UCS749: Conversational AI: Speech Processing & Synthesis

L	T	P	Cr
2	0	2	3.0

Course Objectives: This course will provide students with the overall structure of the Conversational AI pipeline including Speech Processing, Recognition, and Synthesis and building end to end workflows using NeMo and Jarvis SDK.

Introduction: Fundamentals of Speech Processing, Applications of Speech Processing and Deploying NLP, ASR and TTS modules in Jarvis.

Fundamentals of Speech Processing: Introduction to Statistical Speech Processing, HMMs for Acoustic Modeling, WFTS for Automatic Speech Recognition (ASR), Basics of Speech Production, Tied State HMMs, Introduction to NNs in Acoustic Modeling (Hybrid/TDNN/Tandem).
{Papers}

Automatics Speech Recognition (ASR): ASR - DNN models (Jasper, QuartzNet, Citrinet, Conformer-CTC), Open-source Datasets, Language Modelling: N-Gram, Neural Rescoring.
{Survey , Jasper, QuartzNet, CitriNet , Nemo}

Applications of Speech Processing: Speech Commands: Speech Commands Recognition using MatchboxNet. Overview of Noise Augmentation, Voice Activity Recognition and Speaker Recognition.
{Survey, Nemo}

Speech Synthesis: Text Normalization: Preparing Dataset and Text Normalization for input to Speech Synthesis model. Introduction to Text-to-Speech (TTS) Models:- Mel Spectrogram Generator: - Tacotron-2, Glow-TTS, Audio Generators:- WaveGlow, SqueezeWave.
{Papers, Nemo}

Jarvis Deployment: Introduction to Jarvis, Overview of Jarvis ASR, NLU and TTS APIs, Introduction to Jarvis Dialog Manager. Jarvis Deployment:- Nemo model deployment for ASR, NLP and TTS.

Laboratory Work:

- Practical Exercise on Statistical Speech Processing. {Traditional Signal Processing}
- Automatic Speech recognition with NeMo on English Dataset.
- Automatic Speech recognition with NeMo on Indic Language(Hindi) Dataset.
- NeMo Speech Commands Recognition using MatchboxNet, Noise Augmentation, and Speaker Recognition.
- Text to Speech using Tacotron-2 and WaveGlow with NeMo on English Dataset.
- Text to Speech using Tacotron-2 and WaveGlow with NeMo on Indic Language (Hindi) Dataset.

- End-to-End Conversational AI Model (Any Language): ASR/NLP/TTS with NeMo and Jarvis.

Course Learning Outcomes (CLOs) / Course Objectives (COs):

After the completion of the course the student will be able to:

1. Understand Speech Processing pipeline for various applications with accelerated computing. ASR, Speaker Recognition etc.
2. Exploring Speech Synthesis on various data sets.
3. Deep practical hands-on experience from training to deployment of these applications using NVIDIA GPUs and Toolkits:- NeMo, and Jarvis.

Text Books:

4. Jurafsky, Dan and Martin, James, Speech and Language Processing, *Second Edition, Prentice Hall, 2008.*
5. Daniel Jurafsky and James H. Martin, "Speech and Language Processing", 3rd edition draft, 2019 [JM-2019].
6. Mark Gales and Steve Young, The application of hidden Markov models in speech recognition, Foundations and Trends in Signal Processing, 1(3):195-304, 2008.

Reference Books:

4. Daniel Jurafsky and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.
5. Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine, 29(6):82-97, 2012.