Trevor Ashby

Project Proposal

Basic problem I'm trying to address, and which natural languages

The basic problem I am trying to address is the limited ability of a chatbot or service agent in expressing themselves through synthesized speech. I am attempting to improve their expressive abilities by developing a method or model that can leverage prosodic annotations to produce more natural and engaging text-to-speech synthesis. For example, altering rhythm, pitch, or emphasis either simultaneously or individually (TBD). I hypothesize that through this, we will begin to approach TTS systems capable of mimicking practices such as sarcasm, and enhanced storytelling. I plan to develop with English being the primary language but will experiment with additional languages after completing the entire pipeline.

Overall approach

1. Dataset Manipulation/Addition

   My overall approach is twofold. First, I will begin by collecting/manipulating TTS datasets to contain the desired features I wish to explore. The Wavelet prosody analyzer will allow for me to analyze and label my TTS datasets to provide me with data samples consisting of {text, waveform, prosody annotations} triples. These annotations will provide F0, energy, and duration of each word present in the waveform. The dataset now being in the form of the triples will conclude the first portion of my project. I expect this part of the project to be the most time intensive. Depending on the quality of the annotations, additional methods might need to be used to supplement the annotations to provide as much information as possible to the generative model.

2. Audio Generation

   Second, I want to develop a model (most likely following the Tacotron2 architecture) that can generate novel waveforms given both text and prosody annotations. These waveforms when converted to audio files (most likely using an architecture like that of WaveGlow), are intended to contain the spoken representations of the provided text, following the style and prosody of the provided annotations. Thus, annotations provided alongside text will represent a form of "style-transfer" that should occur in the Mel spectrogram.

What technologies

I am certain that a suite of technologies is going to be needed to complete this project. In this suite, I plan to include audio generation and processing technologies such as Tacotron2 and WaveGlow (as previously mentioned).

1. Tacotron2

   My plan for Tacotron2 is to serve as the generative model architecture when generating the Mel spectrograms. I plan to modify this architecture slightly to allow for the prosody annotations to be used as additional input.

2. WaveGlow

I plan to use an un-modified WaveGlow to convert my generated Mel spectrograms into functional audio.

What tools or toolkits

Obviously, I will plan to use PyTorch throughout this whole project along with additional packages such as numpy, matplotlib, etc. However, I have found some helpful tools with respect to step 1 in my overall approach. Two of these tools are listed below:

- Automatic detection of prosodic boundaries in spontaneous speech: Automatic detection of prosodic boundaries in spontaneous speech | PLOS ONE
- Wavelet Prosody Toolkit: asuni/wavelet_prosody_toolkit (github.com)

My hope is that using these tools I can properly extract and create some prosody annotations for all datasets that I choose to use. A concern that I have is that the annotations generated using these tools might not be sufficient for what I am hoping to do (i.e., certain prosody annotations might not get recognized using the tools), but it is a start, nonetheless.

What knowledge sources, corpora, and/or datasets

I have found a plethora of speech datasets with accompanying text. My plan is to use some samples from each in hopes that it'll add robustness to my system. Specifically, the M-AILABS dataset will be interesting to use, because it contains additional languages other than English. Once I have a working and testable pipeline, my plan is to gauge its performance with additional languages. A list of the datasets I'm currently planning on using are listed below:

- The LJ Speech Dataset: The LJ Speech Dataset (keithito.com)
- The M-AILABS Speech Dataset: The M-AILABS Speech Dataset – caito
- The Helsinki Prosody Corpus: Helsinki-NLP/prosody: Helsinki Prosody Corpus and A System for Predicting Prosodic Prominence from Text (github.com)

Evaluation

1. Prosody Annotation

   An evaluation that I am going to do will be checking the quality of prosody annotation between data samples. Are words properly annotated? Are the annotations consistent? Additionally, are the annotations complete?

2. Generated Audio

   I think that a human evaluation of the generated text with prosody manipulation vs. without would be a very necessary study to conduct. Do people enjoy the prosody manipulated text more? Does it seem more personal? On the other hand, is the modified audio even understandable? I also think an interesting evaluation would be unnatural manipulations (i.e., emphasizing each word). Another test that I think would be fun to see would be, can natural spoken language be manipulated in such a way that the audio sounds more like singing instead of simply talking?

3. Alternate Languages

   Finally, I would like to run both evaluation 1 and 2 (listed above) on different languages to see how versatile my pipeline is on languages other than English.