NATHAN DONG PRESENTS

SPEAKER ADAPTATION

USING CASCADE PIPELINES FOR SPEECH TO SPEECH TRANSLATION

PROJECT DESCRIPTION

GOAL: IMPLEMENT AND EVALUATE A SPEAKER-ADAPTED SPEECH GENERATION PIPELINE THAT PRODUCES SPANISH SPEECH WHILE MATCHING A TARGET SPEAKER'S VOICE.

QUESTION: DOES FULL FINE-TUNING OF SPEECHT5 ON A TARGET SPEAKER IMPROVE SPEAKER SIMILARITY OVER ZERO-SHOT CONDITIONING, WITHOUT SIGNIFICANTLY HARMING INTELLIGIBILITY OF THE SPANISH OUTPUT?

HYPOTHESIS: FINE TUNING SPEECHT5 WILL HAVE HIGHER SPEAKER FEDELITY AND TRANSLATION ACCURACY

INSPIRATION FOR MY PROJECT

METHODOLOGY

Approach One: Zero Shot

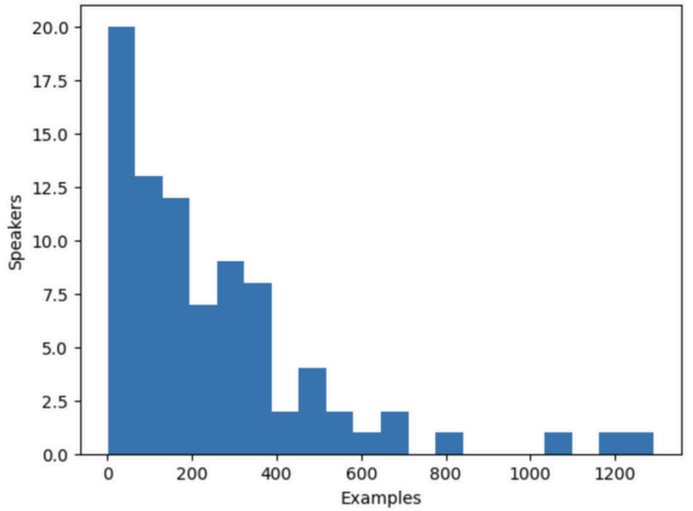
- Source english audio is transcribed using whisper.
- Embeddings are pulled from the reference speaker.
- Google T5 translates en text to es text
- SpeechT5 from Microsoft generates synthetic spanish audo using the translated text and reference speaker embeddings

Approach Two: Fine Tune

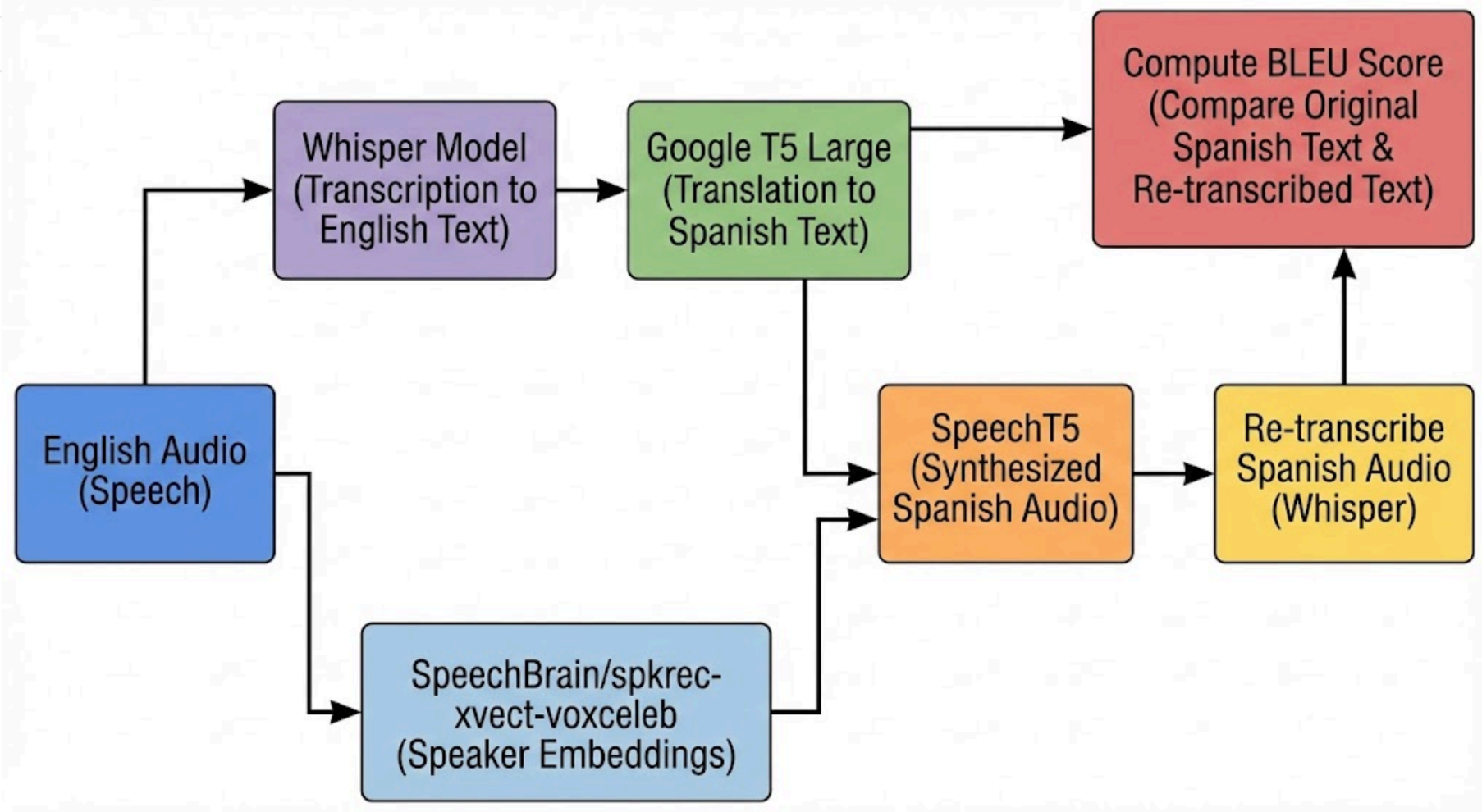
- Similar pipeline to the zero-shot except the SpeechT5 model is fine tuned on multiple clips of the reference speaker

Datasets Used

- **facebook**/**voxpopuli** dataset from hugging face
  - Speaker id's on the audio segments
  - Transcripts provided
- **Helsinki-NLP**/**opus-100** dataset from hugging face
  - en-es text pairs



SYSTEM


English Audio (Speech)
Whisper Model (Transcription to English Text)
Google T5 Large (Translation to Spanish Text)
Compute BLEU Score (Compare Original Spanish Text & Re-transcribed Text)
SpeechBrain/spkrec-xvect-voxceleb (Speaker Embeddings)
SpeechT5 (Synthesized Spanish Audio)
Re-transcribe Spanish Audio (Whisper)

EVALUATION METHOD

Automatic Metrics

- Synthetic spanish audio is back transcribed to text and then compared to reference spanish audio
- BLEU Score
- Cosine Similarity → Measures angle between to vectors (speaker embeddings)
    - How similar are the speaker embeddings?

Human Evaluation

- Five English/Spanish speakers
- Each reviewed three speakers
    - Listened to a 4-second clip of reference audio
    - Then, the target Spanish text is shown.
    - Evaluated zero-shot and fine-tuned synthesized Spanish speech
    - Five sentences for each of the three reference speakers

QUANTITATIVE RESULTS

SpeakerID	ZeroShot BLEU	FineTuned BLEU	ZeroShot Cosine	FineTunedCosine
1055	0.4736	**0.4895**	0.9502	**0.9508**
28165	**0.6297**	0.5141	0.9100	**0.9510**
124992	**0.2442**	0.2111	0.8866	**0.9510**

NOTES

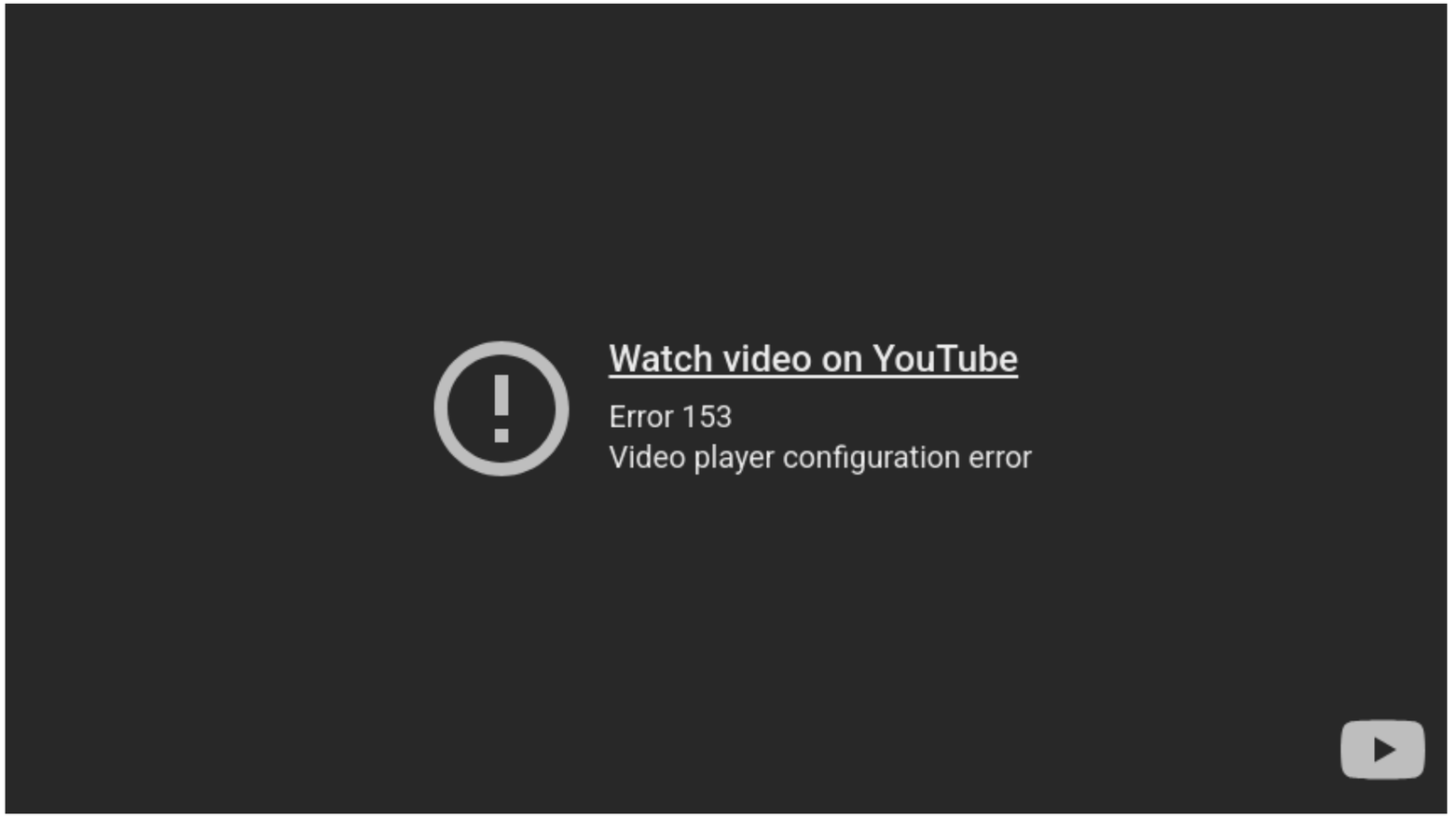
- **Is the model overfitting?**
  - Higher cosine similarity may mean the model is focusing on speaker fidelity rather than quality output.
- **We need to account for the Whisper model back-transcribing synthesized Spanish speech audio.**
  - Does BLEU capture clarity of audio as well?

HUMAN EVALUATION

speaker_id	Fine Tune Acc	Zero Shot Acc	Fine Tune Speaker Match	Zero Shot Speaker Match	preferred_method
1055	**2.77**	2.64	**3.09**	2.86	fine_tune
28165	**3.13**	2.94	**3.19**	2.44	fine_tune
124992	2.95	2.95	**3.25**	3.00	fine_tune

*(*NOTE THAT ACCURACY AND MATCH ARE BOTH REPORTED BY TRANSLATOR)*

EXAMPLES


Watch video on YouTube
Error 153
Video player configuration error

CONCLUSION

- FINE TUNING S2S TRANSLATION MODELS **DOES INCREASE** SPEAKER FIDELITY
- FOR ZERO-SHOT PROMPTING, SEGMENTS OF ABOUT **4-6 SECONDS** PRODUCE BEST BLEU SCORES
- ASR NOISE WAS A BOTTLENECK FOR MY EXPERIMENT. FUTURE WORK MAY INVOLVE **TARGET SPANISH AUDIO** RATHER THAN BACK-TRANSCRIBING TO TEXT

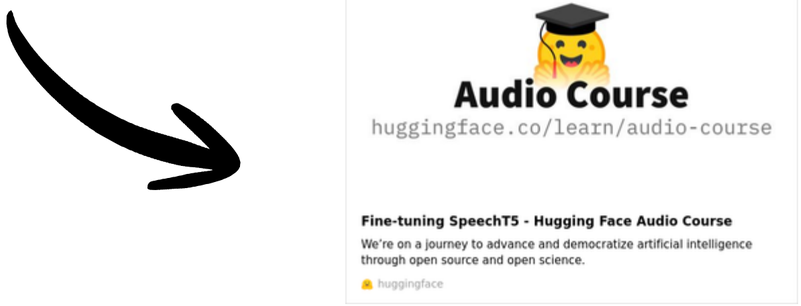
SPIRITUAL INSIGHTS

“Listen to the still small voice”

My original project was focused on the optimal duration of reference speech embeddings for zero-shot prompting.

Half way through my project I had a feeling one day to do some more research on fine tuning SpeechT5.

I came across this website that changed my whole focus for the project. I am so glad I listened to that prompting.



THANK *You!*