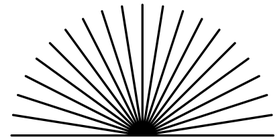
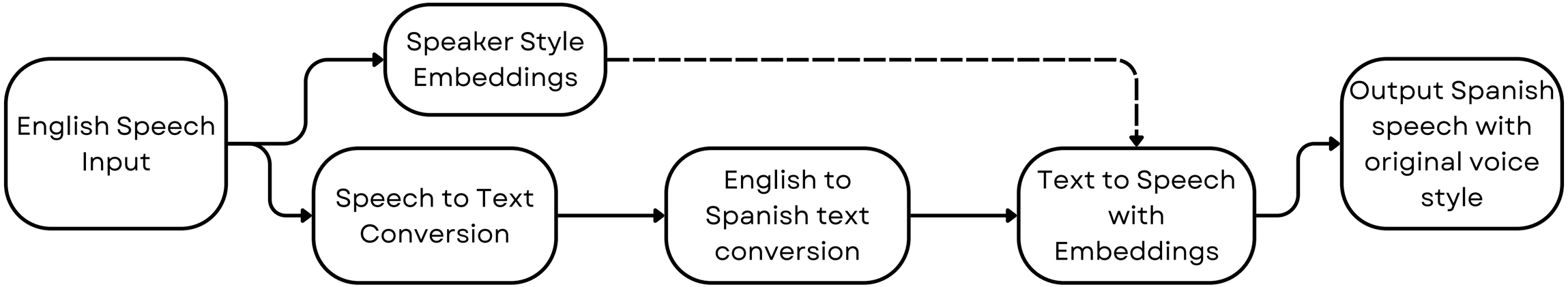
STYLE-PRESERVING ENGLISH-TO-SPANISH SPEECH TRANSLATION USING SPEAKER EMBEDDINGS

NATHAN DONG



Topic / Thesis

Can a speaker’s voice style be carried across languages by feeding an input embedding into a speech to text to speech translation pipeline?



Background Papers

1) Translatotron 2: High-quality direct speech-to-speech translation with voice preservation

- https://arxiv.org/abs/2107.08661
- Researchers at Google develop a *direct* speech-to-speech translation model that preserves style.

2) ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification

- https://arxiv.org/abs/2005.07143
- Instead of using a trained NN to extract speaker representations, researchers used a new model called TDNN (Time Delay Neural Network) to extract embeddings.
- Shows state-of-the-art capabilities of extracting speech embeddings to capture style.

3) YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone

- https://arxiv.org/abs/2112.02418
- This paper shows that high quality speech output from a MT model can be produced with speech data less than 1 minute long for style transfer.

Dataset & Pipeline

Datasets

multilingual_librispeech

- Open source and public
- Accessible using Hugging Face
- Multiple distinct voices which will be valuable for evaluating style transfer

MuST-C

- English Speech
- English Transcript
- Spanish Transcript
- All with distinct voices across pipeline

Code Pipeline

Speech to Text

- Whisper model from OpenAI
- Inputs wav data and converts it into txt
- Could also use speech to text implementation from class

Speach Embeddings

- Some form of wrapper for speech embedding models
- TDNN from the research paper or wav2vec

Text Translation English → Spanish

- Helper class to convert English text to Spanish text
- Marian
- mBART
- Some other pretrained model

Text to Speech with Speach Embeddings

- Using Microsoft **SpeechT5** model, pass in spanish text and source speech embeddings to output audio
- May attempt to fine tune TTS model

Evaluation

Translation Quality

- Using either of the two open source speech datasets, I will split the dataset into a test and a train set
- Because we are working with speech output, I'll integrate a speech-to-text pipeline on the workflow for the output Spanish
- Both datasets provide reference translations. Using those I can compute the BLEU and COMET scores on my backprocessed speech output.

Translation Quality

- Using five people, we will have them listen to 20 sample sources and corresponding output examples of speech. They will fill out the ruberic.

Style Evaluation (5 People)

Example	Translation Quality (1-10)	Language Fluency (1-10)	Voice Similarity (1-10)
sample 1			
...			
Sample 20			

Fine Tuning A Model

- Depending on how long my project takes. I may add a the capability to produce embeddings from the output spanish audio and do a cosine similarity comparison from the source speech to the output speech embeddings. This could be used as a loss function for fine tuning the text to speech portion!

Serving Gods Children

1) In a general sense, being able to translate between languages effectively helps us spread the gospel to the four corners of the earth. While translation in most medium to high-resource languages has been achieved, being able to preserve the style and tonality of the original speaker can increase the quality of the translation in those languages.

2) Speech-to-speech models have a particular potential to positively impact the world in the sense that they increase our ability to communicate openly. Conversations typically do not occur over text, but rather face-to-face. Anything we can do to advance this area of development further improves our ability to connect with others.

3) Lastly, and most importantly, I think that being able to translate speech with some degree of voice cloning can have an immense impact on areas of the church with relatively low membership levels. In my mission, there were many wards where there weren't enough Spanish members to form a branch, so they had to attend the English ward. Having low-latency, high-quality speech-to-speech models could help those minority language members connect with their existing congregations.