Voice Style Cloning for Chinese Speech

https://youtu.be/ruzTRWM8K0Y

Ziqi Chen Haiyun Wang Luoyi Yang

Civil and Environmental Engineering

CS230 Project

Motivation

Mimicking or cloning a person's voice style sounds interesting, and we also find it populer as many YouTube videos using celebrity voices to make new audio clips. However, the quality of newly generated audio clips are usually not so good. More importantly, the majority of these are in English. Therefore, we decided to look into modern text-to-speech systems that can be trained to produce Chinese speech in a specific person's vocal style.

Dataset

Open-source online dataset from data-baker.com: A file called Chinese Standard Mandarin Speech Copus (10000 Sentences) containing 100000 (approximately 10 hours) wave audios in which Chinese sentences are read by a single female Chinese broadcaster.

Figure 1: Dataset samples

Data Preprocessing

We format audio label for each audio in our dataset in the form of audio number plus Chinese phonetic alphabet for that audio. Then we generate <label, audio wav> pairs as our training data labels.

Methods

Tacotron structure:
The architecture of Tacotron is displayed in Figure 2. Tacotron can be divided into three main parts: encoder, decoder and post-processing net. Encoder extracts sequential representation of the text by applying a series of non-linear transformations to the character embeddings. Later on, the encoder representation is passed to a tanh attention decoder where the attention module is applied to each step of the decoding. The decoder uses a fully connected output layer to predict the output targets. At last, post-processing synthesizes targets generated by the decoder to a spectrogram using a Griffin-Lim synthesizer.

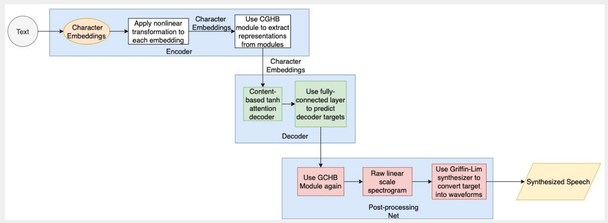

Figure 2: Framework of the Tracotron

Loss Function:

Loss function L used for training Tacotron is a combination of two simple L1 loss functions:

$$L = L_{mel-scale\ spectrogram} + L_{linear-scale\ spectrogram}$$

Where the Lmel-scale spectrogram represents loss from the seq2seq decoder; Linear-scale spectrogram represents loss from the post-processing net.

Training Steps:

We first label our data and generate <text, audio> pairs as our training data. We then move on to training Tacotron with the first dataset Chinese Standard Mandarin Speech Copus (10000 Sentences). Due to limitation in time and computational power, we chop the dataset and train the model with different input data size for 5000 steps. After obtaining recognizable Chinese audio with 1000 training examples, we moved on to tuning hyperparameters and changing optimization methods. We set the original Tacotron as baseline. To improve Tacotron's performance, we test combinations of different initial learning rates, batch sizes and optimization methods. We double the learning rate to 0.004 with the expectation of faster loss convergence during training. We also increase the batch size to 64 attempting to accelerate the overall training process. Besides, we try to implement RMSprop Optimization instead of Adam Optimizer to foster faster gradient descent.

Baseline

We set original Tacotron as baseline, which adopts initial learning rate (LR) of 0.002 with decay learning rate, batch size (BS) of 32 and Adam Optimizer. Hyper-parameter Details are summarized in Table 1.

Table 1: Baseline Model Parameters

Baseline Model	
Batch Size	32
Optimizer	Adam
Initial Learning Rate	0.002
Decay Learning Rate	True

Results & Analysis

Loss Evaluation

Table 2: Loss Comparison

Optimizer		LR = 0.002	LR = 0.004
Adam	BS = 32	0.19281 (baseline)	0.18998
	BS = 64	0.19286	0.1656
RMSprop	BS = 32	0.23615	Not Trained

* Loss at training step 500

Losses are evaluated for all five models at the training step of 500. Model with LR=0.004, BS=64 and Adam Optimizer outperforms all other models in terms of loss. RMSprop Optimizer does not work well as the loss is much higher than any other other models. Therefore, RMSprop is not considered for further investigation.

Synthesized Audio Performance Evaluation

Two approaches are utilized to evaluate the performance of the synthesized audios from models trained to the step of 10000.

Approach 1: Machine Evaluation
We modified Resemblyer to evaluate 12 real speech and obtain training target of 0.81. Then 6 synthesized audios from each model are evaluated using Resemblyer.

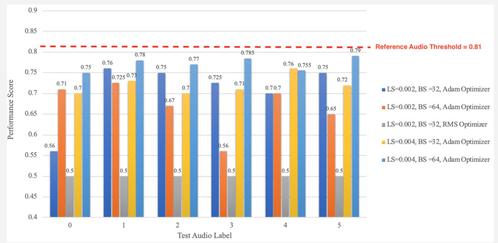

Figure 3: Machine Evaluation Chart

As shown in Figure 3, among the five methods, we find combination 5 (LR=0.004, BS=64, Adam Optimizer) being the best, combination 3 (LR=0.002, BS=32, RMSprop Optimizer) being the worst and the rest fall in the middle. Although synthesized audio from combination 5 does not reach the target threshold, we have made substantial improvement compared to the baseline model.

Approach 2: Human Evaluation

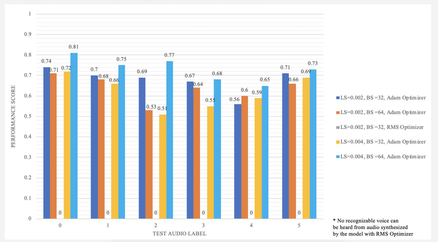

Figure 4: Human Evaluation Chart

Evaluation surveys are distributed to 10 people and they are asked to score the audios from 0 to 1 based on speech naturalness, style, and correctness. Then, we take the average of the 10 scores for each audio clip as the subjective evaluation score. As shown in Figure 4, human score ranking generally matches the machine evaluation results. We confirm that combination 5 is the best training approach among all the combinations that we have tried so far.

Conclusion and Future Work

In this project, we have researched on improving Tacotron, a text-to-speech model, to achieve voice cloning. The five sets of synthesized audios are being evaluated both subjectively and objectively against the target training goal. After finding the idealist setting (LR=0.004, BS=64, Adam Optimizer), we have tried to improve the model performance to reach our target threshold by increasing training steps and datasets.

For future work, we plan to continue our model training to reach the target score of 0.81. Besides, we would further improve the functionality of Tacotron and improve it to a real-time voice cloning model which takes voice recorded by any person in real time.