text-to-Speech and speech-to-text using Machine LEarning (SMILE)

GOALS

- The primary goal was to build and train a Text-to-Speech (TTS) Machine Learning (ML) model with our own professionally actor generated audio tour content
- Our secondary goal was to build a service that utilises the open source DeepSpeech framework, which is a Speech-To-Text engine based on a model trained by the recurrent neural network (RNN) machine learning technique using a large library of voice data.

CHALLENGES

- Recording Content with Human Actors in recording facilities, etc has proven to be inflexible and expensive
- Inability to Easily Modify Content after creation. The digital audio content is very difficult, time consuming, and costly to change or update
- Providing Interactive Content to Tourist Users, such as answering questions in real-time over Wi-Fi/4G/5G networks is impossible.

DEMO SETUP

- We utilised the Iris testbed with 3 x VMs for the experiments.
- We also used two GPU laptops to train the TTS models.
- All necessary libraries were installed here for data processing.
- We developed tools to support processing audio data into a suitable for Text-to-Speech (TTS) training
- It was very easy to install and use the Speech-to-Text framework.

RESULTS - Obj 2

The instructions in this screenshot show how to process WAV audio files using the DeepSpeech - (Speech to Text) framework on the Iris testbed. This tool proved very useful during the project supporting Objective 1.

MORE RESULTS - Obj 1

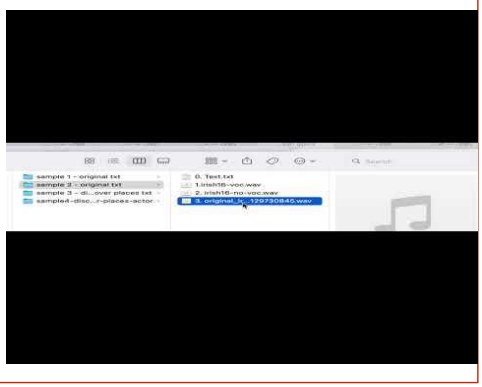
The following YouTube Video shows two main things: (https://youtu.be/AmX3z5amPvY):

The first part shows how to create text to speech using the SMILE trained models

The second part of the video shows examples of the audio content produced by the Text-to-speech models trained during the SMILE project

- Sample 1, Sample 2, and Sample 3 were trained using the Irish_english_male: irm_03397 dataset from the crowdsourced high-quality UK and Ireland English Dialect speech data sets
- Sample 4 was created using the discover places real actor dataset

Results (MOS Score 2.45):We proposed using the mean opinion score (MOS) test (score 1(bad)-5(excellent)) to quantify the quality of the ML models trained. The MOS test is used in the quality of experience domain and can be employed anywhere human subjective experience and opinion is useful.



CONCLUSIONS

We found it very difficult to train the Text-To-Speech models (GLow-TTS, and Multiband-melgan) using the THNP dataset. We edited the dataset many times, and restarted GPU training, with minimal success. In our non-expert linguistics opinion, this was due to the actor not pronouncing all letters and words clearly for all text (Hiberno-English).

Large datasets (with 30,000+ WAV files) are required to get good TTS models.

Recordings need to be professionally completed, and actors need to enunciate every word clearly.

POST MORTEM

We believe that training new research models in this space might be interesting to pursue in future work.

To meet our short-term business objectives, we will utilise commercial TTS tools developed by companies such as Amazon, which are excellent.

As company we learned a lot about Machine Learning, and open source libraries available in this space, which we intend to utilise in future Discover Places products.