Cross Lingual Speaker Adaptation for TTS Applications

Software Project-2021/2022

Rasul, Anna, Claésia, Sharmila

Outline

- Recap
  - Problems Faced
- New Addition
  - Speaker Reduction
  - Training
  - Objective Evaluation (WER)
  - Objective Evaluation (Speaker Similarity)
  - Subjective Evaluation
- What’s Left

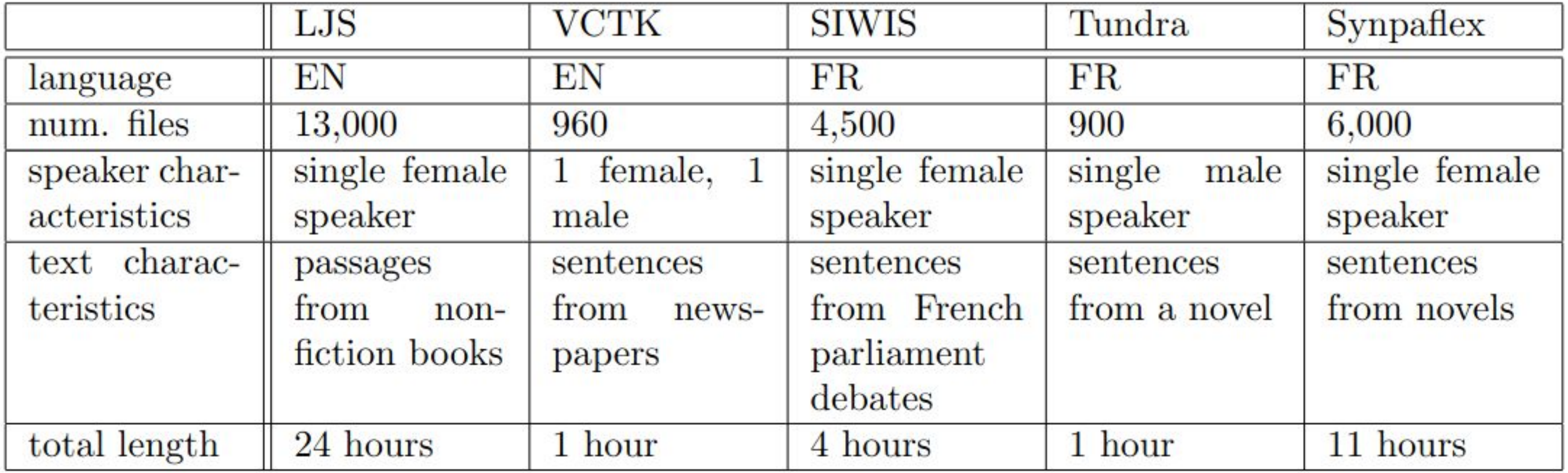
Recap: Problems

- Trouble with speaker recognition model: it was not distinguishing speakers properly and similarity score for different speakers were same.
- WER/CER rate high for some models.

New Addition: Speaker Reduction

- After analysing cosine result, we tried listening and analysing.
- Realised that we could not so we referred to the paper
- Then we went through the paper of synpaflex and realised that it was actually grouped according to the content rather than speaker.
- Hence, we decided that we should now combine these groups and remove some of the english speaker for data balancing.

New Addition: New Dataset



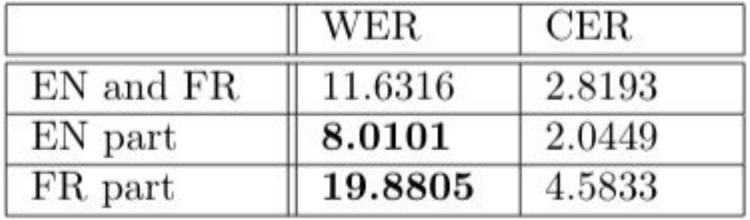
	LJS	VCTK	SIWIS	Tundra	Synpaflex
language	EN	EN	FR	FR	FR
num. files	13,000	960	4,500	900	6,000
speaker characteristics	single female speaker	1 female, 1 male	single female speaker	single male speaker	single female speaker
text characteristics	passages from non-fiction books	sentences from newspapers	sentences from French parliament debates	sentences from a novel	sentences from novels
total length	24 hours	1 hour	4 hours	1 hour	11 hours

Table 2: Main characteristics of used data.

New Addition: New Dataset (Analysis)

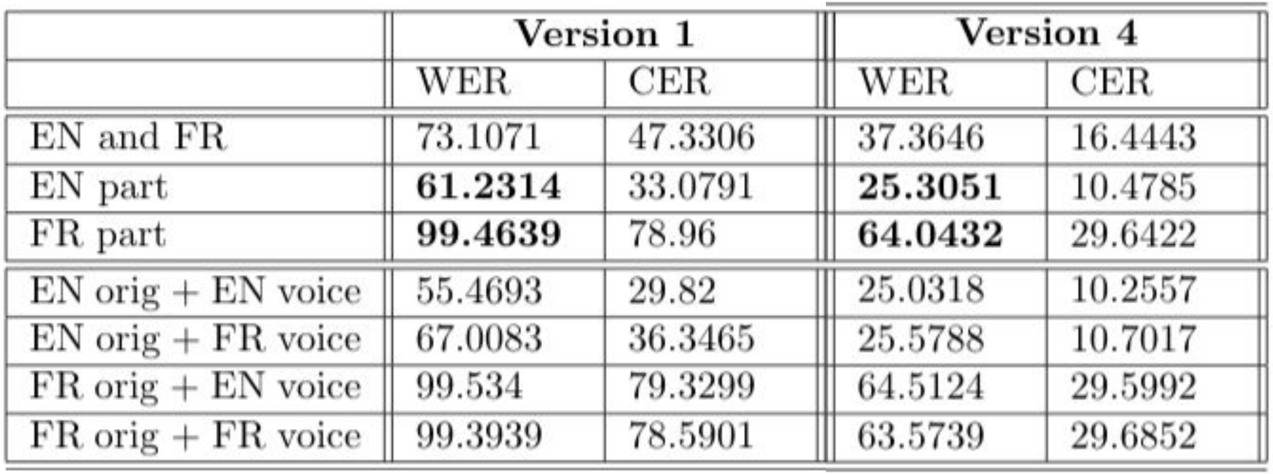
- Data imbalance: French total length = 16 hours while english 25-26 hours
- Phoneme distribution: Almost balanced for english and french dataset except out of 49206 pairs of bigrams and trigrams phonemes of test set ~30 are not present in the training data.
- Test, train and validation: 500 samples for test, 100 for validation and remaining for training

WER results: standard



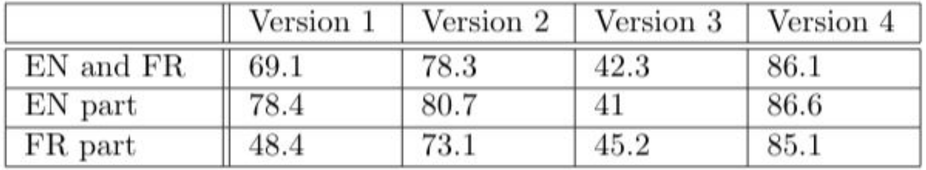
	WER	CER
EN and FR	11.6316	2.8193
EN part	**8.0101**	2.0449
FR part	**19.8805**	4.5833

Word and character error rate for reference audio files



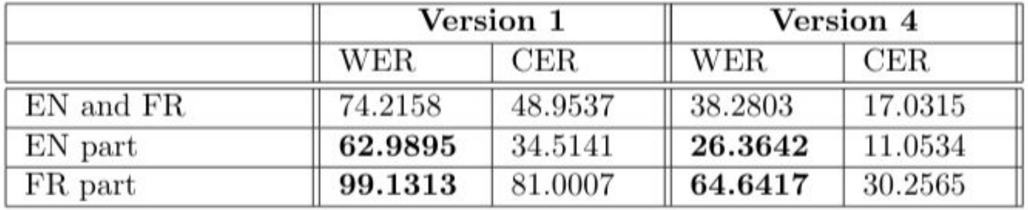
	Version 1		Version 4	
	WER	CER	WER	CER
EN and FR	73.1071	47.3306	37.3646	16.4443
EN part	**61.2314**	33.0791	**25.3051**	10.4785
FR part	**99.4639**	78.96	**64.0432**	29.6422
EN orig + EN voice	55.4693	29.82	25.0318	10.2557
EN orig + FR voice	67.0083	36.3465	25.5788	10.7017
FR orig + EN voice	99.534	79.3299	64.5124	29.5992
FR orig + FR voice	99.3939	78.5901	63.5739	29.6852

WER results: experiment with lengths



	Version 1	Version 2	Version 3	Version 4
EN and FR	69.1	78.3	42.3	86.1
EN part	78.4	80.7	41	86.6
FR part	48.4	73.1	45.2	85.1

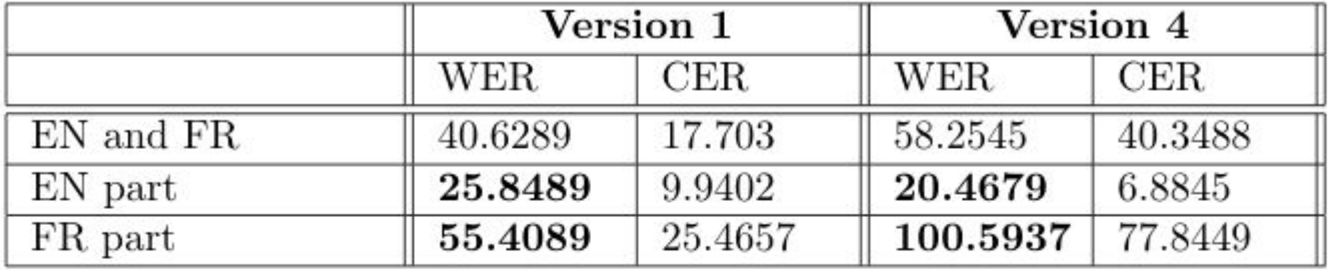
Table 7: Length differences between generated and reference audios (per cent)



	Version 1		**Version 4**	
	WER	CER	WER	CER
EN and FR	74.2158	48.9537	38.2803	17.0315
EN part	**62.9895**	34.5141	**26.3642**	11.0534
FR part	**99.1313**	81.0007	**64.6417**	30.2565

Table 8: Word and character error rates on slowed-down audios for models 1 and 4

WER results: experiment with good data



	Version 1		Version 4	
	WER	CER	WER	CER
EN and FR	40.6289	17.703	58.2545	40.3488
EN part	**25.8489**	9.9402	**20.4679**	6.8845
FR part	**55.4089**	25.4657	**100.5937**	77.8449

Table 9: Word and character error rates for models 1 and 4 trained on LJS and SIWIS

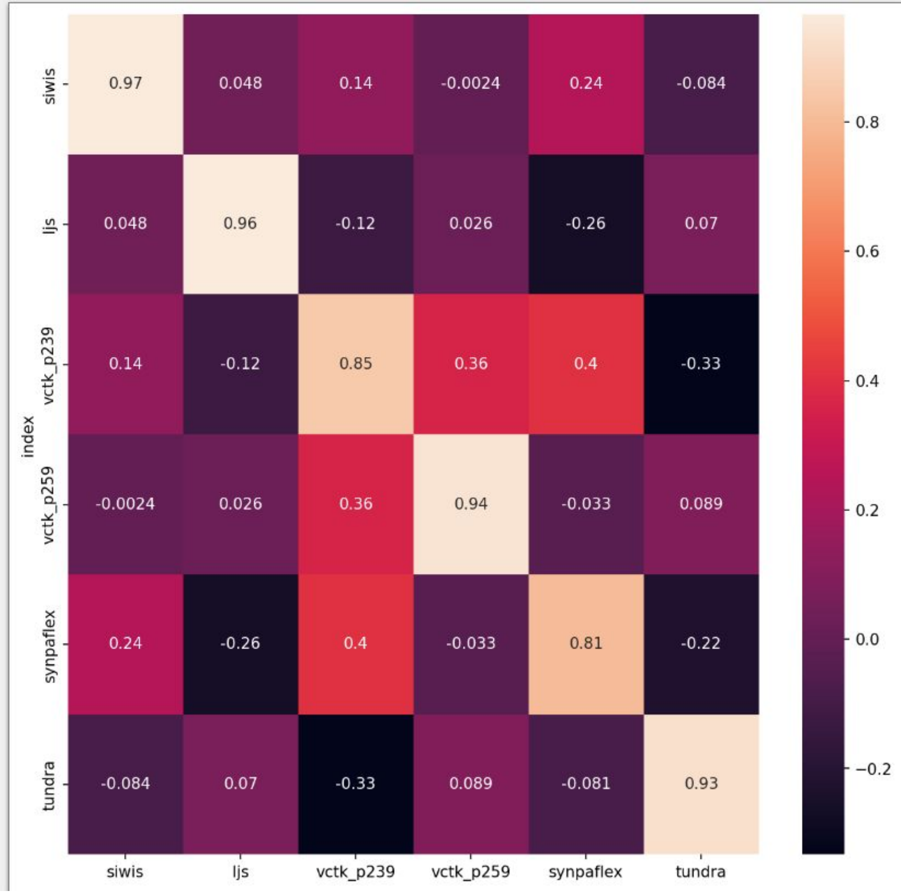
Speaker Similarity results

- ECAPA-TDNN Finetuning summary
- Reference Speaker vs Generated

  Speaker

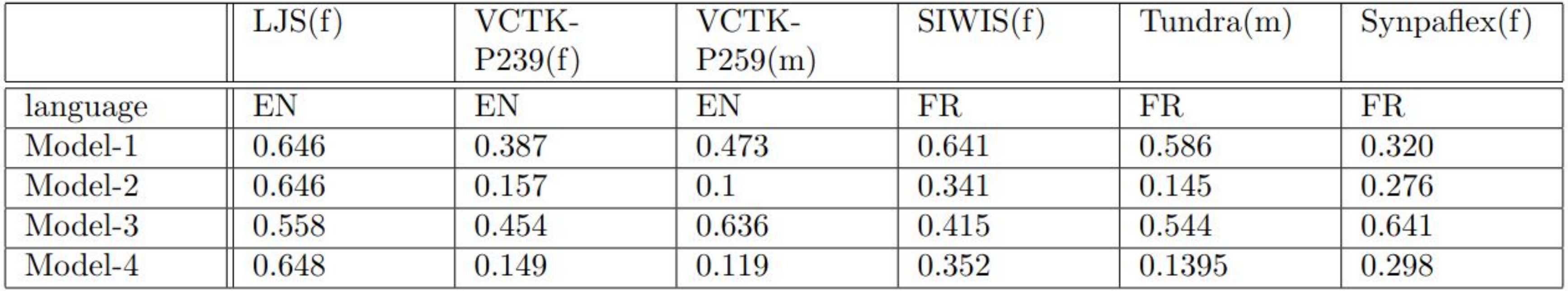
Synpaflex is the combination of audios from same speaker but for different content and it shows slightly less similarity

(Observation): Do they sound different? may be noise in some data? recording environment different?



	siwis	ljs	vctk_p239	vctk_p259	synpaflex	tundra
siwis	0.97	0.048	0.14	-0.0024	0.24	-0.084
ljs	0.048	0.96	-0.12	0.026	-0.26	0.07
vctk_p239	0.14	-0.12	0.85	0.36	0.4	-0.33
vctk_p259	-0.0024	0.026	0.36	0.94	-0.033	0.089
synpaflex	0.24	-0.26	0.4	-0.033	0.81	-0.22
tundra	-0.084	0.07	-0.33	0.089	-0.081	0.93

Cosine similarity results

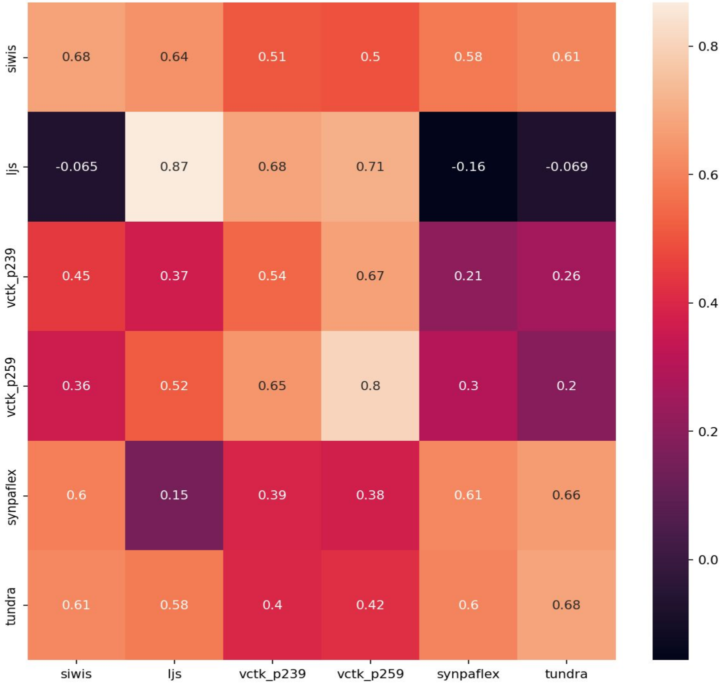


	LJS(f)	VCTK-P239(f)	VCTK-P259(m)	SIWIS(f)	Tundra(m)	Synpaflex(f)
language	EN	EN	EN	FR	FR	FR
Model-1	0.646	0.387	0.473	0.641	0.586	0.320
Model-2	0.646	0.157	0.1	0.341	0.145	0.276
Model-3	0.558	0.454	0.636	0.415	0.544	0.641
Model-4	0.648	0.149	0.119	0.352	0.1395	0.298

Best: model version 3 (0.54), but looks like there was some error in the training as there were only mumbling sounds. Hence, model training was not proper.
Retraining this version to see if it is due to issue with architecture or something else.
Otherwise model 1 has around 0.51

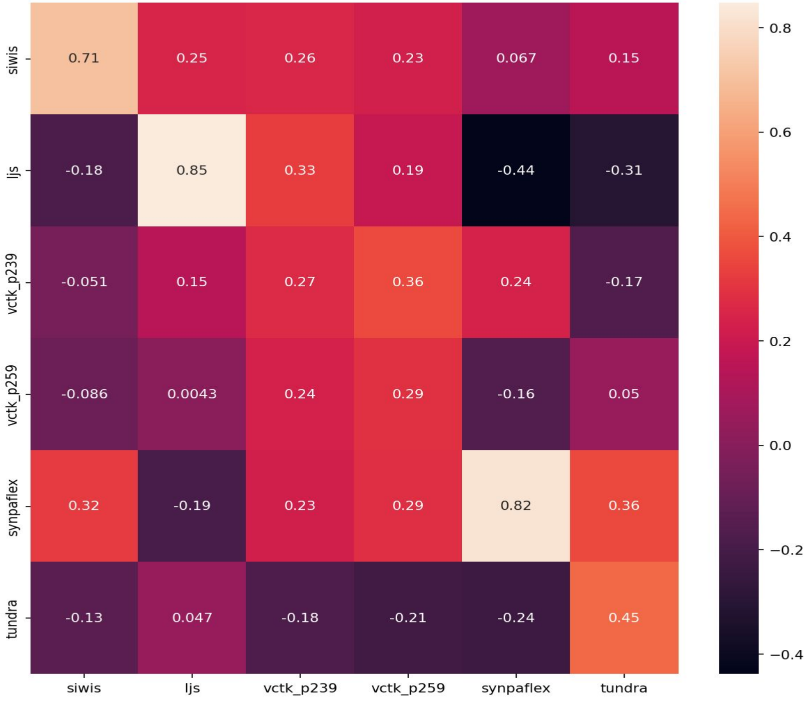
Cosine similarity Breakdown

Version 1



	siwis	ljs	vctk_p239	vctk_p259	synpaflex	tundra
siwis	0.68	0.64	0.51	0.5	0.58	0.61
ljs	-0.065	0.87	0.68	0.71	-0.16	-0.069
vctk_p239	0.45	0.37	0.54	0.67	0.21	0.26
vctk_p259	0.36	0.52	0.65	0.8	0.3	0.2
synpaflex	0.6	0.15	0.39	0.38	0.61	0.66
tundra	0.61	0.58	0.4	0.42	0.6	0.68

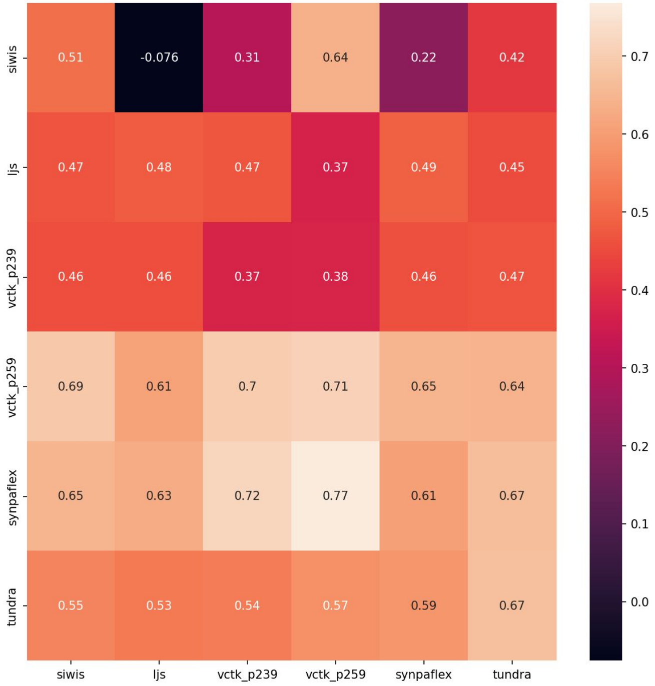
Version 2



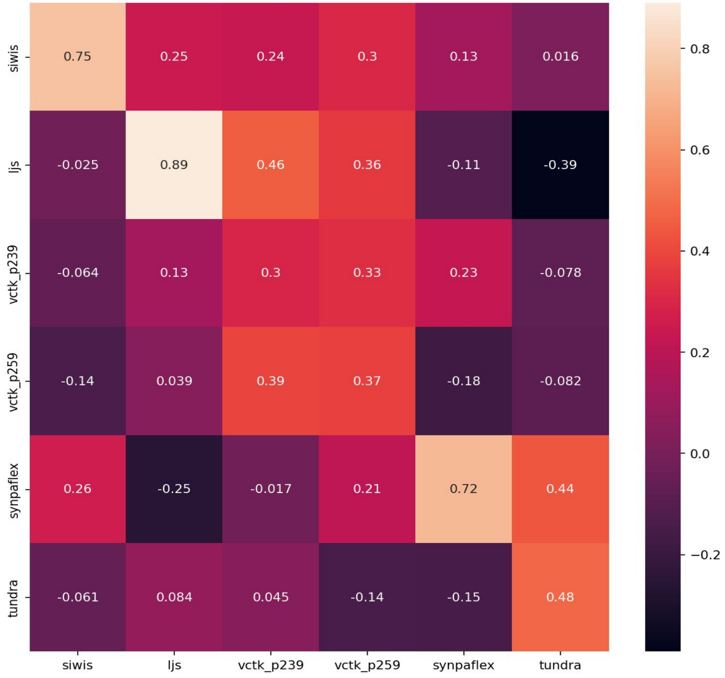
	siwis	ljs	vctk_p239	vctk_p259	synpaflex	tundra
siwis	0.71	0.25	0.26	0.23	0.067	0.15
ljs	-0.18	0.85	0.33	0.19	-0.44	-0.31
vctk_p239	-0.051	0.15	0.27	0.36	0.24	-0.17
vctk_p259	-0.086	0.0043	0.24	0.29	-0.16	0.05
synpaflex	0.32	-0.19	0.23	0.29	0.82	0.36
tundra	-0.13	0.047	-0.18	-0.21	-0.24	0.45

Cosine similarity Breakdown

Version 3



Version 4



WER and Speaker Similarity Tradeoff

- Model version 4 better based on WER
- Model version 1 better based on Speaker Similarity
- Tradeoff between them
- Decided to do few more experiments
- First: with 2 speaker, where hypothesis is both language and speaker representation learning from 2 variation of data does not work
- Result: WER [slide 9] was far more better but speaker similarity was worst being around **0.2**

New Experiment: Merging Phonemes

- Initially, separate tokens for English and French phonemes
- Redundant - 2 parallel TTS systems rather than one generalized system
- English and French have many similarities and several differences:
  - Similar consonants (aspiration!)
  - Nasal vowels
  - Lexical stress
- Idea: Same representation for similar sounds but keep distinct representations for unique sounds

ARPAbet

- Initial models used letters for English and phonemes for French (SAMPA)
- To merge phonemes, we need to have comparable input
- ARPAbet - analogous to IPA for English (ASCII)
- Reduced the overall number of input IDs by about 100

New Experiment

- 4 speakers(ljs, siwis, vctk and synpaflex)
- Version 4 and 1 training on the text represented as combined phonemes.
- English as Arpabet
- 3 set of experiments currently running
- To this date 1 complete (combined phonemes and english as characters)

Result WER   Result Cosine Similarity

Subjective Evaluation

- Due to time constraints, we will be only performing evaluation on Model version 1(speaker similarity), Version 4 (generation quality based on wer).
- Evaluation 1: audio interpretability for french and english audio (each 18 samples)
- Evaluation 2: speaker voice similarity (4 samples for each speakers, 6*4 samples)

Link: https://grad-tts.herokuapp.com/

What's left

- Finalizing the report by adding current running experiments
- Adding results of Subjective evaluation
- Proofreading and wrapping the report
- Hosting the website
- Selecting models that should be shown in the web application