PES UNIVERSITY

UE17CS490A – Capstone Project Phase – 1

End Semester Assessment

Project Title : Voice Cloning Using Deep Learning
Project ID : PW21VRB06
Project Guide : Mr. V.R. Badri Prasad
Project Team : 1511_1523_1551

Agenda

- Problem Statement
- Abstract and Scope
- Literature Survey
- Suggestions from Review – 3
- Design Approach
- Design Constraints, Assumptions & Dependencies
- Proposed Methodology / Approach
- Architecture
- Design Description
- Technologies Used
- Project Progress
- References

Problem Statement

- Designing an artificially intelligent system that learns to mimic a person's voice by analyzing speech recordings and the corresponding text transcripts.

- The system will be developed using the techniques of text encoding followed by Convolutional and Transposed Convolutional neural networks and a Mel to Audio converting algorithm i.e, present a neural text-to-speech model that learns to synthesize speech directly from (text, audio) pairs in customized voices.

Abstract and Scope

- Voice cloning is the replication or creation of a person's voice using Deep Neural Networks. This is a speech synthesis project as it involves producing human speech artificially. The goal of this project is to build a Customized Text to Speech model which can generate natural speech for a variety of speakers in a efficient manner.

- In this project, we try to implement a neural network model for speech synthesis directly from text. We try to implement a Convolutional and Transposed Convolutional neural network model. The network architecture will also consist of Recurrent Neural Network that will be used for sequence to sequence feature prediction.

Literature Survey - Paper 1

[1] A¨aron van den Oord Sander Dieleman Heiga Zen Karen Simonyan Oriol Vinyals Alex Graves Nal Kalchbrenner Andrew Senior Koray Kavukcuoglu

WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

arXiv:1609.03499v2 [cs.SD] 19 Sep 2016

- WaveNet is a deep neural network for generating raw audio waveforms which uses probabilistic and autoregressive methods combined with convolutional layers.

- To deal with long-range temporal dependencies needed for raw audio generation, they use new architectures based on dilated convolutions, which exhibit very large receptive fields.

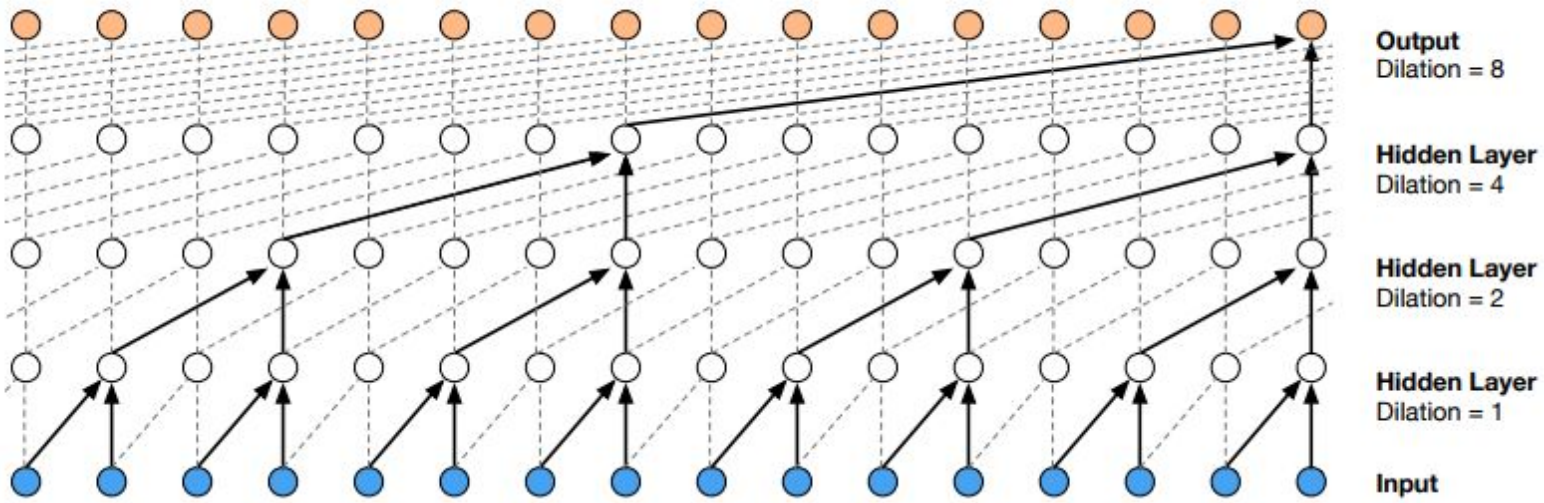



Literature Survey

- WaveNets provide a generic and flexible framework for tackling many applications that rely on audio generation such as TTS, music, speech enhancement, voice conversion, source separation.
- There are no pooling layers in the network, and the output of the model has the same time dimensionality as the input.
- For images, the equivalent of a causal convolution is a masked convolution which can be implemented by constructing a masked filter and doing an elementwise multiplication of this mask with the convolution kernel before applying it. For 1-D data such as audio we can more easily implement this by shifting the output of a normal convolution by a few timesteps.
- Since models with causal convolutions do not have recurrent connections, they are typically faster to train than RNNs, especially when applied to very long sequences of data.

Literature Survey

- At training time, the conditional predictions for all timesteps can be made in parallel because all timesteps of ground truth x are known.
- When generating with the model, the predictions are sequential because after each sample is predicted, it is fed back into the network to predict the next sample.






Literature Survey

- Stacked dilated convolutions enable networks to have very large receptive fields with just a few layers, while preserving the input resolution throughout the network as well as computational efficiency.

- Raw audio is typically stored as a sequence of 16-bit integer values (one per timestep), a softmax layer would need to output 65,536 probabilities per timestep to model all possible values which is computationally expensive.

- Therefore, a μ-law companding transformation is applied to the data to quantize it to 256 possible values and the reconstructed signal after quantization sounded very similar to the original.

- Both residual and parameterised skip connections are used throughout the network to speed up convergence and enable training of much deeper model.




Literature Survey

- They conditioned the model on text inputs in two different ways: global conditioning and local conditioning.
- Global conditioning preserves speaker embeddings in a TTS model whereas Local conditioning preserves linguistic features in a TTS model.
- They used the North American English dataset which contains 24.6 hours of speech data, and the Mandarin Chinese dataset which contains 34.8 hours both of which were spoken by professional female speakers.
- However, WaveNets achieved 5-scale MOSs in naturalness above 4.0, which were significantly better than the then existing methods of Text to Speech systems.




Literature Survey - Paper 2

[2] Hideyuki Tachibana, Katsuya Uenoyama and Shunsuke Aihara.

Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention.

arXiv:1710.08969v1 [cs.SD] 24 Oct 2017

- This paper describes a novel text-to-speech technique based on deep convolutional neural networks, as well as a technique to train the attention module rapidly.

- This paper describes a text-to-speech (TTS) technique based on deep convolutional neural networks (CNN), without any recurrent units.




Literature Survey

- The objective of this paper is to show an alternative neural TTS system, based only on CNN, that can alleviate these economic costs of training.
- The proposed Deep Convolutional TTS can be sufficiently trained only in a night (~15 hours), using an ordinary gaming PC equipped with two GPUs, while the quality of the synthesized speech was almost acceptable.
- The purpose of this paper is to show Deep Convolutional TTS (DCTTS), which is fully convolutional.
- An audio waveform can be mutually converted to a complex spectrogram by linear maps called STFT and inverse STFT.




Literature Survey

- DCTTS model consists of two networks: Text2Mel and Spectrogram Super-resolution Network (SSRN).
- The Text2Mel model consists of four submodules: Text Encoder, Audio Encoder, Attention and Audio Decoder. The objective of this module is to synthesize a mel spectrogram from an input text.
- The second module Spectrogram Super-resolution Network (SSRN) converts a coarse mel spectrogram to the full STFT spectrogram.
- The results are quite satisfying. The use of attention module, synthesized mel and full spectrograms shows that the method can almost work correctly can synthesize quite clear spectrograms.




Literature Survey

- The audio quality generated by the model is far from perfect this can be improved by tuning some of the hyper-parameters.
- This simple TTS model can be extended to other versatile purposes, such as emotional/non-linguistic/personalized speech synthesis, singing voice synthesis, music synthesis.
- To train the networks, they have used LJ Speech Dataset [33], a public domain speech dataset consisting of ~13K pairs of text & speech, ~24 hours in total.
- MOS (95% confidence interval) was 2.71±0.66 (15 hours training)




Literature Survey - Paper 3

[3] YuxuanWang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weissy , Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengioy, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous.

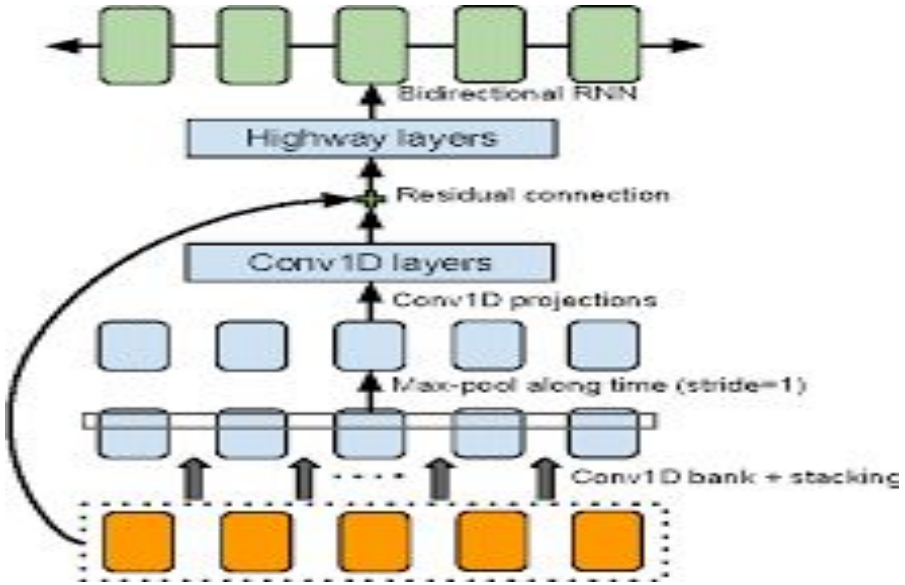
TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS.

arXiv:1703.10135v2 [cs.CL] 6 Apr 2017

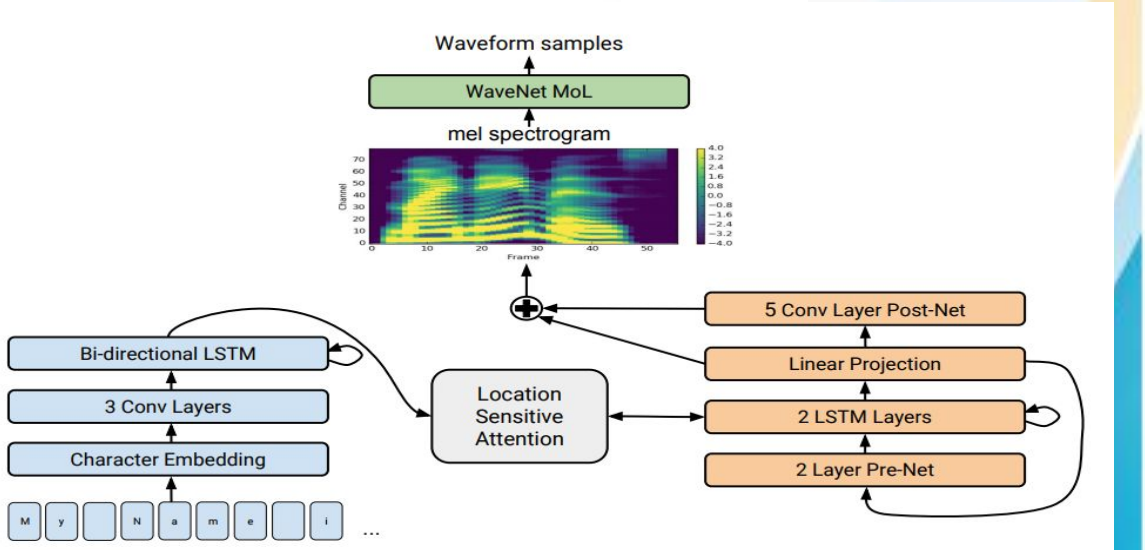
- Tacotron proposes an end-to-end generative TTS model based on the sequence-to-sequence with attention paradigm.
- A vanilla seq2seq model does not work well for character-level inputs.
- At a high-level, the model proposed in this paper takes characters as input and produces spectrogram frames, which are converted to waveforms.

Literature Survey

- The tacotron model is a seq2seq model with an attention module. It includes an encoder, an attention-based decoder, and a post-processing net.



CBHG Module of Tacotron



Architecture of Tacotron




Literature Survey

- Tacotron uses North American English dataset, which contains about 24.6 hours of speech data spoken by a professional female speaker.
- In the architecture the first module is CBHG. It stands for 1-D convolutional filters, Highway Network and a bidirectional gated recurrent unit (GRU) recurrent neural net (RNN). CBHG is a powerful module for extracting representations from sequences.
- The output from the 1-D convolutional is fed into the Multi-layer highway network to extract high-level features.




Literature Survey

- The last layer of this CBHG module is the bidirectional GRU RNN. This final layer is to extract sequential features from both forward and backward context.
- The next module of the architecture is an encoder. The goal of the encoder is to extract robust sequential representations of text.
- The input to the encoder is a character sequence, where each character is represented as a one-hot vector and embedded into a continuous vector.
- In this encoder a set of non-linear transformations called “pre-net” is applied to each embedding.
- A CBHG module transforms the prenet outputs into the final encoder representation used by the attention module.

Literature Survey

- The final layer is the decoder. The content-based tanh attention decoder is used in this layer.
- The output of the attention decoder is passed onto a post processing net which learns to predict the mel spectrogram which is also a CBHG module. The mel spectrogram is then passed on to the Griffin Lim algorithm to generate the audio waveform.
- Tacotron achieves a 3.82 MOS score on US English.
- Many aspects of the model are yet to be investigated. Many early design decisions have gone unchanged. The output layer, attention module, loss function are all up for improvement.





Literature Survey - Paper 4

[4] Jonathan Shen1, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly,Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis and Yonghui Wu.

TACOTRON-2 : NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS.

arXiv:1712.05884v2 [cs.CL] 16 Feb 2018

- It is an AI-powered speech synthesis system that can convert text to speech.
- Tacotron 2's neural network architecture synthesises speech directly from text. It functions based on the combination of convolutional neural network (CNN) and recurrent neural network (RNN).
- The sequence-to-sequence model that generates mel spectrograms has been borrowed from Tacotron.




Literature Survey

- The generative model synthesising time domain waveforms from the generated spectrograms has been borrowed from WaveNet.
- WaveNet is an audio generative model. It takes a sequence of audio samples as input and predicts the most likely following audio sample. However WaveNet is not an end-to-end TTS model.
- Tacotron 2 is made is up of three major components : Tacotron-made mel-spectrogram + WaveNet Vocoder – Griffin-Lim Algorithm.
- The Griffin-Lim Algorithm first appeared in Tacotron 1. It iteratively attempts to find the waveform whose STFT magnitude is closest to the generated spectrogram.




Literature Survey

- The first part of this model is an Encoder which converts the character sequence into word embedding vector.
- The embedding vector representation is later consumed by the Decoder to predict spectrograms.
- The decoder is an autoregressive recurrent neural network which predicts a mel spectrogram from the encoded input sequence one frame at a time.
- Tacotron 2 can also effortlessly distinguish between the meanings of heteronyms and pronounce them based on the usage.
- The researchers claim that the system can be trained directly from data without relying on complex feature engineering.

Literature Survey

- In addition to architectural differences, the important bit is that Tacotron2 uses Wavenet instead of Griffin-Lim to get back the audio signal which makes for very realistic sounding speech.
- The loss function is the Summed mean squared error (MSE).
- The loss is typically computing the residual loss between the model's predictions and the ground truth and returning the absolute value as is in Tacotron 1 and the loss in Tacotron 2 however squares this error for each sample instead of simply returning the difference loss.
- Tactron 2 has an achieved a MOS of 4.53.
- This system can be trained directly from data without relying on complex feature engineering, and achieves state-of-the-art sound quality close to that of natural human speech.




Literature Survey - Paper 5

[5] Sean Vasquez Mike Lewis
MelNet: A Generative Model for Audio in the Frequency Domain
arXiv:1906.01083v1 [eess.AS] 4 June 2019

- MelNet, a generative model for spectral representations of audio.
- MelNet models spectrograms, which are time-frequency representations of audio.
- This generative model for audio can capture longer-range dependencies than existing end-to-end models.
- MelNet aims to model the frequency content of an audio signal.
  Existing generative models for audio have aimed to directly model time-domain waveforms




Literature Survey

- A spectrogram is a two-dimensional time frequency representations which contain information about how the frequency content of an audio signal varies through time.
- spectrogram models can generate speech and music samples with consistency over multiple seconds
- Producing high-fidelity audio has been challenging for existing spectrogram models.
- To reduce information loss, high-resolution spectrograms are modelled.




Literature Survey

- Time-frequency representations highlight how the tones and pitches within an audio signal vary through time.
- To further align the representations with human perception, the frequency axis of the spectrogram is transformed to the Mel scale.




Literature Survey

- To model the distribution over spectrograms, they have devised a highly expressive model which synthesizes many recent advances in autoregressive modelling.
- Autoregression is a time series model that uses observations from previous time steps and decay factor as an input to the next time step.
- MelNet estimates a distribution element-by-element over the time and frequency dimensions of a spectrogram.
- One drawback of autoregressive models is that they tend to learn local structure much better than global structure.




Literature Survey

- The first is a simple time-major ordering which proceeds through each spectrogram frame from low to high frequency, before progressing to the next frame.
- The second is a multiscale ordering.
- A multiscale model which generates spectrograms in a coarse-to-fine order is used.




Literature Survey

- A low-resolution, subsampled spectrogram that captures high-level structure is generated initially, followed by an iterative upsampling procedure that adds high-resolution details.
- By generating spectrograms in this manner, it is possible to decouple the tasks of learning local and global structure.
- In comparison to previous works which model time-domain signals directly, MelNet is particularly well-suited to model long-range temporal dependencies.
- MelNet combines a highly expressive autoregressive model with a multiscale modelling scheme to generate high-resolution spectrograms with realistic structure on both local and global scales.




Literature Survey

They have trained MelNet to generate audio unconditionally using three diverse datasets:

- **Music:** they utilize the MAESTRO dataset, which consists of over 172 hours of solo piano performances.
- **Single-Speaker:** A speech dataset [Blizzard 2013 dataset] consisting of a single speaker reading audiobooks in a quiet environment.
- **Multi-Speaker:** A multi-speaker, multilingual speech dataset [VoxCeleb2 dataset] which contains speech from speakers of 145 different nationalities, covering a wide range of accents, ages, ethnicities and languages.

Suggestions from Review - 3

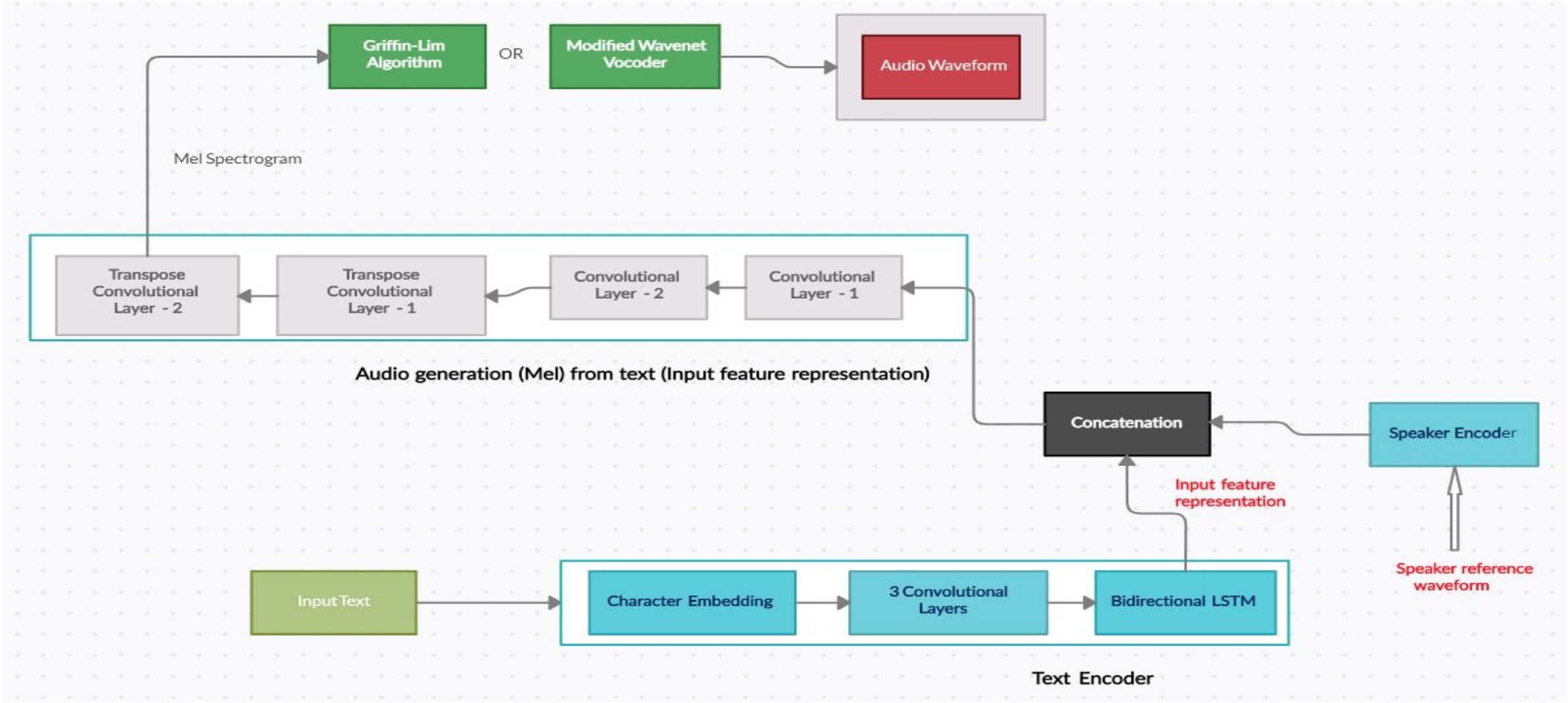
- Some changes in the content of Abstract and Introduction parts.

- To provide more details on the design architecture proposed.

Proposed Methodology / Approach

- The major design aspect of our model approach is the usage of Encoders and Convolutional and Transposed Convolutional Neural Networks.
- The proposed method may be efficient compared to other models as it includes Convolutional layers which doesn't need as much time to train compared to RNNs.
- The major advantage of using this is that without any human supervision the model will be able to learn complex, important and distinct features.
- The extraction of complex features and feature engineering is done automatically and most of the complex feature fedelations of speakers from datasets are combined to improve different possible words in feature generation.

Design Architecture



Design Explanation

- The Input text is first fed into an Encoder part where it is converted into one hot encoders and appended to the continuous vector which is passed on to a series of 3 convolutional layers followed by Bi-directional LSTM, to get the input feature representation which will have the linguistic features along with phonetic information.

- Another pre-trained encoded feature representation of several different possible phonemes and input text features from several available English datasets and is concatenated to the input features of the speaker in context.

- Following which is a series of Convolutional and Transposed Convolutional Layers which converts the input feature representation to Mel spectrogram based on the features it would have acquired from the training involved before.

- Now, the task of converting Mel spectrogram to raw audio format involves a bit of phase loss, but still can produce high quality audio using existing algorithms like Griffin-Lim Algorithm or Modified Wavenet Vocoder, after which we will have the final output of the audio corresponding to the speaker's voice features.

Design Constraints, Assumptions & Dependencies

- **Constraints** : To build and train the model we will depend on English language. This is a major constraint as our model will be only able to produce English speech and read text in English. The model will not be trained for other languages except english.

- **Hardware or software environment :** Since we are training our model on a complex and large audio dataset we need a high end machine with a significant amount of GPU RAM.

- **End-user environment :** Since our project is more focused on developing an efficient and accurate model in which the output should sound like an actual human voice, hence the scope for the end-user environment is minimal. Once the development is completed a basic end-user environment will be developed.

Design Details

1. **Novelty** : We are using transposed Convolutional Neural Network and Convolutional Neural Network.
2. **Innovativeness** : In comparison to various existing models none of the models have used Transposed CNN in their architecture which is a main component in our proposed model.
3. **Performance** : The performance of the model depends on the perception of the end user and the speed of the model prediction.
4. **Reliability** : The model will be able to perform well after an extensive training along with the best hyperparameters.
5. **Reusability** : Once the model is trained on a good dataset and all the hyper parameters of the neural network are tweaked and the best parameters and best design is chosen, the model can be improved further.
6. **Application compatibility** : The model needs an high end machine only for training, the trained model can be deployed on any platform.
7. **Resource utilization** : We need a high end machine to train the model and memory to store huge data sets.

Design Description

1. **Swimlane Diagram**

   The swimlane diagram is shown in the next slide.

2. **User Interface Diagrams**

   Our project is more focused on developing an efficient and accurate model in which the output should sound like an actual human voice, hence the scope for the end-user environment is minimal. Once the development is completed a basic end-user environment will be developed.

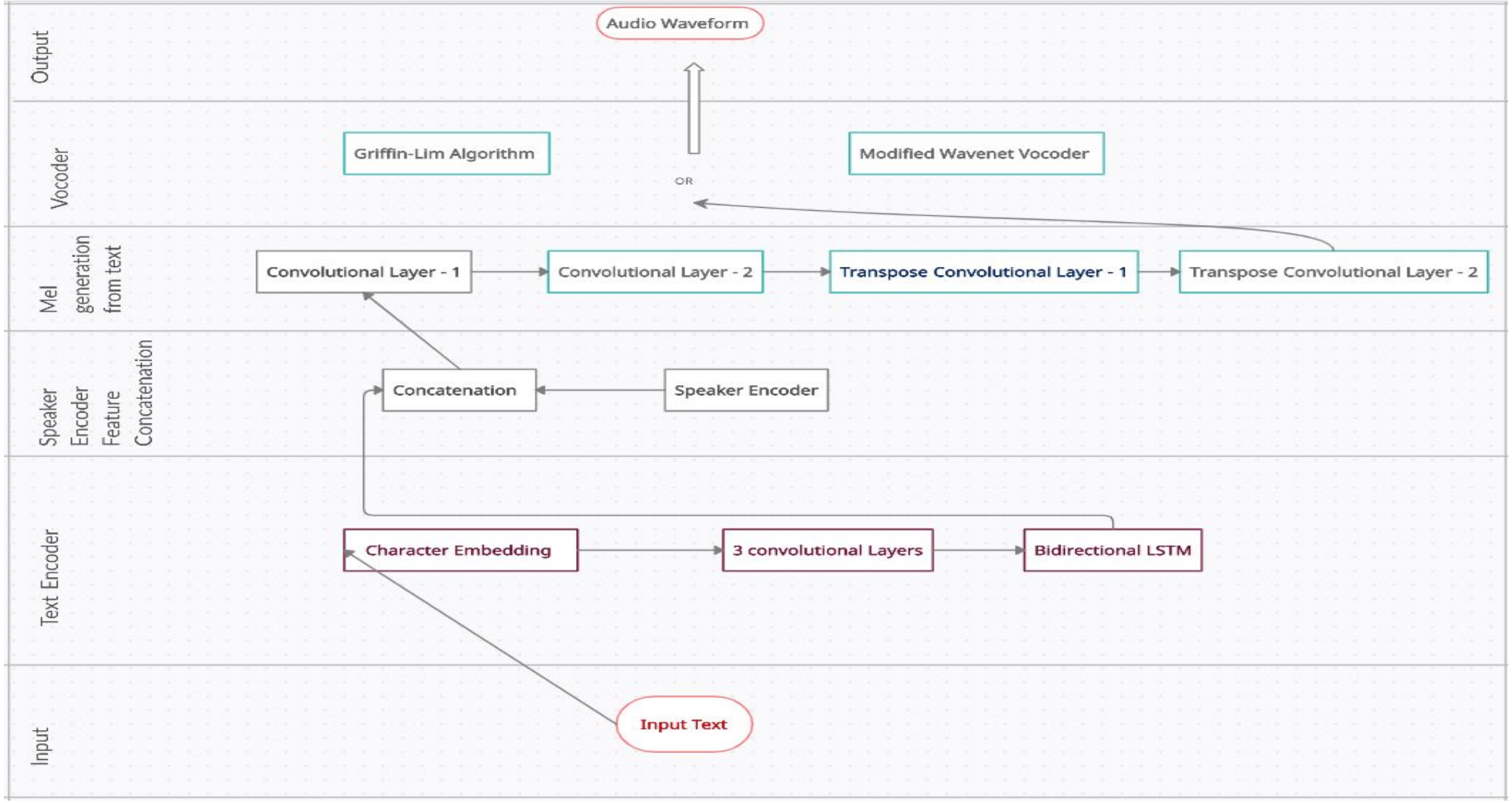
3. **External Interfaces**

   We plan to use the model with a flask application and the user can use a web interface as a platform to interact with our tool.

4. **Packaging and Deployment**

   Since the model is trained on Google Colab, for prediction we may require the usage of GPU, so we are yet to plan on the packaging part which will be decided accordingly in phase-2.

Swimlane Diagram



Datasets Used

These are the datasets that we will be using to train our model.

All Datasets contain English US Language only and middle aged.

- https://keithito.com/LJ-Speech-Dataset/ *(Major)*
- https://commonvoice.mozilla.org/en/datasets
- http://www.festvox.org/cmu_arctic/
- https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset

Technologies Used

These are the technologies which we have used and likely to use in the next phase.

- Python
- Tensorflow
- Pytorch
- Keras
- Librosa
- Numpy
- Matplotlib
- Since a highly efficient and powerful computing machine is required, we are planning to use Google Colab which offers free GPU access with a considerable amount of RAM and also GPU machines provided from the College.

Project Progress

- Phase-1 of the project was successfully completed according to the plans.
- We have studied in depth of the audio features and also successfully completed the research and literature survey of the various existing models and found out the differences between them.
- We have collected relevant datasets for this project and have also completed the initial part of the data preprocessing. The high level design of the model is completed.
- We have also started the coding part where we have analyzed different audio features and also generated Mel spectrograms.

Project Progress

- The major part of the Capstone Phase-2 of our project is the coding and implementation of the architecture design along with some modifications in the design if required.
- The main part is the training of the model with tweaking of the hyperparameters to fine tune the model, the time required and system specifications required itself is a lot. It will take days to train with several checkpoints in between to retain the trained intermediate models.
- We will also do validation and testing, followed by evaluation of the model like MOS.

References

[1] Aaron van den Oord, Sander Dieleman,Wavenet: A generative model for raw audio. arXiv, Sep 2016 - [1609.03499] WaveNet: A Generative Model for Raw Audio.

[2] Hideyuki Tachibana, Katsuya Uenoyama and Shunsuke Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. arXiv:1710.08969v1 [cs.SD] 24 Oct 2017

https://arxiv.org/abs/1710.08969

[3] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weissy, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengioy, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous. TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS - https://arxiv.org/abs/1703.10135

[4] Jonathan Shen1, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly,Zongheng Yang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis and Yonghui Wu. TACOTRON-2 : NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS. https://ieeexplore.ieee.org/abstract/document/8461368

[5] Sean Vasquez, Mike Lewis, Melnet : A Generative Model for Audio in the Frequency Domain ,June 2019 - https://arxiv.org/pdf/1906.01083.pdf

References

[6] Audio Data | Audio/Voice Data analysis Using Deep Learning

[7] Generating Celebrity Voices & Music

[8] Deepest-Project/MelNet: Implementation of "MelNet: A Generative Model for Audio in the Frequency Domain"

[9] Audio Data Analysis Using Deep Learning with Python (Part 1)

[10] Audio samples from "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis"

[11] Voice Cloning Using Deep Learning | by Mohit Saini | The Research Nest

[12] You can now speak using someone else's voice with Deep Learning

[13] Audio Data Analysis Using Deep Learning with Python (Part 1)

[14] Getting to Know the Mel Spectrogram | by Dalya Gartzman

[15] https://pnsn.org/spectrograms/what-is-a-spectrogram

[16] https://sjvasquez.github.io/blog/melnet/

Thank You