Costly Speech: an introduction

Who am I

- Mathieu Duponchelle
- Working at Centricular for a few years now
- Active on several GStreamer components

Costly Speech: an introduction

- Shout-out Ray Tiley (cablecast)
- In which the dog learns a new trick

Demo !

- Live spanish conversation

Transcription

- A transcription element:
  - Takes audio speech as input
  - Outputs text, each word a different buffer
  - Operates live with a latency

AWS backend

- Mistakes were made
- `awstranscriber2`
- `awstranslate`

AWS backend

- Solid word accuracy
- Explicit result stability
- Limited speaker diarization

Speechmatics backend

- Solid word accuracy
- Can be run on premise
- Supports translations with no extra roundtrip
- max-delay argument
- Best speaker diarization

Deepgram backend

- Noise-sensitive word accuracy
- No explicit mechanism for word stabilization
- Can go very low-latency
- Good speaker diarization

Recent transcriber refactoring

- Lessons learned
- The B word
- Interchangeability

textaccumulate

- Complete sentence: `["I", "love", "you", ".", "And"]` -> `["I", "love", "you", "."]`
- On deadline:
  - Clause: `["I", "love", "you", ",", "And"]` -> `["I", "love", "you", ","]`
  - Otherwise: `["I", "love", "you", "and"]` -> ["I", "love", "you", "and"]

awstranslate

- Span tokenization

Speech synthesizers

- New class of elements

awspolly

- Large-ish library of voices
- max-duration
- scaletempo shenanigans

elevenlabssynthesizer

- Very large library of voices, clonability
- "realtime" vs. POST API
- `overflow=compress`
- `max-overflow`

elevenlabsvoicecloner

- Instant Voice Cloning
- `cloner ! transcriber ! .. ! synthesizer`
- `speaker=`

Future development

- Free backend
- Voice fingerprints
- Conversational elements
- Paraphrasing
- Very large spectrum of applications opening up

Questions