Simultaneous Speech Translation in Google Translate

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

Google Research

Agenda

01 Overview

02 Long-form Audio Input

03 Streaming Translation

04 Streaming Text-to-Speech

05 Putting It Together

Google Research

01

Overview

Google Research

Conversational Turn-taking

2011

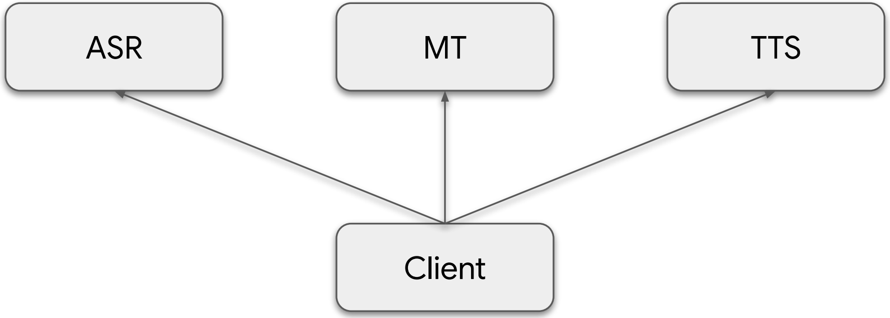
Components

- ASR
- MT
- TTS

Model Orchestration

Google Research

Client-based Model Orchestration

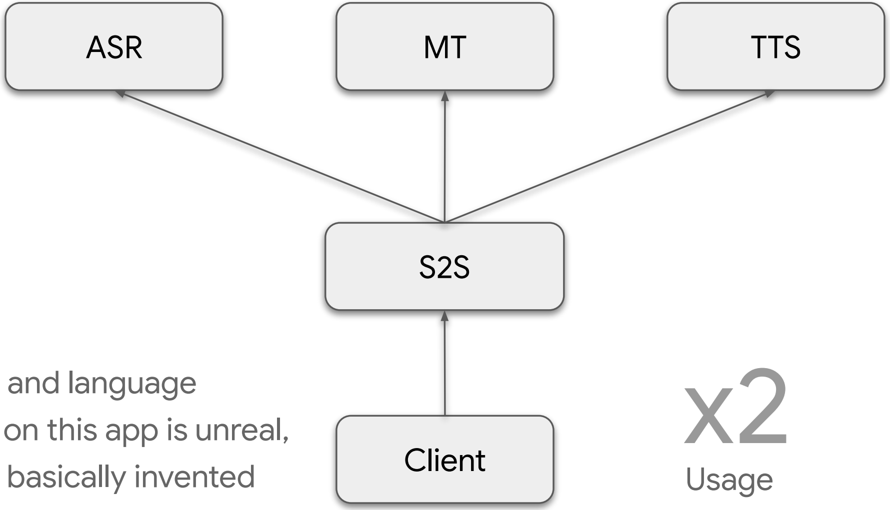


Google Research

(Low) Latency is a feature.

Google Research

Server-based Model Orchestration

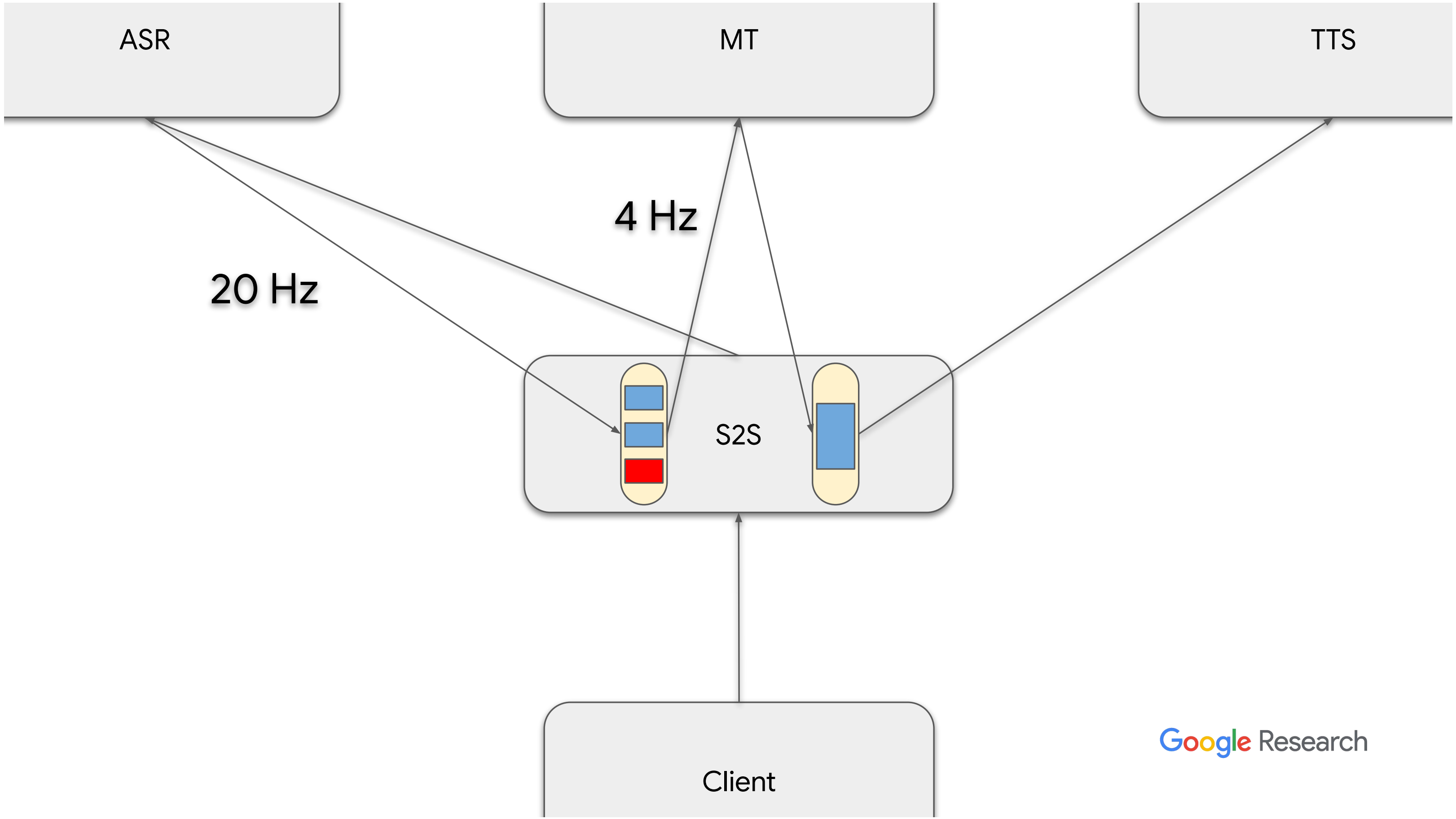


"The speech to text and language translation feature on this app is unreal, I think you lot have basically invented the babelfish

x2

Usage

Google Research


ASR
MT
TTS
4 Hz
20 Hz
S2S
Client
Google Research

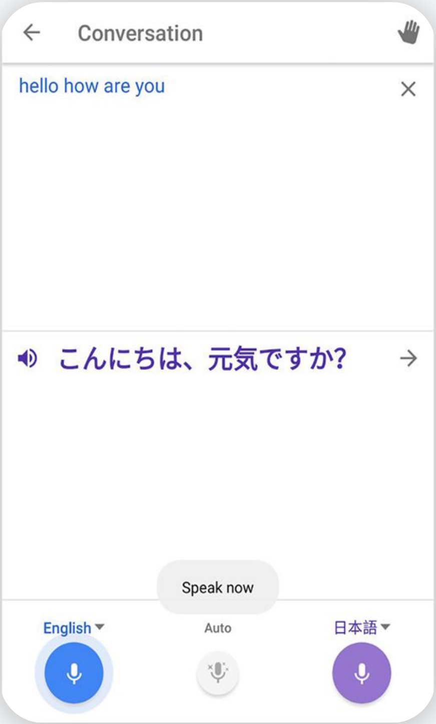
3100	950	500
milliseconds at 95%, previously	milliseconds at 95%, now	milliseconds at 90%, now

Google Research

User experience

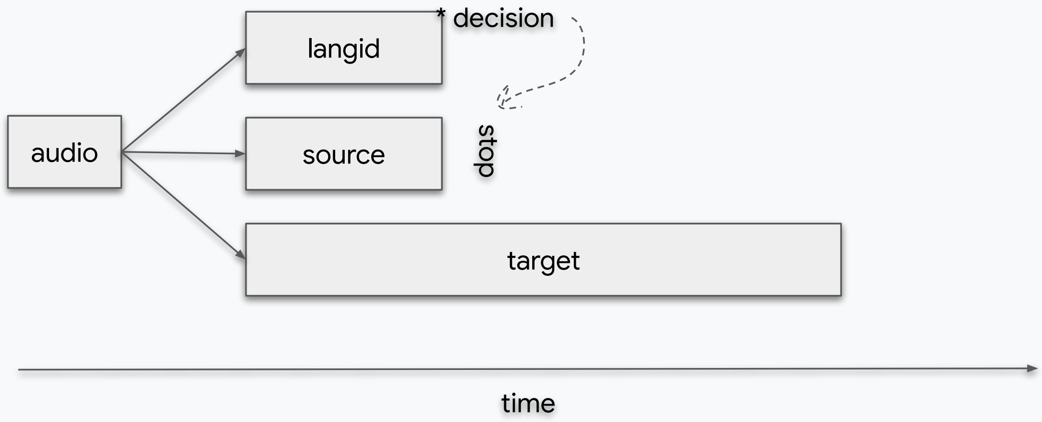
Input interactions

- Tap and hold
- Quick tap
- Auto mic



Google Research

Auto Mic



Google Research

What if we kept the microphone on?

Google Research

02

Long-form Audio Input

Codecs

The Timeout

ASR Model training

Google Research

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

Google Research

The Timeout

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

Solution[1]: Maintain audio buffer on client to stitch sessions together.

1. live-transcribe-speech-engine

Google Research

ASR Model

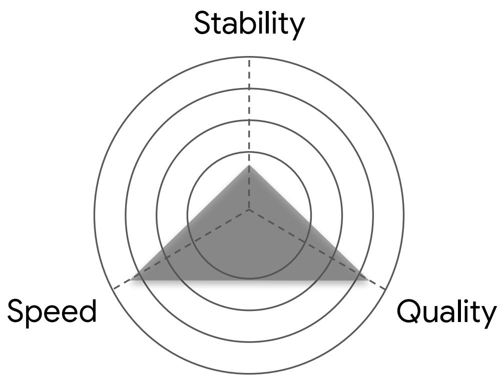
Key insight was to move to models trained on long-form audio.

Google Research

03

Streaming Translation

Google Research



Google Research

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.

Research[1] has shown that audiences get uncomfortable if results take too long.

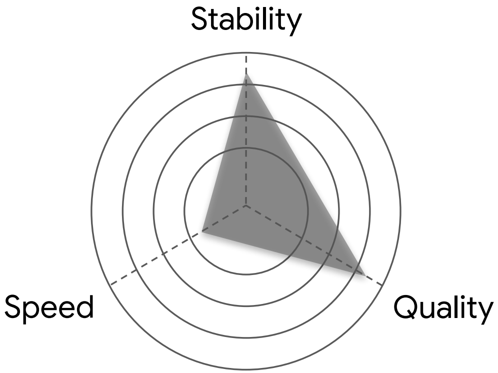
1. Lee, T.-H. 2002. “Ear voice span in English into Korean simultaneous interpretation.” Meta 47 (4): 596–606.

“The sentence continues to change while I’m reading it and it is making me nervous.”

Participant

Use case	ASR	TTS	Overall experience
Lecture	⚠️	✔️	⚠️
Museum tour	⚠️	❌	⚠️
Walking city tour	⚠️	❌	⚠️
Boat / Bus tour	❌	⚠️	❌
Airport	❌	✔️	❌

Google Research



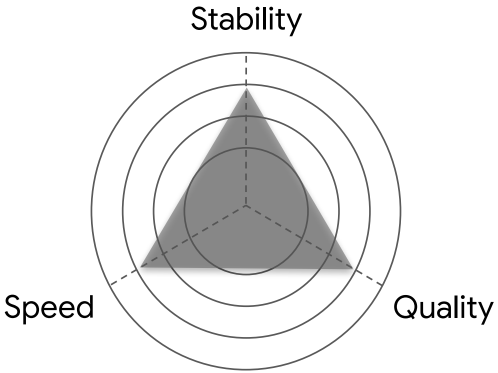
Google Research

Tech. Research

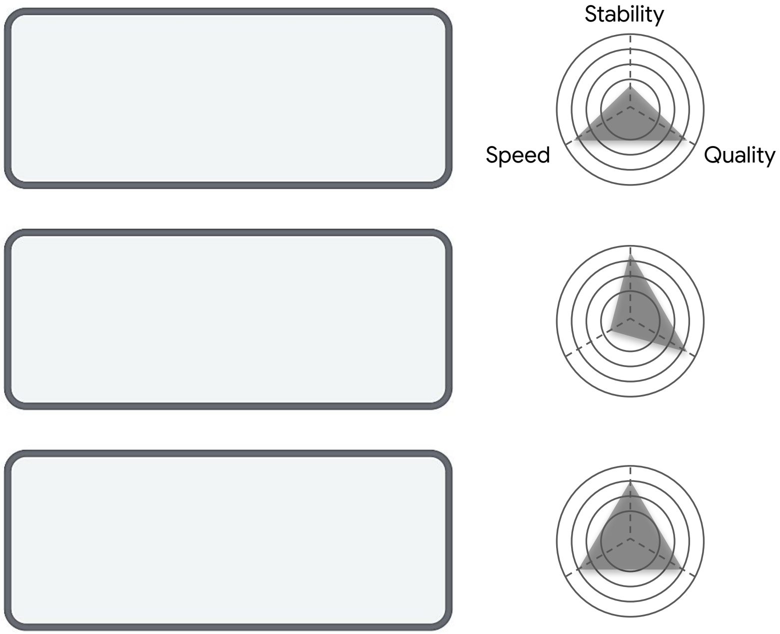
We can re-use off-the-shelf ASR and NMT systems by using edit distance heuristics to stabilize prefixes.

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).

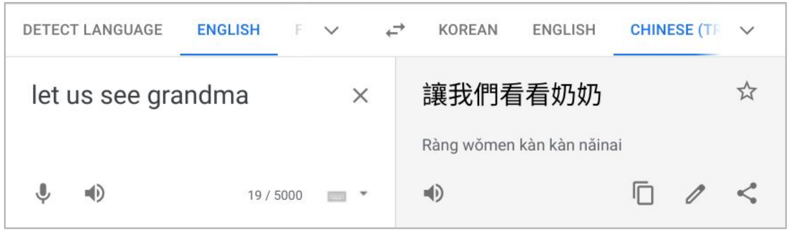


Google Research



Google Research

Unspoken Punctuation



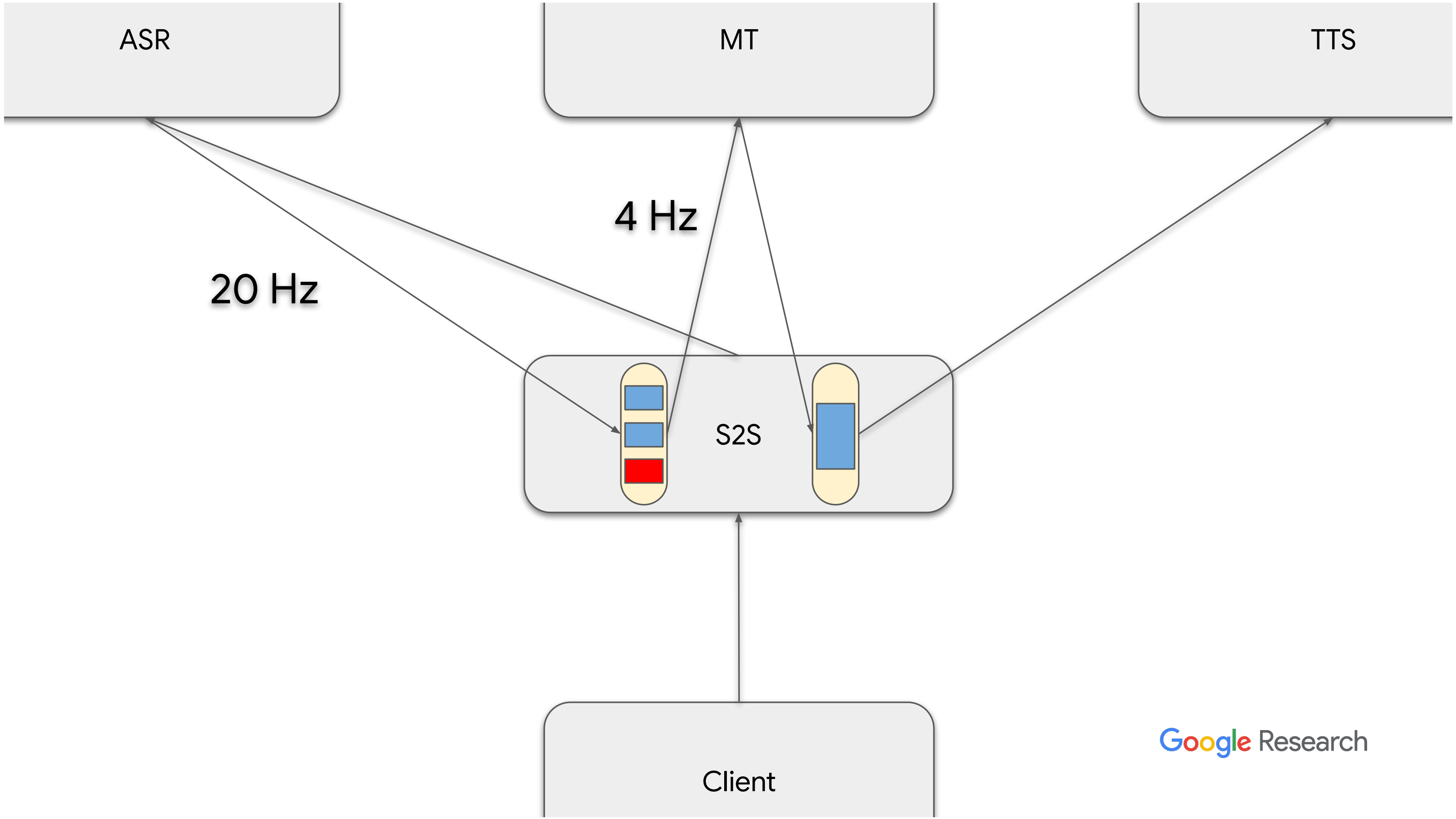
+8 BLEU

Google Research

04

Streaming Text-to-Speech

Google Research


ASR
MT
TTS
4 Hz
20 Hz
S2S
Client
Google Research

Goals

Voice-to-ear Latency

Prosody

Pure VUI?

Google Research

Voice-to-ear

Slow finality of ASR results

Short-form ASR models

TTS Speed

Google Research

Prosody

TTS Speed

Length limitations

Google Research

Pure Voice UI

Quality

Navigation

Google Research

05

Putting It Together

Evaluation

Results

Google Research

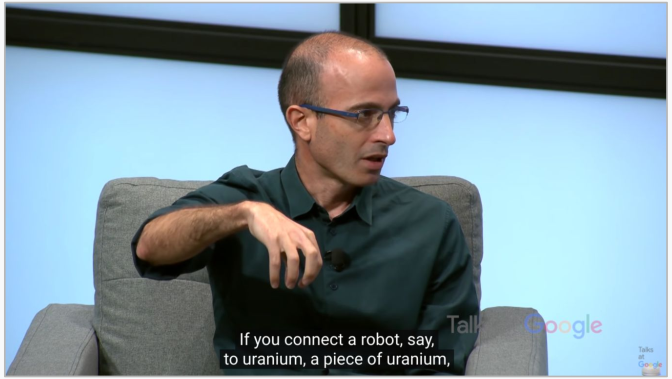
Evaluation

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

Google Research

Initial setup

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



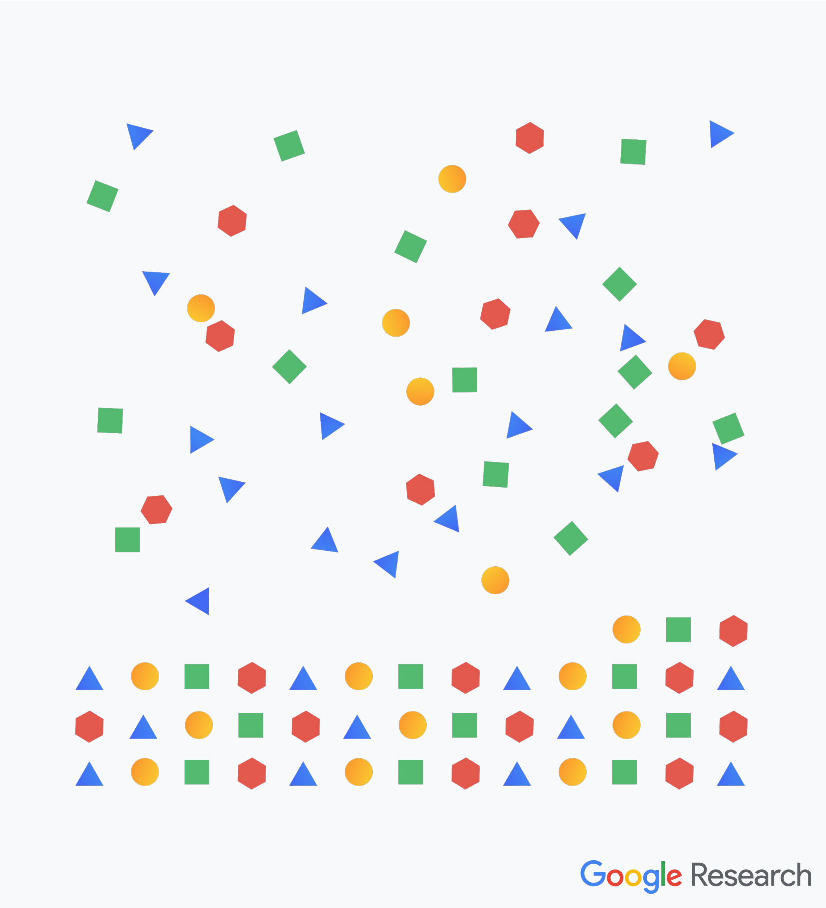
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent quis dolor lacus. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. In eu mi placerat, facilisis tellus vitae, efficitur nisi. Nulla placerat placerat sem, tempor vulputate libero suscipit sed. Mauris sit amet massa eu justo dignissim pharetra. Praesent sapien tortor, ornare et leo nec, aliquet suscipit nisi. Aenean egestas mauris eget hendrerit finibus. In eleifend ex pharetra tellus dignissim.

Google Research

Test set

~100 1-minute publically available videos.

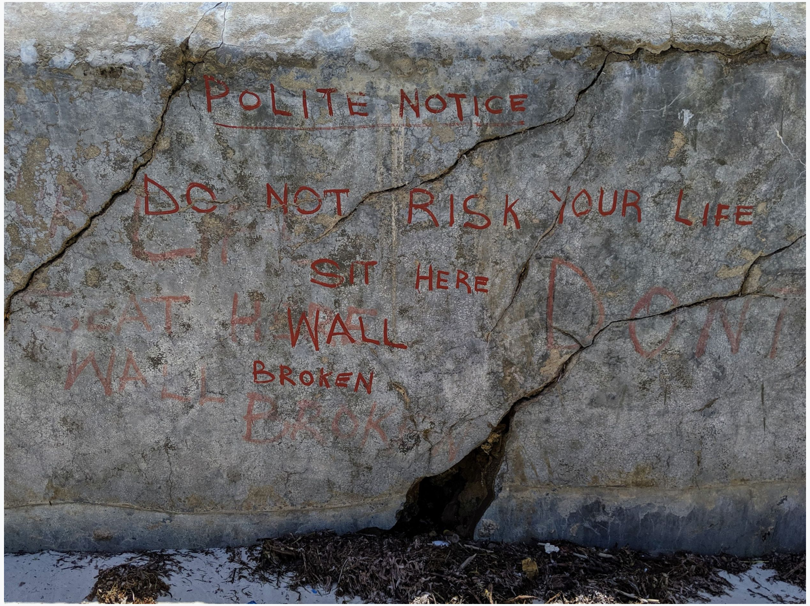
Focused on clean audio with 1 person speaking.



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.



Google Research

Improvements

Minimized video selection bias with better QC

Minimized bilingual bias by using a monolingual template

Ground truth

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent quis dolor lacus. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. In eu mi placerat, facilisis tellus vitae, efficitur nisi. Nulla placerat placerat sem, tempor vulputate libero suscipit sed. Mauris sit amet massa eu justo dignissim pharetra. Praesent sapien tortor, ornare et leo nec, aliquet suscipit nisi. Aenean egestas mauris eget hendrerit finibus. In eleifend ex pharetra tellus dignissim.

System output

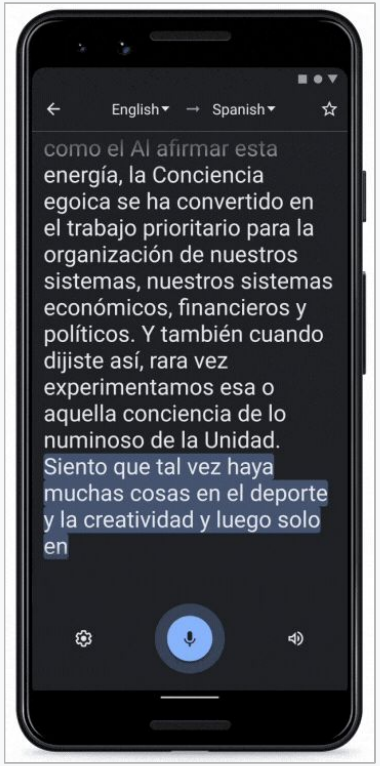
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent quis dolor lacus. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. In eu mi placerat, facilisis tellus vitae, efficitur nisi. Nulla placerat placerat sem, tempor vulputate libero suscipit sed. Mauris sit amet massa eu justo dignissim pharetra. Praesent sapien tortor, ornare et leo nec, aliquet suscipit nisi. Aenean egestas mauris eget hendrerit finibus. In eleifend ex pharetra tellus dignissim.

Google Research

Results

Launched support for 10 languages.

Launched streaming TTS support for Pixel Buds.



Google Research

++

What's next?

Google Research

Advancing Speech Translation

Long-form Audio Input

Streaming Translation

Streaming Text-to-Speech

Evaluation

Google Research

Thank You

Jeff Pitman
Senior Staff Engineering Manager

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

Google Research