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

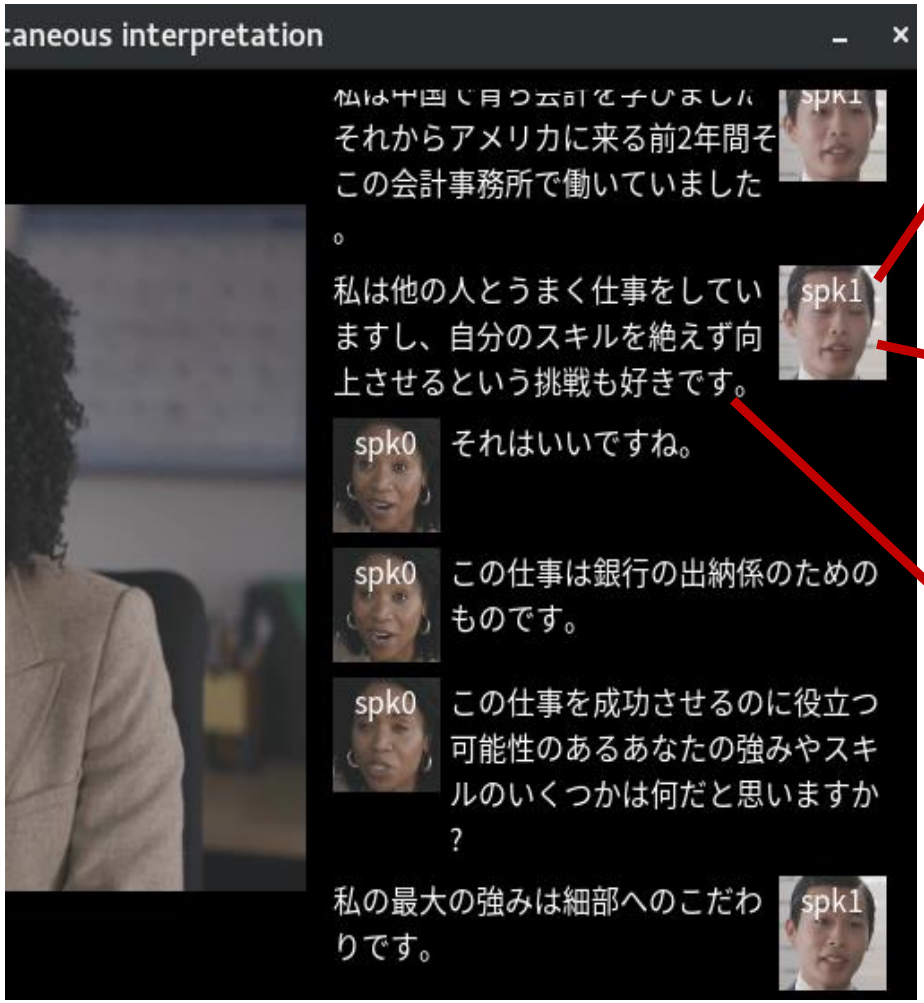
Video 1

Readable Way of Simultaneously Interpreting Video Streams

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

[Video 2]

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*

Post-Editing with Speaker Tags

- 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

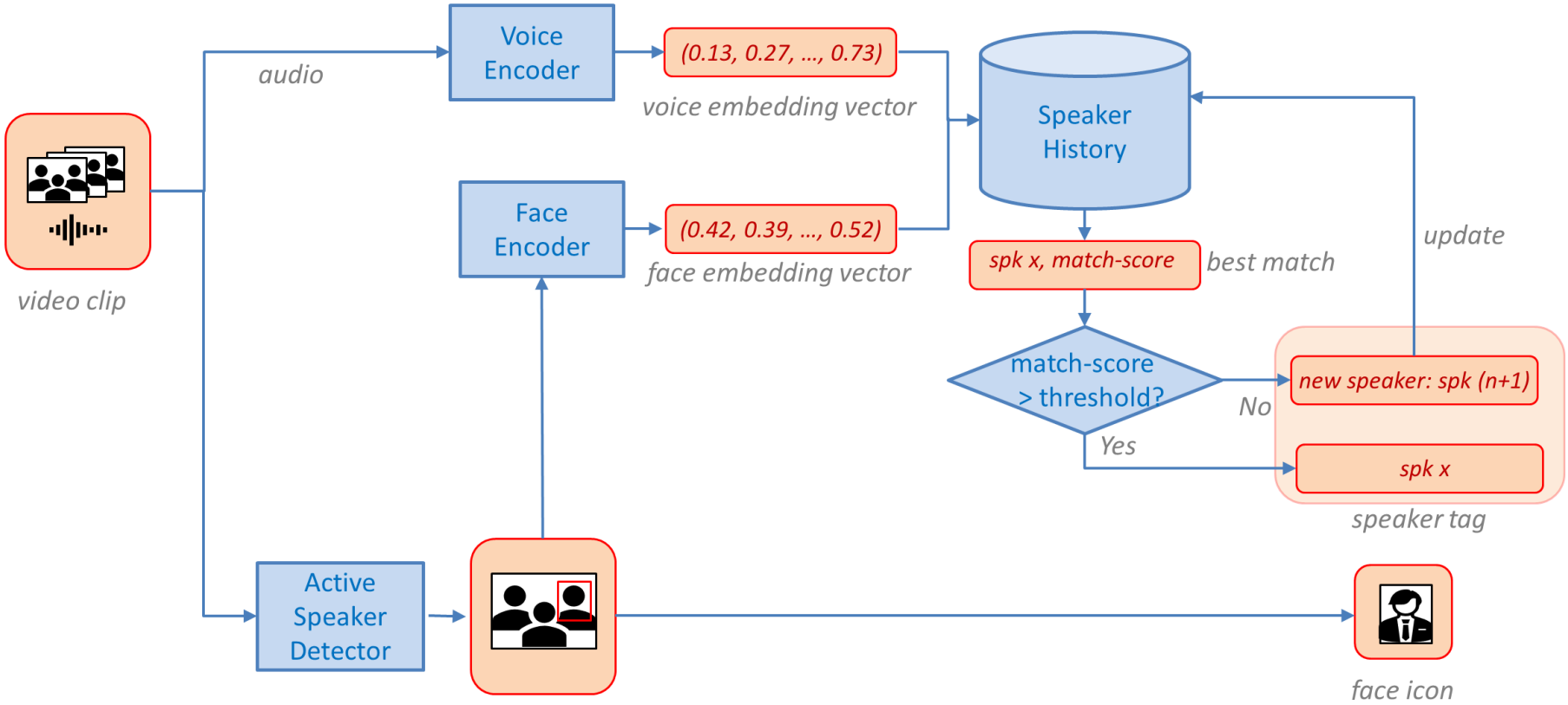
- Annotated transcript

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

Implementations

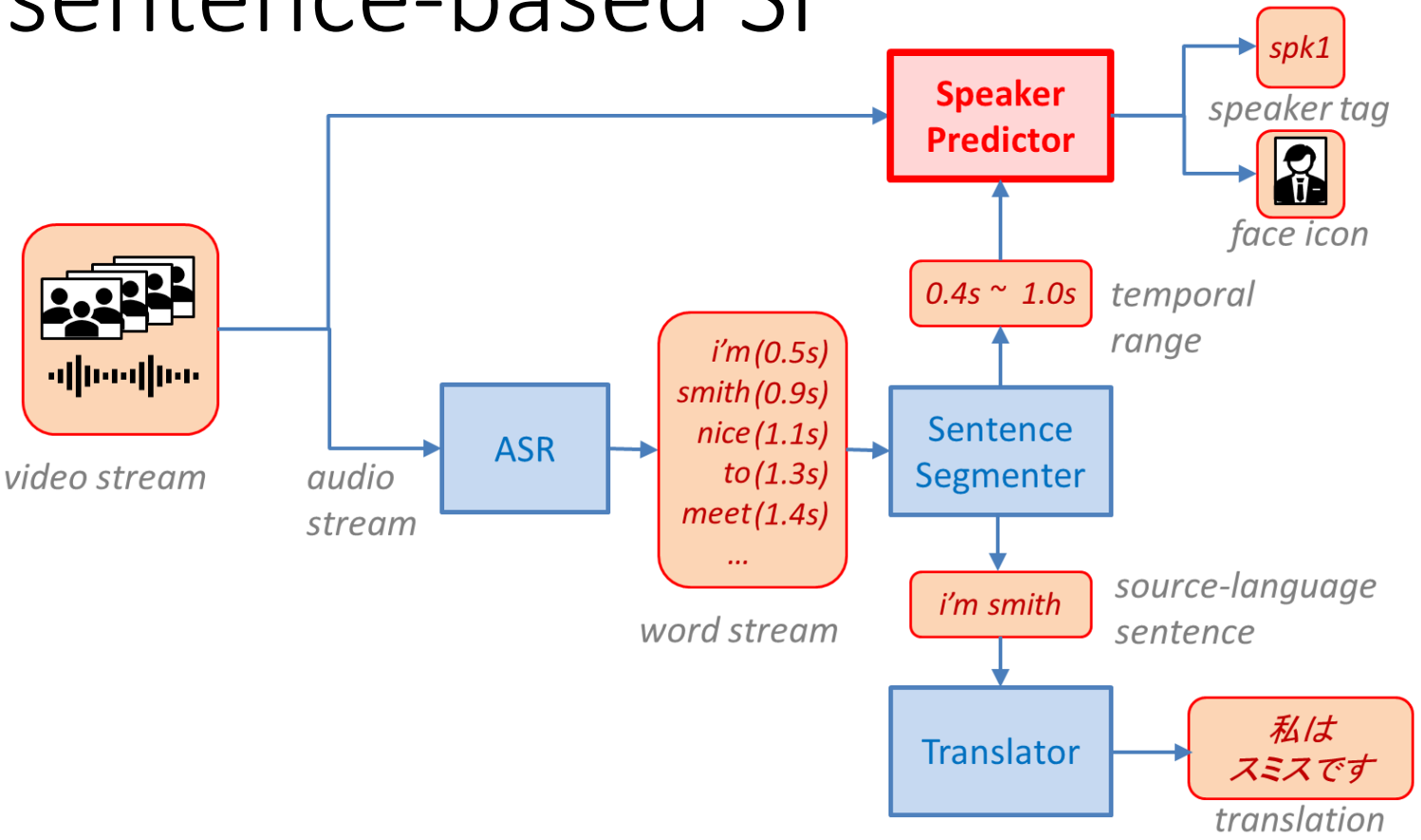
- Challenges
  - Recognize speakers from video streams
  - Maintain low latency for interpretation
- Solution:
  - Multimodal speaker recognition
  - Integrate speaker recognition with sentence-based SI

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)