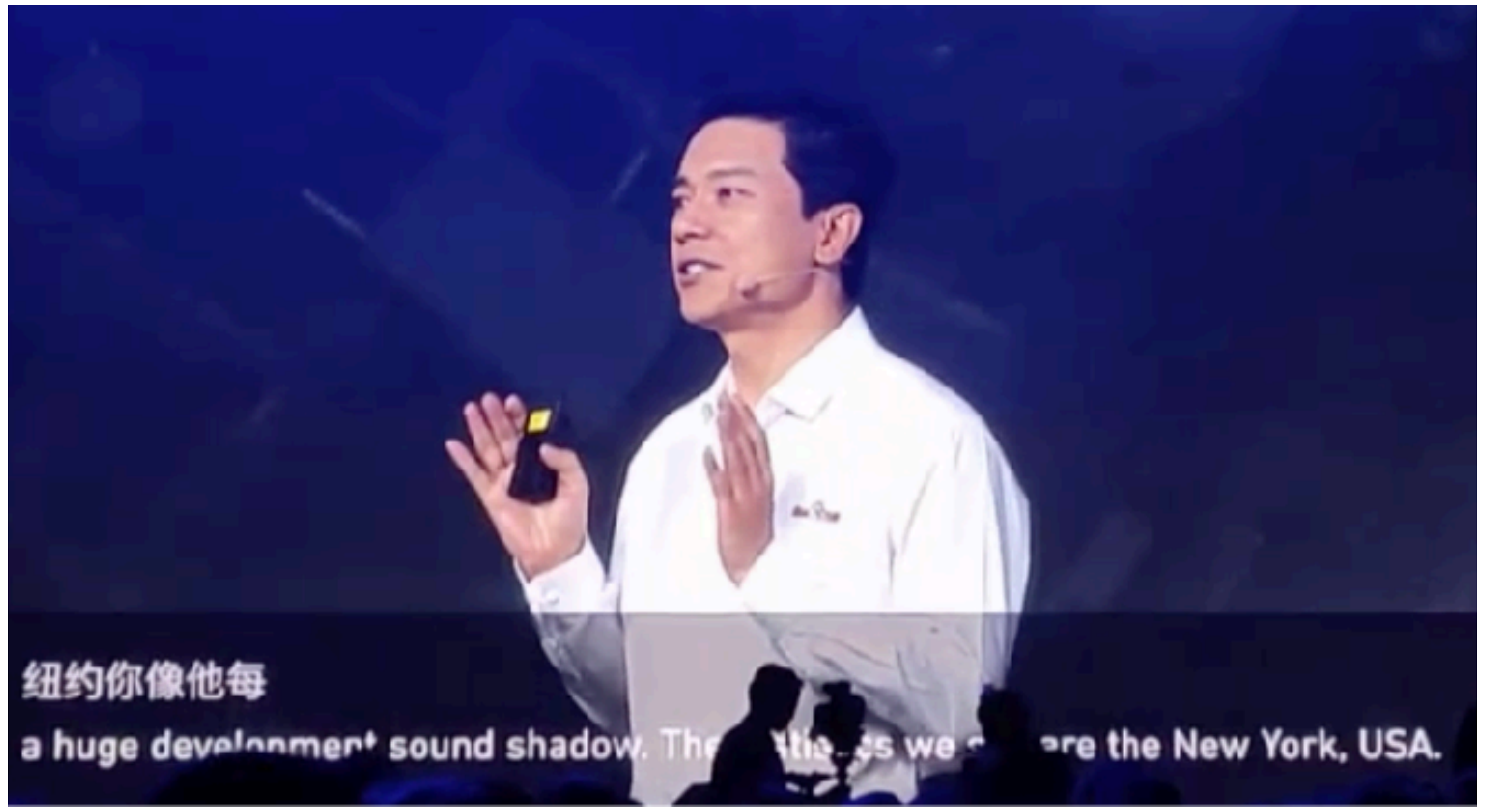
CVPR VIRTUAL JUNE 19-25

Simultaneous Translation: Breakthrough and Recent Progress



Liang Huang

Baidu Research USA and Oregon State University

includes joint work with Mingbo Ma, Renjie Zheng, Baigong Zheng, Junkun Chen, Kaibo Liu, Zhongjun He, et al.

Baidu Research

CVPR 2021 Invited Talk, Virtual Conference, June 2021

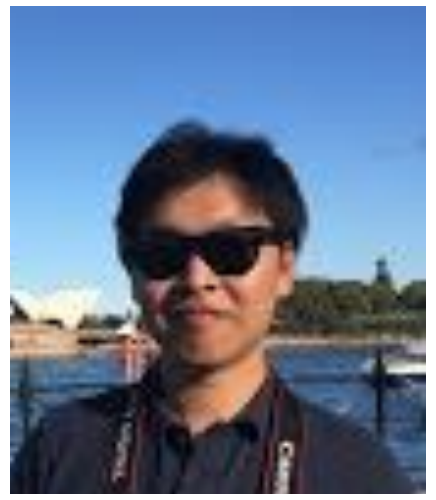
Oregon State University

Co-authors and Collaborators


Mingbo Ma


Renjie Zheng


Junkun Chen


Kaibo Liu


Baigong Zheng*


Ken Church


Jiahong Yuan

Baidu Research & Oregon State University

Baidu Research


Zhongjun He


Hao Xiong*


Chuanqiang Zhang


Ruiqing Zhang


Hua Wu


Haifeng Wang

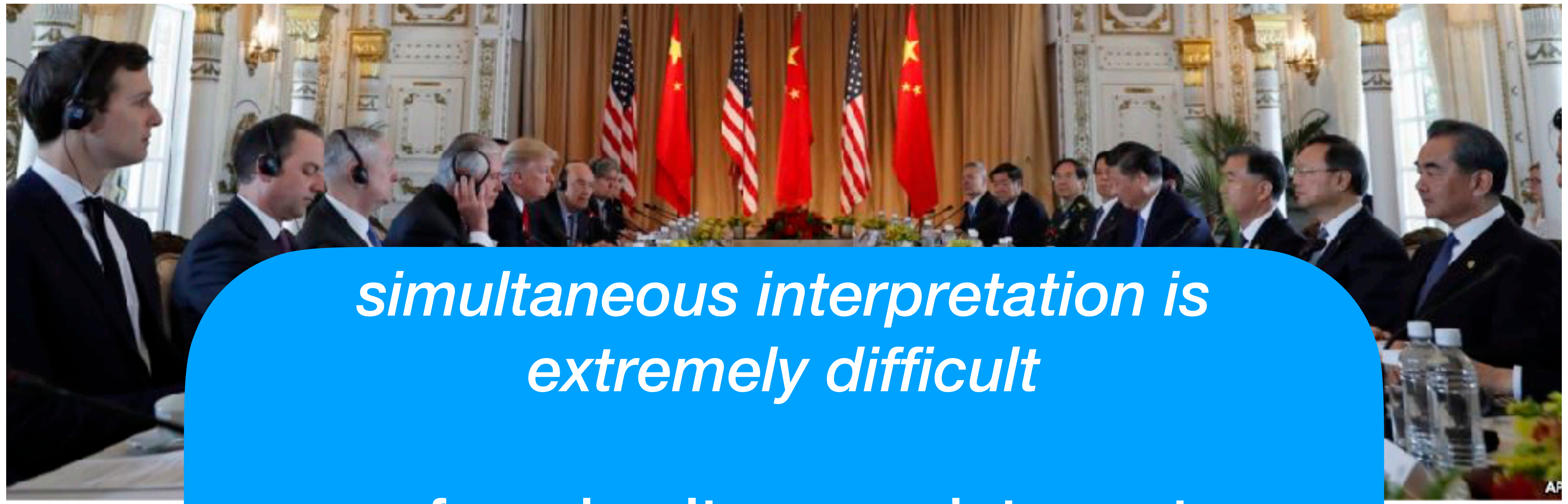
Baidu 百度

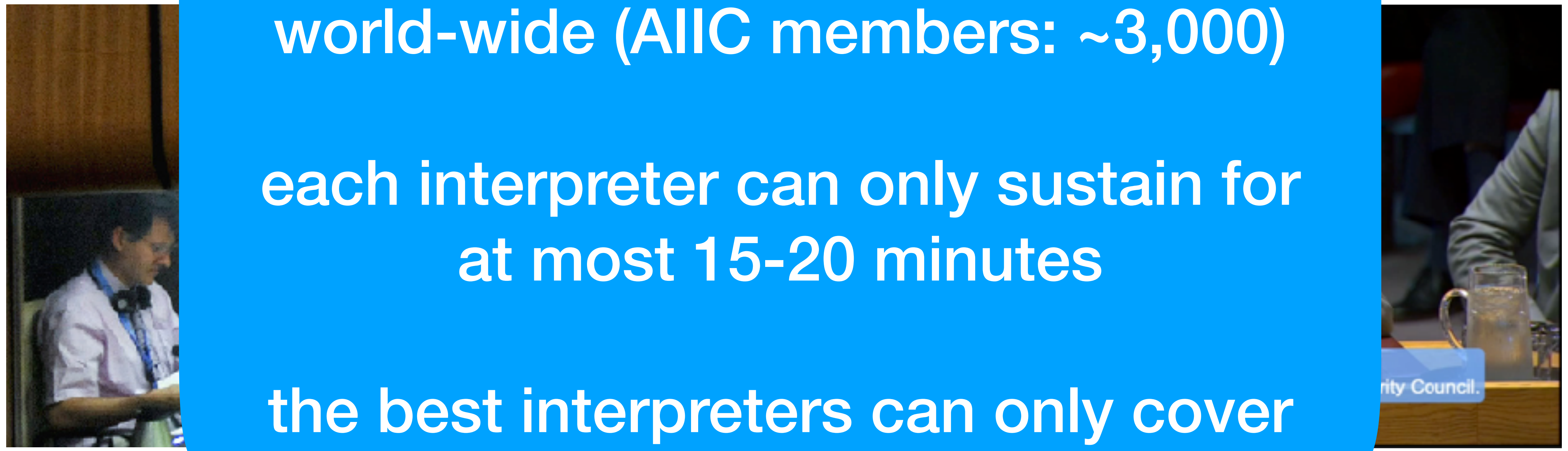
*former members

Consecutive vs. Simultaneous Interpretation

consecutive interpretation
multiplicative latency (x2)

simultaneous interpretation
additive latency (+3 secs)





simultaneous interpretation is extremely difficult

very few simultaneous interpreters world-wide (AIIC members: ~3,000)

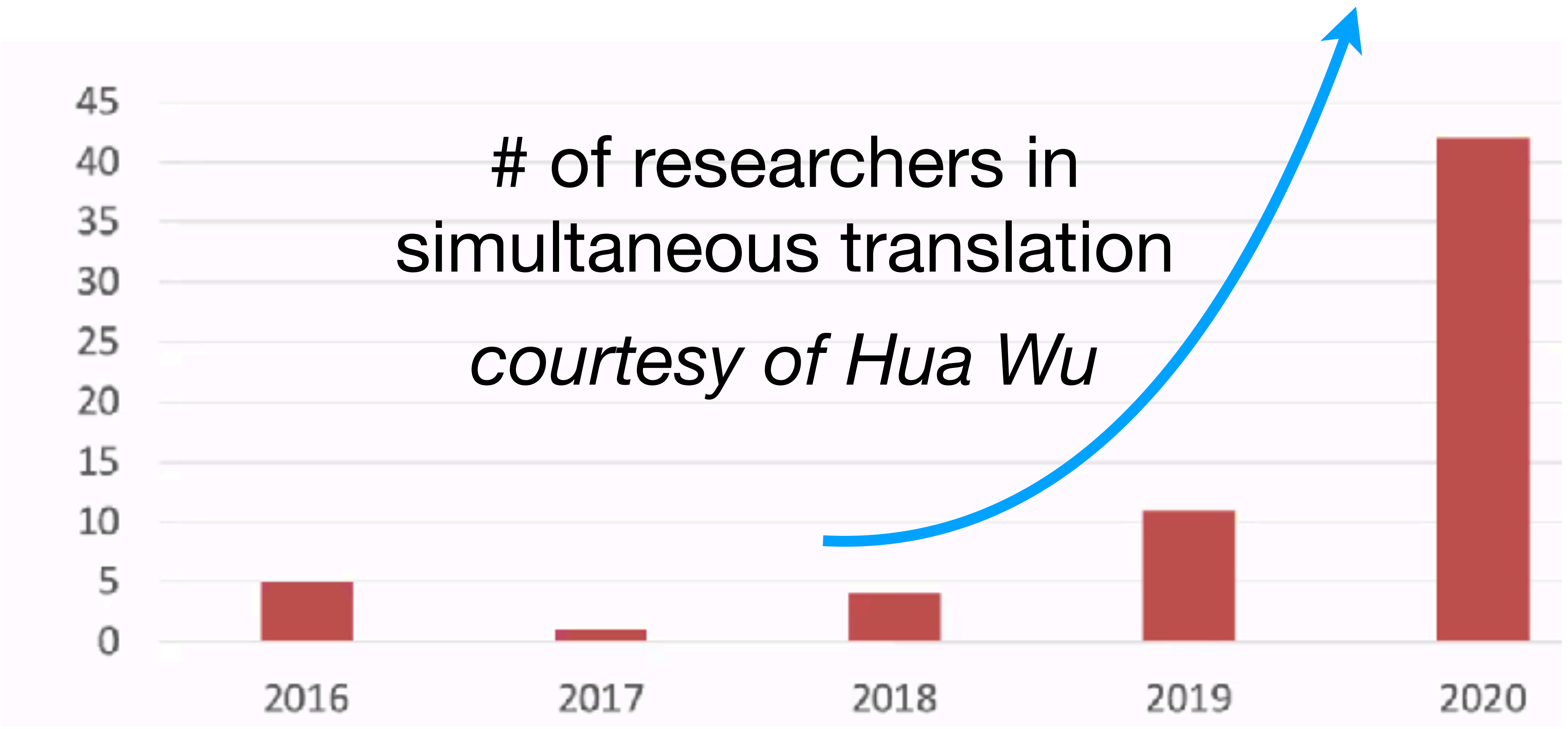
each interpreter can only sustain for at most 15-20 minutes

the best interpreters can only cover ~60% of the source material

Tradeoff between Latency and Quality

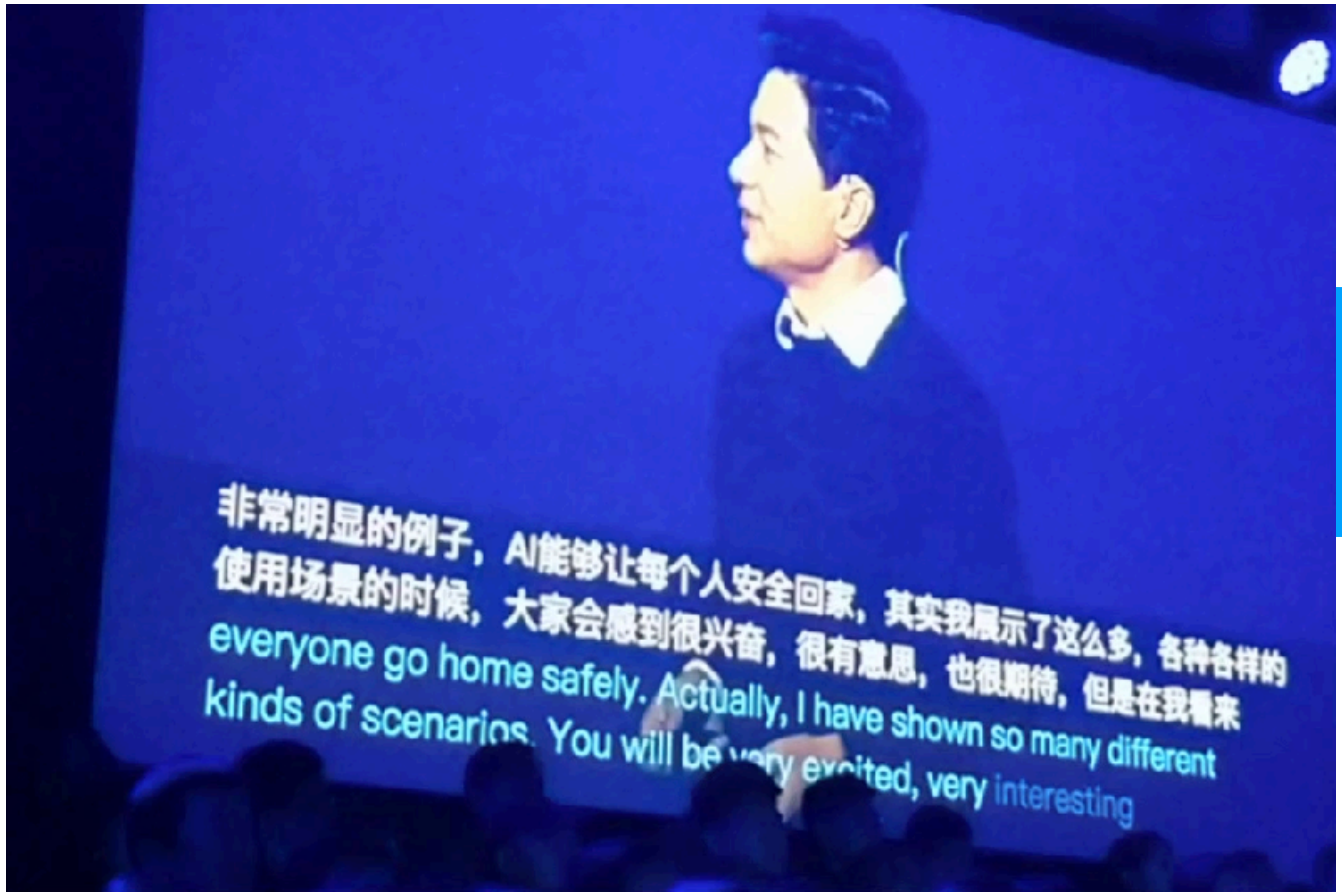
Outline

- Background on Simultaneous Interpretation
- Part I: Text-to-Text Simultaneous Translation
  - Our Breakthrough in 2018: Prefix-to-Prefix Framework
  - Flexible (Adaptive) Translation Policies
- Part II: Towards Speech-to-Speech Simultaneous Translation
  - (Pipelined) Speech-to-Speech Simultaneous Translation
  - Direct Simultaneous Speech-to-Text Translation
- Part III: Multimodal Models
  - Multimodal Speech/Text Pretraining
  - Multimodal Vision/Text Simultaneous Translation



Our Breakthrough in 2018

Baidu World Conference, Nov. 2017
full-sentence translation (latency: 10+ secs)



our work

Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

Media coverage:
IEEE Spectrum, MIT Technology Review, CNBC, VentureBeat, siliconANGLE, Synced (AI Technology & Industry Review), South China Morning Post, engadget, FORTUNE, Programmer, The Register (Biting the hand that feeds IT), lowyat.net (malaysia's largest online community), Packt, Red Pulse, Flipboard, China Knowledge

Main Challenge: Word Order Difference

- e.g. translate from Subj-Obj-Verb (Japanese, German) to Subj-Verb-Obj (English)
  - German is underlyingly SOV, and Chinese is a mix of SVO and SOV
  - human simultaneous interpreters routinely "anticipate" (e.g., predicting German verb)

ich	bin	mit	dem	Zug	nach	Ulm	**gefahren**	
I	am	with	the	train	to	Ulm	**traveled**	
I	(...... *waiting* ......)						**traveled**	by train to Ulm

Grissom et al, 2014

Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*	*yǔ*	*Éluósī*	*zǒngtǒng*	*Pǔjīng*	*huìwù*
布什	总统	在	莫斯科	与	俄罗斯	总统	普京	会晤
Bush	*President*	*in*	*Moscow*	*with*	*Russian*	*President*	*Putin*	*meet*

President Bush meets with Russian President Putin in Moscow

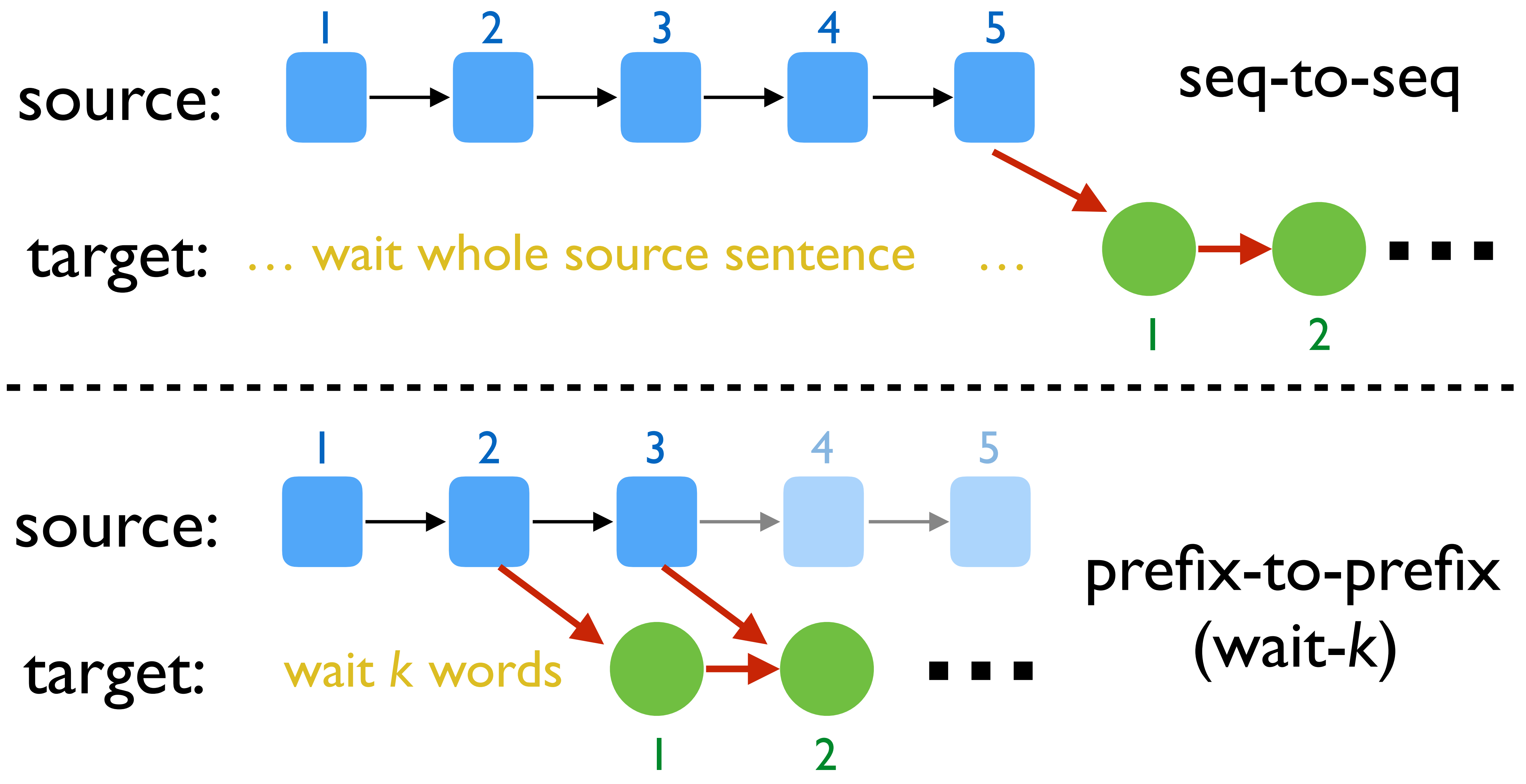
non-anticipative: President Bush (...... *waiting* ......) meets with Russian ...

anticipative: President Bush meets with Russian President Putin in Moscow

Our Idea (2018): Prefix-to-Prefix & Wait-*k*

- standard seq-to-seq is only suitable for conventional full-sentence MT
- we proposed prefix-to-prefix framework tailed to tasks with simultaneity
  - special case: wait-*k* policy: translation is always *k* words behind source sentence
  - decoding this way => controllable latency
  - training this way => implicit anticipation *on the target-side*

$p(y_i \mid x_1 \ldots x_n, y_1 \ldots y_{i-1})$



$p(y_i \mid x_1 \ldots x_{i+k-1}, y_1 \ldots y_{i-1})$

Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*	*yǔ*	*Éluósī*	*zǒngtǒng*	*Pǔjīng*	*huìwù*
布什	总统	在	莫斯科	与	俄罗斯	总统	普京	会晤
Bush	*President*	*in*	*Moscow*	*with*	*Russian*	*President*	*Putin*	*meet*

wait 2 President Bush meets with Russian President Putin in Moscow

Mingbo Ma

Research Demo

江泽民对法国总统的来华

jiang zemin expressed his appreciation

jiāng zémín	duì	fǎ guó	zǒng tǒng	d e	l á i huá	fǎng wèn	biǎo shì	gǎn xiè	
江 泽民	对	法国	总统	的	来华	访问	表示	感谢	。
jiang zemin	to	French	President	’s	to-China	visit	express	gratitude	

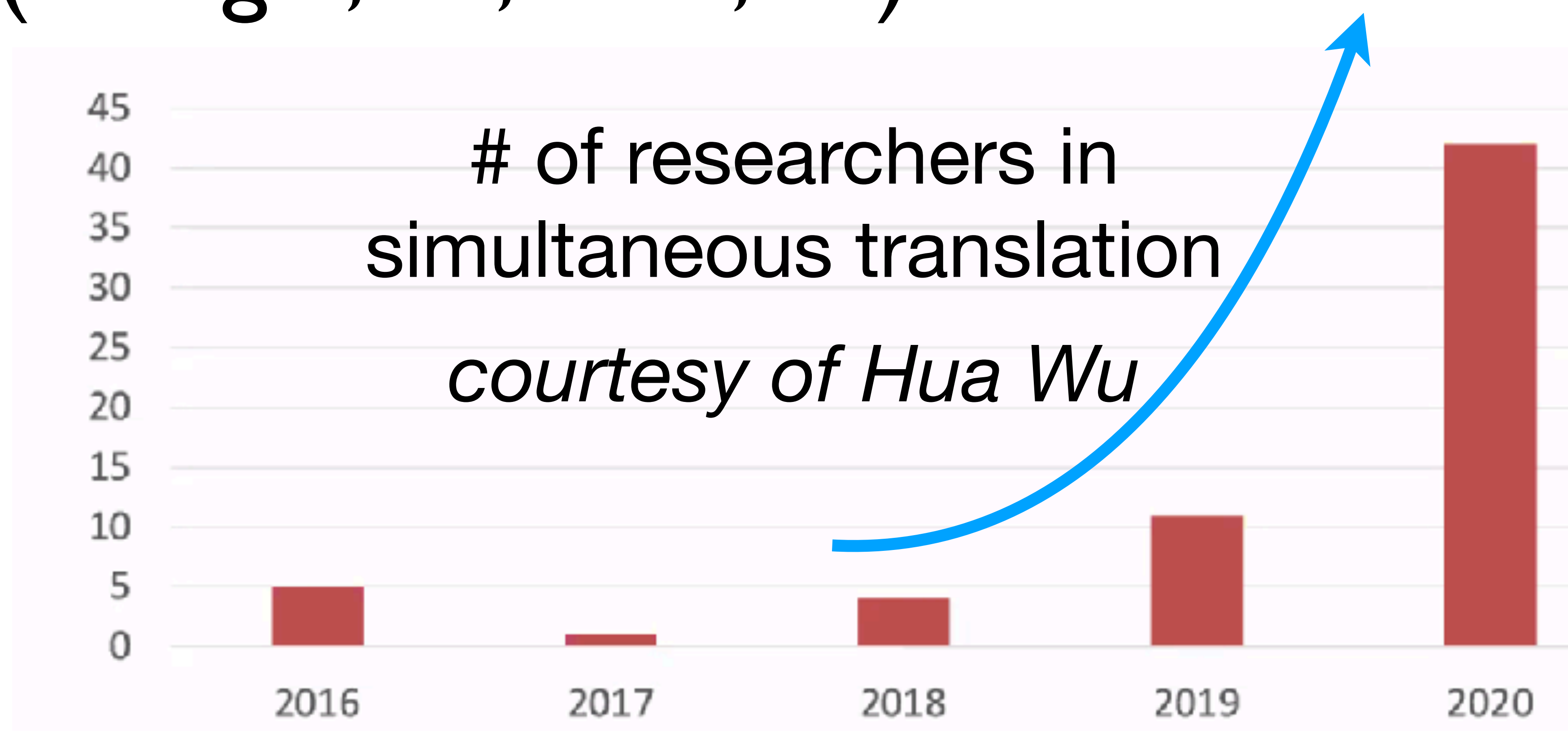
jiang zemin expressed his appreciation for the visit by french president .

Bai du Research This is just our research demo. Our production system is better (shorter ASR latency).

Summary and Roadmap

- "prefix-to-prefix" is the first framework tailed to simultaneity (incremental on both sides)
  - first *genuinely* simultaneous translation model (rather than full-sentence model)
  - very easy to train; scalable and replicable; quickly became the standard approach, replacing RL
  - prefix-to-prefix is very general; can be used in other tasks with simultaneity

- simultaneous translation: "out of reach" =2018=> "commercializable"
  - ACL 2019 Keynote; ACL 2020: 1st AutoSimTrans workshop; IWSLT 2020: shared task
  - simultaneous translation is now a hot problem, esp. in industry (Google, FB, MSR, …)
- since 2018: many active research areas
  - adaptive translation policies
  - simultaneous speech-to-text and speech-to-speech translation



Part 1(b): Towards Adaptive Translation Policies

Baigong Zheng

(B. Zheng, et al., ACL 2020)

(B. Zheng, R. Zheng, et al., ACL 2019)

(B. Zheng, R. Zheng, et al., EMNLP 2019)

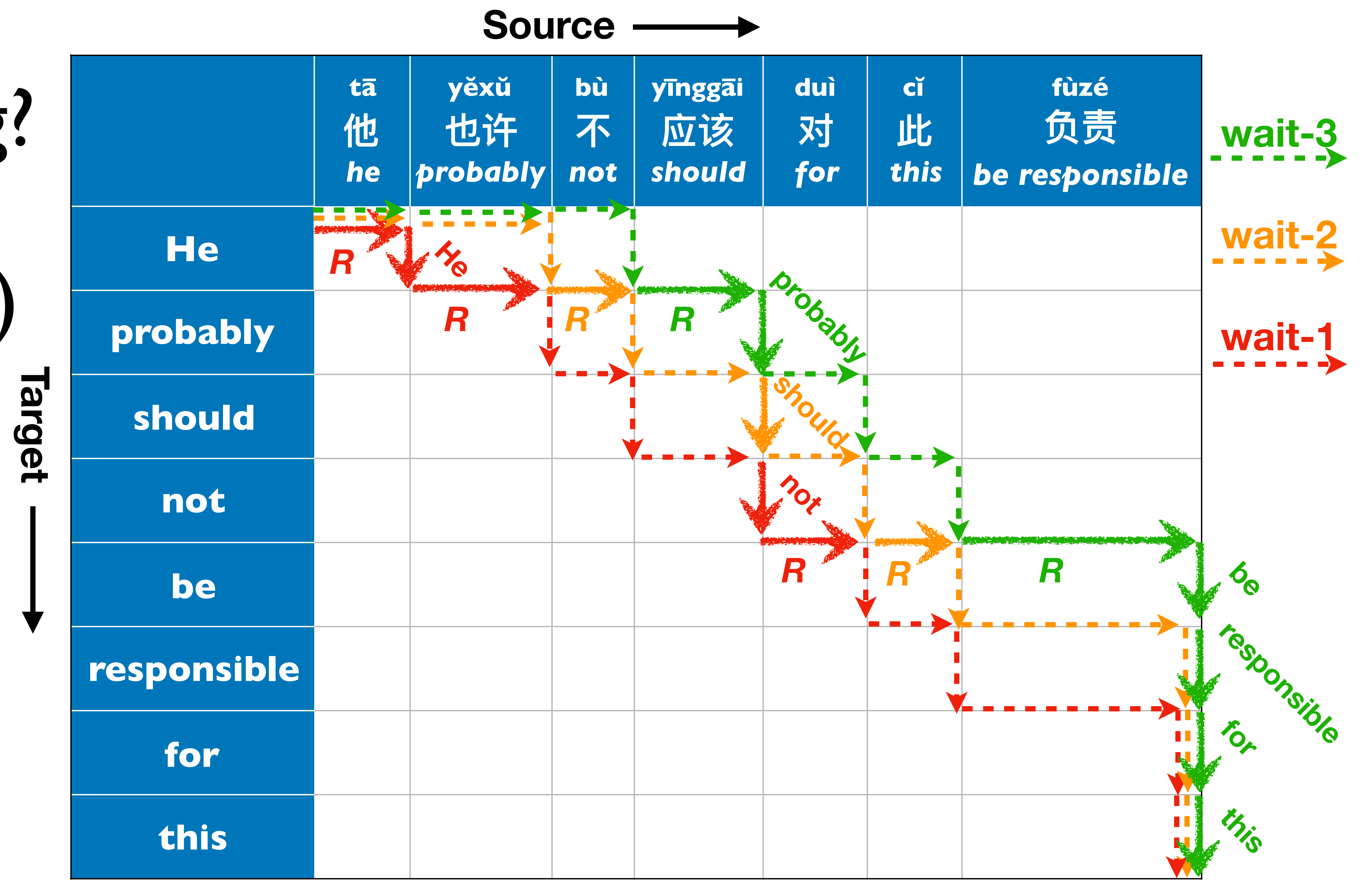
Renjie Zheng

Latency-Accuracy Tradeoff of Wait-*k* Policies

- smaller *k*: faster (lower latency) but could be too aggressive (lower quality)
- larger *k*: slower (higher latency) but more conservative (higher quality)
- Q: what's the optimal *k*? What about adaptively change this *k*?

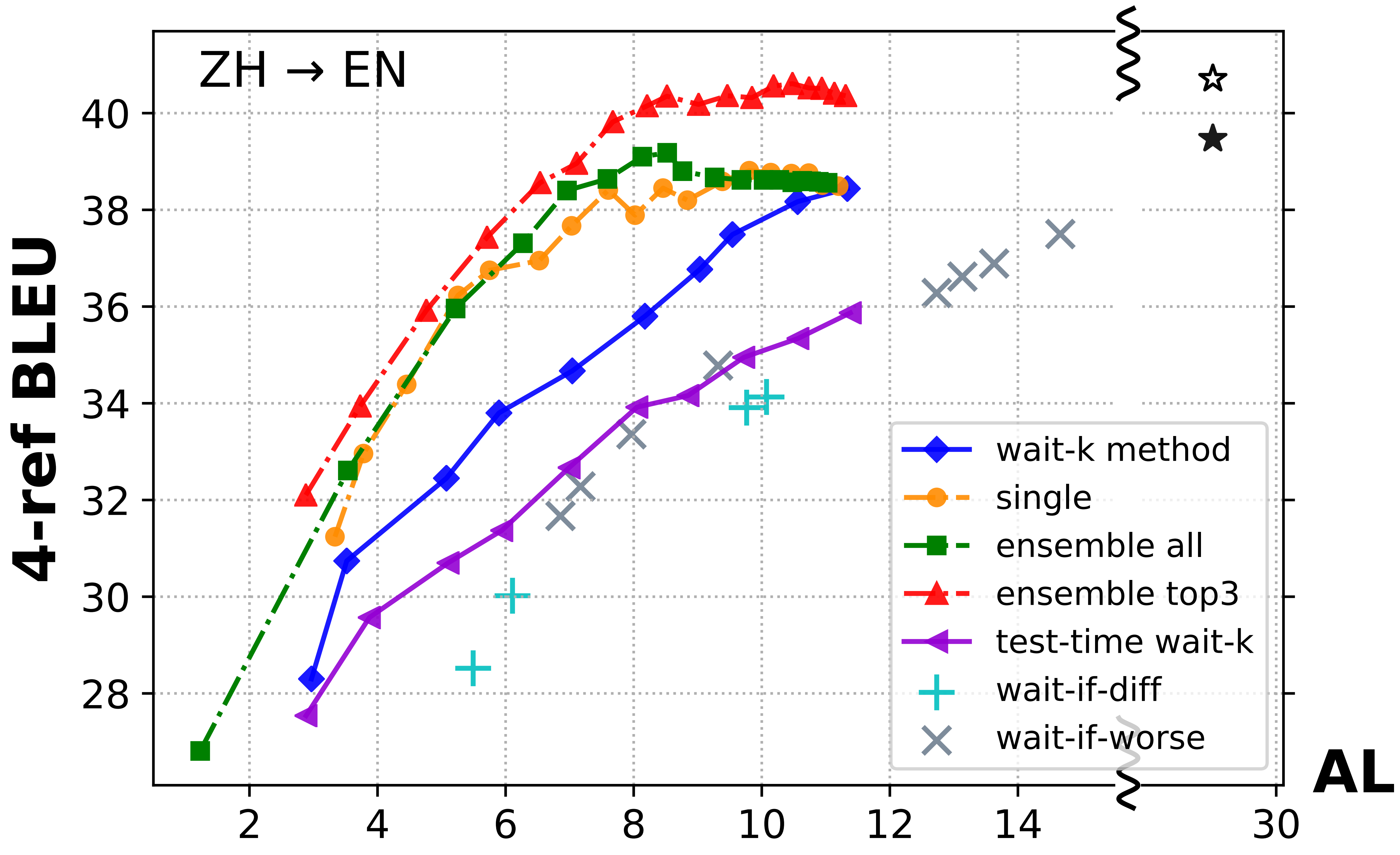
Idea 1 (ACL 2020): wait-*k* with adaptive *k*

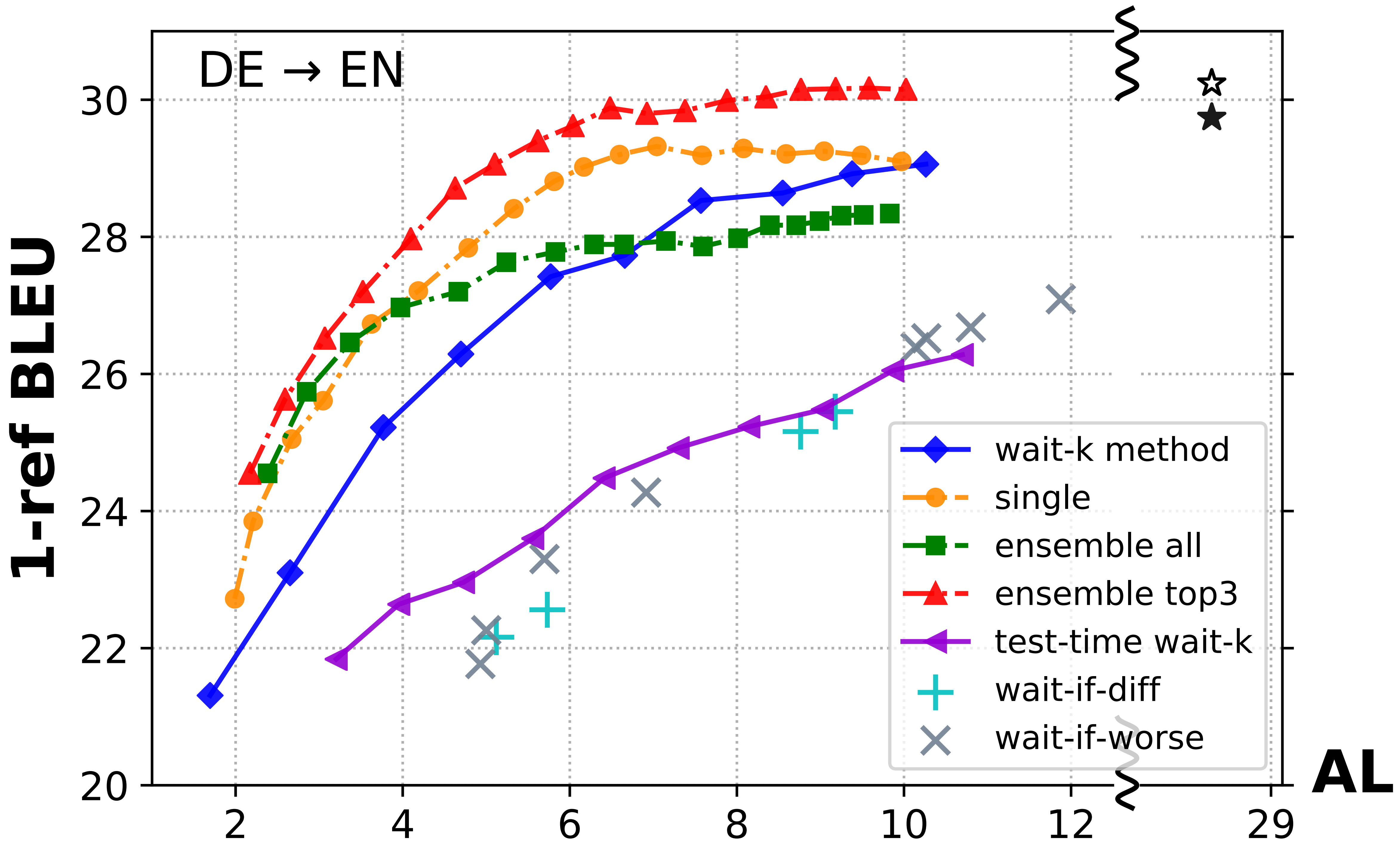
- wait-*k* policies are simple and effective
- can we change *k* dynamically in decoding?
- READ (wait) or WRITE (commit output) based on model confidence
  - prob. of the top-1 word > threshold?
  - if confident enough, WRITE (*k*--)
  - not confident enough, READ (*k*++)



(B. Zheng et al., ACL 2020)

Experiments

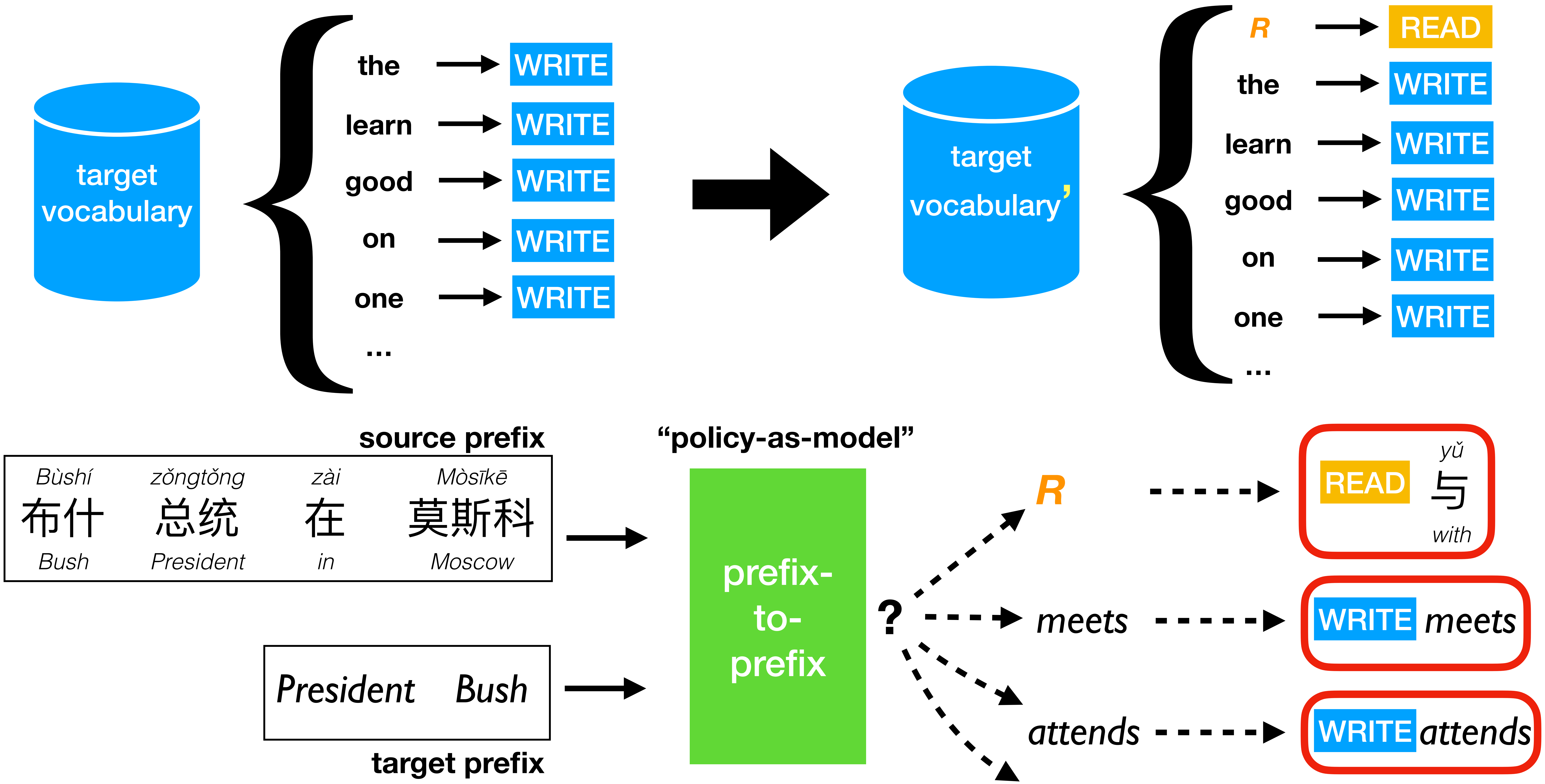




pinyin	*wǒmén xiàng shòuhàizhě de jiāshǔ biǎoshì zuì chéngzhì de tóngqíng hé āi dào*
input	“ 我们 向 受害者 的 家属 表示 最 诚挚 的 同情 和 哀- 悼 . ”
gloss	we to victim 's family express most sincere 's sympathy and condolence
wait-3 (AL=3.72)	“ we have offered our best wishes to the families of the victims , ” he said .
ensemble top-3 $\rho_1 = 0.4, \rho_{10} = 0$ (AL=2.8)	“ we express the most sincere sympathy to the families of the victims . ”

wrong anticipation

Idea 2 (ACL 2019): Policy as Model ("READ" as a word)



(B. Zheng, R. Zheng, et al., ACL 2019)

Summary on Fixed and Adaptive Policies

- most previous work uses RL, but is found to be not replicable
- we introduced four simple and effective approaches

	fixed-latency policies	*adaptive policies*
full-sentence MT model	Dalvi et al. (2018); **test-time wait-*k* (Ma et al. 2018)**	Grissom et al. (2014); Cho & Esipova (2016); Satija & Pineau (2016); Gu et al. (2017); Alinejad et al (2018); …
simultaneous MT model	**wait-*k* (Ma et al. 2018)**	Arivazhagan et al. (ACL 2019) **idea 2: B. Zheng et al. (ACL 2019)** **idea 1: B. Zheng et al. (ACL 2020)**

Part II: Towards Simultaneous Speech-to-Speech Translation

(a) Pipelined Approach


Mingbo Ma


Baigong Zheng


Renjie Zheng

(M. Ma, B. Zheng, et al., EMNLP 2020 Findings)

(R. Zheng, M. Ma, et al., EMNLP 2020 Findings)

(b) Direct Approach


Junkun Chen


Renjie Zheng


Mingbo Ma

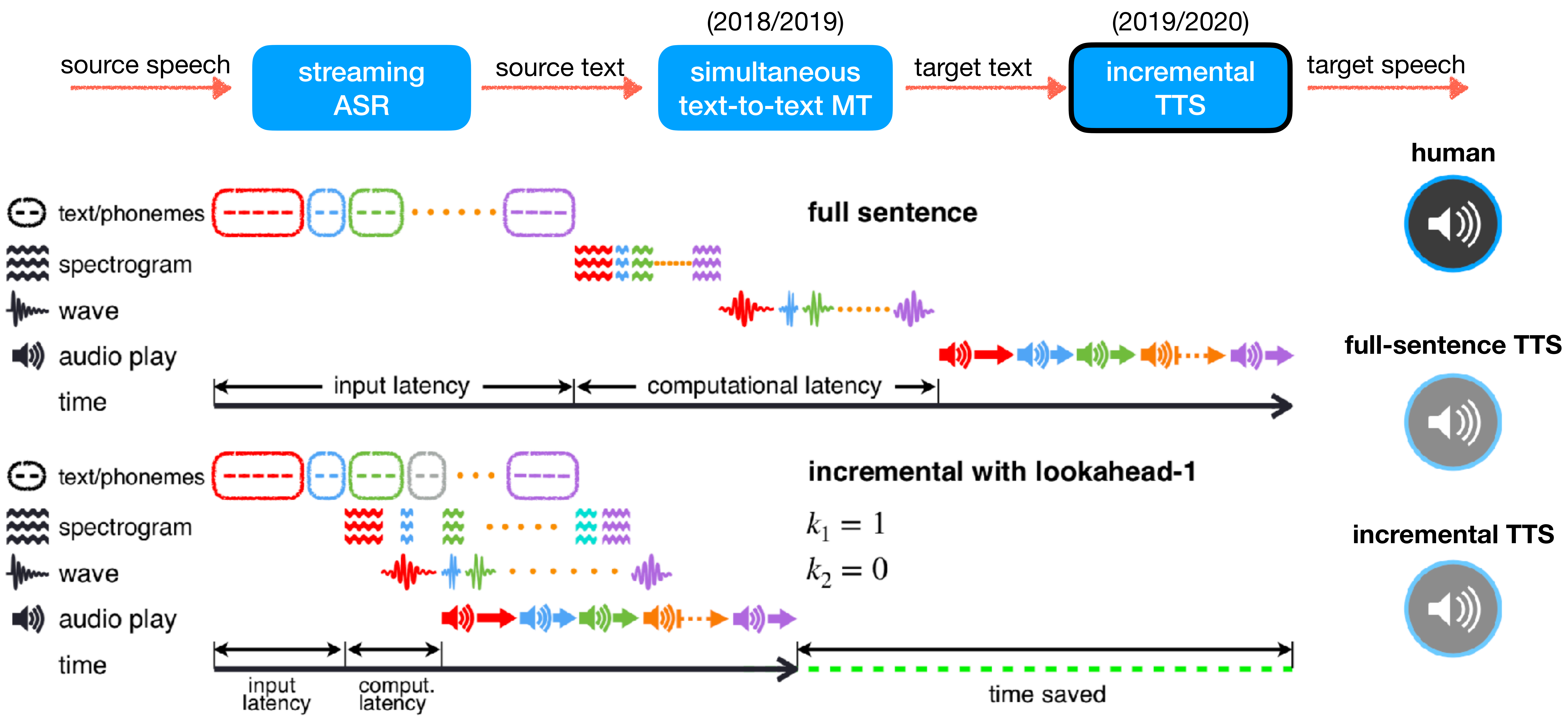
(J. Chen, M. Ma, et al., ACL 2021 Findings)

(R. Zheng, J. Chen, et al., ICML 2021)

Part II(a): Speech-to-Speech Simul. Trans. Pipeline

- text-to-text simultaneous MT is a toy problem; should be speech-to-speech
- all three modules (ASR, MT, TTS) need to be incremental/simultaneous
  - streaming ASR is widely available as APIs
  - we just made simultaneous MT possible
  - need incremental (streaming) TTS
- major challenge in making the whole pipeline work: latency (cf. Will's talk)
  - latency will accumulate across sentence boundaries (lagging more & more)
  - need to automatically "summarize" when falling behind

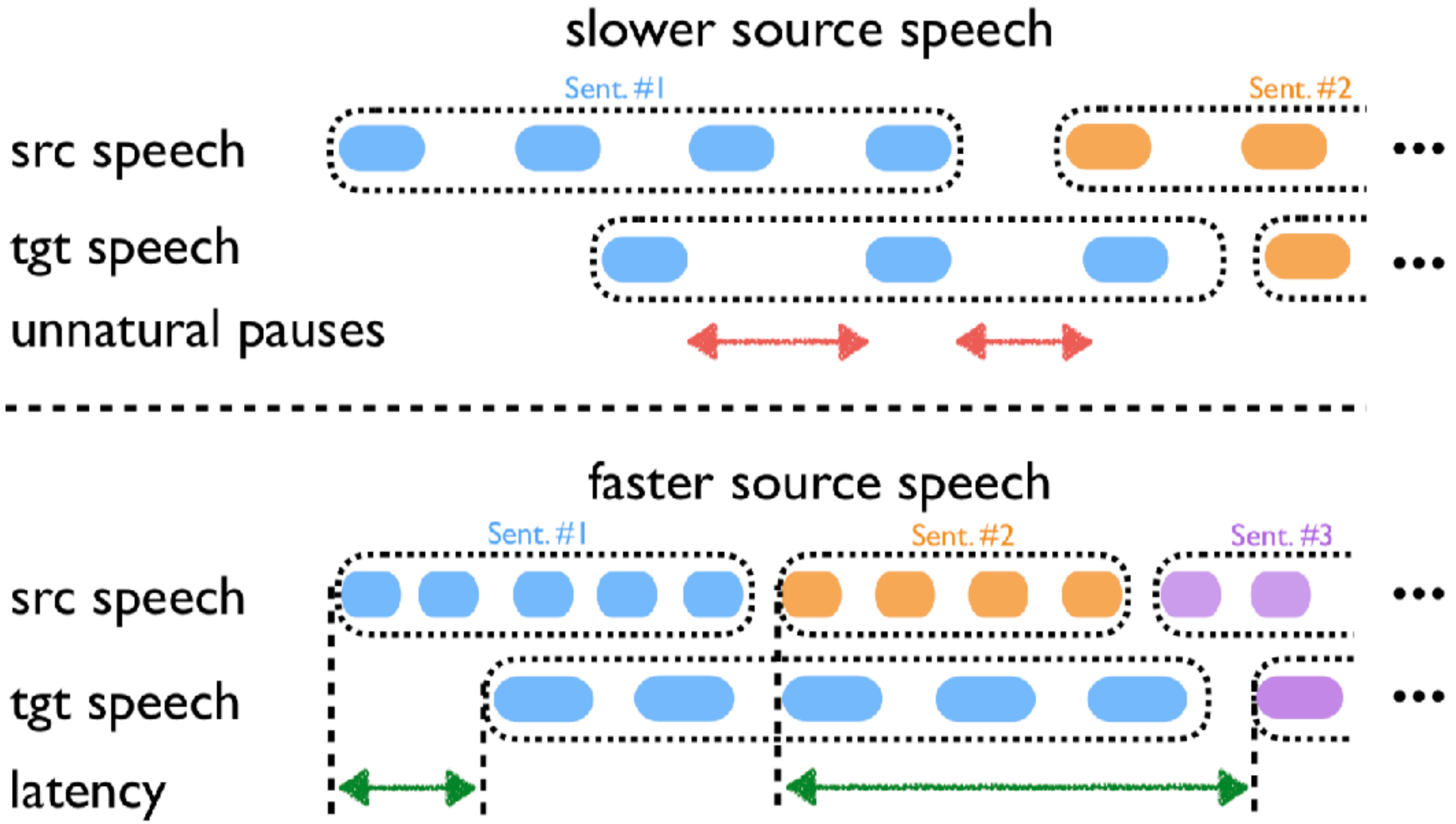
Incremental Text-to-Speech (TTS)



(Ma et al., EMNLP 2020 Findings)

Challenges in Simultaneous Speech-to-Speech

- fixed wait-k is problematic in both slow and fast speeches
  - slow speech: introduce unnatural pauses
  - fast speech: accumulating latencies across sentences, lagging more & more behind



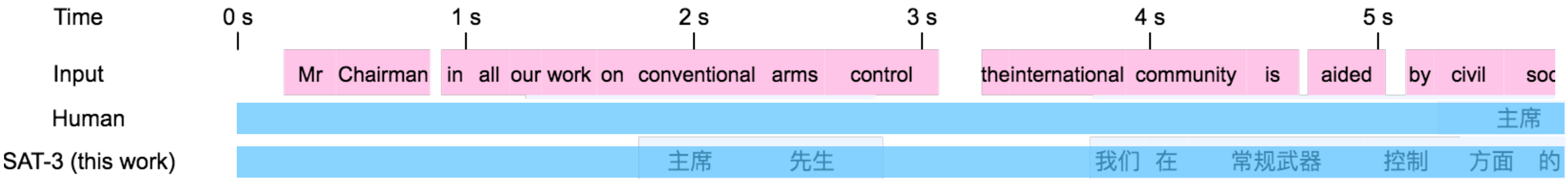
adjusting TTS speech rate is not a good idea!

Speech Rate	MOS
$0.5\times$	2.00 ± 0.08
$0.6\times$	2.32 ± 0.08
$0.75\times$	2.95 ± 0.07
Original	4.01 ± 0.08
$1.33\times$	3.34 ± 0.08
$1.66\times$	2.40 ± 0.09
$2.0\times$	2.06 ± 0.04

(R. Zheng et al., EMNLP 2020 Findings)

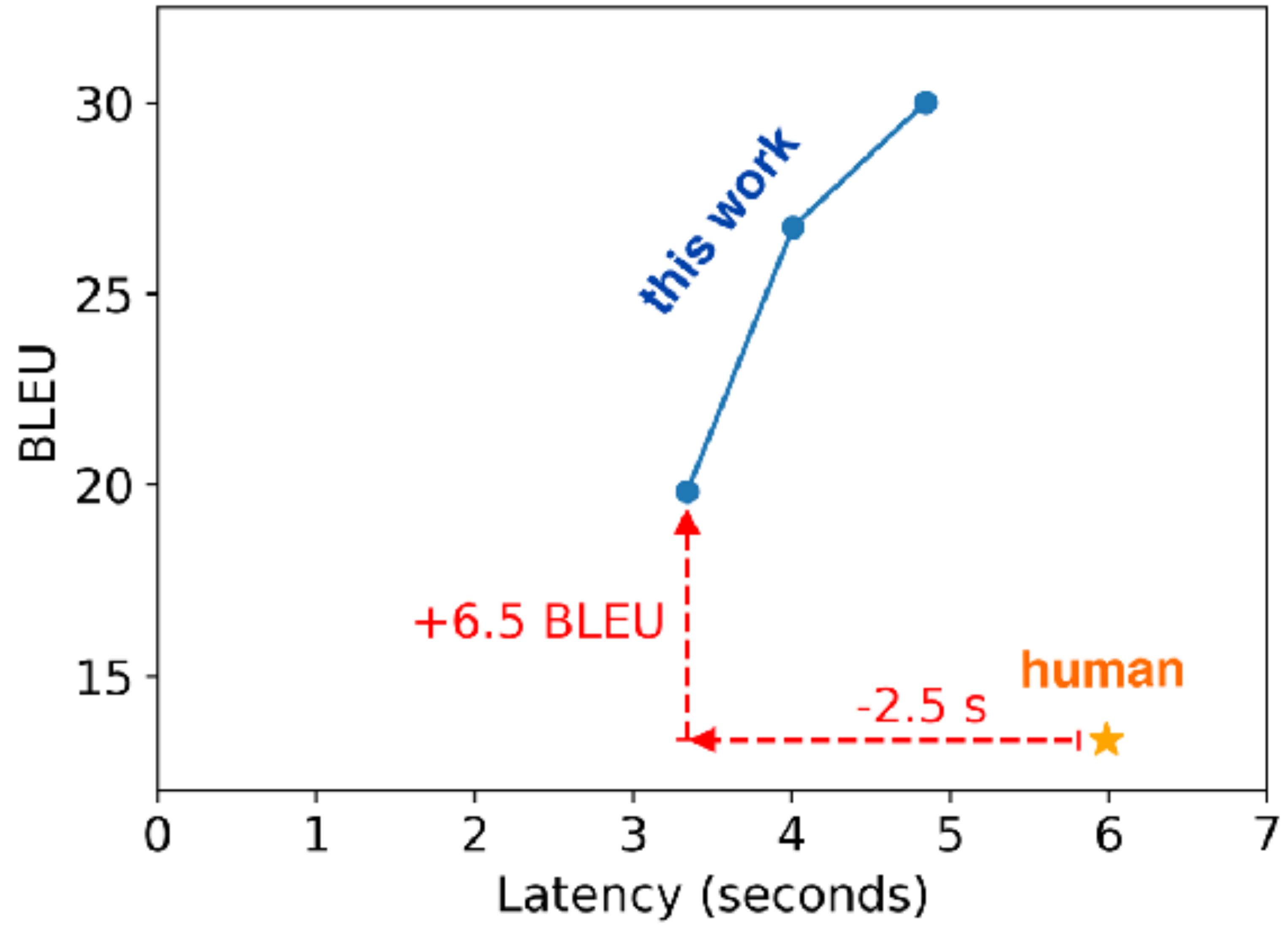
Self-Adaptive Speech-to-Speech Simultaneous Translation

- our speech-to-speech system achieves much lower latency and higher quality than professional simultaneous interpreters in the UN (En=>Ch)



human interpreter

our system



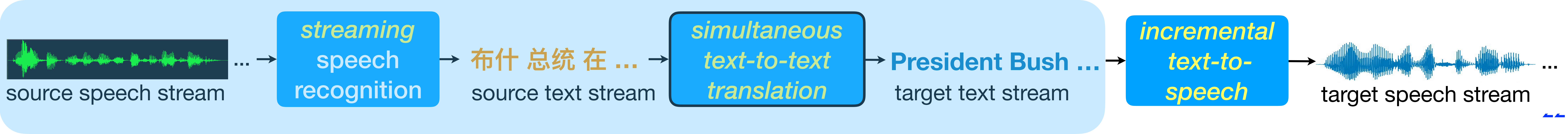
Part II(b): Direct (*non-pipelined*) Speech-to-Text

- Streaming ASR still causes the vast majority of errors in the pipeline
  - streaming ASR is fundamentally more challenging than offline ASR (no bidirectional models!)
  - esp. in code-switching: (zh-to-en) “to-B” =ASR=> “土逼” =MT=> “earth-forcing”
  - recovering from ASR errors (esp. homophones); directly speech-to-speech w/o text-to-text?
- Simultaneous Direct Speech-to-Text Translation
  - avoid error propagation
  - reduce latency (single model instead of two)
  - challenge: how to segment source speech?


Junkun Chen

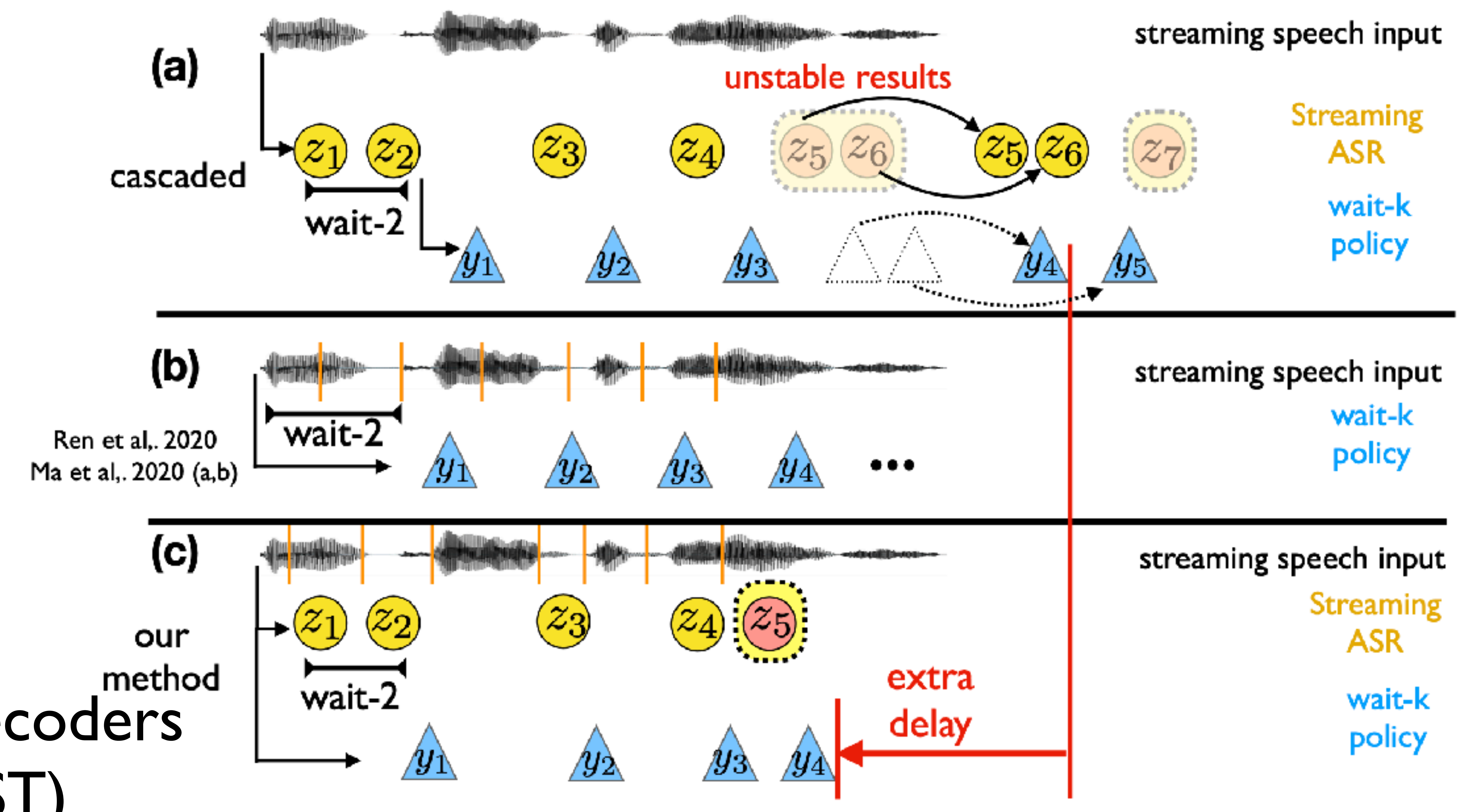

Mingbo Ma


Renjie Zheng



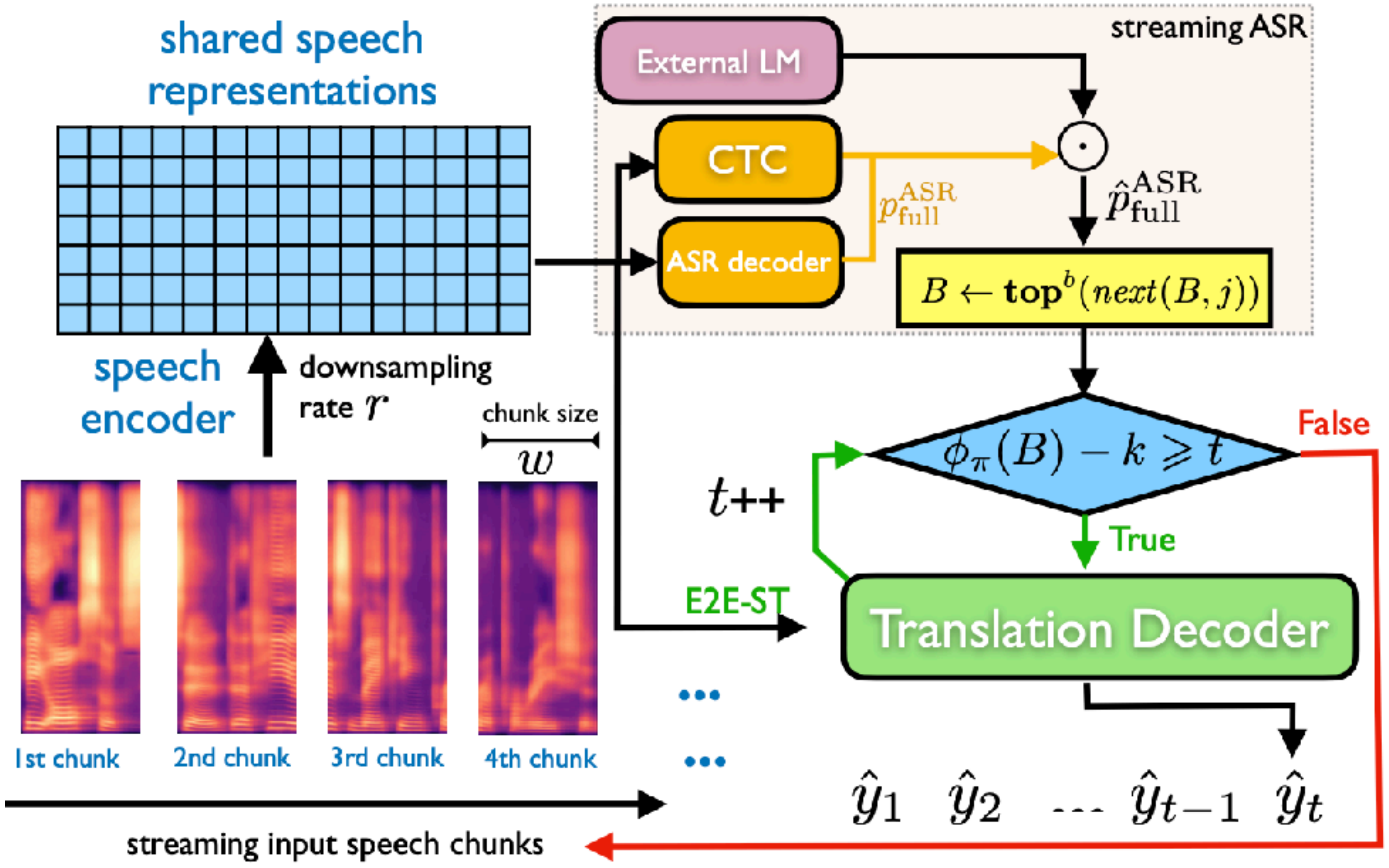
Direct Speech-to-Text Simultaneous Translation

- challenge: speech segmentation
- previous work
  - assume fixed # of words within a certain # of speech frames
  - or use CTC-based segmenter
- our work
  - two separate but *synchronized* decoders (streaming ASR & simultaneous ST)
  - streaming ASR beam search to guide, *but not feed as input to*, simultaneous ST
  - streaming ASR result also useful (caption)



(Chen et al., ACL 2021 Findings)

Decoding Policy: Streaming ASR-guided Wait-*k*



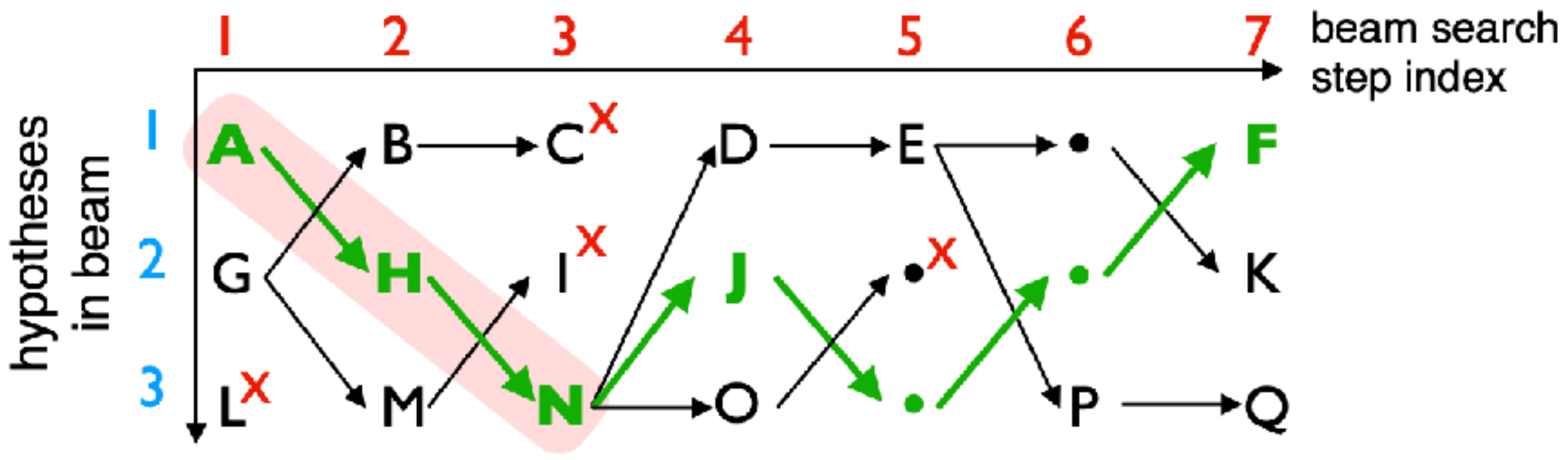


Figure 3: An example of streaming ASR beam search with beam size 3. LCP is shaded in red ($\phi_{\mathrm{LCP}}(B_7) = 3$); SH is highlighted in bold ($\phi_{\mathrm{SH}}(B_7) = 5$). We use • to represent empty outputs in some steps caused by CTC.

two sub-policies:
(a) Longest Common Prefix (LCP) — more conservative
(b) Shortest Hypothesis (SH) — more aggressive

- wait-*k* needs to know # of source "words" in speech: ask ASR beam search
- speech-to-text decoder does *not* depend on ASR output (only "# of words")

En-to-Zh Example

English ASR:	but then two weeks later she called me she said did you know that if you move to states you change your name and gendermarker
Stable (LCP):	but then two weeks later she called me she said did you know that if you move to states you change your name and gendermarker
Gold transcription:	but then two weeks later she called me she said did you know that if you move to the united states you could change your name and gender marker
LCP wait-1 Translation:	但是 两周 后 ， 她 叫 我 说 ， " 你 知道 这 是不是 在 美国 搬走 你 的 名字 和 性别 标记 ？ "
LCP wait-5 Translation:	但是 两周 后 ， 她 说 ， " 你 知道 吗 ， 如果 你 搬 到 美国 ， 你 可以 改变 你 的 名字 和 性别 标记 吗 ？ "
SH wait-1 Translation:	但是 两周 后 ， 她 叫 我 说 ， " 你 知道 这 是不是 搬 到 美国 ， 你 可以 改变 你 的 名字 和 性别 标记 ？ "
SH wait-5 Translation:	但是 两周 后 ， 她 说 ， " 你 知道 吗 ， 你 搬 到 美国 ， 你 可以 改变 你 的 名字 和 性别 标记 吗 ？ "
Cascade wait-1 Translation:	但是 ， 接着 两个 星期 后 ， 她 打电话 给 我 ， 她 说 ， " 你 知道 吗 ， 如果 你 搬 到 国家 ， 你 改变 你 的 名字 和 性别
Cascade wait-5 Translation:	但是 ， 两周 后 ， 她 打电话 给 我 ， 她 说 ， " 你 知道 吗 ， 如果 你 搬 到 国家 ， 你 改变 你 的 名字 和 性别 标记
fs-Cascade Translation:	但 两周 后 ， 她 打电话 给 我 ， 她 说 ， " 你 知道 吗 ， 如果 你 搬 到 国家 ， 你 改变 你 的 名字 和 性别 标记 吗 ？
fs-End2End-ST Tranlation (beam 1):	但是 两周 后 ， 她 说 ， " 你 知道 吗 ， 如果 你 搬 到 美国 ， 你 可以 改变 你 的 名字 和 性别 标记 吗 ？ "
Gold Reference:	两周 之后 她 又 打电话 给 我 ， 她 说 ， " 你 知道 如果 你 移居 到 美国 ， 你 个 可以 换 一个 名字 ， 并且 改变 你 的 性别 标识 么 ？

- "the united states" (美国) =ASR=> "states" =MT=> 国家
- another ASR (not shown): "the united states" =ASR=> "united" =MT=> 曼联
- SH policies faster than cascaded

En-to-De Example

English ASR:	can i be on this i don 't love that question
Stable (LCP):	can i be on this i don 't love that question
Gold transcription:	can i be honest i don 't love that question
LCP wait-3 Translation:	Kann ich ehrlich sein ? Ich liebe diese Frage nicht .
SH wait-3 Translation:	Kann ich ehrlich sein ? Ich liebe diese Frage nicht .
Cascade wait-3 Translation:	Kann ich da sein ? " Ich liebe diese Frage nicht .
Full-sentence-Cascade Translation:	Kann ich da sein ? Ich liebe diese Frage nicht .
Full-sentence-E2E-ST Tranlation (beam 1):	Kann ich ehrlich sein ? Ich liebe diese Frage nicht .
Gold Reference:	Darf ich ehrlich sein ? Ich mag diese Frage nicht .

chunk index	1	2	3	4	5	6	end
Gold transcript	can I be	**honest**	*SIL*	I don 't **love**	**that question**	*SIL*	
Gold translation	Darf ich **ehrlich** sein ?			Ich **mag diese Frage** nicht .			
Streaming ASR	can I			be **on this** I don 't	**love that question**		
simul-MT wait-3				Kann ich **da** sein ? '	Ich **liebe**		**diese Frage** nicht .
SH wait-3				Kann ich **ehrlich** sein ? Ich **liebe**	**diese Frage**		nicht .
LCP wait-3				Kann ich **ehrlich** sein ? Ich			**liebe diese Frage** nicht .

- ASR error ("honest" => "on this") propagated to MT
- direct system is also faster (lower latency) in generating "Ich liebe diese Frage"

Part III: Multimodal Models for Simultaneous Translation

multimodal pretraining for speech translation



Renjie Zheng Junkun Chen Mingbo Ma

(R. Zheng, J. Chen, et al., ICML 2021)

vision-aided simultaneous translation

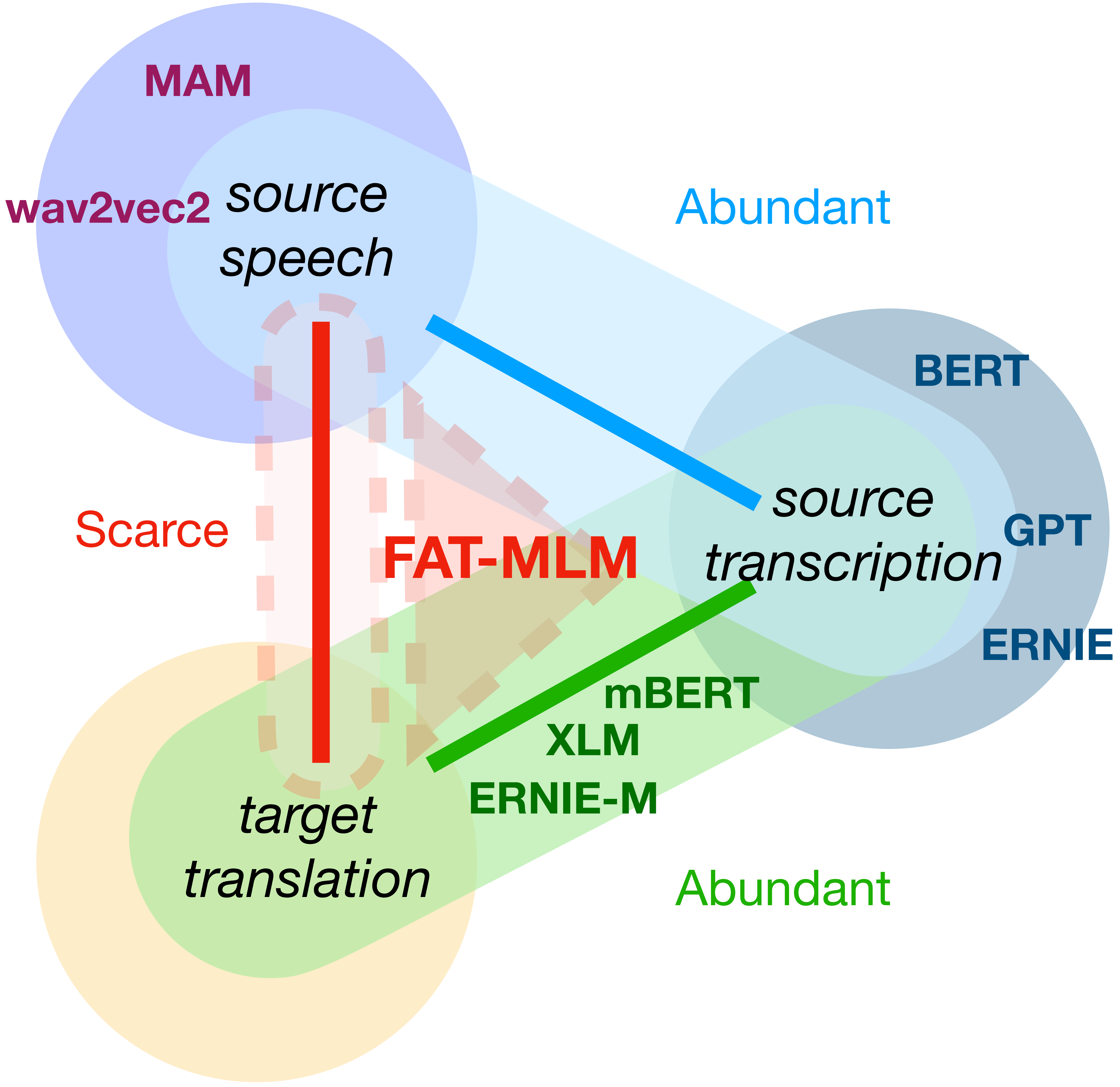
courtesy of L. Specia
NOT MY WORK!

Lucia Specia's group

(Caglayan et al., EMNLP 2020)

Part III(a): Multimodal Pretraining for Speech Translation

- here: direct full-sentence speech-to-text translation
  - next: direct simultaneous speech-to-text translation
- limitation of direct speech translation
  - large-scale parallel speech translation data is rare
  - but abundant data for ASR and text MT
- we propose a Fused Acoustic and Text Masked Language Model (FAT-MLM)
- encode source speech and bilingual text into a **unified representation** with self-supervision
- first speech-and-text multi-modal pretraining



Example 1

Source		those are their expectations of who you are not yours
Reference		那是他们所期望的你的样子而不是你自己的期望
Cascade	ASR	those are **there** expectations **to do** you are not yours
	Translation	那些都是希望**做到的，你不是你的**。
FAT-ST		这些是他们对你的期望，而不是你的期望。

Example 2

Source		she is not welcomed neither by father nor by mother
Reference		她不受欢迎,无论是父亲还是母亲
Cascade	ASR	she’s not welcomed neither by father **narby** mother
	Translation	她不欢迎父亲**纳尔比·**母亲。
FAT-ST		她并不欢迎父亲,也不属于我的母亲。

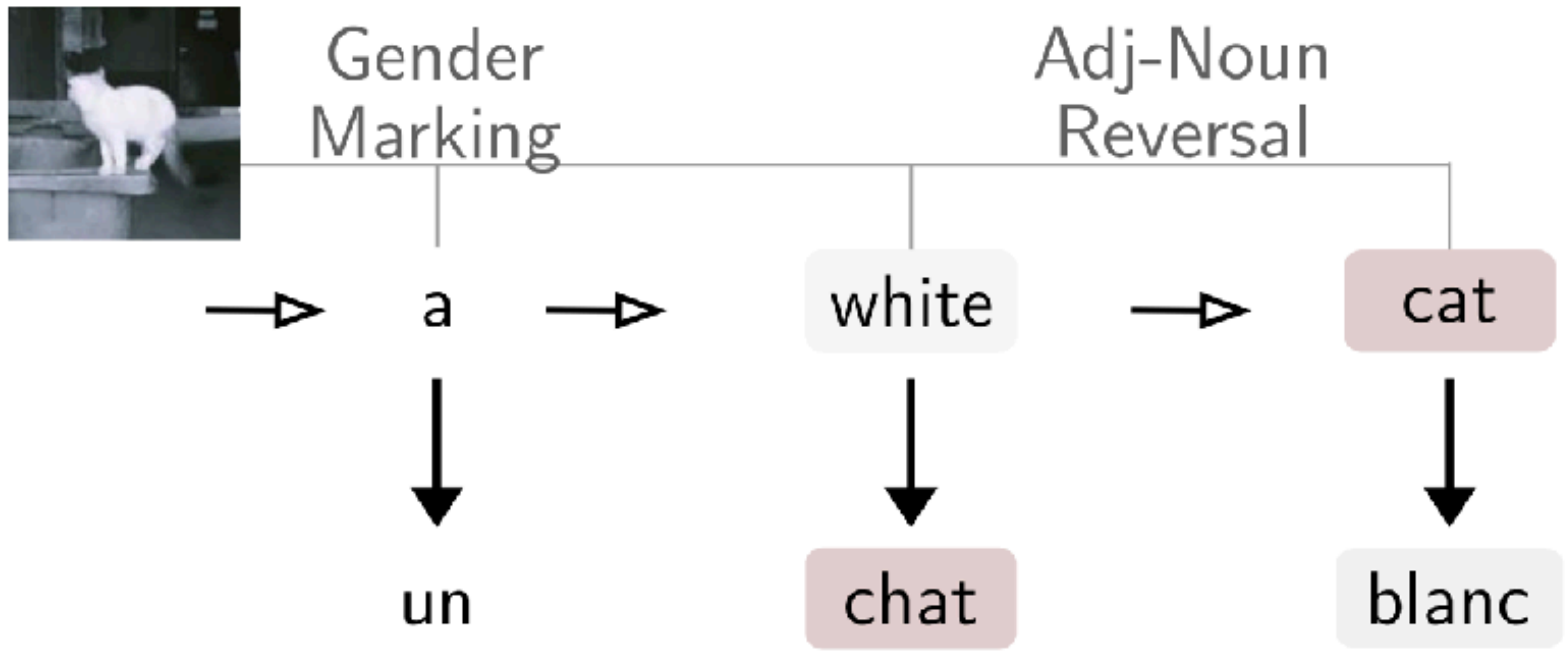
Part III(b): vision-aided simultaneous translation

- English=>French/German/… simul translation needs to anticipate
  - gender marking of the pronoun (un/une; ein/eine)
  - the head noun (a big house => una casa grande)
  - almost impossible for small *k* in wait-*k*
  - idea: image can help you anticipate!

Lucia Specia's group

(Caglayan et al., EMNLP 2020)

wait-1 English=>French



wait-1 English=>German

SRC : a young brunette woman ...
NMT : ein junger **brünette frau ...**
MMT : **eine junge brünette frau ...**

SRC : a black and white bird ...
NMT : **un** chien (dog) **noir et blanc ...**
MMT : **un oiseau (bird) noir et blanc ...**

wait-1 English=>French

Conclusions

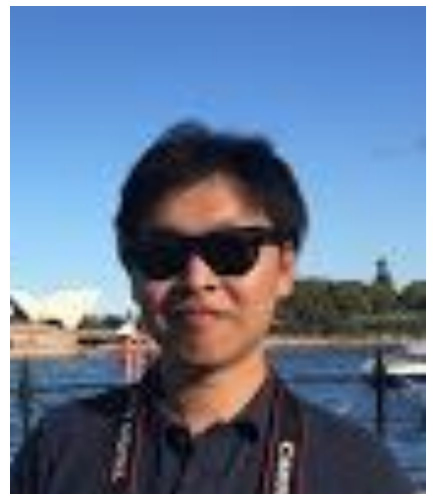
- prefix-to-prefix framework (esp. wait-*k* policy) is an easy & effective solution
  - turned simultaneous translation from obscurity to a hot topic
- adaptive (flexible) policy can improve latency and quality
- making the first steps towards simultaneous speech-to-speech pipeline
  - can surpass professional simultaneous interpreters in latency and quality
- direct simultaneous speech-to-text translation
  - avoids error propagation from streaming ASR, and reduces latency
  - speech translation guided by, but not using input from, streaming ASR beam search
- multimodal pretraining addresses data scarcity for direct speech-to-text
- vision can help you anticipate in simultaneous translation!

Co-authors and Collaborators


Mingbo Ma


Renjie Zheng


Junkun Chen


Kaibo Liu


Baigong Zheng*

Baidu Research & Oregon State University


Ken Church


Jiahong Yuan

Baidu Research


Zhongjun He


Hao Xiong*


Chuanqiang Zhang


Ruiqing Zhang


Hua Wu


Haifeng Wang

Baidu百度

*former members

非常 感谢 您　　来　听　我　　的 演讲

Thank you very much for listening to my speech

Baidu Research

Oregon State University