Exploration of Novel Training Methods for 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 and also requires far less training data for the speech recognition network component of the system (x115 times less). The proposed approach has a similar word-error rate (WER) to a previous SOTA approach, both being around 68%.

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 analysis highlighting what type of dataset is most useful for the approach.

EMG augmentation via data synthesis is also explored, however this approach is less successful but is left in the report to help other researchers avoid the issues this project encountered.

Ultimately this dissertation provides a novel and efficient method for greatly improving the training time, dataset size requirements 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

I would like to say thank you to my supervisor, Dalin, for his continued support over the last 2 years.

And I would also like to say thank you to David Gaddy for his amazing papers "Digital Voicing of Silent Speech" and "An Improved Model for Voicing Silent Speech".

After spending a long time researching and experimenting with the methods within both papers, I have come to appreciate how significant the contributions of both papers are, not only for EMG based silent speech, but possibly others as well.

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 Aims

The overall aim of this project is to contribute to open-source research concerning sEMG based silent speech systems. Originally this meant creating an improved model for sEMG silent speech and using an EMG dataset acquired from myself and other participants and then open-sourcing the new model and dataset.

However after I decided to not pursue the data acquisition part of my project, the aim of my project was instead, still focused on the research and development of novel methods to improve sEMG silent speech systems.

1.3 Project Objectives

As eluded to in the project initiation document, there are three objectives for this project, with one primary objective, one supplementary objective and a another objective which was mainly reliant on the primary objective.

The first and most important objective to achieving the aim of this project was to discover methods within the EMG silent speech literature and intuitions gained from analysing an open-source silent speech dataset to determine how to improve on existing methods. Formally, my objective here was to improve on existing methods by outperforming a state-of-the-art model on a task based on at least one evaluation metric. Before I commenced this project, I found the following two metrics were of particular interest within the silent speech domain: 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 a model) and training dataset sizes (number of hours of data required to train a system).

The second objective for my project was to purchase an EMG data acquisition device and then create a new EMG silent speech dataset from at least myself, and if possible from willing participants. The purpose of this was to address the gap I found when trying to find open-source EMG silent speech datasets.

The third objective of this project was to release an EMG silent speech dataset to the public which could further be used by other researchers to accelerate progress in the silent speech research domain. The other part of this objective was to release the improved approach for silent speech into the public domain as an open-source repository on GitHub. This would make it available to any future researchers to build upon.

Chapter 2

Literature Review

The purpose of this literature review is to familiarise readers with basic concepts, key to understanding silent speech systems. Also the background research for each of the approaches is briefly explored. However, each approach also contains a further critical review of the literature for the chosen method.

2.1 Key Terms

A list of silent speech and speech recognition domain specific key terms are provided to familiarise the reader with certain concepts to help the reader digest this report.

- *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-Spectrogram:* A regular spectrogram is a visualisation of the frequency spectrum of a signal. Whereas a mel spectrogram is the same thing but the frequency values use the mel scale rather than their raw values. The mel scale is a scale of pitches judged by humans to be equal in distance from another and is commonly used in speech processing systems as the scale normalises the raw frequency values based on the scale of pitches. This makes it easier to produce speech synthesis or speech recognition systems for machine learning models.

- *Transduction/Transducer:* A transducer is an electronic device that converts energy from one form to another. In the context of this project, this refers to converting signals recorded using electrodes from a persons body into speech signals.

2.2 Silent Speech Transduction

There are two primary landmark papers released in the silent speech domain which this project takes large inspiration from. These are "Digital Voicing of Silent Speech" (Gaddy and Klein, 2020) and "An Improved Model for Voicing Silent Speech" (Gaddy and Klein, 2021).

This direction of research is the first successful attempt at transducing speech directly from EMG signals recorded when a person is silently articulating speech.

The original paper discusses how training silent speech models purely on biosignals recorded during the silent articulation of speech leads to a poor training signal due to differences between silent and vocalized articulation of speech when transducing speech from this data. The main reason for this is the silent speech data has no time-aligned audio to use as a training target.

The original paper addresses this issue by using a set of utterances recorded during silent and vocalized speech and finding alignments between the two recordings and finding an alignment between both instances, using Dynamic Time Warping.

The authors also released their dataset used in their research to enable other authors to research silent speech. This makes it very suitable for this project as not only is it free to access, it has also been proven to be useful as the authors are able to use it to create the first state-of-the-art transduction model for EMG silent speech.

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%.

The new state-of-the-art performance is established using 3 different methods compared to the previous paper.

- *Replace LSTMs with Transformer-Encoder:* Model bi-directional triple stacked 1024-unit LSTM layer replaced with 6 layer Transformer-Encoder layer to predict EMG features found using a ResNet inspired block 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 as an additional output of the self-attention transformer-encoder. This self-supervisory 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 a 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 Data Augmentation via EMG Synthesis

2.3.1 Data Augmentation Background

Data augmentation refers to techniques which are used to either increase the amount of available training data by adding slightly modified copies of the original data or to synthesize new examples entirely. The purpose of this is to regularise the training and help reduce overfitting when training a machine learning model (Shorten and Khoshgoftaar, 2019).

Many methods exist to generate augmented data for machine learning. One of them is to apply geometric transformations such as: translations, rotations, cropping, flipping, scaling, etc. (Shorten and Khoshgoftaar, 2019).

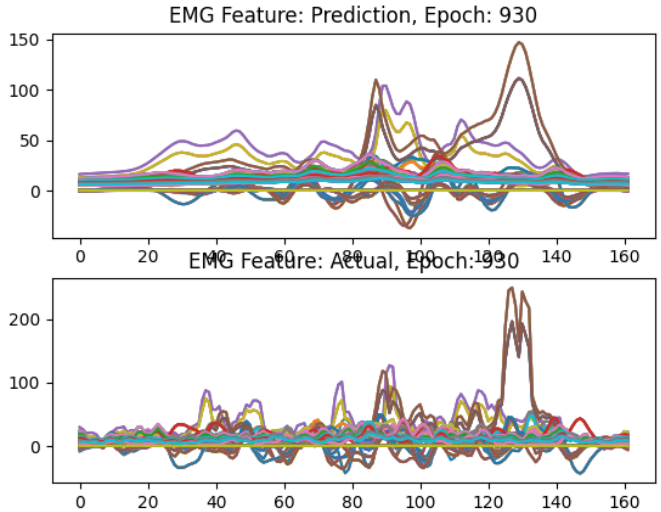
This method is commonly applied in the computer vision domain and leads to substantial improvements in performance. For instance, when cropping was applied to an image recognition dataset for the Caltech 101 dataset, it lead to a Top-1 score increase of 13.82% (Taylor and Nitschke, 2017).

2.3.2 Applicability to EMG Domain

However unlike images, EMG data is a collection of non-stationary time-series data which has been collected from different electrodes and also contains many other artefacts such as baseline noise, powerline interference and other contamination (De Luca et al., 2010). Moreover, simple geometric transformations are not necessarily appropriate for EMG as they may impact time domain features as EMG signals are 1-dimensional.

Further to this, in computer vision projects, it is easy to determine whether an augmented dataset is similar to the original signal. However for EMG signals, it is difficult to determine if an augmented dataset has still retained key properties of the original dataset. Figure 2.1 provides a sample of an original EMG data recording during vocalised speech and the synthesized EMG data for an utterance from this project, to demonstrate this difficulty. Although both of the data samples appear similar, when they are used in a transduction model to produce speech, they sound noticably different from each other because of the difference within the smaller time windows.

Figure 2.1: Original vs Synthesized Vocal EMG Data from EMG Synthesis Approach



Despite these issues with EMG synthesis, this work attempts different methods to address these issues.

2.4 Silent Speech Classification

In recent years, there has been increasing research into systems which can classify silent speech signals into text. One major approach is the MIT AlterEgo (Kapur et al., 2018) which details a system which records biosignals recorded by electrodes placed on a person's face and throat while a user is silently articulating speech (EMG).

Then these EMG signals are preprocessed, filtered and then processed using a convolutional neural network (CNN) model (Albawi et al., 2017).

Although CNN models are typically used to process 2-dimensional data, they are also very well suited to processing 1-dimensional data and can be very efficient. One paper found that using CNNs instead of deconvolution algorithms to process digital signals results in a 65x speedup (Fortino et al., 2022).

The final accuracy of the MIT AlterEgo method was 92% mean word accuracy across multiple users. However, the vocabulary set represented used in this dataset was very limited. This motivates the need to explore EMG silent speech recognition for larger datasets, especially datasets with larger vocabulary sizes.

2.5 Summary

From this literature review we can see that there are many promising approaches for improving silent speech systems. This literature review only broadly covers the silent speech literature and related topics. The approaches section contain more detailed reviews which specifically deal with the chosen approach and it's related literature.

Unless specified otherwise, silent speech dataset in this project refers to the open-source Digital Voicing (Gaddy and Klein, 2020) dataset with the phoneme information released with the second Digital Voincing paper (Gaddy and Klein, 2021).

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 Informal Research Goals

One central idea to the research experiments conducted in this project is reproducibilty. Every single experiment for the silent speech recognition fine tuning was run with the same random seed as was the experiments for the EMG synthesize.

Random seeds are important for machine learning projects because they are used to initialise the weights of neural networks. Different random seeds can drastically change the performance of a neural network, with some authors finding that the standard deviation of a model's performance can vary by as much as 72% (Madhyastha and Jain, 2019). They are also used to determine the order in which training datasets are shuffled and arranged into mini-batches for training and many other countless implementation details.

The following random seeds were used during research and were used to intialise the random seed values of the NumPy, PyTorch (Paszke et al., 2019) and default Python random number generator.

Table 3.1: Random Seed Values per Approach

Approach	Random Seed Value
EMG Synthesis	1
ASR Fine Tuning	7

3.2 Full List of Research Questions with Hypotheses

3.2.1 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 trained with augmented EMG data versus without

Following this hypothesis having been disproved, my general research into the low level features which silent speech systems used along with my side research into ASR systems lead me to make a hypothesis about training silent speech recognition systems.

3.2.2 Fine Tuning Speech Recognition on Transducer Predictions

The basis for this hypothesis comes from my literature review and experiments conducted during the EMG synthesis experiments. I noticed how the mel spectrograms from the regular transduction model looked blurrier, less precise and out of sync with the ground truth models.

From this observation I looked into speech recognition models and tried to determine if speech recognition models which were trained on clean datasets suffered drastic accuracy loses when evaluated on noisy datasets. As it turns out this is true (Amodei et al., 2015). This line of reasoning is explored further in the ASR fine tuning approach.

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

The end-to-end inference time of the system is not measured in this approach. The reason for this is because, as the approach removes one entire component compared to previous approaches, it was determined that measuring this difference is redundant as the system is guaranteed to be faster. Measuring this difference will be conducted in future research.

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.

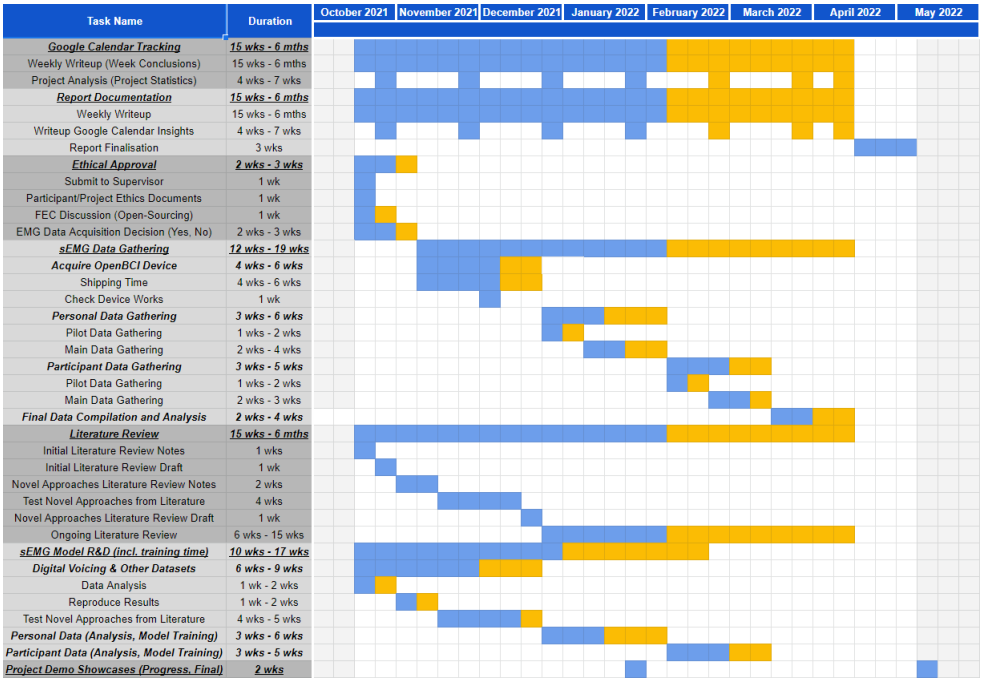
Then when I found an approach which I believed was promising, I conducted initial small-scale experiments to either validate or disprove my initial assumption, and then scaled up my experiments iteratively to verify whether or not my hypothesis was valid.

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

4.2 Research Plan

The research plan for my project is orignally based on my Project Intiation Document (PID), but was modified as the data acquisition task for this project was cancelled.

The original research plan was based around finding improvements to silent speech methods by using the Digital Voicing dataset (Gaddy and Klein, 2020) open-source dataset and using the data acquired from myself and participants. Unfortunately due to financial constraints on the project and the difficulty in acquiring the EMG data acquisition device which was originally proposed in the PID, this objective of the project was no longer achievable. The original project plan was segmented based along two of the stated objectives of this project, create an improved model / method for silent speech and data acquisition.



As I only decided to not pursue one of these objectives and the main objective of my project was still attainable, I simply stuck to the original project plan and allocated more time to the research and development for silent speech

methods. However, in-between the first approach and the second approach of my research, I conducted smaller scale experiments which have not been mentioned in this report for brevity.

They are included in the full time line in the appendix but they are more sporadic and less structured than the two main approaches explored.

4.3 Project Milestones

This section contains a table which lists the major milestones which were achieved during this project to give an overview of what was achieved, when it was achieved and how well the project was managed.

For context, OTR Approach in the table below refers to off the record experiments. Not every experiment or approach attempt was recorded in this report due to the number of approaches which did not work, but they are included in the project milestone table to provide context for how the final documented approaches were arrived at.

Table 4.1: List of Project Milestones with Dates

Date	Type	Milestone
19/10/2021	Admin	PID Submission
October 2021	Lit Review	General silent speech background
November 2021	Exploration	Explored EMG Dataset for Interesting Patterns
November 2021	Approach	Discovered EMG Augmentation
November 2021	Synth Experiment	Initial Bi-LSTM Model Experiment
December 2021	Create	Created the EMG synthesis private repo on GitHub
December 2021	Synth Experiment	Experimented with PyTorch AMP training
December 2021	Synth Experiment	Experimented with learning rate decay
December 2021	Synth Experiment	Speech generation
December 2021	Admin	Email supervisor with status update
December 2021	Synth Experiment	Switched vocoder to WaveGlow
January 2022	Admin	Email supervisor with status update
February 2022	Break	No significant milestones in February
March 2022	Approach	Considered silent speech recognition
March 2022	Lit Review	Researched typical speech recognition models
March 2022	Lit Review	Researched deepspeech2 and conformer models
March 2022	Lit Review	Researched CTC loss function and decoder
March 2022	OTR Approach	Researched ASR on direct EMG signals
05/05/2022	Publish	Switch EMG synthesis repo to GitHub publicly
05/05/2022	Publish	Released silent speech ASR on GitHub publicly
05/05/2022	Admin	Switched project to research on SUMS and new ethics

4.3.1 Off-the-record (OTR) approaches

Speech Recognition on Raw EMG signals

This approach proved to be unsuccessful because of the complexity of converting the raw EMG signals into a suitable latent representation which could then be decoded into text. I tried using the hand crafted EMG features from (Gaddy and Klein, 2020) and the 5-layer RNN layer from the DeepSpeech2 model from this project but to no avail. However, this project did make me realise that perfoming ASR directly on the predicted mel spectrograms from the transducer model would be a worthy direction for research.

Reconstruction Loss for Transducer Model

Research was conducted into adding an auxilliary loss signal to the state-of-the-art Digital Voicing transducer model (Gaddy and Klein, 2021) based on an approach used in a sequence-to-sequence voice reconstruction for silent speech paper (Li et al., 2021a). I tried the mean absolute error used in the paper with similar weighting and mean loss error but unfortuately rather than regularise training, it destablised training.

Chapter 5

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.

For both approaches, 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 primary reason for this was because the pre-trained vocoder (speech features to audio waveform) model I was using to listen to predicted audio used mel spectrograms.

5.1 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.

5.1.1 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).

5.1.2 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.

My research design involved experimenting with different hyperparameters for my proposed network to see which values produced the lowest loss value on the evaluation dataset. This metric was chosen as it was the most objective measure of how the model was able to generalise it's ability to produce novel EMG signals given a mel spectrogram input.

The loss function chosen for this task is mean squared error. This was chosen for two reasons: firstly the original transduction model from the Digital Voicing (Gaddy and Klein, 2020) paper uses this loss function to transduce mel-frequency cepstrum co-efficients (MFCCs). MFCCs are another common method of representing speech features.

Another reason mean squared error was used was that recent research into

silent speech transduction uses an auxilliary loss function where the transducer is given an additional task to predict the features of the vocalised EMG input signal representation while it is presented with a silent EMG input signal. The loss functions which the authors chose was mean absolute error, however, they likely chose absolute error instead of squared error because it has less of a drastic effect on their overall loss calculation. For this task, we only care about predicting the vocalised EMG signals given speech features so mean squared error should provide a stronger signal to the overall network (Li et al., 2021b).

5.1.3 Proposed Method

My proposed method involves taking the transduction model from the original Digital Voicing (Gaddy and Klein, 2020) paper, and reversing the input and the output of the model. The model from the paper is simply a 3-layer, 1024 hidden dimension, bi-directional LSTM.

The EMG inputs in this approach are preprocessed using the same technique described in the paper. One key detail about the synthesized EMG signals in this approach is that it is possible to use the raw EMG signals as the output for this model. This reason this was not done was because it would have added more complexity to the decoding of the EMG signals and the purpose of these initial experiments was to determine who feasible EMG synthesis for during vocalised speech.

In other words feeding in a ground truth mel spectrogram into the model and having it predict the EMG signals, in a sequence to sequence manner.

The intuition behind this is that, if the model is able to take the featurised EMG inputs and predict either MFCCs or mel spectrograms reliably, it may be possible to do the reverse.

5.1.4 Hyperparameter Tuning

Different hyperparameter values were experimented with to determine what produced the lowest loss value on a validation set. Initial experiments showed that using an LSTM model with less than 3-layers performed far worse than the 3-layers so I settled on 3-layers. Therefore, I only experimented with the hidden dimension size of the LSTM.

The results of those hyperparameter experiments for the best performing experiments are recorded:

Table 5.1: EMG Synthesis Model Hyperparameter Results

LSTM Hyperparameters	No. of Epochs	Test Loss
3-layer, 128	100	191.56
3-layer, 128	1000	104.30
3-layer, 256	1000	102.266
3-layer, 256	1000	98.64

Interestingly enough, even though this approach was unsuccessful, the hyperparamters which performed the best were very similar to a related task from another paper, which was only discovered after the fact which made it seemed to add credibility to this task.

5.1.5 Results and Visualisations

For the results, GT refers to the ground truth dataset and SYNTH refers to a synthesized version of the groud truth datset.

Table 5.2: Test Loss for Different Datasets

Dataset	LSTM Hyperparameters	Test Loss
GT	3-layer, 1024 dim	2.513
GT + SYNTH	3-layer, 1024 dim	4.829

As can be seen from the Figure **??**, the EMG synthesis model is able to learn the temporal relationship between the input mel spectrograms and the target EMG signal. The synthesis model is also able to learn the peaks of the EMG

Figure 5.1: EMG Synthesis Sample

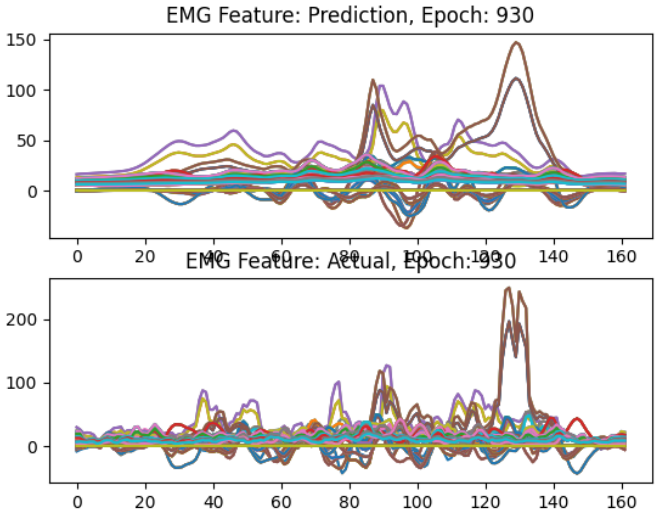
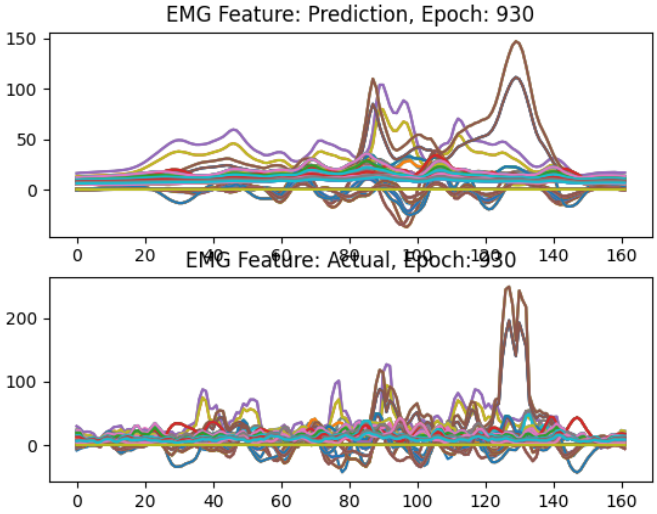


Figure 5.2: Mel Spectrogram from Transducer Trained with Real and Synthesized EMG



signals across time. However, because the model is only learning a smoothed representation of the underlying signal, it is not able to accuracy reproduce the EMG signals which distinguish between different pitches. This can be seen in Figure **??**.

5.1.6 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 which hurts the models ability to accurately reproduce the fine tuned parts of the EMG signal which correspond to the pitch of the input signal.

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. I also could have used a convolutional layer to generate the EMG signal.

5.2 Fine Tuning Speech Recognition on Transducer 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 different proportions 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.

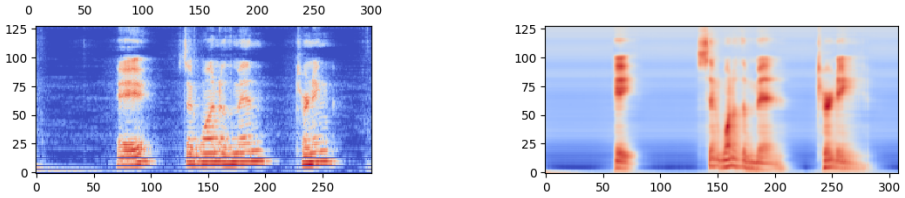
This proposed method skips over the vocoder entirely which improves the end-to-end performance as there is now one less step involved whilst also requiring far less training data, and therefore training time, for the entire silent speech recognition pipeline.

5.2.1 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 effect on evaluation performance on a noisy dataset. For this reason, this paper researches the effect of providing predicted speech from the SOTA silent speech transduction model to a speech recognition model to close the gap between the WER when evaluted on the ground truth audio features



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

Figure 5.3: Comparison of Normalised Ground Truth and Predicted Mel Spectrograms

From both of these mel spectrograms, we can see the low level features of the predicted mel spectrogram appear blurrier (the horizontal red lines with blurry edges). Not pictured in the figures, is the text transcription for the utterances which is: '"Eh?" said one of the men, turning.'.

and the predicted audio features (Amodei et al., 2015).

5.2.2 Relation to EMG Silent Speech Classification

5.2.3 Proposed Method

My proposed method involves pre-training the state-of-the-art EMG silent speech transduction model from the second Digital Voicing paper (Gaddy and Klein, 2021), and then saving the predicted mel spectrogram speech features from the model as a custom dataset (refer to Figure 5.3 for an example).

Then this dataset will be used along with the ground truth audio files to train a regular speech recognition model. Based on my literature review and intuition from exploring the audio outputs viewing the mel spectrogram visulisations from the model, I was confident this approach would be successful.

The speech recognition model used for this approach is the DeepSpeech2 model. I chose this model for two reasons: firstly it was easy to find a well-optimised implementation as the model has been around for years and the model is known to have good performance and secondly, the Digital Voicing

paper (Gaddy and Klein, 2020) uses, although an albeit older, version of the DeepSpeech architecture for evaluation. This means that the improved performance of my approach was less likely the speech recognition model architecture, and more because of the proposed training regime.

The description of the exact model used can be found in the appendix along with the hyperparameters and audio preprocessing techniques used. The source code for the model, hyperparameters, pre-trained weights and audio preprocessing techniques used can all be found at the open-source release for this project [1].

5.2.4 Evaluation Criteria

As mentioned in the Research chapter, word-error rate (WER) is the main criteria used to evaluate the performance of the EMG silent speech recognition system. However, character-error rate (CER) is used to provide additional context for each systems predictions.

A good example of this is the 70% WER for the model trained and evaluated on the closed vocabulary portion of the silent speech dataset. Although it's WER is high, it's CER is only 40%. This indicates that the model is moreso struggling to spell out the words in the dataset rather than recognising individual parts of words. This would indicate for future research that language model would be particularly effective at reducing the WER.

5.2.5 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

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

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

Table 5.3: 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.

5.2.6 Open Vocabulary (Parallel-Only) Dataset

The next dataset used is the parallel voiced and silent mel spectrogram predictions. This is used because it's smaller than using the entire open vocab-

ulary 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%. Once again, we would not expect our speech recogntion model to a lower WER rate than this.

Table 5.4: 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 parallel 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.

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

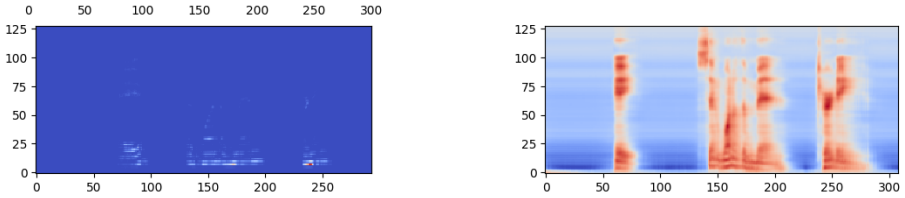
This dataset is the full entire dataset, not including the closed vocabulary dataset. The experiments for this dataset include the same experiments for the previous two datasets. However, extra experiments are added for the additional non-parallel vocal EMG mel spectrogram predictions. The paralell silent 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 5.5: DeepSpeech2 Open Vocabulary Full Dataset Finetuned Results

Dataset	**CER**	**WER**
Ground Truth (GT)	61.99	100.74
Parallel Voiced	176.37	104.37
Silent	35.89	74.15
Silent, Parallel Voiced	35.20	72.31
Parallel Voiced, GT	34.37	71.36
Silent, GT	32.48	70.84
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



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

Figure 5.4: Comparison of Real Ground Truth and Predicted Mel Spectrograms The raw mel spectrogram of the raw ground truth audio shows less distinct features as the range of values is exponentially higher than the log-melspectrogram of the transducer models predictions. Not pictured in the figures, is the text transcription for utterances which is: '"Eh?" said one of the men, turning.'.

characters but it's ability to predict words has been slightly reduced.

5.2.8 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 5.4. Comparing both of the mel spectrograms side by side, we can see that for the raw-mel spectrogram, the loudness (represented as brightness) of the lower pitched values is much higher than the other ones. Whereas the log-mel spectrogram treats the loudness of values at every pitch the same. It is difficult to determine what effect this would have had on the fully trained silent speech recognition system proposed in this paper.

The main difference is that the mel spectrogram values from the transduction model are log-mel spectrogram values. 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 represented as raw mel spectrograms mixed with the log-mel spectrograms of the transducer predictions together still results in better performance.

This means that performance of this model may be greatly increased in the future by fixing this bug. Unfortunately this bug was discovered towards the end of the project where it would be too time consuming to explore the effect of fixing this bug.

5.2.9 Limitations

Their are two primary limitations of this approach which relate to the design of the speech recognition model, and the implementation of the design of the plan. The first limitation is the use of greedy decoding and lack of a language model when the speech recognition model produces it's final text transcript. Speech recognition models typically use a technique called Beam Search (Tillmann and Ney, 2003) which uses a breadth-first search over the predicted tokens for a transcription to find the most likely transcription.

Instead the greedy decoder used in this approach simply returns the most character which the speech recognition model has determined has the highest probability. Not only that but recent speech recogntion models use language models to improve the intelligibility of the final transcript. This is typically done by training a recurrent neural network (RNN) or other model to learn which mistakes the initial speech recognition model makes and correct them (Tam et al., 2014).

However rather than being a negative sign for this approach, this has a positive finding. When this approach uses a completely unoptimised method for transcribing text from silent speech signals, it can still perform as well an older state-of-the-art method (Gaddy and Klein, 2020). This means that if these simple techniques were applied to this approach, the WER would almost certainly be dramtically reduced even further.

5.2.10 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 6

Conclusion

Overall this project was highly successful in achieving it's primary objective, which was to find a method of improving either silent speech classification or transduction based on at least one evaluation metric. A novel training method is proposed which achieves the same WER as a prior state-of-the-art (SOTA) system, 68% for the Digital Voicing approach (Gaddy and Klein, 2020) and 68.26% for the proposed approach from this project. However the proposed method requires 115 times less training data to train the ASR component of the system making the training data requirements and training time both drastically lower. Also the the vocoder component required by the old approach is also removed which leads to a faster end-to-end system.

Secondly, the objective to open-source the method and dataset used for this approach was also successful with the successful speech recognition approach [1] and less successful EMG synthesis approach [2] both being open-sourced before the submission of this project.

However the additional objective to create an open-source sEMG silent speech dataset and release it to the public was unfortunately not achieved. This is unfortunate as there are very few available silent speech datasets available

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

[2] https://github.com/MiscellaneousStuff/semg_silent_speech_py

and the creation and release of such a dataset would of been highly beneficial for researchers.

In conclusion, this project ultimately provides a method which reduces the required data and machine learning system components for silent speech recognition whilst still achieving competitive performance.

Chapter 7

Future Work

7.1 Implementation Errors

7.1.1 Data Augmentation via EMG Synthesis

One major flaw with the model used for synthesizing EMG signals was that the input mel spectrograms for the synthesis model were then directly fed into the LSTM network. However all modern machine learning approaches which process speech features such as mel spectrograms either process the mel spectrograms into hand-crafted features or use a convolutional neural network to learn features in an end-to-end model. Although it's unclear how this may have affected the proposed approach in this project, this is a fundamental flaw with the approach that would need to corrected in any future approaches.

7.1.2 Fine Tuning Speech Recognition on Transducer Predictions

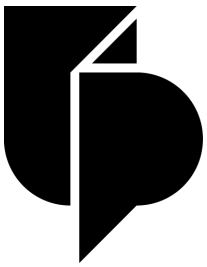
The experiments conducted for the fine tuning ASR approach contained a major flaw where the ground truth ASR model was trained using the ground truth audio data, which was transformed into raw mel spectrograms. How-

ever, the ASR models in the experiment were then refined using predicted log-mel spectrograms from the transduction model.

A visual demonstration of the difference between these representations was presented in Figure 5.4. Future work on this method should explore whether using log-mel spectrograms for both, the ground truth and predicted mel spectrograms improves performance.

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

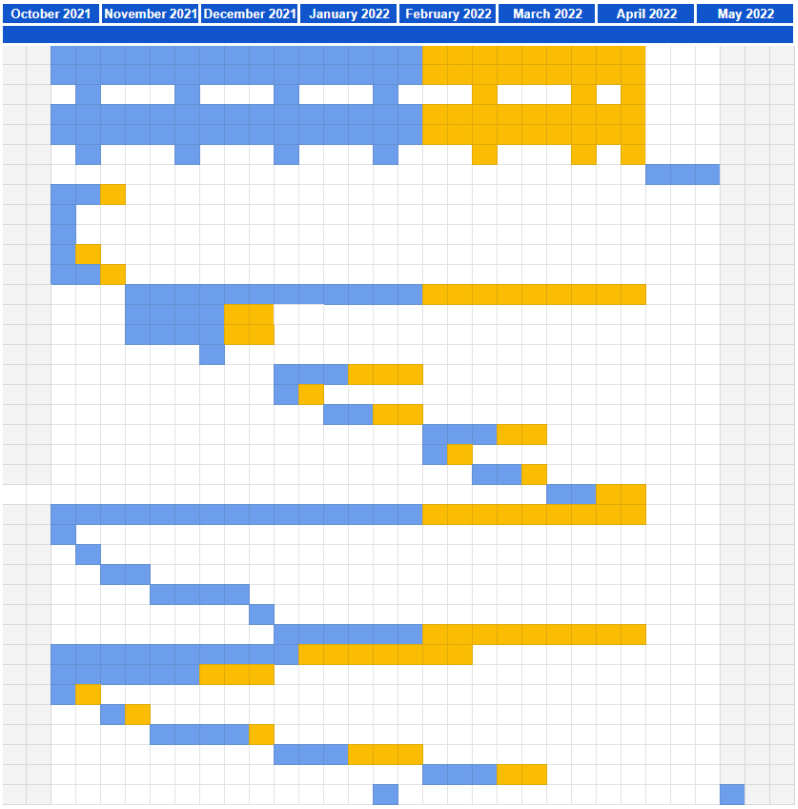
Description	Impact	Likelihood	Mitigation	First indicator
COVID-19 infection risk between research team and participants	*Severe*	*Likely*	*Ask participants if they have any symptoms of COVID-19: high temperature, cough, loss of taste*	*Either the research team starts displaying COVID-19 systems or a 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

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network, In *2017 international conference on engineering and technology (icet)*. https://doi.org/10.1109/ICEngTechnol.2017.8308186

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

De Luca, C. J., Donald Gilmore, L., Kuznetsov, M., & Roy, S. H. (2010). Filtering the surface emg signal: Movement artifact and baseline noise contamination. *Journal of Biomechanics*, *43*(8), 1573–1579. https://doi.org/https://doi.org/10.1016/j.jbiomech.2010.01.027

Fortino, G., Zamora, J., Tamayose, L., Hirata, N., & Guimarães, V. (2022). Digital signal analysis based on convolutional neural networks for active target time projection chambers. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *1031*, 166497. https://doi.org/10.1016/j.nima.2022.166497

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

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

Kapur, A., Kapur, S., & Maes, P. (2018). Alterego: A personalized wearable silent speech interface, In *23rd international conference on intelligent user interfaces*, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3172944.3172977

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.

Li, H., Lin, H., Wang, Y., Wang, H., Zhang, M., Gao, H., Luo, Q. A. Z., & Li, G. (2021a). Sequence-to-sequence voice reconstruction for silent speech in a tonal language. arXiv. https://doi.org/10.48550/ARXIV.2108.00190

Li, H., Lin, H., Wang, Y., Wang, H., Zhang, M., Gao, H., Luo, Q. A. Z., & Li, G. (2021b). Sequence-to-sequence voice reconstruction for silent speech in a tonal language. arXiv. https://doi.org/10.48550/ARXIV.2108.00190

Madhyastha, P., & Jain, R. (2019). On model stability as a function of random seed. arXiv. https://doi.org/10.48550/ARXIV.1909.10447

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. arXiv. https://doi.org/10.48550/ARXIV.1912.01703

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

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data*, *6*, 60. https://doi.org/10.1186/s40537-019-0197-0

Tam, Y.-C., Lei, Y., Zheng, J., & Wang, W. (2014). Asr error detection using recurrent neural network language model and complementary asr, In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. https://doi.org/10.1109/ICASSP.2014.6854012

Taylor, L., & Nitschke, G. (2017). Improving deep learning using generic data augmentation. arXiv. https://doi.org/10.48550/ARXIV.1708.06020

Tillmann, C., & Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational linguistics*, *29*(1), 97–133.