Cross Lingual Speaker Adaptation for TTS Applications

Software Project-2021/2022

Rasul, Anna, Claesia, Sharmila

Outline

- Big recap
  - Inspiration & models
  - Problems we solved
- Evaluation
  - Subjective
  - Objective
- What’s next?

Big Recap: Inspiration & Models

Inspiration:

GradTTS for TTS and expressivity transformation between speaker.

Our Contribution:

Single model setup and evaluation for two collective tasks:

1. **Multilingual TTS**: Two languages i.e. english and french
2. Projecting the work done in expressivity transformation into **speaker voice transformation across language**

Deep Learning model pipelines, extensive evaluation on the newly proposed task.

Big Recap: Problems we solved

- phonemes in English and French where some overlapping some are not
- changing the available resources according to our nead i.e. modifying models formation or architecture
- formulating 4 different variations to analyse the effects of different embedding layers for speakers and language transformation and how these embeddings represents speakers voice and language

Last Presentation: Where we were

Done:

- 4 models
- Inference (414*12*4 samples)
- Site

Left to do:

- Evaluation
- Put site online & improve design
- Write report

Evaluation

We have to evaluate the **quality of speech produced** based on text and **proximity to the target speaker's voice** in the audio.

1. Objective
2. Subjective

Evaluation: Subjective

Mean Opinion Score – through the site that will be distributed to evaluators.

Native speakers (EN and FR) will rate the speech samples from 1 to 5 in terms of their oral quality.

12 speakers, 4 models, 2 samples per speaker = 96 in total

Note: We did an intermediate subjective evaluation for French audio (6*4 samples) to get the brief outlook. Result: score of 3 for **model 1, 3 and 4**. But for **model 2** it's 2 only.

Evaluation: Objective

Idea: measure interpretability. Metrics:

- Mel cepstral distortion
- Word & character error rate
- Cosine similarity (speaker embeddings)

Mel Cepstral Distortion

Measures how different two sequences of mel cepstra are. Assesses the quality of speech synthesis: the smaller the MCD between synthesized and natural mel cepstral sequences, the closer the synthetic speech is to reproducing natural speech.

$$\frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t}\sqrt{\sum_{i}^{25}2||mc(t,i)-mc_{synth}(t,i)||^2}$$

Word & Character Error Rate

- Use NEMO nvidia model to receive transcriptions
  - QuartzNet15x5Base-En for EN
  - Stt_fr_quartznet15x5 for FR
- Compute WER and CER for synthesized & original audios
- Compare the scores

$$WER = \frac{S + I + D}{N}$$

S: # substitutions
I: # insertions
D: # deletions
N: # words in the reference

Word & Character Error Rate: Results

References	WER	CER
overall	11.6316	2.8193
en part	**8.0101**	2.0449
fr part	**19.8805**	4.5833

Version2	WER	CER
overall	57.9951	36.6522
en part	**29.8603**	13.1075
fr part	**123.8212**	91.7393
en orig + en voice	29.3885	12.88
en orig + fr voice	30.3321	13.335
fr orig + en voice	124.0779	91.6201
fr orig + fr voice	123.5645	91.8586

Version3	WER	CER
overall	72.1965	42.2821
en part	**61.2242**	33.081
fr part	**97.1889**	63.2402
en orig + en voice	57.7705	30.5631
en orig + fr voice	64.678	35.5988
fr orig + en voice	99.4427	66.2038
fr orig + fr voice	94.9351	60.2767

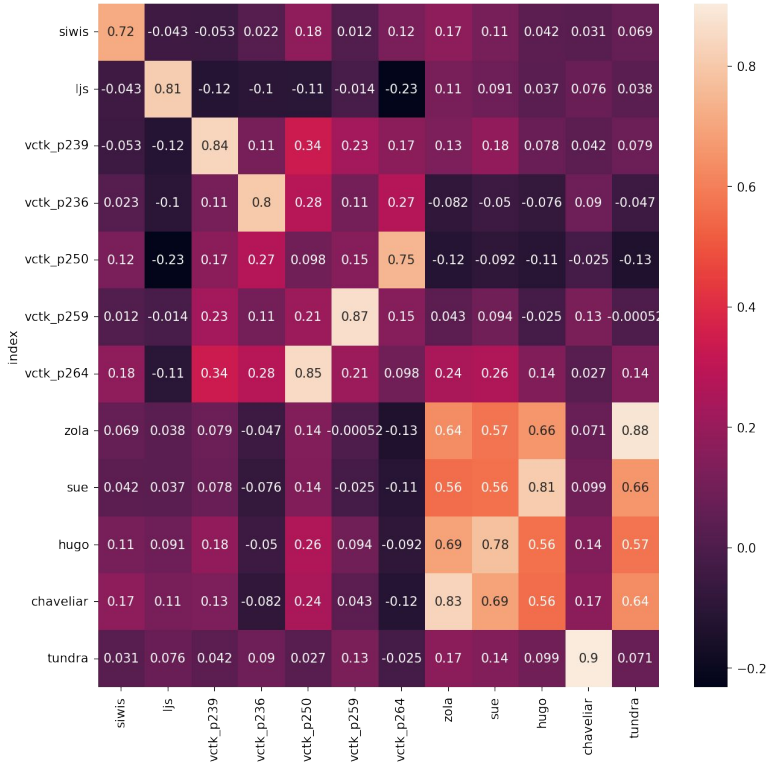
Cosine Similarity (Speaker)

Idea: see how well the voice is transferred from one language to another.

- Extract embeddings for the synthesized & original audios
  - Use ECAPA-TDNN model that is state of art for speaker variation
  - The number of nodes in the final fully-connected layer is 192 (embedding of speaker)
  - Pretrained model, result for 12 speakers in next slide
  - Finetuned the model the for 12 speakers (for better result)

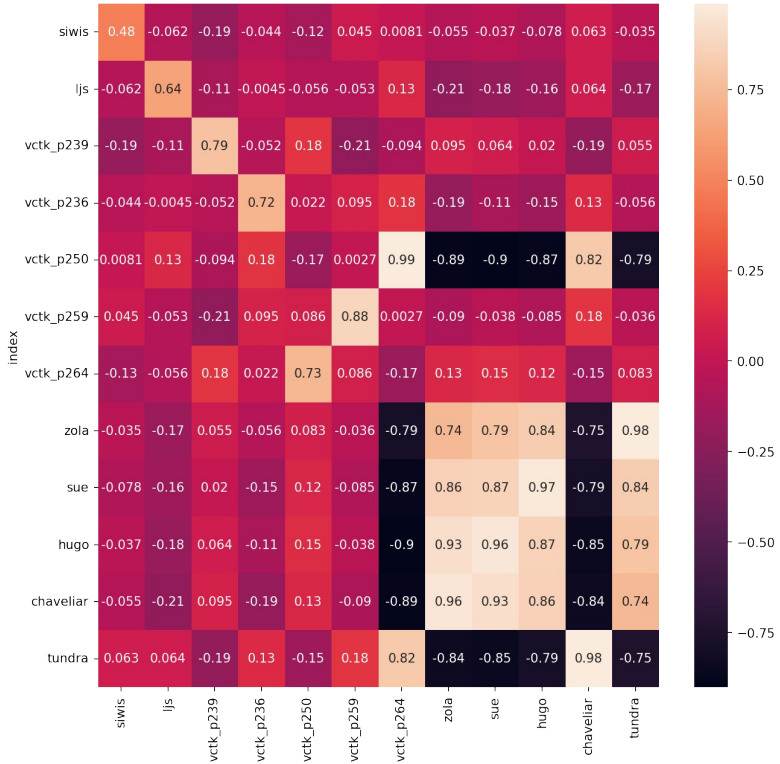
- Compute cosine score between the two audio to get the similarity between speakers voice

Cosine Similarity



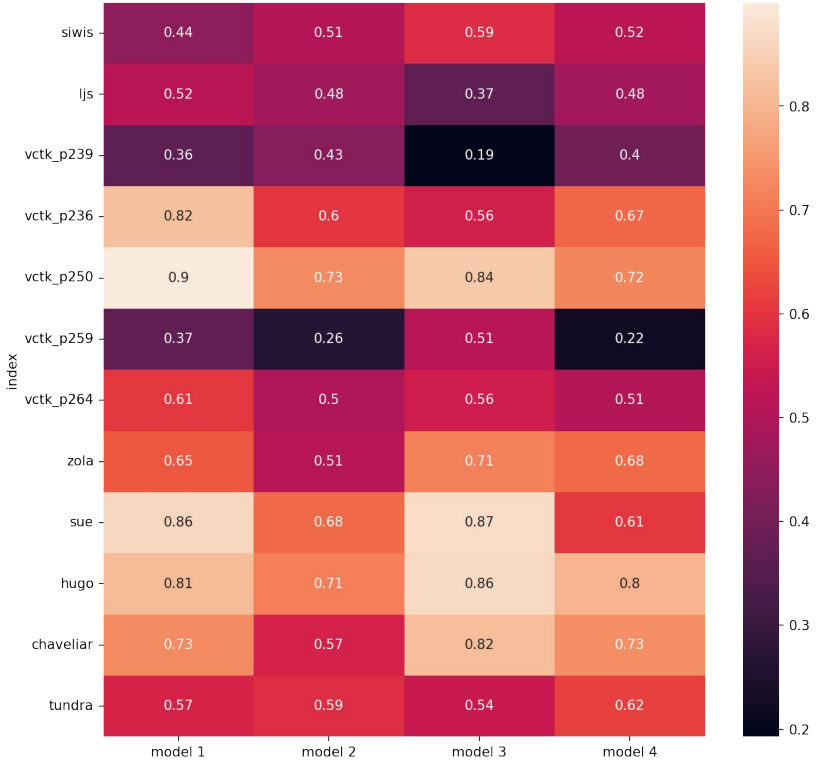
- **Pretrained ECAPA-TDNN** model embedding extracted and cosine similarity calculated between speakers.
- Speakers from same dataset : synpaflex is high for different speakers. **“hugo:sue:zola”**
- Hence, decided to finetune the model to learn the clear distinction of speaker

Cosine Similarity



- **Fine Tuned ECAPA-TDNN** the model for speaker verification
- However, the similarity is much more worse for the hugo, sue and zola.
- Needs to be analysed and trained more

Cosine Similarity



Calculated the mean similarity for 4 (**small sample**) audios for all speaker

Result are subject to the performance of finetuned model. Hence, the given result is not the final.

What's next?

1) Finish evaluation

Website for subjective evaluation, collect the answers from speakers.

Compute WER for models 1 and 4, and MCD for others.

2) Report

Finish with the report. Wrap up and conclusion from the project.

Thank you! Any questions?