Software Requirements Specification

For

Text to speech and audio cloning web application

07-01-2022

Prepared by

Specialization	**SAP ID**	**Name**	**Roll No.**
B.Tech CSE – IoT & SC	*500067932*	*Dhairya Kumar Sharma*	*R164218023*
B.Tech CSE – IoT & SC	*500067886*	*Muskan Sawa*	*R164218048*
B.Tech CSE – IoT & SC	*500068004*	*Akash Raj*	*R164218007*

Department of Systemics
School Of Computer Science
UNIVERSITY OF PETROLEUM & ENERGY STUDIES,
DEHRADUN- 248007. Uttarakhand

Table of Contents

Revision History

Date	Change	Reason for Changes	Mentor Signature
18-11-2021	Initial draft	For making initial draft	
07-01-2022	Final draft	Adding new specifications for the web application	

1	INTRODUCTION	
	1.1 Purpose of the Project	Studying and implementing different neural models and architecture for text to speech conversion and comparing the results with mean opinion score (MOS). And building a TTS feature which will take text as input and use any of our models for speech synthesis.
	1.2 Target Beneficiary	Individuals with visual and perusing weaknesses were the early adopters of TTS. It bodes well: TTS facilitates the insight for the 1 out of 5 individuals who have dyslexia, low education pursuers and others with learning disabilities by eliminating the pressure of perusing and introducing data in an ideal configuration. Our venture plans to gather an AI based arrangement of text-to-speech (TTS) amalgamation that can create speech sound in the voice of various speakers, including those inconspicuous during preparation.
	1.3 Project Scope	Main objectives: ● Text to speech ● TTS with proper human like articulation ● Speech Synthesis to Voice Conversion, using transfer learning If time permits: ● Multiple language support ● Make a website on which we can pass out text and then it calls our server which will return audio as a response
	1.4 References	[1] J. Shen and R. Pang, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in (ICASSP):(pp. 4779-4783)IEEE, Canada, 2018 [2] K. Kumar, "MelGAN: Generative Adversarial Networks for," in arXiv preprint:1910.06711, Mila, 2019. [3] G. Yang, "Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech," in IEEE: Spoken Language Technology Workshop (SLT), China, 2020. [4] Y. Ren, "FastSpeech: Fast, Robust and Controllable," in arXiv preprint:1905.09263, China, 2019. [5] Y. Ren, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in arXiv preprint:2006.04558, China, 2020.
2	PROJECT DESCRIPTION	
	2.1 Reference Algorithm	1. Tacotron/ Tacotron 2 2. FastSpeech / FastSpeech 2

		3. MelGAN / MelGAN – STFT / Multi band MelGAN
	2.2 Characteristic of Data	*NA*
	2.3 SWOT Analysis	Strengths: Realtime audio comparison with different features given by different models. Weaknesses: All models have different strengths, the models which are faster have lesser accuracy, the slower models have a better output Opportunities: comparing new Gan models like StyleGan with our current text2mel models can yield better results *Refer to SWOT attached alongside this document.*
	2.4 Project Features	Main Objective: ● To create a web application and compare the strengths and weaknesses of different text to mel models with different vocoders and deriving the best MOS model to be used for audio books
	2.5 User Classes and Characteristics	The application is divided into 3 main features: ● Training models – training the text to mel models, vocoders ● Training auto processor for text to sequence input ● Synthesis Function for deriving mel outputs and output durations for plotting spectrograms and giving audio output for comparison Python with flask framework and MySQL workbench was used for backend development of the web application for a better comparison of the featured models The frontend of the application is made using HTML and CSS.
	2.6 Design and Implementation Constraints	We used PyCharm for our project and the project was made with TensorFlow V-2.6.0 which has compatibility issues with the current version of Cuda and Cudnn thus Cuda v11.2 was used along with Cudnn v8.1
	2.7 Design diagrams	*Use Case Diagram, Flowchart, Process Diagram of Wavenet, Tacotron2, FastSpeech and MelGAN, IPC* *See diagrams attached in appendix D at the end of this document.*
	2.8 Assumption and Dependencies	The project doesn’t have any dependencies or constraints which can’t be worked around.
3	SYSTEM REQUIREMENTS	

	3.1 User Interface	The user interface is needed for the following components: 1. Spectrogram comparison 2. Audio Comparison 3. Text Input 4. Mel and Vocoder Selection
	3.2 Software Interface	Windows 10+
	3.3 Database Interface	MySQL workbench
4	NON-FUNCTIONAL REQUIREMENTS	
	4.1 Performance requirements	All the libraries must be properly installed according to the provided versions. Previously installed libraries like Keras may clash when importing TFAutoModels from Tensorflow so it is better to create a new environment
	4.2 Security requirements	Although the security of the web application is not necessary, there is a scope of adding a user login system in the web application for the sake of preserving individual audio samples
	4.3 Software Quality Attributes	**Adaptability:** In our web application, it is extremely easy to change vocoder models to suit your taste **Correctness:** Different Text to mel models can provide different speech clarification and thus the suitable one can be derived easily **Flexibility:** Pairing Tacotron model with multi band is for a better-quality build whereas you can always rely on FastSpeech for quick processing **Interoperability:** Our components are designed to be interoperable with each other. **Reliability:** The database used in the application is secure and doesn't have permission over deleting saved audio files so they can be easily recovered if lost **Reusability:** Previously stored files and graphs can be used for comparisons
Appendix A: Glossary		**GAN:** Generative adversarial networks **STFT:** Short-time Fourier transform **MOS:** Mean Opinion Score **TTS:** Text to Speech
Appendix B: Analysis Model		*Refer to diagrams attached alongside this document.*

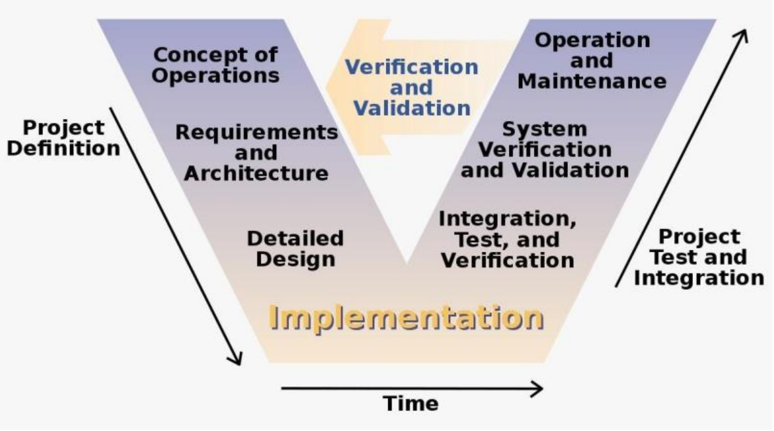
<table>
<tr>
<td>Appendix C: Issues List</td>
<td>

A better comparison factor like graphs for data and validation loss for generator and discriminator can bring a better visualization of the results.

Improve diagrams.

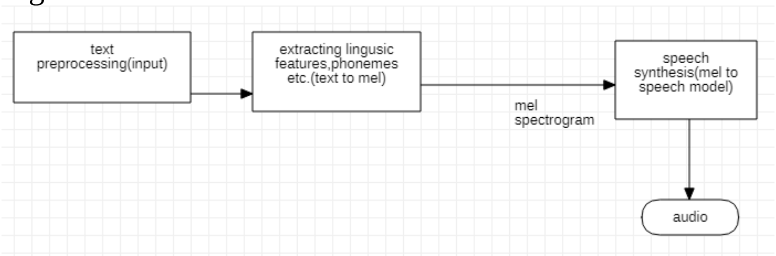
</td>
</tr>
<tr>
<td>Appendix D: Diagrams and illustrations</td>
<td>

Figure 1. Steps of the SDLC model used in the project.



Concept of Operations
Verification and Validation
Operation and Maintenance
Project Definition
Requirements and Architecture
System Verification and Validation
Detailed Design
Integration, Test, and Verification
Project Test and Integration
Implementation
Time

Figure 2. Flowchart



text preprocessing(input)
extracting lingusic features,phonemes etc.(text to mel)
mel spectrogram
speech synthesis(mel to speech model)
audio

Figure 3. Processing flow diagram of wavenet

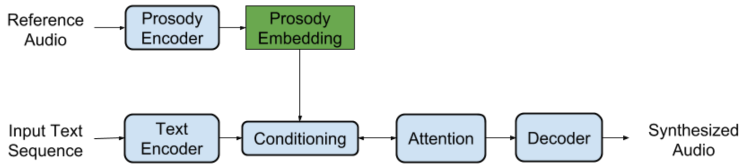


Figure 4. Process Diagram of tacotron2

</td>
</tr>
</table>

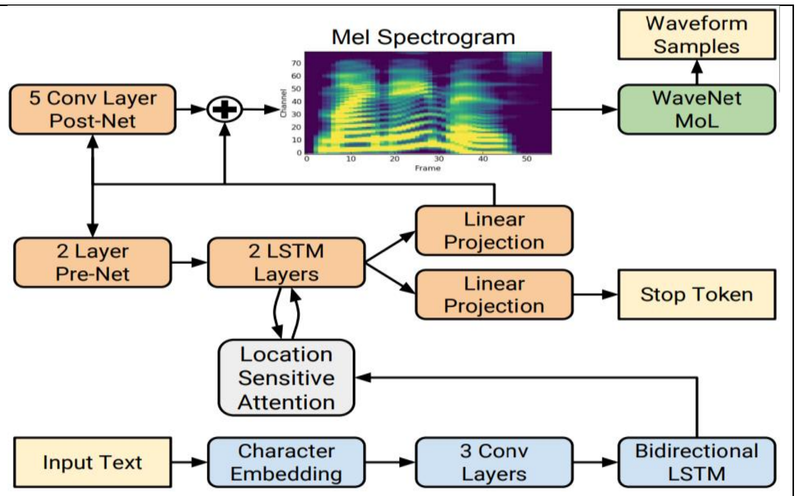


Figure 5. Process Diagram of FastSpeech

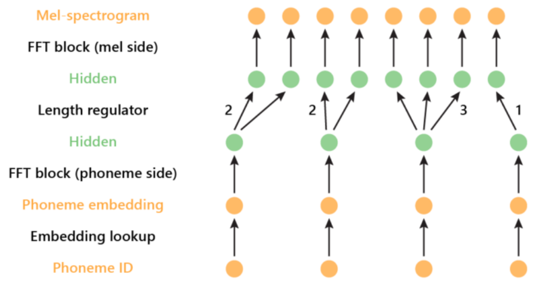


Figure 6. Process Diagram of MelGAN

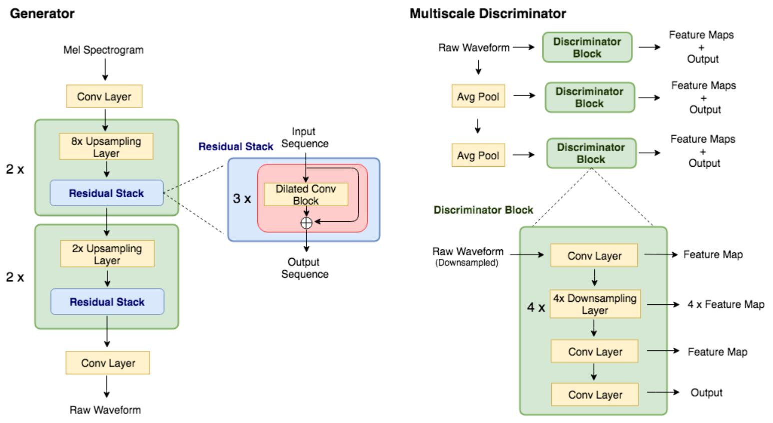


Figure 7. IPC

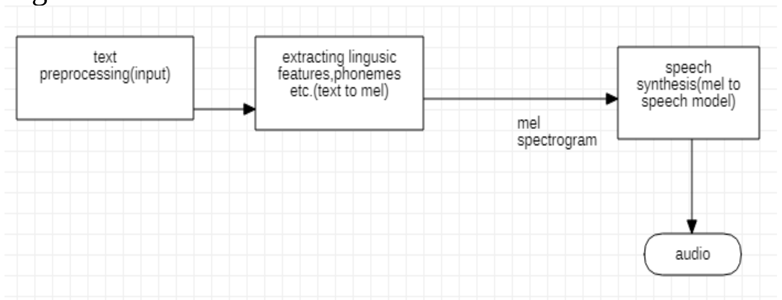
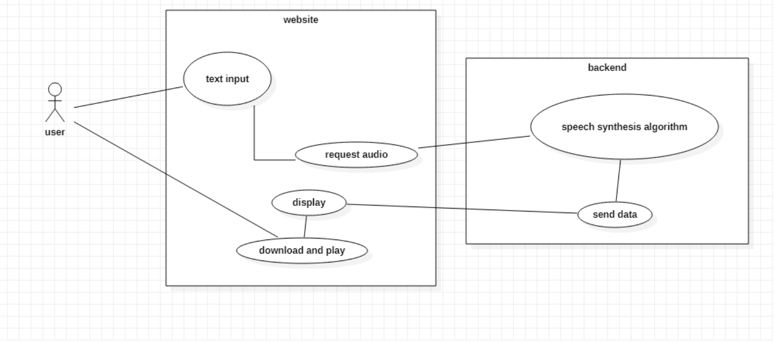

text preprocessing(input)
extracting lingusic features,phonemes etc.(text to mel)
mel spectrogram
speech synthesis(mel to speech model)
audio

	Figure 8. Use Case Diagram  website text input user request audio display download and play backend speech synthesis algorithm send data