A multiplexed star electrode array for session-independent silent speech recognition

Sudharshan Sundaramahalingam

3rd Year Project Final Report

Department of Electronic &
Electrical Engineering

UCL

Supervisor: Dr Arsam Shiraz

5 April 2024

I have read and understood UCL's and the Department's statements and guidelines concerning plagiarism.

I declare that all material described in this report is my own work except where explicitly and individually indicated in the text. This includes ideas described in the text, figures and computer programs.

I acknowledge the use of Chat-GPT-3.5 of OpenAI, https://chat.openai.com to summarise my initial notes and proofread my final draft.

This report contains 28 pages (excluding this page and the appendices) and 10773 words.

Signed: DocuSigned by: Sudharshan Sundaramahalingam FA85EB729F8A477... Date: 4/4/2024

(Student)

A multiplexed star electrode array for session-independent silent speech recognition

Sudharshan Sundaramahalingam

Voice-based human-machine interfaces are experiencing a resurgence, shown by an annual growth rate of 20% in the global voice interface market. However, adoption has been slowed by issues of privacy, ambient noise interference, and accessibility. This study addresses these concerns by developing a surface electromyography (sEMG) based silent speech interface with a novel multiplexed 8-channel star electrode array. The study investigates the use of configurable electrode orientations as a simpler and more practical alternative to high-density electrode arrays (HD-sEMG) for silent speech recognition systems. The goal is to create a system that shows robust performance without requiring large quantities of user-dependent training data, especially in the presence of cross-session variability. A state-of-the-art silent speech recognition system was first replicated and assessed, revealing a substantial 34% drop in accuracy due to cross-session variability. Subsequently, the novel star electrode array was designed, capable of adjusting the orientation of its electrodes for optimal muscle-electrode alignment. Tests confirmed the array's efficacy, significantly enhancing the signal-to-noise ratio (SNR) of the sEMG signal by an average of 6.7 dB across channels. Most notably, the array displayed a remarkable 41% relative improvement in cross-session classification accuracy over the control. The study concludes that the multiplexed star electrode array offers a tangible advancement over current HD-sEMG systems. With the convenience and portability of cheek placement and improved cross-session accuracy, the star electrode array demonstrates significant potential for practical applications in silent speech interfaces. Future work would involve larger participant groups across multiple sessions to solidify the findings and potentially integrate adaptive electrode orientation adjustments for real-time SNR optimization.

Contents

1 Introduction

Voice-based human-machine interfaces are experiencing a resurgence shown by an annual growth rate of 20% in the global voice interface market [1], due to the rising capabilities of natural language processing systems. But while voice interfaces have the lowest cognitive load compared to other human-machine interfaces [2], concerns of user experience such as a lack of privacy and susceptibility to ambient noise interference have kept its adoption rates low [3].

1.1 Silent speech interfaces

A silent speech interface measures non-audible signals correlated with the quiet mouthing or imagining of speech to achieve speech recognition. Therefore, silent speech is private, unaffected by ambient sound and accessible to people who are unable to vocalize, thereby improving the user experience of voice interfaces. These features have led to compelling applications of silent speech in defense [4], accessibility [5][6] and human-computer interfacing [7][8].

1.2 Different modalities for silent speech recognition

There are several detection modalities for silent speech [9], of which the most common are outlined in Table 1. The ultrasound modality utilizes echos caused by unvoiced lip and tongue movement for speech recognition, and has been demonstrated on a small vocabulary of 30 words at a 96% accuracy [10]. Similarly, the use of inertial measurement units to classify jaw movement patterns during mouthing has been shown to achieve a 91% word classification accuracy over a 100 word vocabulary [11]. Electroencephalography (EEG) based techniques have also been successfully deployed to decode imagined speech using invasive cortical implants, achieving a 70% word classification accuracy with a 1000 word vocabulary [12][13] and allowing patients with locked-in syndrome to communicate. Finally, surface electromyography (sEMG) is a silent speech modality which measures electrical activity in the face and neck muscles during mouthed speech. Of the aforementioned modalities, sEMG offers a balance of wearability, accuracy, and non-invasiveness which makes it a popular choice for silent speech interfaces [9].

Modality	Reference(s)
Ultrasound Mapping	[10][14]
Motion and Inertial Measurement	[11]
Electroencephalography (EEG)	[12][13]
Surface Electromyography (sEMG)	[8][15][4]

Table 1: Summary of silent speech modalities.

1.3 sEMG-based silent speech recognition

While sEMG is a compelling silent speech modality, sEMG measurements are dependent on environmental conditions like skin impedance and muscle-electrode alignment [16][17]. This results in a large cross-session and cross-user variability in sEMG-based speech recognition accuracy [18]. In particular, cross-session variability implies that the system must be trained prior to every use, which severely affects its practical viability.

There are two approaches to reducing cross-session variability. The first approach is high density sEMG (HD-sEMG), which utilizes a dense 2D array of electrodes to minimize dependence on the muscle-electrode alignment of any individual electrode [19][20]. However, HD-sEMG increases the wiring complexity and the size of silent speech systems. As these systems must be mounted proximal to the target face and neck muscles, the size of HD-sEMG and their associated measurement systems makes them hard to deploy in practice. The alternative is to create a lower density electrode array which is capable of actively modifying the orientation of each electrode.

1.4 Purpose of the study

This study investigates the feasibility of configurable electrode orientations to achieve muscle-electrode alignment, and its impact on cross-session accuracy. Furthermore, a novel multiplexed 8-channel electrode star array is proposed and shown to be a viable, lower-complexity and more compact alternative to HD-sEMG to minimize session dependence in silent speech recognition.

Following this, Section 2 outlines the requisite background regarding sEMG-based silent speech recognition. Then, Section 3 conducts a literature review and presents state-of-the-art work on achieving session-independent silent speech recognition. After that, Section 4 describes the methods and materials used in this study. Next, Section 5 highlights the results and discusses experimental outcomes. Finally, Section 7 concludes and summarizes the work presented in this study, and suggests potential future work.

2 Background

2.1 Speech muscle anatomy

Since sEMG-based silent speech recognition is based on muscle activity, it is important to identify the key muscle anatomy involved in mouthed speech. Speech is a highly coordinated motor action consisting of over a 100 muscles, orchestrated by the Broca's area and the ventral sensorimotor cortex [21]. The articulators involved in speech are distributed around the face, pharynx, larynx and vocal cavity, and are precisely controlled for the desired sound[8]. These regions are illustrated in Figure 1.

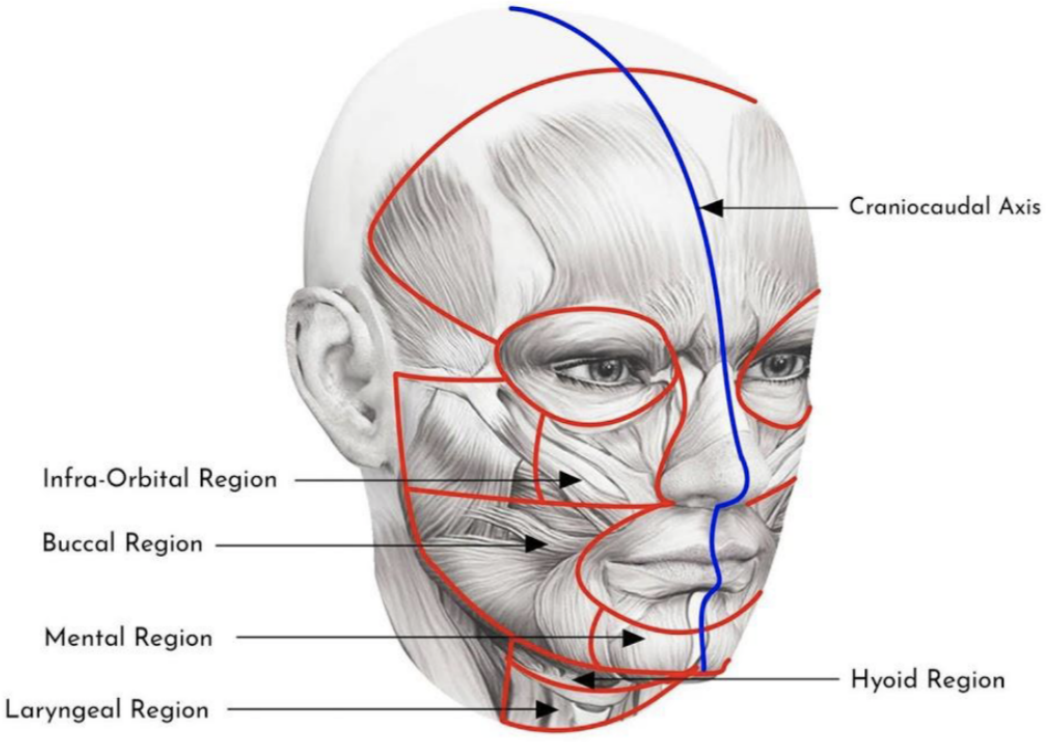


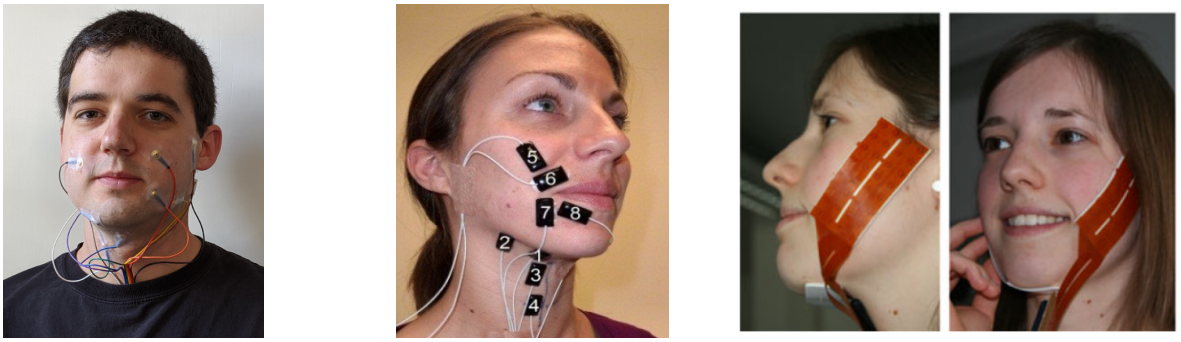
Figure 1: Musculature involved in human vocalization [8].

During innervation, action potentials move along a muscle fiber, creating potential

difference patterns that can be detected on the skin. sEMG measures these potential difference patterns using a pair of electrodes placed on an user's body [22].

2.2 Signal acquisition

An sEMG electrode setup can be broadly categorized into two configurations. In the monopolar configuration, one electrode is placed on the target muscle and a reference electrode is placed on electrically neutral tissue [23]. In the bipolar configuration, two electrodes are placed on the target muscle, orientated longitudinally along the muscle. In both cases, the sEMG signal is measured as the potential difference between the two electrodes [23]. The electrodes can be deployed either as individual channels, or arranged in a 2D array [22]. When deployed as individual channels, the electrodes are placed precisely over a set of target muscles that are monitored. In contrast,, an array of electrodes can simply be placed in the general target muscle region [22].



(a) Monopolar [24] **(b)** Bipolar [4] **(c)** HD-sEMG [25]

Figure 2: Different electrode configurations for silent speech recognition.

In addition to the electrode positioning, the electrode type also varies across different silent speech recognition setups. Electrodes are typically made of polarization-resistant metals such as platinum, silver and silver-silver-chloride (Ag-AgCl) [26]. Wet electrodes combine the metal with a conductive gel, to decrease skin-electrode impedance and allow for better signal pickup [26]. Conversely, dry electrodes utilize direct tissue contact for convenience at the cost of higher skin-electrode impedance and greater susceptibility to motion artifacts [26]. For the silent speech task, 10 mm gold-cup EEG electrodes filled with conductive gel are commonly used for their small size, which makes them easy to attach on the face [24][8]. An example of 10 mm gold-cup electrodes is shown in Figure 2a.

2.3 Signal processing

The sEMG signals acquired from the electrodes are small and require amplification and filtering. The polarization at the electrode-tissue interface also introduces an additional offset to the electrode potential [26]. This galvanic cell potential changes with local electrolyte concentration and contributes to movement artifacting, especially in dry electrodes [26]. Due to the high impedance looking into the electrode, the electrode is also susceptible to powerline (50/60 Hz) interference [27] which is superimposed on the signal. To properly extract the sEMG signal, a differential amplifier must be deployed to remove sources of common-mode noise like powerline interference. Thus, electrode type matching

across the electrodes in the system is important to minimize common-mode-to-differential conversion [26].

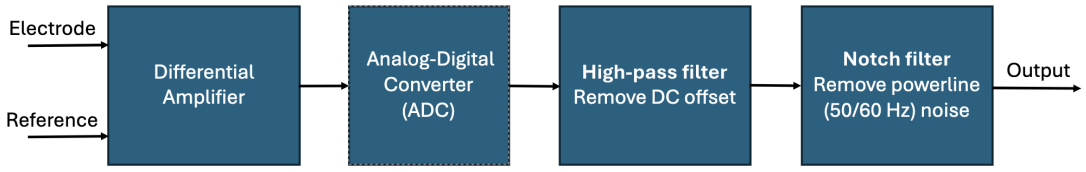


Figure 3: Digitally implemented sEMG signal processing flow.

Post-amplification, the sEMG signal must be filtered either in the analog and the digital domain [27]. Due to the availability of digital bio-signal acquisition ICs like the ADS1299[1], many silent speech recognition systems choose to use a digital implementation for its flexibility and lower component costs. A typical signal processing flow is illustrated in Figure 3. The ADS1299 provides a high-resolution 24-bit delta-sigma ADC, with a programmable gain differential amplifier and configurable sample rate. The high-resolution ADC works with a low-gain differential amplifier to avoid amplifier saturation due to DC offsets and thereby removes the need for AC-coupling circuitry prior to the amplification stage. Instead, the DC removal can be implemented digitally through an Infinite-Impulse Response (IIR) high-pass filter with a cutoff in the range of 0.5-4 Hz [27]. Similarly, powerline interference can also be removed digitally through the use of a IIR notch filter centered at 50/60 Hz [27]. Finally, the spectral range of sEMG signals ranges from approximately 5-550 Hz as shown in Figure 4. Thus, respecting Nyquist Sampling Theorem, the sampling frequency for the ADC must be greater than 1 kHz [28]. However, since much of the spectral power is below 125 Hz, silent speech recognition systems use sampling frequencies as low as 250 Hz [8] due to the lower data transmission bandwidth requirements.

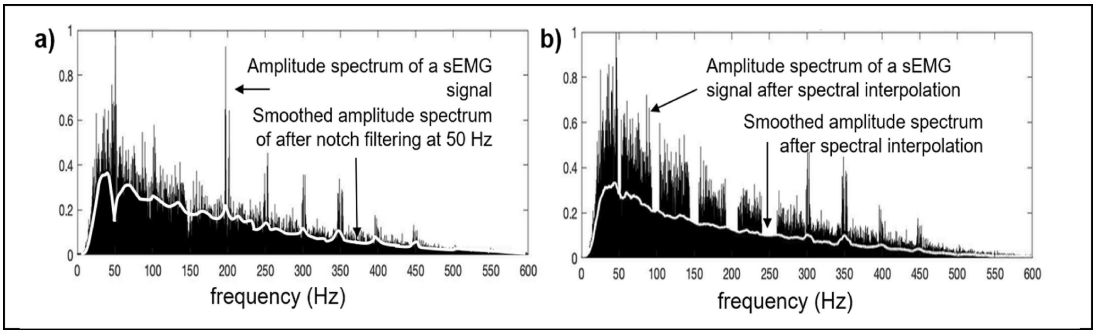


Figure 4: Amplitude spectrum of sEMG signal [28].

2.4 Feature extraction

For many silent speech recognition systems, the signal is further processed through feature extraction. Some commonly used time series features are Mean Absolute Value, Waveform Length, Slope Sign Change and Zero Crossing Rate [29][27] as shown in Table 2.

[1]https://ti.com/lit/gpn/ads1299

Feature Extraction	Mathematical Equation
Mean Absolute Value (MAV)	$MAV = \frac{1}{N}\sum_{n=1}^{N}\|x_n\|$
Waveform Length (WL)	$WL = \sum_{n=1}^{N-1}\|x_{n+1} - x_n\|$
Slope Sign Change (SSC)	$SSC = \sum_{n=2}^{N} f(\|(x_n - x_{n-1}) \cdot (x_n - x_{n+1})\|)$ where $f(x) = \begin{cases} 1, & \text{if } x \geq \text{threshold} \\ 0, & \text{otherwise} \end{cases}$
Zero Crossing (ZC)	$ZC = \sum_{n=1}^{N-1}[\text{sgn}(x_n \cdot x_{n+1}) \wedge \|x_n - x_{n+1}\| \geq \text{threshold}]$ where $\text{sgn}(x) = \begin{cases} 1, & \text{if } x \geq \text{threshold} \\ 0, & \text{otherwise} \end{cases}$

Table 2: Mathematical expressions for common time-series features.

Similarly, many frequency-based features such as the Short-time Fourier Transform, Wavelet Transform and Mel-frequency cepstral coefficients (MFCCs) have been tested for silent speech recognition [15]. Of these, MFCCs been empirically shown to have the most robust classification performance [15]. MFCCs are widely used for speech and audio processing, and are designed to simulate human ear's frequency response to audio stimuli [30]. This process involves taking the Fourier transform of a signal and mapping the powers of the spectrum obtained onto the mel scale, followed by taking the logs, the discrete cosine transform, and then the amplitudes of the resulting spectrum [30].

In recent literature, classification without feature extraction has also become feasible due to neural network approaches [8][31][19]. However such systems require larger datasets or dataset augmentation for the networks to identify the proper representations for the sEMG data.

2.5 Machine learning

There are three main machine learning approaches to silent speech recognition. The first is word classification, where models are trained to predict discrete words given the sEMG signal [8][7][29]. The second is phoneme classification, where models can have unlimited vocabulary size once all the requisite phonemes are learnt. The word classification task is easier, but is restricted to a smaller vocabulary as the number of classes grows linearly with the vocabulary size. Conversely, phoneme classification is more challenging due to the shorter duration of each class ($\leq$ 1 s), which means that each class is harder to discriminate. The third approach is EMG-to-speech, where a sequence-to-sequence approach is taken to directly convert sEMG signals into audio signals [25][19]. Then, conventional speech recognition can be used to convert the speech signal into a transcription. This approach is used with neural network architectures and larger datasets.

A wide variety of models are deployed for silent speech recognition. Decision Trees [29] and Linear Discriminant Classifiers (LDAs) are popular for small vocabulary and small dataset word classification tasks. Hidden Markov Models are popular and show good performance [32][15] for medium to large vocabulary word/phoneme classification tasks. Lastly, Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM) models are popular for large-vocabulary decoding [24] and EMG-to-speech tasks [19].

Metric	**Mathematical Equation**
Word Classification Accuracy	$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100$
Word Error Rate (WER)	$\text{WER} = \frac{\text{S} + \text{D} + \text{I}}{\text{N}} \times 100$
Mel Cepstral Distortion (MCD) [33]	$\text{MCD} = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\sum_{n=1}^{N} \left(c_n^{(t)} - \hat{c}_n^{(t)} \right)^2}$

Table 3: Common metrics for evaluating model performance.

The metrics to evaluate the performance of a model vary based on the desired task. The common metrics are shown in Table 3. Word classification accuracy is used in this study and in many silent speech recognition systems which perform small vocabulary word classification. The WER metric is calculated as the sum of substitutions (S), deletions (D), and insertions (I) divided by the number of words in the reference (N), multiplied by 100 to get a percentage. It is typically used for larger vocabulary or phoneme based classification tasks. Lastly, the MCD metric measures the distortion between two sequences of MFCC values. MCD is typically used in sequence-to-sequence tasks like EMG-to-speech.

3 Literature Review

In the mid-1980s, the concept of silent speech interfaces through sEMG emerged simultaneously in works by Morse *et al.* [34] and Sugie *et al.* [35]. Since then, the viability of silent speech recognition has been demonstrated over a range of vocabularies [8][24], most notably by Meltzer *et al.* who achieved a 91.10% accuracy over a large vocabulary recognition task in 2018 [15]. However, as displayed in Table 4, many of the studies reporting impressive results are evaluated over just a single session. This means that the classification models are trained and tested on one round of data collection per participant, ignoring the variation in skin-electrode impedance and muscle-electrode alignment across sessions. Therefore, single-session results does not account for significant cross-session variability [18] and can be misleading regarding the viability of silent speech interfaces in practice.

Author	**Electrode Configuration**	**Vocabulary**	**Model**	**Metrics**
Lai *et al.* [36] (2023)	3-ch Bipolar (Dry)	26 words	CNN	85.90% ACC (Single) N/A (Cross)
Gaddy *et al.* [24] (2022)	8-ch Monopolar (Wet)	67 words	CNN	96.00% ACC (Single) N/A (Cross)
Kapur *et al.* [8] (2019)	8-ch Monopolar (Wet)	30 words	CNN	91.00% ACC (Single) N/A (Cross)
Meltzner *et al.* [15] (2018)	8-ch Bipolar (Dry)	2100 words	HMM	91.10% ACC (Single) N/A (Cross)
Diener *et al.* [19] (2018)	40-ch HD-sEMG (Dry)	Unlimited	CNN	5.05 MCD (Single) 6.23 MCD (Cross)
Maier-Hein *et al.* [18] (2005)	8-ch Monopolar (Wet)	10 words	HMM	97.10% ACC (Single) 87.10% ACC (Cross)

Table 4: Review of recent silent speech interfaces and their setups. (ACC is classification accuracy, MCD is mel cepstral distortion score.)

The attempts to tackle session dependence in silent speech recognition thus far, primarily involves software approaches. Maier-Hein *et al.* [18] demonstrated that training concurrently on multiple sessions of data, and the use of feature mean variance normalization improves cross-session classification accuracy from 76.2% to 87.1% on a 10 word vocabulary. Then in 2010, Wand and Schultz [37] demonstrated the use of maximum likelihood linear regression techniques and bundled phonetic features [32] to create speaker and session adaptive models. Both studies were conducted with HMMs using feature extraction and managed to achieve modest gains in accuracy over a medium sized vocabulary [32][37].

Recently, HD-sEMG has been shown to provide rich spatial-temporal data [22], which could be exploited for session-independence. Diener *et al.* [19] combined the spatial data from HD-sEMG with CNNs, to create an EMG-to-speech model which demonstrated a 10% improvement in cross-session performance. However, compared to single-session performance, a 20% decrease in MCD for cross-session evaluation was still reported after the improvements [19]. Another approach is to use HD-sEMG with channel selection to identify electrode positions with the most information for silent speech recognition [38]. This approach can be extended to dynamically perform electrode-muscle alignment on each session to minimize cross-session variability, however that is not explored in detail

by the study [38]. Expanding on that, Deng *et al.* [20] used transfer learning with a CNN model to train on a 64 channel HD-sEMG setup and adapt it to 8 selected channels. By adding rich prior context about the spatio-temporal sEMG patterns to the model, the study showed that severe electrode shift in the 8 selected channels only contributed to a 2.09% decrease in classification accuracy [20] and a mild electrode shift contributed a small 0.96% decrease in accuracy [20]. This is promising and shows that HD-sEMG contains spatial information that can be used to counteract session dependence. Additionally, the study demonstrates that the 8 channel setup with transfer learning outperforms the setup that utilizes all 64 channels during inference [20]. This shows the importance of channel selection to avoid cross talk and inactive muscle regions which may contribute to model confusion [20]. Unfortunately, the study does not make an attempt to directly quantify the cross-session performance of the system.

Overall, as seen from Table 4, there is a lack of recent literature reporting the cross-session accuracy of silent speech recognition systems. The work by Deng *et al.* [20], demonstrates the potential of exploiting electrode selection to optimize cross-session performance. Additionally, the study indicates that a smaller set of selected channels outperforms HD-sEMG [20]. Therefore, it is plausible that an electrode array with fewer channels and the capability to change its electrode orientations could mimic the effect of channel selection to minimize cross-session variability. By reducing the channel count of the array, the system can additionally be made easier to deploy as compared to complex, and bulky HD-sEMG setups.

4 Materials and Methods

This section is split into two parts. Due to the lack of recent literature on cross-session performance, in Section 4.1, a procedure is outlined to replicate the state-of-the-art in silent speech recognition and quantify the impact of cross-session variability on classification accuracy. Then in Section 4.2, the motivation and development of the 8-channel electrode star array with configurable electrode orientations is outlined. Additionally, the experimental procedure to evaluate the performance of the star electrode array is described.

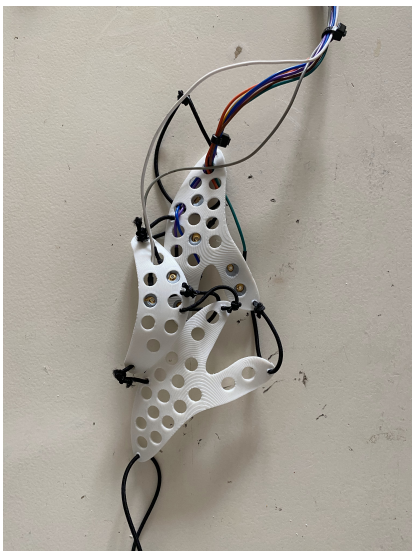
4.1 Investigating impact of cross-session variability on state-of-the-art

The basis for the state-of-the-art silent speech recognition system closely follows the studies conducted by Kapur *et al.* [8], Gaddy *et al.* [24] and Maier-Hein *et al.* [18]. The study will be conducted through testing classification accuracy over a 3-word and 30-word vocabulary across 2 sessions. This allows both single-session and cross-session accuracy to be determined over both vocabulary sizes.

4.1.1 Electrode configuration

In line with Kapur *et al.* [8] and Gaddy *et al.* [24], a 8-channel monopolar electrode configuration is adopted for this setup. The electrodes used in the setup were 10 mm diameter gold cup EEG electrodes which were filled with Ten20 conductive paste. The electrode positions selected for the study were replicated from Kapur *et al.* [8], and as illustrated in Figure 1, target the mental, submental, infraorbital, buccal, hyoid and laryngeal regions of musculature. A reference electrode is placed on the left earlobe, and a bias cancelling electrode is placed on the right earlobe. To conduct experiments in a repeatable fashion, a 3D printed brace is designed and shown in Figure 5. The brace

contains cutouts for possible electrode positions, and is attached to the user's face with elastic cord over the ears. The gold cup electrodes are adhered firmly to the 3D printed brace with bluetack, and can be repeatedly filled with the Ten20 conductive gel without being removed.



(a) Design of the electrode brace

(b) User wearing the brace using elastic cords

Figure 5: The brace was designed to hold the electrodes at a fixed location, such that studies could be conducted repeatably.

4.1.2 Signal acquisition

An eight-channel EMG acquisition system, the OpenBCI Cython[2] board, is used for signal acquisition. The data is captured simultaenously across the 8 channels at 2 MHz, digitally decimated, wirelessly streamed to a laptop and stored at a final sample frequency of 250 Hz. The Cython board is configured to the maximum possible gain configuration of 24 V/V.

4.1.3 Data collection

As mentioned in Section 4.1, two distinct rounds of data collection are required for a 3 word vocabulary (Dataset 1) and a 30 word vocabulary (Dataset 2). The summary of the datasets is shown in Table 5. The chosen vocabulary for the classification task derives from the Talon phonetic alphabet[3]. Talon is a voice computing system for people with repetitive strain injuries, and its phonetic alphabet consists of short and phonetically dissimilar words which code for each word in the alphabet. The phonetic dissimilarity makes it more differentiable within silent speech classification, and the alphabet coding allows a smaller vocabulary size to be leveraged to any communication task.

For Dataset 1, a participant was instructed to wear the brace and silently mouth the words 'air', 'bat' and 'cap' which represents letters 'a', 'b' and 'c' in the Talon phonetic alphabet. For each word, the participant is given 3 seconds to mouth the word, and is

[2]https://docs.openbci.com/Cyton/CytonLanding/
[3]https://talonvoice.com/

Dataset	Vocabulary	Total Samples	Participant Count	Sessions
Dataset 1	3 words	300	1	2
Dataset 2	30 words	600	1	2

Table 5: Summary of collected datasets.

allowed to rest between words. Each word was mouthed 50 times in a randomized order to minimize the effect of external noise on classifier accuracy. To provide visual feedback and instructions to the participant, a helper tool is written using Python and Matplotlib. The helper tool informs the participant of the word to mouth and simultaenously monitors the sEMG datastream and ambient sound level to ensure that the user is quietly mouthing the word. It also paces the participant to the 3 second data collection period and helps them check that their recording was accurately captured in the final sEMG waveform. A screenshot of the data collection tool is shown in Figure 6.

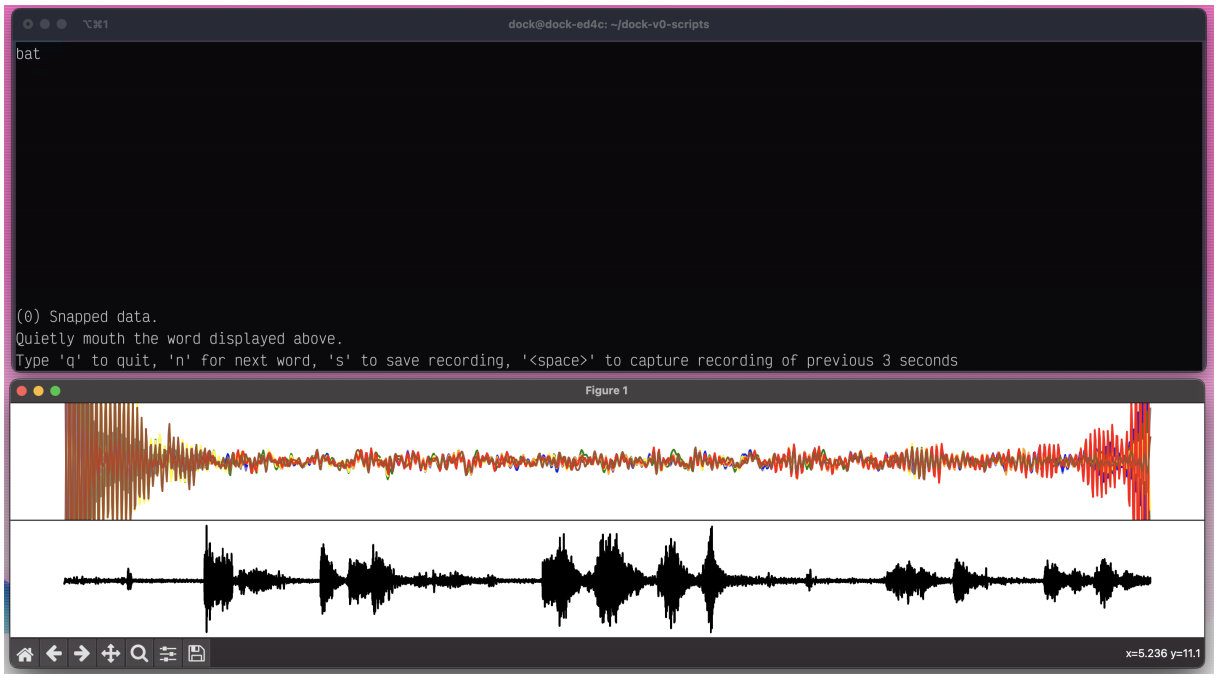


Figure 6: A screenshot of the data collection tool with a relaxed participant.

For Dataset 2, the original vocabulary is expanded by adding the remaining letters of the phonetic alphabet as shown in Appendix B. The 26 words in the phonetic alphabet are then augmented with the four control words 'enter', 'space', 'end' and 'delete' to derive the full 30-word vocabulary. Each word is also mouthed 10 times instead of 50 times. Otherwise, the data collection procedure is kept identical to Dataset 1.

4.1.4 Signal processing and feature extraction

The signal processing for the 3-word and 30-word dataset are identical. First, the mean value of the signal is removed from itself to prevent artifacts due to digital filters. Lastly, as described in Section 2.3, a high-pass filter is applied at 2 Hz to minimize DC bias and movement artifacts. A 50 Hz digital notch filter is used to remove power-line interference.

For the HMM model, the sample is now windowed with a window length of 400 ms

and a window stride of 100 ms. This allows for a sequence of features to be extracted from the sample. Otherwise, for the LDA model, the entire sample is considered as one window. Features were then extracted from the windows, namely MAV, WL, ZC, SSC, and the first 6 MFCC coefficients. This resulted in 80 features per window (10 features per channel across 8 channels). Then, 4 combinations of features from the feature set were tested on the 3-word dataset to determine the best feature set to use in the star electrode experiments outlined in Section 4.2.6. The feature sets are described below:

- **All:** This feature set includes MAV, WL, SSC, and ZC alongside the first 4 MFCC coefficients.
- **Time-Only:** This feature set only includes time series features, which are MAV, WL, SSC, and ZC.
- **Frequency-Only:** This feature set only includes frequency features, which are the 4 MFCC coefficients.
- **Time-Frequency:** This feature set includes MAV and ZC, alongside the first 2 MFCC coefficients. This retains the same number of feature dimensions as the Time-Only and Frequency-Only feature sets, but combines time and frequency information.

For the 30-word dataset, due to the high number of output classes, more features are included to increase the dimensionality of the input. Thus, MAV, WL, and 6 MFCC coefficients are included for the features in the 30-word dataset.

4.1.5 Classification

Classification is conducted using both LDAs and HMMs. The LDA classification model is used with the 3-word dataset as the dimensionality of the output classes is small. For the LDA model, an 80/20 train-test split is used to measure classifier performance. Validation is done using a 5-fold cross validation scheme.

For the 30-word classification task, a 5-state left-to-right HMM is deployed in line with the work from Maier-Hein *et al.* [18] and Meltzner *et al.* [15]. The feature set with 80 feature dimensions is reduced to 10 components using a LDA model. Finally, after dimensionality reduction, the HMM is trained for 100 iterations. Similarly to the LDA model, the HMM is tested with an 80/20 test-train split.

To evaluate single session accuracy, the model is trained and tested independently on each session within the dataset, and the mean accuracy of both sessions is computed. Cross session accuracy is obtained by training the model on one session of data, and testing it on the held-out session, and calculating the mean of both possible permutations. And finally, a combined session test accuracy is computed by pooling data from both sessions before applying the 80/20 test-train split, as inspired by Maier-Hein *et al.* [18]. Finally, in all experiments, word classification accuracy is reported by the following metric.

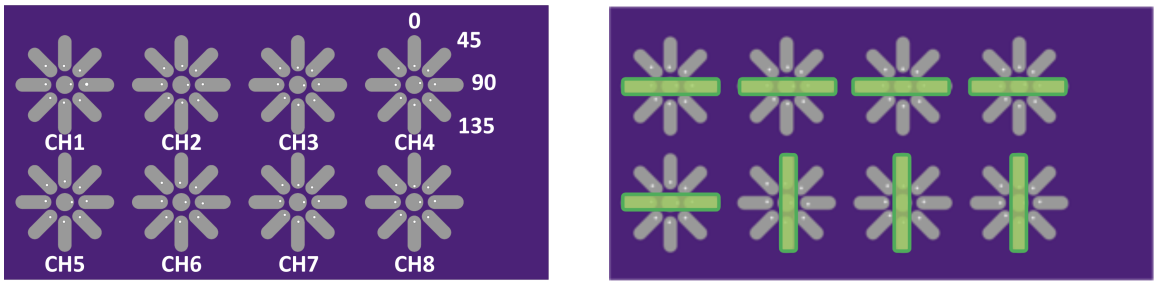
$$\text{Accuracy} = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}} \times 100\% \tag{1}$$

4.2 Development and evaluation of 8-channel electrode star array

Based on the results from the previous experiments, as illustrated in Section 5.1, the significant impact of cross-session variability was verified. This section of the study is concerned with the development and investigation of an 8-channel star electrode array which is able to change the orientation of its electrodes to minimize cross-session variability.

4.2.1 Developing the star electrode array and control array

The star electrode array is designed to be a flexible electrode array that is capable of assuming 4 different electrode orientations, namely, [0,45,90,135] degrees. By switching from a circular 10 mm gold cup electrode to rectangular electrodes, it is hypothesised that the electrodes could align themselves orthogonally to small facial muscle fibers, thereby increasing the signal-to-noise ratio (SNR) of the obtained sEMG signal. To allow the electrode to switch orientations, a central circular node is created within each electrode, and fins are formed around the node. In any given configuration, 2 fins which are diametrically opposed may be connected to a central node, which in turn is always connected to one channel of the signal acquisition circuitry. By changing which 2 fins are connected to the central node, it is possible to rotate the orientation of the approximately rectangular electrode into the desired 4 electrode orientations. This gives the electrode array its characteristic star shape as shown in Figure 7.



(a) Illustrating channels and orientations of star electrode array

(b) Hypothetical electrode configuration, green highlighting selected fins

Figure 7: 8-channel star electrode array, showing orientations, channels and potential configuration

The star array is designed with a center-to-center inter-electrode distance (IED) of 13 mm and has a total tip-to-tip fin length of 10 mm. The 4x2 array configuration is chosen to allow for area coverage of all the important musclature shown in Figure 1. Because the array is significantly different in electrode placement from the prior setup in Section 4.1, a control circular array must be developed for comparative purposes. The control array is designed with a 10 mm diameter, and a 13 mm IED to match the specifications of the star array. Both arrays are shown in Figure 8.

Both star and control arrays are fabricated on flexible polyimide PCBs, with an ENIG coating on the electrodes. Since the facial region has a high curvature, and electrode contact with the skin can only be maximised with a high degree of flexibility, there are no electrical components placed on the back of the arrays. However, since the electrodes orientation always co-selects diametrically opposing fins, all diametrically opposing fins are hard-wired together to effectively half the number of wires leaving the array. Then, a

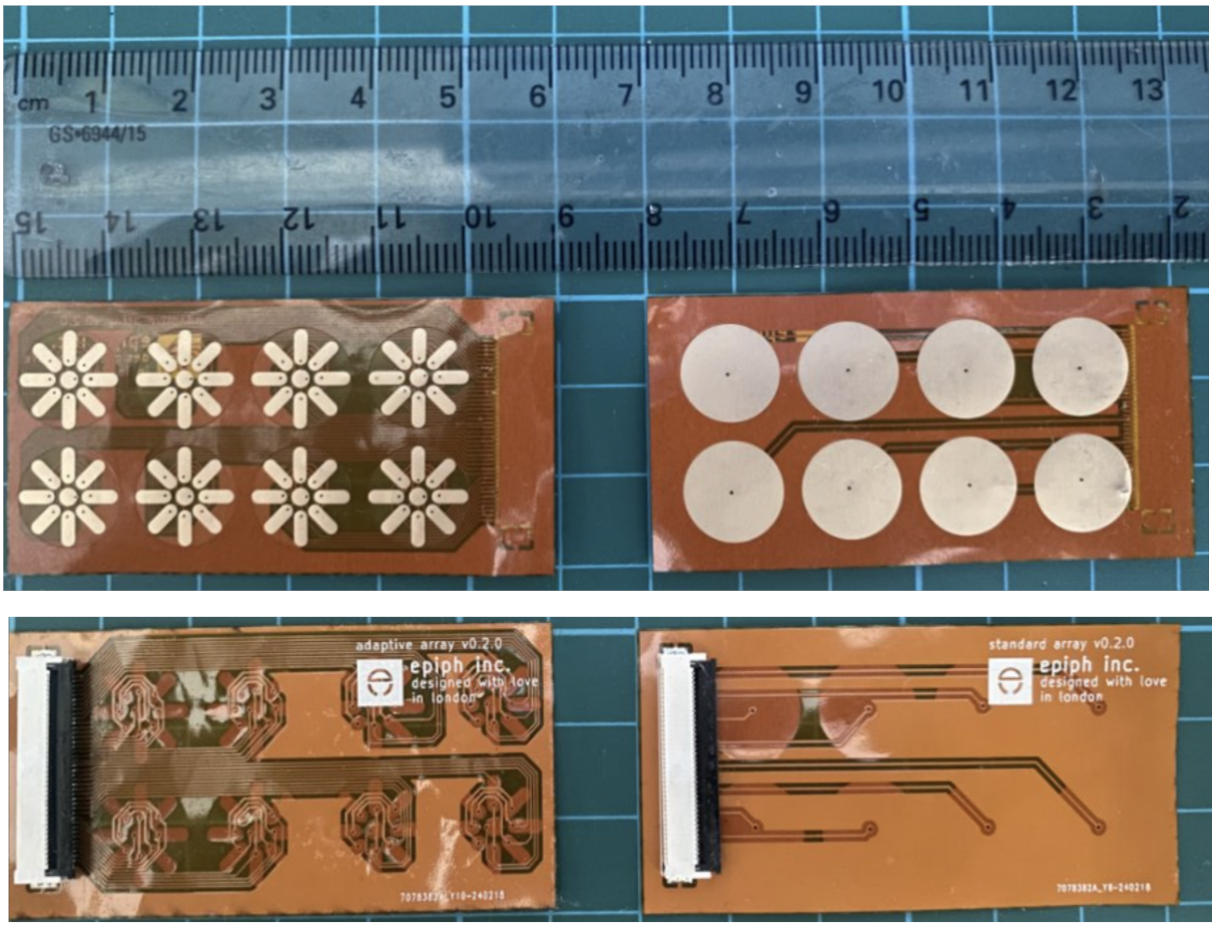


Figure 8: A side by side of the front and back of the star and control arrays

40-pin Flat Flexible Cable (FFC) connector (8 channels × 4 pairs of fins + 8 central nodes) is used to interface the electrode contacts with the signal acquisition and control circuitry described in Section 4.2.2. The full layout of both arrays can be found in Appendix C.

4.2.2 Developing the base unit

The star and control arrays are connected to a base unit that performs two critical functions. The first is signal acquisition, where the small sEMG signal must be amplified, digitized and transmitted to a computer for processing and storage. The second is control, where the electrode fins and central nodes are configured for the different desired electrode orientations. These two functions are implemented on two separate PCBs, which are combined using stackable 2.54 mm headers to form the base unit. The overall system diagram of the base unit is shown in Figure 9 and the fully connected base unit is shown in Figure 10.

Figure 9: Full system diagram of the base unit

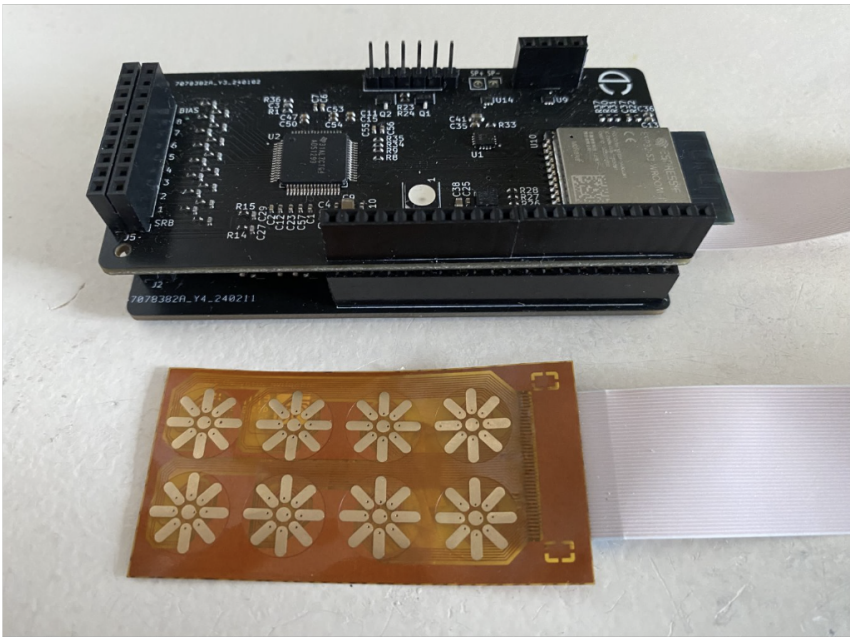


Figure 10: The stacked and fully connected base unit, with the star electrode array

4.2.3 Developing the signal acquisition circuitry

The signal acquisition board is implemented similarly to the process described in Section 2.3. It is also influenced by the design of the OpenBCI Cython board[4]. The sEMG signals are first filtered using an anti-aliasing filter set to 72.3 kHz before being passed to the ADS1299 ADC IC. Internally, the ADS1299 is configured to apply a 24 V/V gain using its programmable gain amplifier, and to sample the 8-channels simultaneously at 2 MHz. The ADS1299 then digitally decimates the signal to 250 Hz before transmitting the samples over a 2 MHz SPI bus to an ESP32-S3[5] micro-controller. The ESP32-S3 is a 240 MHz micro-controller (MCU), with built in WiFi and Bluetooth support. This allows the ESP32-S3 to send the acquired sEMG signal to a computer over the network using an User Datagram Protocol (UDP) connection. The signal acquisition board is shown in Figure 11. Full schematics and layout can be seen in Appendix D.

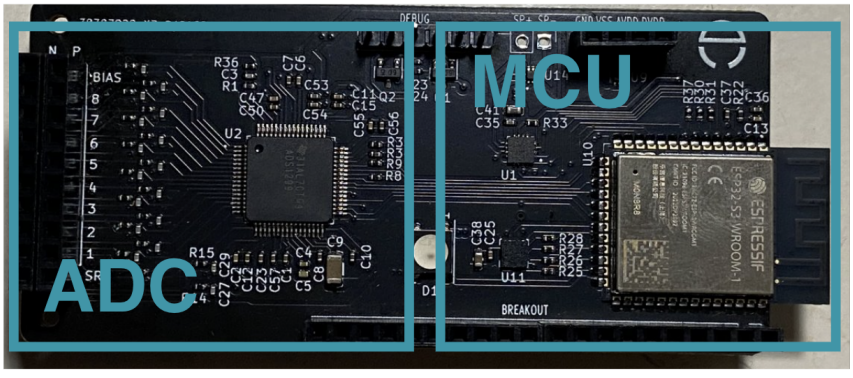


Figure 11: Labelled image of the signal acquisition circuitry.

[4]https://docs.openbci.com/Cyton/CytonLanding/
[5]https://www.espressif.com/en/products/socs/esp32-s3

4.2.4 Developing the control circuitry

The control board is responsible for connecting the correct fins to the correct channels of the ADS1299, to provide the desired electrode orientation as described in Section 4.2.1. To accomplish this, the control board is equipped with 4 ADG715[6] analog switches, where each ADG715 consists of 8 SPST switches. To achieve the multiplexing for each electrode, 4 SPST switches should be connected such that if one switch is triggered, one pair of diametrically opposed fins is connected to the central node and the corresponding signal acquisition channel. Then, each switch in the 4 switch set configures the electrode to a particular orientation. This is illustrated in Figure 12.

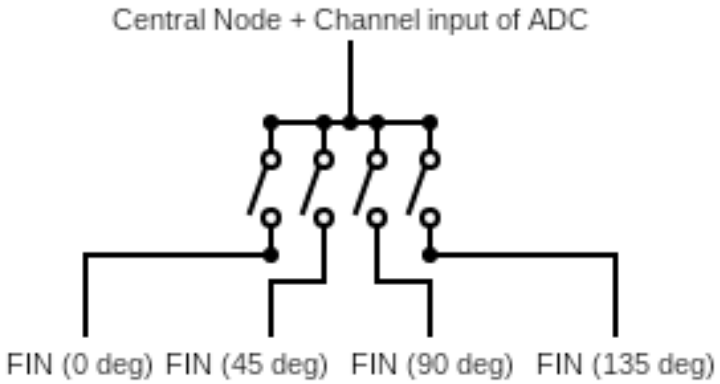


Figure 12: Multiplexing circuit diagram for a single channel

Thus, each ADG715 is able to control 2 electrode channel orientations allowing 4 ADG715 ICs to control an entire array. The control board interfaces the 4 ADG715 IC over a I2C bus to the ESP32-S3, which issues the control signals to configure the star electrode array. Since the analog frontend of the ADS1299 IC has a 2.5V dual-supply power setup, the control board passes a ± 2.5 V power supply from the signal acquisition board to the ADG715 ICs to maintain proper electrode bias. It also includes logic-level shifters to convert the 3.3V logic of the ESP32-S3 to 2.5V for the ADG715. The control board circuit is shown in Figure 13. Full schematics and layout can be seen in Appendix E.

4.2.5 Investigating muscle-electrode alignment with star electrode array

First, the feasibility of muscle-electrode alignment with the star electrode array must be determined. It is expected that when the electrodes are perfectly orthogonal to the direction of muscle fibers underneath the electrode, then the SNR observed on the corresponding sEMG signal is maximal. Therefore, the level of muscle-electrode alignment for a given electrode can be identified by looking at the SNR of the corresponding sEMG channel. Thus, the optimal electrode orientation can be defined as the orientation that maximises the SNR of its corresponding sEMG channel.

[6] https://www.analog.com/en/products/adg715.html

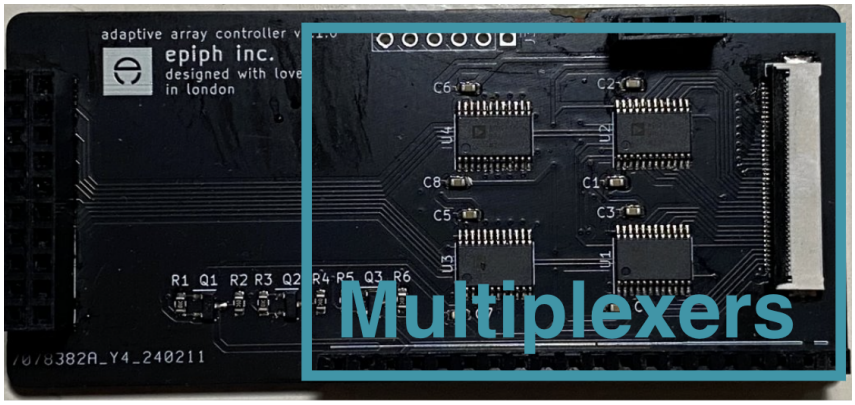


Figure 13: Labelled image of the control circuitry

One complication is that given the silent speech recognition task, different mouthing patterns corresponding to different phonemes and therefore activate different parts of the facial musculature. Thus, a static electrode orientation may be optimal for some mouthing patterns but sub-optimal for others. In other words, there may not be a single electrode orientation that maximises SNR across every possible phoneme. Instead, a sufficient condition of optimality can be proposed as the electrode orientation that has the maximum average SNR across all possible phonemes. Therefore, the efficacy of the star array in achieving muscle-electrode alignment can be proven if there is an optimal electrode orientation for each channel that has a greater average SNR across all phonemes as compared to the corresponding channel on the control array. Additionally, it is important to validate that the optimal electrode orientation changes across sessions. Otherwise, it would be sufficient to have a static electrode array with rectangular electrodes in the optimal orientation.

The experimental procedure for validating muscle-electrode alignment in the star array is as follows. First, a line is drawn on a participant's face, from the earlobe to the tip of the chin. The midpoint of the line is marked, and aligned with the center of the star electrode array, with the array longitudinally parallel to the jawline. Once the position of the array is determined, it is fixed in place with bio-safe double-sided adhesive around the edges. This is shown in Figure 14. Then, the array is connected to the base unit and the base unit is connected to the laptop. After that, the array is configured such that all the channels are set to the 0 degree orientation, and the data stream is started. First, one minute of data is recorded with the participant remaining relaxed and silent without moving their mouth.

$$\overline{x_i^2} = \frac{1}{N}\sum_{n=0}^{N} x_i[n]^2 \tag{2}$$

The noise power can then be established per channel using Equation 2, where there are N samples of data and $x_i[n]$ is the n^{th} sEMG data sample for channel i. Then, the participant is instructed to silently mouth a phonetically balanced passage and the sEMG data is recorded again. For this experiment, The Rainbow Passage, a phonetically balanced passage commonly used in speech evaluation [39], is mouthed. The second recording can be used to establish the signal power for each channel using Equation 2,

Figure 14: A participant after the star array has been attached to the cheek.

similar to the noise power. Finally, the SNR for channel i using the 0 degree configuration can be written as:

$$SNR_i = 10 \log_{10} \left(\frac{\overline{x_i^2}}{\overline{n_i^2}} \right) \tag{3}$$

where $\overline{n_i^2}$ is the noise power and $\overline{x_i^2}$ is the signal power. This SNR measurement must be repeated for all 4 possible orientations of the star array and for the control array which will results in an SNR value for every channel in every possible electrode orientation and the control configuration. Then, the optimal electrode orientation for a given channel and it's corresponding optimal SNR can be identified by selecting the orientation that maximises the SNR for the channel. When the optimal electrode orientations for all channels are identified, they can be considered the optimal electrode configuration for the star array, with an associated optimal SNR for each channel. This can be compared with the control configuration and its SNR per channel, to determine if the star electrode array is able to achieve superior SNR and by extension better muscle-electrode alignment. This whole process must then be repeated over two sessions to understand the change of optimal electrode orientation across sessions.

4.2.6 Investigating cross-session accuracy with star electrode array

Once the effectiveness of the 8-channel star electrode array in achieving muscle-electrode alignment is determined, it is important to establish the impact that maximal SNR has on cross-session silent speech recognition accuracy. This experiment extends the SNR anal-

ysis by further establishing the difference in cross-session classification accuracy between optimal electrode configurations on the star array, and the control array. For simplicity, only the 3-word classification problem is considered in this experiment. The experimental procedure is as follows.

First, the participants attaches the 8-channel star electrode array and identifies the optimal electrode configuration as described in Section 4.2.5. Then, with the optimal electrode configuration, the 3-word dataset consisting of 50 samples each from the [air, bat, cap] vocabulary is collected following the procedure in Section 4.1.3. This is repeated with the control array as well and the whole process is repeated across a total of two sessions for both arrays.

After the data collection, there are two datasets of 300 samples containing two sessions each. The datasets corresponds to either the optimal electrode configuration with the star array, or the control configuration. The single and cross-session accuracy is computed for each configuration using the datasets and a LDA classifier. The classification procedure used is identical to the one outlined earlier in Section 4.1.4 and 4.1.5.

5 Results

5.1 Investigating impact of cross-session variability on state-of-the-art

First, the results for the feature selection and the single/cross session accuracies are presented in Figure 15. As noted in Section 4.1.5, the single and cross session accuracies are reported as mean values. The learning curve of the model for the highest accuracy case (Time-Frequency, Single Session) is shown in Figure 16.

	All	**Time-only**	**Frequency-only**	**Time-Frequency**
Single Session	86.5%	86.8%	76.5%	**90.5%**
Cross Session	56.5%	56.7%	47.3%	**59.0%**
Combined Session	**81.9%**	81.9%	67.2%	77%

Figure 15: Test accuracy for 3-word dataset using LDA model.

While the Time-Frequency feature set is the best performing feature set for both single and cross session evaluation, the All feature set performs optimally for the combined session task. It is also clear that there is a large accuracy decrease of 30% between the single session and cross session classification tasks regardless of the features that are deployed. Training on both session jointly, as illustrated in the combined session task, has a significant boost in classification accuracy in line with the results from Maier-Hein *et al.* [18]. Importantly, the Time-Frequency features are chosen for the star-electrode experiments, due to their superior classification performance on the single and cross session tasks.

The results from the 30-word classification task are presented in Figure 17. The issue of cross-session accuracy is once again highlighted in the drastic 40% decrease in classification accuracy from the single-session case.

5.2 Investigating muscle-electrode alignment with star electrode array

Following the procedure outlined in Section 4.2.5, the optimal electrode configurations for the star array across two sessions are identified and shown in Figure 18.

From Figure 19, it is clear that the optimal electrode configuration consistently delivers

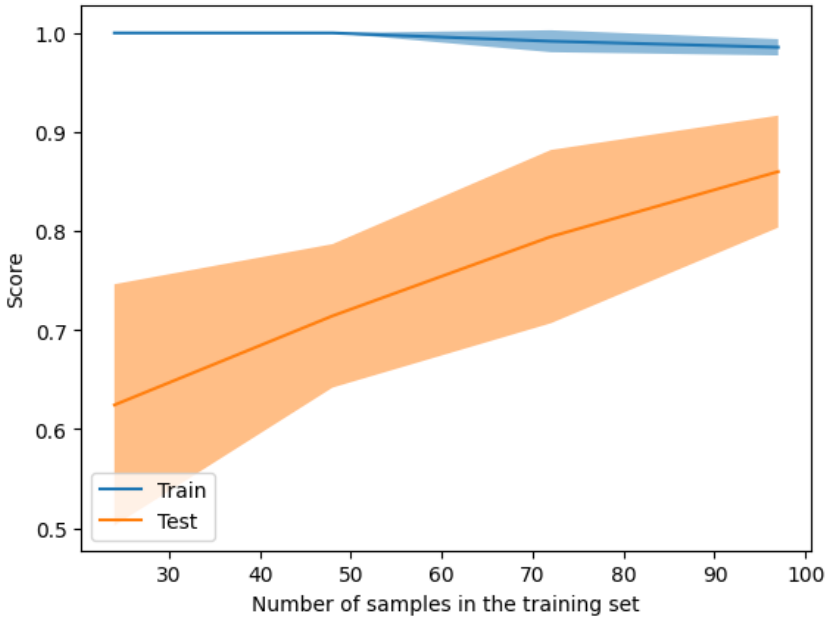
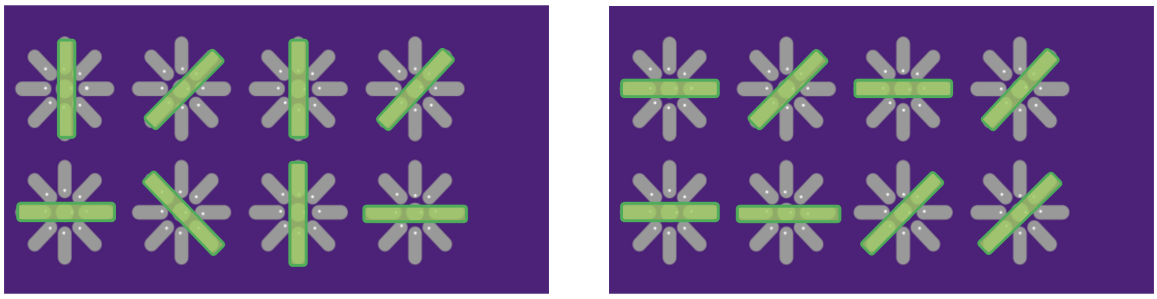


Figure 16: The learning curve for the Time-Frequency features with the LDA on Session 0.

Single Session	Cross Session	Combined Session
80.5% ± 6.5%	38.5% ± 4.5%	60.0% ± 4.0%

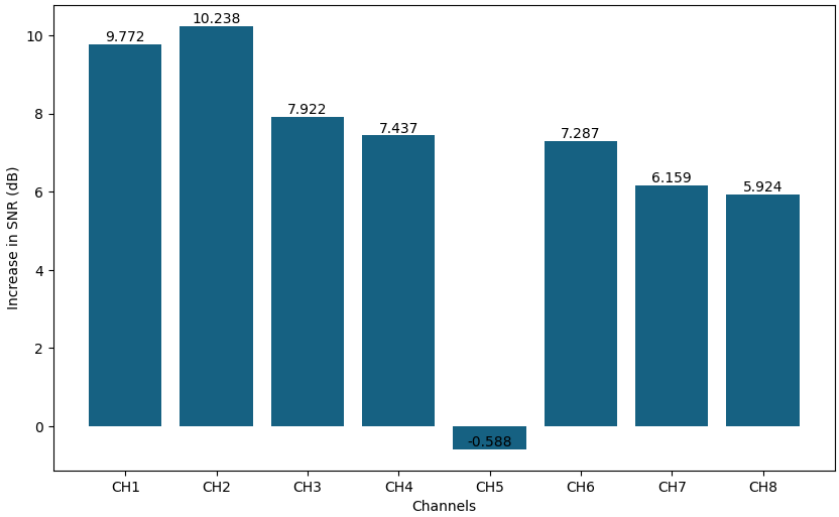
Figure 17: Test Accuracy for 30-word dataset using HMM model.

a significant increase to SNR as compared to the control array. There is however an outlier in Channel 5, which shows decreased performance on the star electrode array. This is addressed in greater detail in Section 6.2. From Figure 18, it is noted that the optimal electrode configuration changes significantly across sessions which is a promising indication for the utility of the electrode star array's configurable electrode orientations.

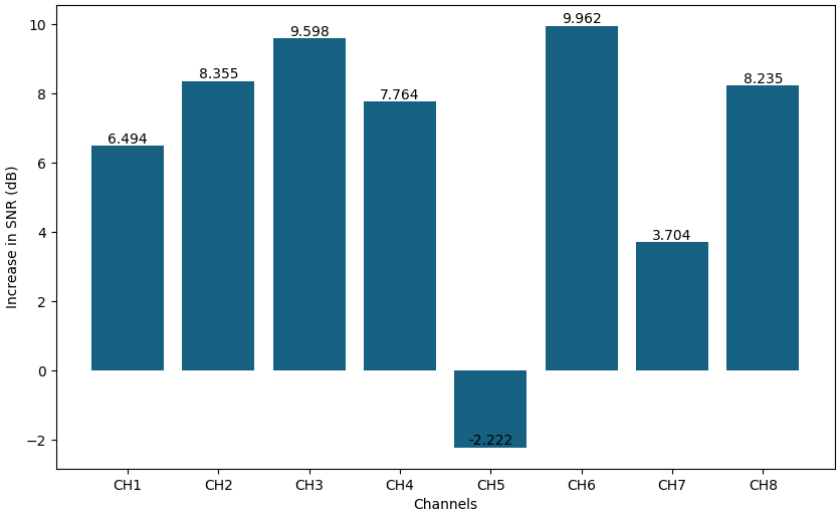


(a) Session 1 (b) Session 2

Figure 18: Visualization of optimal electrode configuration over both sessions.



(a) Session 1



(b) Session 2

Figure 19: Increase in SNR (dB) when using optimal electrode configuration compared to control configuration, in both sessions.

5.3 Investigating cross-session accuracy with star electrode array

The results for the classification accuracy using the star electrode array and control array are presented in Table 20 using the same format as in Section 5.1 for consistency. Overall, the results are overwhelmingly positive, with the star electrode array outperforming the control setup across single-session, cross-session and combined-session classification. Notably, the baseline classification accuracy of the control array is significantly greater than the in the previous experiment using the electrode brace as shown in Figure 15. Furthermore, it is shown that the star electrode array increases the cross-session classification accuracy by 41% (absolute 21.7%) which is strong evidence in favour of its effectiveness.

	Single Session	**Cross Session**	**Combined Session**
Control Array	94.9%	52.6%	98.3%
Star Array	**98.3%**	**74.3%**	**100%**

Figure 20: Test accuracy for 3-word dataset using LDA model on star electrode array and control electrode array.

6 Discussion

6.1 Impact of cross-session variability on state-of-the-art

As indicated in Table 15, the 3-word classification accuracy of 90.5% in the setup approaches the state-of-the-art. Additionally, the impact of cross-session variability on classification accuracy is significant as seen from both Table 15, 17 and in agreement with the literature [18]. The effectiveness of combined session training to tackle cross-session variability is also promising. It may be possible to increase cross-session classification performance, by scaling the dataset to include many sessions from the same participant. However, given the speaker-dependent nature of silent speech recognition systems, collecting a large dataset for every user would be impractical. Instead, an approach where a small speaker-specific dataset can be used to achieve a session-independent model is desirable. This supports the approach that is enabled by the 8-channel star electrode array, which minimizes cross-session variability using hardware.

There is one limitation to this experiment. In both the 3-word and 30-word classification task, the accuracy is lower than the best silent speech recognition systems outlined in the literature [8][15]. This could be due to the relatively small dataset size of this study, consisting of only 50 and 10 samples per word per session in Dataset 1 and Dataset 2 respectively. As seen in Table 4, the state-of-the-art systems have much larger datasets collected across multiple participants. Additionally, many of them use neural network architectures to achieve higher classification accuracies. However, given the small size of the dataset in this study, it was found that neural networks did not perform as well as the simpler models that were deployed. Therefore it is likely that the achievable classification accuracy using the setup could be higher than the reported values if the dataset size is increased. Due to the small dataset size, concerns about the impact of over-fitting on the decrease in cross-session accuracy must be addressed. In the 3-word classification task, the Time-Frequency feature set is intentionally kept small (4 features per channel), and the LDA with 2 components is also a simple classifier which minimizes over-fitting. Additionally, the learning curve of the models as shown in Figure 16 do not show any signs of over-fitting.

Since the effects of the small dataset size were mitigated, the results support the hypothesis that there is an opportunity to improve the cross-session performance of silent speech recognition systems without massively scaling dataset size requirements. This motivates the development of the 8-channel star electrode array.

6.2 Muscle-electrode alignment with the star electrode array

The results of SNR improvements through muscle-electrode alignment shown in Section 5.2 are very positive. It is shown in Figure 19 that a large increase in SNR is consistently achieved by the star electrode array across multiple sessions. Additionally, Figure 18 indicates that there is a change in optimal electrode orientation across sessions, which supports the configurable electrode orientations which are enabled by the proposed star electrode array. One interesting result is that channel 5 is shown to have consistently worse SNR than the same channel on the control electrode array as seen in Figure 19. Referring to Figure 1, channel 5 is attached to the lower buccal region of the face and one plausible explanation is that it is resting over a muscle attachment region or another region with little sEMG activity. Further study and analysis is required before the issue can be conclusively identified.

These results could be validated with greater certainty by increasing the number of sessions and participants, as currently the experiment only contains 2 sessions from a single participant. By incorporating a greater diversity of participants, the optimal electrode configuration could be analyzed across different facial musculature, and mouthing patterns. Nevertheless, both results support further testing of the star electrode array, and show strong indication of its utility in maximising the SNR for silent speech recognition systems.

6.3 Impact of star electrode array on cross-session accuracy

Following the results of the previous section, the impact of the star electrode array on classification accuracy is extremely positive. Firstly, from Figure 20, it is noted that both the control and star array display significantly higher single and combined session performance as compared to the results on the electrode brace (Section 5.1). This is surprising as it suggests that an electrode array placed on the cheek can obtain more information relevant for silent speech recognition, as compared to the sparsely distributed electrodes across the cheek, submental region and throat in the case of the electrode brace. It is possible that an electrode array on the cheek is able to more accurately capture the activity of the densely-spaced and small facial muscles involved in mouthing words. This is a promising outcome, as a single electrode array on the cheek is far simpler to wear and less obtrusive as compared to the electrode brace in Section 4.1.1.

Additionally, the star electrode array outperforms the control array on single-session, cross-session and combined-session accuracy. This is likely attributed to the greater SNR observed in the sEMG signals from the star electrode array with the optimal electrode configuration. Due to the small size of the test dataset (30 samples), it is hard to conclude if the difference in the single-session and combined-session accuracy is significant. However, the cross-session accuracy of the star-electrode array is significantly better than the control array, with a 41% relative increase. This provides strong evidence for the utility of the star electrode array and suggests that a low-density electrode array capable of modifying the orientation of its electrodes can be used to increase the cross-session accuracy in silent speech interfaces.

7 Conclusion and Future Work

This study aimed to investigate the use of a low-density electrode array capable of modifying the orientation of its electrodes to combat cross-session variability in silent speech recognition systems. First, the adverse effects of cross-session variability was demonstrated on a state-of-the-art silent speech recognition system, showing a 34% relative decrease in classification accuracy.

Then, a novel multiplexed star electrode array was proposed, with the ability to modify the orientation of its electrodes. The array's configurable electrodes, designed to optimize muscle-electrode alignment, proved successful in significantly improving SNR in the sEMG signal by an average of 6.7 dB across all channels. Additionally, it was shown that the optimal electrode configuration changes across sessions and the adaptability of the star electrode array is justified. Consequently, the star electrode array achieved a notable 41% relative increase in cross-session accuracy compared to the control array.

The outcomes of this study affirm the potential of the star electrode array as a viable, simpler, and more practical alternative to HD-sEMG for silent speech interfaces. Its implementation on the cheek is portable and much easier to deploy than other HD-sEMG systems. The considerable boost in cross-session accuracy underscores the star electrode array's capacity to mitigate the challenges posed by variability across sessions.

For future work, expanding the number of sessions and participant diversity is critical. This would further validate the star electrode array's effectiveness across a wider set of facial structures and mouthing patterns. Additionally, exploring the integration of deep learning models, trained on larger datasets, could harness the increased SNR for even higher accuracy. Implementing an automatic electrode orientation adjustment during live operation, based on real-time SNR analysis, could also be an intriguing direction to achieve session independence in practice. Furthermore, applying the findings to other EMG-based applications, such as prosthetic control or rehabilitation devices, could expand the impact of this research beyond silent speech recognition.

References

[1] Grand View Research, "Voice user interface market size, share trends analysis report by offering, by application, by industry vertical, by region, and segment forecast, 2023 - 2030." `https://www.grandviewresearch.com/industry-analysis/voice-user-interface-market-report`, 2023. Accessed: 2024-03-27.

[2] Z. Wang, H. Wang, H. Yu, and F. Lu, "Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 5, pp. 524–534, 2021.

[3] A. M. Klein, K. Kölln, J. Deutschländer, and M. Rauschenberger, "Design and evaluation of voice user interfaces: What should one consider?," *Lecture Notes in Computer Science*, p. 167–190, 2023.

[4] Y. Deng, G. Colby, J. T. Heaton, and G. S. Meltzner, "Signal processing advances for the mute semg-based silent speech recognition system — ieee conference publication — ieee xplore," 11 2012.

[5] Y. Wang, T. Tang, Y. Xu, Y. Bai, L. Yin, G. Li, H. Zhang, H. Liu, and Y. Huang, "All-weather, natural silent speech recognition via machine-learning-assisted tattoo-like electronics," Aug 2021.

[6] C. E. Stepp, J. T. Heaton, R. G. Rolland, and R. E. Hillman, "Neck and face surface electromyography for prosthetic voice control after total laryngectomy," Apr 2009.

[7] C. Jorgensen and K. Binsted, "Web browser control using emg based sub vocal speech recognition," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pp. 294c–294c, 2005.

[8] A. Kapur, S. Kapur, and P. Maes, "Alterego: A personalized wearable silent speech interface," in *23rd International conference on intelligent user interfaces*, pp. 43–53, 2018.

[9] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, pp. 270–287, 04 2010.

[10] R. Zhang, K. Li, Y. Hao, Y. Wang, Z. Lai, C. Zhang, and F. Guimbretière, "Echospeech: Continuous silent speech recognition on minimally-obtrusive eyewear powered by acoustic sensing," vol. 18, 2023.

[11] T. Srivastava, P. Khanna, S. Pan, P. Nguyen, and S. Jain, "Muteit," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 6, pp. 1–26, 09 2022.

[12] S. L. Metzger, K. T. Littlejohn, A. B. Silva, D. A. Moses, M. P. Seaton, R. Wang, M. E. Dougherty, J. R. Liu, P. Wu, M. A. Berger, and et al., "A high-performance neuroprosthesis for speech decoding and avatar control," Aug 2023.

[13] F. R. Willett, E. M. Kunz, C. Fan, D. T. Avansino, G. H. Wilson, E. Y. Choi, F. Kamdar, M. F. Glasser, L. R. Hochberg, S. Druckmann, and et al., "A high-performance speech neuroprosthesis," Aug 2023.

[14] N. Kimura, M. Kono, and J. Rekimoto, "Sottovoce," *arXiv (Cornell University)*, 05 2019.

[15] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Development of semg sensors and algorithms for silent speech recognition," *Journal of Neural Engineering*, vol. 15, p. 046031, 06 2018.

[16] D. Hewson, J. Duchêne, and J.-Y. Hogrel, "Changes in impedance at the electrode-skin interface of surface emg electrodes during long-term emg recordings," in *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4, pp. 3345–3348, IEEE, 2001.

[17] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, and G. Rau, "Development of recommendations for semg sensors and sensor placement procedures," *Journal of electromyography and Kinesiology*, vol. 10, no. 5, pp. 361–374, 2000.

[18] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," 12 2005.

[19] L. Diener, G. Felsch, M. Angrick, and T. Schultz, "Session-independent array-based emg-to-speech conversion using convolutional neural networks — vde conference publication — ieee xplore," 08 2018.

[20] Z. Deng, X. Zhang, X. Chen, X. Chen, X. Chen, and E. Yin, "Silent speech recognition based on surface electromyography using a few electrode sites under the guidance from high-density electrode arrays — ieee journals magazine — ieee xplore," 02 2023.

[21] D. Conant, K. E. Bouchard, and E. F. Chang, "Speech map in the human ventral sensory-motor cortex," Feb 2014.

[22] I. Campanini, A. Merlo, C. Disselhorst-Klug, L. Mesin, S. Muceli, and R. Merletti, "Fundamental concepts of bipolar and high-density surface emg understanding and teaching for clinical, occupational, and sport applications: Origin, detection, and main errors," *Sensors*, vol. 22, p. 4150, May 2022.

[23] A. B. Dobrucki, P. Pruchnicki, P. Plaskota, P. Staroniewicz, S. Brachmański, and M. Walczyński, "Silent speech recognition by surface electromyography," Jul 2016.

[24] D. Gaddy and D. Klein, "Digital voicing of silent speech," 2020.

[25] M. Janke and L. Diener, "Emg-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2375–2385, 2017.

[26] H. Tankisi, D. Burke, L. Cui, M. de Carvalho, S. Kuwabara, S. D. Nandedkar, S. Rutkove, E. Stålberg, M. J. van Putten, and A. Fuglsang-Frederiksen, "Standards of instrumentation of emg," *Clinical Neurophysiology*, vol. 131, p. 243–258, Jan 2020.

[27] R. H. Chowdhury, M. B. I. Reaz, M. A. B. M. Ali, A. A. A. Bakar, K. Chellappan, and T. G. Chang, "Surface electromyography signal processing and classification techniques," *Sensors (Basel, Switzerland)*, Sep 2013.

[28] R. Merletti, M. Besomi, A. Botter, C. DeLuca, H. Hermens, D. Hewson, R. Merletti, E. Spinelli, L. Bareket, and et al., "Tutorial. surface emg detection, conditioning and pre-processing: Best practices," Jun 2020.

[29] A. Abdullah and K. Chemmangat, "A computationally efficient semg based silent speech interface using channel reduction and decision tree based classification," *Procedia Computer Science*, vol. 171, pp. 120–129, 2020. Third International Conference on Computing and Network Communications (CoCoNet'19).

[30] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022.

[31] R. Song, X. Zhang, X. Chen, X. Chen, X. Chen, S. Yang, and E. Yin, "Decoding silent speech from high-density surface electromyographic data using transformer," *Biomedical Signal Processing and Control*, vol. 80, p. 104298, 02 2023.

[32] T. Schultz and M. Wand, "Modeling coarticulation in emg-based continuous speech recognition," *Speech Communication*, vol. 52, pp. 341–353, 04 2010.

[33] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, pp. 125–128 vol.1, 1993.

[34] M. S. Morse and E. M. O'Brien, "Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes," *Computers in Biology and Medicine*, vol. 16, pp. 399–410, 01 1986.

[35] N. Sugie and K. Tsunoda, "A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production — ieee journals magazine — ieee xplore," 07 1985.

[36] W. Lai, Q. Yang, Y. Mao, E. Sun, and J. Ye, "Knowledge distilled ensemble model for semg-based silent speech interface," in *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*, pp. 117–122, 2023.

[37] M. Wand and T. Schultz, "Speaker-adaptive speech recognition based on surface electromyography," *Communications in computer and information science*, pp. 271–285, 01 2010.

[38] T. Kubo, M. Yoshida, T. Hattori, and K. Ikeda, "Towards excluding redundancy in electrode grid for automatic speech recognition based on surface emg," *Neurocomputing*, vol. 134, pp. 15–19, 06 2014.

[39] J. S. Sevitz, B. R. Kiefer, J. E. Huber, and M. S. Troche, "Obtaining objective clinical measures during telehealth evaluations of dysarthria," *American Journal of Speech-Language Pathology*, vol. 30, p. 503–516, Mar 2021.

Appendices

A Code, documentation and datasts

All the relevant code, cad files and datasets are freely available on Github [7] for replication.

B Full 30-word vocabulary

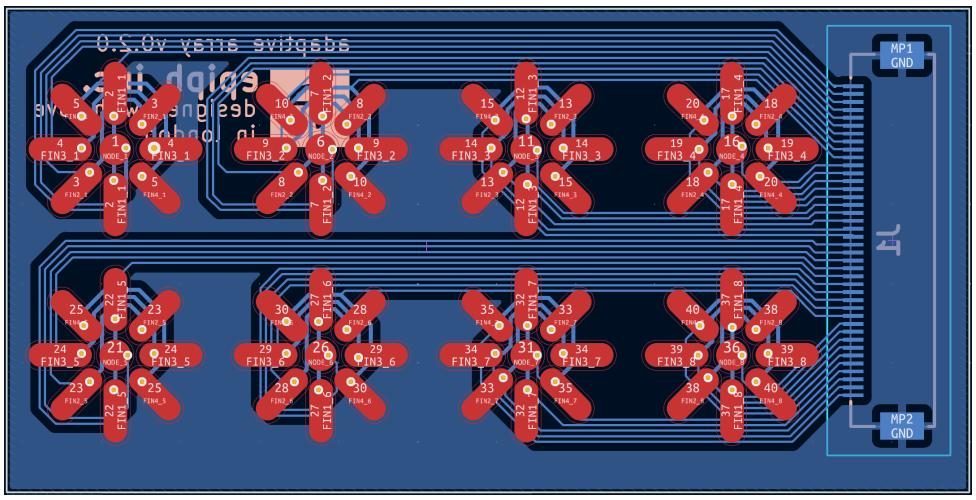
The full 30 word Talon phonetic alphabet is given as follows.

Listing 1: Talon Phonetic Alphabet

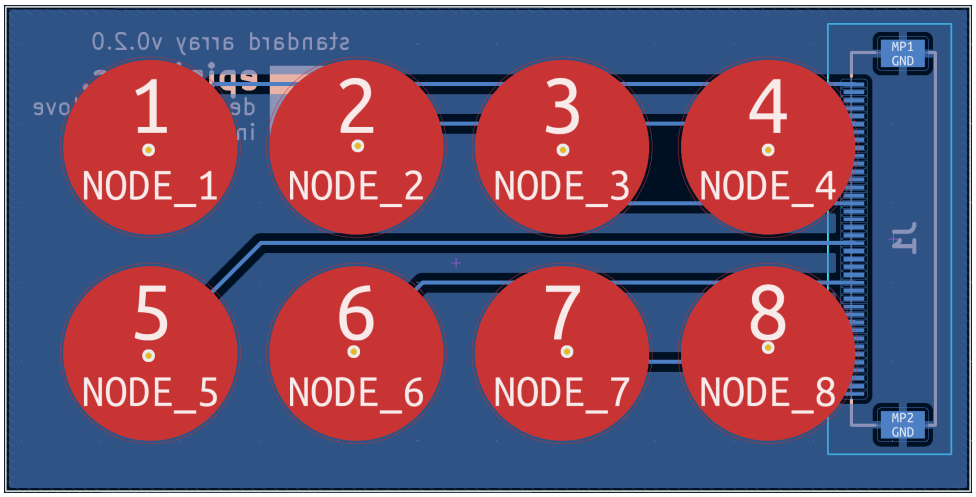
```
air     a
bat     b
cap     c
drum    d
each    e
fine    f
gust    g
harp    h
sit     i
jury    j
crunch  k
look    l
made    m
near    n
odd     o
pit     p
quench  q
red     r
sun     s
trap    t
urge    u
vest    v
whale   w
plex    x
yank    y
zip     z
```

[7] https://github.com/solderneer/nexus-silent-speech

C Electrode array layouts



(a) Star electrode array layout



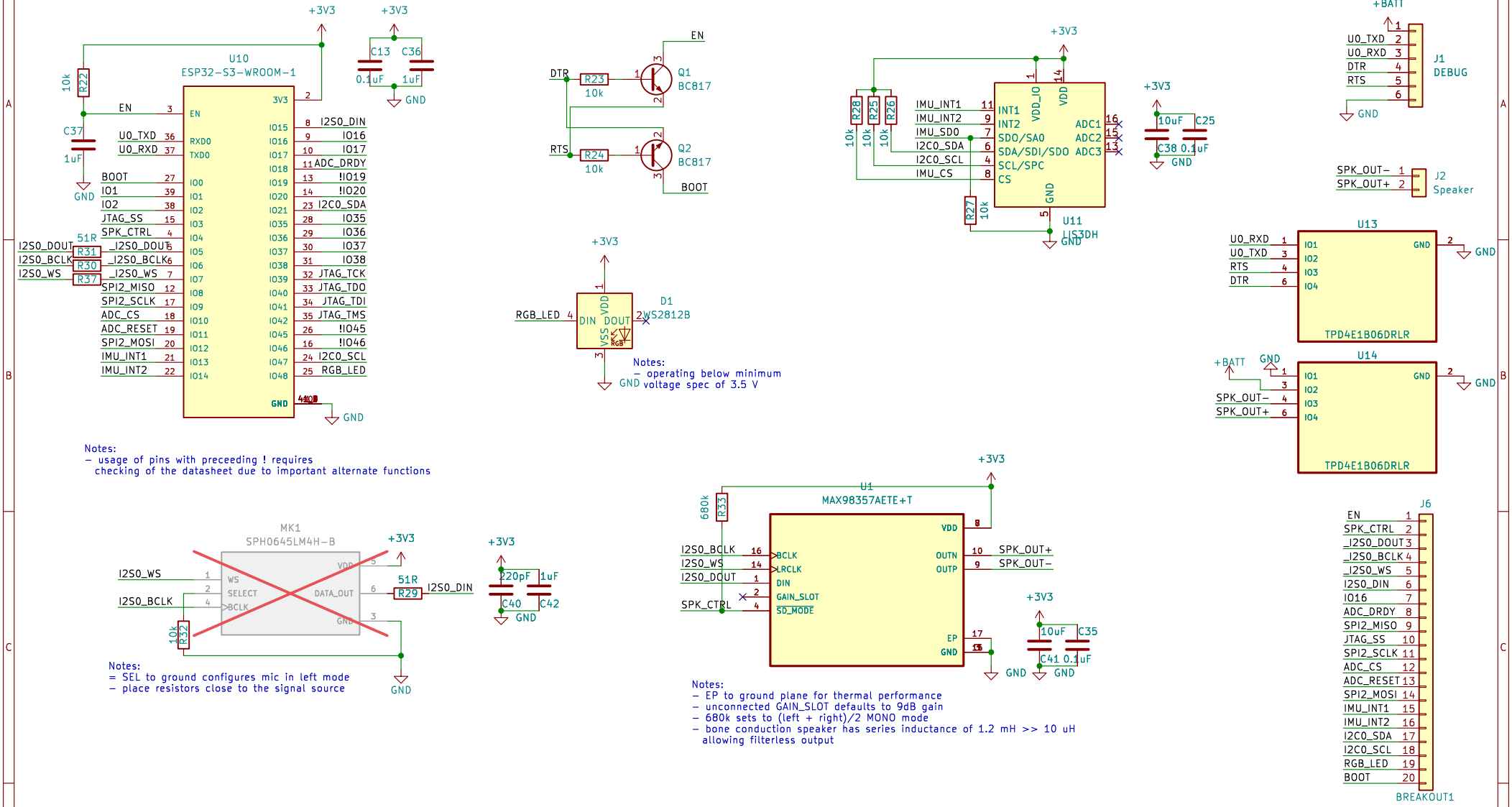
(b) Control electrode array layout

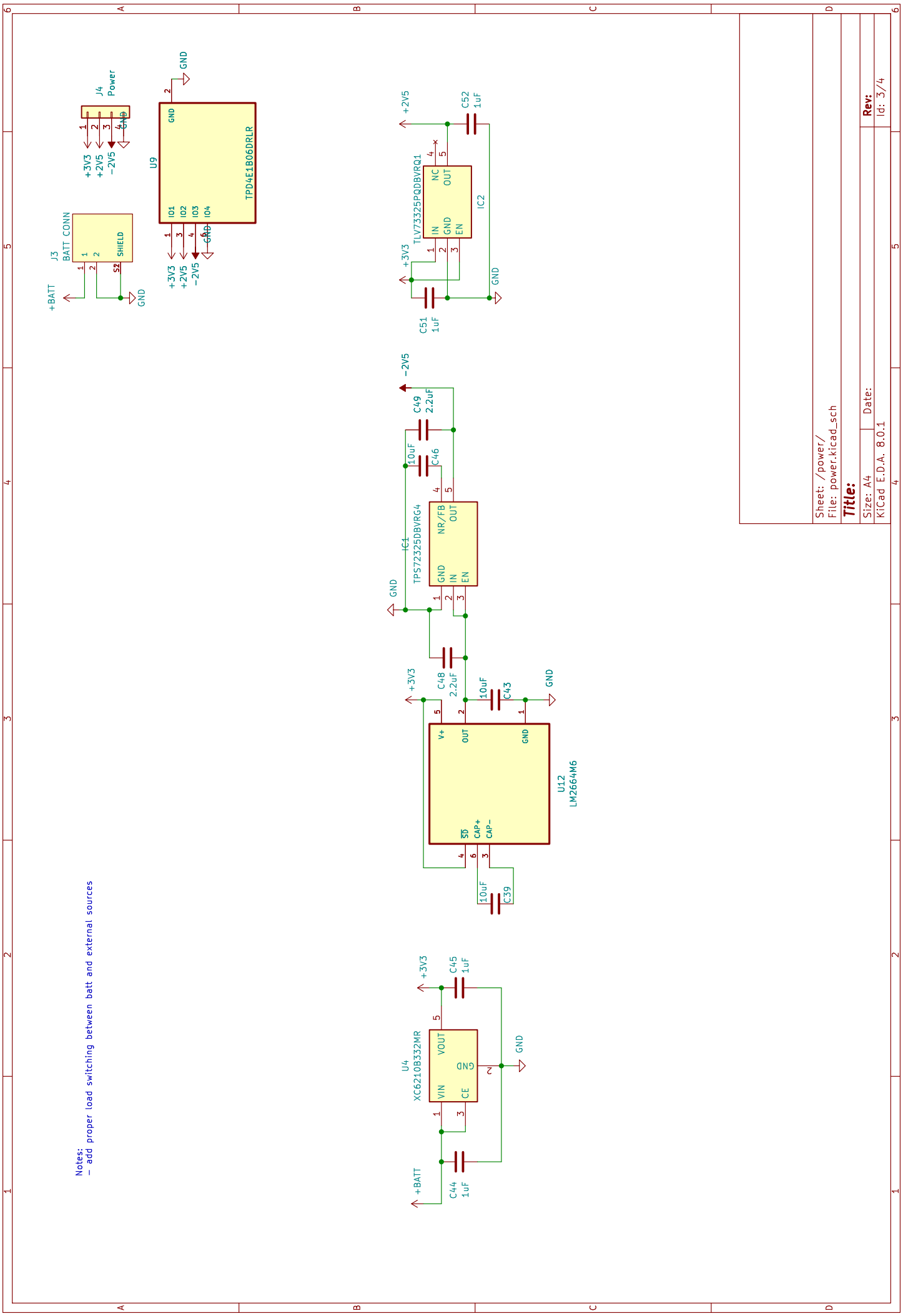
Figure 21: PCB layout for the electrode arrays

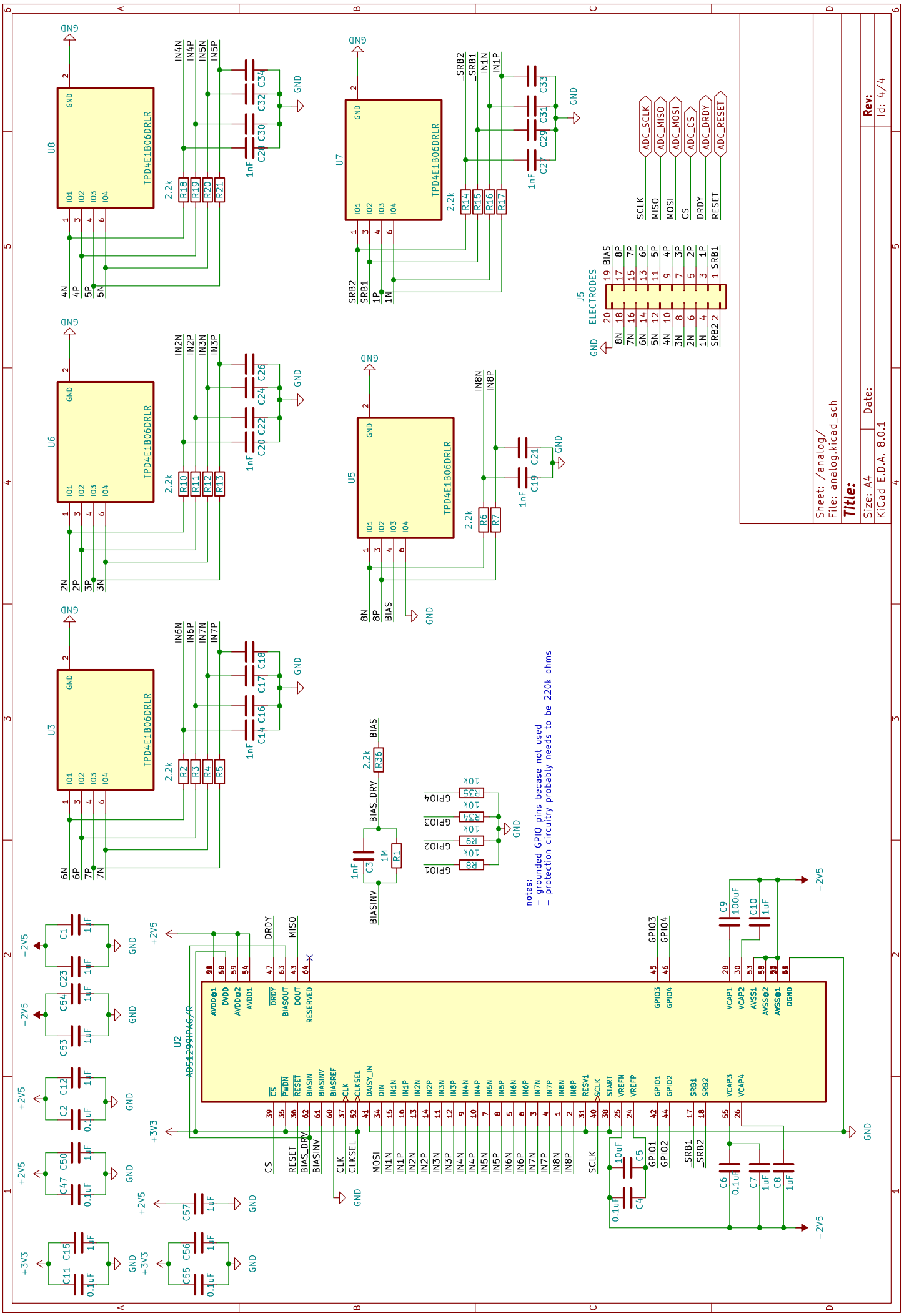
D Signal acquisition board

4.1 Schematic diagram

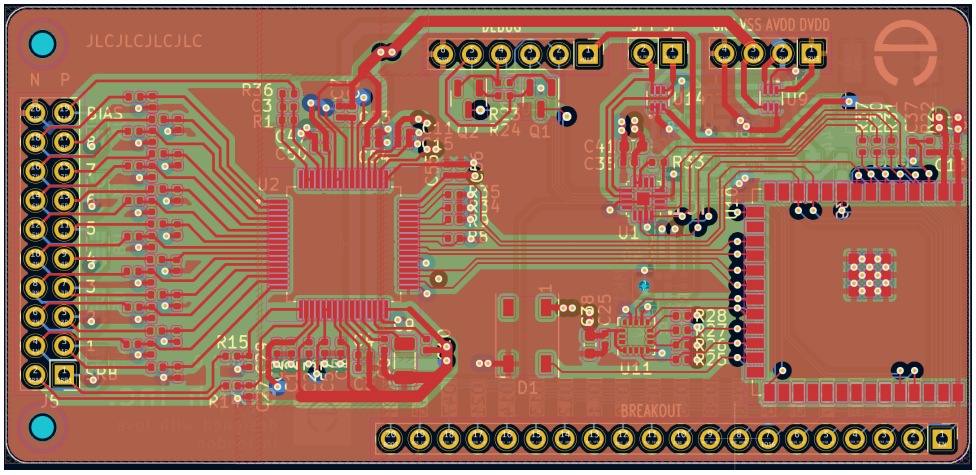
Shown on the following page.



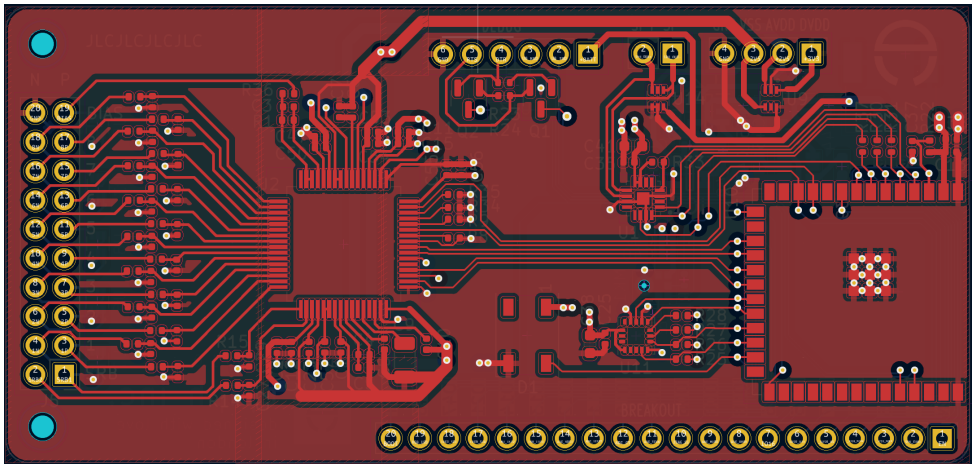




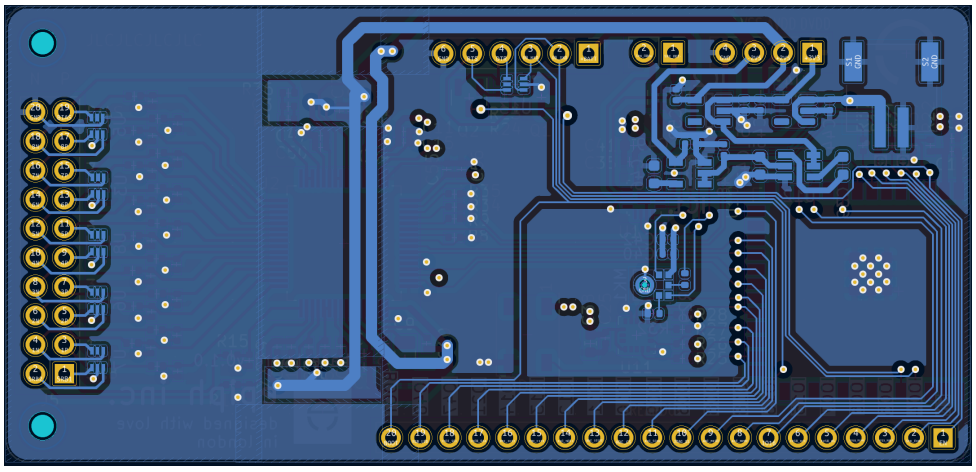
4.2 Layout



(a) Full layout



(b) Top layer



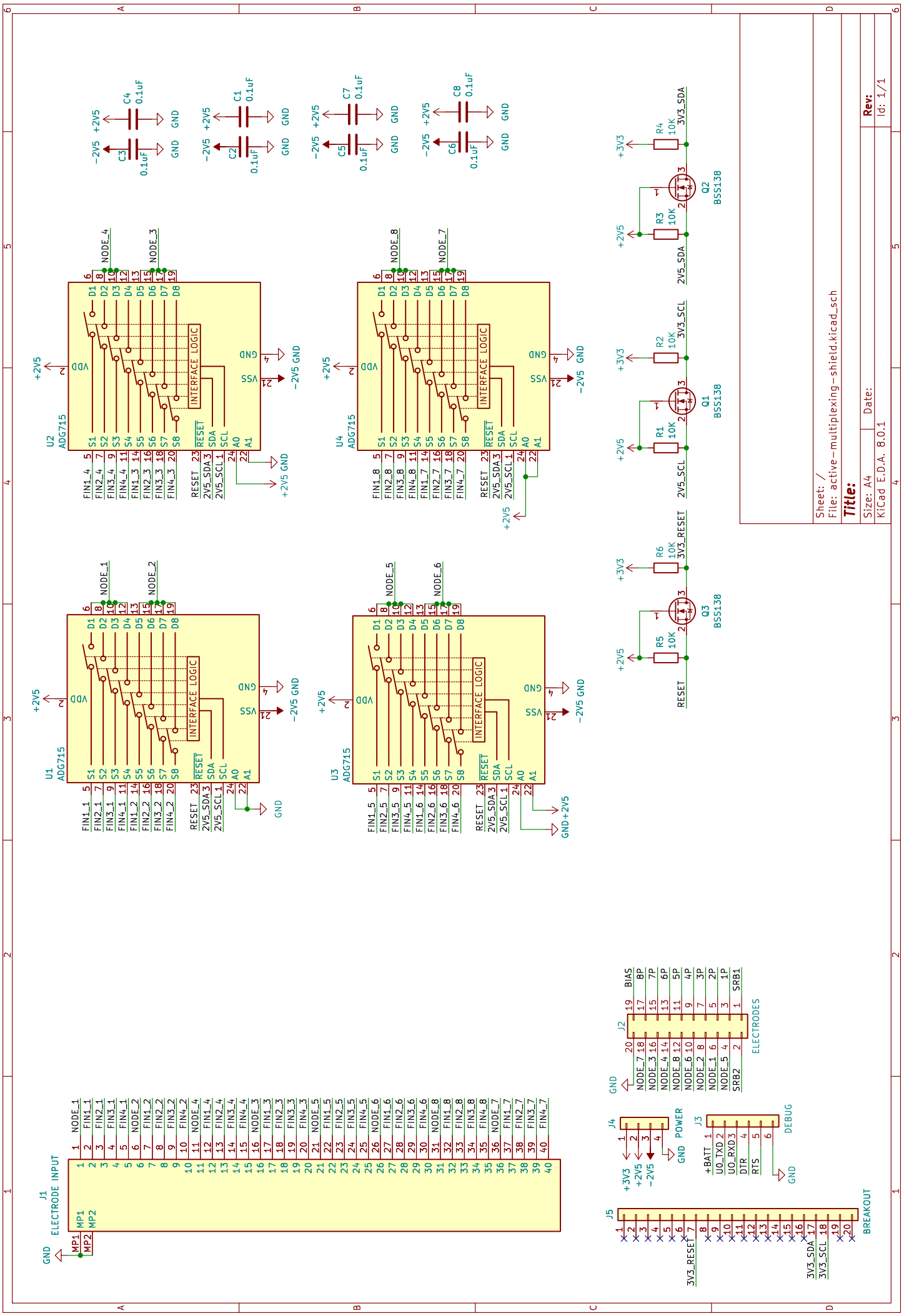
(c) Bottom layer

Figure 22: PCB layout for the signal acquisition board

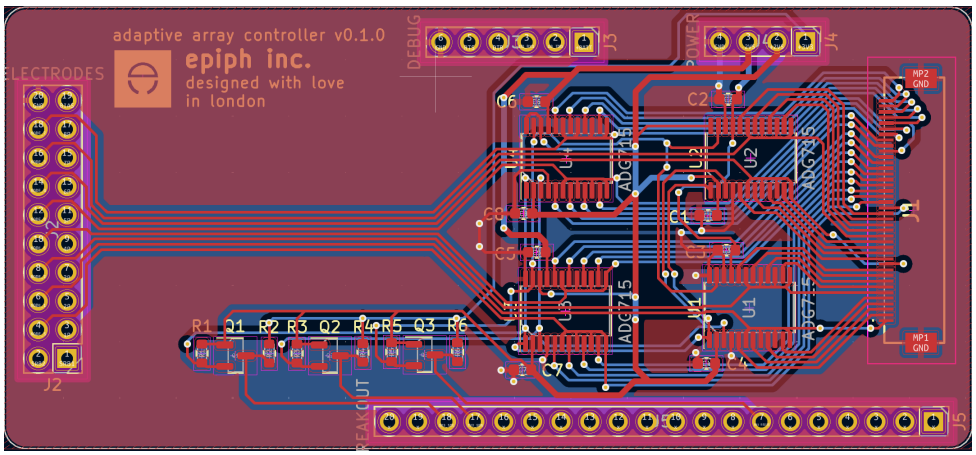
E Control board

5.1 Schematic diagram

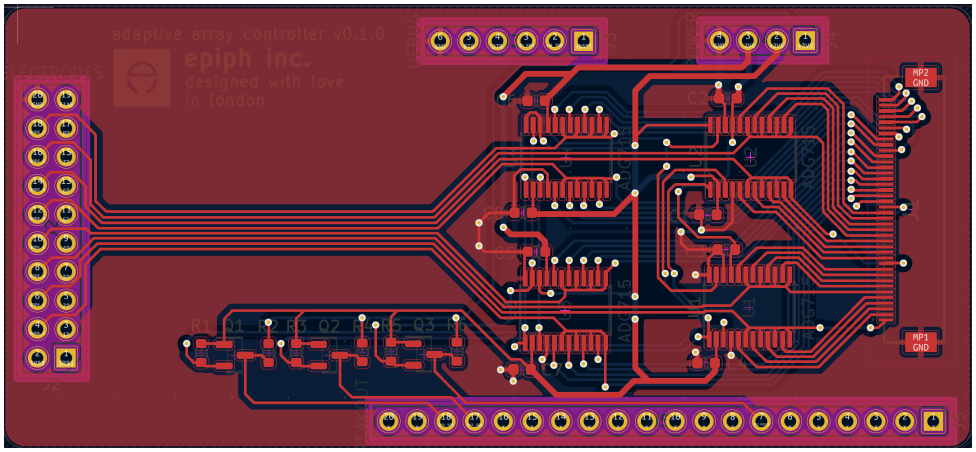
Shown on the following page.



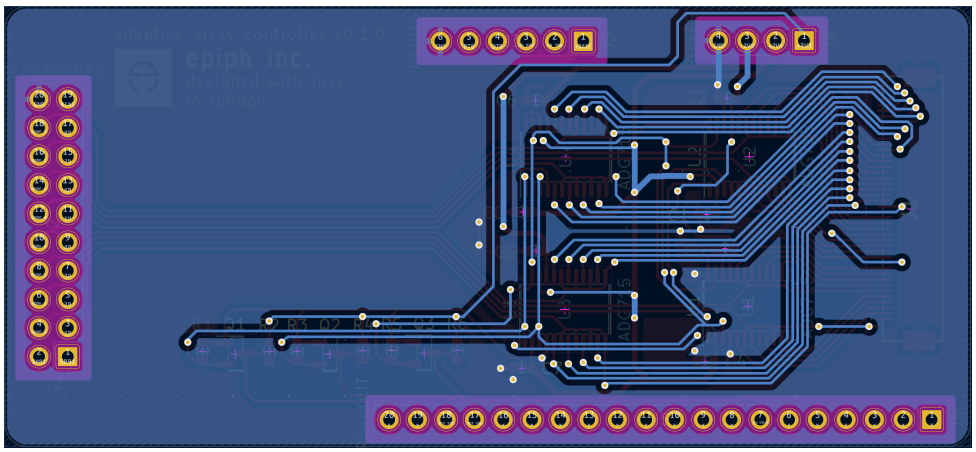
5.2 Layout



(a) Full layout



(b) Top layer



(c) Bottom layer

Figure 23: PCB layout for the control board