Pushto Text To Speech System

Aimen Wadood
P180002@nu.edu.pk

Muhammad Iftikhar
P180054@nu.edu.pk

Hamza Majeed
P180128@nu.edu.pk

Supervisor:
Dr. Taimoor Khan

Table of Contents

Introduction

Text to speech systems take input text , dilect and will convert it into desired speech.

Applications

- Learning new language.

Applications

- Learning new language.
- Used for people with learning disability.

Applications

- Learning new language.
- Used for people with learning disability.
- Online Reading.

Applications

- Learning new language.
- Used for people with learning disability.
- Online Reading.
- Used by Visually Impaired People.

The Development of Pashto Speech Synthesis System [1]

Problem:

- Articulatory Synthesis.
- Formant Synthesis.

Solution:
Concatenative Synthesis.
Limitations:
The problem of acronyms, abbreviations, and out of vocabulary words are not discussed in this paper.

Deep Pashto Text-to-Speech [2]

Problem:

- Wavenet
- Tactron

Solution:

- Two stage DNN model development.
- Two stage RNN-LSTM.

Limitations:

- The scope of the model is limited to the standard dialect of Pashto language. Other dialects will not be considered.
- Deep learning models require a large amount of data but data was very limited.

Literature Review

Name	Problem	Solution	Limitations
Text to Speech System for Urdu Language [*3*]	As in the unit selection base synthesis pre-recorded units are combined to obtainthe speech of given Text but development of corpus is very difficult thing.	HMM Base Model Synthesis is a statistical parametric bases speech Synthesis that it stored the statistics rather then the waveform.	The problem of acronyms, abbreviations, and out of vocabulary words is not considered very efficiently.

Literature Review

Name	Problem	Solution	Limitations
An Arabic TTS System Based on the IBM Trainable Speech Synthesizer [4]	Formant synthesizers, controlled by rules, have the advantage of small footprints but the synthesized speech doesn't sound natural.	They have constructed a system using a state-of-the-art IBM trainable unit selection based concatenative speech synthesizer.	There was lack of large well discretized and POS tagged Arabic corpus. The database recorded was not large enough.

Literature Review

Name	**Problem**	**Solution**	**Limitations**
Development of An Arabic Text-To-Speech System [*5*]	Formant and concatenative Synthesizers have their own limitations.	They have built a hybrid model where the formant and concatenative models have been applied parallel to phonemes where they are most suitable.	There was lack of large well discretized and POS tagged Arabic corpus. The database recorded was not large enough.

Literature Review

Name	Problem	Solution	Limitations
Glow TTS [*6*]	Fast Speech and Para Net have been proposed to generate mel-spectograms from text in parallel. It cannot be trained autoregressive TTS models as their external aligners.	Based on flow based model for parallel TTS, Glow TTS, do not require any external aligner. It models the conditional distribution of mel-spectrograms.	Neural TTS (Glow) models could sometimes synthesize undesirable speech with slurry or wrong pro-nunciations.

Literature Review

Name	Problem	Solution	Limitations
Semi-supervised Learning for Multi-speaker Text-to-speech Synthesis Using Discrete Speech Representation [7]	Former multi speaker TTS requires a large amount of paired high-quality speech and text data which is unavailable under low resources due to expensive data collection.	Unpaired data is accessible Therefore, a semi supervised training TTS,SeqRQ-AE was introduced. It is trained for unpaired audio-text pairs.	Not suitable for cross-lingual.

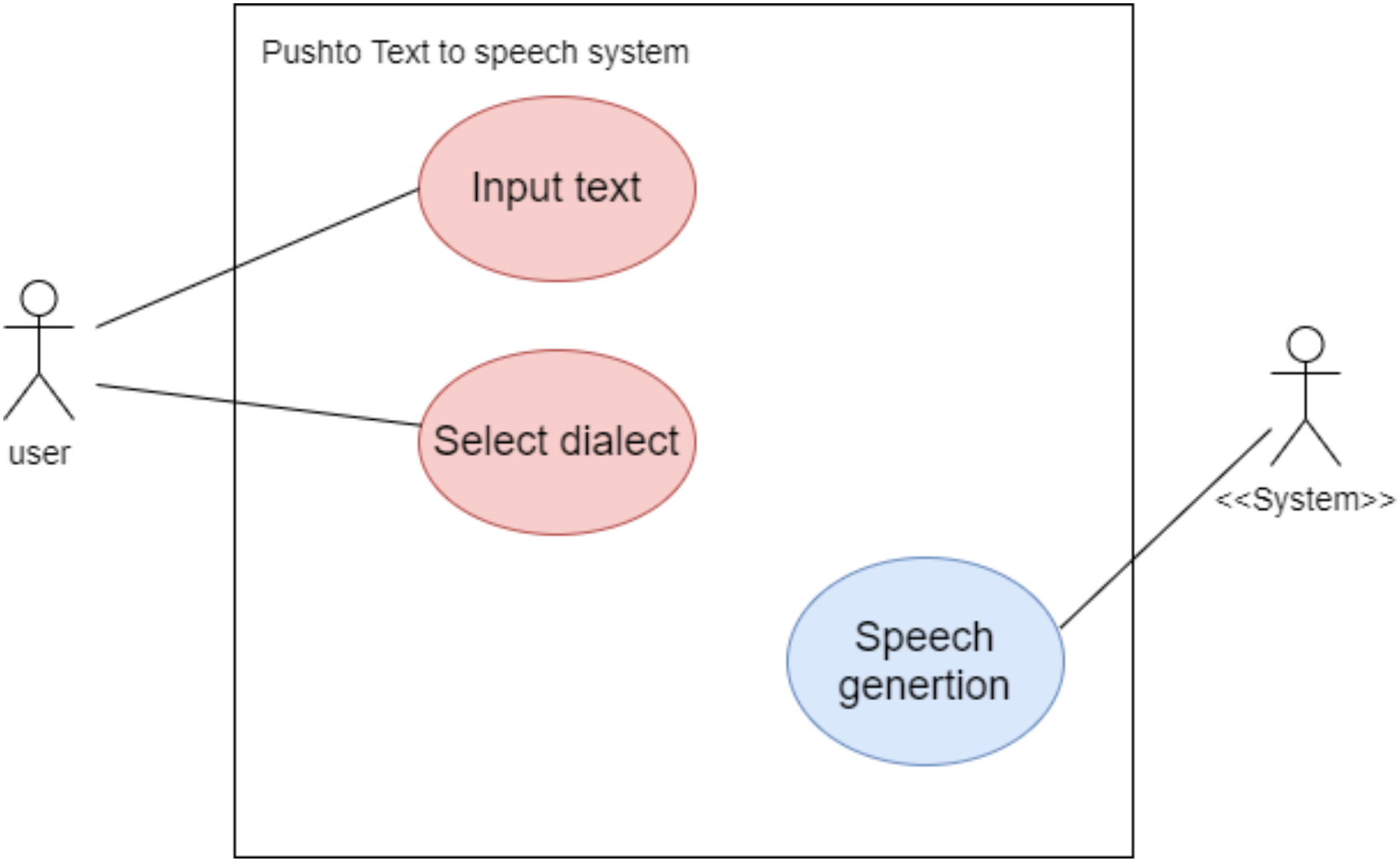
Literature Review

Name	**Problem**	**Solution**	**Limitations**
Fast Speech: Fast, Robust and Controllable Text to Speech [*8*]	Slow interface speed for mel-spectogram generation.	Through paralle mel-spectrogram generation, Fast Speech greatly speeds up the synthesis process.	The parallel TTS models cannot be trained without autoregressive TTS models as their external aligners.

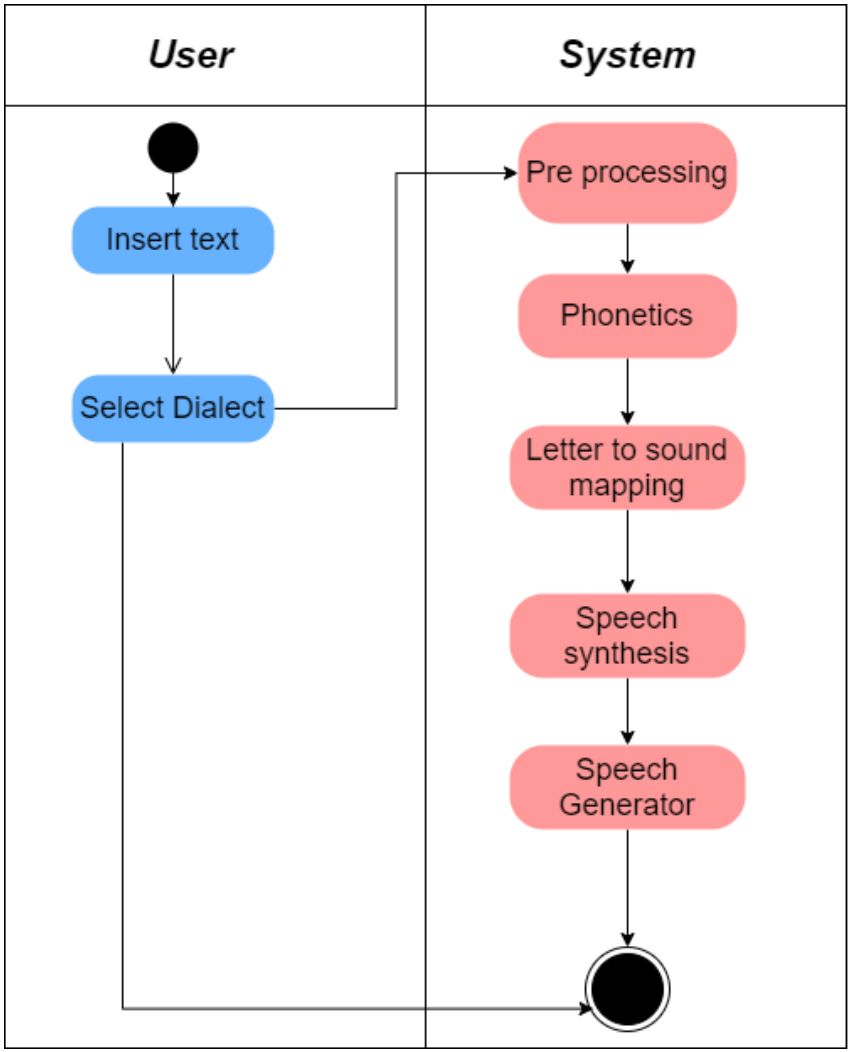
Literature Review

Name	Problem	Solution	Limitations
Pre-trained Text Embeddings for Enhanced Text-to-Speech Synthesis [*9*]	Factorization of TTS systems allows training each module separately, but results in errors propagating from one component to subsequent components.	An E2E-TTS system has two modules: A feature generation module, A waveform synthesis module. They proposed two models; subword-level model, phrase-level model.	It only works on the pre-trained Bert text embeddings.

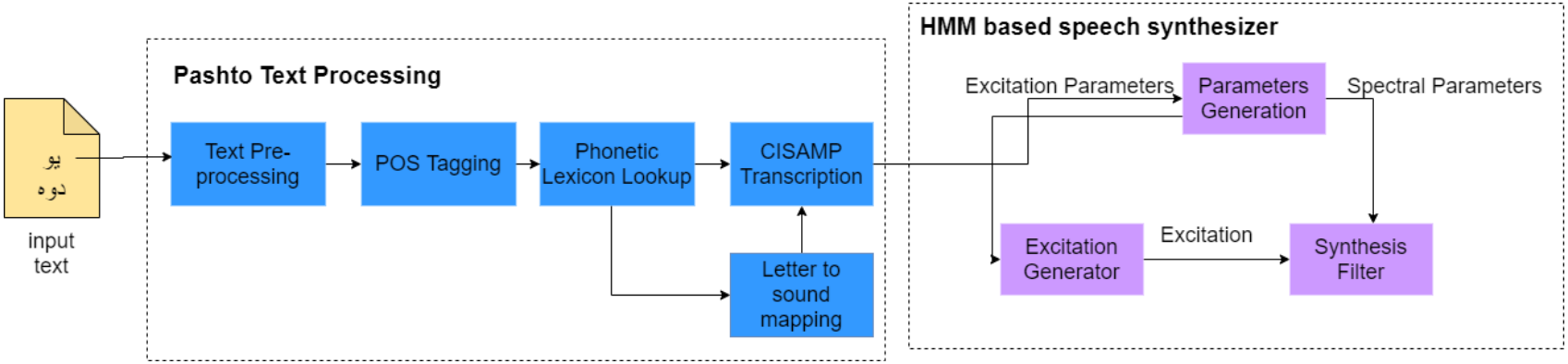
Use Case Diagram



Activity Diagram



Methadology



References

- Deep Pashto Text-to-Speech, Sarmad HussaiAbdul Rahman Safi, 2019. [1]
- The Development of Pashto Speech Synthesis System ,Muhammad Akbar Ali Khan, Sahibzada Abdur Rehman Abid, Fatima Tuz Zuhra, 2013. [2]
- Text to speech system for urdu language, Sarmad Hussain, 2019. [3]
- An Arabic TTS System Based on the IBM Trainable Speech Synthesizer, Ossama Emam, Amr Youssef, 2004. [4]
- Development of An Arabic Text-To-Speech System, Mustafa Zeki, Othman O. Khalifa, A. W. Naji, 2010.[5]

References

- Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search, Jungil Kong, Sungroh Yoon, Sungwon Kim, Jaehyeon Kim, 2020[6]
- Semi-supervised Learning for Multi-speaker Text-to-speech Synthesis Using Discrete Speech Representation, Tao Tu, Yuan-Jui Chen, Alexander H. Liu, Hung-yi Le, 2020.[7]
- FastSpeech: Fast, Robust and Controllable Text to Speech, Yi Ren, Yangjun Ruan, Xu Tan, 2019 [8]
- Pre-Trained Text Embeddings for Enhanced Text-to-Speech Synthesis Hayashi, Tomoki and Watanabe, Shinji and Toda, Tomoki and Takeda, Kazuya and Toshniwal, Shubham and Livescu, Karen 2019. [9]

Questions!!