[Proposal-ML] Mimicking Humanity: A synthetic data-based approach to voice cloning in Text to Speech Systems

Michael Wang
DCST
Tsinghua University
2024280208

Algazinov Aleksandr
DCST
Tsinghua University
2024280037

Joydeep Chandra
DCST
Tsinghua University
2024280035

1 Background

While text-to-speech (TTS) technology has greatly improved in recent years, traditional TTS models often rely on large datasets of recorded human speech to produce accurate and natural synthetic speech. However, obtaining high-quality data at scale presents challenges such as high costs, user privacy concerns, lack of linguistic diversity, and data paucity.

Speech data remains inaccessible for many languages and dialects, leading to disparities in TTS quality across languages and accents.

This project aims to train a voice cloning model for use with TTS systems primarily using synthetic training data. The objective is to create a voice cloning model that can generalize well without reliance on human-produced training data.

The model is intended to be small enough to permit local inference. This will allow users to generate speech with their voice without granting legal rights to their data to a third party, as is currently the case with commercially available voice generation platforms. Furthermore, if generalizable, this approach of using synthetic training data can reduce the cost of acquiring training data when applied to the training of other categories of ML models.

2 Related Work

Several important TTS models include WaveNet [12], Tacotron [22], and FastSpeech [2]. WaveNet is an autoregressive model capable of producing high-quality, natural-sounding speech. However, since it is an autoregressive model, it is computationally inefficient. Tacotron is a seq2seq model with attention, making it faster compared to WaveNet. However, the model still faces issues with robustness and controllability, and it is still computationally demanding. The FastSpeech model shows better performance compared to Tacotron 2 (in terms of the listed metrics), while being faster.

While all of these models are capable of generating high-quality speech, they are limited to voices associated with pre-trained speakers. Zero-shot TTS systems aim to add the ability to generate speech for unknown speakers with only a short voice sample.

VALL-E [18] and VALL-E 2 [19] are zero-shot synthesis models that include direct TTS capability. Unfortunately, the source code associated with these models is not publicly available. OpenVoice

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

[20], which we take inspiration from for our system architecture, is a newer model that aims to handle the voice cloning aspect without needing to directly handle any TTS function.

TTS models typically use datasets derived from LibreVox, which is a public domain audiobook repository, for training. However, synthetic data [21] has been successfully used for training speech recognition models, which suggests that synthetic data can be used as part of training other types of models as well.

3 Proposed Method

The primary focus is on developing a model that can take any speech audio and modify it so that the modified audio sounds as if it was spoken by some target speaker.

We intend for the model to have two inputs. The first is the audio to be transformed. Typically this will be audio generated by some TTS model, but human-spoken audio should also be a valid input. The second is a voice sample of the target speaker, which is used to extract unique characteristics of the target speaker's voice, such as pitch, tone, and timbre. Ideally, this should not need to be longer than a few seconds.

To clone the target speaker's voice, the model generates an embedding that represents the features extracted from the voice sample. Then, the input audio is passed through a voice conversion block, which modifies the input audio to match the speaker's embedding characteristics.

The final output is speech audio that sounds like it was spoken by the target speaker, retaining the linguistic content of the original synthesized audio but transformed to the voice characteristics of the provided speaker sample.

Synthetic data will be used to pre-train the model, reducing the need for extensive real-world recordings. By pre-training with a broad range of synthetic voices, we expect our model to generalize better to a variety of speaker profiles with minimal fine-tuning.

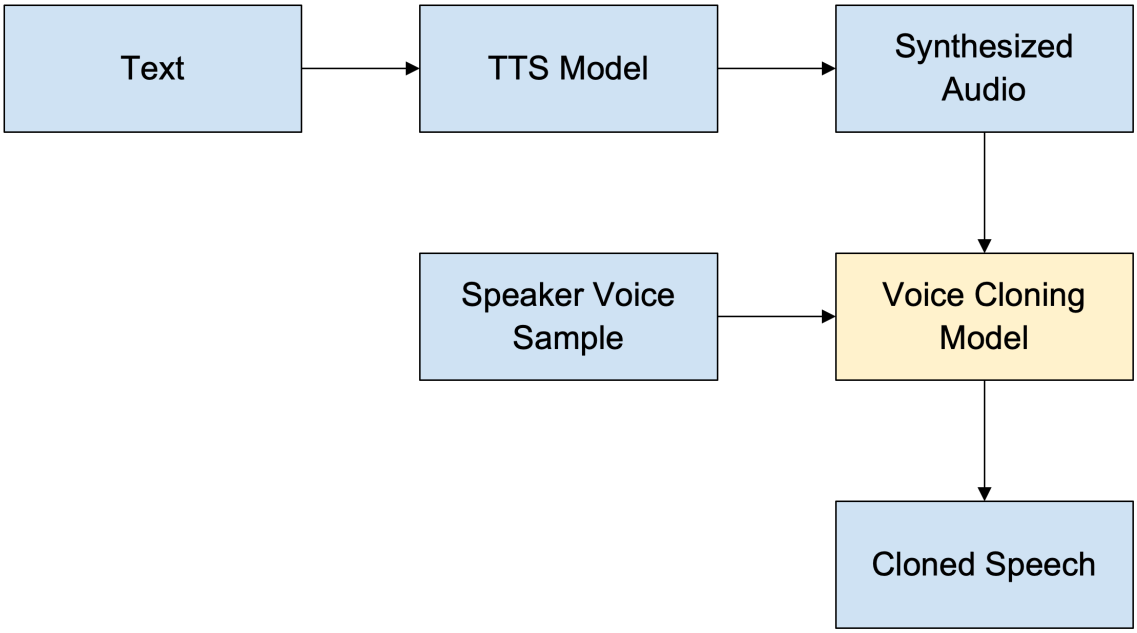

Figure 1: Architecture of a voice cloning system. The system this project is focused on is marked in yellow.

References

[1] Explaining Neural Scaling Laws https://arxiv.org/html/2102.06701v2

[2] FastSpeech: Fast, Robust and Controllable Text to Speech https://arxiv.org/pdf/1905.09263

[3] Wac2Vec: Unsupervised Pre-training For Speech Recognition https://arxiv.org/pdf/1904.05862

[4] Robust Speech Recognition via Large-Scale Weak Supervision https://cdn.openai.com/papers/whisper.pdf

[5] Unsupervised Speech Decomposition via Triple Information Bottleneck https://arxiv.org/pdf/2004.11284

[6] Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron https://arxiv.org/pdf/1803.09047

[7] On the Problem of Text-To-Speech Model Selection for Synthetic Data Generation in Automatic Speech Recognition https://arxiv.org/abs/2407.21476

[8] Attention Is All You Need https://arxiv.org/pdf/1706.03762

[9] SynthASR: Unlocking Synthetic Data for Speech Recognition https://arxiv.org/abs/2106.07803

[10] SYNT++: Utilizing Imperfect Synthetic Data to Improve Speech Recognition https://ieeexplore.ieee.org/abstract/document/9746217

[11] Generating Synthetic Audio Data for Attention-Based Speech Recognition Systems https://ieeexplore.ieee.org/abstract/document/9053008

[12] Wavenet: a generative model for raw audio https://arxiv.org/pdf/1609.03499

[13] Improving Speech Recognition using GAN-based Speech Synthesis and Contrastive Unspoken Text Selection http://www.interspeech2020.org/uploadfile/pdf/Mon-2-2-4.pdf

[14] Expressive Neural Voice Cloning https://proceedings.mlr.press/v157/neekhara21a/neekhara21a.pdf

[15] Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis https://arxiv.org/pdf/1806.04558

[16] Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning https://arxiv.org/abs/1907.04448

[17] Multilingual Speech Synthesis for Voice Cloning doi:10.1109/BigComp51126.2021.00067 https://ieeexplore.ieee.org/document/9373282

[18] Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers https://arxiv.org/pdf/2301.02111

[19] VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers https://arxiv.org/pdf/2406.05370

[20] OpenVoice: Versatile Instant Voice Cloning https://arxiv.org/pdf/2312.01479

[21] SynthASR: Unlocking Synthetic Data for Speech Recognition https://arxiv.org/pdf/2106.07803

[22] Tacotron: Towards End-To-End Speech Synthesis https://arxiv.org/pdf/1703.10135