Simultaneous Speech Translation in Google Translate

Jeff Pitman <jrp@google.com>
For animations, see: t.co/mz6oZiLEP4
Agenda

01 Overview

02 Long-form Audio Input

03 Streaming Translation

04 Streaming Text-to-Speech

05 Putting It Together
Overview
Conversational Turn-taking

2011

Components
- ASR
- MT
- TTS

Model Orchestration
Client-based Model Orchestration

Client

- ASR
- MT
- TTS
(Low) Latency is a feature.
"The speech to text and language translation feature on this app is unreal, I think you lot have basically invented the babelfish"
milliseconds at 95%, previously

milliseconds at 95%, now

milliseconds at 90%, now
User experience

Input interactions

- Tap and hold
- Quick tap
- Auto mic
Auto Mic

- audio
- langid
- source
- target

Time

Decision

Stop
What if we kept the microphone on?
Long-form Audio Input

Codecs

The Timeout

ASR Model training
Codecs

AMR-WB\(^1\) only worked well with clean recording environments and at close distance to the microphone.

Opus\(^2\) @24kbps performed just as well as uncompressed audio. Ended up using 32kbps.

1. Adaptive Multi-rate Wideband
2. Opus
The Timeout

**Problem:** ASR limited to 30 second sessions. But, anything could cause a disconnection.

**Solution:** Maintain audio buffer on client to stitch sessions together.

1. live-transcribe-speech-engine
ASR Model

Key insight was to move to models trained on long-form audio.
Streaming Translation
UX Research

Participants thought that the instability of the text results were disruptive.

Without preparation, professional interpreters are roughly 60% to 70% accurate in simultaneous interpretation.

Research¹ has shown that audiences get uncomfortable if results take too long.

---

We can re-use off-the-shelf ASR and NMT systems by using edit distance heuristics to stabilize prefixes.

We can further improve stabilization by making NMT prefix-aware. Beam search is then constrained on prefixes.

We evaluate performance using a metrics triple of BLEU, Voice-to-eye Latency, and Erasure (flickering rate).
## Unspoken Punctuation

<table>
<thead>
<tr>
<th>DETECT LANGUAGE</th>
<th>ENGLISH</th>
<th></th>
<th>KOREAN</th>
<th>ENGLISH</th>
<th>CHINESE (TR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>let us see grandma</td>
<td>×</td>
<td></td>
<td>讓我們看看奶奶</td>
<td>✯</td>
<td>Ràng wǒmen kàn kàn nǎinai</td>
</tr>
<tr>
<td><img src="play" alt="Audio" /></td>
<td>19 / 5000</td>
<td></td>
<td><img src="play" alt="Audio" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DETECT LANGUAGE</th>
<th>ENGLISH</th>
<th></th>
<th>KOREAN</th>
<th>ENGLISH</th>
<th>CHINESE (TR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Let us see, grandma.</td>
<td>×</td>
<td></td>
<td>奶奶，讓我們看看。</td>
<td>✯</td>
<td>Nǎinai, ràng wǒmen kàn kàn.</td>
</tr>
<tr>
<td><img src="play" alt="Audio" /></td>
<td>21 / 5000</td>
<td></td>
<td><img src="play" alt="Audio" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

+8 BLEU
Streaming Text-to-Speech
Goals

Voice-to-ear Latency

Prosody

Pure VUI?
Voice-to-ear

Slow finality of ASR results

Short-form ASR models

TTS Speed
Prosody

TTS Speed

Length limitations
Pure Voice UI

Quality

Navigation
Putting It Together

Evaluation

Results
Evaluation

We wanted to see if human judgement in a controlled environment can help make launch decisions.
Initial setup

Asked 3 bilingual raters to watch original video, read final and static NMT output, answer adequacy/fluency and gist questions.

Test set

~100 1-minute publically available videos.

Focused on clean audio with 1 person speaking.
Problems

Domain of test sets misaligned across languages

Raters were not trustworthy .. understanding source language was a bias .. just answering yes to everything was a bias.
Improvements

Minimized video selection bias with better QC

Minimized bilingual bias by using a monolingual template

Ground truth


System output

Results

Launched support for 10 languages.

Launched streaming TTS support for Pixel Buds.
What's next?
Advancing Speech Translation

Long-form Audio Input

Streaming Translation

Streaming Text-to-Speech

Evaluation
Thank You

Jeff Pitman
Senior Staff Engineering Manager

Deck Props: Shilp Vaishnav, Tom Small, Kannu Mehta, Mengmeng Niu, Bryan Lin, Naveen Ari, Colin Cherry