**PROJECT PROPOSAL**

|  |
| --- |
| **Date of proposal:**  **01/10/2022** |
| **Project Title:**  Intelligent Speech Generation System |
| **Group ID (As Enrolled in LumiNUS Class Groups):**  **Group 3**  **Group Members (name, Student ID):**  **Ma Xin A0261775W**  **Zhang Zhimu**  **Sun Wenyuan**  **Wang Zhiyuan A0261827Y** |
| **Sponsor/Client:** *(Company Name, Address and Contact Name, Email, if any)*  *No* |
| **Background/Aims/Objectives:**  **Background**  Due to the rapid development of new media and personal video creation, the speech generation system has become more significant to assist new media operators. For example, the people who want to dub for videos they have produced sometimes need to use a speech generation system, transferring their text to speech. Besides, a speech generation system can generate a particular speech obtained from some celebrity with a beautiful voice, which can enhance the attraction of created video. In addition, a speech generation system is also necessary for the network broadcast field. Sometimes, an anchor needs to interact with audiences through the bullet screen. But if the anchor does not notice what audiences express on the bullet screen, the speech generation system can transform text on the bullet screen to voice to remind the anchor, which is quite convenient and effective. Therefore, based on its potential business value, we try to construct a speech generation system, which also satisfies the requirement for the PRS practice module.  **Aims**  Design and generate an intelligent speech generation system, transferring various text datasets to specific speeches to satisfy a series of usage requirements.    **Objectives**  Determine the practice module requirements.  Carry out a literature review.  Capture ideal text datasets from online text libraries from different fields.  Design the front end to interact with users.  Construct a speech generation system algorithm to generate speech from texts.  Design the back end to achieve data interaction.  Integrate the front end, system algorithm, and back end.  Test and improve the system. |
| **Project Descriptions:**  The project will consist of a frontend and a backend with a well-trained model inside the back end.  For the frontend, we will provide a text input box where the user can input the text which he wants the speech to generate from, and a voice selector where the user can choose a voice at his own preference. Once the user has done his selection and text input, the frontend will interact with the backend and play the generated speech.  As for the backend, which is designed to get the text input together with the voice selected information send from the frontend, will process the text into speech through inference of the model and then return the generated speech to the frontend.  Our TTS model is based on VAE and GAN, which is a hybrid model called VITS proposed in the article Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. The model mainly consists of a posterior encoder, prior encoder, decoder, discriminator, and stochastic duration predictor. For the training procedure, we use the linear spectrogram obtained from raw audio waveforms through the STFT as the input of the posterior encoder to train the encoder and decoder, which make up to be the VAE. The phonemes of the input texts will be encoded by the prior encoder, and then we will proceed normalizing flow to the encoded phonemes and the latent variables of VAE to do the monotonic alignments search (MAS), which will optimize the evidence lower bound of the model. The stochastic duration predictor calculates the duration token output from the MAS procedure to train a deterministic duration predictor which helps generate human-like rhythms of speech. And the discriminator is designed to enhance the output generated by decoder, it distinguishes between the decoder output and the ground truth waveform in order to make the generated output more natural and less mechanical.  The discriminator and the posterior encoder will not appear in the inference procedure, as when inferencing, we just feed the extracted phonemes to the prior encoder and by the efforts of stochastic duration predictor and the well-trained decoder, the waveform of speech will be generated and be output. |