# ELL-888 Assignment 3 - Sequential models

April 10, 2018

## 1 Introduction

This assignment is regarding processing sequential data using recurrent neural networks. You will be given a fairly open problem which can be solved using multiple techniques. However, you are expected to use DNNs which are known to be strong machinery for these kinds of tasks.

## 2 The Problem statement

You will be given speech recordings (wav files) of several hundred speakers uttering many English sentences. Each wav file will be associated with a word file (.wrd file) that would contain the words that are there in the sentence along with their temporal boundaries. Table 1 is an example of a .wrd file. The temporal boundaries are in number of samples with the sampling rate being 16000 Hz. Figure 1 shows an example wave file with corresponding word boundaries marked. The problem is to determine the word boundaries in a new wav file given the list of words that are spoken. Since this is the final assignment, the expectation is that you come up with the approach as well as the solution. However, it is recommended to use sequential models. There are two specific possible problems that could be solved:

| start of the word | end of the word | Word |
|---|---|---|
| 3050 | 5273 | She |
| 5723 | 10337 | Had |
| 9190 | 11517 | your |
| 11517 | 16334 | dark |
| 21199 | 22560 | suit |
| 22560 | 28064 | greasy |
| 28064 | 33360 | wash |
| 33754 | 37556 | water |

Table 1: Example wrd file

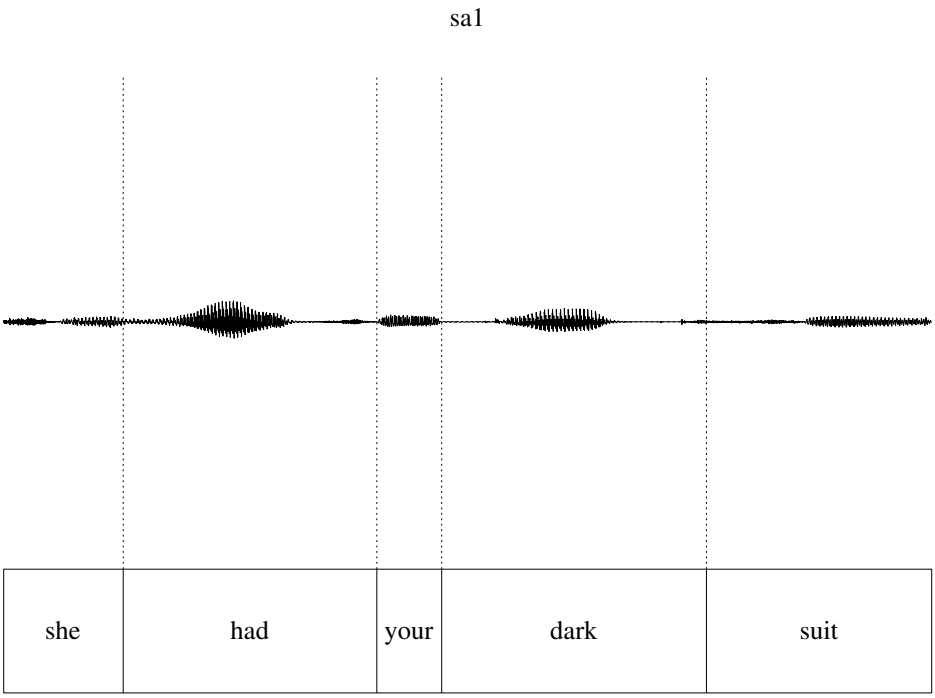


Figure 1: Speech and corresponding word boundaries.

1. Using the word boundary information (all columns in Table 1) during training.

2. Not using the word boundary information but only the sequence of words (only the the last column in Table 1) during training.

The former problem can be solved using standard supervised learning with one of the standard cost functions, the latter problem needs one to use specialized cost and approaches such as connectionist temporal cost. Students can solve either of the problems or both. As usual, the test data will be released 48 hrs a-priori to the submission deadline.

# 3 Evaluation metrics

Given a wav file and corresponding wrd file, a word-window is defined as the duration between two successive true word boundaries. The following metrics that are to be computed on each word window.

1. A correct detection - If there is exactly one detection in a word window.

2. False detection - If there are more than one detection within one word window.

3. Missed detection - If there is no detection in a word window.

4. Detection rate - total number of correct detections divided by the total number of word windows averaged over the entire test data.

5. Miss rate - total number of missed detections divided by the total number of word windows averaged over the entire test data.

6. False detection rate - total number of false detections divided by the total number of word windows averaged over the entire test data.

7. Accuracy histograms - For every correct detection, compute the deviation (in samples) between the actual boundary (left-boundary) and the detected boundary. Then compute the histograms of all such deviations for the entire test data and plot the histograms. Also, obtain the summary statistics (mean and standard deviations) of this histogram.

8. If the algorithm is parameterized by a critical parameter (may be a threshold), plot the ROC curves a and compute precision, recall and F-score.