Natural Speech Synthesis using Modular Neural Network

Sajjad Hussain
BRAC University
Dhaka, Bangladesh
sajjad.hussain2@g.bracu.ac.bd

Abstract

This document addresses the prospect of developing modular neural networks that combines the power of various existing models related to speech processing, along with nuanced-phonetic transcriptors, to generate highly natural conversational speech. The framework for the model closely mimics the human brain.

1 Introduction

Speech synthesis has been around for quite some time and has been regularly used for tasks such as ChatBots or Read-Out-Loud features for the visually challenged. However, the fidelity of such synthesizers remains limited due to the monotone or robot-like nature of the generated voice. On the other hand, text generators have far more applications currently due to massive leaps in improvement in the field of neural networks. One may argue that these improvements in text generators essentially lead to improvement of speech synthesizers as the latter is implemented on the output of the former. However, it is still a far cry from proper conversational speech. This is because conversational speech has non-standard phonetic attributes, or nuances, that can come from dynamic factors such as context of the conversation or emotion of the speaker.

In this paper we will look at a novel prospect of using modular neural network for speech generation that derived from the way the different regions of the brain responsible different functions work together to carry out 'nuanced' conversation.

2 Literature Review

In 2016, a deep neural network for generating raw audio waveforms called WaveNet was proposed by van den Oord et al. (2016) based on Convolutional Neural Network(CNN). These waveforms include realistic breaths and lip smacks – but do not conform to any language.

In 2018, Li and Mandt (2018) stated that they were able to use WaveNet with DeepMind that allowed *content swapping*, where the voice on any given audio recording could be swapped for any other pre-existing voice while maintaining the text and other features from the original recording. This also made WaveNet capable of distinguishing between the spoken text and modes of delivery (modulation, speed, pitch, mood, etc.)

3 Modular Neural Network Framework

The task of 'nuanced' speech generation can be broken down into the following sub-tasks that can be carried out by individual neural network models:

- Prompt recognition -
  Automatic Speech Recognition (ASR)
- Expression or nuance analysis -
  Nuanced-phonetic Transcriptors
- Natural language processing -
  Language Models (LM)
- Response generation -
  Text To Speech (TTS)

The first key here is the expression or nuance analysis of the input speech to recognize the speaker's voice attributes such as excitement, urgency, sadness, etc.. Once analyzed, an appropriate *response tune* can be applied to the outgoing response.

The second key is to have these individual models *talk* to each other through interfacing mechanisms to further improve cohesion between the models. This may also encourage the models to generate intuitive casual speech if done efficiently.

4 Potential Challenges

One of the daunting task would be creating those appropriate interfaces between models in the most efficient way. This will most definitely require modifying the existing models themselves in order to accommodate the additional information.

Additionally, while phonetic transcriptors do exist, nuanced-phonetic transcriptors do not. In fact, they will probably have to be generated through unsupervised training since nuances are not standardized and can have millions of variations.

Perhaps the biggest challenge in this case would be acquiring the appropriately annotated corpus to train the model as standard speech data-sets would be useless in providing the enormous variety of nuances of the same words/sentences.

5 Conclusions

We have discussed the novel prospect of using various neural networks in a modular fashion to generate more human-like conversational speech that is derived from the architecture of the human brain. Potential use-cases for such models can range from Speech-tutoring for foreigners to audio therapy for Alzheimer's disease patients.

Success in the implementation of Modular Neural Networks may lead to advancement in performing more dynamic tasks in other fields of AI. Because it mimics the mechanism of the human brain, it has the potential to open the door to creation of a synthetic brain.

References

Yingzhen Li and Stephan Mandt. 2018. Disentangled sequential autoencoder.

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