**Research Proposal**

Generating speech which mimics human voice has always been a challenging problem in machine learning (ML). Current speech synthesizers are not able to express emotions. The goal for the summer research project is to develop a text-to-speech (TTS) system which can generate speech in accordance with the emotions expressed in the text.

TTS will comprise of two deep learning models- Recurrent Neural Network (RNN) and Generative Adversarial Network (GAN). RNN will be used for sentiment analysis of the text. The sentiment deduced will be used by GAN to generate speech. There are many ML algorithms for implementing sentiment analysis, but RNN will be the most suited model for TTS due to the variable length of text [1]. The pivot of the TTS is GAN, which comprises of a generator and its adversary differentiator [2] competing against each other. GAN utilizes a game theoretic approach by training on a minimax objective function[3]. After training GAN, the Generator will be able to learn the transformation from simple to complex and high dimensional training distribution. Thus, producing realistic samples. Since its inception, GANs have been used for generating images, but research on its application for generating speech has been limited. The current works on Speech synthesis using GAN can only mimic human voice to a certain extent but not the sentiments. My project will be incorporating recent research works on GAN for speech, most notably Speech Enhancement GAN (SEGAN)[4] which produces noise free human speech from samples of speech in different noisy environments. The research aim on developing a GAN model for incorporating emotions in speech. Research will be also focused on testing the proposed model with different objective loss functions such as L1, SoftMax cross entropy, JS divergence etc. After devising an optimum training mechanism for GAN which includes designing of generator, discriminator and loss function, it will be integrated with RNN. This will provide a fully functional TTS which will be able to express emotions while generating speech. TTS using GAN will have a significant edge over DeepMind’s state of the art WaveNet because GANs are designed to be able to generate results in parallel, yielding greater generation speed. Vectors representing a particular sentiment incorporated speech will be analysed using Interpretable vector math[5], in order to observe the connection between emotions and speech.

|  |  |
| --- | --- |
| Sub Tasks | Approximate Time |
| Data Pre-processing and Discriminator model | 1 week |
| Development of generator model | 1 week |
| Testing with different training objectives | 3 weeks |
| Sentiment Analysis using RNNs | 1 week |
| Testing | 1 week |
| Interpretable vector math for new insights | 1 week |

References

[1] Hong, J. and Fang, M., 2015. *Sentiment analysis with deeply learned distributed representations of variable length texts* (pp. 655-665). Technical report, Stanford University.

[2] Bengio, Y., Goodfellow, I.J. and Courville, A., 2015. Deep learning. *Nature*, *521*, pp.436-444.

[3] Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep learning*. MIT press.

[4] Pascual, S., Bonafonte, A. and Serrà, J., 2017. SEGAN: Speech Enhancement Generative Adversarial Network. *arXiv preprint arXiv:1703.09452*.

[5] Goodfellow, I., 2016. NIPS 2016 tutorial: Generative adversarial networks.

[6] Radford, A., Metz, L. and Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.