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 two adversaries- generator and differentiator [2] competing against each other. GAN utilizes a game theoretic optimization technique. Generator and differentiator are involved in a min-max game, thus enhancing the ability of the generator to fool the discriminator and produce realistic samples in accordance with the training dataset. Since its inception, GANs have been used for generating images [3], but research on its application for generating speech has been limited. The current works on Speech synthesis using GAN can only mimic human voice but not the sentiment [4]. But recent research work on Speech Enhancement GAN (SEGAN) [5], utilized 20 different noise conditions and was able to successfully generate a noise-free human voice using the generator. The research will be focussed in this direction, with an aim of developing a SEGAN inspired GAN model. Alterations in the SEGAN’s model will be required, for example, noise condition samples will be replaced by speech samples with different emotions for training GAN. After implementing the GAN, research will be focussed on the loss function which is integral to the success of GAN. There exists a spectrum of loss functions [6] and finding the suitable function requires a significant amount of testing. 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 the current state of the art such DeepMind’s WaveNet.

Expected time for completion of research and development for the TTS is eight weeks. First two weeks will be dedicated to data pre-processing and development of the discriminator. Three weeks will be required for the development of the generator and research on the loss function. Integration of the generator with differentiator and training will require a week. So, by the sixth week, a GAN which can generate speech using the sentiments (using the training data) will be operational. Development of RNN model and integration with GAN will require another week. Last week will be for testing and presentation of the research work.

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.

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

[3] Yang, S., Xie, L., Chen, X., Lou, X., Huang, D. and Li, H., 2017. Statistical Parametric Speech Synthesis Using Generative Adversarial Networks Under A Multi-task Learning Framework. *arXiv preprint arXiv:1707.01670*.

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

[5] Arjovsky, M. and Bottou, L., 2017. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*.