Cross Lingual Speaker Adaptation for TTS Applications

Software Project-2021/2022

Final presentation

Rasul, Anna, Claésia, Sharmila

Outline

1. Motivation & Task Formulation
2. Model
3. Data
4. Experiments
5. Evaluation
   a. Objective
   b. Subjective
6. Demo
7. Conclusion

Motivation & Task Formulation

Motivation

- TTS - making machines talk
  - Screenreaders
  - Automatic phone operators
- Longterm focus on intelligibility
- Deep-learning based gains in naturalness -> new problems
- Expressive, humanistic TTS
  - + Giving people their voice back
  - + Dubbing movies in the actors voice
  - - - Increased risk for fraud/defamation

Task Formulation

- Transfer the voice of a monolingual speaker into a new language
- Primary Objective: Intelligibility
  - Word Error Rate, Character Error Rate
- Secondary Objective: Naturalness
  - Speaker similarity
- Tertiary Considerations:
  - Reusability (Low resource language - Louisiana Creole)
  - Resource use
  - Consistency/reliability
  - Safety: Is it clear this is not a real person?

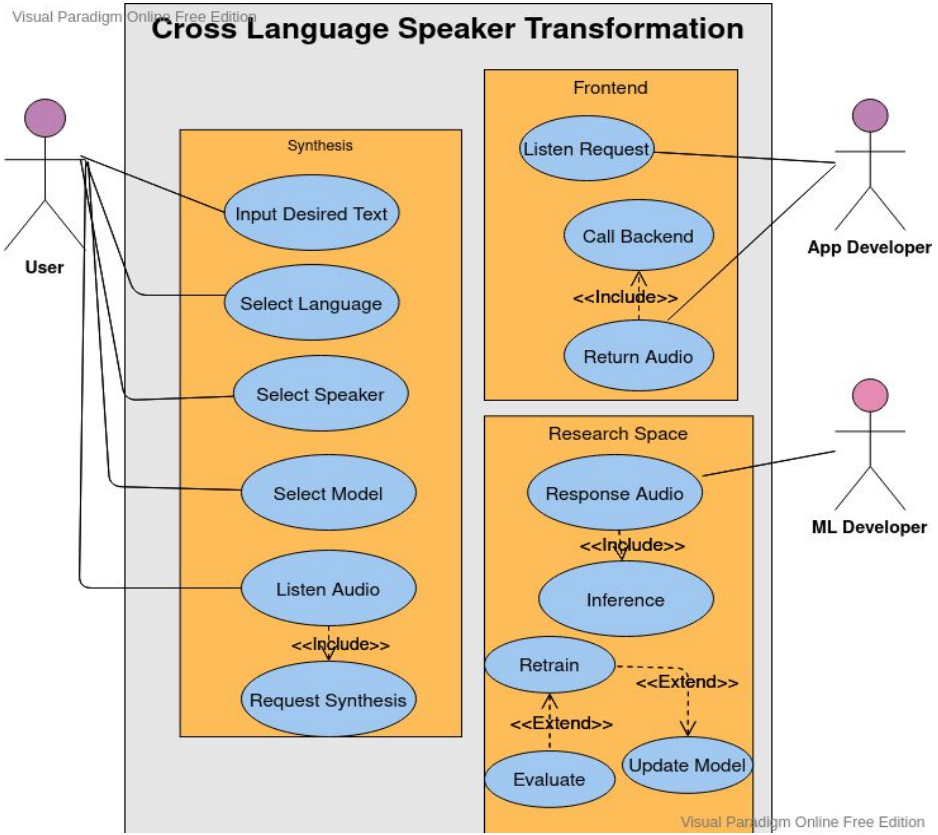
Software (Usecase Diagram)

- Actors: User, Developers
- SubSystem: Application, TTS Pipeline

Technical Details:

Flask, Jquery, Python, Pytorch

Software Development Cycle: **Iterative**



Model, Data, Experiments

Model

Model: based on Grad-TTS.

What we've changed:

- different input data
- multi-gpu support
- training from checkpoints

	language representation	speaker representation
Version 1	id	id
Version 2	id	MEL features → embedding network
Version 3	MEL features → embedding network	id
Version 4	MEL features → embedding network	MEL features → embedding network

Table 1: Model inputs

Data

	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

- Data imbalance (length, gender)
- Different quality of datasets

Experiments

1. Four model versions, standard data
   Results not perfect; generated are shorter
2. Slowed down audios
   No improvement; maybe the quality of data?
3. Two best datasets
   No improvement; different sets of parameters for each language?
4. Merging phonemes
   a. Only consonants
   b. Both consonants and vowels
5. Louisiana Creole
   Zero-shot synthesis

Evaluation

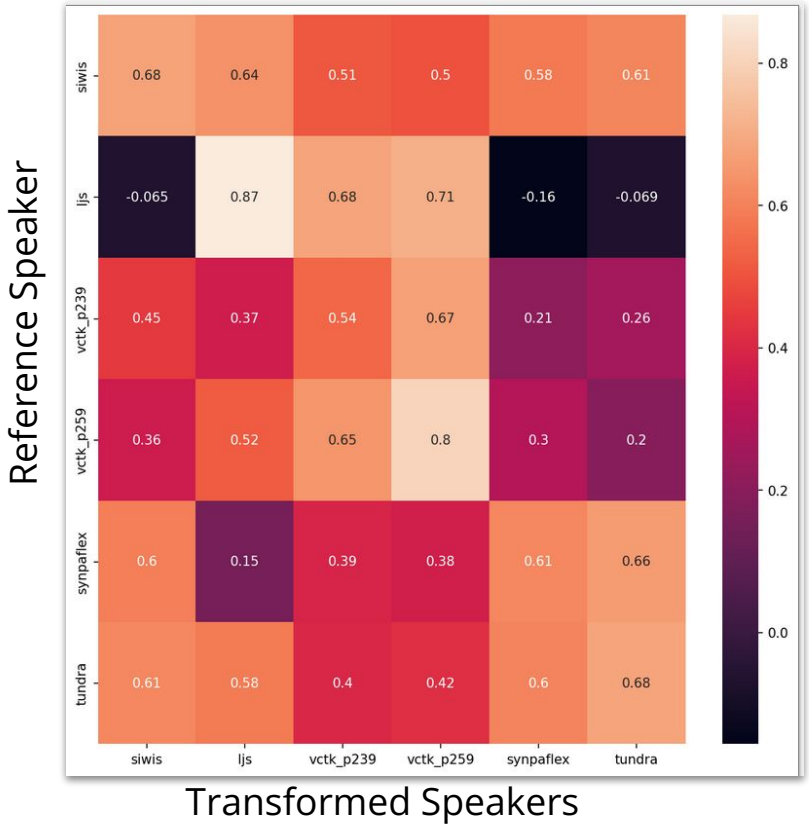
Objective Evaluation: Cosine Similarity

Speaker Embedding- ECPA TDNN

	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

Table 6: Speaker Cosine Similarity

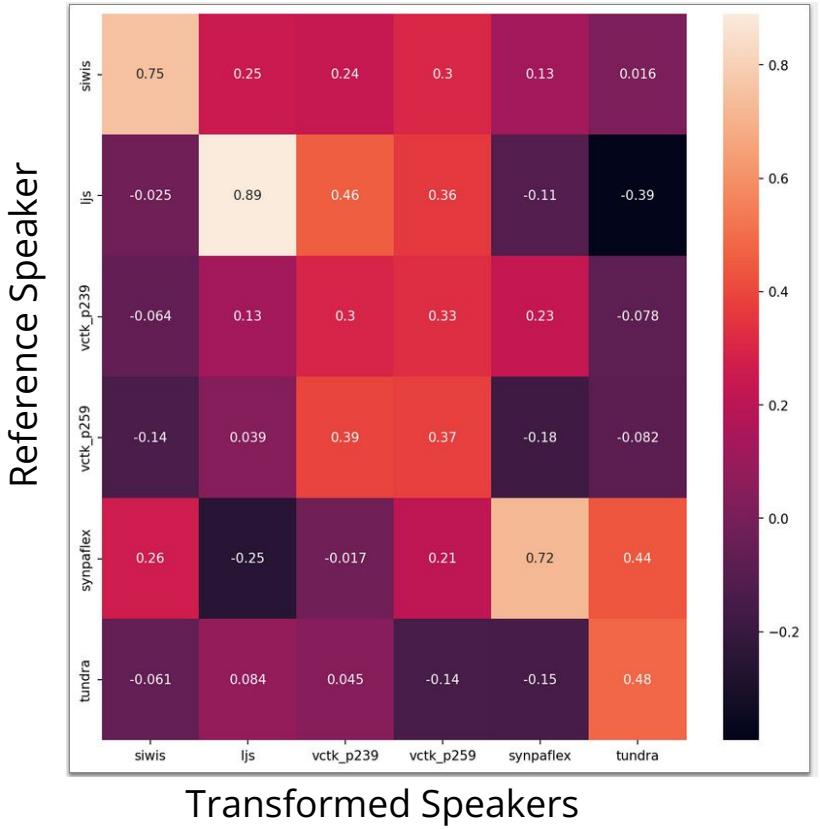
Objective Evaluation: Cosine Similarity(Version 1)



- From English to french speaker- bad result(for ljs)
- Average similarity around 0.51
- French to English better
- WER low, hence disregarded for subjective evaluation

Conclusion: speaker representing as id only. Hence, less complex and more efficient for speaker transformation.

Objective Evaluation: Cosine Similarity(Version 4)



- From English to french speaker- bad result(for all speakers)
- Average similarity around 0.28
- French to English better than other way
- WER better, hence selected for subjective evaluation

Conclusion: speaker representing as embedding network. Hence, more complex, less efficient for speaker transformation.

Objective Evaluation: Trade-off

- Best cosine: model 1
- Best WER: model 4
- Trade-off between TTS and voice characteristics' transfer
- MCD – the same models

	Version 1	Version 2	Version 3	Version 4
EN and FR	**5.864**	5.955	6.106	5.937
EN part	5.834	5.832	6.165	**5.794**
FR part	**5.93**	6.228	5.974	6.254

Table 3: Mel cepstral distortion for each model and language

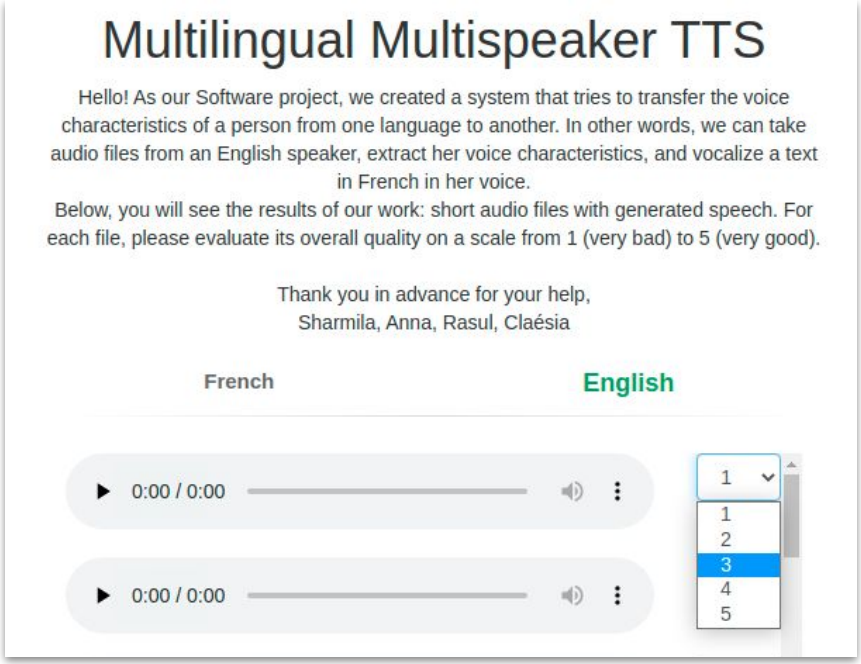
Maybe first do TTS, then add voice transfer

	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

Table 5: Word and character error rate for all model versions

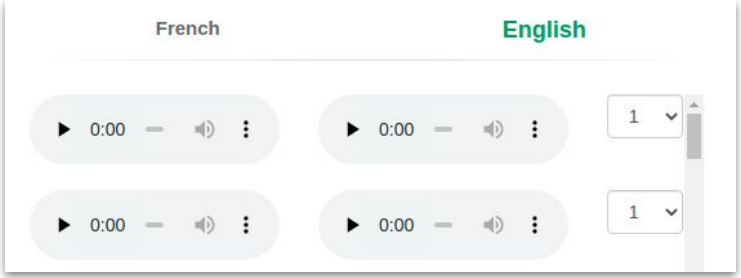
Subjective Evaluation

- Only model 4
- Two evaluation forms
  - Interpretability (18 and 16 respondents)
  - Voice transfer (9 and 9 for EN and FR)
- Scale 1-5, Mean Opinion Score

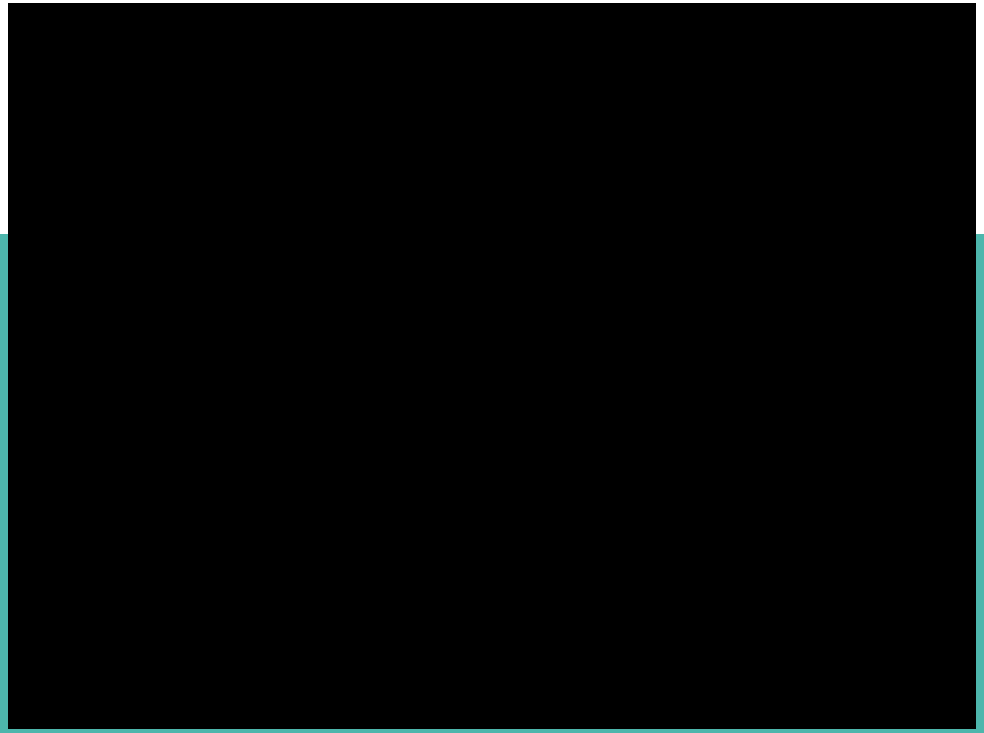


	Speech quality	Voice similarity
EN part	2.703	1.751
FR part	2.027	1.888

Mean opinion score for each language and evaluation setting



Demo



Conclusion

- Multilingual TTS system for transferring voice characteristics between speakers of French and English
- Four model architectures based on Grad-TTS with differentiated representations for languages and speakers
- Several experiments and their evaluation, subjective and objective

Possible improvements:

- Experiments with better and more balanced data
- Learning TTS first, and then speaker transformation

Thank you!
Questions?

Sharmila Upadhyaya

Anna Kriukova

Rasul Dent

Claésia Costa