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.
Co-authors and Collaborators

Mingbo Ma  Renjie Zheng  Junkun Chen  Kaibo Liu  Baigong Zheng*

Ken Church  Jiahong Yuan

Zhongjun He  Hao Xiong*  Chuanqiang Zhang  Ruiqing Zhang  Hua Wu  Haifeng Wang

*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 worldwide (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

- **high quality**
  - word-by-word translation
  - one of AI's holy grails
  - needs fundamentally new ideas!

- **low quality**
  - simultaneous interpretation

- **low latency**
  - ~3 seconds
  - streaming speech recognition
  - source speech stream

- **high latency**
  - consecutive interpretation
  - full-sentence machine translation
  - seq-to-seq is already very good

- **written translation**
  - target speech stream
  - target text stream
  - incremental text-to-speech

- **simultaneous work in simultaneous translation**
  - previous foundational work & Aims 1-3

- **President Bush**
  - written translation

- **source text stream**
  - text-to-text translation

- **target speech stream**
  - text-to-speech

- **source speech stream**
  - streaming speech recognition
  - text-to-text translation

- **incremental text-to-speech**
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

# of researchers in simultaneous translation courtesy of Hua Wu
Baidu World Conference, Nov. 2017
full-sentence translation (latency: 10+ secs)

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

Media coverage:
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)

Grissom et al, 2014

\[ \text{ich bin mit dem Zug nach Ulm gefahren} \]
\[ \text{I am with the train to Ulm traveled} \]
\[ \text{I (...... waiting......) traveled by train to Ulm} \]

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 tailored 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

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

source: \[1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5\]  seq-to-seq

target: \[\ldots \text{wait whole source sentence} \ldots\]

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

source: \[1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5\]  prefix-to-prefix (wait-k)

target: \[\text{wait } k \text{ words}\]

President Bush meets with Russian President Putin in Moscow
Research Demo

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

江泽民对法国总统的来华访问表示感谢。

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

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.
“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

- # of researchers in simultaneous translation
  courtesy of Hua Wu
Part I (b): Towards Adaptive Translation Policies

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

![ZH → EN Diagram](image)

- **4-ref BLEU**
  - wait-k method
  - single
  - ensemble all
  - ensemble top3
  - test-time wait-k
  - wait-if-diff
  - wait-if-worse

- **1-ref BLEU**
  - wait-k method
  - single
  - ensemble all
  - ensemble top3
  - test-time wait-k
  - wait-if-diff
  - wait-if-worse

---

**pinyin| input gloss**

<table>
<thead>
<tr>
<th>wén</th>
<th>tián</th>
<th>shănghăizhè</th>
<th>de</th>
<th>júshā</th>
<th>bīngsī</th>
<th>zài</th>
<th>chéngzhī</th>
<th>de</th>
<th>tòngqìng</th>
<th>hé</th>
<th>āi</th>
<th>dào</th>
</tr>
</thead>
<tbody>
<tr>
<td>“我们向受害者的家属” 表示最诚挚的同情和哀悼。”</td>
<td>we to victim's family express most sincere sympathy and condolence.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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**

---

14
Idea 2 (ACL 2019): Policy as Model ("READ" as a word)

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

<table>
<thead>
<tr>
<th></th>
<th>fixed-latency policies</th>
<th>adaptive policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>full-sentence MT model</td>
<td>Dalvi et al. (2018); test-time wait-$k$ (Ma et al. 2018)</td>
<td>Grissom et al. (2014); Cho &amp; Esipova (2016); Satija &amp; Pineau (2016); Gu et al. (2017); Alinejad et al (2018); …</td>
</tr>
<tr>
<td>simultaneous MT model</td>
<td>wait-$k$ (Ma et al. 2018)</td>
<td>Arivazhagan et al. (ACL 2019)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>idea 2: B. Zheng et al. (ACL 2019)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>idea 1: B. Zheng et al. (ACL 2020)</td>
</tr>
</tbody>
</table>
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)

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!

<table>
<thead>
<tr>
<th>Speech Rate</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5×</td>
<td>2.00 ± 0.08</td>
</tr>
<tr>
<td>0.6×</td>
<td>2.32 ± 0.08</td>
</tr>
<tr>
<td>0.75×</td>
<td>2.95 ± 0.07</td>
</tr>
<tr>
<td>Original</td>
<td>4.01 ± 0.08</td>
</tr>
<tr>
<td>1.33×</td>
<td>3.34 ± 0.08</td>
</tr>
<tr>
<td>1.66×</td>
<td>2.40 ± 0.09</td>
</tr>
<tr>
<td>2.0×</td>
<td>2.06 ± 0.04</td>
</tr>
</tbody>
</table>

(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)

<table>
<thead>
<tr>
<th>Time</th>
<th>0 s</th>
<th>1 s</th>
<th>2 s</th>
<th>3 s</th>
<th>4 s</th>
<th>5 s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Mr Chairman in all our work on conventional arms control</td>
<td>the international community is aided by civil society</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>主席</td>
<td>主席 先生</td>
<td>我们 在 常规武器 控制 方面 的</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT-3 (this work)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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!)
- Especially 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?

Mingbo Ma
Renjie Zheng
Junkun Chen
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

- **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”)

Figure 3: An example of streaming ASR beam search with beam size 3. LCP is shaded in red ($\phi_{LCP}(B_7) = 3$); SH is highlighted in bold ($\phi_{SH}(B_7) = 5$). 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
“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


#### Full-sentence-E2E-ST


<table>
<thead>
<tr>
<th>Gold Reference</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>end</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold transcript</td>
<td>can I be</td>
<td>honest</td>
<td>SIL</td>
<td>I don 't love</td>
<td>that question</td>
<td>SIL</td>
<td></td>
</tr>
<tr>
<td>Streaming ASR simul-MT wait-3</td>
<td>can I be</td>
<td>on this</td>
<td>I don 't love</td>
<td>that question</td>
<td>Ich liebe diese Frage nicht .</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SH wait-3</td>
<td>Kann ich ehrlich sein ? Ich liebe diese Frage nicht .</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCP wait-3</td>
<td>Kann ich ehrlich sein ? Ich liebe diese Frage nicht .</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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

vision-aided simultaneous translation
courtesy of L. Specia


Lucia Specia’s group

(Caglayan et al., EMNLP 2020)
• 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**

那些是他们所期望的你的样子而不是你自己的期望

**ASR**

those are there expectations to do you are not yours

**Cascade**

那些都是希望做到的，你不是你的。

**Translation**

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

ASR

she's not welcomed neither by father narby mother

Cascade

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

Gender Marking

a

Adj-Noun Reversal

white

un

cat

chat

blanc

wait-1 English=>German

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

wait-1 English=>French

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!
Thank you very much for listening to my speech.