Etude 14

Super Size It

Date: 9th February 2020

Dimitar Bankov

Josh Casey

Chen Lu

Usman Shah

Introduction

Sentences makes up speech, words makes up sentences, and syllables makes up words. It is not unexpected, that when we try to get the amount of syllables from the spelling of a word alone, things starts to fall apart. Exceptions are added after exceptions, new rules are introduced as new words are made, it drives people to madness, attempting to get the perfect score, on the Etude No1 - Syllable Slam. How can we improve this etude?How can we make it more fun to do, with less rules making, and more problem solving?How can we make it more unique, and hence offer a more diverse experience/skill acquired for the students? Here we propose to alter several steps of the etude, as one potential approach to super size it: Specifically, to change the etude from getting the amount of syllables from text data, to attempting to parse the amount of syllables in an audio data.

Changes Made

We have made changes to the way the data is inputed. Instead of reading in a string of text/a word in alphabets, students will be required to read in a spoken audio clip, of an english word, transform it into a manipulatable way, and attempt to figure out the amount of syllables that are present within the word. Furthermore, the students will be given a dataset, while the lecturers will hold a secret testing dataset of a similar structure. Improving the accuracy on the public dataset, but also ensuring for a generalization (hence not simply fitting on to the given dataset) will be required.
Though the students will be welcomed to generate their own data, using whatever tools/strategy available.

Additional Work

Firstly, students will be required to learn to wrangle with the data format. Some audio may be in an “mp3” format, while others in “wav”. They may need to learn to use suitable libraries, or understand the format standards.
Secondly, Students will be required to think of the problem of syllable parsing in terms of analysing speech. This is a vastly different way of thinking, as it is no longer the process of constructing a finite state machine, but rather smart feature detectors, that will likely be finally successful in accomplishing high accuracy in the given dataset.

Lesson Learnt

Students will learn how to wrangle with audio data. This does not only give good benefit on encouraging an elaborated understanding of the various data formats, pipelines, and existing solutions;But will also be of value, in terms of developing one's ability to find out about an arbitrary format of data and various interfaces/standards to communicate/work with it.

Students will learn how to analyse audio data. This includes understanding the concept of waveforms, frequencies, sampling frequencies and other specifications of what constitutes as a sound/speech, and use these newly gained knowledge to come up with creative approaches to solve the problem.

Students will learn about the idea of testing on a dataset, when 100% accuracy is not guaranteed.

By introducing a specific testing dataset, not only does it add clarity to the task, but also simulates the real world environment,when datasets acquired to develop tricky algorithms are always a reflection of the real world, but not the entirety of it.

Course Objectives emphasized

1.2 1.3 1.4

2.1 2.2 2.3 2 2.6 2.8 2.9

3.2 3.5 3.6

4.1 4.2 4.3 4.4 4.5

More details:

Can the task be accomplished?

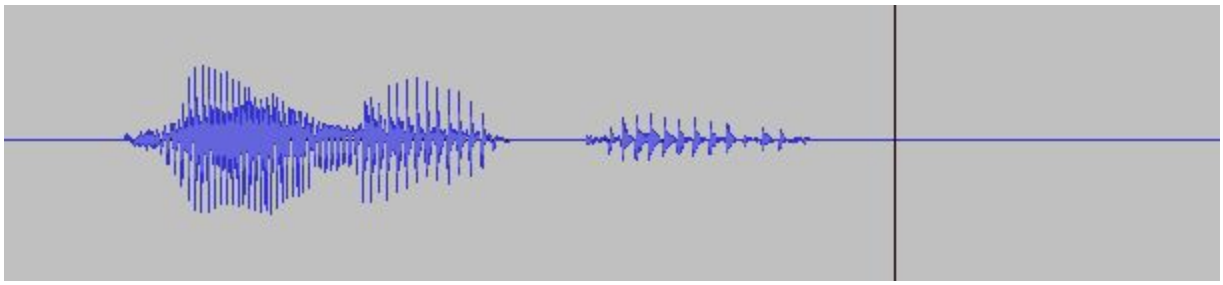

Fig1.1 The wave form of the word “pineapples”, which contains 3 syllables. Generated by a digital TTS service based on artificial neural networks.

The task of parsing the the word based on waveforms seems to be daunting, however, since it is possible to identify certain phoneme based on wave forms, and phoneme counting is the most accurate way to count syllables, we believe that this task is in fact solvable.

To solve it, however, would certainly still require clever work arounds, and identifying key features, or creating automatic feature detectors/machine learning algorithms.

Although it is possible to varify the statistical correlation, between obvious features (such as number of waves) to syllables, in order to give a good estimate on the difficulty of the task. It is beyond the focus of this report.

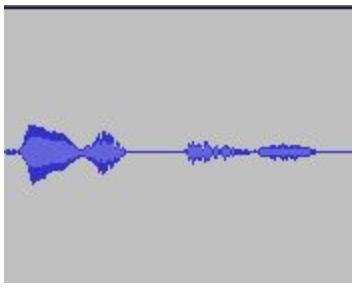
Fig1.2 (The same word “pineapples”, generated by a slightly less well made synthesizer, it places too much stress on the ending ‘s’, making it not possible to identify using the wave peak counting approach, but perhaps there are other ways…)

Dataset and metrics

The dataset should be defined prior to the etude hand out. And should consist of a varied distribution of diverse cases of words and waveforms, the dataset is then split into two portions of a similar structure, so that if the student were to be able to find some robust method to adapt to the public dataset, it will also do well on the secret dataset. This introduction of a well defined dataset is optional, and mostly to mimic the construction standards of machine learning challenges/datasets, which has proven successful to provide at least a fun and concise problem definition. The lecturer should figure out a reasonable accuracy threshold, which when surpassed, passes the submission.

Generating Audio?

As an option for the lecturer to decide, the task of generating the audio can be either done by the lecturers, or left as an exercise for the students.

The lecturers can use any text to speech methods, but we would suggest to use a neuronetwork based synthesizer, for the clarity of the audio created.

References:

https://en.wikipedia.org/wiki/Speech_synthesis

https://github.com/Kyubyong/dc_tts

https://arxiv.org/abs/1710.08969(Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention)