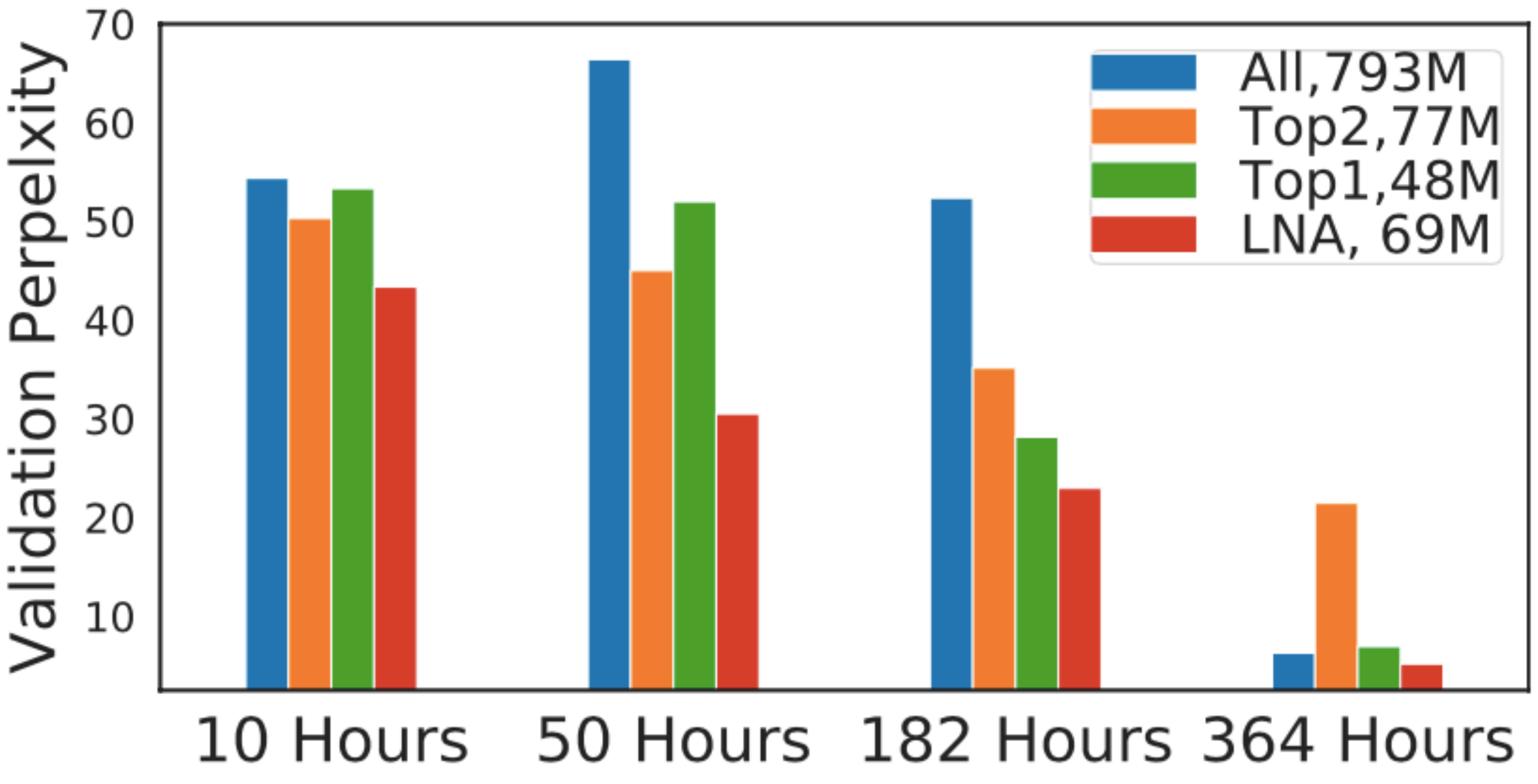
From https://github.com/facebookresearch/covost

Training

- Cross-Modal Efficient Finetuning
- Encoder initialized with word2vec 2.0 pretrained on unlabelled English
- mBART pretrained with monolingual data for 50 languages
  - Further trained with Bitext for 49 X-EN, 49 En-X languages
- LNA fine-tuning to encoder, decoder, both using X-En, En-X
- Joint training
  - Last 12 layers of wav2vec encoder replaced with 12 mBART encoder layers

Results

- Comparing fine-tuning all parameters and just LNA-Minimalist (69M)
- LNA-Minimalist generalizes better and improves training efficiency



Results

- Source-side results (speech)
  - Achieves SOTA on Portuguese

				Train					Zero-shot
	Enc	Dec	Params.	Fr	De	Es	Ca	It	Pt
LNA-E,D	LN+SA	LN+EA	170.7M	**32.4**	**24.9**	**31.6**	**28.6**	**24.0**	**8.2**
LNA-D	All	LN+EA	384.8M	31.6	23.7	31.0	27.8	23.2	7.6
Finetune All	All	All	793.0M	27.1	17.7	27.8	21.7	18.9	5.1
ASRPT+Multi				23.1	15.3	21.2	19.9	14.9	4.4
Supervised (Multi) SOTA (Wang et al., 2020b)				26.5	17.6	27.0	23.1	18.5	6.3

Results

- Target-side results (text)
  - 1.3 off of SOTA for Japanese

				Train				Zero-shot
	Enc	Dec	Params.	De	Fa	Tr	Zh	Ja
LNA-E,D	LN	LN+EA	69.4M	22.1	17.7	13.4	29.2	22.9
LNA-E,D	LN+SA	LN+EA	170.7M	23.8	19.2	14.2	30.6	29.2
LNA-D	All	LN+EA	384.8M	**24.9**	**19.8**	15.2	**32.7**	**30.6**
LNA-E	LN+SA	All	477.6M	22.0	18.1	14.2	29.5	0.8
Finetune All	All	All	793.0M	24.1	19.6	**15.6**	32.4	0.4
ASRPT+Multi				9.5	10.9	6.8	23.5	0.0
Supervised (Multi) SOTA (Wang et al., 2020b)				17.3	14.5	10.7	28.2	31.9

Results

- Multilingual results
  - Often achieves E2E SOTA in most languages
  - Surpasses cascaded SOTA in 8/10 languages
  - LNA-D is best performing

Results

	High Resource				Low Resource					
→ **En**	Fr	De	Es	Ca	It	Ru	Pt	Nl	Sl	Sv
Train Hours	264	184	113	136	44	18	10	7	2	2
Scratch-BL	24.3	8.4	12.0	14.4	0.2	1.2	0.5	0.3	0.3	0.2
+ ASR PT	26.3	17.1	23.0	18.8	11.3	14.8	6.1	3.0	3.0	2.7
+ Multi.	26.5	17.5	27.0	23.1	18.5	4.7	6.3	5.0	0.7	0.5
+mBART	28.1	19.7	28.1	24.0	19.9	2.7	6.2	8.1	0.5	1.4
LNA-E,D (170.7M)	**33.8***	**26.7***	**34.0***	**29.5***	**26.1***	**21.1**	**19.2**	**14.1***	**4.6**	**5.9**
LNA-D (384.8M)	**35.0***	**28.2***	**35.2***	**31.1***	**27.6***	**22.8**	**24.1***	**14.2***	**5.0**	**5.0**
Finetune All (793.0M)	**33.0***	**24.5***	**33.6***	**28.0***	**25.2***	**20.2**	**19.5**	**9.4**	**4.6**	**4.8**
Joint Training (1.05B)	**33.5***	**28.6***	**33.5***	**30.6***	**26.6***	**17.6**	**12.0**	**15.0***	**3.9**	2.6
+ Extra MT Data	**34.4***	**29.6***	**34.4***	**30.6***	**27.7***	**27.7***	**14.6**	**14.5***	**5.2**	**3.4**
Prev. E2E SOTA	27.0	18.9	28.0	24.0	11.3	14.8	6.1	8.4	3.0	2.7
Cascade SOTA	29.1	23.2	31.1	27.2	22.9	25.0	22.7	10.4	7.0	11.9

→ **En**	Fa	Zh	Tr	Et	Mn	Ar	Lv	Cy	Ta	Ja	Id	Avg.
Train Hours	49	10	4	3	3	2	2	2	2	1	1	
ASR (WER)	62.4	45.0	51.2	65.7	65.2	63.3	51.8	72.8	80.8	77.1	63.2	
Baseline	1.9	1.4	0.7	0.1	0.1	0.3	0.1	0.3	0.3	0.3	0.4	
+ ASR PT	3.7	5.8	3.6	0.1	0.2	4.3	2.5	2.7	0.3	1.5	2.5	
+ Multi.	2.4	5.9	2.3	0.6	0.1	0.4	0.6	1.9	0.1	0.1	0.3	7.0
+ mBART	3.3	5.4	2.4	0.7	0.2	0.5	0.6	1.4	0.1	0.2	0.2	7.3
LNA-E,D (170.7M)	**4.0**	**6.2**	**5.5**	**1.3**	**1.0**	3.7	**4.6**	2.8	**0.7**	**1.7**	**2.9**	12.5
LNA-D (384.8M)	3.6	**6.0**	**4.8**	**1.5**	**0.9**	2.8	**4.9**	2.3	**0.8**	**1.7**	**3.7**	12.6
Finetune All (793.0M)	3.7	**6.5**	**4.0**	**1.4**	**1.0**	3.3	**4.9**	2.1	**0.5**	**2.1**	**3.4**	11.2
Joint Training (1.05B)	**6.1***	5.4	3.3	0.7	0.2	0.8	**2.7**	1.0	0.1	0.3	0.5	10.7
+ Extra MT Data	**5.0**	**6.2**	**4.0**	0.8	0.3	1.0	**3.6**	1.1	0.2	0.5	0.5	11.7
Prev. SOTA	3.7	5.9	3.7	0.9	0.2	4.3	2.5	3.3	0.3	1.5	2.5	
Cascade	5.8	11.4	9.3	3.8	1.0	12.3	7.2	7.4	0.4	3.8	11.8	

Results

- For very high resource languages (18+hrs, 1m+ sentences), joint training improves singly trained model
- Performs well on zero-shot Europarl translation

		Target					
		De	En	Es	Fr	It	Pt
Source	De		12.8/**20.6**	10.2/**13.8**	11.6/**14.9**	6.6/**8.6**	10.4/**13.0**
	En	13.1/**22.5***		23.1/**32.3***	22.1/**30.0***	14.9/**21.5**	20.7/**28.4**
	Es	9.2/**12.1**	18.9/**26.0**		19.0/**21.8**	13.3/**15.4**	20.0/**21.9**
	Fr	9.8/**13.6**	19.8/**27.9***	18.6/**21.7**		13.8/**15.2**	19.7/**21.4**
	It	10.1/**11.9**	19.8/**25.6**	18.8/**20.8**	19.1/**20.0***		19.8/19.2
	Pt	9.0/**11.4**	19.0/**24.1**	19.8/19.6	18.1/**18.6**	15.6/**16.1**	

Takeaways

- Combined two large pre-trained single-modality models
- Fine-tuning can be very parameter efficient (10-20% fine-tuned)
- Achieved E2E SOTA and SOTA over cascaded models in some cases
- Can provide strong zero-shot results

Discussion Questions

- What advantages and disadvantages does this have compared to other E2E models?

Discussion Questions

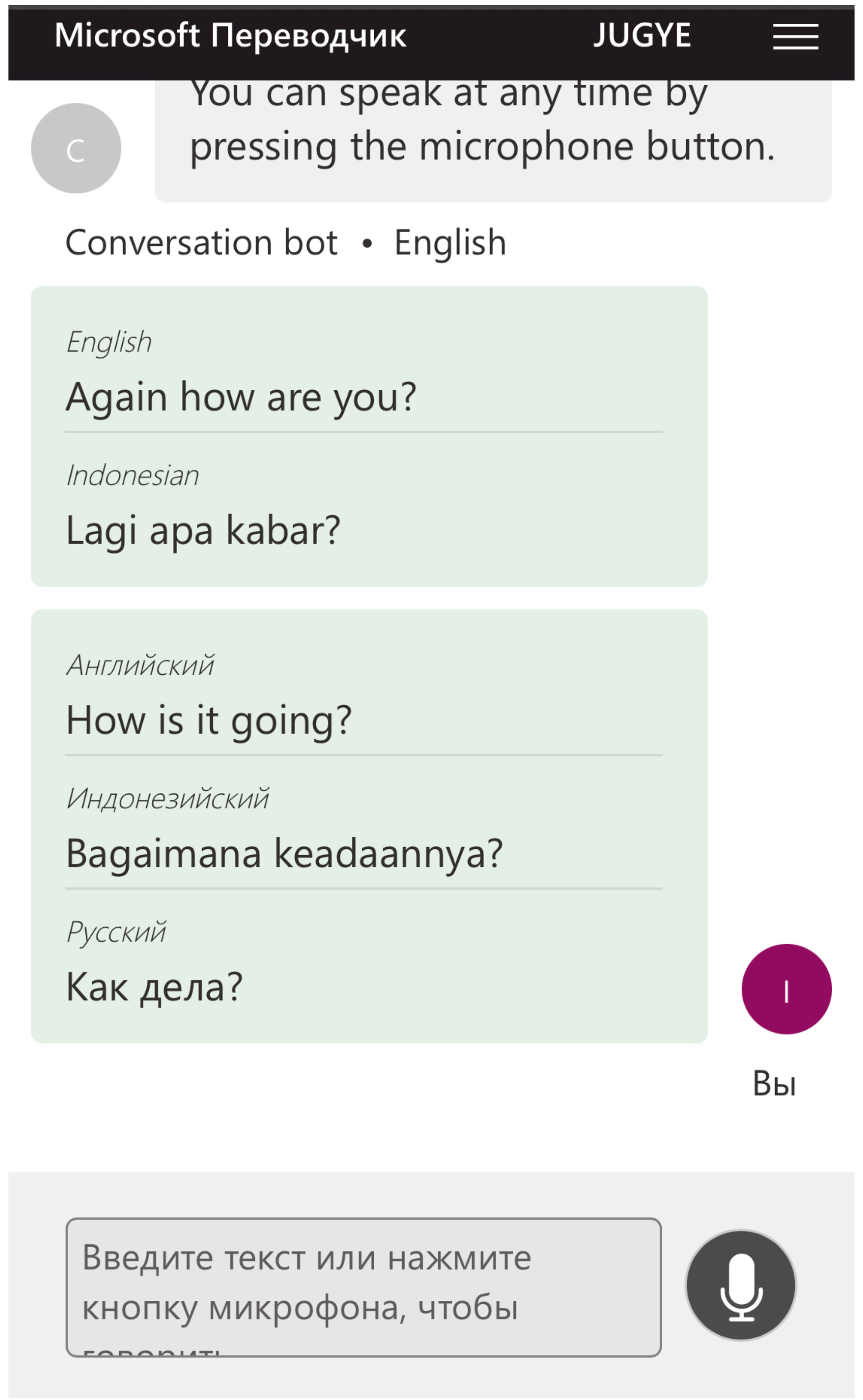
- What advantages and disadvantages does this have compared to other E2E models?
- Could this be adapted for speech-speech translation?

Conclusions

- Complicated, exciting research area
- Trade offs between data efficiency and modeling power
- To get around lack of data for E2E, we may undercut the benefits of E2E
- Future research
  - Low-resource models
  - Simultaneous models
  - Multilingual models
  - Extracting more from the audio

Resources

- https://st-tutorial.github.io/overview/
- https://pythonrepo.com/repo/dqqcasia-awesome-speech-translation-python-natural-language-processing
- https://iwslt.org/
- Many demos of cascaded ST: e.g. https://translator.microsoft.com/chatroom/



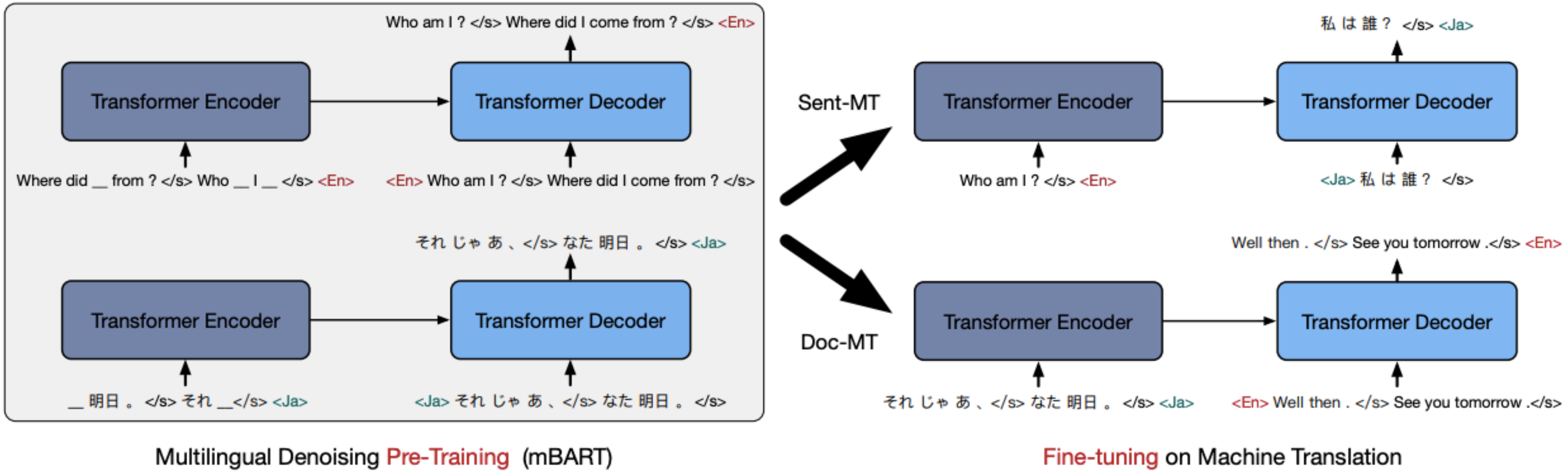
Citations

- Assets:
- Sound waveform by Oleksandr Panasovskyi from the Noun Project
- Diagrams inspired by and/or remade from Jan Niehues: Spoken Language Translation; Interspeech 2019
- All non-explicitly cited screenshots from the respective paper the section is about

- Invaluable resources to making this presentation:
- Introduction following content in Jan Niehues: Spoken Language Translation; Interspeech 2019
- Other content from ST Tutorial: https://st-tutorial.github.io/materials/
- Paper content from https://arxiv.org/abs/1904.06037, https://arxiv.org/abs/2107.08661, https://aclanthology.org/2021.acl-long.68.pdf

mBART

- Diagram from https://arxiv.org/pdf/2001.08210.pdf



Paper 2:
Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation

Renjie Zheng, Junkun Chen, Mingbo Ma, Liang Huang (Baidu Research; 2021)

https://arxiv.org/abs/2102.05766