Cross-lingual speaker adaptation for TTS applications

Sharmila Upadhyaya, Anna Kriukova, Rasul Dent, Claésia Costa

7 February, 2022

Abstract

This paper describes a multilingual TTS system for transferring voice characteristics between speakers of French and English. In this project, we modified the grad-TTS architecture [14] to implement four architectures with differentiated representations for languages and speakers. The results were evaluated objectively (with four metrics) and subjectively using Mean Opinion Score (MOS). The evaluations considered speaker transformation, speech interpretability and resemblance with the original audio. We present an analysis of the results and discuss ideas for future research. Despite results for French audio not being very competitive, some models produced decent results for speaker transformation.

Keywords: Multilingual TTS, Speaker Transformation, Grad-TTS

1 Introduction

For several decades, automatic speech synthesis, also known as text-to-speech (TTS), has been studied as part of the broader discipline of speech processing. In many regions, speech synthesis is already used to provide general services, such as automatic telephone operators. Moreover, TTS technologies like screen-readers have made digital materials more accessible to people whose needs are not met by strictly visual interfaces. Although there have been numerous advances in the quality and linguistic coverage of such systems in recent years, largely due to the diffusion of deep-learning based synthesis models, many of the best performing models are only trained to produce one language. This practice indirectly contributes to the marginalization of linguistic communities who may not have the resources to pursue language-specific speech technologies.

Similarly, despite considerable advances in encoding expressivity, synthesized speech is often perceived as flat and distinctly non-human. Voice cloning, a relatively new subfield of speech synthesis, aims to create intelligible speech while reproducing the characteristics of a particular voice. By combining methods from multilingual text-to-speech and voice cloning, it is possible to transfer elements of a person's voice into speech in a language that the person does not speak.

In this paper, we have introduced an adaptation of grad-TTS: Diffusion Probabilistic Model for Text-to-Speech [14] to generate expressive, speaker-specific voices in both English and French. We experimented with multiple representations for language and speaker and evaluated the performance of our models using both objective and subjective criteria[1]. Additionally, we explored representing text as language-specific tokens, as well as joint English-French phonemes. Finally, we expanded our model to produce speech in Louisiana Creole, an unseen language, by reusing phonemes from French and English.

2 Related work

2.1 Text to Speech

Text-to-speech, also known as speech synthesis, is a subfield of speech processing that aims to automatically convert graphical representations of text into comprehensible audio. Although concatenative and parametric synthesis were the standard for many decades, end-to-end models have become the new state of the art in the last few years [14]. As in parametric synthesis, end-to-end speech synthesis consists of two principle subprocesses: conversion of raw text into machine-readable acoustic data (feature generation) and the subsequent generation of audio files from said

[1]The code can be found in our github repo.

acoustic data via a vocoder. Wavenet, one of the most well-known neural vocoders, was trained to model speech autoregressively through convolutional neural networks (CNN) [21]. The success of Wavenet prompted the development of models that first predict mel-frequency spectrograms during the feature generation stage, such as Tacotron2, which combines CNNs and Long Short-Term Memory (LSTM) networks to represent the relationship between text and mel-frequencies through time [16]. More recently, flow-based vocoders like WaveGlow and generative adversarial networks such as Hifi-GAN have been shown to improve upon the performance of autoregressive models while requiring less time for inference [15, 10]. Similarly, the development of Glow-TTS and other flow-based monotonic alignment models has considerably reduced the temporal requirements for feature generation while maintaining high Mean Opinion Scores [9] Our work is primarily concerned with extending Grad-TTS, a multispeaker end-to-end system built on a novel diffusion-based monotonically-aligned feature generator and the Hifi-GAN vocoder, into the realm of multilingual speech synthesis [14].

2.1.1 Multilingual TTS, Voice Cloning, and Ethics

Whereas many non-neural models intentionally limited speaker variability as much as possible, the success of end-to-end models has contributed to the growth of multilingual TTS, a subfield of focused on using one model to generate speech in multiple languages, as well as voice cloning, where there is an additional goal of reproducing the vocal characteristics of specific speakers. Although these specializations have different aims, they share many techniques and researchers are increasingly focusing on cross-lingual voice cloning, which necessarily relies on advances in both areas [24]. In same-language scenarios, both speaker adaptation, where the parameters of a known speaker are adjusted to resemble a new speaker, and speaker embeddings, which directly model a given speaker, have been shown to been explored as viable methods for voice cloning [1].

As with other forms of media generation and editing, speech synthesis can potentially be misused to create misleading and harmful content [22]. In the case of voice cloning, a heightened susceptibility to fraud resulting from the ability to imitate real people has already been identified as a serious threat [18]. More specifically, deepfakes based on speech synthesis using deep neural networks have been shown to be capable of evading detection by both humans and production-level speaker recognition software [23]. Although there remain many open questions regarding how to prevent abuse and whether the positive use cases are worth the risk of abuse, mitigation strategies can be built directly into text-to-speech architectures. For example, one can design models which are produced speech that is similar to the original speaker, but more similar to other synthetic speech [8].

2.1.2 Low Resource TTS and Louisiana Creole

In consideration of the aforementioned risks, it is especially important to consider how multilingual voice transfer can positively impact society. Low resource TTS is one possible path, since multilingual speech synthesis models have been shown to be suitable for developing speech synthesis systems for low-resource languages. [4]. Moreover, developments in noise-robust TTS have created the possibility to fine-tune models for low-resource languages even when only relatively noisy data is available [3]. In light of these observations, we explore the possibility of including Louisiana Creole in our synthesis model.

Louisiana Creole, also known as Kouri-Vini, is a critically-endangered language spoken primarily in the Gulf Coast region of the United States. Like with many other creole languages, there are numerous lexical and phonological similarities between Louisiana Creole and its principle lexifier language, French [20]. Additionally, due to the influence of both regional French languages during the colonial period and prolonged contact with English in the last century, both Louisiana French and Louisiana Creole share some sounds with English that are not common in prestige varieties of European French [12]. Despite the potential feasibility of transfer learning, Louisiana Creole and other creole languages have long been overlooked in the fields related to speech and language processing [11].

3 Model architecture

3.1 Grad-TTS

Grad-TTS is an acoustic feature generator model proposed by [14] utilizing the concept of diffusion probabilistic modelling. The Grad-TTS is mainly composed of a diffusion-based decoder that converts Gaussian noise parameterized with the encoder output into mel-spectrogram, using Monotonic Alignment Search for alignment. The

model consists of three modules: encoder, duration predictor, and decoder. The proposed model allows to control the trade-off between inference speed and synthesized speech quality by varying the number of decoder steps at inference.

The models we implement in our research are based on Grad-TTS. We modified the architecture so that the inputs to our models are text, mel features of the reference audio, language, and speaker. All models learn to predict a mel-spectrogram for the given text, language, and speaker combination, which is later used to generate an audio. The models differ in the way how language and speaker are represented in their input: either by a simple id, or by an embedding. This gives four possible combinations, and they are illustrated in the Table 1.

	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

3.2 Front End

To deploy our models, we built a Python Flask app. The homepage consists of a simple entry form where the user can enter a text and specify which model, language and speaker combination they would like to listen to. After the necessary information has been submitted, the text and parameters are passed to the inference script, which runs the appropriate model from the local environment. In the case of Louisiana Creole, the orthographic text is first transcribed to IPA using a custom grapheme-to-phoneme module which relies on Epitran [13]. Furthermore, we provide the option to compare the audio produced by our models with audio from the gTTS Python package [5].

4 Experiments

4.1 Datasets

We used several datasets for French and English language. For English, we used LJS dataset [7] with a single female speaker (*approx 24 h*), and VCTK dataset [2] with one female and one male speakers' audios (*approx 1 hour*). Similarly, for French we used SIWIS dataset [6] with a single female speaker(*approx 4 h*), Tundra [19] with a single male speaker (*approx 1 h*), and Synpaflex [17] dataset with a single female speaker (*approx 11 h*). The datasets' overview is illustrated in the Table 2. The sampling rate for all of the audios is *22050 Hz*. These datasets were divided into training, test and validation dataset. We took 500 samples of short audios as test data while 100 samples as validation data. These data were prepared to include all speaker's sample according to their proportion in the total dataset.

Both English and French texts were converted into a sequence of digits representing phonemes. We performed experiments firstly without overlapping of the common phonemes (treating English and French phonemes as different ones), and then with overlapping either consonants or vowels.

4.2 Experiment setups

We performed 5 different experiments. The first one included the 4 versions of the model trained on the whole dataset. After noticing that the generated audios were in general shorter than the reference ones, we slowed down the generated ones according to calculated coefficients and evaluated two best models once more.

	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

The initial experiment did not show outstanding results, thus we updated the strategy and kept only the best datasets for the next experiment: LJS and SIWIS (for English and French respectively). Our hypothesis was that the imperfect quality of the data could influence the results.

Additionally, we noticed that the French audio tended to cut out more sounds than the English audio. Reasoning that the system might have been learning entirely different sets of parameters for each language, rather than using than learning to share general paramaters and adapt the output via the language control mechanism, we decided to experiment with using the same symbols (and consequently IDs) for similar English and French phonemes. Since the initial English system used characters rather than phonemes as input, we first implemented subsystem to convert English to ARPAbet symbols, and then made two tables mapping English and French phonemes onto new symbols in the combined system: one where only some consonants were merged, and another where both vowels and consonants were merged.

The results of these experiments are detailed in the evaluation section.

4.3 Louisiana Creole

As noted in Section 2, Louisiana Creole is a critically endangered language that is closely related to French. After training models that theoretically covered nearly the entire phonemic inventory of Louisiana Creole, we sought to explore whether our models could accurately produce speech in this language without relying on any Louisiana Creole specific training data (zero shot synthesis). If successful, we would provide a method which could be repeated for other closely related low-resource and very-high resource language/variety pairs, and especially other creole languages.

To adapt our multilanguage models to produce Louisiana Creole, we created a dictionary that maps phonemes directly to labels, which would facilitate mixing English and French phonemes in the same sequence. To obtain phoneme sequences, we first passed strings through a previously-built rule-based grapheme-phoneme converter.

In broad qualitative terms, the quality of synthesis for Louisiana Creole is closely tied to the quality of synthesis for French. Due to the scarcity of high-quality recordings available for research, we decided to forego objective evaluation. Similarly, due to the historically unequal and extractive relationship between academic linguists and the speech community, we elected to postpone requesting subjective evaluation until after resolving the issues noted in our French models.

5 Evaluation

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

5.1 Objective evaluation

For objective evaluation, we computed several metrics that are often used in the field of speech synthesis.

5.1.1 Mel cepstral distorion in initial experiments

Mel cepstral distortion measures how different two sequences of mel cepstra are and therefore how close the synthetic audio is to natural speech.

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

Where T is the number of frames in the shorter audio and mc is the mel-frequency cepstral coefficients.

The results of the evaluation are presented in Table 3. As we can see, the difference between the results is not that significant. However, we can clearly see that the version 3 is worse than all other models. It will be confirmed by the WER results in the next section as well. Overall, the results are better in version 1 (for the whole data "EN and FR" and for "FR" part), whereas for the English they are better in version 4. As we will show later, the best models indeed turn out to be 1 and 4.

	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

5.1.2 WER and CER in initial experiments

Word and character error rate assesses how interpretable the synthesized speech is, taking into account the number of word and character mistakes in transcriptions by an ASR system. For this reason, we used NVIDIA ASR models: QuartzNet15x5Base-En for English and Stt_fr_quartznet15x5 for French. Before computing the scores for the outputs of our models, we assessed how the ASR models work on the originals of our data.

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

Table 4: Word and character error rate for reference audio files

As can be seen, English model performs much better than the French one. However, all the values are a little lower than the ones reported in the models' descriptions.

The results for our models are illustrated in the Table 5. As can be seen, we computed the mean values for each model, as well as for English and French texts separately. "EN part" in the table includes the files where the original audio and text are in English, but both EN and FR speakers are taken for speaker transformation. We expected the results for EN and FR parts to differ a little, with EN being better, because of the different performance of ASR models (ref. to 4). However, they differ much more than we initially thought they would. One of the possible explanations is that the data in training was imbalanced, with much more EN samples. Overall, the results of the model version 4 are the best ones in terms of WER and CER, and we will use it as one of the two best models later; the second one will be the model version 1 because it receives the best scores in the cosine similarity (it is discussed in the next session). The worst results, however, are obtained by the model version 3. Listening to the audios generated by it, we discovered that they indeed consisted of indistinguishable sounds; however, we couldn't find any explanation for that. The only conclusion we can make in such a case is that the model architecture, namely, language input as an embedding together with speaker as an id, is the least promising combination.

We also wanted to see if the language of the speaker influences the metrics, so computed mean WER and CER scores for four cases: "EN orig + FR voice" (when the original audio and text are in English but the speakers for transformation are French), "EN orig + EN voice", etc. Our hypothesis was that the results would be visibly better for the pairs of the same language, but as we can see, it is so only for the EN part of model 1, whereas all other models and combinations have approximately the same values. This is why in the following calculations we distinguish only between EN and FR parts but not between original and speaker languages.

	Version 1		Version 2	
	WER	CER	WER	CER
EN and FR	73.1071	47.3306	52.6077	29.5042
EN part	**61.2314**	33.0791	**33.0253**	14.8118
FR part	**99.4639**	78.96	**96.1243**	62.1541
EN orig + EN voice	55.4693	29.82	33.3336	14.9849
EN orig + FR voice	67.0083	36.3465	32.717	14.6387
FR orig + EN voice	99.534	79.3299	96.5818	62.5462
FR orig + FR voice	99.3939	78.5901	95.6668	61.7619
	Version 3		**Version 4**	
	WER	CER	WER	CER
EN and FR	98.8535	79.3263	37.3646	16.4443
EN part	**98.705**	78.618	**25.3051**	10.4785
FR part	**99.1834**	80.9002	**64.0432**	29.6422
EN orig + EN voice	98.3506	78.7617	25.0318	10.2557
EN orig + FR voice	99.0594	78.4744	25.5788	10.7017
FR orig + EN voice	99.2792	82.373	64.5124	29.5992
FR orig + FR voice	99.0876	79.4275	63.5739	29.6852

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

5.1.3 Cosine similarity

Cosine similarity between speaker embeddings shows how close the transferred voice is to the original speaker's voice. To extract the embeddings from the synthesized and original audios, we used the ECAPA-TDNN model that is state of art for speaker diarization and recognition. We experimented with the pretrained model and it did not classify our speakers properly as it was trained on external speakers and for the better result finetuning was one viable solution. Consequently, we first finetuned the model for speaker recognition on our 6 speakers and then extracted the speaker embeddings which are later used as reference embeddings in the speaker similarity. Additionally, we then took all the generated audios for each speaker, extracted embeddings for all of the audios and measured the similarity with the corresponding reference embedding. All the similarity scores between reference and generated audios for each speaker are averaged and presented below in the Table 6.

	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

In the Table 6, the speaker's sex is mentioned as either female or male. Since we are using an external finetuned model for speaker embedding extraction, the result of our speaker similarity depends on the shortcomings of this model. Hence, below in Figure 1 we present the similarity matrix of all the speakers on a gold dataset. For each speaker, 4 reference audio were selected and similarity score was calculated. The scores shown in the table are the average cosine similarity scores of each speaker against all other speakers.

As shown in Table 6, speaker transformation for model 1 where both speaker and language are represented as id) seems better than the rest (0.51, the average score for all speakers). However, the best model according to WER and CER scores was model 4 (with embedding networks as representation for speaker and language), which has poor results in terms of cosine similarity. Thus, we suppose that there's a trade-off between good speaker similarity

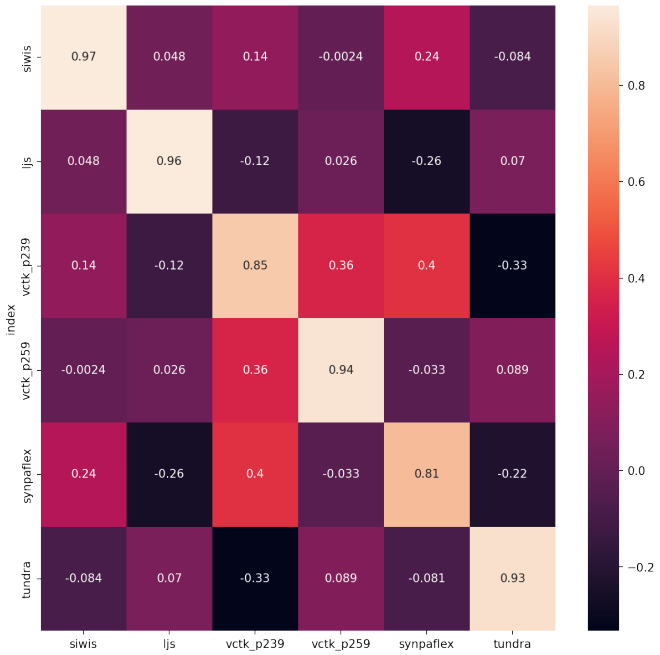


Figure 1: Speaker similarity between reference speakers using finetuned ECAPA-TDNN model

result and better speech quality (measured by WER). When the model learns to transfer voice characteristics, it hinders speech interpretability, whereas better quality of speech leads to fewer similarity with the reference audios.

Because of this, further experimentation to analyse the effects of the number of speakers was carried out. While reasoning the selection of datasets, we realized that the quality of data for French audio was not optimal. Hence, we chose 2 speakers: one for French (LJS) and one for English (SIWIS), and trained the model version 1 and model version 4 hoping to get a better result. However, the observed speaker similarity score was worse in that case. From French speaker to English transformation, similarity we got was 0.23 and from English to French it was around -0.0098. The results of the model version 1 were very similar to those of the version 4.

The conclusion we made was that with 2 speakers and 2 languages the model was unable to differentiate between the language features and speakers voice's features. Hence, it is required to have more than two speakers for the model to distinguish the audio based on language and speaker.

Furthermore, Figure 2 shows the breakdown of speaker transformation score for each model and speaker in the initial experiment with 6 speakers. The speaker transformation score for model version 3 is higher than other versions. Model version 1 is the second based except the transformation from English to French speaker is low. Rest of the versions do not show good result. Both version 2 and 4 models have speaker represented as an embedding, and differ in their language representation (either id or embedding), and we assume that the complexity of speaker representation does not improve the transformation. We can infer that the representation of speakers as id is less complex and provides better result which is a important finding for improvements and further experimenting.

5.1.4 WER and CER in further experiments

Seeing that WER and CER results were not good, we tried to find possible reasons for that. Firstly, we noticed that generated audios were shorter than the corresponding references. We couldn't come up with the reason for that but decided to see if the results would change if we stretch the generated audios. For that, we calculated the mean proportion of lengths between generated and reference audios (Table 7).

For example, we can see that for the FR part of the model 1 the generated audios are two times shorter on average than the reference ones.

We then stretched the generated audios according to these coefficients and calculated WER and CER again (Table 8). Unfortunately, it didn't influence the results much, they even became a little worse.

The next idea we had was that the results were not good because of the quality of the training data. LJS and

	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

SIWIS datasets are known for better quality, and we decided to see if the results can be enhanced by training only on them, as was mentioned in the section describing cosine results. We trained model 1 and model version 4 on the new dataset with only 2 speakers and calculated WER and CER again (Table 9). The error rates here decreased for the EN part but went up for the FR one. This can be explained by the imbalance in the training data: with 24 hours of EN speech and only 4 hours of FR. Together with poor results of the cosine similarity, we could clearly see that this setup was not helping our model.

	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

Finally, we performed two more experiments, this time trying to merge phonemes for English and French. To our surprise, using ARPAbet with the symbol mergers had a particularly ruinous effect on the accuracy of both TTS. Upon manual review, the audio was mostly incomprehensible. This confirmed our suspicions that the model was not effectively learning how to separate the languages. Even so, the high error rates were shocking because we might expect output with mixed English and French characteristics, not unintelligible noise.

	Merged consonants		**Arpa**	
	WER	CER	WER	CER
EN and FR	61.2942	33.0286	98.9164	80.5975
EN part	44.5812	21.7021	98.6807	80.5139
FR part	83.4213	48.0242	99.2263	80.7075

Table 10: Word and character error rates for model 1 in experiments with merged phonemes

5.2 Subjective evaluation

To evaluate how the produced speech is perceived by speakers, we performed subjective evaluation. We chose the model version 4 for this evaluation, because even though the version 1 had better scores of cosine similarity between speakers, its generated speech had significantly larger Word Error Rate, meaning that the audios were not interpretable enough.

We created two online forms where we asked English and French speakers to rate the given speech samples. In the first form, they needed to evaluate the overall quality of audios, whereas in the second one we provided pairs of

reference and generated audios, and the task was to evaluate how similar they were in terms of voice characteristics. For both evaluations, the scale from 1 to 5 was used, and Mean Opinion Score was calculated.

In the first form, we provided 6 audio samples per each speaker (there are 3 speakers for EN and 3 for FR), meaning that for each language there were 18 samples to rate. In the second form, we provided 31 pairs of audios that included one generated sample and one speaker reference.

We received 34 answers in the first form, where 18 answers were for English language and 16 for French one. In the second form, we received 18 answers, 9 per language. The evaluators consisted mostly of friends and classmates. The results of this experiment can be seen in Table 11.

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

Table 11: Mean opinion score for each language and evaluation setting (scale from 1 to 5)

As we can see from the results, somehow the subjective evaluation confirms the results of objective evaluation. Considering the MOS metric, we received poor evaluation results for both forms. Some evaluators gave us feedback about the audios, most of them talked about the effort to understand the generated audios. In the first form, in the French part some of them claim how it "seems like words are cut in the middle", and in the English part some people reported that they rated the quality solely based on the words they could understand. In the second form, in the French part the feedback concentrated on how the voices in the audios do not match and how the generated samples are "more loud and less highs" than the reference ones. For the English part in the second form we did not received any feedback. All of this leaves space for further experiments to improve both speaker similarity and quality of the produced speech.

6 Future work

We have identified several possible sources of error which could explain the relatively low quality of our synthesized speech, especially in the case of French. Firstly, we deployed the same HiFi-GAN vocoder checkpoint that was provided with Grad-TTS. Since the checkpoint was only trained on English speech, it is possible that our changes to the spectrogram generator did encode useful information, but at the expense of degrading the performance of the vocoder. By hypothesize that fine-tuning the vocoder on our datasets could result in significant gains in intelligibility. Secondly, the speaker embedding module showcases unusual results but since the speech generated are not better for French audio, focusing on TTS first then only performing experiment for transformation would be the best option.

7 Conclusion

In conclusion, we successfully implemented an application version of our cross language speaker transformation for Text-to-Speech in French and English language project idea. This process included the implementation of gradtts model with the representation of texts as phoneme integer, language as either id or embedding network and speaker as id or embedding network. The usage of only 2 speakers does not contribute in the distinction of language and speaker features and hence the variation of more than 2 speakers is required in order to distinguish the language and speaker representation. Using our web application, users can input a text in French or English, select the language of the text and speaker whose voice they want to be played the audio in. After generation, user will receive a response from the system in the form of an audio file that can be played in the application itself. The generated audios did not achieve evaluation scores that are comparable to those of previous works for speech generation. However, the speaker transformation results are quite competitive and this project opens an opportunity and direction for us to work further in future to solve this problem.

References

[1] Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples, 2018.

[2] publisher = University of Edinburgh. The Centre for Speech Technology Research (CSTR) year = 2017 Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, title = SUPERSEDED - CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit.

[3] Dongyang Dai, Li Chen, Yuping Wang, Mu Wang, Rui Xia, Xuchen Song, Zhiyong Wu, and Yuxuan Wang. Noise robust tts for low resource speakers using pre-trained model and speech enhancement, 2020.

[4] Marcel de Korte, Jaebok Kim, and Esther Klabbers. Efficient neural speech synthesis for low-resource languages through multilingual modeling, 2020.

[5] Pierre Nicolas Durette. Google Text-to-Speech.

[6] Pierre-Edouard Honnet, Alexandros Lazaridis, Philip Garner, and Junichi Yamagishi. The siwis french speech synthesis database – design and recording of a high quality french database for speech synthesis. 01 2017.

[7] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[8] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis, 2019.

[9] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search, 2020.

[10] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020.

[11] Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 58–71, Online, November 2021. Association for Computational Linguistics.

[12] Oliver Mayeux. *Rethinking decreolization: Language contact and change in Louisiana Creole.* PhD thesis, University of Cambridge, 2019.

[13] David R. Mortensen, Siddharth Dalmia, and Patrick Littell. Epitran: Precision G2P for many languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May 2018. European Language Resources Association (ELRA).

[14] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[15] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis, 2018.

[16] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018.

[17] Aghilas Sini, Damien Lolive, Gaëlle Vidal, Marie Tahon, and Elisabeth Delais-Roussarie. SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018.

[18] Sergey S. Sokolov, Oleg M. Alimov, Dmitry A. Tyapkin, Yury F. Katorin, and Aleksandr I. Moiseev. Modern social engineering voice cloning technologies. In *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 513–516, 2020.

[19] Adriana Stan, Oliver Watts, Y. Mamiya, Mircea Giurgiu, Robert Clark, and Junichi Yamagishi. Tundra: A multilingual corpus of found data for tts research created with light supervision. 08 2013.

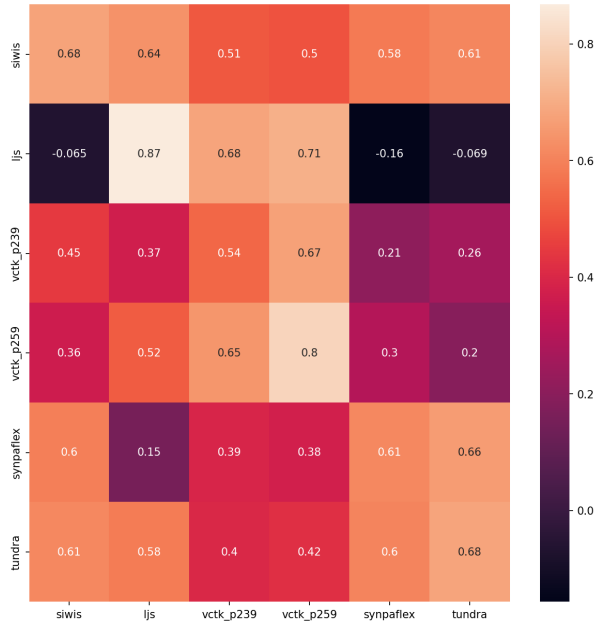
[20] Albert Valdman. *French and Creole in Louisiana*, chapter "The Structure of Louisiana Creole". Plenum Press, New York, NY, 1997.

[21] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

[22] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

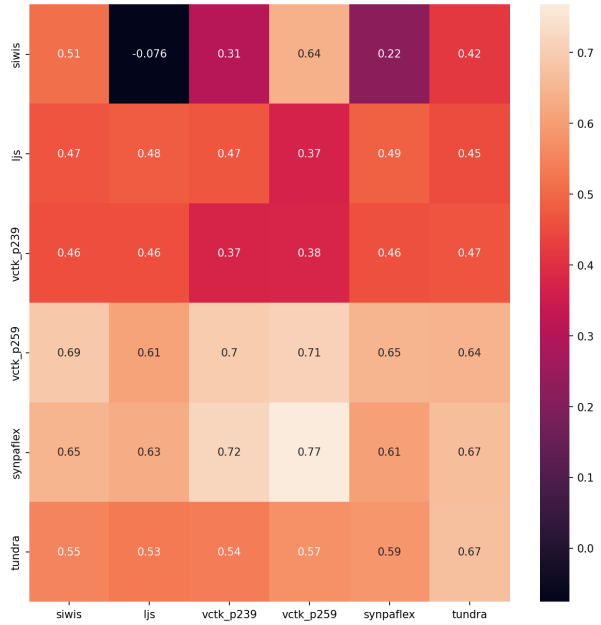
[23] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y. Zhao. "hello, it's me": Deep learning-based speech synthesis attacks in the real world. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, Nov 2021.

[24] Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning, 2019.

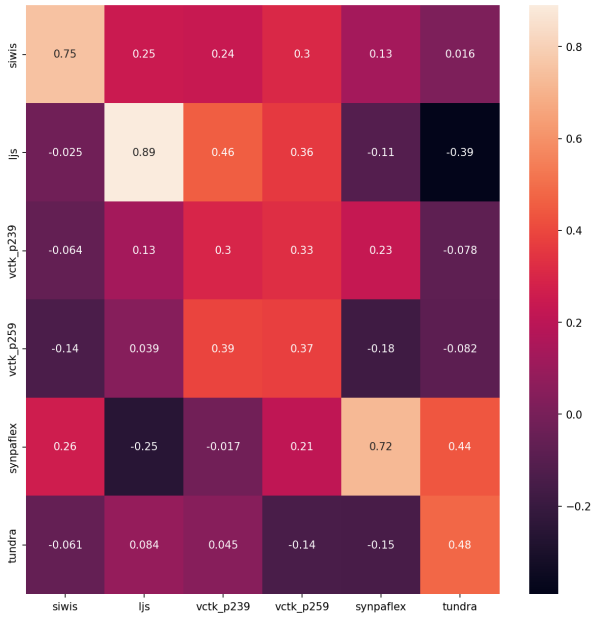


(a) Speaker similarity between reference speaker to target speaker for model version 1

(b) Speaker similarity between reference speaker to target speaker for model version 1



(c) Speaker similarity between reference speaker to target speaker for model version 1



(d) Speaker similarity between reference speaker to target speaker for model version 1

Figure 2: Speaker Similarity breakdown