A Multimodal Simultaneous Interpretation Prototype: Who Said What

Xiaolin Wang, Masao Utiyama and Eiichiro Sumita
{xiaolin.wang, mutiyama, eiichiro.sumita}@nict.go.jp
Advanced Translation Research and Development Promotion Center
National Institute of Information and Communications Technology, Japan
Conventional Way of Simultaneously Interpreting Video Streams

- Subtitle ↔ Translation

Video 1

So at some point I still hate my life so far, but I really have to decide to do what I want to do if I really have only one time.
Readable Way of Simultaneously Interpreting Video Streams

- Chat Log ↔ Speaker + Translation (Who Said What)

---

I thought I'd come and do what I liked, but after a while I thought I couldn't do it, so I gave up.

But after three years, I still want to do something in the world of cooking, so I quit my job and now I'm working at a restaurant, and there are quite a few people who want to open a restaurant after five years.

So at some point I still hate my life so far, but I really have to decide to do what I want to do if I really have only one time.
User Interface : Chat Log

- **Speaker Tag**
  - Plain text of “spk n”
  - facilitate post-editing

- **Face Icon**
  - Speakers’ sentiments
  - Double check speaker tag

- **Translation**
  - Content of speech
### Plain transcript

- wonderful
- and what would you say are some of your weaknesses
- one of my biggest weaknesses is asking for help when I need it
- I'd like to do better at that
- I appreciate your honesty mister wang
- what can you tell me about some of your goals over the next few years
- my primary goal is to gain more work experience
- so a position like this would help me meet that goal

### Annotated transcript

<table>
<thead>
<tr>
<th>spk0</th>
<th>wonderful</th>
</tr>
</thead>
<tbody>
<tr>
<td>spk0</td>
<td>and what would you say are some of your weaknesses</td>
</tr>
<tr>
<td>spk1</td>
<td>one of my biggest weaknesses is asking for help when I need it</td>
</tr>
<tr>
<td>spk1</td>
<td>I'd like to do better at that</td>
</tr>
<tr>
<td>spk0</td>
<td>I appreciate your honesty mister wang</td>
</tr>
<tr>
<td>spk0</td>
<td>what can you tell me about some of your goals over the next few years</td>
</tr>
<tr>
<td>spk1</td>
<td>my primary goal is to gain more work experience</td>
</tr>
<tr>
<td>spk1</td>
<td>so a position like this would help me meet that goal</td>
</tr>
</tbody>
</table>
Implementations

• Challenges
  • Recognize speakers from video streams
  • Maintain low latency for interpretation

• Solution:
  • Multimodal speaker recognition
  • Integrate speaker recognition with sentence-based SI
Multimodal speaker recognition

- Voice embedding
- Face embedding
- Active speaker detector: find the faces of speakers
Integrate speaker recognition with sentence-based SI

- Launch speaker predictor on video clips (video stream + temporal range)
  - Assemble large batches for GPU efficiency
- Parallelize translator with speaker predictor
  - Low latency
Weakness & Future

• Weakness
  • Highly dependent on textual-based sentence segmentor
  • Performs poorly on video with difficult speeches
    • Poor sentence segmentation
    • Poor translation quality

• Future: robust multimodal SI
  • Fuse audio and visual input with
    • Sentence segmenter
    • Translator
    • ASR

Thank you! (Q&A)