Cross-lingual speaker adaptation for TTS applications

Sharmila Upadhyaya, Anna Kriukova, Rasul Dent, Claésia Costa

January 2022

Abstract

This paper describes a multilingual TTS system for transferring of voice characteristics between speakers in French and English. We modify grad-TTS and experiment with four different model architectures. The results are evaluated both objectively with four metrics and subjectively using MOS for assessing speaker transformation and speech quality. We provide analysis of the results and discuss possible further directions of research. The system demo is available online on (link).

1 Introduction

Over years, research based on end-to-end neural Text to Speech models has been increasing and has shown state-of-the art performance. Moreover, most of the work done in language-dependent system has provided optimal result. However, development of language-independent TTS models is more challenging since some languages have no overlapping character sets or phonemes. In the case with English and French, they have similar character sets but different phonemes.

Voice cloning is another field in Speech Synthesis, it is the audio is generated to clone the voice features of a particular speaker. Our goal is to setup experiments and research analysis for the multilingual TTS along with speaker variation across the languages.

In this paper, we have introduced the implementation of grad-TTS: Diffusion Probabilistic Model for Text-to-Speech [reference] to generate the audio in particular language, for particular speaker. During inference, the model takes the language, text and speaker as inputs and produces the speech in the specified language, in the voice of the specified speaker. During the experiments, representations for language and speaker were varied and the performance of models are evaluated using both objective and subjective evaluation. Similarly, the text representation is done as phoneme units where we combined the phonemes for French and English language without overlapping. The representation of each phoneme is controlled by the language due to the highly similar alphabets with varying sounds for same alphabets. The main contribution of our work is as follow:

1. Multilingual Text-to-Speech representation for French and English language.
2. Speaker transformation across same or different languages.
3. Evaluating the effects of varying representations of language and speaker.
4. Evaluating the effects of adding speakers on the performance.

2 Related work

2.1 Text to Speech

Text-to-speech, also known as speech synthesis, is a subfield of speech processing which 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 [8]. 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 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) [14]. The success of Wavenet prompted the development of models which 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[10]. 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 [9, 6]. 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 [5] 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 [8].

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 voicing cloning, which necessarily relies on advances in both areas [15]. 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 generative paradigms in artificial intelligence, speech synthesis can potentially be misused to create harmful and offensive content. In the case of voice cloning, the potential damages of misuse are particularly high because such content would be attributed to a real person. There remain many open questions regarding how to prevent abuse and whether the positive use cases are worth the risk of abuse. []

2.1.2 Low Resource TTS and Louisiana Creole

Louisiana Creole, also known as Kouri-Vini, is a critically-endangered language spoken primarily in the Gulf Coast region of the United States. Like many other creole languages, there are numerous lexical and phonological similarities between Louisiana Creole and its principle lexifier language, French [13]. 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 [7].

3 Model architecture

1 page
Grad-TTS[]
4 model versions

	language embedding	speaker embedding
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.

3.1 Front End

The system is available online on ¡site¿.

4 Experiments

1/2 pages

4.1 Datasets

We used several datasets for French and English language. For English, we used LJS dataset [4]; single speaker female speaker (*approx 24 h*), from VCTK dataset [2] one female's (approx) and one male speaker's(approx needtofillup) audio. Similarly, for French we used siwis dataset [3]; single female speaker(*approx 4 h*), tundra [12]; male speaker (*approx 2 h*) and synpaflex [11] dataset; single female speaker audio (*approx 10 h*). The sampling rate for all of the audios is needtofillupkhz. These dataset were divided into training, test and validation dataset. We took 500 sample of short audios as test data while 100 samples as validation data. These data were prepared to include all speaker's sample according to there proportion in the total 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

Both English and French texts were converted into a sequence of digits representing phonemes. Instead of overlapping the common phonemes, we treated English and French phonemes as different ones. This leaves a window to work on the overlapping in the future. In total, there were .. phonemes.

4.2 Louisiana Creole

After training models which theoretically covered nearly the entire phonemic inventory of Louisiana Creole, we sought to determine if our models could accurately produce speech in this language. To adapt our multilanguage models to produce Louisiana Creole, we created a dictionary which maps phonemes directly to labels, which facilitated 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.

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

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.

	Version 1	Version 2	Version 3	Version 4
EN and FR	5.864	5.955	0	5.937
EN part	5.834	5.832	0	5.794
FR part	5.93	6.228	0	6.254

Table 3: Mel cepstral distortion for each model and language

ANALYSIS OF MCD RESULTS

5.1.2 WER and CER

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 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 thought they would. This can be explained by the fact that the data in training was imbalanced, with much more EN samples. Overall, the results of the model 4 are the best ones, and we will use it later.

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' (it means that the original audio and text are in English but the speakers for transformation are French), 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	0	0	37.3646	16.4443
EN part	0	0	**25.3051**	10.4785
FR part	0	0	**64.0432**	29.6422
EN orig + EN voice	0	0	25.0318	10.2557
EN orig + FR voice	0	0	25.5788	10.7017
FR orig + EN voice	0	0	64.5124	29.5992
FR orig + FR voice	0	0	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 diarizarion and recognition. We experimented with pretrained and finetuned models. COnsequently, we first finetuned the model for speaker reccognition on our 6 speakers then extracted the embeddings for speakers which is used as reference embedding in the speaker similarity. Similarly, we then took all the generated audio for each speaker, extracted embeddings for all of the audio and measured the similarity with the reference embedding. All the similarity score between reference and generated audio for each speaker was 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, speakers gender is mentioned as either female or male. Since we are using a external finetuned model for speaker embedding extraction, the result of our speaker similarity depends on the shortcoming of this model. Hence, below in table 7 we have presented the similarity matrix of all the speakers on gold dataset. For each speaker, 4 reference audio was selected and similarity score was calculated. The below shown score is the average cosine similarity score of speaker against all speakers. As shown in table 5, speaker transformation for model 1 seems better than the rest. However, due to the tradeoff between good speaker similarity result and better wer score, further experimentation to analyse the effects of number of speaker was carried out. While reasoning the selection of dataset, we realized that the quality of synpaflex for french audio is not that great. Hence, we chose 2 speakers: one for french and one for english and trained the model 1 and model 4 to find a better tradeoff. The observed speaker similarity score was worst, from french speaker to english around 0.23 and english to french around -0.0098. Similar was the result for model 1. The conclusion we came for this was as we have 2 speakers and our languages also 2, the model cannot quite differentiate between the language feature and speakers voice feature as two speakers are also from different languages. Hence, it is required to have more speakers for the model to distinguish the audio based on language and speaker both.

5.1.4 WER and CER in further experiments

Seeing that the 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).

	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)

The results mean that for example 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.

	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

The next idea we had was that the results were not good because of the quality of the training data. LJS and SIWIS datasets are known for better quality, and we decided to see if the results can be enhanced by training only on them. We trained model 1 and model 4 on them 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, again, can be explained by the imbalance in the training data: with 24 hours of EN speech and only 4 hours of FR.

5.2 Subjective evaluation

To evaluate how the produced speech is perceived by native speakers, we perform subjective evaluation. We created online forms where we asked native speakers (EN and FR) to rate the given speech samples on a scale from 1 to 5 in terms of their oral quality. We then used the Mean Opinion Score as the metric.

In the form, we provided 10 audio samples per each speaker (there are 3 speakers for EN and 3 for FR). We selected two of our models with the optimal objective scores: one with the best speaker cosine similarity between

	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

the speakers and the other with the best Word Error Rate. Our main observation from the objective evaluation was that the better speaker transformation is shown by the model 1 (where both speaker and language are represented as id). Similarly, model 4 (with embedding networks as representation for speaker and language) performs better for general multilingual TTS. Hence, our hypothesis is that model 1's poor performance of speech synthesis (measured by WER) might be due to the speaker transformation that affected the general comprehensibility of audio. In order to further investigate this, we decided to calculate the MOS score for speech comprehensibility separately for:

1. the audios that were transformed to different speakers i.e. the corresponding reference audios were not available for those speakers.

2. audios that were transformed to same speaker i.e. the corresponding reference audios were available for those speaker.

With 2 models, it resulted in 60 speech samples in total, half of which French, the other half English. Since evaluating 48 audio samples in one experiment would require too much time, effort, and concentration from a native speaker, we decided to create 2 forms for each language. As a result, each subject needed to evaluate 24 audios in their native language.

The results of this experiment can be seen in Table 10. Inter-rater agreement was ¡later¿.

	Version 1	Version 2	Version 3	Version 4
EN and FR	0	0	0	0
EN part	0	0	0	0
FR part	0	0	0	0

Table 10: Mean opinion score for each model and language.

As we can see from the results, ..¡analysis later¿

6 Future work

Mention transfer learning

7 Conclusion

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] 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.

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

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

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

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

[8] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech, 2021.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.