# Voice Cloning

*A B.Tech Project Report Submitted*
*in Partial Fulfillment of the Requirements*
*for the Degree of*

Bachelor of Technology

*by*

**Chinmaya K**
(170121013)

**DEPARTMENT OF PHYSICS**
**INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI**
**GUWAHATI - 781039, ASSAM**

April 2021

# CERTIFICATE

*This is to certify that the work contained in the project report titled "**Voice Cloning**" by **Chinmaya K (Roll No. 170121013)**has been carried out under my supervision in the Department of Electrical and Electronics and Engineering, and that this work has not been submitted elsewhere for a degree.*

Signature of Supervisor

Name:

Department:

Date:

# DECLARATION

*This is to declare that the project report titled* ***Voice Cloning*** *submitted by me to the department of Physics, Indian Institute of Technology Guwahati, for the partial fulfilment of the requirement for the degree of bachelor of technology is a bonafide work carried out by me under the supervision of* ***Dr.Prithwijit Guha****. The content of this work, in full or in parts, has not been submitted elsewhere for the award of any other degree or diploma. I also declare that this report is based on my personal study and/or research and I have acknowledged all materials and resources used in its preparation.*

________________________________

Signature of Supervisor

Name:

Roll no:

Department:

Date:

# Acknowledgements

Deep Learning and its related fields is an exciting and increasingly relevant field in our present time. Therefore I feel fortunate to have been doing my BTP under esteemed Professor Dr.Prithwijit Guha and I would like to express my grattitude for giving me an opportunity to work under his guidance. Also his Research group in Multimedia Analytics Lab is always brimming with activity and involvement. I am greatly indebted to them for their invaluable advice and support.

I also express my gratitude towards Prof. Perumal Algaraswamy (HOD, Physics department) and Sayan Chakrabarti (DUPC secretary) for giving me the opportunity to present this project.

# Abstract

Artificial production of human speech is known as speech synthesis,also known as text-to-speech (TTS).The main task here is to convert text input to speech output. This technology utilizes works from various major fields(broadly speaking) such as computer science,linguistics,machine learning,signal processing and mathematics.

It was first developed to aid the visually impaired by offering a computer-generated spoken voice that would "read" text to the user. The newest applications in speech synthesis are in the area of multimedia.

Here the neural-network that is to be described uses an encoder to break down voice input into tangible data,to be fed into the TTS system for output.Basically, if we give a audio dataset of a reference speaker,this neural network can clone the voice and can synthesise custom speech, that is inputted as text, in the voice of this speaker.Optimizing the underlying processes of this neural network should make the synthesised voice almost indistinguishable from the speaker i.e in terms of pitch,stress voice and other features that make the voice more authentic and natural.

# Contents

# List of Figures

# Chapter 1

# Introduction

Text-to-Speech synthesis,as the term suggests is a method that employs cutting-edge technology to convert analog or digital textual input to sound output which is understandable to the user.

When it comes to the sound output part many properties must be taken into consideration: the linguistic features of the output,acoustic features, context, emotion, expression etc. From the data input ,with the help of linguistics one has to capture the various phonemes,syllables,phrases and other feature that one overlooks so often.

Many approaches have been used to integrate all the above variables to give meaningful solutions,of which Deep Learning has shown the most promise.

Deep learning (DL) is a new research direction in the machine learning area in recent years. It can effectively capture the hidden internal structures of data and use more powerful modeling capabilities to characterize the data .DL-based models have gained significant progress in many fields such as handwriting recognition , machine translation , speech recognition and speech synthesis .To address the problems existing in speech synthesis, many researchers have also proposed the DL-based solutions and achieved great improvements.Here we emphasise on these solutions.

## 1.1 Organization of The Report

This chapter provides a background for the topics covered in this report.We begin with a introduction to the topic of speech synthesis and its origins as mentioned above.

Then we move onto the traditional methods used for speech synthesisafter which we address the Deep Learning methods,which is what is being used most popularly in the field .Here we briefly explain the evolution of the methods that were used for TTS,namely, the Restrictive Boltzmann Machines(RBM),Tacotron,Deep Belief Networks,Google's Wavenet and its associated parts.

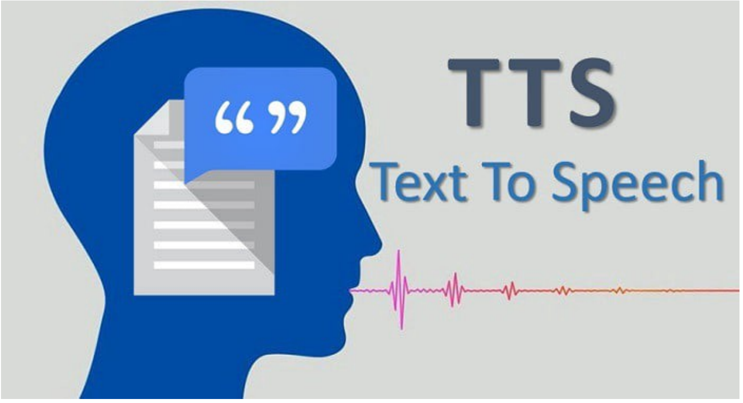
I have implemented a voice cloning model using custom datasets,mined from the internet.The results of the model when compared to the input speech would be described ,along with the Neural Network Architecture that was used,in great detail.This would be followed by a conclusion,future applications and an Appendix that explains technical terms for the readers understanding.

# Chapter 2

# Literature Review

## 2.1 Overview

Speech synthesis, more specifically known as text-to-speech (TTS), is a comprehensive technology that involves many disciplines such as acoustics, linguistics, digital signal processing and statistics [4].This is actually a very important tool used in the fields of communication and information processing.



After the widespread availability of computing resources became available,at around 1990,that there were multiple contributions and continues improvement to this technology.We can classify all the methods of this time into two approaches :concatenative approach and parametric approach.

## 2.2 Concatenative Approach

In the concatenative method, speeches from a previously prepared large database is used to produce new speech.As the word ”concatenate” suggests,it links together speech waveforms from a given database and outputs a speech stream. The databases are structured in such a way that one can obtain custom fit audio based on accent,nationality,gender and so on. For concatenative synthesis there are two methods : one is based on linear prediction coefficients (LPCs) [1] , the other is based on PSOLA. LPC reduces storage capacity and is a simple method,and PSOLA could adjust the ***prosody*** of the speech units ,so one can understand the context of the spoken words.

***Prosody*** is the the stress and intonation patterns of an utterance.

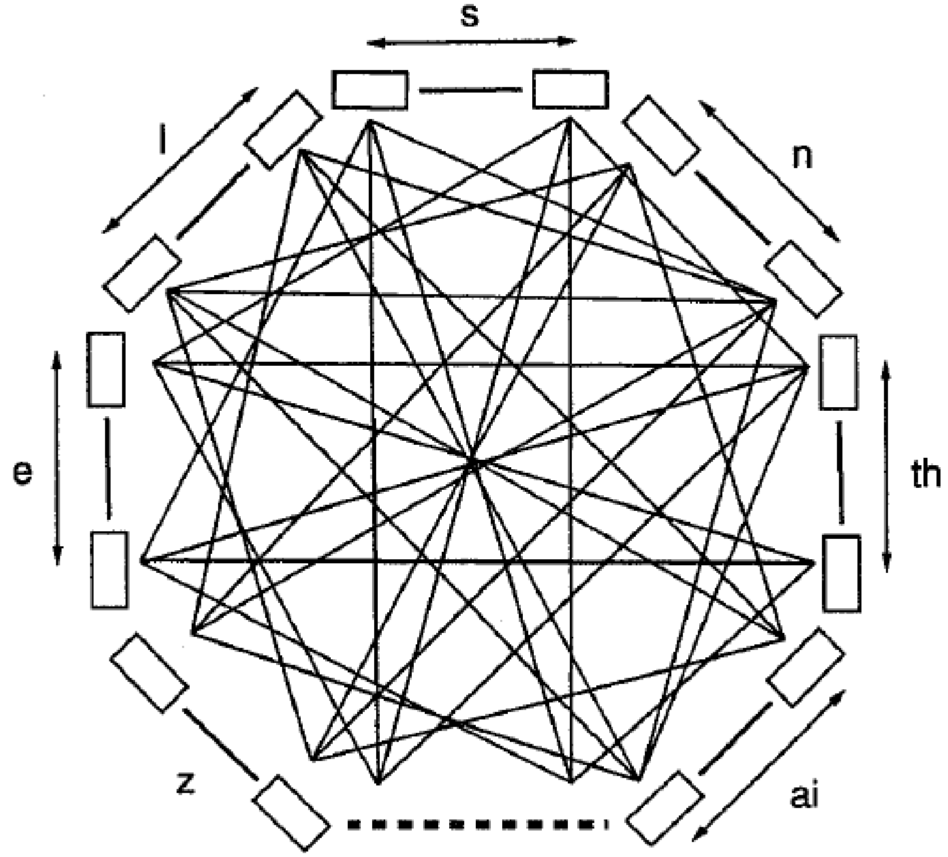


**Fig. 2.1**: The figure below shows a Phoneme Network for a particular database

## 2.3 Parametric Approach

The parametric approach uses a recorded subjects voice and makes the characteristic features of his voice,as a function of parameters/variables that can be adjusted to suit the situation.This technique considers the human vocal process as a soundbox that can produce voiced and unvoiced sounds over time. Voice sounds are basically the sounds generated by vibrations of the vocal cords.(larynx On contrary, the sound produced at the pronunciation of the letters such as ‘l’, ‘r’ or ‘t’,’s is known as unvoiced sounds,which are produced by air flow through lips and teeth(oral cavity)

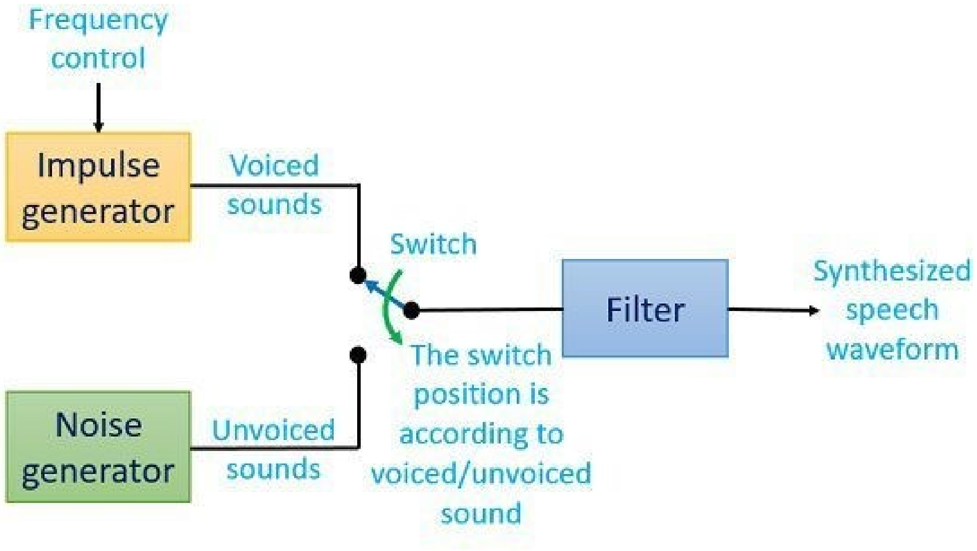


**Fig. 2.2**: Simplified view of human vocal process through electronic circuits

With improvement in technology,this came to be called Statistical Parametric Speech Synthesis(SPSS).It consisted of three components.1)Text Analysis.2)Parametric prediction and finally 3) Speech synthesis.

These two approaches represent the old way of doing speech synthesis. The new and popular approaches incorporate statistics,linguistics,phonetics and subsequently deeplearning.

## 2.4 Deep Learning

**Deep Learning** methods such as Restrictive Boltzmann Machines(RBM),Deep Bi-directional LSTM's ,Tachotrons have increased intelligibility(capability of being understood by a layman) ,expressiveness and so many other attributes that the synthesised speech is almost indistinguishable from recorded speech.

However, the powerful representation capabilities of DL-based models have also brought some new problems. Since DL methods use RNN's and CNN's,to achieve better results, the models need more hidden layers and nodes, which will undoubtedly increase the number of parameters in the network, and the time complexity and space complexity for network training.

However Deep Learning has paved the path for state-of-the-art performance and can certainly promote the development of speech synthesis in the future.
everything below 2.4 comes under Deep Learning methods.

### 2.4.1 Restrictive Boltzmann Machines and Deep Belief Networks

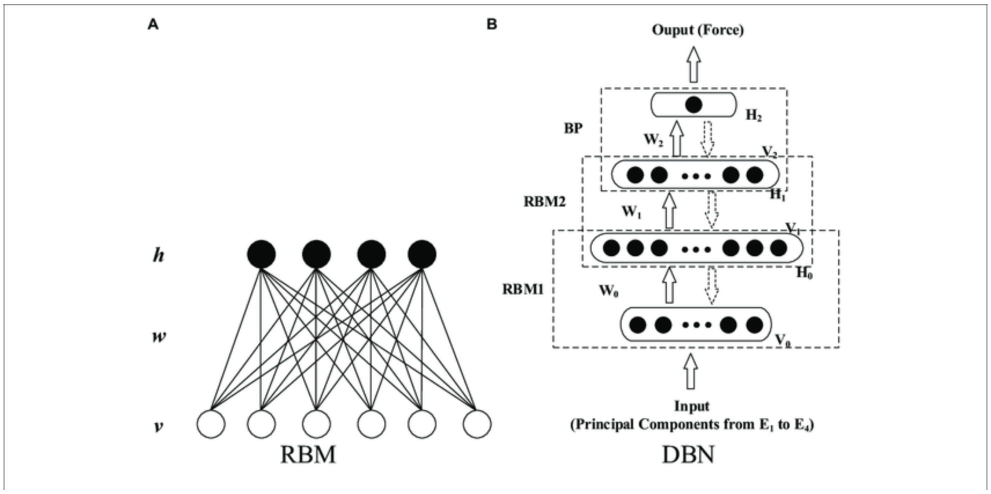


**Fig. 2.3**: The figure below shows a visual comparison of layers in a RBM and DBN

A Restrictive Boltzmann Machine(RBM) is a kind of bipartite undirected graphical model (i.e., Markov random field) which is used to describe the dependency among a set of random variables using a two-layer architecture [2].

A deep belief network (DBN) is a probabilistic generative model which is composed of many layers of hidden units.. In this model, each layer captures the correlations among the activities of hidden features in the layer below.

The Hidden Markov Model(HMM) based synthesis method is a parametric speech synthesis method,that models the spectrum,F0,and other acoustic features using Markov Layers. To overcome the problems faced by this technique the above two computational tools or network architecture was used .

### 2.4.2 End to End Speech Synthesis

A regular TTS system consists of text analyser,parameter modeller and speech synthesiser.Since these components are independent from each other an error from each component would compound and affect the end result.Therefore to alleviate these problems all the components are unified under one roof ,thus elimination all these errors .Below are the two most important methods that employ this.

#### Wavenet

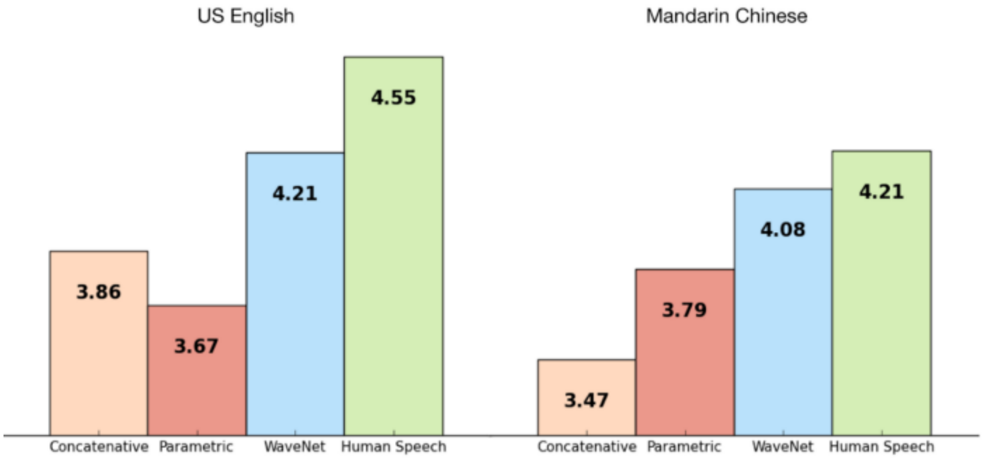


**Fig. 2.4**: The figure below shows the quality of waveNets on a scale of 1–5 compared to other methods

WaveNet is an audio generating model based on the PixelCNN [5]. It is able to produce high quality sound almost identical to a human.

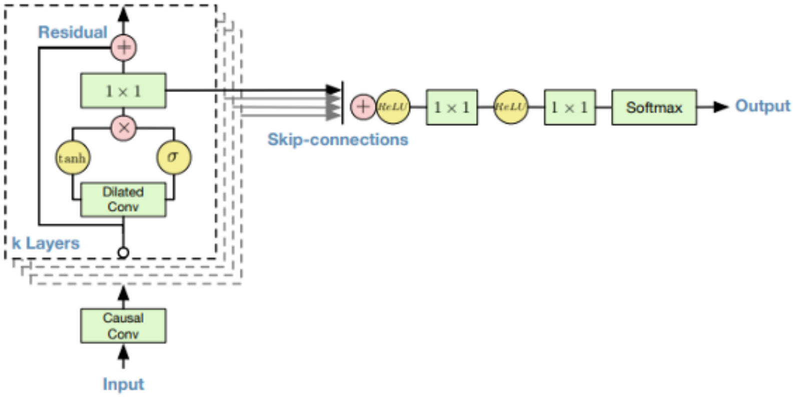


Figure 4: Overview of the residual block and the entire architecture.

Here Convolutional Neural Networks are employed.They are a type of Neural network that have proved very effective in image recognition and classification. In the audio generative model each sample is audio is compared to the previous one,or conditioned.The conditional probability is determined by convolutional layers

In this architecture the output of the model has the same dimension as the input.This conforms with the way Matrices work and thus a predicted voice sample can be fed back to the network to assist in in predicting the next one.

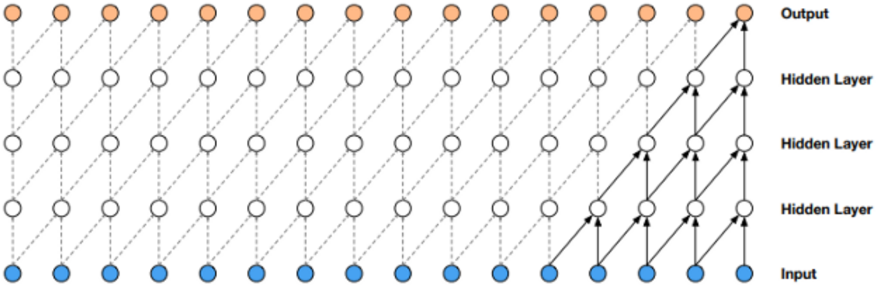
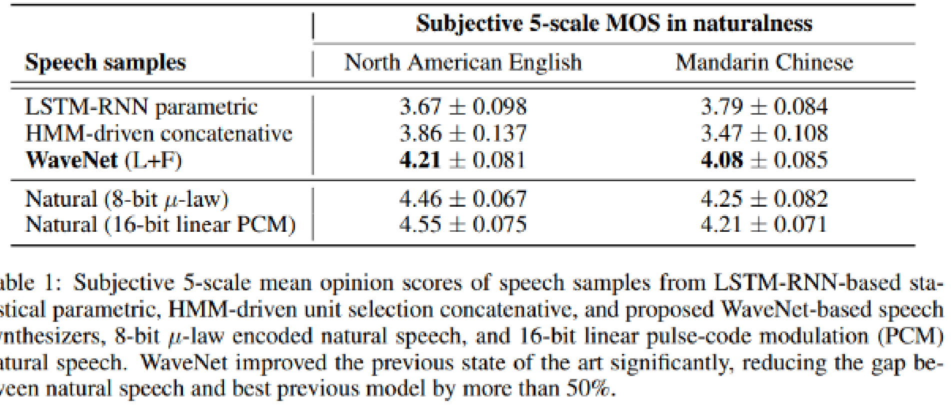


Figure 2: Visualization of a stack of causal convolutional layers.

These are faster to train than RNN's.

The ***MOS (Mean Opinion Score)*** is used for this testing. It measures the quality of voice. It's basically the opinion of a person about the voice quality. It is a number between one and five, with five being the best quality.



| | Subjective 5-scale MOS in naturalness | |
|---|---|---|
| **Speech samples** | North American English | Mandarin Chinese |
| LSTM-RNN parametric | 3.67 ± 0.098 | 3.79 ± 0.084 |
| HMM-driven concatenative | 3.86 ± 0.137 | 3.47 ± 0.108 |
| **WaveNet** (L+F) | **4.21** ± 0.081 | **4.08** ± 0.085 |
| Natural (8-bit $\mu$-law) | 4.46 ± 0.067 | 4.25 ± 0.082 |
| Natural (16-bit linear PCM) | 4.55 ± 0.075 | 4.21 ± 0.071 |

Table 1: Subjective 5-scale mean opinion scores of speech samples from LSTM-RNN-based statistical parametric, HMM-driven unit selection concatenative, and proposed WaveNet-based speech synthesizers, 8-bit $\mu$-law encoded natural speech, and 16-bit linear pulse-code modulation (PCM) natural speech. WaveNet improved the previous state of the art significantly, reducing the gap between natural speech and best previous model by more than 50%.

### Tacotron

Tacotron is an end-to-end TTS model that synthesizes speech directly from [text,audio] pairs. Tacotron generates speech at frame-level and is, therefore, faster than sample-level autoregressive methods [7]. As defined above about MOS,Tacotron has a mean opinion score of 3.82 in English.

The model is trained on audio and text pairs, which makes it very adaptable to new datasets. Tacotron has a seq2seq model that includes an encoder, an attention-based decoder, and a post-processing net. As seen in the architecture diagram below, the model takes characters as input and outputs a raw spectrogram. This spectrogram is then converted to waveforms.

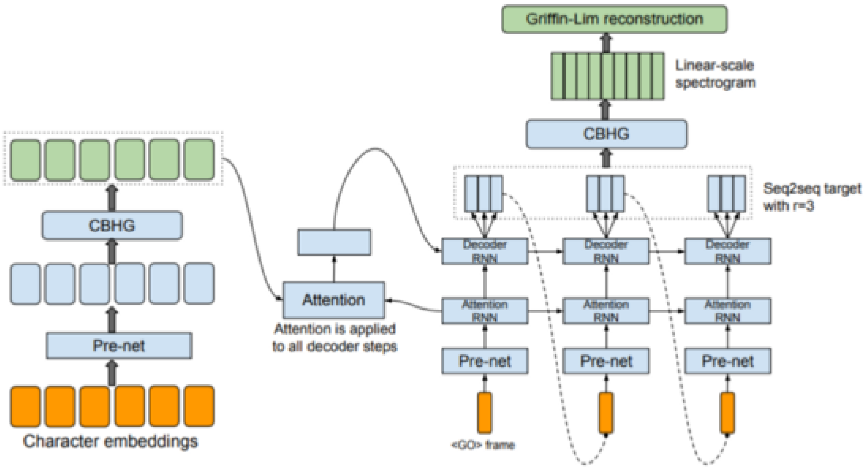


Figure 1: *Model architecture. The model takes characters as input and outputs the corresponding raw spectrogram, which is then fed to the Griffin-Lim reconstruction algorithm to synthesize speech.*

The figure below shows what the CBHG module looks like. It consists of 1-D convolution filters, highway networks, and a bidirectional GRU (Gated Recurrent Unit).

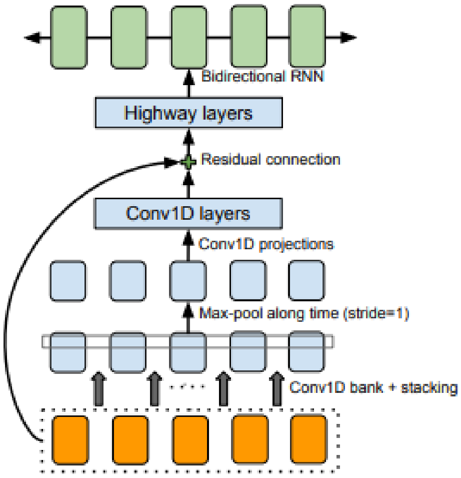


Figure 2: The CBHG (1-D convolution bank + highway network + bidirectional GRU) module adapted from Lee et al. (2016).

A character sequence is fed to the encoder, which extracts sequential representations of text. Each character is represented as a one-hot vector and embedded into a continuous vector. Non-linear transformations are then added, followed by a dropout layer to reduce overfitting. This, in essence, reduces the mispronunciation of words.

The decode used is a tanh content-based attention decoder. The waveforms are then generated using the Griffin-Lim algorithm [3].

The figure below shows the performance of Tacotron compared to other alternatives.

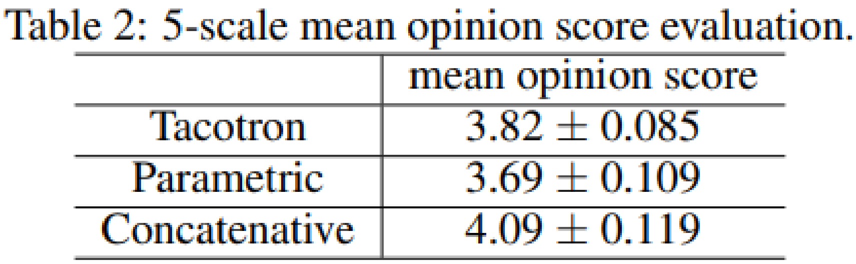


Table 2: 5-scale mean opinion score evaluation.

| | mean opinion score |
|---|---|
| Tacotron | $3.82 \pm 0.085$ |
| Parametric | $3.69 \pm 0.109$ |
| Concatenative | $4.09 \pm 0.119$ |

# Chapter 3

# Dataset Description

As suggested by my supervisor,audio samples pertaining to an NDTV(New Delhi Television Ltd ) news anchor and a WION news anchor were collected.For diversifying the voice output,one of the anchors was male-Prannoy Roy and the other was female-Palki Sharma,both of them being reputed anchors and fluent in the English Language.

(a) Prannoy Roy

(b) Palki Sharma

**Fig. 3.1**: Images of the male and female anchors

## 3.1 Pre-Processing

An hour worth of audio was datamined for each of the anchors.It was done by downloading .mp4 files of the anchors reporting news,through YouTube.Then using Python libraries such as moviepy it was converted to audio files such as .mp3 and .wav files ,so we can input them into the voice cloner.Further preprocessing meant removing other speaker's sounds and background noise from the files for an optimal output.For this the usual Python libraries consisting of Matplotlib,numpy(all part of scipy)was employed along with Librosa.Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.

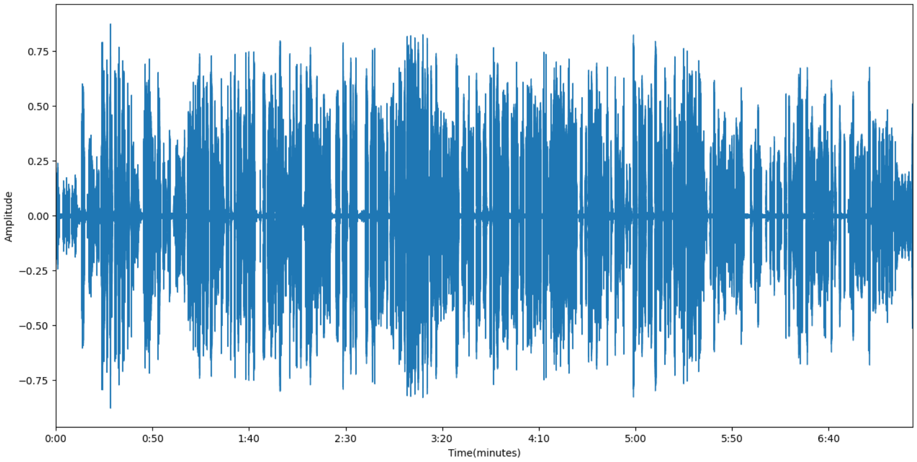


**Fig. 3.2**: Waveform generated from one of the audio clips(7min.wav)

In order to take out the silence and the pauses in the audio,the librosa.effect.split function was used.

After taking into consideration sampling rate and number of frames,the time segements of the desired speakers were obtained and trimmed using librosa.

# Chapter 4

# End-to-End Multi-Speaker TTS model Architecture

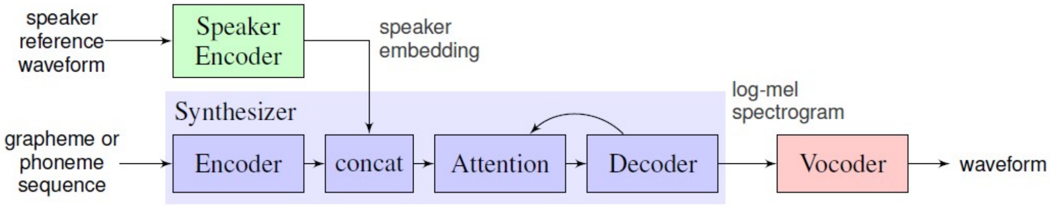


**Fig. 4.1**: A sketch of the model

## 4.1 Encoder

In order for Neural Networks to correct their parameters and give accurate results after a number of iterations ,it has something known as the loss function.This computes how wrong the output result was from the desired result,and updates the parameters accordingly.Here the generalized end-to-end loss training function is used.The Generalized end-to-end (GE2E) training is based on processing a large number of utterances at once, in the form of a batch that contains N speakers, and M utterances from each speaker in average. [6]

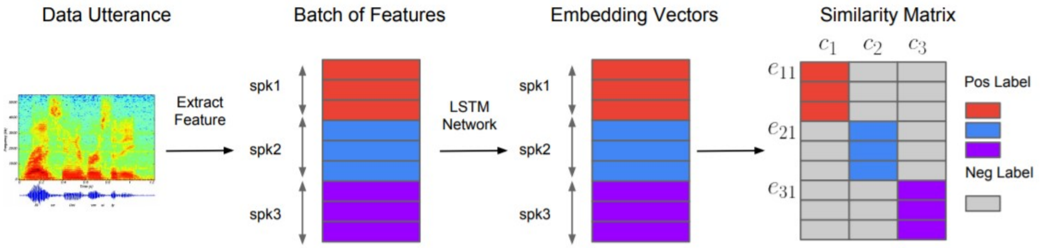


**Fig. 4.2**: Pipeline of Encoder.Different colors indicate utterances/embeddings from different speakers

So what a good Encoder does is it takes the input speech signal or reference speech signal and captures the important characteristics of it,such as Fundamental frequency,pitch,Amplitude,the varying Harmonics and so on.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.

## 4.2 Synthesizer

Here the synthesizer uses a ***Tachotron 2*** architecture as shown in the figure.As explained previously this Architecture is best for converting text to speech waveforms from the reference speaker (when connected with the encoder)

The synthesizer is trained beforehand on pairs of text transcript(like subtitles of a movie) and target audio.At the input, we map the text to a sequence of ***phonemes***.

## 4.3 Vocoder

A vocoder (short for voice encoder) is a synthesis system, which was initially developed to reproduce human speech. Vocoding is the cross synthesis of a musical instrument with

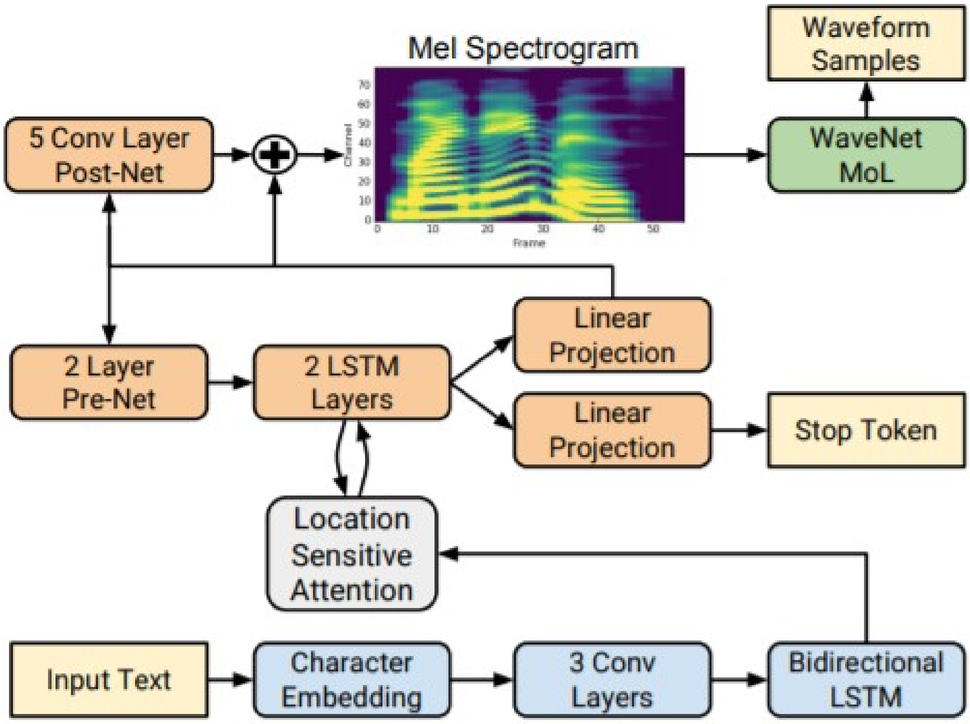


**Fig. 4.3**: what

voice. It was called the vocoder because it involved encoding the voice (speech analysis) and then reconstructing the voice in accordance with a code written to replicate the speech (speech synthesis).

We improve Tacotron by introducing a post-processing neural vocoder, and demonstrate a significant audio quality improvement.We use the sample-by-sample autoregressive WaveNet [19] as a vocoder to invert synthesized mel spectrograms emitted by the synthesis network into time-domain waveforms.

The mel spectrogram predicted by the synthesizer network captures all of the relevant detail needed for high quality synthesis of a variety of voices, allowing a multispeaker vocoder to be constructed by simply training on data from many speakers.

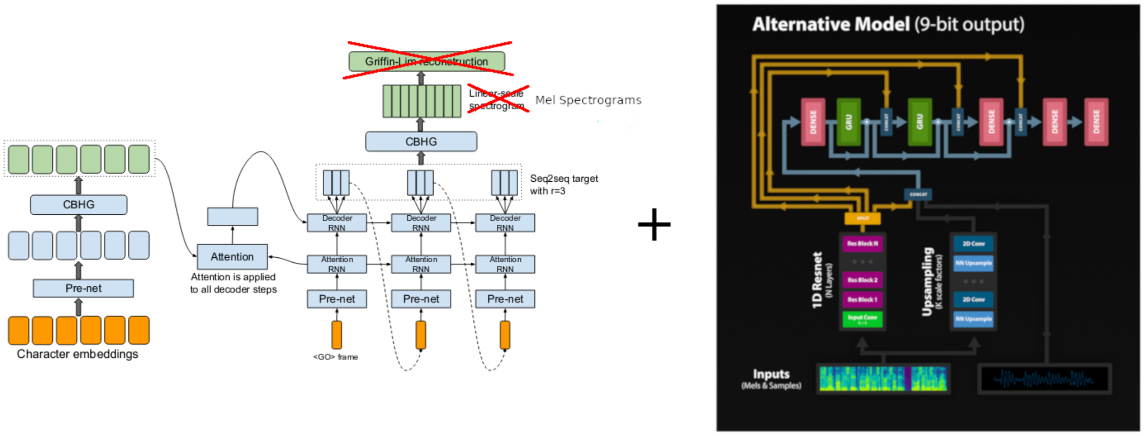


**Fig. 4.4**: Wave RNN Model

# Chapter 5

# Conclusion

Thus speech synthesis is one of the many fields that has been invigorated by the field of Deep learning,but with giving due credit to its predecessors that refined its acoustic,linguistic and mathematical variables. It adds great value to consumer-based products that are currently heavily based on multimedia.We can see these example in our everyday life such as Amazons Alexa,Google's Voice Assistant,Apple's Siri to name a few.Many other major corporations such as Walmart,Tesla and Netflix are looking forward to employing this technology.
On a lighter side it is also aiding the visually impaired and sensory disabled citizens of the world.At present a major portion of this research has been in English and Chinese Mandarin,But soon many other languages will also be included.

Voice cloning ,has been seeing increasing number of practical applications.Voices are a significant part of our identity. Mimicking the voices of friends, enemies, and celebrities have long been an element of comedy and entertainment.Here are some of the use cases:

1.It gives people with conditions such as Huntington's disease, autism, strokes, or traumatic brain injuries the ability to speak naturally.This can be made possible if people can find a comprehensive dataset of them speaking,or can use audio corpora available .

2. Speed up and ease the dubbing process.This can revolutionise the film industry ,when movies are being dubbed in regional languages.Here the performers can commercialise their speech datasets ,which can then be synthesised in other languages using voice cloners. This can also be used in the gaming industry,online education,marketing and advertisements and so on(in short,Multimedia Applications).

One must also keep in mind the ethical considerations of a flawless voice cloner. It can be used to impersonate people's voice ,which would threaten their privacy and financial security.It could also be used to create DeepFakes of popular icons spreading false propaganda which could lead to political instability.

As Deep Learning and AI technology improves it can find solutions to these issues and potentially have a positive impact on society.

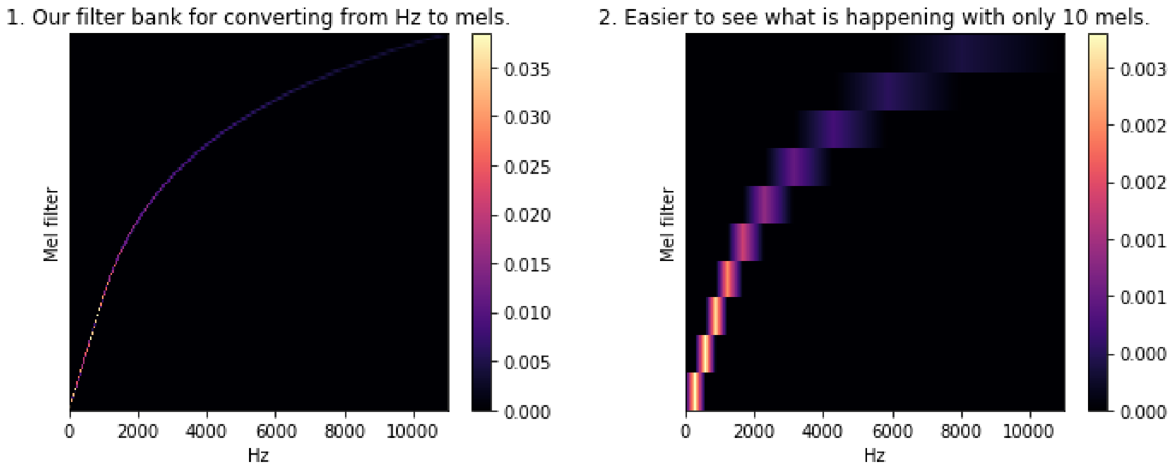
# Chapter 6

# Appendix A:The Mel Spectrogram

## 6.1 The Mel Scale

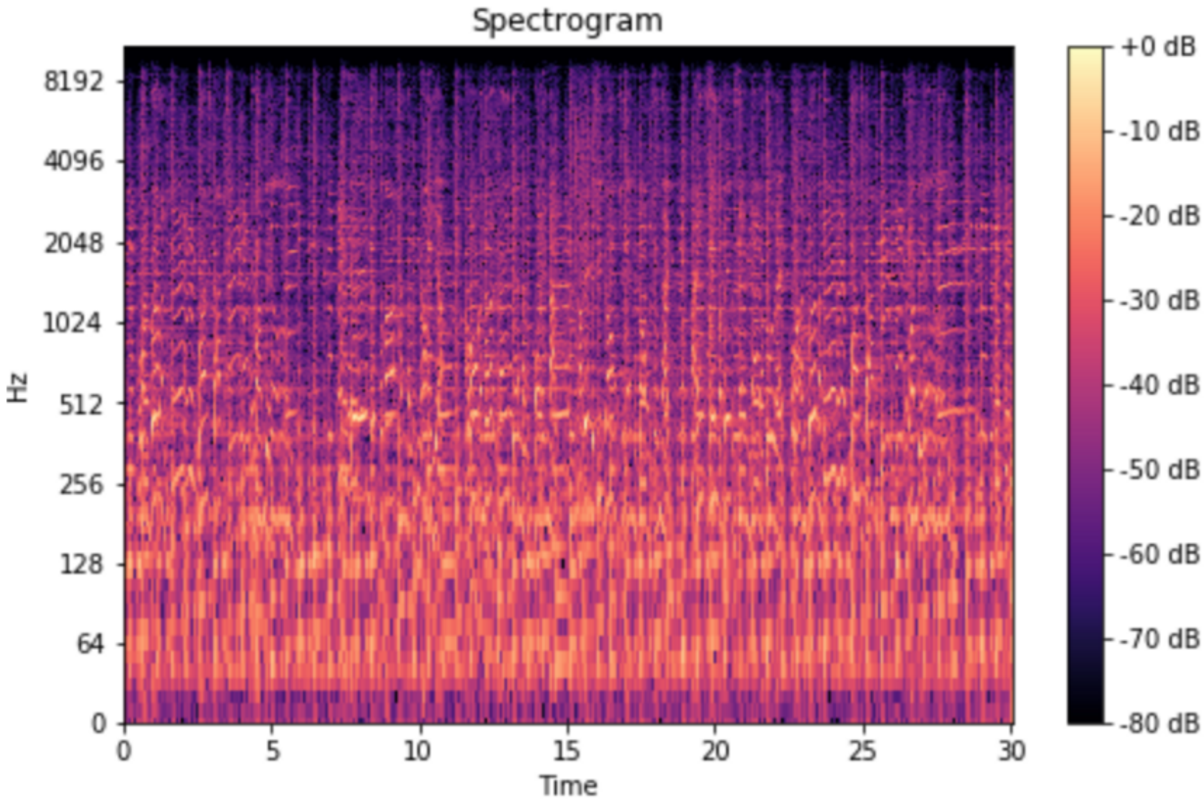
Through studies and practical experience it has been proved that humans don't perceive sound the same way its linearly graphed on a scale.It can be shown that humans can easily make out the difference between sounds of pitch/frequency 500Hz and 1000Hz but not that of 6000Hz and 6500Hz even though the difference in pitch(500Hz) is the same.Please note that humans have a hearing range of 20Hz-20kHz.

Therefore for practical reasons the linear scale was not really useful for making conclusive results.Therefore the mel scale was introduced in 1937 by Stevens, Volkmann, and Newmann.They proposed a unit of pitch such that equal distances in pitch sounded equally distant to the listener. This was called the mel scale.



## 6.2 Spectrogram

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. When applied to an audio signal, spectrograms are sometimes called sonographs, voiceprints, or voicegrams.

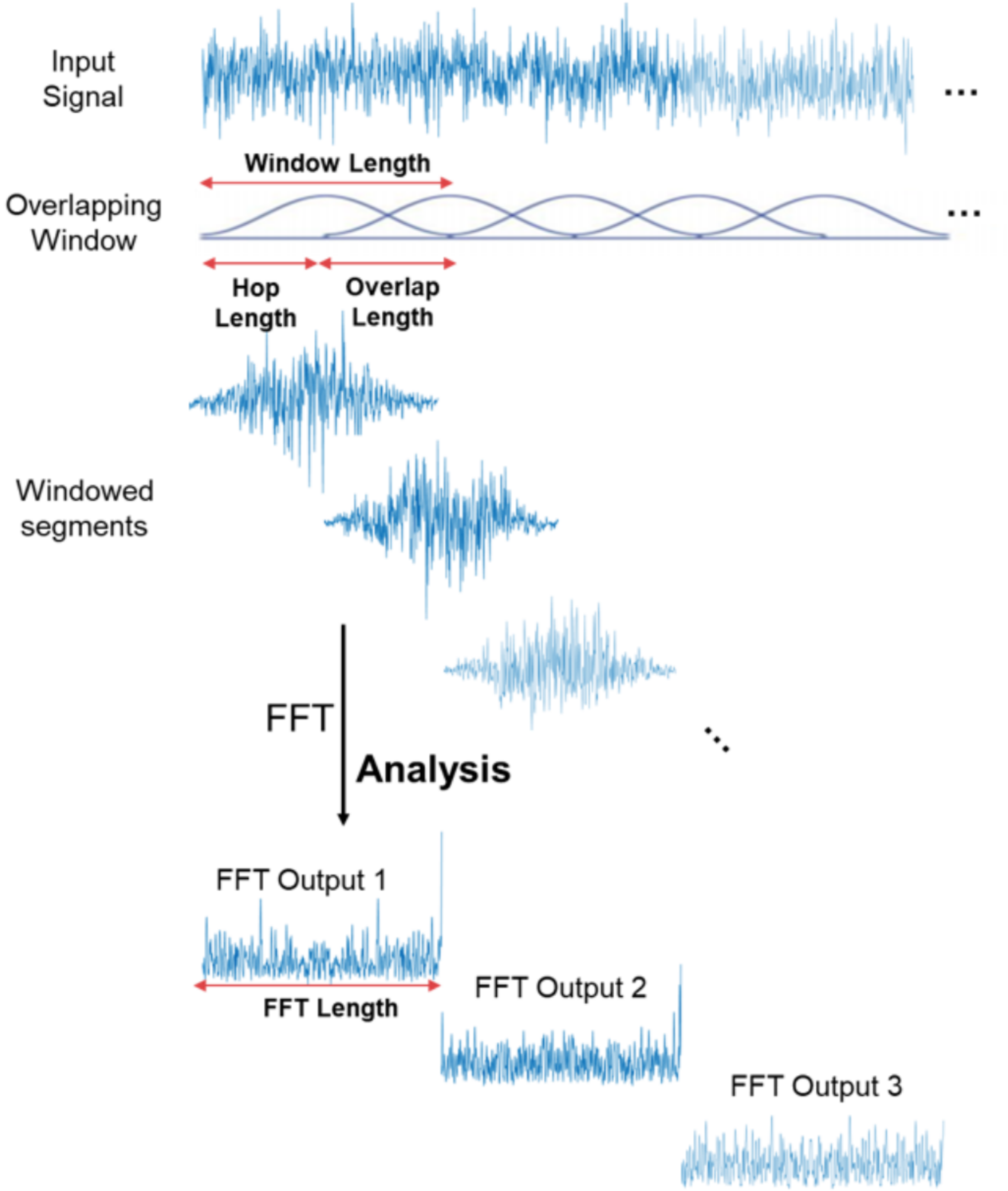


Below are some acoustic parameters that are predicted and their definitions for the sake of understanding:

F0-is the ***Fundamental frequency*** of a speech signal,but practically refers to the approximate frequency of periodic wave.This is high for females and children and low for males.

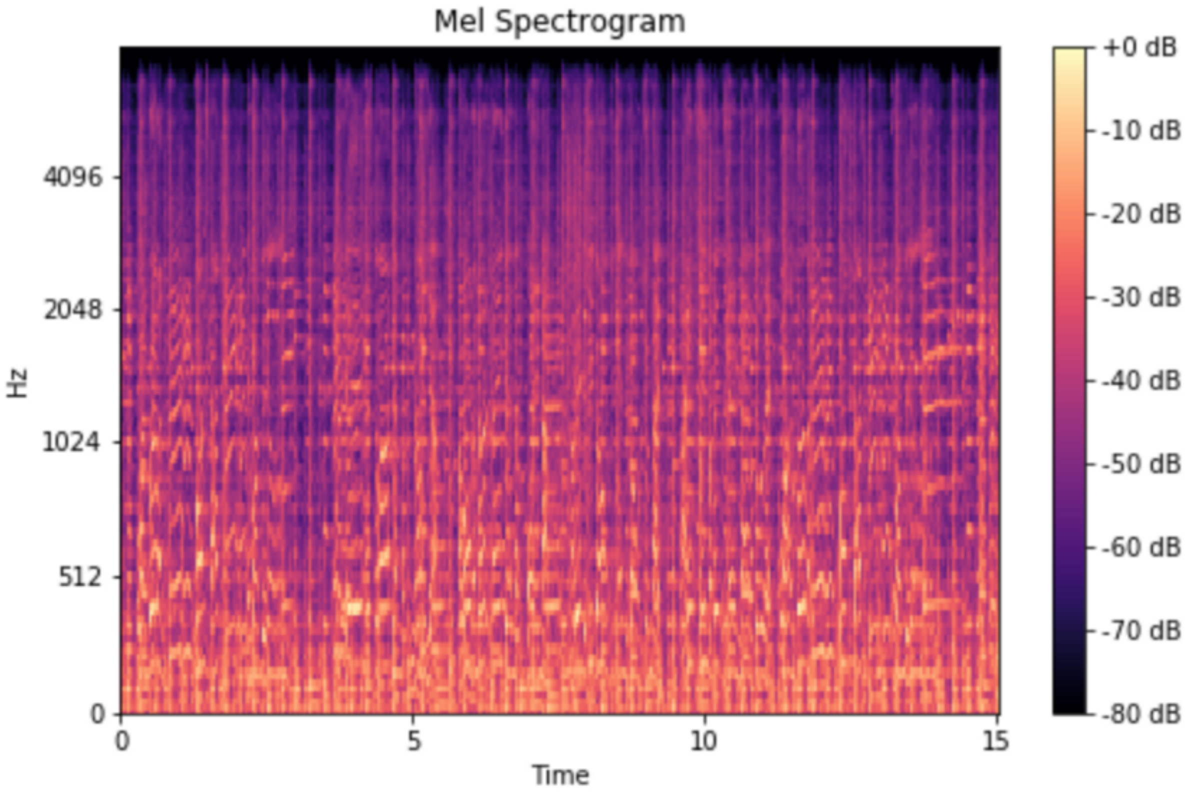
***Sampling rate*** defines the number of samples per second taken from a continuous signal to make a discrete or digital signal.A ***signal*** is the modification/change in a certain quantity with respect to time. For sound signals, the quantity that varies is air pressure. In order to obtain this information digitally one has to take samples of the air over time. We can sample data at our own pace but the standard rate is 44.1kHz.In total this is a waveform which can be studied ,tweaked and studied.

***Fourier Transform*** is a mathematical concept that can decompose a signal into its constituent frequencies. The fast Fourier transform (FFT) is an algorithm that can optimally compute the Fourier transform,on a computer. Using this algorithm one can take segments of the wave,apply Fourier Transform on it and obtain the frequency content of the signal.But as segments vary over time the FFT output also changes over time(which is the case for most audio signals) Here we use another algorithm known as ***short-time Fourier transform*** ,which performs FFT on audio segments. The FFT is computed on overlapping windowed segments of the signal, and we get what is called the ***spectrogram***. The below image obtained from Matlab documentation summarizes this.



## 6.3 Mel Spectrogram

The mel spectrogram that has been used so frequently in Research literature pertaining to DSP,Speech synthesis and so on;is nothing but the spectrogram but wherein the y-axis uses a mel scale along with a decibel(dB) scale as a heat map to show intensity.



# References

[1] B. S. Atal and Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655, 1971.

[2] Z. Ling, L. Deng, and D. Yu. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2129–2139, 2013.

[3] Yoshiki Masuyama, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada. Deep griffin-lim iteration, 2019.

[4] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19):4050, Sep 2019.

[5] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders, 2016.

[6] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification, 2020.

[7] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017.