Group 6

[Email address]

Abstract

To train a Synthetic Automated Speech Recognizer, that is trained directly using text, which in turn is transcribed through a multi speaker TTS to feed ASR training audio files

Training Automatic Speech Recognition Systems Using Text-to-Speech Synthesized Data

* DLFA Capstone Project

Contents

[Problem Statement: 2](#_Toc163743901)

[Background Information: 2](#_Toc163743902)

[Motivation for Selection of the Project: 2](#_Toc163743903)

[Detailed Dataset Description and Dataset Source: 3](#_Toc163743904)

[Current Benchmark: 3](#_Toc163743905)

[Proposed Plan: 3](#_Toc163743906)

[Approaches: 3](#_Toc163743907)

[Stages with Defined Deliverables: 4](#_Toc163743908)

[Methodology: 4](#_Toc163743909)

[Preliminary Exploratory Data Analysis: 4](#_Toc163743910)

[Expected Outcomes: 5](#_Toc163743911)

[Project Demonstration Strategy (Tentative Plans): 5](#_Toc163743912)

## Problem Statement:

*The problem revolves around training an automatic speech recognition (ASR) system using only text data. This unconventional approach involves leveraging a pre-trained multi-speaker text-to-speech (TTS) model to generate synthetic speech paired with corresponding text. The synthesized speech-text pairs are then used to train the ASR system. The ultimate goal is to evaluate the trained ASR's performance in transcribing unseen sentences for both seen and unseen speakers from the multi-speaker TTS system, with performance measured in terms of Word Error Rate.*

## Background Information:

*The domain of automatic speech recognition (ASR) has garnered significant interest due to its wide-ranging applications, including virtual assistants, transcription services, and language translation. Traditional ASR systems rely on paired data of speech and text for training, which can be time-consuming and expensive to collect, especially for multiple speakers. This problem statement proposes an innovative solution by utilizing a pre-trained multi-speaker text-to-speech (TTS) model to generate synthetic speech from text data.*

## Motivation for Selection of the Project:

*The motivation behind this project lies in addressing the challenge of acquiring large amounts of paired speech-text data for training ASR systems, especially in scenarios where such data is scarce or costly to obtain. By leveraging TTS technology to synthesize speech from text, we can potentially overcome this limitation and enable the training of ASR systems using readily available text corpora. This solution has multiple uses e.g., automating testing of ASR system, benchmarking WER. It can also help quickly to train ASR systems on multiple voices and potentially in multi[le lamguages.*

## Detailed Dataset Description and Dataset Source:

*The dataset for this project consists of text data, which serves as input for the pre-trained multi-speaker text-to-speech (TTS) model to generate synthetic speech. The text data is sourced from corpora English corpus. It should encompass a diverse range of vocabulary and linguistic patterns to ensure robust training of the ASR system. This data is then converted into Audio files at 44100 Hz sampling rate using Nemo TTS with 20 speakers comprising of both male and female voices. We are also considering training of conqui TTS with multiple Indian English speakers, time permitting.*

## Current Benchmark:

*As of the current state, there might not be a specific benchmark for training ASR systems solely using text-to-speech synthesized data. However, benchmarks for ASR systems trained on traditional speech-text paired data can serve as a reference point for evaluating the performance of the proposed approach.*

## Proposed Plan:

### Approaches:

*- Utilize a pre-trained multi-speaker text-to-speech (TTS) model (Nemo TTS from NVIDIA) to generate synthetic speech from text data.*

*- Use English Text Corpus from Corpora with 3000 sentences*

*- Use 20 Speaker IDs from Nemo-TTS to randomly generate audio fies from these 3000 sentences*

*- Train an automatic speech recognition (ASR) system using the synthesized speech-text pairs from 18 spearer IDs.*

*- Evaluate model using speech-text pairs from 18 Speaker IDs and the 2 Speaker IDs we kept aside for WER, to cover unseen sentences for both seen and unseen speakers from the TTS system.*

### Stages with Defined Deliverables:

*1. Data Collection and Preprocessing:*

*- Gather diverse text data for synthesis.*

*- Preprocess text data to ensure consistency and quality.*

*2. Speech Synthesis:*

*- Utilize the pre-trained multispeaker TTS model to generate synthetic speech.*

*- Pair synthesized speech with corresponding text.*

*3. ASR Model Training:*

*- Train the ASR model using synthesized speech-text pairs.*

*4. Evaluation:*

*- Evaluate the performance of the trained ASR system on unseen sentences for seen and unseen speakers from the TTS system.*

### Methodology:

#### Packages and Tools: Python, Nemo TTS, TensorFlow, Keras, PyTorch, Librosa,.

#### Algorithms: MFCC for feature extraction, RNN, Bi-directional LSTM, GRU based models for automatic speech recognition.

#### Metrics: Word Error Rate (WER) for evaluating ASR performance.

#### Deployment Plan: Once the ASR model is trained and evaluated, explore deployment options such as integration into existing systems or standalone applications.

## Preliminary Exploratory Data Analysis:

*- Analyze the distribution of text data in terms of vocabulary size, sentence lengths, and linguistic diversity.*

*- Assess the quality of synthesized speech in terms of naturalness and intelligibility.*

## Expected Outcomes:

*- A trained ASR system capable of transcribing speech from unseen sentences, utilizing text-to-speech synthesized data for training.*

*- Reduction in dependency on paired speech-text data for training ASR systems.*

*- Insights into the effectiveness and limitations of training ASR using synthesized speech-text pairs.*

## Project Demonstration Strategy (Tentative Plans):

*- Demonstrate the ASR system's performance on a set of unseen sentences, showcasing its ability to accurately transcribe speech from various speakers synthesized by the TTS model.*

*- Present comparative analysis with ASR systems trained on traditional speech-text paired data, highlighting the advantages and potential challenges of the proposed approach.*

*This document outlines a comprehensive plan for training an automatic speech recognition (ASR) system using text-to-speech synthesized data, emphasizing the novelty, methodology, and expected outcomes of the project.*