

# Homework 10 – Speech

Arthur J. Redfern  
[arthur.redfern@utdallas.edu](mailto:arthur.redfern@utdallas.edu)  
Apr 10, 2019

---

Note: this set of homework solutions is incomplete

## 0 Outline

- 1 Logistics
- 2 Reading
- 3 Theory
- 4 Practice

## 1 Logistics

Assigned: Wed Apr 10, 2019  
Due: Wed Apr 17, 2019  
Format: PDF uploaded to eLearning

## 2 Reading

1. To reinforce the speech to text transduction material presented in class, read the following overview paper covering CTC, RNN transducer and attention based models

A comparison of sequence-to-sequence models for speech recognition  
[https://www.isca-speech.org/archive/Interspeech\\_2017/pdfs/0233.PDF](https://www.isca-speech.org/archive/Interspeech_2017/pdfs/0233.PDF)

2. Beam search is a critical component of accurate speech to text transduction (and language to language translation), but was not covered in detail in the slides. To address this, read the following paper that shows how to incorporate an external language model with speech to text transduction networks using beam search

First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs  
<https://arxiv.org/abs/1408.2873>

3. Embedded devices are an important target for speech to text transduction systems. Read the following blog post and paper on a deployed speech to text transduction system that includes many items we've discussed: RNN transducer models, weight quantization, beam search, ... and a number of items we haven't.

An all-neural on-device speech recognizer  
<https://ai.googleblog.com/2019/03/an-all-neural-on-device-speech.html>

Streaming end-to-end speech recognition for mobile devices  
<https://arxiv.org/abs/1811.06621>

4. To improve your understanding of text to speech transduction, it's worthwhile to read the WaveNet blog post and paper

WaveNet: a generative model for raw audio  
<https://deepmind.com/blog/wavenet-generative-model-raw-audio/>  
<https://arxiv.org/abs/1609.03499>

## 3 Theory

None

## 4 Practice

5. Work through the TensorFlow command recognition tutorial

Simple audio recognition  
[https://www.tensorflow.org/tutorials/sequences/audio\\_recognition](https://www.tensorflow.org/tutorials/sequences/audio_recognition)

If interested, also check out a recent command recognition model that uses CNN layers, bi directional LSTM layers, NN layers and attention

A neural attention model for speech command recognition  
<https://arxiv.org/abs/1808.08929>

<https://github.com/douglas125/SpeechCmdRecognition>

6. Attempt to run a DeepSpeech or DeepSpeech2 based speech to text transduction system (this is going to be somewhat involved / time consuming, but is ultimately good practice for working with 3rd party models). For DeepSpeech see

Project DeepSpeech

<https://github.com/mozilla/DeepSpeech>

<https://progur.com/2018/02/how-to-use-mozilla-deepspeech-tutorial.html>

and for DeepSpeech2 see

DeepSpeech2 model

[https://github.com/tensorflow/models/tree/master/research/deep\\_speech](https://github.com/tensorflow/models/tree/master/research/deep_speech)

based on

A PaddlePaddle implementation of DeepSpeech2 architecture for ASR

<https://github.com/PaddlePaddle/DeepSpeech>

Feel free to use a different implementation if you prefer.