Automatic Speech Recognition for sEMG Silent Speech

Tada Makepeace

School of Computing
Final Year Research Project

May 6, 2022

Abstract

In this dissertation, I present my work on a novel training regime which greatly reduces the amount of training data required for electromyography (EMG) silent speech classification systems. This approach also decreases the end-to-end inference time of the classification by removing a key component requried in previous approaches. The proposed approach has a similar word-error rate (WER) to a previous SOTA approach, both being around 68%, but requires far less training data (x115 times less) and inference speed is improved.

EMG augmentation via data synthesis is also explored, however this approach is less successful. Nonetheless, the findings from the research is included in this report to help future researchers avoid fruitless methods in the future.

Experimental results, visualisations, findings and ablation studies are provided for the silent speech classification task, along with intuitions for how the approach was discovered along with ablation studies highlighting what type of is most useful for the approach.

Ultimately this dissertation provides a novel and efficient method for greatly improving the training and runtime efficiency of future EMG silent speech classification systems.

Consent to share

I consent for this project to be archived by the University Library and potentially used as an example project for future students.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Thanks.

Chapter 1

Introduction

1.1 Project Background

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 silently articulated speech devices are a human computer interface which offer 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 it's 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 skin of the user.

The above examples will benefit all users of 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. (Kapur et al., 2020)

1.2 Project Objectives

The primary objectives for this project are threefold:

- *Improve SOTA Approach:*
- *EMG Data Acquisition:*
- *Open-Source Release:*

The deliverables for this research project are an open-source sEMG silent speech machine learning system which outperforms the current SOTA approach (text classification or speech synthesis) based on a key evaluation metric of that system.

Chapter 2

Literature Review

This section of the report identifies the early research directions which were considered, along with explaining basic terms in the EMG silent speech domain. The goal of this section is to briefly familiarise the reader with the silent speech domain and silent speech interfaces in general. Then the two key approaches to silent speech systems are explored, silent speech to audio and silent speech to text.

2.1 Key Terms - Silent Speech

- *Word Error Rate (WER):* Metric used in speech recognition systems which accounts for differences in length of a recognised word sequence and a reference word sequence. WER is derived from Levenshtein distance which is a string metric for measuring the difference between two sequences. The Levenshtein distance between two words is essentially the minimum number of single-character edits (including insertions, deletions or substitutions) required to change one word into another. Another term for it is the "edit distance".

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

Where each term means:
S: Number of substitutions
D: Number of deletions
I: Number of insertions
C: Number of correct words
N: Number of words in the reference (N = S+D+C) (Levenshtein, 1966)

- *Mel-Scale:* Perceptual scale of pitches judged by listeners to be equal in distance from one another.

$$m = 2595 \log_{10} \left(1 + \frac{f}{700}\right)$$

(O'Shaughnessy, 1987)

- *Silent Speech Interface (SSI):* "A silent speech interface (SSI) is a system enabling speech communication to take place where an audible acoustic signal is unavailable. By acquiring sensor data from elements of the human speech production process, an SSI produces a digital representation of speech which can be synthesized directly, interpreted as data or routed into a communications network" (Denby et al., 2010).

- *Surface Electromyography (sEMG):* Non-invasive, computer-based technique that records the electrical impulses placed on the surface of the skin overlying the nerve at rest (i.e. static) and during activity (i.e. dynamic).

2.2 Machine Learning Background

2.2.1 David Gaddy and Dan Klein

David Gaddy and Dan Klein have published two papers relating to the transduction of silent speech. These two papers are of particular interested for silent speech research as they describe a novel method of transducing silent speech into speech features. This makes it easier to produce speech synthesis

and text classifiction systems for silent speech research. Also, the original paper also contains an accompanying open-source dataset which is comparable to smaller speech recognition datasets such as LJSpeech (Ito and Johnson, 2017). This makes the two papers and their open-source dataset critical to the success of this dissertation, as only on in the project it was decided to not pursue the EMG data acquisition.

Original Paper

The paper digitally voices silent speech from silently mouthed words which are converted to audible speech based on EMG sensor measurements which capture muscle movements. The novel contribution of this paper is a method of transducing speech features directly from EMG recordings of a person's face during silently articulated speech. The main innovative approach suggested in this paper is to transfer audio targets from vocalised EMG recordings to silent EMG recordings by learning the alignment between silent EMG recordings and ground truth audio features.

For each utterance in the dataset, silent EMG and vocalised EMG recordings are recorded in two different sessions. However, only the vocalised EMG recording also has an associated ground truth audio file (as silent speech involves not actually producing noise when speaking). Then the silent EMG recording is aligned with the vocalised EMG recording. This makes it possible to directly transduce silent EMG recordings into speech.

(Gaddy and Klein, 2020)

An Improved Model for Voicing Silent Speech

Evolution of Digital Voicing paper which improves model performance compared to previous work. Model learns its own input features directly from EMG signals instead of hand-crafted features in prior work. New model uses convolutional layers to extract features from the signals. Transformer-encoder layers used to propagate signals across longer distances instead of

bi-directional LSTM layers. Additional signal introduced during learning by introducing auxillary task of predicting phoneme labels in addition to predicting speech audio features. On open vocabulary intelligibility evaluation, the new model improves the absolute WER by 25.8%. (Gaddy and Klein, 2021)

Establishes new SOTA performance via 3 main improvements:

- *Replace LSTMs with Transformer-Encoder:* Model bi-directional triple stacked 1024-unit LSTM layer replaced with 6 transformer-encoder layer to predict EEG features found using CNN over 1 second to predict speech along with phoneme loss signal. Ablation shows WER go from 42.2% to 45.2% when this is removed.
- *Introduce Phoneme Loss:* Introduced auxillary phoneme prediction task to as additional output of self-attention transformer-encoder. This signal is used to improve the performance of the network by learning to predict the phoneme of the speech which is being detected and also regularises training. Ablation shows WER go from 42.2% to 51.7% when this is removed.
- *Replace hand-designed features with convolution features:* Uses CNN block based on ResNet to extract features from EMG signals which are end-to-end trainable instead of using hand-crafted features to improve the representational power of the network. Ablation shows WER go from 42.2% to 46% when this is removed.

(Gaddy and Klein, 2021)

2.3 Silent Speech Interfaces for Speech Restoration: A Review

2.4 Proposed Directions for Research

2.4.1 Text Classification

2.4.2 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.

One possible data augmentation method for sEMG based silent speech is to reverse a SOTA transduction model. Existing transduction models for EMG to speech use either EMG features or raw EMG data and transduce it into speech features (typically mel spectrograms or MFCCs). However, it may be possible to predict the EMG data from the mel spectrograms.

2.4.3 Improving Existing Models

Model Improvements

Recently there has been more research into silent speech models which can directly produce text from muscle activity. One paper introduces a sequence-to-sequence voice reconstruction model and training regime called *SSRNet* to address what they call the sEMG2V problem in a tonal language.

combines multiple methods into one to achieve strong performance in synthesizing speech in Mandarin Chinese from EMG signals. The authors of this paper propose a complex architecture, where the model is based on the FastSpeech (Ren et al., 2019) neural network architecture.

The FastSpeech neural network architecture is to convert text, it it's phonemic representation into mel spectrograms. Compared to other speech syn-

thesis models, FastSpeech includes novel neural network components. These include a Feed-Forward Transformer, FFT Block, Length Regulator and Duration Predictor. The purpose of these neural network blocks is to improve the abiltiy of the model to correctly predict the duration of a phoneme. This is particularly important because for text to speech models, it is difficult to determine the alignment of the inputs to the outputs as the only information which is provided is text, which does not contain timing information.

2.5 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. It was originally proposed in a 2006 paper (Graves et al., 2006) and specifically dealt with training recurrent neural networks (RNN) (Sherstinsky, 2020).

Chapter 3

Research Overview

This section details the research which was conducted during this project at a high level. The individual approaches futher on contain the datasets, experimental conditions, assumptions, related work and evaluation metrics along with justification for each individual approach.

3.1 Full List of Research Questions with Hypotheses

3.1.1 Data Augmentation via EMG Synthesis

Question: How feasible is the synthesis of silent speech EMG signals as a data augmentation technique for training EMG silent speech systems?
Hypothesis: Training an EMG synthesis model based on a small EMG dataset will improve the final performance of transduction model
Measurable: Test loss of the final transduction system when trained with augmented EMG data versus without
Measurable: Visual inspection of the synthesized EMG signals
Measurable: Visual inspection of the predicted mel spectrograms when traind with augmented EMG data versus without

3.1.2 Fine Tuning Speech Recognition on Model Predictions

Question: Can fine tuning an ASR model on a silent speech transduction models mel spectrogram predictions improve performance for text classification?

Hypothesis: Fine tuning an ASR model with the predicted mel spectrograms from a pre-trained silent speech transduction model can improve inference time and reduce the required training dataset while maintaining competitive accuracy

Measurable: WER of final system compared to SOTA for text classification

Measurable: Dataset size required to train the ASR system in hours

Measurable: Estimated end-to-end inference time for the system

Chapter 4

Methodology

4.1 Software Development Life Cycle (SFLC)

The methodology selected for this project was a mixture of the incremental and iterative software development lifecycles. This is a typical methodology selected for machine learning projects as the exploratory nature of many machine learning projects necessitates a management method which allows for the flexible changing of approaches throughout the course of the project.

At a high level, the project is managed using an incremental approach whereby I found promising ideas from my on-going literature review and personal exploration of the EMG silent speech dataset.

Then when I found an approach which I believed was promising, I conducted initial small-scale experiments to either validate or disprove my initial assumption, and then scaled up my experiments iteratively to verify whether or not my hypothesis was valid.

The evaluation metric for each incremental approach was selected based on the common evaluation metric for that task as reported in the literature.

Chapter 5

Datasets

This section is provided to explain the origin of the datasets for this thesis, why they were chosen, the particular slices of the dataset that were used for the different experiments and datasets that were generated for the purpose of this project.

The datasets used for this project are based on the open-source release of data accompanying the Digital Voicing of Silent Speech (**gaddy2020improved**) paper.

Chapter 6

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.

6.1 Data Preprocessing

6.1.1 Audio Speech Features

For the purposes of the fine tuning experiments, the ground truth audio files are preprocessed into mel spectrograms rather than MFCCs as in the original two papers (Gaddy and Klein, 2020, Gaddy and Klein, 2021). The reason for this is because mel spectrograms contain more raw data than MFCCs which makes them more appropriate for speech recognition and also the vocoder used for this paper uses mel spectrograms as the input, as do most SOTA vocoders.

6.2 Data Augmentation via EMG Synthesis

One approach for improving the performance of any machine learning model is to synthesize more data. This is particularly useful if the original dataset is small and there are methods to synthesize more data.

6.2.1 Related Work

Previous approaches have used various deep learning techniques to synthesize more EMG data to train EMG deep learning models. One approach (J. J. Bird et al., 2021) uses a GPT-2 (Radford et al., 2018) like model to synthesis EMG signals for simple action recognition such as grasp and release (actions common to robotic prosthetics and manipulators). The inclusion of synthesized EMG data during the training process improved the overall gesture recognition accuracy from 68.29% to 89.5%.

Another related paper from the same author experiments with LSTM and GPT-2 models for synthesizing more speech for a speaker recogntion task. The best model found by the authors for this task was a 3-layer, 128 hidden dimension LSTM network (J. Bird et al., 2020).

6.2.2 Research Design

My formal hypothesis for this section is that it is possible to train an EMG augmentation model for voiced EMG data which can improve the performance of a regular transduction model by training on the ground truth voiced EMG data and the synthesized voiced EMG data.

6.2.3 Summary

In summary, my hypothesis was disproved. My initial hypothesis was that it was possible to simply reverse the transduction network, which was introduced in the Digital Voicing paper for transcribing from EMG features into speech features, but instead reverse the features. From my findings I found

that this wasn't true because the low level features of the EMG data are difficult for the LSTM network to correctly reproduce.

In hindsight my approach could have been improved by being more selective in the early stages. I could have determined which electrode contributed the most in the transduction task, and then tried to just synthesise the signals for that particular channel and then tried to synthesise an increasing number of electrode channels.

6.3 Fine Tuning Speech Recognition on Model Predictions

This section details an improved approach for training a silent speech recognition system by pre-training a DeepSpeech2 model directly on the predictions of a chosen portion of the dataset.

This is better than performing speech recognition on the final waveform generated in the Digital Voicing (Gaddy and Klein, 2020) paper.

Although the method used in that paper isn't used directly for speech recognition, the end-to-end evaluation procedure can be used for speech recognition by transducing from from the silent EMG data into speech features, then using a vocoder to go from speech features into an audio waveform and then using a pretrained ASR model to predict speech from the waveform.

The proposed method skips over the vocoder entirely which improves the end-to-end performance as there is now one less step involved.

6.3.1 Related Work

The intuition behind fine tuning a speech recognition model on the predicted mel spectrograms from the transduction model comes from the literature for speech recognition. Typical speech recognition models suffer from a loss of accuracy when they are only trained on clean audio data and then evaluated

on a noisy dataset (Amodei et al., 2015). For example, in the DeepSpeech2 paper the authors trained their speech recongition model on different fractions of a dataset.

Their results showed that the performance of a model trained on a clean dataset, when evaluated on a dataset of clean audio compared to noisy audio, showed a larger gap when trained on less data. When the authors trained their model on 1% of the entire dataset (120 hours), the WER for the clean dataset was 29.23% whereas for the noisy dataset it was 50.97%. However when they trained on the full dataset (12,000 hours), the clean audio WER was 8.46% compared to 13.59% for the noisy dataset.

The results from their paper show a strong relationship between datasize and affect on evaluation performance on a noisy dataset. For this reason, this paper researches the affect of providing predicted speech from the SOTA silent speech transduction model to a speech recogntion model to improve it's performance for speech recognition and speech synthesis (Amodei et al., 2015).

6.3.2 Relation to EMG Silent Speech Classification

6.3.3 Research Question

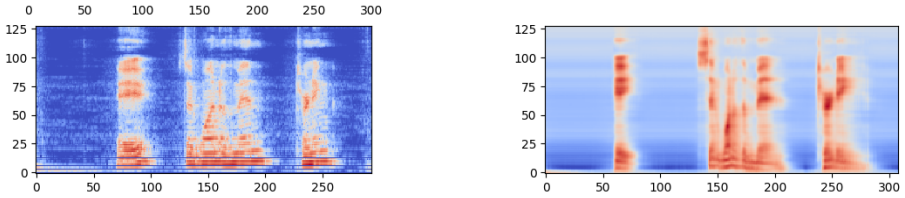
6.3.4 Hypothesis

My hypothesis based on my literature review,

6.3.5 Closed Vocabulary Dataset

The closed vocabulary dataset is the first dataset used to determine how much better a speech recognition model can be improved by training on the predicted mel spectrograms from an already trained transduction model.

The WER of the closed vocabulary ASR model trained only on the training dataset of the closed vocab dataset is 37%. This means that we would not



(a) Ground Truth Mel Spectrogram **(b)** Predicted Mel Spectrogram

Figure 6.1: Comparison of Ideal Ground Truth and Predicted Mel Spectrograms From both of these mel spectrograms, we can see the the low level features of the predicted mel spectrogram appear blurrier (the curvy red lines). This is because the transduction model uses a simple linear layer (MLP) between the final transformer layer and the output. This limits the ability of the model to produce more fine tuned results compared to using a convolutional layer. Not pictured in the figures, is the text transcription for the utterances which is: '"Eh?" said one of the men, turning.'.

expect the WER of the model, when it's evaluated on transduced examples, to perform better than the ground truth word-error rate.

Table 6.1: DeepSpeech2 Closed Vocab Finetuned Results

Dataset	**CER**	**WER**
Ground Truth (GT)	87.10	100.50
Voiced	51.27	87.33
Silent	38.80	78.10
Silent, Voiced	**35.26**	75.33
Silent, Voiced, GT	35.72	**70.83**

The above results do not use phonemeic prediction and only use greedy decoding in the decoder. DeepSpeech2 is also not the SOTA model for speech recognition which also reduces the performance. This means that the best WER of 70.83% could be improved upon a lot more.

For the combined silent, voiced and GT condition, the model is pre-trained on the ground truth model and and then fine-tuned on the predictions of the silent and voiced utterances.

Here we can see that training on predictions from both modalities from scratch is better than training on a single modality only. However the model is only evaluated on silent EMG text classifications which means that for this experiment, training on silent and voiced predicted speech improves the performance when evaluating on only the silent EMG predictions.

This may be because both of the individual datasets are small (only 400 utterances) so doubling the dataset to 800 utterances gives the model far more examples to learn from, even though the voiced examples diverge from the silent examples.

6.3.6 Open Vocabulary (Parallel-Only) Dataset

The next dataset used is the parallel voiced and silent EMG data. This is used because it's smaller than using the entire open vocabulary condition dataset in one go which makes training the individual ASR models and the transduction model faster.

The WER of the ASR model trained only on the training dataset of the open vocabulary parallel voiced audio is 64.47%.

Table 6.2: DeepSpeech2 Open Vocabulary Parallel Finetuned Results

Dataset	**CER**	**WER**
Ground Truth (GT)	66.70	110.49
Voiced	156.29	100.00
Silent	44.21	84.19
Silent, Voiced	42.11	85.30
Voiced, GT	43.06	84.65
Silent, GT	41.24	83.97
Silent, Voiced, GT	**41.02**	**81.69**

Due to the results of the closed vocabulary training condition and time constraints, training on the voiced portion of the open vocabulary paralell mel spectrogram predictions wasn't conducted.

There is one interesting difference in this training run compared to the closed

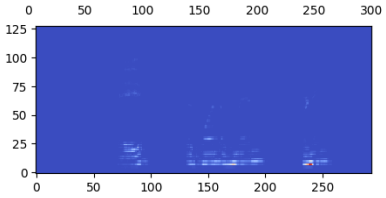
vocabulary dataset. Training on the silent EMG and voiced EMG conditions together reduces performance. This may be because training on both together was better on training on just the silent EMG signals for the closed vocabulary dataset because the silent EMG dataset only contained 400 data samples so doubling the dataset to 800 data samples, even though the vocalised samples are sub-optimal, is better. However, here the silent EMG dataset is comprised of 2,778 data samples which means that the difference between the silent EMG and voiced EMG predicted mel spectrograms is reducing the performance of the ASR model more than having a larger overall dataset is beneficial.

6.3.7 Open Vocabulary Full (Parallel and Non-Parallel) Dataset

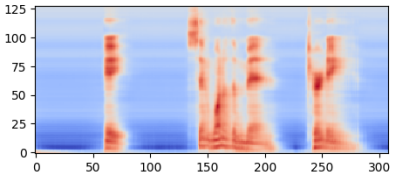
This dataset is the full entire dataset, not including the closed vocabulary dataset. The experiments for this dataset include the same experiments for the previous two datasets. However, extra experiments are added for the additional non-parallel vocal EMG mel spectrogram predictions. The paralell silent and voiced mel spectrogram predictions are included here to show how predicted speech features from a model trained with more data improve the predictions of the same dataset. This has the effect of making the predicted mel spectrograms closer to the ground truth mel spectrograms which means the model is better able to classify the text as the dataset for the entire training regime is closer to the same underyling distribution.

The WER for the ground truth model evaluted on the audio is 45%. This means that we wouldn'êxpect the model to have a lower WER than this value.

Table 6.3: DeepSpeech2 Open Vocabulary Full Dataset Finetuned Results



(a) Ground Truth Mel Spectrogram



(b) Predicted Mel Spectrogram

Figure 6.2: Comparison of Real Ground Truth and Predicted Mel Spectrograms From both of these mel spectrograms, we can see the the low level features of the predicted mel spectrogram appear blurrier (the curvy red lines). This is because the transduction model a simple linear layer (MLP) between the final transformer layer and the output. This limits the ability of the model to produce more fine tuned results. Not pictured in the figures, is the text transcription for utterances which is: '"Eh?" said one of the men, turning.'.

Dataset	**CER**	**WER**
Ground Truth (GT)	61.99	100.74
Parallel Voiced	–	-
Silent	-	-
Silent, Parallel Voiced	-	-
Silent, Parallel Voiced, GT	32.48	**68.26**
Silent, Parallel Voiced, Non-Parallel Voiced, GT	**31.16**	68.51

Here we can see that when the transduction model is trained on the full dataset, it produces the best final WER. However there is also another interesting finding, fine tuning the ground truth model on the silent, parallel voiced and non-parallel voiced harms the overall WER but the CER continues to improve. This means that the model is better able to predict individual characters but it's ability to predict words has been slightly reduced.

6.3.8 Implementation Errors

One interesting implementation error that was found after conducting the experiments was that the ground truth audio files have not had the exact same preprocessing techniques applied as the target mel spectrograms which

the transduction model uses as a ground truth target during training, which can be seen in Figure 6.2.

The main difference is that the mel spectrogram values from the transduction model has had the log() function applied. This is a common preprocessing technique in machine learning used to make it easier for a regression model to predict a real valued output. This finding may confound the results for the ground truth results in the above experiments. Interestingly, training the ASR model on the grouth truth values with a different range still results in better performance. This makes it unclear whether the ground truth values being within a different range actually helps or hurts the performance of the final model.

6.3.9 Summary

In summary, this approach proves to be very successful for producing an efficient EMG silent speech recognition system which can match the performance of an old transducer model, re-purposed for ASR, while the speech recognition portion of the new proposed system is only trained on around 33 hours of data versus 3,817 (Mozilla, 2020) of data.

Chapter 7

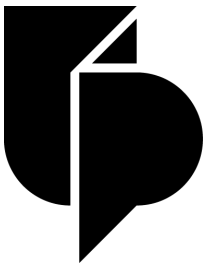
Conclusion

Chapter 8

Future Work

Appendix A

Project Initiation Document



UNIVERSITY OF PORTSMOUTH

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 it's 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

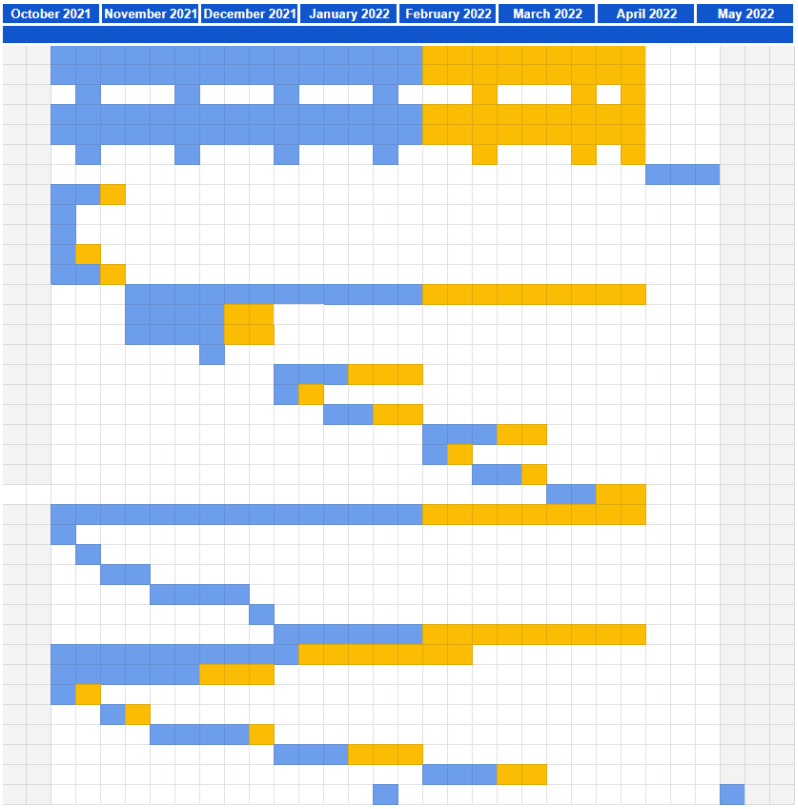
Description	Impact	Likelihood	Mitigation	First indicator
COVID-19 infection risk between research team and participants	*Severe*	*Likely*	*Ask participants if they have any symptoms of COVID-19: high temperature, cough, loss of taste*	*Either the research team starts displaying COVID-19 systems or a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

			at rest backups for all of the data I gather both locally and on external storage providers	*my gathered data or there is a change in size of the original dataset*
OpenBCI hardware is damaged during testing (electrodes, main device, etc.)	*Severe*	*Unlikely*	*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.*	*Device is either unusable or shows other signs of reporting faulty data or working improperly.*
Ten20 Conductive Paste electrode paste allergy	*Severe*	*Depends on the participant*	*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.*	*Any persistent redness, soreness, burning, itching, or swelling on the skin.*

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
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 - Automatic Speech Recognition (ASR)

Name:	Tada Makepeace	**User ID**:	904749	**Application date**:	05/05/2022 12:45:02	**ER Number**:	TETHIC-2022-103171

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 **Haythem Nakkas, David Williams**

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?: No
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
Will your project or project deliverables be relevant to defence, the military, police or other security organisations and/or in addition, could it be used by others to threaten UK security?: 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**:** 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)
  )
)
```

Appendix D

Hyperparameters

Throughout the course of this research, many different machine learning models were trained to evaluate their performance on tasks. During these experiments, the hyperparameters of the individual models were tweaked to improve the performance of the models on the individual tasks. This section lists the final hyperparameters of the individual networks.

D.1 DeepSpeech2 Audio Hyperparameters

Table D.1: DeepSpeech2 Audio Model Hyperparameters

Hyperparameter	Description	Value
n_cnn_layers	Number of CNN block layers	3
n_rnn_layers	Number of RNN block layers	5
n_class	Number of model classes	Vocabulary size
n_feats	Number of CNN block features	128
stride	CNN stride value	2
dropout	Dropout value	0.1
learning_rate	Initial learning rate	5e-4
batch_size	Examples per single minibatch	5

D.2 Silent Speech Transduction Hyperparameters

D.2.1 Improved Voicing Transducer Model used in Fine Tuning ASR

The hyperparameters for the transduction model were reduced from the original SOTA model due to hardware constraints. In the Additional Reproducability Information of the second Digital Voicing (Gaddy and Klein, 2021) paper, they state that to train the full model takes roughly 12 hours on an Nvidia Quadro RTX 6000 which has 24GB VRAM. The GPU I was experimenting on only had 8GB VRAM which meant that I had to halve the number of transformer layers and reduce the hidden dimension size. For reference, on an Nvidia RTX 3060 with Automated Mixed Precision enabled on PyTorch, training the transduction model with these settings on the full dataset took 4 hours and 49 minutes.

Table D.2: SOTA Silent Speech Transduction Model

Hyperparameter	Description	Value
n_transformer_layers	Number of Transformer Encoder layers	6 to 3
stride	ResNet Block stride value	1
learning_rate	Initial learning rate	1e-3
batch_size	Examples per single minibatch	32
epochs	Number of full dataset training iterations	80
model_size	Size of hidden dimension for transformer encoder	768 to 512

Appendix E

Silent Speech - ASR Examples

The top 10 most accurate transcriptions predicted by the silent speech recognition system are provided below. These results can be reproduced by following the evaluation instructions on [1]

1.
Target:
Prediction:
WER: 0.000000 CER: 0.000000
2.
Target: iv
Prediction: iv
WER: 0.000000 CER: 0.000000
3.
Target: where are you going i asked
Prediction: where are you going i ast
WER: 0.166667 CER: 0.063830
4.

[1] https://github.com/MiscellaneousStuff/semg-asr

Target: the place was impassable
Prediction: the place was impassedabel
WER: 0.250000 CER: 0.275862
5.
Target: and so forth
Prediction: and so farth
WER: 0.333333 CER: 0.058824
6.
Target: that was it
Prediction: that was dede
WER: 0.333333 CER: 0.250000
7.
Target: then suddenly the white flashes of the heat-ray came leaping towards me
Prediction: then suddenly the white flashes of the headray came at leaping towaeasman
WER: 0.333333 CER: 0.131579
8.
Target: i saw astonishment giving place to horror on the faces of the people about me
Prediction: i saw u stonishment kaving macsed order on the faces of the people abou me
WER: 0.400000 CER: 0.170732
9.
Target: the shell burst clean in the face of the thing
Prediction: the chill maghs geleane and the face of the thing
WER: 0.400000 CER: 0.235294
10.
Target: he turned stared bawled something about crawling out in a thing like a dish cover and ran on to the gate of the house at the crest
Prediction: he tone stared palled something about crawling out in tae think

lckid dish comber and ran on to the gade of the house at the grest
WER: 0.407407 CER: 0.143750

Appendix F

Audio Preprocessing

This section contains the mel spectrogram options applied to all of the audio files, whether they were transduced or just produced from the ground truth audio files.

Table F.1: DeepSpeech2 and Transducer Mel Spectrogram Hyperparameters

Setting	Description	Value
sample_rate	Input audio sampling rate	16000
n_mels	Number of mel-filterbank bins	128 bins
hop_length	Length of non-intersecting portion of window length	160 samples
win_length	Number of samples to consider for each window	432 samples
n_fft	Number of Fast Fourier Transform bins	512 bins

The preprocessing pipeline for the training set of the ground truth audio models included extra data augmentation steps. These techniques are regularly used in ASR models as they have been shown to regularise the training of the networks.

Figure F.1: Ground Truth Audio Trainset Preprocessing Pipeline

```
train_audio_transforms = nn.Sequential(
    torchaudio.transforms.MelSpectrogram(
        sample_rate=16_000,
        n_mels=128,
        hop_length=160,
        win_length=432,
        n_fft=512,
        center=False),
    torchaudio.transforms.FrequencyMasking(
                                    freq_mask_param=15),
    torchaudio.transforms.TimeMasking(time_mask_param=35)
                                    )
```

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

Bird, J., Faria, D., Ekárt, A., Premebida, C., & Ayrosa, P. (2020). *Lstm and gpt-2 synthetic speech transfer learning for speaker recognition to overcome data scarcity.*

Bird, J. J., Pritchard, M., Fratini, A., Ekárt, A., & Faria, D. R. (2021). Synthetic biological signals machine-generated by gpt-2 improve the classification of eeg and emg through data augmentation. *IEEE Robotics and Automation Letters*, *6*(2), 3498–3504. https://doi.org/10.1109/LRA.2021.3056355

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., & Brumberg, J. (2010). Silent speech interfaces [Silent Speech Interfaces]. *Speech Communication*, *52*(4), 270–287. https://doi.org/https://doi.org/10.1016/j.specom.2009.08.002

Gaddy, D., & Klein, D. (2020). Digital voicing of silent speech.

Gaddy, D., & Klein, D. (2021). An improved model for voicing silent speech.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, In *Proceedings of the 23rd international*

conference on machine learning, Pittsburgh, Pennsylvania, USA, Association for Computing Machinery. https://doi.org/10.1145/1143844.1143891

Ito, K., & Johnson, L. (2017). The lj speech dataset.

Kapur, A., Sarawgi, U., Wadkins, E., Wu, M., Hollenstein, N., & Maes, P. (2020). Non-Invasive Silent Speech Recognition in Multiple Sclerosis with Dysphonia (A. V. Dalca, M. B. McDermott, E. Alsentzer, S. G. Finlayson, M. Oberst, F. Falck, & B. Beaulieu-Jones, Eds.). In A. V. Dalca, M. B. McDermott, E. Alsentzer, S. G. Finlayson, M. Oberst, F. Falck, & B. Beaulieu-Jones (Eds.), *Proceedings of the machine learning for health neurips workshop*, PMLR. http://proceedings.mlr.press/v116/kapur20a.html

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*, 707.

Mozilla. (2020). Deepspeech 0.7.0 release. GitHub. https://github.com/mozilla/DeepSpeech/releases/tag/v0.7.0

O'Shaughnessy, D. (1987). *Speech communication: human and machine*. Addison-Wesley.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. arXiv. https://doi.org/10.48550/ARXIV.1905.09263

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306