

Novel Transformer Architecture for sEMG Silent Speech Recognition

Tada Makepeace

School of Computing
Final Year Research Project

April 12, 2022

Abstract

No more than 300 words summarizing this dissertation.

Table of Contents

Abstract	i
Acknowledgements	vi
1 Introduction	1
1.1 A section	1
1.1.1 A sub-section	1
1.2 Citations	1
1.3 Figures	1
2 Literature Review	2
2.1 Silent Speech Interfaces for Speech Restoration: A Review . .	2
2.2 Proposed Directions for Research	2
2.2.1 Text Classification	2
2.2.2 Data Augmentation	2
2.2.3 Improving Existing Model	2
2.3 Connectionist Temporal Classification (CTC)	3
3 Approaches	4
3.1 Dataset	4
3.1.1 Dataset Selection and Justification	4
3.1.2 Feature Selection	5
3.2 Models	5
3.2.1 DeepSpeech2 Model	5

3.3	Baseline ASR Testing	5
3.3.1	Datasets	5
3.3.2	Implementation	6
3.3.3	Results	6
A	Project Initiation Document	8
B	Ethics Review	19
C	DeepSpeech2 ASR Model	22
	References	25

List of Tables

3.1	DeepSpeech2 Audio ASR Baseline Results	7
-----	--	---

List of Figures

Acknowledgements

Thanks.

Chapter 1

Introduction

A gentle reminder not to get this chapter perfect until the dissertation is nearing its completion. . .

1.1 A section

1.1.1 A sub-section

A sub-sub-section

1.2 Citations

1.3 Figures

Chapter 2

Literature Review

2.1 Silent Speech Interfaces for Speech Restoration: A Review

2.2 Proposed Directions for Research

2.2.1 Text Classification

2.2.2 Data Augmentation

It might be possible to use GANs (Generative Adversarial Networks) or VAEs (Variational Auto Encoders) to create a model which can generate more data samples.

2.2.3 Improving Existing Model

Use RL-Learning to find a model which can outperform the original model (problem is small dataset size which means that the benefits of using RL to create a model might not be as applicable to this problem).

Use CapsNet on top of the CNNs to improve the generalisability of the

detected EEG features (there are papers which show that these are feasible).

2.3 Connectionist Temporal Classification (CTC)

For any speech recognition task, a model must know the alignment between the input (e.g. audio, EMG features, etc.) and the target transcription. On the surface, this makes training any speech recognition model difficult.

Without having the alignments between the input and the transcription, simpler approaches aren't available to us, such as mapping a single character to a fixed number of inputs. This is because people's rates of speech vary, regardless of the input modality (e.g. audio, EMG features, etc). Another option is to hand-label the alignments between the input and the text transcriptions. This would produce a performant model, however, hand-labeling the alignments is time consuming, especially for larger datasets.

Connectionist Temporal Classification (CTC) is a method to train a model without knowing the alignment between the input and the output and is especially well suited to labeled datasets such as speech recognition.

Chapter 3

Approaches

This chapter will describe the main dataset used for this research, why it was chosen, the main EMG signals of interest from the dataset, methods of feature selection and visualisation and the machine learning approaches used to transcribe the EMG data into a text transcription.

3.1 Dataset

3.1.1 Dataset Selection and Justification

The dataset which is used throughout this research project is the open-source surface electromyography silent speech (sEMG silent speech) dataset released by David Gaddy along with his paper, Digital Voicing of Silent Speech (Gaddy and Klein, 2020). The paper describes a novel method of transcribing aligned silent speech data directly into speech features along with the largest open-source sEMG silent speech dataset.

This dataset was chosen for this research as it is the largest, high quality open source sEMG silent speech dataset.

3.1.2 Feature Selection

For this research there were two primary ways of selecting features. The first method was to use the same feature processing methods described in the original (Gaddy and Klein, 2020) paper. The second method was to use a convolutional neural network (CNN) architecture to automatically learn features from the dataset, in an end-to-end manner.

3.2 Models

3.2.1 DeepSpeech2 Model

The DeepSpeech2 model was the initial automatic speech recognition (ASR) machine learning model which was considered for experimentation as it is relatively simple to implement and is known to have good performance, even on smaller datasets, achieving a 3.10 WER on the WSJ eval’92 dataset (Amodei et al., 2015).

3.3 Baseline ASR Testing

To get a realistic baseline for the possible performance of the silent speech models, standard audio based ASR models were trained on different slices of the Digital Voicing datasets audio and text transcriptions only. The intuition for this was that an ASR model trained on the audio and text transcriptions should in theory perform better than on the purely vocalised EMG data or final silent EMG data and their respective text transcriptions. The results of the baseline ASR models is provided below.

3.3.1 Datasets

Three main datasets were chosen for the baseline ASR tests. The first dataset was the LJSpeech dataset (Ito and Johnson, 2017) which is an open-source dataset primarily used in speech synthesis and also speech recognition. The

remaining two datasets are comprised of the audio and text transcriptions of vocalised utterances from the Digital Voicing dataset release (Gaddy and Klein, 2020). The first dataset from this data comes from utterances from the closed vocabulary set of recordings during the vocalised condition from the dataset. The second dataset from this data comes from all of the audio and text transcriptions during the vocalised condition across the entire dataset. The three datasets are referred to as *ASR-LJSpeech*, *ASR-SilentSpeechVocal-Closed* and *ASR-SilentSpeechVocal-Full*, respectively.

The *ASR-LJSpeech dataset* was chosen as it is a standard open-source dataset used for speech related machine learning tasks. It also has similar, but greater quantity of important properties than the *ASR-SilentSpeechVocal-Full* dataset. These include the vocabulary size, average length of recordings and duration of the entire dataset. This intuitively means that the DeepSpeech2 ASR model trained on the *ASR-LJSpeech dataset* should have a lower WER rate, so better performance, than trained on the *ASR-SilentSpeechVocal-Full dataset*. Whereas the *ASR-SilentSpeechVocal-Closed* dataset was chosen because it has a far lower duration and vocabulary size than the *ASR-SilentSpeechVocal-Full* dataset which makes it a lot faster to experiment with as the training time is lower.

3.3.2 Implementation

3.3.3 Results

The following results table shows the word-error rate for each dataset which the DeepSpeech2 model was trained on. Each dataset used a slightly different vocabulary. The vocabulary for each dataset is the list of valid characters of the text transcriptions which the model considers, any other characters are encoded as <UNK>. A different vocabulary was chosen for the *ASR-SilentSpeechVocal-Closed* dataset because it contains far more numbers than the other datasets and including numbers in it’s vocabulary means that the transcriptions would be useful, rather than filled with the unknown place-

holder value.

Table 3.1: DeepSpeech2 Audio ASR Baseline Results

Dataset	Encoding Vocabulary	WER
ASR-LJSpeech	" abcdefghijklmnopqrstuvwxyz-"	0.55
ASR-SilentSpeechVocal-Closed	" abcdefghijklmnopqrstuvwxyz0123456789-"	0.33
ASR-SilentSpeechVocal-Full	" abcdefghijklmnopqrstuvwxyz-"	0.45

Appendix A

Project Initiation Document



School of Computing Project Initiation Document

Tada Makepeace

**sEMG Silent Speech Research
Research Project**

1. Basic details

Student name:	Tada Makepeace
Draft project title:	sEMG Silent Speech Research
Course:	BSc Computer Science
Project supervisor:	Dr Dalin Zhou
Client organisation:	N/A
Client contact name:	N/A

2. Degree suitability

This project satisfies the criteria for my course as it involves using computer science to solve a novel problem. This project relies on using a custom electronic device, i.e. a surface electromyography (sEMG) device to record electrical signals from a person's facial muscles to be able to convert their speech articulations to either digitally voiced speech (reproducing their voice) or text, without the user having to produce sound or strongly move their facial muscles. This project provides a novel way for humans and computers to interact with one another, in other words, it proposes a novel brain-computer interface (BCI).

This project uses a computer based information processing architecture to process the signals acquired from the BCI device by preprocessing multiple channels of sEMG signals which are recorded from the device and converting the data into a format which is suitable for processing. This is performed by removing the noise from the signal and converting the new data into features which can be used in a machine learning transduction (sEMG to digitally voiced speech) or classification architecture (sEMG to text). Then the features extracted from the processed multi-channel sEMG signal can be used to either reproduce the user's voice as they are speaking, or to just classify the signal into text.

The entire project from end to end uses computer science methods to acquire surface electromyography (sEMG) signals from a user's facial muscles, preprocess the data into a suitable representation as features for a machine learning system and then either feed these features into a transduction model or classification model.

3. Outline of the project environment and problem to be solved

The problem I will investigate is how to classify electrical signals from a person's facial muscles captured using non-invasive surface electromyography (sEMG) into text without them speaking (i.e. silently articulated speech). I believe this project is worth working on because non-invasive silent speech devices are a brain-computer interface which offers unique benefits which other methods do not. A non-exhaustive list of examples is provided below:

Privacy of conversation: Typical speech recognition systems have users broadcast what they're saying to the environment (e.g. issuing commands to an Amazon Alexa device) and therefore privacy is not maintained. A silent speech device does not require the user to say anything aloud, rather they're only required to slightly move their facial muscles to mouth out what they would like to say.

Eavesdropping: Voice interfaces are always listening in on conversations, when not desired, only to be visibly activated later on by a specific trigger-word such as "Ok Google". A silent speech device could avoid this entirely by providing a physical mechanism for the user to start recording what they would like to say instead.

Attention requiring: Existing voice interaction devices have low usability as they require the user to pay full attention to what they're saying and how the device is responding. Also, proximity to the device is required as using any voice system from far away reduces its effectiveness. Silent speech systems would avoid this issue as they would be directly recording what the user is intending to say directly from the electrodes which are attached to the surface of the skin of the user.

The advantages above will benefit all users of sEMG silent speech systems. However, these systems could also uniquely benefit users who have medical issues which make regular speech difficult such as people who suffer from Multiple Sclerosis with Dysphonia. There has already been [promising research](#) into using silent speech devices to help individuals with varying levels of speech impairments which can benefit from silent speech systems, and further research into healthy or unhealthy people will benefit all future users of silent speech systems.

4. Project aim and objectives

The overall aim of this project is to contribute to open-source research concerning sEMG based silent speech systems. This means creating an improved model for sEMG silent speech and acquiring an EMG dataset from participants, including myself, and then open-sourcing the acquired data from each participant who consents to having their data shared in public.

The first and most important objective which will lead to achieving this objective is a combination of using the literature review and open-source code with their datasets to achieve a machine learning model which outperforms the state-of-the-art according to at least a single metric. My initial literature review before starting this project has highlighted the following metrics as being of interest: inference time (milliseconds between the model receiving an input and producing an output) and WER (word-error-rate, a raw measure of the accuracy of the model).

So my initial objectives will be replicating the results of existing sEMG silent speech approaches which are open-source, then creating a model which can either produce a better WER, or to create a model with a similar WER and decrease the inference time of the model. The inference time of the model is important as it can become very uncomfortable for a user using a silent speech system if the system takes more than a few hundred milliseconds to respond. The metrics which I'm optimising along may change depending on how the pilot studies and initial research progress or the direction of research may change entirely.

The second objective is to acquire an OpenBCI device and use it to first acquire EMG data from myself as I will be able to acquire the most data from myself, and then additionally, to acquire EMG data from other participants. The purpose of acquiring data from other participants is to increase the robustness of the model against different people speaking, and this may lead to discoveries for generalising sEMG silent speech technology across different users.

My last objective will be to release my final model and EMG dataset (from consenting participants) as open-source. This means releasing the code and weights of my final model on GitHub, and the EMG dataset using a reliable storage provider who will provide access to the dataset long-term. The specifics regarding the open-sourcing of the model need to be determined as the project progresses, as if the participant refuses to consent to sharing their data, the weights within the trained model would still implicitly contain their data. Related to this objective, I aim to make the model and dataset releases as reproducible as possible by including the code, model, data (consenting participants data, parameters (along with hyperparameters) and environment setup (i.e. using containerisation, etc.) available.

5. Project deliverables

Information system artefacts: Open-source sEMG Silent Speech Machine Learning System
Datasets: Open-source sEMG Silent Speech Dataset (My own sEMG dataset and participants who consent to share their data publically), closed-source sEMG Silent Speech Dataset (Participants who do not consent to share their data publically)
Documentation: Project report

For this research project, the information system artefact will be the machine learning system itself trained with user data and used in conjunction with an OpenBCI device attached to the skin of a user. And the documents which will be produced are the project report which describes the research undertaken, relating to the information system artefact and acquiring data from participants.

6. Project constraints

The key constraint for this project is that it is dependent on using an sEMG device to record EMG data from participants. This requires participants to be physically attached to the device (non-invasively) to make the recordings which means that data can only be gathered from one participant at a time. Also, the data collection process from each participant is time consuming as it may require hours of data from each participant (not including setup, briefing and other mandatory stages). This will limit how many participants can be used within the research because it will take a lot of each individual's time. I will mitigate this by strongly focusing on collecting data from myself in the early stages. This will guarantee I have enough data to train my system and will reduce the amount of data required from other participants.

7. Project approach

The background research I will do for this project will involve how to acquire and set up a surface EMG device as I have no prior experience with EMG / EEG acquisition. However, once I have completed this step, my previous experience in digital signal processing (working with audio datasets with ASR), will help along with looking at the [Digital Voicing of Silent Speech](#) code repository to figure out the rest.

I will establish the research direction of this project by finding prior research during my literature review and identifying directions for future research, as well as trying to replicate previous findings for open-source projects which have code and data available. From these open-source projects and the literature, I attempt to implement new methods after reproducing old ones, either by reimplementing methods from one domain into sEMG silent speech, or by innovating entirely new methods. I will also be performing ablation studies (removing components of a system to determine their contribution) during my project to justify the contribution of individual components and the summary of this will be included in the writeup with full details included in the appendix.

After this, I will use my findings from the previous stage in combination with the data I will acquire from participants to create a new model which improves on the metrics found within my literature review. Which metrics I choose will be guided mainly by the initial model prototyping stage as well as the data acquisition stage. The data acquisition stage will be guided by the literature review and initial attempts at improving on different metrics (word-error rate, inference time in ms, etc).

The methodology I will be using in this project will be a combination of the iterative and incremental approaches as that is the current recommended methodology for MLOps (machine learning operations) projects. Although this project is more of an academic research project than an industry service focused one, I still need to create a working demo of my final model (ideally with the OpenBCI device).

This means that I will need a stable working model as I am going along and as I complete the data acquisition stage and parts of my training infrastructure are finalised (e.g. feature selection, training pipeline, data quality checking, etc.), I will be creating automated systems to allow me to just adjust my model or other parts of the system and then start new training runs. The key idea behind the management of this project's artefacts is CACE (changing anything changes everything). This means that tracking code, datasets and models (along with model parameters) is imperative along with combinations of those three is imperative.

8. Literature review plan

The main starting point for my research is the [Digital Voicing of Silent Speech](#) paper by David Gaddy and Dan Klein and their further paper [An Improved Model for Voicing Silent Speech](#) which improves upon their initial paper. This paper provides the greatest depth of explanation out of any paper along with an accompanying dataset and working code. I will also refer to meta-review papers which sample the entire landscape for sEMG Silent Speech, and silent speech research in general (including fMRI research as taking signals directly from the scalp may also be beneficial).

I will also refer to other seminal works in this field such as the [MIT AlterEgo](#) paper released by Arnav Kapur and colleagues as Kapur also released a 60 page thesis which goes into great detail about the biological, biosensing and machine learning aspects of the problem. The paper released by Kapur and colleagues also describes a Convolutional Neural Network (CNN) based model, which is typically used in complex Digital Signal Processing (DSP) problems such as automated speech recognition (ASR) as CNN based models are highly suitable for DSP due to the convolution theorem.

From these papers, I will research papers which are related in the fields of ASR (due to the high overlap between EMG to speech transduction), encoding data into latent space (due to several recent sEMG related papers finding success with this and particularly performant

general models which encode information into latent space and can be used to perform multiple tasks by querying this latent space representation of data).

After reviewing these papers and related fields and methods, I will collate the most relevant methods and any other relevant findings, and I will trial different methods on the open-source dataset released in the [Digital Voicing of Silent Speech](#) paper to find what is successful. From this, I will use this information to inform the construction of my own machine learning model which I will use firstly on the [Digital Voicing of Silent Speech](#) dataset, then on my own acquired data.

9. Facilities and resources

The main hardware I need for this project is an OpenBCI device which is a relatively cheap, open-source device which is used to record EEG signals and can also be used for recording EMG signals. Depending on which device is used for research, the cost of the device along with electrodes, will be around £500, give or take £150. However, I will need to get this device shipped from either China or the USA, which means I need to account for shipping time as well during this project. The estimated shipping times are around 1 month so this device will need to be ordered as soon as possible in case the shipping is delayed, or the device and it's related components have an issue.

The second most important hardware I need is a modern GPU as I will be training machine learning models using 10s of GBs worth of data. Fortunately I have an RTX 3060 Ti at home which is suitable for at least prototyping machine learning models and I have also built my own library for utilising preemptible (interruptible) cloud processing devices, which means I can use cloud computing for 20% of the price. However, my local GPU should suffice for this project.

10. Log of risks

Description	Impact	Likelihood	Mitigation	First indicator
<i>COVID-19 infection risk between research team and participants</i>	<i>Severe</i>	<i>Likely</i>	<i>Ask participants if they have any symptoms of COVID-19: high temperature, cough, loss of taste</i>	<i>Either the research team starts displaying COVID-19 systems or a participant does after gathering data</i>
<i>EMG dataset data loss</i>	<i>Severe</i>	<i>Unlikely</i>	<i>I will set up automatic encrypted</i>	<i>I am unable to access a part of</i>

			<i>at rest backups for all of the data I gather both locally and on external storage providers</i>	<i>my gathered data or there is a change in size of the original dataset</i>
<i>OpenBCI hardware is damaged during testing (electrodes, main device, etc.)</i>	<i>Severe</i>	<i>Unlikely</i>	<i>I will personally store the OpenBCI device in my home and only take it out when acquiring data. I will also acquire more electrodes than necessary for backup.</i>	<i>Device is either unusable or shows other signs of reporting faulty data or working improperly.</i>
<i>Ten20 Conductive Paste electrode paste allergy</i>	<i>Severe</i>	<i>Depends on the participant</i>	<i>I will not accept any participants who carry a risk of an allergic reaction to the electrode paste. This includes anyone with a history of skin allergies or a history of sensitivity to cosmetics or lotions.</i>	<u><i>Any persistent redness, soreness, burning, itching, or swelling on the skin.</i></u>

My plan for reviewing risks will be identifying what risks have already occurred and what new risks may occur based on changes between different revisions of the risk assessment checklist (e.g. if I am unable to acquire a certain electrode paste and have to get a different brand, checking with the manufacturers guidance on skin irritation, etc.)

11. Project plan

Task Name	Duration	October 2021	November 2021	December 2021	January 2022	February 2022	March 2022	April 2022	May 2022
Google Calendar Tracking	15 wks - 6 mths								
Weekly Writeup (Week Conclusions)	15 wks - 6 mths								
Project Analysis (Project Statistics)	4 wks - 7 wks								
Report Documentation	15 wks - 6 mths								
Weekly Writeup	15 wks - 6 mths								
Writeup Google Calendar Insights	4 wks - 7 wks								
Report Finalisation	3 wks								
Ethical Approval	2 wks - 3 wks								
Submit to Supervisor	1 wk								
Participant/Project Ethics Documents	1 wk								
FEC Discussion (Open-Sourcing)	1 wk								
EMG Data Acquisition Decision (Yes, No)	2 wks - 3 wks								
sEMG Data Gathering	12 wks - 19 wks								
Acquire OpenBCI Device	4 wks - 6 wks								
Shipping Time	4 wks - 6 wks								
Check Device Works	1 wk								
Personal Data Gathering	3 wks - 6 wks								
Pilot Data Gathering	1 wks - 2 wks								
Main Data Gathering	2 wks - 4 wks								
Participant Data Gathering	3 wks - 5 wks								
Pilot Data Gathering	1 wks - 2 wks								
Main Data Gathering	2 wks - 3 wks								
Final Data Compilation and Analysis	2 wks - 4 wks								
Literature Review	15 wks - 6 mths								
Initial Literature Review Notes	1 wks								
Initial Literature Review Draft	1 wk								
Novel Approaches Literature Review Notes	2 wks								
Test Novel Approaches from Literature	4 wks								
Novel Approaches Literature Review Draft	1 wk								
Ongoing Literature Review	6 wks - 15 wks								
sEMG Model R&D (incl. training time)	10 wks - 17 wks								
Digital Voicing & Other Datasets	6 wks - 9 wks								
Data Analysis	1 wk - 2 wks								
Reproduce Results	1 wk - 2 wks								
Test Novel Approaches from Literature	4 wks - 5 wks								
Personal Data (Analysis, Model Training)	3 wks - 6 wks								
Participant Data (Analysis, Model Training)	3 wks - 5 wks								
Project Demo Showcases (Progress, Final)	2 wks								

12. Legal, ethical, professional, social issues (mandatory)

The main security implication of this project is the generation of EMG and possibly audio data unique to each participant. Secure storage and processing of this data is the main security concern and will be achieved by only processing the data on my local device or a trusted cloud computing provider. The data itself will only be named with a unique number instead of any personally identifiable information so if the data is lost, it is not easily identifiable. This will ensure that only I know whose data the unique numbers correspond with. This security obligation is also consistent with the data protection legal requirement.

However, as I intend to open-source as much of this project, including the code and dataset, I will be releasing the data from participants who consent to having their data released publically. This will be added as an additional option in the <> document provided to participants. Participants are fully within their rights to not consent to having their data shared and it will not be included in the open-source model or dataset release. At a minimum I will be releasing the dataset and model which I record of myself, which will likely comprise a large portion of the collected data.

Appendix B

Ethics Review

Certificate of Ethics Review

Project title: sEMG Silent Speech Research

Name:	Tada Makepeace	User ID:	904749	Application date:	14/10/2021 16:55:45	ER Number:	TETHIC-2021-101306
--------------	----------------	-----------------	--------	--------------------------	------------------------	-------------------	--------------------

You must download your referral certificate, print a copy and keep it as a record of this review.

The FEC representative(s) for the **School of Computing** is/are [Philip Scott](#), [Matthew Dennis](#)

It is your responsibility to follow the University Code of Practice on Ethical Standards and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers including the following:

- [University Policy](#)
- [Safety on Geological Fieldwork](#)

It is also your responsibility to follow University guidance on Data Protection Policy:

- [General guidance for all data protection issues](#)
- [University Data Protection Policy](#)

Which school/department do you belong to?: **School of Computing**

What is your primary role at the University?: **Undergraduate Student**

What is the name of the member of staff who is responsible for supervising your project?: **Dr Dalin Zhou**

Is the study likely to involve human subjects (observation) or participants?: Yes

Will peoples` involvement be limited to just responding to questionnaires or surveys, or providing structured feedback during software prototyping?: No

Will the study involve National Health Service patients or staff?: No

Do human participants/subjects take part in studies without their knowledge/consent at the time, or will deception of any sort be involved? (e.g. covert observation of people, especially if in a non-public place): No

Will you collect or analyse personally identifiable information about anyone or monitor their communications or on-line activities without their explicit consent?: No

Does the study involve participants who are unable to give informed consent or in are in a dependent position (e.g. children, people with learning disabilities, unconscious patients, Portsmouth University students)?: No

Are drugs, placebos or other substances (e.g. food substances, vitamins) to be administered to the study participants?: No

Will blood or tissue samples be obtained from participants?: No

Is pain or more than mild discomfort likely to result from the study?: No

Could the study induce psychological stress or anxiety in participants or third parties?: No

Will the study involve prolonged or repetitive testing?: Yes

Will financial inducements (other than reasonable expenses and compensation for time) be offered to participants?: No

Are there risks of significant damage to physical and/or ecological environmental features?: No

Are there risks of significant damage to features of historical or cultural heritage (e.g. impacts of study techniques, taking of samples)?: No

Does the project involve animals in any way?: No

Could the research outputs potentially be harmful to third parties?: No

Could your research/artefact be adapted and be misused?: No

Does your project or project deliverable have any security implications?: No

I confirm that I have considered the implications for data collection and use, taking into consideration legal requirements (UK GDPR, Data Protection Act 2018 etc)

I confirm that I have considered the impact of this work and and taken any reasonable action to mitigate potential misuse of the project outputs

I confirm that I will act ethically and honestly throughout this project

Supervisor Review

As supervisor, I will ensure that this work will be conducted in an ethical manner in line with the University Ethics Policy.

Supervisor's signature: **Dalin Zhou**

Date: 19/10/21

Faculty Ethics Committee Review

Faculty Ethics Committee Member's signature:

Date:

Appendix C

DeepSpeech2 ASR Model

```
SpeechRecognitionModel(  
  (cnn): Conv2d(1, 32, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1))  
  (rescnn_layers): Sequential(  
    (0): ResidualCNN(  
      (cnn1): Conv2d(32, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
      (cnn2): Conv2d(32, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
      (dropout1): Dropout(p=0.1, inplace=False)  
      (dropout2): Dropout(p=0.1, inplace=False)  
      (layer_norm1): CNNLayerNorm(  
        (layer_norm): LayerNorm((64,), eps=1e-05, elementwise_affine=True)  
      )  
      (layer_norm2): CNNLayerNorm(  
        (layer_norm): LayerNorm((64,), eps=1e-05, elementwise_affine=True)  
      )  
    )  
    (1): ResidualCNN(  
      (cnn1): Conv2d(32, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
      (cnn2): Conv2d(32, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))  
      (dropout1): Dropout(p=0.1, inplace=False)  
      (dropout2): Dropout(p=0.1, inplace=False)
```

```

(layer_norm1): CNNLayerNorm(
  (layer_norm): LayerNorm((64,), eps=1e-05, elementwise_affine=True)
)
(layer_norm2): CNNLayerNorm(
  (layer_norm): LayerNorm((64,), eps=1e-05, elementwise_affine=True)
)
)
(2): ResidualCNN(
  (cnn1): Conv2d(32, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (cnn2): Conv2d(32, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (dropout1): Dropout(p=0.1, inplace=False)
  (dropout2): Dropout(p=0.1, inplace=False)
  (layer_norm1): CNNLayerNorm(
    (layer_norm): LayerNorm((64,), eps=1e-05, elementwise_affine=True)
  )
  (layer_norm2): CNNLayerNorm(
    (layer_norm): LayerNorm((64,), eps=1e-05, elementwise_affine=True)
  )
)
)
(fully_connected): Linear(in_features=2048, out_features=512, bias=True)
(birnn_layers): Sequential(
  (0): BidirectionalGRU(
    (BiGRU): GRU(512, 512, batch_first=True, bidirectional=True)
    (layer_norm): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False)
  )
  (1): BidirectionalGRU(
    (BiGRU): GRU(1024, 512, bidirectional=True)
    (layer_norm): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False)
  )
  (2): BidirectionalGRU(

```

```

        (BiGRU): GRU(1024, 512, bidirectional=True)
        (layer_norm): LayerNorm((1024,)), eps=1e-05, elementwise_affine=True)
        (dropout): Dropout(p=0.1, inplace=False)
    )
    (3): BidirectionalGRU(
        (BiGRU): GRU(1024, 512, bidirectional=True)
        (layer_norm): LayerNorm((1024,)), eps=1e-05, elementwise_affine=True)
        (dropout): Dropout(p=0.1, inplace=False)
    )
    (4): BidirectionalGRU(
        (BiGRU): GRU(1024, 512, bidirectional=True)
        (layer_norm): LayerNorm((1024,)), eps=1e-05, elementwise_affine=True)
        (dropout): Dropout(p=0.1, inplace=False)
    )
)
(classifier): Sequential(
  (0): Linear(in_features=1024, out_features=512, bias=True)
  (1): GELU()
  (2): Dropout(p=0.1, inplace=False)
  (3): Linear(in_features=512, out_features=29, bias=True)
)
)

```

References

- Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A., Jun, B., LeGresley, P., Lin, L., ... Zhu, Z. (2015). Deep speech 2: End-to-end speech recognition in english and mandarin. arXiv. <https://doi.org/10.48550/ARXIV.1512.02595>
- Gaddy, D., & Klein, D. (2020). Digital voicing of silent speech.
- Ito, K., & Johnson, L. (2017). The lj speech dataset.